TOWARDS RELIABLE BRAIN TUMOR SEGMENTATION IN MRI NEUROIMAGING: INTEGRATING UNCERTAINTY ESTIMATION AND ENSEMBLE METHODS FOR CLINICAL APPLICATIONS

AGATA MOSINSKA

Thesis supervisor
ALFREDO VELLIDO ALCACENA (Department of Computer Science)

Thesis co-supervisor
ESTELA CAMARA MANCHA

Degree
Master's Degree in Artificial Intelligence

Master's thesis

School of Engineering
Universitat Rovira i Virgili (URV)

Faculty of Mathematics
Universitat de Barcelona (UB)

Barcelona School of Informatics (FIB)
Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

14/05/2025

Abstract

This thesis addresses uncertainty in automated brain tumor segmentation using deep learning. An ensemble of SwinUNETR, SegResNet, and Attention U-Net models was developed, incorporating test-time augmentation and test-time dropout for uncertainty estimation. Evaluation on the BraTS 2021 dataset showed that the ensemble achieved statistically significant improvements in segmentation accuracy, particularly for edema, relative to individual models. Uncertainty maps correlated with voxel-wise error and enabled risk-aware prediction. A Streamlit application was prototyped for clinical feasibility, allowing for interactive visualization of segmentation and uncertainty.

The developed application lays the theoretical and technical groundwork for future compliance studies in medical image segmentation, with the potential to accelerate regulatory approval and clinical integration. Further progress will require larger, more diverse models and prospective reader studies to translate these preliminary findings into clinically meaningful improvements.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem statement and motivation of the thesis

The brain is the most complex organ in the human body. It acts as a control center, managing sensory information, coordinating bodily functions, and enabling thoughts and emotions. However, these processes can be significantly disrupted by the presence of brain tumors leading to major impairments in daily lives [1]. Moreover, malignant brain tumors (both primary and other central nervous system tumors) have a five-year relative survival rate of approximately 35.7%, far worse than most other cancer types [2], underscoring the urgent need for effective diagnostic and treatment strategies. As a result, brain tumor analysis is a very active area of research attracting attention from both medical and technical communities.

Accurate segmentation of brain tumor volume is a critical component of both clinical care and neuroscience research. By precisely outlining the extent of tumor involvement, segmentation provides essential quantitative data that functions as a biomarker for surgical planning, tracking disease progression, evaluating treatment response, and comparing outcomes across clinical trials [3]. In clinical settings, volumetric data supports the monitoring of tumor growth, informs treatment decisions, and helps determine whether therapies are effectively managing the disease.

In research, segmented tumor volumes are equally important. They facilitate tumor classification and enable studies that explore how tumor size relates to cognitive decline, neurological function, and patient survival. This volumetric information allows researchers to uncover key

correlations between tumor burden and clinical outcomes, deepening our understanding of brain tumor pathology.

However, the heterogeneous nature of brain tumors with their varying sizes, shapes, and often indistinct boundaries, makes segmentation of tumor regions a tough and time-intensive job performed only by specially trained neuroradiologists [4, 5].

In recent years, scientists have been looking for automatic and robust methods for brain tumor segmentation. Particularly, deep learning methods have revolutionized the area of medical image segmentation yielding promising results [6]. Among these, the U-Net architecture stands out for its precise localization capabilities, making it a popular choice for segmenting brain tumors, despite its limitations in capturing global context essential for distinguishing tumor tissues from healthy brain regions [5]. Advanced models like V-Net extend U-Net to 3D processing, offering better spatial understanding for segmenting magnetic resonance imaging (MRI) scans [7]. However, these methods often require significant computational resources, making them less scalable for large datasets [8]. Meanwhile, more recent hybrid approaches, such as Swin UNETR, combine convolutional neural networks (CNNs) and transformers to integrate global and local features, achieving state-of-the-art results [9]. This variety of available methods reflects the complexity of brain tumor segmentation, highlighting that no single model is universally optimal, as each architecture offers unique strengths and trade-offs tailored to specific medical imaging challenges.

Despite the impressive performance of modern deep learning models, clinicians often find themselves not trusting the model predictions, especially in complex tasks like tumor segmentation [10, 11]. In neuro-oncology, the precise delineation of necrotic core, edema, enhancing tumor, and healthy tissue is often inherently ambiguous: boundaries blend gradually and have heterogeneous appearances that are hard to define with certainty. For algorithms to learn effective segmentation, they rely on expert-annotated ground truth delineations informed by clinical expertise and visual cues that are often subtle or disputed. Conversely, clinicians would greatly benefit from models that not only provide predictions but also communicate their level of certainty, particularly in ambiguous regions. Such uncertainty-aware outputs could serve as valuable decision-support tools, drawing attention to areas where human expertise is most

needed and where automated predictions are less reliable.

In general, concerns about AI reliability and transparency have led to increasing regulatory efforts, including the European Union's Artificial Intelligence Act (the "AI Act"), which came into force on August 1, 2024. It introduces a risk-based framework for AI deployment [12]. AI-driven medical imaging systems, classified as high-risk, must meet strict requirements to ensure safety, accountability, and trustworthiness.

The EU parliament recognized the importance of addressing the uncertainty in AI systems, particularly in healthcare. The EU's Scientific Foresight Unit emphasized that healthcare AI must estimate uncertainty to help clinicians assess prediction confidence [13]. By identifying high-risk predictions that require human oversight, uncertainty estimation aligns with the AI Act's risk mitigation framework. Furthermore, it enhances transparency and trust by clarifying the limits of AI systems, thereby reinforcing the Act's overarching principles of safety and reliability.

In such context, this thesis aims to develop an uncertainty-aware ensemble approach for brain tumor segmentation, incorporating uncertainty estimation to improve the reliability and interpretability of automated segmentation methods in clinical practice. Despite the strong performance of modern deep learning architectures, their lack of uncertainty measures remains a major barrier to adoption in clinical settings. This thesis bridges the gap between state-of-the-art segmentation methods and their clinical applicability by developing an ensemble-based segmentation method that integrates uncertainty quantification. By leveraging the strengths of different model architectures through an ensemble approach and accounting for predictive uncertainty, this project aims to enhance segmentation reliability, provide uncertainty estimates for clinical decision-making, and align AI-based medical imaging analysis with regulatory standards.

1.2 Structure of the thesis

Chapter 1 — Introduction states the problem of uncertainty-aware glioma segmentation, places it in the context of the EU AI Act, and outlines the remainder of the document.

Chapter 2 — Clinical Context reviews the biology of brain tumors, MRI fundamentals, and

the BraTS benchmark, thereby establishing the clinical motivation and data source used throughout the study.

Chapter 3 — State of the Art in Brain-tumor Segmentation surveys segmentation techniques from classical methods to modern hybrid CNN–Transformer models, then covers ensemble learning and uncertainty quantification strategies relevant to this work.

Chapter 4 — Research Gaps & Thesis Objectives identifies limitations in current literature—such as inconsistent sub-region performance and limited use of voxel-wise uncertainty—and formulates the specific objectives addressed by the thesis.

Chapter 5 — Materials and Methods details the BraTS-2021 dataset, preprocessing pipeline, four base architectures, data augmentation, loss functions, hyper-parameter tuning, uncertainty estimation techniques, ensemble fusion strategy, evaluation metrics, and the design of an interactive Streamlit application.

Chapter 6 — Experiments and Results reports cross-validation and test-set performance of individual models and ensembles, calibration analyses, uncertainty–error correlations, out-of-distribution tests, and timing results for the prototype application.

Chapter 7 — Discussion interprets the empirical findings, contrasts them with prior work, analyses the marginal gains from ensembling and uncertainty estimation, discusses study limitations, and sketches directions for future research and clinical validation.

Chapter 8 — Sustainability Analysis and Ethical Implications analyzes the project from the environmental, economic, social, and ethical perspectives, giving a closer examination of the broader impact of this research.

Chapter 9 — Conclusions distils the main contributions, reflects on how the objectives were met, and summarises the broader clinical and regulatory implications of voxel-wise uncertainty for high-risk medical AI.

Appendices provide auxillary material, including job scripts, extended hyper-parameter tables, and additional quantitative results.

Chapter 2

Clinical Context

The objective of this chapter is to situate the thesis in a clinical context and outline the motivations behind the methods proposed later in the project. It begins with an overview of brain tumors highlighting their clinical significance and associated challenges. This is followed by a detailed discussion of magnetic resonance imaging (MRI), a key imaging method in neuro-oncology, underscoring its role in brain tumor diagnosis and segmentation. Finally, the chapter introduces the BraTS Challenge which is the source of the dataset used in this project.

2.1 Brain tumors

Cancer is a group of diseases characterized by abnormal and uncontrollable cell growth that can invade surrounding tissues and spread to distant parts of the body [14]. Among these, brain tumors represent a particularly challenging subtype, originating in the brain, surrounded by a rigid, bony skull, and producing significant neurological symptoms before treatment or cure [15]. These tumors represent a significant clinical challenge due to their complexity, aggressiveness, and limited treatment options which are often very invasive.

2.1.1 Types of brain tumors

Brain tumors include several types of primary and secondary tumors, differing in origin, behavior, diagnostic, and prognosis. Most brain tumors are secondary, meaning they originate in other

parts of the body and spread to the brain, a process known as metastasis [16]. The most common types of primary brain tumors, i.e. those originating in the brain itself, are meningiomas and gliomas [16]. Meningiomas arise in the meninges, which are membranes protecting the brain and spinal cord. They are usually slow-growing and have a favorable patient survival prognosis [16]. Gliomas, in contrast, originate in glial cells, which support and regulate neuronal function, and are often more aggressive and clinically challenging compared to meningiomas [17].

Gliomas account for almost 80% of all malignant tumors in adults, presenting a big challenge in treatment due to their potential aggressiveness as well as high recurrence rates [18]. This thesis will focus only on gliomas, hence, the following sections will provide a more detailed discussion of this tumor type.

2.1.2 Glioma sub-regions

Glioma tumors show significant histological heterogeneity. Its cells are partitioned into several sub-regions that reflect different tumor behaviors and responses to treatment [19]:

- **Necrotic core (NCR)**: The area of cellular death inside tumors. It appears due to the inability of the blood vessels to adequately supply oxygen and nutrients to the tumor's growing interior [20].
- **Edema (ED)**: The region surrounding the tumor. It occurs when plasma leaks into functional brain tissue due to impaired brain blood vessels [21].
- **Enhancing tumor (ET)**: This sub-region contains active and aggressive tumor cells, making it a focal point for medical treatments. It represents the areas where the tumor is most likely to grow and invade surrounding tissues [22].

These sub-regions are depicted in Figure 2.1, which visualizes their spatial distribution in a glioma tumor using T1-weighted contrast-enhanced (T1CE) MRI.

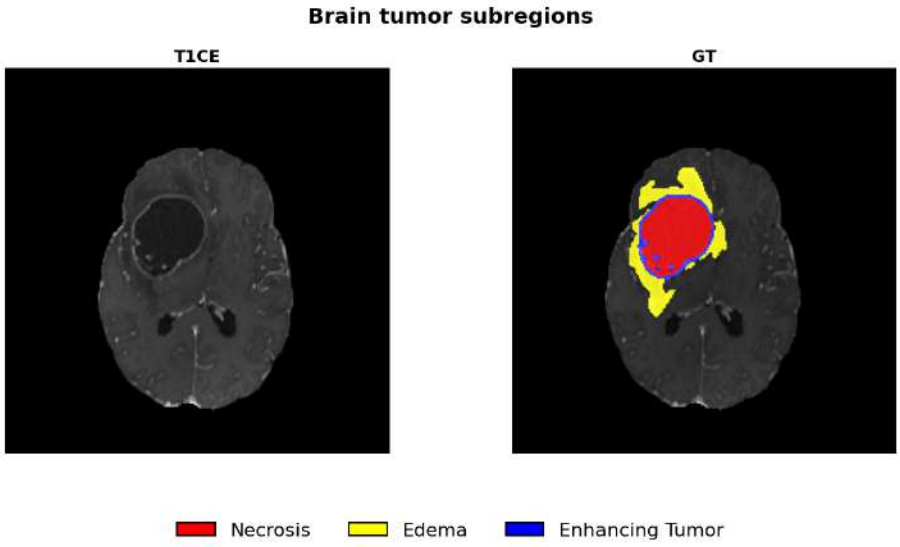


Figure 2.1: Visualization of glioma tumor sub-regions in T1CE MRI. The example scan comes from the BraTS 2021 Adult Glioma Challenge dataset.

2.1.3 WHO grading of gliomas

To classify gliomas and guide treatment decisions, the World Health Organization (WHO) established a grading system based on histological characteristics and clinical behavior of tumors [23, 24]. This system currently divides gliomas into four grades, with Grades I and II considered low-grade gliomas (LGG) and Grades III and IV classified as high-grade gliomas (HGG). Table 2.1 summarizes the characteristics, possible treatments, and disease progression associated with each of these grades.

Table 2.1: WHO classification of brain tumor grades [23, 24].

	Grade	**Characteristics**	**Treatment/disease progression**
Low grade	I	- Low proliferation - Non-infiltrative	- Can be cured via resection alone
	II	- Low proliferation - Slightly infiltrative	- Relatively slow growing - Can recur as higher grade
High grade	III	- Malignant - Infiltrative - Nuclei of tumor cells look abnormal - Rapid growth	- Patients are often treated with additional radiation therapy and/or chemotherapy
	IV	- Most malignant - Widely infiltrative - Rapidly growing - Necrosis-prone	- The disease often progresses rapidly before and after surgery, typically leading to a fatal outcome

Overall, LGGs tend to grow slowly, often lack necrosis, and have a better prognosis, whereas HGGs are aggressive, rapidly growing, and exhibit pronounced necrosis and enhancement. These differences between LGG and HGG, as illustrated in Figure 2.2, pose unique challenges for tumor segmentation, as LGGs often have indistinct boundaries, while HGGs display greater heterogeneity. Moreover, HGGs are significantly more common than LGGs [25], further underscoring the need to develop robust segmentation methods that are tailored to different types of glioma.

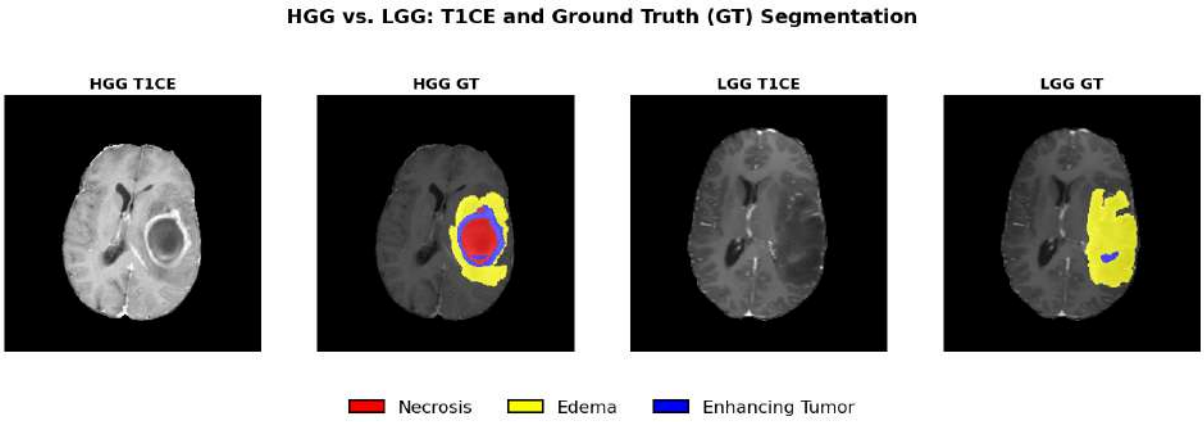


Figure 2.2: Comparison of HGG and LGG using the T1CE modality. The left column shows the raw T1CE scans and the right column shows the corresponding ground truth (GT) segmentation overlays. In HGG, the tumor exhibits a well-defined enhancing region (blue), a necrotic core (red), and edema (yellow), demonstrating its aggressive and heterogeneous nature. In contrast, LGG appears more homogeneous and exhibits minimal enhancement, no necrosis, and moderate edema, indicating slower progression. These differences highlight the challenges in distinguishing glioma sub-regions and delineating boundaries, particularly in LGG. The example scans come from the BraTS 2021 Adult Glioma Challenge dataset.

As described above, glioma tumors are highly heterogeneous and complex, and their accurate diagnosis and effective treatment remain challenging. Advanced neuroimaging techniques, particularly MRI, have become essential tools in neuro-oncology. MRI visualizes the structure of the tumor and can provide insights into the disease's progression and present plausible therapy options. The ability to segment tumor sub-regions accurately using MRI scans is pivotal for guiding surgical resections, planning radiotherapy, and monitoring treatment effects. The following section will further elaborate on the applications of MRI in neuro-oncology, focusing on its role in brain tumor segmentation.

2.2 The role of MRI in neuro-oncology

Medical images are extensively used in oncology, from diagnosis and treatment planning to monitoring patient outcomes. Neuro-oncologists use medical imaging to locate tumors and analyze their characteristics, such as their size or the presence of different tumor sub-regions.

2.2.1 How MRI works

MRI relies on strong magnetic fields and radiofrequency pulses to generate detailed images of the brain [23, 26]. During the procedure, the patient is positioned inside a strong magnet, where hydrogen nuclei present in the water molecules of the body align with the magnetic field. Then, radiofrequency pulses are applied, causing the nuclei to emit signals that are detected by the MRI scanner and reconstructed into cross-sectional images [26].

Figure 2.3 provides a schematic view of an MRI scanner, showing its main components, including the magnet, gradient coils, and radiofrequency coil, as well as a real-world example of an MRI machine.

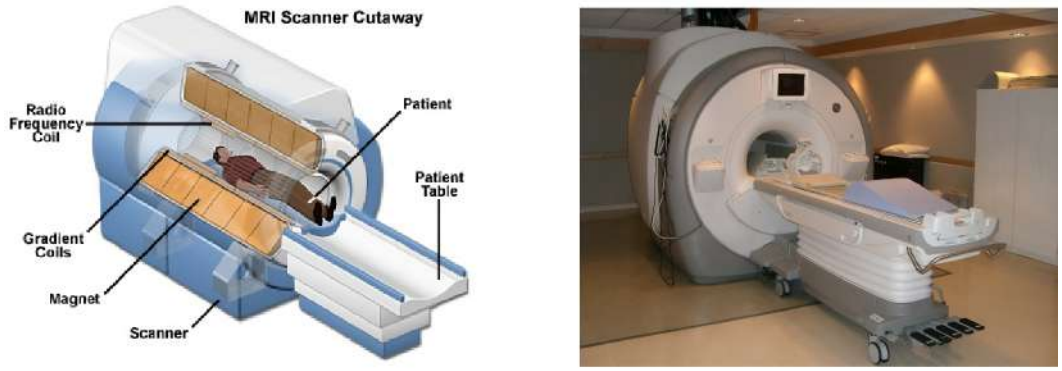


Figure 2.3: Schematic diagram (left) and real-world example (right) of an MRI scanner. Source: [27].

2.2.2 Advantages of MRI in neuro-oncology

MRI is currently the leading imaging modality in patients with primary brain tumors [28]. It is a non-invasive imaging technique relying on the magnetic spin properties of the hydrogen nuclei, abundant in the human body, to produce high-resolution images [23]. In contrast to

imaging techniques such as computed tomography (CT) and positron emission tomography (PET), MRI does not use ionizing radiation, making it safer for repeated use [29] - this is particularly important in neuro-oncology where repeated scanning is needed for monitoring tumor progression and response to treatment.

Additionally, MRI provides superior contrast between different soft tissues, making it ideal for visualizing complex brain structures and distinguishing healthy brain tissue from the different tumor sub-regions (such as NCR, ED, and ET) - crucial for accurate diagnosis and treatment planning [30].

2.2.3 Common MRI modalities in neuro-oncology

In clinical MRI settings, several sequences, often called MRI modalities, are usually acquired. MRI modalities refer to the different MRI sequence acquisition settings used to enhance particular tissues in the brain [31]. For brain tumor patients, the following MRI modalities are typically acquired:

- **T1-weighted (T1)**: It is a standard MRI sequence that highlights anatomical structures in the brain and provides contrast between different brain tissues such as white matter, gray matter, and cerebrospinal fluid (CSF). It is useful for analyzing overall brain anatomy and identifying gross tumor boundaries [32].
- **Post-contrast T1-weighted (T1CE)**: It is a T1-weighted MRI sequence that involves the injection of a gadolinium-based contrast agent, highlighting the areas with a disrupted blood-brain barrier function. T1CE is particularly useful for delineating the ET region [33].
- **T2-weighted (T2)**: It is another standard MRI sequence. It emphasizes the fluid-rich regions allowing for better distinction of healthy vs. tumor-affected brain tissues, particularly the fluid-related changes such as ED [34].
- **T2 Fluid Attenuated Inversion Recovery (FLAIR)**: It is another T2 MRI sequence that suppresses the CSF signal and minimizes contrast between white matter and gray

matter. In effect, it enhances the visibility of the lesions, particularly near ventricles or in periventricular areas [35].

Each modality provides unique information about the brain anatomy and the lesion, aiding in an accurate diagnosis and delineation of tumors. Figure 2.4 shows the 2D MRI views of a patient with a brain tumor, illustrating the differences between the modalities.

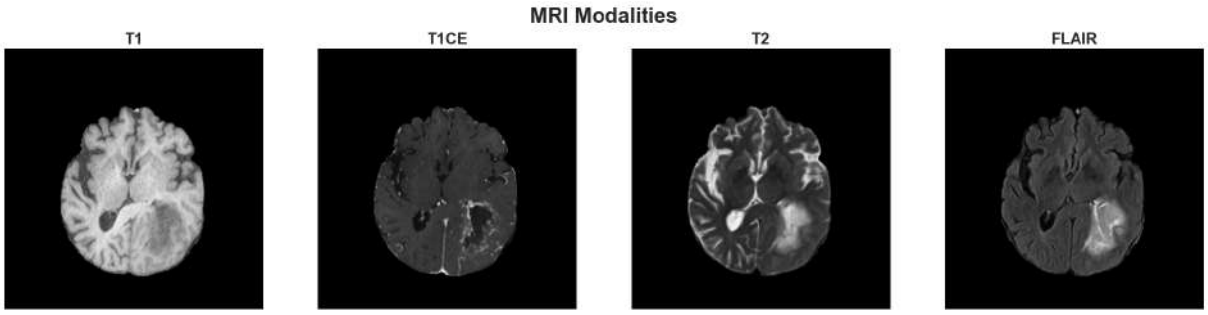


Figure 2.4: Comparison of MRI modalities (T1, T1CE, T2, FLAIR) for the same brain slice. In T1, white matter appears brighter than gray matter. T1CE highlights enhancing regions with bright intensities near the ventricles. T2 shows fluid-rich areas like edema as hyperintense (bright), while FLAIR suppresses CSF signals to make periventricular lesions more visible. The scans come from the BraTS 2021 Adult Glioma Challenge dataset.

2.2.4 Brain tumor segmentation using MRI scans

Neurologists and radiologists rely on MRI to annotate the brain tumor boundaries and delineate its specific regions - a process crucial for the preparation for resection surgery and further treatment planning. Nevertheless, despite the high resolution of the MRI images, manual segmentation of MRI sequences poses several challenges:

- **Tumor heterogeneity**: As mentioned in Section 2.1.3, brain tumors demonstrate significant heterogeneity in terms of shape, size, and the presence of different sub-regions. This variability makes it difficult to establish clear boundaries between tumor sub-regions.
- **Multi-modality analysis**: To effectively annotate all tumor sub-regions, the neurologist needs to review all four MRI modalities - T1, T1CE, T2, and FLAIR - as each modality provides useful information about different tissue characteristics. This further increases the complexity of the task.

- **Artifacts and noise**: MRI scans are prone to motion artifacts, partial volume effects, and signal noise, which can obscure critical details and further complicate the annotation process.
- **Time-intensity**: Annotation of MRI scans is a time-consuming process as it requires significant focus and attention to detail.

To address these challenges, automated segmentation methods are being developed using advanced machine learning techniques [36]. These methods aim to reduce the burden on clinicians while improving accuracy and reproducibility. However, creating and validating these algorithms requires high-quality annotated datasets, which are often limited due to the challenges mentioned above.

In recent years, efforts have been made to curate and make publicly-available datasets to help develop automated methods for tumor segmentation. There are several initiatives in this domain that aim to establish a benchmarking framework for glioma segmentation, providing researchers with a common dataset and evaluation methodology.

2.3 The BraTS challenge

One of the most widely recognized initiatives for benchmarking brain tumor segmentation methods is the Brain Tumor Segmentation (BraTS) challenge. It is an annual competition hosted by the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), designed to advance automated brain tumor segmentation techniques by providing a standardized dataset and evaluation framework for comparing state-of-the-art segmentation models [37].

Since the beginning of the conference in 2012, BraTS has played a pivotal role in developing and validating deep learning-based segmentation methods. By providing expert-annotated multimodal MRI scans, consistent preprocessing pipelines, and standardized evaluation metrics, the challenge allows researchers to benchmark their models on a common dataset. This way, BraTS has driven progress in the development of automated tumor segmentation methods.

Initially centered on adult gliomas, the BraTS challenge has expanded over the years to include multiple tasks, encompassing a variety of tumor types such as pediatric tumors, meningiomas, and metastases. Moreover, recent editions have introduced lesion-wise evaluation metrics to improve clinical relevance [37].

This thesis utilizes the BraTS 2021 Adult Glioma dataset, which contains cases of both high-grade and low-grade glioma patients annotated by expert neuroradiologists, providing a robust foundation for the development of automated segmentation techniques.

2.3.1 The BraTS 2021 Adult Glioma challenge

Among the various BraTS tasks, adult glioma segmentation has been the core focus of the challenge since its inception. This task involves identifying different glioma sub-regions - NCR, ED, and ET - from multimodal MRI scans.

The BraTS 2021 Adult Glioma dataset comprises MRI scans of patients gathered across different institutions in the United States of America. The number of patients in the training set grew over the years until the 2021 edition when it reached 1251 patients. In the 2022 and 2023 editions, the dataset has been kept the same, while the 2024 edition introduced post-operative cases, including patients with resected tumors. This extension aims to address the additional challenges of surgical cavity segmentation and post-treatment changes.

Despite these advancements, this study focuses on pre-operative glioma segmentation, making BraTS 2021 the most appropriate dataset for this research. A more detailed description and discussion of the dataset and its preprocessing can be found in Chapter 4 (Methodology), Section 5.1.

Chapter 3

State-of-the-Art in Brain Tumor Segmentation

This chapter aims to provide an overview of the current state-of-the-art in brain tumor segmentation, reviewing the existing literature on this topic. It begins with an overview of segmentation methods, outlining their evolution from conventional and machine learning-based approaches to modern deep learning architectures. This is followed by a discussion of state-of-the-art deep learning models, highlighting their strengths and limitations. Finally, the chapter explores recent developments in ensemble learning and uncertainty quantification, which are key areas relevant to this thesis.

3.1 Evolution of segmentation techniques

Segmentation methods for brain MRI can be divided into three categories: conventional segmentation methods, traditional machine learning-based methods, and deep learning-based methods. Each category includes a range of techniques, evolving from simple intensity-based thresholding to state-of-the-art deep learning architectures. Figure 3.1 provides an overview of these segmentation approaches and their key algorithms.

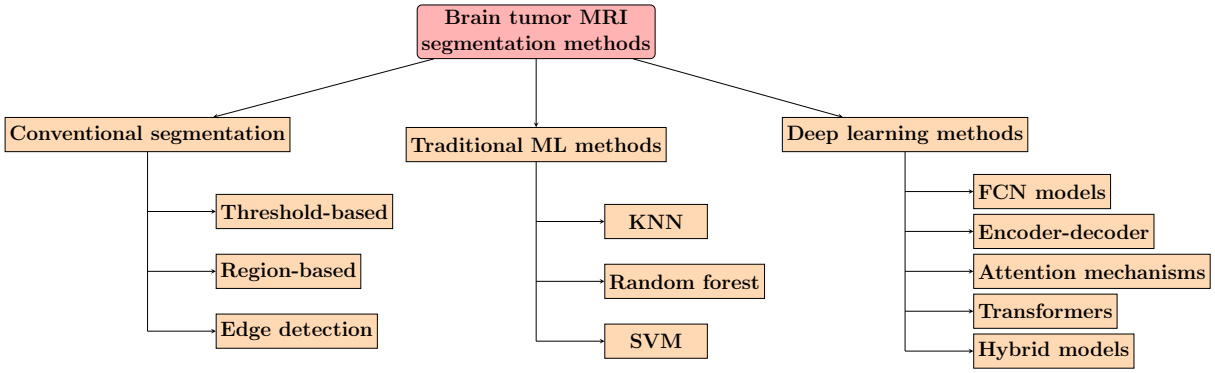


Figure 3.1: Overview of brain tumor MRI segmentation approaches.

Conventional methods, such as threshold-based, region-based, and edge and contour-based methods, laid a foundation for automated brain tumor segmentation by introducing approaches that were simple, interpretable, and computationally efficient. Nevertheless, the reliance of these methods on predefined thresholds or local pixel relationships limits their ability to handle complex tumor characteristics [38]. For instance, threshold-based methods struggle with segmenting tumors showing heterogenous intensity, while region-based and edge and contour-based methods are sensitive to noise and over-segmentation [38]. Although these methods provided early insights into brain tumor segmentation, they were not sophisticated enough to capture more complex tumor morphologies.

Traditional machine learning methods including, among others, K-Nearest Neighbors (KNN), Random Forests (RF), and Support Vector Machines (SVM), introduced statistical learning to improve segmentation accuracy. These methods rely on handcrafted features such as intensity, texture, and gradients to train models for tumor classification. However, although traditional machine learning approaches outperform the performance of conventional methods, they still come with certain shortcomings. In particular, their heavy reliance on manually engineered features and parameter tuning limits their ability to capture complex, hierarchical patterns, limiting their generalizability to large, diverse datasets [39].

The limitations of conventional and traditional machine learning-based methods paved the way for the use of deep learning methods. By automatically extracting complex hierarchical features directly from raw MRI scans, deep learning methods address the challenges of their predecessors, achieving state-of-the-art performance in brain tumor segmentation.

3.2 State-of-the-art deep learning architectures

Deep learning has revolutionized the field of brain tumor segmentation. As previously mentioned, traditional machine learning methods rely heavily on handcrafted features, requiring expert medical knowledge to extract meaningful representations from MRI scans [39]. In contrast, deep learning models learn hierarchical feature representations directly from raw images, enabling them to capture complex tumor characteristics and improve segmentation accuracy [40].

Various deep learning architectures have been proposed to improve segmentation performance, each addressing different challenges in feature extraction, spatial awareness, and computational efficiency. This section explores the state-of-the-art deep learning architectures that have significantly advanced brain tumor segmentation.

3.2.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are one of the most common deep learning architectures for working with image data. Their ability to automatically extract hierarchical features from raw image data has made them highly effective in tasks such as object detection, image classification, and segmentation. In the field of medical imaging, CNNs have been extensively applied to analyze MRI scans, aiding in disease detection and segmentation tasks, including brain tumor segmentation [38].

CNNs utilize convolutional filters to extract spatial features at multiple levels, enabling them to learn both low-level patterns (e.g., edges and textures) and high-level abstractions (e.g., tumor structures) from medical images [41]. To progressively reduce the spatial dimensions while retaining essential information, pooling layers are employed after convolutions [41]. These layers play an important role in downsampling, preventing overfitting, and increasing the receptive field. This process is illustrated in Figure 3.2, where a CNN architecture processes an MRI scan, progressively extracting features through convolutional and pooling layers before classifying the input as either benign or tumor. Additionally, CNNs employ the weight-sharing mechanism which reduces the number of trainable parameters, making them computationally efficient while maintaining solid feature extraction capabilities [38]. These properties have led to the widespread

adoption of CNN-based architectures in brain tumor segmentation, with models such as U-Net, V-Net, and SegResNet achieving state-of-the-art performance.

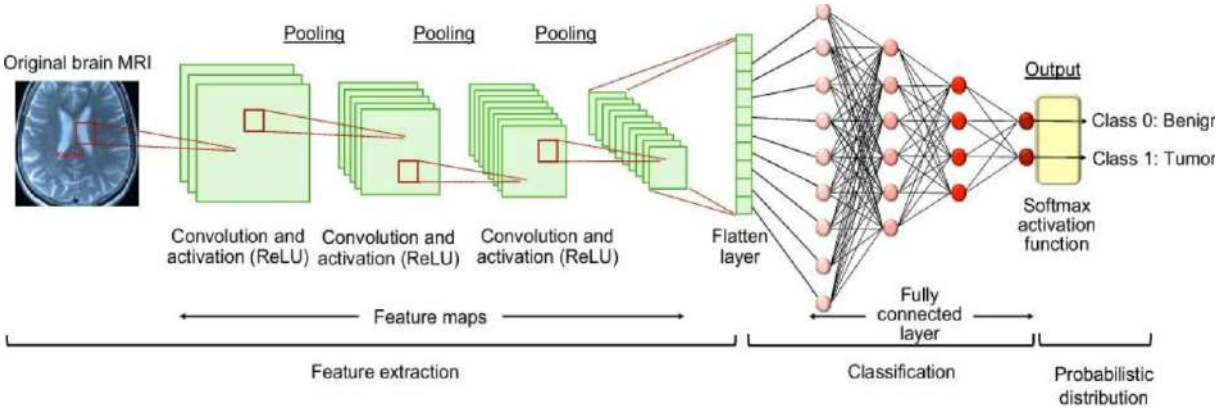


Figure 3.2: Example of a CNN architecture applied to brain tumor classification. The model extracts hierarchical features from an MRI scan through convolutional and pooling layers before passing the extracted features to a fully connected layer for classification. The final output is a probabilistic distribution over two classes: benign and tumor. Source: [42]

In the BraTS challenge, CNNs were first introduced in 2013 through the work of Zikic et al. [43] and Urban et al. [44], who developed deep convolutional neural networks (DCNNs) tailored specifically for brain tumor segmentation, achieving promising results. However, as the depth of DCNNs increased, challenges such as gradient explosion and gradient vanishing appeared during the training process, hindering the performance and convergence of deeper networks.

In order to solve the problem of network degradation, He et al. proposed a deep Residual Network (ResNet) [45], which incorporated residual connections within convolutional blocks. These connections enable the network to effectively learn identity mappings, facilitating the training of much deeper architectures by mitigating gradient-related issues. A residual block from a ResNet architecture is presented in Figure 3.3. ResNet was first used in the BraTS challenge in 2017 by Beers et al. [46], who employed a sequential 3D U-Net framework for biologically-informed tumor segmentation, and Pawar et al. [47], who utilized a cascaded anisotropic CNN approach for hierarchical segmentation of tumor sub-regions.

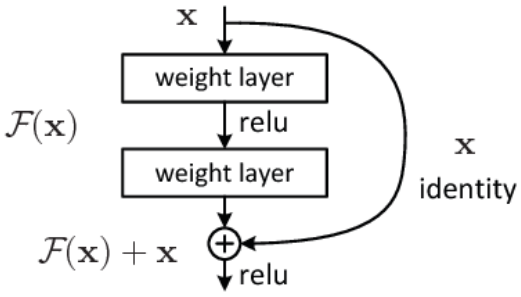


Figure 3.3: A building block with residual connections in ResNet. The identity mapping (x) is added to the output of the residual block ($F(x)$), allowing deeper networks to mitigate gradient-related issues. Source: [45].

While ResNet addressed the challenges of training deep networks, another significant advancement in image segmentation was the introduction of Fully Convolutional Neural Networks (FCNNs), which consist exclusively of convolutional layers and exclude fully connected layers [48]. This architectural design significantly reduces the number of parameters in the network, enhancing computational efficiency. Moreover, unlike traditional CNNs, FCNNs have no fixed input image size requirements and include an upsampling process in the final convolutional layers. This upsampling ensures that the output matches the spatial dimensions of the input image, enabling pixel-wise predictions while preserving spatial information [38]. As a result, FCNNs are highly suitable for pixel- or voxel-level semantic image segmentation, making them suitable for medical image segmentation, where fine-grained and spatially accurate classifications are key.

The best known FCNN is the U-Net, characterized by its fully convolutional encoder-decoder architecture, forming a distinct U-shape, as depicted in Figure 3.4. The encoder is responsible for capturing contextual information by progressively downsampling the input, while the decoder restores spatial details through upsampling [49]. Both modules are interconnected by skip connections that allow the network to merge fine-grained details extracted by the encoder with high-level contextual information from the decoder, improving segmentation accuracy [49].

U-Net's skip connections and encoder-decoder design allow for effective feature fusion, making it suitable for tasks like brain tumor segmentation which require dealing with heterogeneous tumor sub-regions [49]. Indeed, the U-Net architecture has achieved great results in the BraTS challenge. All top-performing participants in BraTS 2019 and 2020 challenges used some vari-

ant of the U-Net architecture [50]. Building on the success of the U-Net, Milletari et al. [51] introduced a volumetric U-Net (V-Net) which adapted the U-Net architecture to process 3D volumetric data, such as the MRI scans used in the BraTS dataset. Unlike 2D U-Net, V-Net processes entire 3D volumes, capturing spatial context between slices, which are critical in the case of segmenting larger and more irregular brain tumors.

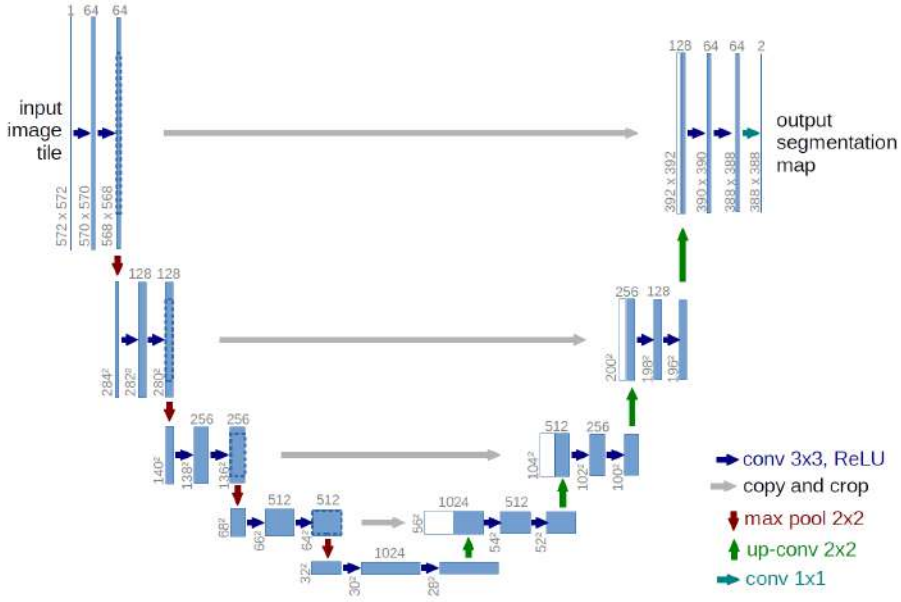


Figure 3.4: The U-Net architecture (illustrated for a 32x32 pixel resolution at the lowest level) is depicted using blue boxes, each representing a multi-channel feature map. The number of channels is indicated above each box, while the dimensions (x-y size) are specified at the bottom-left corner. White boxes correspond to copied feature maps, and the arrows illustrate the various operations performed within the network. Source: [49]

Overall, CNNs have significantly advanced the field of tumor segmentation, providing efficient architectures capable of learning hierarchical features from raw medical images. CNN-based models have achieved state-of-the-art performance in the BraTS challenge, regularly placing in the top places of the competition.

Nevertheless, CNN-based architectures also face certain limitations. Specifically, they often struggle to capture the global context and focus on critical tumor regions, particularly in cases with indistinct boundaries or heterogeneous tumor structures [52]. Moreover, CNN architectures often suffer from domain shift problems, meaning that they perform significantly worse when exposed to data differing from the training dataset [42]. These challenges have led to the

introduction of attention mechanisms, which aim to further enhance segmentation accuracy by enabling networks to focus on the most critical features. The next section will explore the role of attention mechanisms in advancing brain tumor segmentation.

3.2.2 Attention mechanisms

Attention mechanisms in deep neural networks were first introduced by Bahdanau et al. [53] to improve sequence-to-sequence natural language processing. Inspired by human visual attention, these mechanisms enable models to focus on the most relevant parts of an input, ignoring its less relevant parts. They dynamically weigh features, improving the model's ability to capture important details [53].

In medical image segmentation, attention is particularly useful for improving the delineation of tumor regions by selectively highlighting relevant spatial and contextual features, leading to more precise and reliable predictions (see Figure 3.5) [6, 54]. Many segmentation models in the area of medical image analysis incorporate attention modules in the form of squeeze-and-excitation (SE) blocks that capture channel-wise relationships [55] or by including attention gates in the U-Net architecture [56] to emphasize tumor-related information and suppress less relevant information.

In the BraTS challenge, several teams adapted attention mechanisms specifically for brain tumor segmentation. Yuan et al. [57] and Noori et al. [58] inserted attention layers into the encoder to enhance tumor-specific feature extraction, while Xu et al. [59] and Zhou et al. [60] designed cascaded attention mechanisms in the decoder to exploit correlations between different tumor sub-regions.

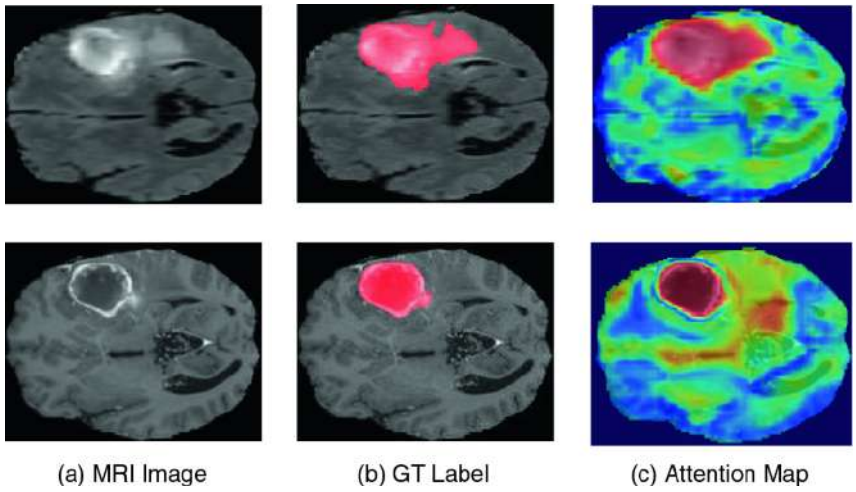


Figure 3.5: Illustration of the role of attention mechanisms in brain tumor segmentation. The first column shows a raw MRI scan, the second column highlights a ground truth with the segmentation in pink, and the last column displays an attention heatmap. Darker colors in the attention heatmap signify regions of higher attention, demonstrating how the model emphasizes tumor regions while ignoring less relevant structures. Source: [59]

With the rise of the Transformer architecture, attention mechanisms have evolved from being enhancements of CNN architectures to being a backbone of deep learning architectures. Transformers, originally introduced by Vaswani et al. [61] for NLP tasks, utilize self-attention to process input data holistically rather than sequentially or locally as in CNNs [61] (see Figure 3.6). Thus, transformers can better capture global contextual dependencies, which is crucial in medical image segmentation where tumors often have complex, irregular shapes that span multiple spatial regions [54].

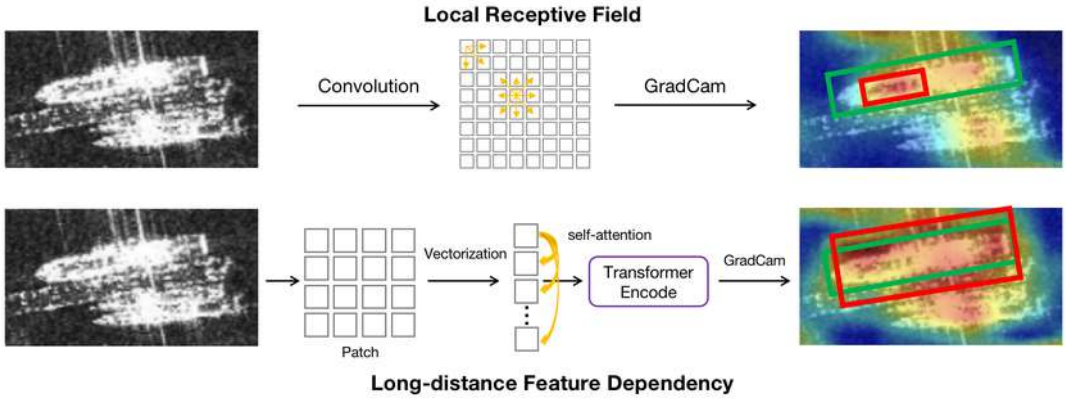


Figure 3.6: Comparison of CNNs and Transformers in processing feature dependencies. While CNNs rely on local receptive fields (top), Transformers use self-attention to capture long-range spatial dependencies (bottom). Source: [62]

Recent works have introduced fully transformer-based models for brain tumor segmentation. For example, the mmFormer model [63] is a multimodal medical Transformer designed to leverage both intra-modal self-attention and inter-modal attention mechanisms to capture correlations across multiple MRI modalities. Another example is the 3D Medical Axial Transformer (MAT), which utilizes axial attention mechanisms to process volumetric MRI data [64]. These models show that transformers can be successfully utilized in brain tumor segmentation.

Nevertheless, fully transformer-based architectures face some limitations in segmentation tasks. One significant bottleneck is the missing inductive bias of locality [54] which is inherently present in CNNs. Transformers lack this built-in prior knowledge which makes them less effective at capturing fine-grained local details like small tumor boundaries, particularly when training data is limited [54]. They also generally require larger datasets and additional computational resources for effective training compared to CNNs [42].

To address this, researchers have increasingly turned to hybrid architectures that combine the strengths of CNNs and transformers. The next section explores these hybrid architectures in more detail, focusing on their role in brain tumor segmentation.

3.2.3 Hybrid architectures

Hybrid architectures combine different approaches to leverage their strengths and suppress their respective bottlenecks. They have emerged as powerful solutions in various domains, including brain tumor segmentation, where they have shown promising results.

In brain tumor segmentation, hybrid architectures typically combine CNNs with other advanced techniques, such as transformers or recurrent networks, to enhance performance. The development of hybrid architectures is driven by the complexity of brain tumor segmentation tasks, the need for capturing both local and global contexts as well as improving computational efficiency.

For instance, transformer-based hybrid models, such as TransBTS [65] and Swin-UNETR [66], integrate convolutional and transformer-based encoders to balance local feature extraction with long-range dependency modeling. Figure 3.7 illustrates the structure of Swin-UNETR,

where hierarchical transformer layers (blue blocks) capture global dependencies while convolutional blocks (orange) refine spatial details. These architectures have been effective in integrating multimodal MRI sequences, allowing for better fusion of spatial and contextual features across different imaging modalities. Another notable hybrid model is H2NF-Net (Hybrid High-resolution and Non-local Feature Network) [67], which integrates high-resolution features with non-local dependencies to enhance segmentation accuracy. H2NF-Net demonstrated its effectiveness by achieving second place in the BraTS 2020 challenge segmentation task. Similarly, 3D-TransUNet [68] extends the popular TransUNet architecture into a 3D framework, effectively capturing both local details and global dependencies in volumetric MRI data. This architecture secured second place in the BraTS 2023 Metastasis challenge.

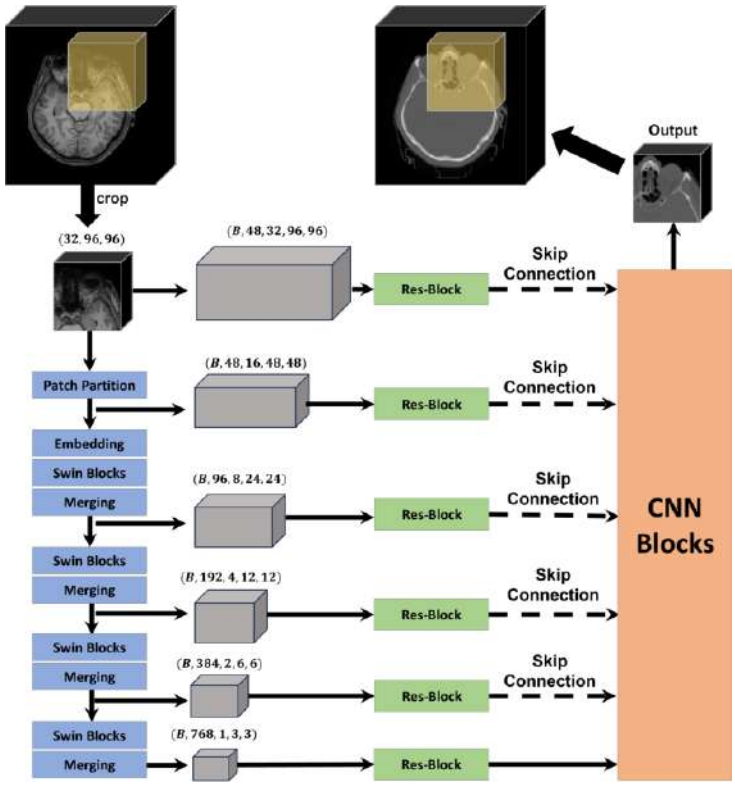


Figure 3.7: Example of hybrid architecture: the Swin-UNETR model. Source: [69]

Overall, hybrid models have demonstrated significant potential in advancing the performance on the brain tumor segmentation task. By combining the strengths of different deep learning architectures, they can effectively capture both local and global contexts, leading to more precise tumor delineation and state-of-the-art performance.

As the field continues to evolve, researchers have been exploring new ways to further enhance these hybrid models, for instance, through the use of ensemble learning techniques. While hybrid models leverage the strengths of different paradigms within a single model, ensembles take it a step further by combining the outputs of multiple models to reduce errors and account for variability in predictions.

3.3 Ensemble learning

Model ensembling is a widely recognized technique in machine learning. It involves training multiple models and then merging their outputs to produce a final segmentation. Figure 3.8 showcases a general framework for creating an ensemble prediction. By combining predictions from multiple models, the ensemble approach minimizes biases and variability of individual models, resulting in more reliable and precise segmentation results [70]. This section will look in more detail at the most commonly used ensemble techniques applied to brain tumor segmentation.

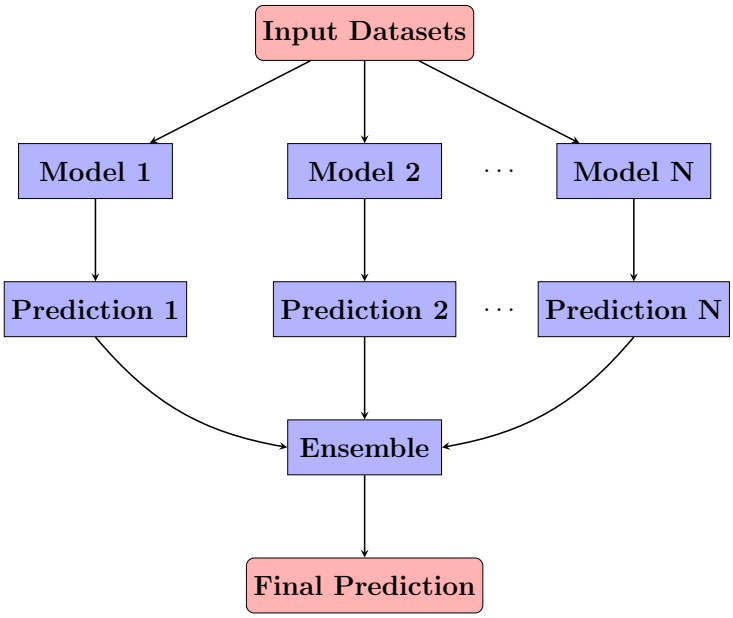


Figure 3.8: General framework for ensemble learning.

3.3.1 Averaging

The most widely used ensembling strategy is **averaging**. It works by taking the average of multiple segmentations to produce a final output. Despite its simplicity, this technique mitigates the impact of inconsistent model errors by smoothing predictions over multiple network instances [71].

A notable example of this approach is the BraTS 2023 Adult Glioma Challenge winners' model. They created an ensemble strategy of multiple nnU-Net models to improve segmentation. Several variations of the nnU-Net architecture were first trained with different initialization seeds and hyperparameters to introduce model diversity. Then, the outputs of these individual models were combined using model averaging to generate the final prediction, combining softmax outputs to reduce inconsistencies and improve robustness. This approach helped reducing variance, mitigating individual model biases, and enhancing generalization across tumor sub-regions [72].

Nevertheless, although averaging is computationally efficient and provides a strong baseline, more advanced methods exist, such as bagging and stacking, which improve on this concept by incorporating diverse models and optimizing their individual contributions.

3.3.2 Bagging and stacking

3.3.2.1 Bagging

In the bagging method, models are trained on different subsets of the data, reducing correlation between individual models and improving robustness (see Figure 3.9) [71]. It has been commonly used in random forests, but it has also been adapted for deep learning-based medical image segmentation. For instance, Shivhare et al. [73] introduced a framework called Tumor Bagging in which they applied the bagging method to combine several segmentation methods based on the Multilayer Perceptron (MLP), achieving a Dice Score of more than 92% for detection of the entire tumor.

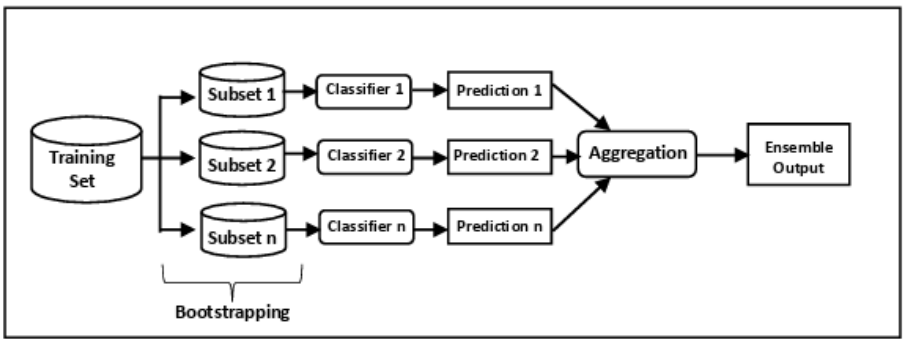


Figure 3.9: Bagging. Source: [74]

3.3.2.2 Stacking

Unlike bagging, stacking uses a meta-learner, such as logistic regression or a shallow artificial neural network, to combine predictions from multiple models and optimize the final segmentation (see Figure 3.10). This approach is particularly useful when models capture complementary features [74]. For example, a recent study proposed a stacking ensemble method (SELDenseNet201) combining DenseNet201 with six diverse base models, achieving a Dice coefficient of 97.43% for brain tumor segmentation [75]. However, a drawback of this approach is that it requires additional validation data to train the meta-learner.

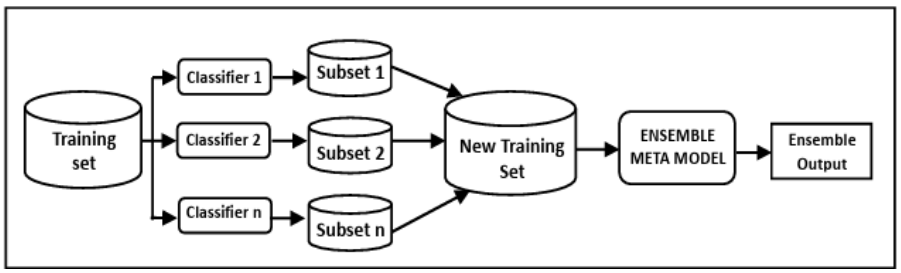


Figure 3.10: Stacking. Source: [74]

3.3.3 Simultaneous Truth and Performance Level Estimation (STAPLE)

Simultaneous Truth and Performance Level Estimation (**STAPLE**) takes ensemble learning further by assigning weights to models based on their sensitivity and specificity. Unlike averaging or stacking, STAPLE dynamically adjusts the contribution of each model, prioritizing those with higher accuracy [76]. By prioritizing models with higher accuracy, STAPLE ensures

an informed fusion of predictions. It has been particularly effective in medical image analysis when integrating outputs from diverse architectures, such as CNNs and vision transformers [72].

An example of a successful application of the STAPLE method was provided by Fadugba et al. [70]. They used the STAPLE method to ensemble three models: 3D U-Net, V-Net, and MSA-VNet. With this ensemble, they were able to generalize from the BraTS 2023 Adult Glioma dataset to the BraTS 2023 Sub-Saharan Africa dataset, achieving promising results.

To provide a general overview of the presented methods, Table 3.1 summarizes each method's approach, highlighting their goals and complexity, and offering suggestions for scenarios in which each method might prove useful.

Table 3.1: Comparison of Ensemble Methods for Brain Tumor Segmentation

Feature	**Method**	**Model Diversity**	**Goal**	**Complexity**
Averaging	Combines predictions from multiple models by taking their average	Can use similar or different models	Reduce variance and improve overall accuracy	Low
Bagging	Creates multiple subsets of training data, trains models on each subset, and averages predictions	Typically uses homogeneous models	Reduce variance and prevent overfitting	Medium
Stacking	Trains diverse base models and uses their predictions as input for a meta-learner	Uses heterogeneous models	Maximize accuracy by leveraging strengths of different models	High
STAPLE	Assigns weights to models based on sensitivity and specificity to combine their predictions	Can use models from different architectures	Improve segmentation accuracy by emphasizing reliable models	Medium

As demonstrated by the studies mentioned above, ensemble approaches have achieved promising results and have become more popular in the area. The reason for this is that ensemble models allow for addressing the limitations of individual models and improve generalizability, making them more reliable on unseen data. As deep learning continues to develop, ensemble models are likely to remain a key strategy for improving accuracy on tasks requiring high reliability, such as medical image segmentation.

3.4 Uncertainty estimation

3.4.1 Significance of uncertainty in medical image analysis

Despite high segmentation accuracy, current segmentation methods based on deep learning still fall short of the robustness required for clinical applications. This limitation arises from several factors, including the significant variability in MRI scans (e.g. due to using different machines and imaging protocols) as well as the inherent heterogeneity of brain tumors themselves [77].

To mitigate these issues, manual quality control of the generated segmentation is recommended before continuing with the analysis [78]. However, it has several shortcomings: it is time-consuming, susceptible to intra- and inter-variability, and typically applied globally to the entire scan [78]. These limitations highlight the need for more efficient and reliable methods for estimating segmentation quality.

Automatic uncertainty estimation provides a promising way to alleviate these issues while reliably quantifying segmentation performance, even on the voxel level. Unlike manual quality control, it provides a more localized and dynamic approach to evaluating model performance, highlighting regions where the model is more uncertain.

In the context of deep neural networks, uncertainty estimation (also referred to as uncertainty quantification) involves predicting the confidence level of model outputs, allowing for quantification of its reliability [79]. By identifying areas of high uncertainty, this approach allows clinicians to focus their attention on critical regions that may require human review and manual correction. This in turn enhances the decision-making process, reduces the risk of wrong predictions, and ensures that medical professionals are involved in critical cases [79].

Moreover, uncertainty estimation can improve the performance of the models (e.g. sensitivity and accuracy) [80]. Additionally, it also allows for better resource allocation as it ensures that uncertain cases go through expert review while confident predictions can be processed automatically. By creating uncertainty-aware models, we can accelerate the incorporation of deep learning-based segmentation systems into clinical practice. Ultimately, uncertainty estimation helps ensure that AI-driven medical imaging systems produce more reliable and transparent

results, bridging the gap between artificial intelligence and practical medical applications.

3.4.2 Types of uncertainty

In the context of deep learning models, uncertainty refers to the model's lack of confidence or ambiguity about determining the correct output for a specific input [79]. There are two main types of uncertainty in deep learning models, as illustrated in Figure 3.11, which are described below:

- **Aleatoric uncertainty (data uncertainty)**: It captures the inherent data uncertainty due to noise and variability in inputs, which cannot be reduced by collecting more data [77, 79]. Aleatoric uncertainty is inherent to the data and cannot be reduced by improving the model or collecting more data, as it stems from intrinsic noise in the imaging process [79].

  In the context of brain tumor segmentation, aleatoric uncertainty can emerge, for example, from the noise in MRI scans. For instance, low-contrast regions can make it difficult to differentiate between healthy tissue and tumor and motion artifacts can introduce blurriness, making segmentation more challenging. Additionally, insufficient spatial resolution or anisotropic voxels—where voxel dimensions differ across axes—can lead to partial volume effects, further reducing the reliability of tissue differentiation and exacerbating segmentation challenges [81].

- **Epistemic uncertainty (model uncertainty)**: This type of uncertainty captures model uncertainty due to lack of knowledge or insufficient model structure [79]. Unlike aleatoric uncertainty, epistemic uncertainty can be reduced by collecting more data, using more complex models, or incorporating regularization methods [79].

  In brain tumor segmentation, epistemic uncertainty can emerge in the segmentation of less common tumor sub-regions (e.g. necrotic core) or tumor types, as the model has seen fewer cases of these during training.

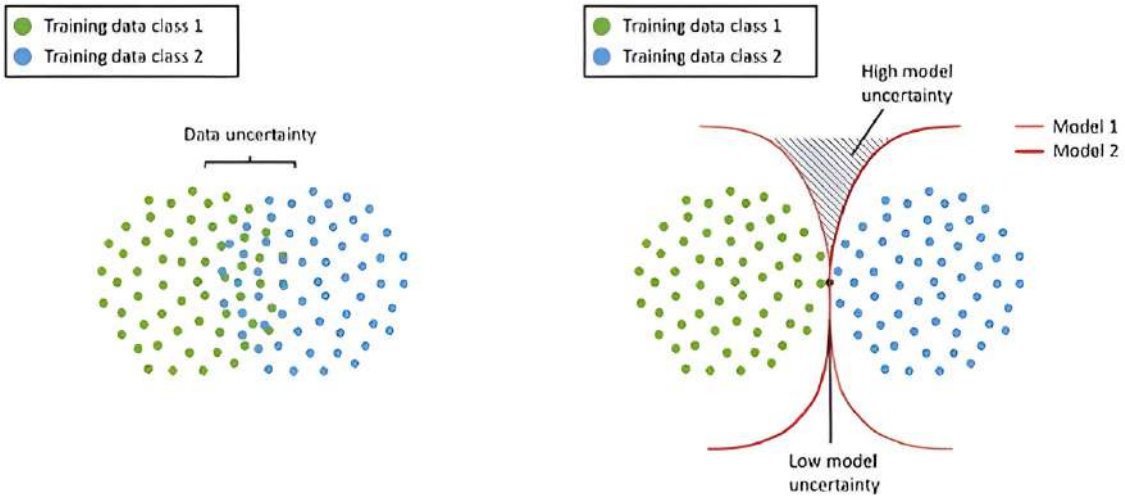


Figure 3.11: Aleatoric vs. epistemic uncertainty. Aleatoric uncertainty is shown as overlapping data regions (left panel), while epistemic uncertainty is represented by areas of high model disagreement or out-of-distribution samples (right panel). Source: [82]

3.4.3 Methods for quantifying uncertainty

There are several different approaches to estimating uncertainty in deep neural networks, broadly categorized into **deterministic** and **probabilistic** methods. Deterministic methods provide a single measure to represent the level of uncertainty from a single forward pass, making them computationally efficient, but often struggling to capture epistemic uncertainty [79, 83]. Meanwhile, probabilistic methods leverage multiple inferences or model ensembles to estimate aleatoric and epistemic uncertainty. The following subsections will outline several of the most commonly used techniques, including deterministic confidence estimation, Bayesian artificial neural networks, test-time dropout, test-time augmentation, and ensemble-based approaches.

3.4.3.1 Deterministic methods

Deterministic uncertainty estimation methods are widely used to estimate uncertainty in deep learning models due to their computational efficiency and simplicity [79]. These methods assume fixed model weights (non-Bayesian) and infer uncertainty from output probabilities or feature representations using a single forward pass [79]. Figure 3.12 illustrates this process.

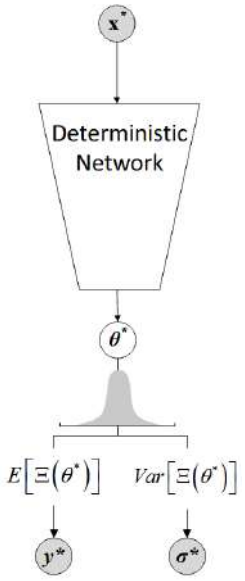


Figure 3.12: Visualization of uncertainty modeling in a deterministic approach. A deterministic network processes an input and derives uncertainty from the variance of predictions. Source: [79]

Below, three common deterministic methods are described.

- **Confidence-based approaches**: These methods estimate uncertainty directly from the model's output probabilities - typically from the softmax layer in segmentation models. A low maximum softmax score is taken as an indication of higher uncertainty.
- **Distance-based approaches**: These methods quantify uncertainty by measuring the distance between an input sample's feature representation and class prototypes or centroids. The underlying assumption is that inputs closer to known examples are predicted more reliably, whereas those further away are less certain.
  - **Example in medical image segmentation**: Judge et al. [84] proposed CRISP (ContRastive Image Segmentation for Uncertainty Prediction). CRISP uses a contrastive learning framework to create a joint latent space for images and their corresponding segmentation maps, where uncertainty is estimated by computing distances within this latent space, resulting in an interpretable and anatomically consistent uncertainty map. They showed the effectiveness of their method on several datasets designed for the segmentation of cardiac and lung diseases.

- **Evidential deep learning**: Evidential deep learning (EDL) models uncertainty by predicting distributions (e.g. Dirichlet distributions) over segmentation outputs instead of relying solely on categorical probabilities. This enables EDL to quantify both aleatoric uncertainty and epistemic uncertainty simultaneously.

  - **Example in medical image segmentation**: Zou et al. [85] introduced DEviS, which integrates evidential deep learning into standard segmentation architectures to improve calibration and robustness while offering efficient uncertainty estimation. DEviS leverages subjective logic theory to explicitly model both probability and uncertainty within a deterministic framework. Their method was evaluated on several medical image segmentation datasets, including the BraTS 2019 dataset.

The main advantage of deterministic methods is that they are simple and computationally efficient and thus practical for real-time applications [79]. However, despite their efficiency, deterministic methods often struggle with capturing epistemic uncertainty, which represents uncertainty due to model limitations or insufficient training data [83]. Moreover, the fact that they rely on a single opinion makes them very sensitive to the underlying network architecture, training procedure, and training data [82].

3.4.3.2 Bayesian Neural Networks

Bayesian Neural Networks (BNNs) are a class of probabilistic deep learning models that incorporate uncertainty by treating the model parameters as random variables rather than fixed values [79]. Unlike deterministic models that produce a single-point estimate, BNNs infer a distribution over possible outputs, enabling the explicit quantification of both aleatoric and epistemic uncertainty [79] and making them particularly valuable in medical image segmentation.

BNNs achieve uncertainty estimation by learning a posterior distribution over network parameters, which is generally intractable and requires approximation techniques such as **Variational Inference (VI)** or **Markov Chain Monte Carlo (MCMC)** [86]. A visualization of a BNN is provided in Figure 3.13, illustrating how multiple forward passes with different parameter samples lead to a predictive distribution, allowing uncertainty estimation.

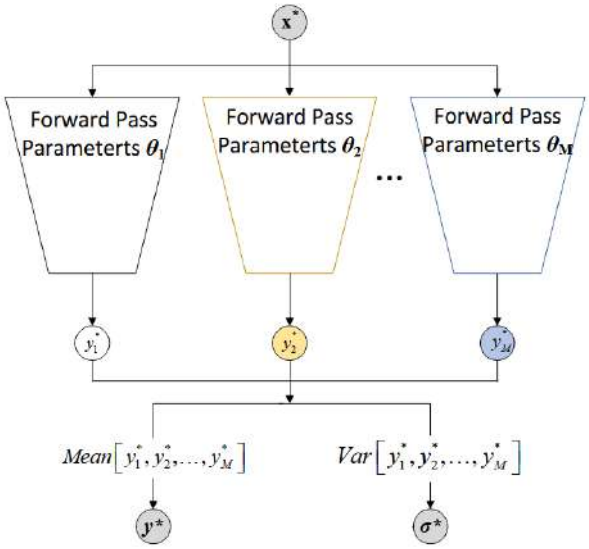


Figure 3.13: A visualization of the basic principles of uncertainty modeling in Bayesian Neural Networks. Source: [79]

VI approximates the posterior distribution by optimizing a simpler distribution to be close to the true posterior, while MCMC methods (e.g. Hamiltonian Monte Carlo) are used to sample from the posterior distribution of network parameters [86]. Although VI is a simpler approach, both methods are computationally expensive and often impractical in practice as they do not scale well with an increasing number of parameters or data points [87]. This limitation has led to the development of efficient approximations, such as **Monte Carlo Dropout** [88], which seek to bring Bayesian uncertainty estimation to more practical applications.

In recent years, several studies have successfully applied BNNs in medical image segmentation. The following are a few examples of such studies:

- Kwon et al. [89] developed a BNN-based framework for ischemic stroke lesion classification and segmentation, effectively quantifying both aleatoric and epistemic uncertainty.
- Hu et al. [90] proposed a BNN architecture designed to estimate inter-rater uncertainty in medical image segmentation. Their method models individual annotator variations, providing a detailed understanding of expert disagreement.
- Konathala [91] introduced BA U-Net, an uncertainty-aware segmentation model that integrates BNNs with attention mechanisms. This approach enhances both segmentation

accuracy and interpretability, demonstrating its effectiveness on MRI scans from the BraTS dataset.

The main advantage of BNNs is that, especially in comparison to deterministic methods, they excel at capturing epistemic uncertainty, which is crucial for detecting out-of-distribution inputs or cases with insufficient training data. However, their adoption in real-world applications is rather limited due to their higher computational cost compared to deterministic approaches.

Recent advancements in Bayesian approximations - such as Monte Carlo Dropout and other scalable alternatives - seek to make Bayesian uncertainty estimation more feasible for real-world applications.

3.4.3.3 Test-time dropout

As discussed in the previous section, exact Bayesian inference is an intractable problem for deep neural networks and it requires approximation methods. One widely used approach is **Monte Carlo Dropout**, introduced by Gal and Ghahramani [88], which leverages dropout - a commonly used regularization technique in neural networks - as a Bayesian approximation method.

Dropout involves randomly deactivating a percentage of neurons in a layer during training, which prevents overfitting and encourages generalization (see Figure 3.14) [92]. In standard practice, dropout is disabled during inference, meaning all neurons remain active. However, **test-time dropout (TTD)** extends dropout to test time and performs multiple stochastic forward passes with different dropout masks, resulting in a distribution of outputs rather than a single deterministic prediction.

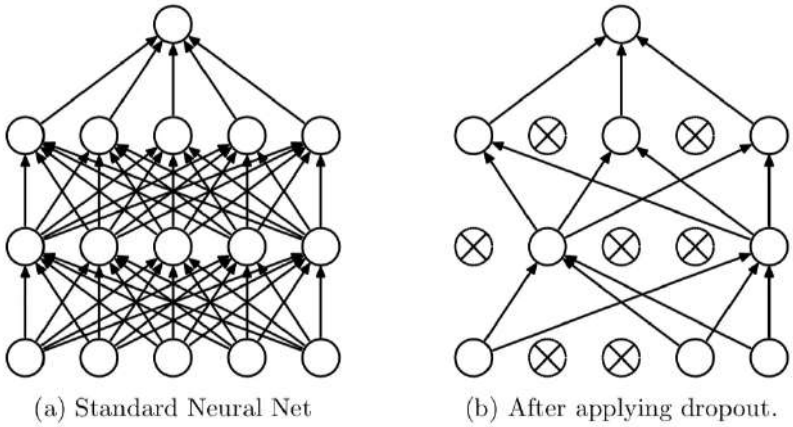


Figure 3.14: Illustration of dropout in neural networks. Source: [92]

The average of all estimations is used to make the final prediction, and the variance of the predictions is calculated to model the uncertainty [93]. This method provides a practical approximation of Bayesian inference, capturing epistemic uncertainty without significantly increasing the computational costs.

In the context of brain tumor segmentation, TTD is particularly valuable for detecting regions with high model uncertainty, such as poorly represented tumor sub-regions or cases with atypical MRI characteristics. For example, Ballestar et al. [93] applied TTD to voxel-wise uncertainty estimation in 3D-UNet-based segmentation models, demonstrating its effectiveness in highlighting ambiguous tumor boundaries.

Overall, TTD provides a scalable and efficient alternative to other Bayesian Approximation methods, such as VI and MCMC, while still effectively capturing epistemic uncertainty. Nevertheless, its effectiveness depends on the number of stochastic passes. Too few iterations may lead to unreliable uncertainty estimates, while a large number of iterations could significantly increase the computation time [93].

3.4.3.4 Test-time augmentation

Test-time augmentation (TTA) is a simple and widely used predictive method of estimating uncertainty. This approach involves generating multiple variations of each test sample using data augmentation techniques. These augmented samples are then evaluated to obtain a predictive

distribution, which serves as a measure of uncertainty (see Figure 3.15). The key idea behind TTA is that by exposing the model to different plausible versions of the same input, it becomes possible to assess the consistency of predictions and identify regions of high uncertainty [82].

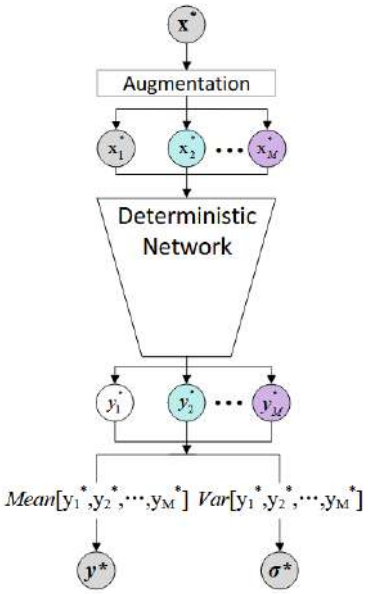


Figure 3.15: Visualization of test-time data augmentation method. Source: [79]

TTA has been widely used in medical image analysis, as this area of research involves heavy use of data augmentation since collecting medical images is costly. Moshkov et al. [94], for example, used TTA to estimate uncertainty in a cell segmentation task by generating multiple test-time variations and aggregating predictions from a UNet or Mask R-CNN architecture using majority voting. Similarly, Ballestar and Vilaplana [93] utilized TTA for voxel-wise aleatoric uncertainty estimation in brain tumor segmentation, demonstrating its effectiveness in highlighting ambiguous regions.

One of the main advantages of TTA is its ease of implementation. It does not require modifying the underlying model, retraining, or incorporating additional data, making it a practical and computationally efficient approach for uncertainty estimation [82].

Nevertheless, several things should be kept in mind when applying this technique. Most importantly, only valid data augmentation techniques should be applied to the data, meaning that the augmentations should not generate out-of-distribution samples. Inappropriate aug-

mentations could lead to artifacts or unrealistic variations which could misrepresent the true distribution of the data and, as a result, lead to unreliable uncertainty estimates [82].

Moreover, several open questions also remain regarding the impact of different data augmentation methods on uncertainty estimation produced by TTA. Different augmentations may contribute unevenly to uncertainty quantification - basic transformations like reflection may add limited uncertainty information, whereas domain-specific augmentations (e.g., shearing or intensity shifts in MRI) may be more informative [82]. Furthermore, an optimal number of data augmentations required for a reliable uncertainty estimation remains an open challenge [82]. This is especially relevant in resource-constrained applications, such as large-scale medical imaging, where inference must be performed at scale with limited computational resources.

Overall, TTA is a simple and efficient method for uncertainty estimation, making it particularly attractive for medical imaging applications. However, its effectiveness depends on careful augmentation selection and aggregation strategies.

3.4.3.5 Ensemble methods for uncertainty estimation

Besides minimizing biases and variability of individual models and improving generalizability, ensemble methods are also increasingly used to model uncertainty on deep neural network predictions [82]. The ensemble approach for uncertainty estimation relies on utilizing multiple models to generate predictions, which are then combined. The variability among these individual predictions is used as an indicator of uncertainty [79]. This process is demonstrated in Figure 3.16.

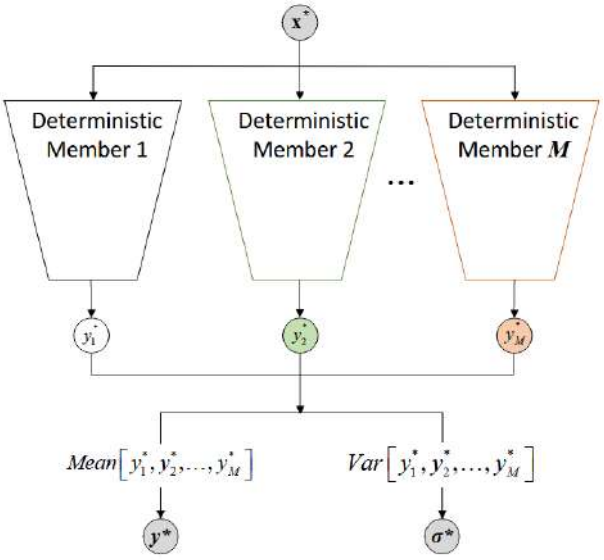


Figure 3.16: An illustrated of the ensemble method for uncertainty estimation. Source: [79]

This method was first shown to be effective by Lakshminarayanan et al. [95], whose approach employs two-headed networks that predict both the model output and an associated uncertainty measure. Expanding on this, Ovadia et al. [96] conducted a comprehensive comparison of uncertainty quantification techniques and found that even a small ensemble of five models significantly improved robustness to dataset shifts compared to other methods.

Despite their advantages, ensemble methods come with increased computational and memory costs, making them less practical for real-world deployment, especially in resource-constrained environments [82]. Training multiple models requires proportional increases in storage, inference time, and power consumption, which can be a limiting factor in clinical and embedded applications.

However, ensembles remain relatively easy to implement, as they do not require fundamental modifications to deterministic models. Furthermore, training can be highly parallelized, as ensemble members train independently, reducing training time when sufficient computational resources are available. Nonetheless, the linear increase in computational demand with each additional model can make ensembles challenging to deploy in scenarios where efficiency is a priority.

In summary, ensemble methods offer a simple yet powerful approach to uncertainty quantifi-

cation, providing better-calibrated predictions and improved robustness to data shifts. However, their practicality depends on balancing computational feasibility with accuracy improvements.

3.4.4 Applications in brain tumor segmentation

Uncertainty estimation is an important aspect of brain tumor segmentation, particularly in order to increase trust in automated segmentation techniques and pave the way to clinical interpretation. Recent studies have been exploring different methods to quantify uncertainty to enhance the reliability of brain tumor segmentation.

The significance of uncertainty estimation is underscored by the inclusion of the QU-BraTS (Quantification of Uncertainty for Brain Tumor Segmentation) sub-task in the BraTS challenges of 2019 and 2020. In this sub-task, uncertainty estimates in brain tumor segmentation were assessed and ranked based on their accuracy [97]. The overarching goal of this task was to improve the reliability of automated segmentation methods and thereby, facilitate their incorporation in clinical practice [97].

In this context, Li et al. [98] developed a region-based EDL model, evaluated on the BraTS 2020 dataset. Their approach interprets neural network outputs as evidence values, parameterizing them as a Dirichlet distribution and treating predicted probabilities as subjective opinions. This framework has been shown to produce reliable uncertainty maps and robust segmentation results.

Uncertainty quantification has also been used to improve the segmentation accuracy. Lee et al. [99] developed a general framework exploiting uncertainty information to improve a baseline segmentation model (U-Net). This approach has shown significant improvements, particularly in the enhancing tumor and tumor core regions in the BraTS 2018 dataset.

Furthermore, Jungo et al. [77] evaluated common approaches to measure uncertainty in brain tumor segmentation, such as TTD, TTA, and ensemble methods, focusing on their calibration and ability to detect segmentation failures. They found that while these methods are generally well-calibrated at the dataset level, they exhibit limitations at the subject level, particularly in error localization.

Despite the recent advances in the area of uncertainty estimation in brain tumor segmentation, the different uncertainty methods are still under-explored. Further research is crucial in order to integrate automated segmentation techniques into clinical practice. Addressing current limitations in uncertainty estimation will be pivotal in accelerating this process.

Chapter 4

Research Gaps & Thesis Objectives

As shown in Chapter 3, significant advancements have already been made to improve algorithmic methods for brain tumor segmentation. This chapter outlines key research gaps identified from the state-of-the-art literature and presents the objectives of this thesis, which aim to address such gaps.

4.1 Research gaps

The existing literature has shown the efficacy of individual deep learning architectures, yet several gaps remain that involve ensuring consistent performance across different tumor sub-regions, handling model uncertainty, and optimizing ensembles of models for improved generalizability.

4.1.1 Inconsistent model performance across tumor sub-regions

Different classes of segmentation models, such as CNNs, attention-based architectures, and hybrid models, exhibit different strengths when applied to distinct tumor sub-regions (NCR, ED, ET). CNN-based models, which excel at capturing local spatial features, often struggle with segmenting highly heterogeneous regions such as ED [52]. In contrast, attention-based architectures leverage global contextual information, making them more effective for sub-regions like ET, where long-range dependencies are critical [52, 100].

Despite these differences, most studies in the literature attempt to address segmentation

challenges primarily through loss function modifications rather than leveraging the complementary strengths of different models [100]. Current ensemble strategies often treat all sub-regions equally, without accounting for these performance variations. Thus, a more adaptive approach that incorporates model-specific strengths per sub-region could improve overall segmentation quality.

4.1.2 Limited integration of uncertainty into ensemble strategies

As shown in Sections 3.3 and 3.4.3.5, ensemble methods have been proven effective in improving segmentation accuracy and they have also been used to estimate uncertainty. Yet, existing methods for model ensembling primarily rely on simple aggregation techniques, such as averaging probabilistic outputs or majority voting. These strategies fail to account for uncertainty-weighted decision-making, where models contributing high-uncertainty predictions should have less influence on the final segmentation. A more sophisticated ensemble framework that integrates uncertainty estimates at a voxel-wise level could possibly enhance segmentation reliability.

4.1.3 Lack of a standardized toolbox for uncertainty-aware brain tumor segmentation

While numerous studies have explored uncertainty estimation in brain tumor segmentation as shown in Section 3.4, there is no toolbox for visualizing segmentation masks including uncertainty estimates. Researchers typically develop methods for specific datasets and tasks, without consulting clinical practitioners. While several toolboxes for brain tumor segmentation exist, for instance, Radionics [101], none of them include uncertainty estimation. A modular and open-source toolbox that integrates uncertainty estimation techniques could facilitate better evaluation of uncertainty-aware segmentation methods through feedback from clinicians.

4.2 Thesis objectives

Addressing the aforementioned gaps, the primary objective of this thesis is to develop an uncertainty-aware ensemble approach for brain tumor segmentation in MRI.

To achieve this, the following sub-objectives are proposed:

1. **Investigate performance of state-of-the-art models accross tumor sub-regions**
   (a) Assess the performance of different state-of-the-art segmentation model architectures, including Swin UNETR, SegResNet, V-Net, and Attention UNet, on different brain tumor sub-regions: necrotic core, edema, and enhancing tumor.
   (b) Identify the strengths and limitations of each model in segmenting these sub-regions to later inform the ensemble strategy.
2. **Develop an uncertainty-aware ensemble-based segmentation model**
   (a) Investigate various uncertainty estimation techniques applicable to medical image segmentation, including test-time dropout and test-time augmentation.
   (b) Select and implement the most suitable methods to quantify both epistemic and aleatoric uncertainties within the ensemble model.
   (c) Integrate the selected uncertainty estimation methods as well as the performance of each model on each sub-region into the ensemble to assess and improve segmentation reliability.
3. **Create a prototype of the clinical toolbox**
   (a) Develop a user-friendly software toolbox that would allow clinicians to apply the ensemble-based segmentation model to MRI data.
   (b) Implement features for visualizing segmentation results and associated uncertainty estimations.

In summary, this thesis aims to bridge research gaps in brain tumor segmentation by developing an uncertainty-aware ensemble and incorporating it into a clinical toolbox for practical use. By leveraging the strengths of different neural network architectures, integrating uncertainty estimation, and designing a user-friendly interface, the approach presented in this thesis aims to improve segmentation reliability and the applicability of automated segmentation methods in practice. The following chapter will detail the methodology of this project used to achieve the outlined objectives.

Chapter 5

Materials and Methods

This chapter presents the methodology used to achieve the project's objectives outlined in the previous chapter. It begins with a detailed description of our use of the BraTS 2021 Adult Glioma dataset, including its characteristics, the applied preprocessing steps, and the data splitting strategy. Next, it outlines the model architectures, training process, and hyperparameter tuning experimental design approach. Finally, it describes the ensemble method and the techniques used to quantify model uncertainty.

5.1 Dataset and preprocessing

The experiments in this thesis are based on the BraTS 2021 Adult Glioma dataset, which remains one of the most widely used benchmarks for glioma segmentation. While more recent editions, such as BraTS 2024, have introduced post-surgery cases, the primary focus of this work is on pre-operative glioma segmentation, where tumor sub-regions (NCR, ED, ET) remain structurally unchanged. BraTS 2021 provides high-quality, expert-annotated MRI scans of pre-operative gliomas, making it a reliable dataset for evaluating the proposed methods and comparing them to existing approaches [37]. Moreover, using BraTS 2021 ensures consistency with prior research, allowing for a more meaningful assessment of model performance in the context of glioma segmentation.

5.1.1 BraTS 2021 Adult Glioma dataset overview

The BraTS 2021 Adult Glioma dataset comprises MRI scans collected across 15 different institutions [37]. This fact contributes to the variability in the dataset, including the variability in acquisition protocols and the exact models of MRI scanners used, reflecting the real diverse imaging practices across different institutions. The exact number of cases included in the dataset from each institution is depicted in Figure 5.1. Notably, the largest contributions come from UPENN-GBM (403 cases) and UCSF-PDGM (263 cases), while other institutions provide far fewer samples. This imbalance highlights the dominance of certain data sources, which should be taken into consideration as it may impact model generalization across underrepresented collections.

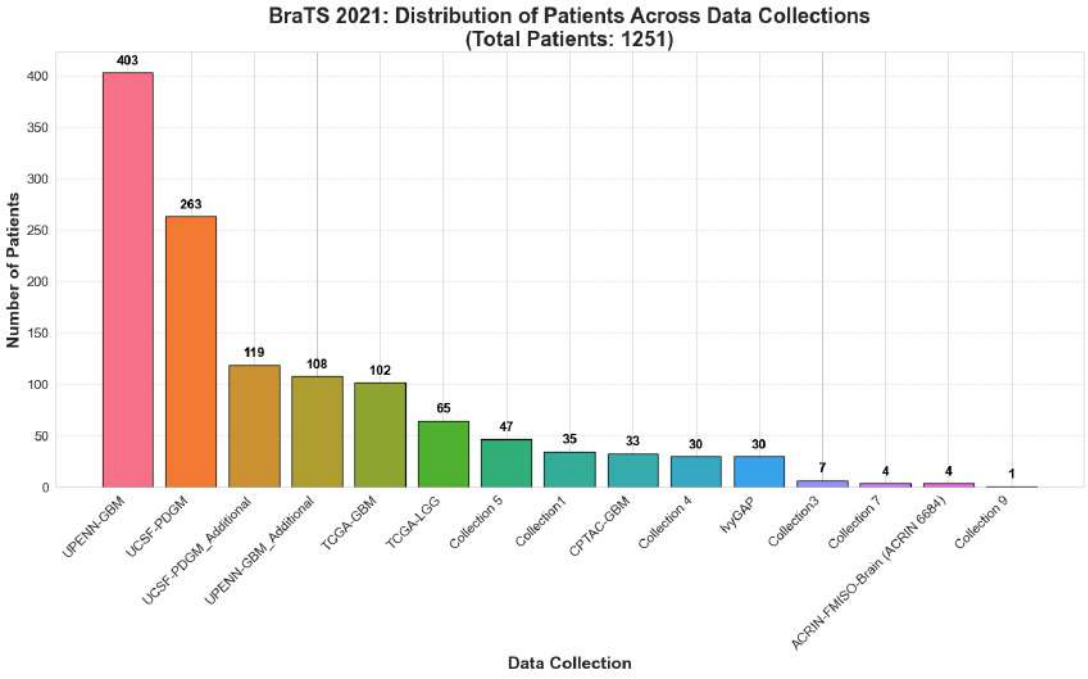


Figure 5.1: Distribution of patients across data collections in the BraTS 2021 dataset, comprising a total of 1251 patients.

The diversity in institutions from which the data comes from and in the imaging acquisition protocols makes BraTS 2021 a well-balanced representation of real-world variability, which is crucial for evaluating segmentation methods. Additionally, its widespread adoption in prior studies ensures that models trained and tested on this dataset can be effectively compared to

existing state-of-the-art approaches.

5.1.2 Sub-region labeling & distribution

The BraTS 2021 dataset provides manual annotations for three tumor sub-regions: necrotic core (NCR), peritumoral edema (ED), and enhancing tumor (ET). The details about their morphological characteristics can be found in Section 2.1.2. These sub-regions exhibit significant variability in size and coverage across patients, as summarized in Table 5.1. ED is the most prevalent and largest sub-region, with a mean volume of 60.21 cm^3, while NCR and ET are smaller, with mean volumes of 14.31 cm^3 and 21.45 cm^3, respectively, and do not appear in some cases.

Table 5.1: Volume Statistics for Brain Tumor Sub-regions in BraTS dataset.

Anatomical Structure	**Necrotic Core (NCR)**	**Edema (ED)**	**Enhancing Tumor (ET)**
No. cases	1208	1250	1218
Coverage (%)	96.56	99.92	97.36
Mean volume (cm^3)	14.31	60.21	21.45
Median volume (cm^3)	7.37	52.31	17.34
Max volume (cm^3)	189.15	216.41	111.25

The patient-level distribution of sub-region volumes is also illustrated in Figure 5.2, which highlights the differences in size, variability, and presence of outliers among the three sub-regions. The ED sub-region has the greatest median volume and also the highest variability, being the most prominent sub-region. NCR, while generally being the smallest sub-region, shows a significant number of outliers, indicating that some patients have unusually large necrosis. In contrast, the ET sub-region, even though larger than NCR, displays a much smaller number of outliers. These distributions emphasize the heterogeneity in tumor sub-regions across patients, which poses unique challenges for segmentation algorithms.

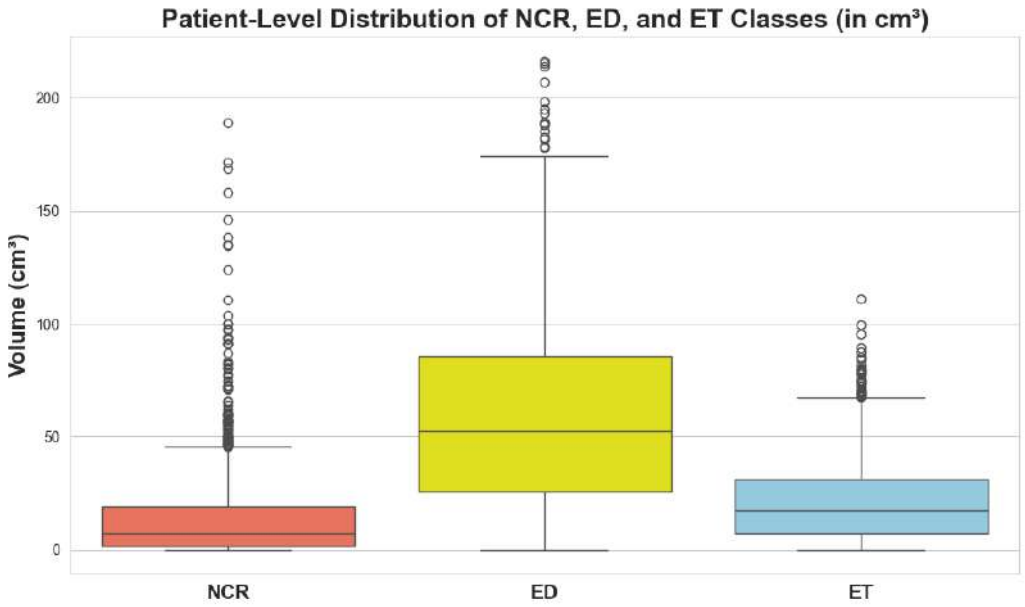


Figure 5.2: Patient-level distribution of tumor sub-region volumes (in cm^3) for NCR, ED, and ET in the BraTS dataset. The box-plots display the median, inter-quartile range (IQR), and variability in sub-region volumes.

While the dataset provides annotations for the different sub-regions of brain tumors, these sub-regions do not always co-occur together. Table 5.2 summarizes the counts for the observed combinations of the tumor sub-regions across patients in the dataset. Most patients exhibit all three sub-regions (NCR, ED, ET), but certain cases lack specific regions.

Table 5.2: Observed tumor sub-region combinations and their prevalence in the dataset. Some region combinations, such as (NCR absent, ED absent, ET present), were not observed in the dataset. Only observed combinations are shown.

NCR	ED	ET	Patient Count
✓	✓	✓	1180
✗	✓	✓	37
✓	✓	✗	27
✗	✓	✗	6
✓	✗	✓	1

The absence of certain tumor sub-regions in the dataset can be attributed to biological, radiological, and annotation-related factors. As mentioned in Section 2.1.3, LGGs typically lack NCR due to their slower growth and better vascularization, while contrast enhancement

variability can weaken or obscure ET, especially in infiltrative margins [102, 103]. Annotation inconsistencies and evolving protocols may further contribute to missing smaller sub-regions[37, 104].

This absence may bias segmentation models, making rare sub-regions like NCR harder to learn and increasing misclassification risks. Uncommon sub-region combinations further hinder generalization, impacting segmentation accuracy and model robustness in real-world applications.

5.1.3 Image preprocessing

The scans included in the BraTS 2021 Adult Glioma dataset consist of four modalities: T1, T1CE, T2, and FLAIR (see Section 2.2.3 for more details). All BraTS 2021 scans have gone through standardized preprocessing, including conversion of the DICOM files to the NIfTI file format, co-registration to the same anatomical template, resampling to a uniform isotropic resolution ($1mm^3$), and skull-stripping [37]. Table 5.3 summarizes the standardized voxel spacing and image dimensions for all scans in the BraTS 2021 dataset. Additionally, Figure 5.3 provides an illustration of a sample case with all modalities and ground truth segmentation.

Table 5.3: Standardized Voxel Spacing and Image Dimensions in the BraTS Dataset.

Dataset Statistics	**Voxel spacing (mm)**	**Image dimensions (voxels)**
min	(1.0, 1.0, 1.0)	(240, 240, 155)
median	(1.0, 1.0, 1.0)	(240, 240, 155)
max	(1.0, 1.0, 1.0)	(240, 240, 155)

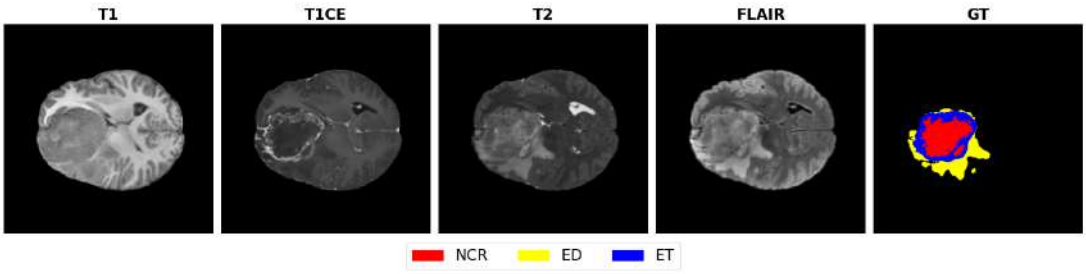


Figure 5.3: Example of BraTS 2021 Adult Glioma case across four modalities (T1, T1CE, T2, FLAIR) and a corresponding expert-annotated, ground truth (GT) tumor segmentation.

To ensure consistency across scans from different MRI scanners, Z-score intensity normalization was applied separately for each MRI modality. This process involved computing the mean and standard deviation only within the brain region (excluding the background) and using these values to normalize the intensity distribution. As a result, each MRI volume was transformed to have zero mean and unit variance, ensuring comparability across subjects. This normalization step mitigates scanner-dependent variations and provides a standardized feature distribution [105], enhancing the robustness of the model to intensity differences across datasets. The effect of this normalization can be observed in Figure 5.4, which shows the Z-score normalized MRI modalities for a sample BraTS 2021 case.

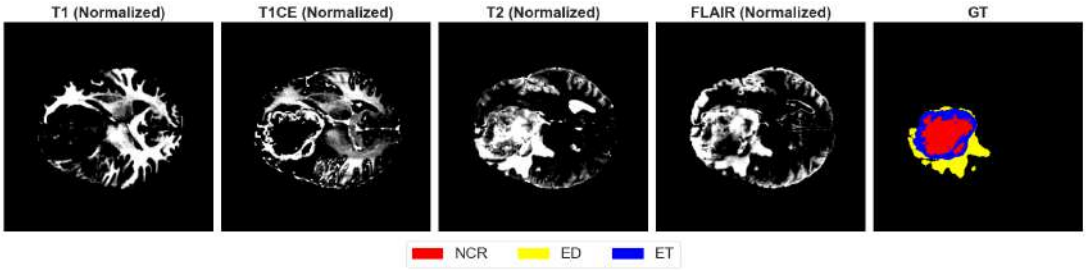


Figure 5.4: Example of Z-score normalized MRI modalities (T1, T1CE, T2, FLAIR) for the same BraTS 2021 Adult Glioma case as shown in Figure 5.3. The intensity normalization was applied separately for each modality to ensure zero mean and unit variance within the brain region.

Lastly, in order to guarantee compatibility with multi-class segmentation loss function, the ground truth labels were converted into one-hot encoded format. Each voxel was assigned a binary vector indicating its class membership, resulting in a four-channel label representation (background, NCR, ED, and ET).

5.1.4 Data splitting

To address the issue of sub-region co-presence imbalance, the dataset was split into training, validation, and test sets using a s**tratified 5-fold cross-validation** strategy. Stratified cross-validation ensures that each subset of the dataset, called fold, maintains the same proportional distribution of sub-region combinations. As a result, it reduces the risk of bias caused by the imbalanced dataset.

The splitting strategy consisted of two main steps: 1) test allocation and 2) training and validation split via stratified 5-fold cross-validation.

5.1.4.1 Test allocation

First, tumor cases were categorized on the basis of the sub-region combinations, as shown in Table 5.2. The test set was formed to ensure proper representation of rare sub-region combinations:

- For the most common combination (all sub-regions present), **15% of the cases** were randomly assigned to the test set, while the remaining cases were retained for training and validation.
- For rare sub-region combinations (e.g. NCR absent, ED present, ET present), **at least five cases** were allocated to the test set, or fewer if availability was limited.
- The most infrequently occurring combination (NCR present, ED absent, ET present) was **assigned entirely to the training set** to maximize learning from limited data.

The test allocation strategy is depicted in Figure 5.5, which shows how tumor cases were divided between training/validation and test sets based on the sub-region co-presence.

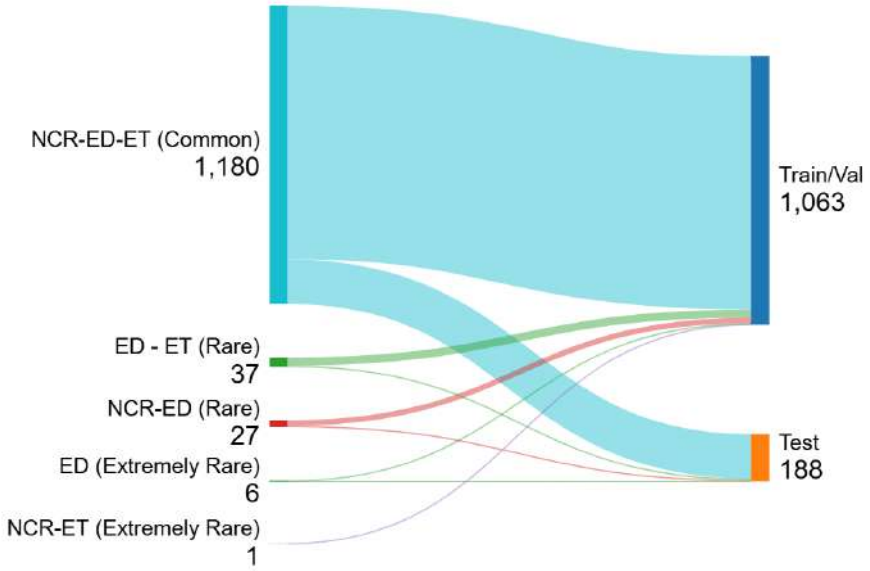


Figure 5.5: Test set allocation strategy based on sub-region co-presence.

5.1.4.2 Training and validation split

After setting aside the test set, the remaining dataset was into training (80%) and validation (20%) while maintaining the same proportional distribution of sub-region combinations. This stratified approach ensured that rare tumor sub-types were well-represented in both subsets, preventing bias.

The stratified training-validation split procedure is shown in Figure 5.6. It shows how data was divided, while maintaining the original class distributions.

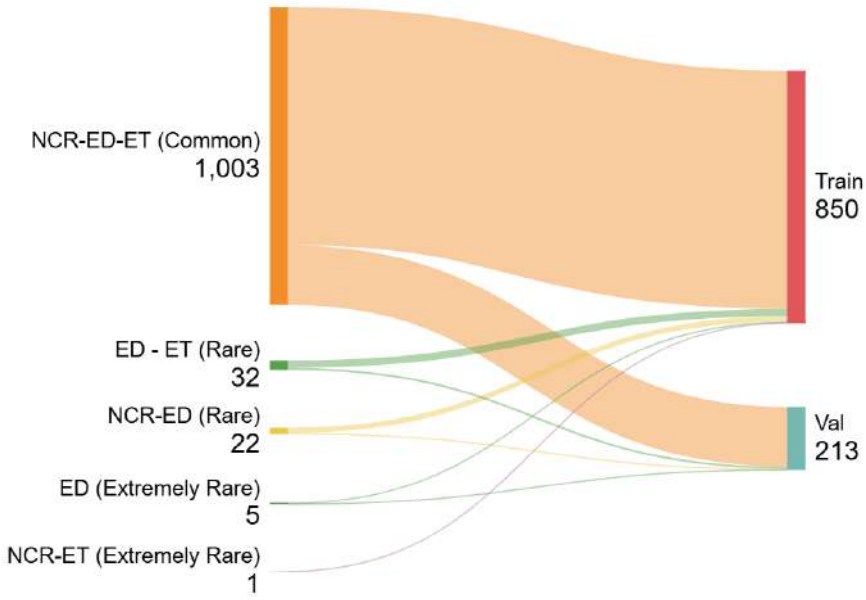


Figure 5.6: This diagram illustrates how the remaining dataset (after test set allocation) was divided into training (80%) and validation (20%) subsets.

5.1.5 Final training split and model finalization

After the completion of hyperparameter tuning via 5-fold cross-validation, a final training-validation split was performed in order to train the definitive model with optimal hyperparameters.

To create the final split, the original test set from the initial allocation was preserved, while the training and validation subsets were redefined. All cases were split into training (90%) and validation (10%) subsets using stratified sampling, similar to the approach described above, preserving the original distribution of sub-region combinations.

With this final split, the model was retrained from scratch using the best hyperparameters found during cross-validation. No further tuning was performed. The final model was then evaluated once on the held-out test set to obtain the reported performance metrics.

This protocol of creating the final training-validation split ensures that there is no data leakage as the test set remains fully independent, providing unbiased estimates of model performance.

5.2 Model architectures

Since one of the primary objectives of this thesis is to develop an uncertainty-aware ensemble of state-of-the-art segmentation models, several deep learning architectures have been selected as ensemble members. Each architecture offers distinct advantages and limitations, contributing to a more robust and reliable segmentation system.

This section will outline each of these selected architectures - its characteristics, advantages, and limitations. This is followed by an overview of the training configuration, data augmentations used, as well as the computational resources utilized for model training.

5.2.1 V-Net

5.2.1.1 Overview of the V-Net architecture

V-Net is a CNN developed by Milletari et al. [51] for volumetric medical image segmentation, making it well-suited for the segmentation of brain tumors in volumetric MRI scans. Unlike U-Net which processes 2D slices (see more details in Section 3.2.1), V-Net employs a 3D encoder-decoder architecture, working directly on entire 3D MRI volumes to preserve spatial information critical for brain tumor segmentation [51].

V-Net consists of a compression path that progressively reduces spatial resolution while extracting deep features and an expansion path that restores spatial resolution to generate the final segmentation mask [51]. In contrast to the traditional CNNs that use max-pooling, V-Net replaces pooling layers with strided convolutions ($2 \times 2 \times 2$) and deconvolutions, which improve memory efficiency and feature learning [51]. Additionally, residual connections are incorporated at each stage to facilitate gradient flow and accelerate convergence [51]. Figure 5.7 presents a schematic representation of the entire V-Net architecture.

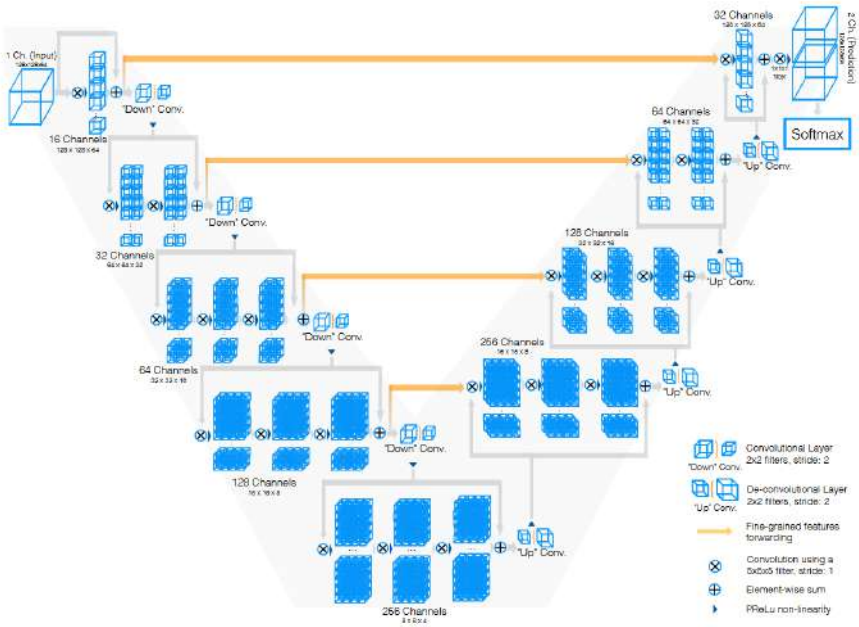


Figure 5.7: Schematic representation of the V-Net architecture. Source: [51]

5.2.1.2 Role in the ensemble

In the ensemble model, V-Net complements architectures such as Attention U-Net and Swin-UNETR which both utilize attention mechanisms. While these models are superior at capturing long-rage dependencies, V-Net focuses on extracting hierarchical volumetric features by utilizing its fully convolutional architecture. As outlined by Milletari et al. [51] and subsequently confirmed by further research, V-Net's key advantage is its ability to effectively segment small and complex sub-regions, such as NCR and ET. These regions often exhibit subtle variations in intensities and intricate spatial arrangements, which V-Net can handle well. This ability to successfully segment smaller tumor sub-regions makes V-Net add a layer of robustness to the ensemble, ensuring better segmentation of potentially under-segmented areas.

5.2.1.3 Implementation details

The V-Net architecture was implemented using the `MONAI` library which is a framework optimized for deep learning in medical imaging [106]. The configuration of the model is summarized in Table 5.4.

Table 5.4: V-Net model configuration used in this thesis.

Parameter	Value
Spatial dimensions	3D
Input channels	4 (T1, T1CE, T2, FLAIR)
Output channels	4 (background, NCR, ED, ET)
Dropout probability (encoding)	0.2
Dropout probability (decoding)	(0.2, 0.2)
Dropout applied across spatial dimensions	3D
Activation function	ELU

The input to the model consists of four MRI sequences: T1, T1CE, T2, and FLAIR. The model outputs a four-channel segmentation map, predicting the background, NCR, ED, and ET. To ensure robust training and segmentation performance, dropout regularization was applied in both the encoder and decoder paths, preventing overfitting. The Exponential Linear Unit (ELU) activation function was chosen to improve gradient flow and prevent vanishing gradients, facilitating stable training.

5.2.2 SegResNet

5.2.2.1 Overview of the SegResNet architecture

The SegResNet architecture was introduced by Andriy Myronenko in his 2018 paper *3D MRI brain tumor segmentation using autoencoder regularization* [107]. His approach won first place in the 2018 edition of the BraTS challenge.

The SegResNet follows a U-Net-like encoder-decoder architecture, with residual blocks inspired by ResNet which improve feature propagation. Furthermore, instead of batch normalization, SegResNet utilizes GroupNorm normalization which is more suitable for smaller batch sizes. For downsampling, SegResNet uses strided convolutions, while upsampling is done using deconvolutions, pixel shuffling, or interpolation [107].

In its original formulation, SegResNet also includes an optional variational autoencoder (VAE) branch [107], which reconstructs the input image from a latent representation to improve regularization. However, in this work, only the segmentation pathway is used, and the VAE

branch is omitted. A schematic representation of the SegResNet architecture is presented in Figure 5.8.

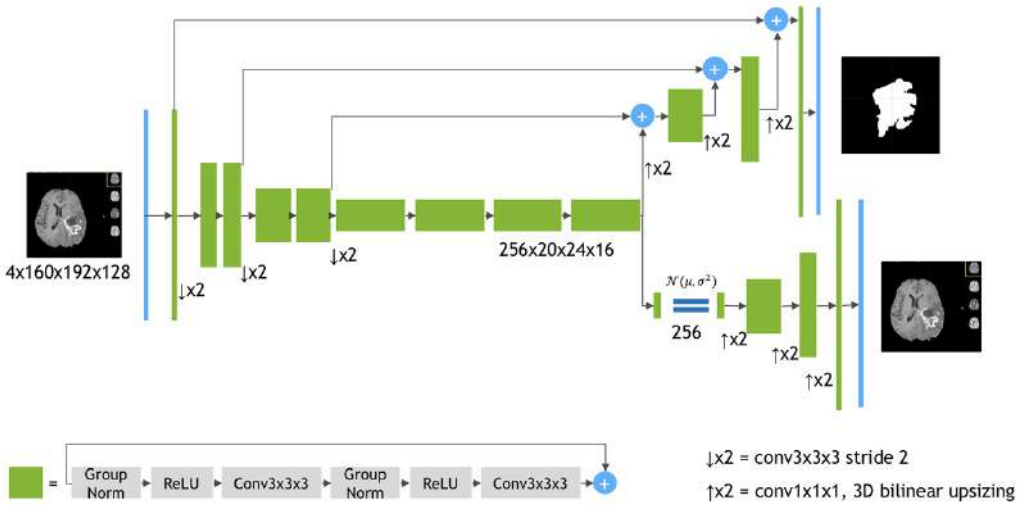


Figure 5.8: SegResNet architecture. Each green block is a ResNet-like block with GroupNorm normalization. The implementation used in this work does not include the VAE branch, which means that only the upper segmentation path is utilized. Source: [107]

5.2.2.2 SegResNet role in the ensemble

SegResNet contributes to the ensemble by leveraging its residual encoder-decoder structure, which enhances feature propagation and preserves spatial consistency. This architectural approach differs from those of V-Net, Attention U-Net, and SwinUNETR, by introducing diversity in feature extraction that enables the ensemble to capture a wider range of tumor morphologies.

Previous research suggests that SegResNet performs well in whole tumor (WT) segmentation, while its generalization to ET and tumor core (TC; consists of NCR and ED) is somewhat weaker [108]. Since WT consists of NCR, ED, and ET, this indicates that SegResNet contributes most effectively to NCR and ED segmentation, where its residual connections help retain structural details.

The inclusion of SegResNet within the ensemble increases architectural diversity, mitigating over-reliance on the attention mechanisms incorporated in other architectures. Additionally, its robust performance in segmenting larger tumor regions facilitates improved overall structural coherence in the final segmentation output.

5.2.2.3 Implementation details

The SegResNet model used in this work is implemented using the `MONAI` library [106], which provides a version optimized for medical image segmentation. The model configuration is summarized in Table 5.5.

Table 5.5: SegResNet model configuration used in this thesis.

Parameter	Value
Input channels	4 (T1, T1CE, T2, FLAIR)
Output channels	4 (background, NCR, ED, ET)
Initial number of filters	16
Downsampling blocks	[1, 2, 2, 4]
Upsampling blocks	[1, 1, 1]
Dropout probability	0.2

Similarly to other models, SegResNet processes multi-modal MRI inputs, utilizing four MRI sequences (T1, T1ce, T2, and FLAIR). The segmentation output consists of four channels, corresponding to the background, NCR, ED, and ET.

The encoder-decoder structure consists of four downsampling blocks and three upsampling blocks, following the `MONAI` SegResNet design. The initial number of filters is set to 16, doubling at each downsampling stage to progressively capture higher-level features. The dropout probability of 0.2 is applied to reduce overfitting while maintaining feature variability.

5.2.3 Attention U-Net

5.2.3.1 Overview of the Attention U-Net architecture

The Attention U-Net architecture was introduced by Oktay et al. in 2018 for pancreas segmentation [56]. The key innovation of this model is the integration of attention gates (AGs), which allow the network to focus on relevant regions while suppressing irrelevant background information.

Unlike standard U-Net, where all skip connections directly transfer features from the encoder to the decoder, Attention U-Net modulates these connections by applying spatial attention

mechanisms. This helps the model refine feature maps and enhance the segmentation of target structures, while reducing false positives [56]. Figure 5.9 presents a schematic representation of the Attention U-Net architecture that illustrates how attention gates refine feature propagation through skip connections.

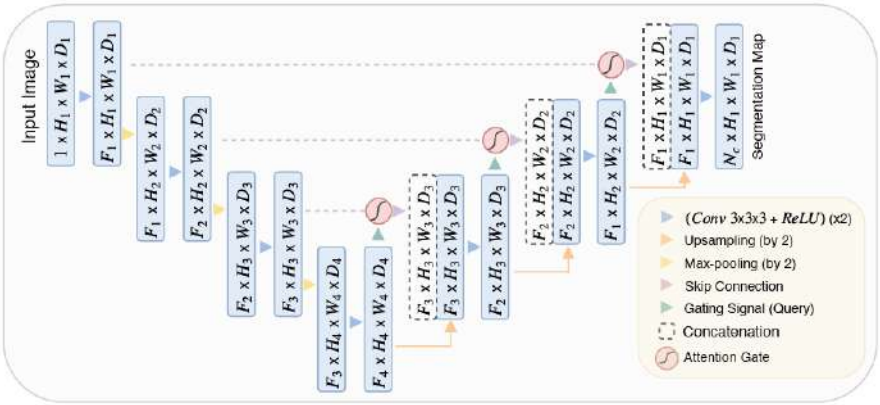


Figure 5.9: Architecture of the Attention U-Net model. The input image is progressively filtered and downsampled in the encoder, while attention gates (AGs) refine the features propagated through skip connections. AGs use contextual information from coarser scales to enhance feature selectivity. Source: [56]

5.2.3.2 Role in the ensemble

The attention mechanisms integrated into the Attention U-Net architecture have been shown to play a critical role in the model's ability to focus on relevant regions within medical images, thereby improving segmentation accuracy [56]. Previous research shows that 3D Attention U-Net is particularly effective at tumor boundary delineation, achieving a superior Hausdorff distance score which measures the performance at boundary delineation [109]. A recent study by Talukder et al. [110] further confirms this finding showing that U-Net enhanced with attention mechanisms improves segmentation performance, particularly in regions where tumor boundaries are indistinct or complex.

The primary advantage of including Attention U-Net in the ensemble is its ability to enhance precision in tumor boundary delineation, complementing models such as V-Net, SegResNet, and SwinUNETR, which focus on capturing global tumor structures and optimizing overall overlap between predictions and ground truth. By leveraging attention-based refinement, the

ensemble benefits from improved segmentation consistency, particularly in challenging regions with ambiguous boundaries.

5.2.3.3 Implementation details

The Attention U-Net model used in this work was implemented using the `MONAI` library [106], which provides an optimized version of the architecture for medical image segmentation tasks. The model configuration is summarized in Table 5.6.

Table 5.6: Attention U-Net model configuration used in this thesis.

Parameter	Value
Spatial dimensions	3D
Input channels	4 (T1, T1CE, T2, FLAIR)
Output channels	4 (NCR, ED, ET, background)
Feature channels	(32, 64, 128, 256, 512)
Strides	(2, 2, 2, 1)
Dropout probability	0.2

As was the case in the other models, Attention U-Net operates on 3D volumetric MRI scans, utilizing four input channels (T1, T1CE, T2, FLAIR). The output consists of four channels, corresponding to background, NCR, ED, and ET. The model consists of a multi-scale feature extraction pipeline, where feature channels progressively increase from 32 to 512, enabling the network to capture low-level spatial details as well as high-level contextual information. The strides $(2, 2, 2, 1)$ control the downsampling process, ensuring that deep features are preserved at the final resolution stage. Dropout regularization with a probability of 0.2 is applied to prevent overfitting, improving the generalization of the model.

5.2.4 SwinUNETR

5.2.4.1 Overview of the SwinUNETR architecture

The SwinUNETR architecture is the most recent model among the four models incorporated into the ensemble. It was introduced in 2022 by Hatamizadeh et al. [66] for brain tumor

segmentation in MRI scans. SwinUNETR combines the strengths of the Swin Transformer and the U-Net architecture, enhancing performance in semantic segmentation tasks, particularly in the field of medical imaging.

SwinUNETR builds upon the previously introduced UNETR architecture [9] by replacing its plain transformer encoder with a Swin Transformer encoder, which employs shifted window self-attention. This design enables efficient modeling of long-range dependencies while maintaining a hierarchical feature representation [66].

Unlike traditional U-Net-based architectures, which rely solely on convolutional operations, SwinUNETR processes 3D image patches as input tokens and applies multi-scale self-attention mechanisms. This approach allows for better contextual feature extraction, making it particularly effective in segmenting tumors with variable sizes and complex boundaries. The extracted multi-resolution features from the transformer-based encoder are then fused with a CNN-based decoder via skip connections, ensuring both global contextual awareness and fine-grained segmentation accuracy [66].

Figure 5.10 provides a schematic representation of the SwinUNETR architecture, illustrating its patch-based encoding, hierarchical self-attention, and U-Net-style decoder.

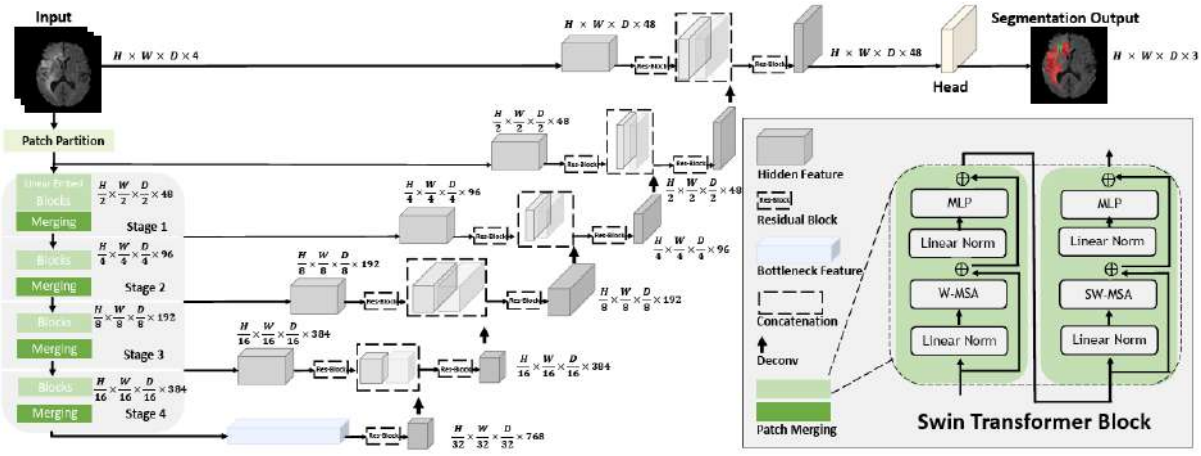


Figure 5.10: SwinUNETR architecture. The model partitions the input image into non-overlapping patches and processes them using a patch partition layer to generate windows for self-attention computation. The encoded feature representations from the Swin transformer are then fused with a UNet-like decoder via skip connections at multiple resolutions. Source: [66].

5.2.4.2 Role in the ensemble

The SwinUNETR architecture has the capability of learning multi-scale feature representations and modeling long-range dependencies [66]. Hatamizadeh et al. [66] demonstrated that SwinUNETR achieves state-of-the-art performance on the BraTS 2021 challenge, outperforming many CNN-based approaches. A key strength of SwinUNETR is its ability to differentiate between peritumoral edema (ED) and enhancing tumor (ET) due to its context-aware feature fusion, which helps resolve ambiguous boundaries between tumor sub-regions. Furthermore, its robustness to variations in tumor size and shape makes it a valuable addition to the ensemble, ensuring better generalization across different patient cases [66].

In the ensemble, SwinUNETR complements the strengths of other models. While V-Net and SegResNet effectively extract volumetric and hierarchical features, they rely purely on convolutions, which limits their ability to capture global dependencies. On the other hand, Attention U-Net refines local feature selection, but SwinUNETR expands this further with hierarchical self-attention, ensuring more stable segmentation across regions with variable contrast and morphology.

5.2.4.3 Implementation details

Similarly to other models, the SwinUNETR model used in this work was implemented using the `MONAI` library [106], which provides an optimized version of SwinUNETR for medical image segmentation. The model configuration is summarized in Table 5.7.

Table 5.7: SwinUNETR model configuration used in this thesis.

Parameter	**Value**
Image size	(96, 96, 96)
Input channels	4 (T1, T1CE, T2, FLAIR)
Output channels	4 (NCR, ED, ET, background)
Feature size	48
Dropout rate	0.2
Attention dropout rate	0.2
Dropout path rate	0.2
Gradient checkpointing	Enabled

SwinUNETR processes 3D volumetric MRI scans, utilizing four input channels corresponding to T1, T1CE, T2, and FLAIR sequences. The model outputs a four-channel segmentation map, predicting the background, NCR, ED, and ET.

The encoder of SwinUNETR employs a Swin Transformer, which partitions the input image into non-overlapping patches and processes them using shifted window self-attention to capture multi-scale contextual features. The feature size of 48 ensures a balance between computational efficiency and representational capacity.

To enhance model regularization and generalization, dropout regularization is applied at multiple levels: 1) the dropout rate of 0.2 prevents overfitting in the network's dense layers, 2) the attention dropout rate of 0.2 improves robustness by randomly deactivating attention units during training, and 3) the dropout path rate of 0.2 is used in stochastic depth regularization, which randomly skips residual connections in the transformer blocks, further stabilizing training.

Lastly, gradient checkpointing is enabled, reducing memory consumption during training and allowing for training on larger volumes without exceeding GPU memory limitations.

5.3 Data augmentation

Data augmentation is a technique for increasing the diversity and amount of data available for a deep learning model without collecting new samples [111]. This can be accomplished through various image processing techniques applied to the original data or by generating new samples using generative models [112]. Given the limited size of the annotated datasets with medical images, data augmentations allow for increasing variability in training data, thereby improving generalization and mitigating overfitting. Consequently, deep learning models for medical image segmentation significantly benefit from data augmentations.

This thesis employs a set of spatial and intensity-based data augmentations, aiming to improve model robustness while maintaining computational feasibility. Due to the high memory demands of 3D medical image processing, patch-based training was employed, enabling the model to learn from smaller sub-regions of MRI scans instead of full volumes. This approach follows the methodology of Ballestar and Vilaplana [93], increasing the number of training samples while

preserving crucial anatomical structures.

The chosen augmentations were implemented using the `MONAI` library [106] and were applied during training but **not** during validation or testing to ensure unbiased evaluation.

5.3.1 Patch-based training

Since full multi-modal MRI images are too large to process due to memory constraints, patch-based training was adopted. Instead of using whole-brain volumes, random spatial cropping (`MONAI`: `RandSpatialCropd()`) was applied to extract fixed-size patches of the region of interest (ROI) size of $(96 \times 96 \times 96)$ from the input MRI scans and corresponding segmentation masks. The three dimensions correspond to width, height, and depth (number of slices), respectively. This ROI size was selected to balance memory limitations while providing sufficient contextual information for model training. Figure 5.11 illustrates the patch-based training approach. The original brain MRI scan is shown alongside a cropped patch, demonstrating how smaller regions are sampled for model training.

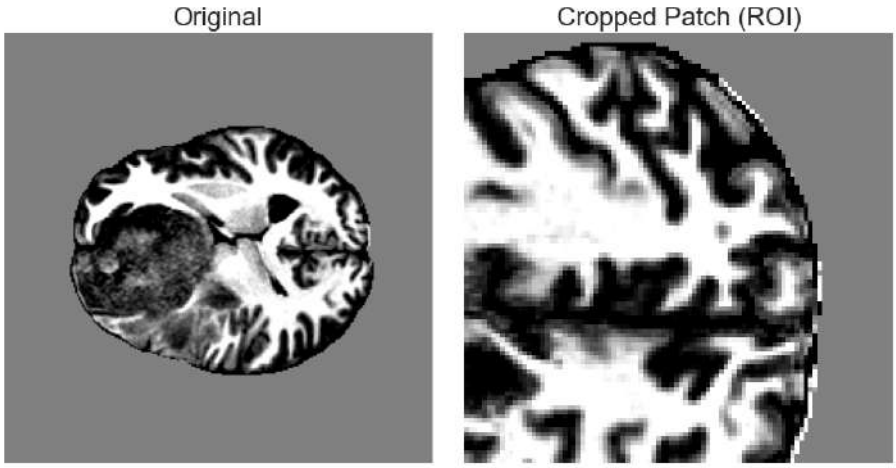


Figure 5.11: Example of patch-based training using random spatial cropping. The original, normalized MRI scan (left) is cropped into a $96 \times 96 \times 96$ patch (right) to enable memory-efficient training.

Overall, this strategy provides the following benefits:

- Ensures efficient memory usage, allowing larger batch sizes during training.

- Increases the number of training samples by generating multiple patches from a single MRI volume.
- Exposes the model to different tumor locations, preventing positional bias and improving segmentation accuracy across varying anatomical structures.

5.3.2 Spatial augmentation

To enhance spatial invariance, random flipping was applied independently along all three anatomical axes (axial, coronal, and sagittal) with a probability of 50% per axis. It was implemented using `MONAI RandFlipd()` function. Figure 5.12 demonstrates random flipping in the x-axis and y-axis. The z-axis was omitted from visualization since images are displayed in 2D.

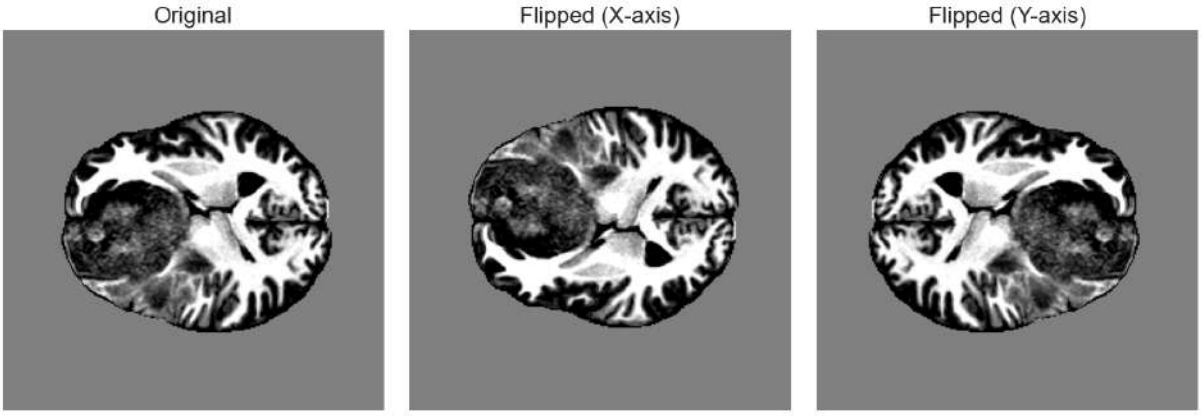


Figure 5.12: Illustration of random flipping applied to brain MRI scans. The original, normalized MRI scan (left) is shown alongside its flipped versions across the x-axis (middle) and y-axis (right).

This augmentation strategy ensures that the models do not learn positional bias from the training dataset. It improves model generalization by accounting for the arbitrary orientation of brain tumors, which lack a fixed position. Additionally, it simulates real-world variations in tumor locations across different patients. As a result, the model becomes more robust to anatomical variability, leading to improved segmentation performance across diverse cases.

5.3.3 Intensity-based augmentations

MRI intensities can vary significantly due to differences in scanners, acquisition protocols, and patient-specific factors. To enhance model robustness to these variations, intensity scaling and intensity shifting were applied using `MONAI RandScaleIntensityd()` and `RandShiftIntensityd()` functions.

In intensity scaling, each patch's intensity values were randomly scaled by a factor within ±10%. This simulates variations in contrast and brightness, accounting for scanner-dependent intensity differences. Figure 5.13 presents an example MRI scan with scaled intensity alongside the visualization of the difference between the original image and the augmented image.

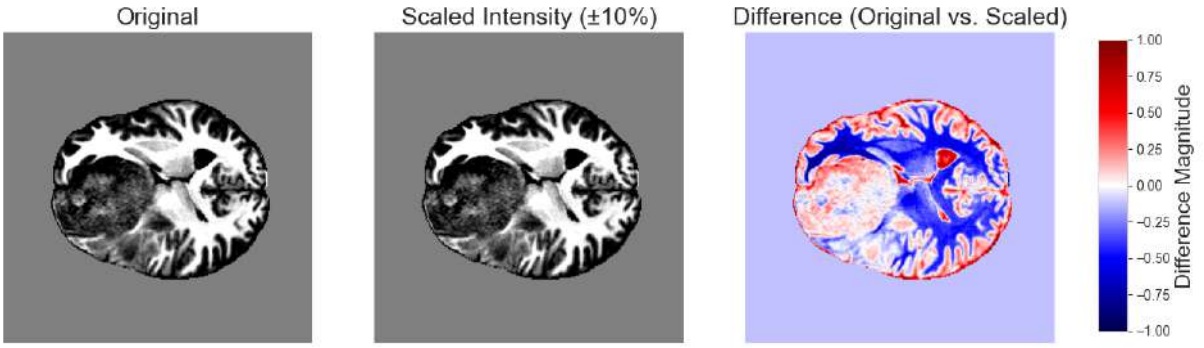


Figure 5.13: Effect of intensity scaling on an MRI scan. The original image (left) is compared to the scaled intensity image (middle), and the difference between the two is visualized using a heatmap (right). The heatmap represents pixel-wise intensity differences, scaled between -1 and 1, where red indicates increased intensity, blue indicates decreased intensity, and white represents no change.

Additionally, a random intensity shift within ±10% was applied. This accounts for lighting variations in the MRI scans and helps prevent the model from overfitting to a specific intensity distribution. Figure 5.14 illustrates the effects of intensity shifting, showing the original image, intensity-shifted image, and the difference between them.

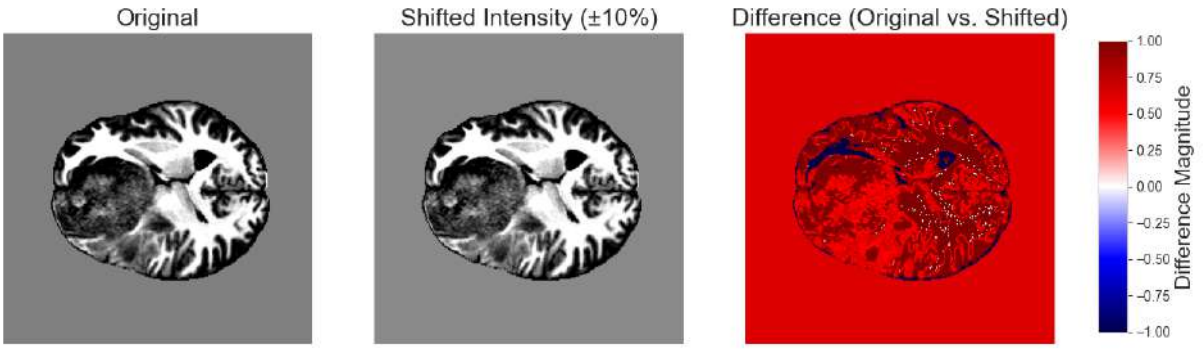


Figure 5.14: Effect of intensity shifting on an MRI scan. The original image (left) is compared to the intensity-shifted image (middle), with the difference visualized in the heatmap (right). The heatmap is scaled between -1 and 1, where red represents areas where intensity has increased, blue highlights decreased intensity, and white denotes no significant change.

These intensity-based augmentations ensure that the model learns structural features rather than relying on absolute intensity values, which can differ across imaging protocols. This significantly improves generalization to unseen datasets.

5.3.4 Summary

The selected augmentations provide a controlled increase in variability while maintaining efficient training. The balance between spatial transformations and intensity modifications ensures that the model generalizes well without imposing excessive computational demands. Since hyperparameter tuning was conducted using this augmentation setup, no additional transformations were introduced later in training to avoid disrupting the optimization process.

5.4 Loss function

The loss function employed in this thesis is designed to address the class imbalance problem highlighted in Section 5.1.2. While the task involves multi-class segmentation of three tumor sub-regions, each class is treated separately using one-hot encoding, effectively making the computation binary per class channel. This ensures the model learns to segment each tumor subregion independently, without bias toward the dominant background class. Hence, the following equations will present the loss functions for binary cases.

5.4.1 Generalized Dice Loss

One of the most commonly used loss functions in medical image segmentation is the Dice Loss, which optimizes the Dice Similarity Coefficient (DSC) directly. Originally proposed by Crum et al. [113] for multi-class segmentation evaluation, **Generalized Dice Loss** was first introduced as a CNN training loss by Sudre et al. [114]. Unlike standard Dice Loss, which may be biased toward larger classes, GDL introduces a class-specific weighting scheme to balance contributions from both small and large classes. The `MONAI` implementation follows the formulation from Sudre et al., where the GDL is computed as:

$$\mathrm{GDL} = 1 - 2\frac{\sum_{l=1}^{2} w_l \sum_n r_{ln} p_{ln}}{\sum_{l=1}^{2} w_l \sum_n (r_{ln} + p_{ln})}, \tag{5.1}$$

where:

- w_l is the class weighting factor, ensuring that smaller tumor regions contribute more to the loss;
- r_{ln} represents the ground truth for class l at voxel n;
- p_{ln} represents represents the predicted probability for class l at voxel n.

To counteract severe class imbalance, the squared inverse volume weighting scheme is applied to each class:

$$w_l = \frac{1}{\left(\sum_n r_{ln}\right)^2} \tag{5.2}$$

This weighting ensures that classes with fewer voxels receive higher importance, preventing the model from neglecting them. As a result, GDL reduces the under-segmentation of smaller structures, making the model more robust.

5.4.2 Focal Loss

While Generalized Dice Loss addresses class imbalance, it does not differentiate between easy and hard-to-classify voxels. In segmentation tasks, the model tends to quickly learn and confidently classify easy regions while struggling with ambiguous boundaries, small structures, and underrepresented classes [115].

To address this issue, **Focal Loss** was introduced by Lin et al. [115] in the context of object detection to address the class imbalance problem in scenarios where there are significantly more background samples than foreground samples. It modifies the standard Binary Cross Entropy (BCE) loss by adding a focusing term that down-weights well-classified samples and amplifies the contribution of harder, misclassified samples. The formulation of FL is:

$$\mathcal{L}_{\mathrm{FL}} = -\sum_{n} \alpha(1 - p_n)^{\gamma} \log(p_n), \tag{5.3}$$

where:

- p_n is the predicted probability for voxel n;
- α is an optional balancing factor, used to adjust the relative weight of positive and negative examples;
- γ is the focusing parameter, which controls how much emphasis is placed on hard-to-classify voxels.

Higher values of γ increase the down-weighting effect for easy-to-classify samples, making the model focus on difficult cases. In this work, the γ parameter is set to 2.0 following the work of Lin et al. [115]. The balancing factor α was not applied in this work since GDL already introduces a class-weighting mechanism through w_l factor, which assigns higher importance to underrepresented tumor sub-regions.

5.4.3 Generalized Dice Focal Loss

Both GDL and FL address different aspects of the class imbalance issue:

- GDL mitigates class imbalance by weighting underrepresented classes more heavily.
- FL focuses on difficult samples, ensuring hard-to-classify voxels contribute more to the loss.

However, neither method fully resolves both issues at the same time. To leverage the strengths of both, this thesis employs Generalized Dice Focal Loss (GDFL), which combines the benefits of GDL and FL into one function by computing their weighted sum. The combined loss is formulated as:

$$\mathcal{L}_{\text{GDFL}} = \lambda_{GDL}\mathcal{L}_{\text{GDL}} + \lambda_{FL}\mathcal{L}_{\text{FL}} \tag{5.4}$$

where

- $\mathcal{L}_{\text{GDL}}$ is the Generalized Dice Loss;
- $\mathcal{L}_{\text{FL}}$ is the Focal Loss;
- λ_{GDL} and λ_{FL} are trade-off weights determining the contribution of each loss component.

By combining both class-balancing tackled by GDL and voxel difficulty awareness addressed by FL, GDFL improves segmentation performance across all tumor sub-regions, especially for smaller and more ambiguous tissues.

5.5 Hyperparameter tuning

The performance of deep learning models relies on well-tuned hyperparameters. To ensure that the differences between the models in the ensemble can be primarily attributed to the differences between their architectures and not sub-optimal hyperparameters, a small but meaningful set of hyperparameter configurations was tested. The learning rate, optimizer, and weight decay were selected as the primary tuning parameters, as they significantly influence training stability and generalization.

5.5.1 Learning rate, optimizer, and weight decay selection

Three key hyperparameters were selected for tuning: learning rate, optimizer, and weight decay. The following sections will describe the role of these hyperparameters in training deep learning models. Finally, the tested configurations of hyperparameters will be presented.

5.5.1.1 Learning rate

The learning rate determines the step size of weight updates during backpropagation. An appropriate learning rate is crucial to ensure efficient convergence:

- A high learning rate accelerates convergence but increases the risk of overshooting the optimal solution, leading to training instability.
- A low learning rate promotes stable learning but may result in slower convergence, requiring more epochs to reach an optimal solution.

Since medical image segmentation involves complex spatial structures and high class imbalance, selecting an appropriate learning rate is necessary to balance convergence speed and model stability. In this work, two learning rates were tested:

- 1e-3: a higher learning rate for faster convergence.
- 1e-4: a lower learning rate for more stable learning.

5.5.1.2 Optimizer

The optimizer determines how model parameters are updated based on the computed gradients. Two widely used optimization algorithms were tested:

- **AdamW**: An adaptive learning rate optimizer that dynamically adjusts learning rates per parameter, improving convergence stability and weight decay handling [116].

- **SGD with momentum**: A traditional gradient descent-based optimizer that smooths updates using past gradients, often leading to better generalization but requiring careful tuning [117]. The value of the momentum was set to 0.9.

AdamW is commonly used in medical image segmentation due to its efficient parameter updates, whereas SGD is sometimes preferred for its superior generalization in deep learning models.

5.5.1.3 Weight decay

Weight decay is a form of L2 regularization that prevents overfitting by penalizing large weight values in the model. This is particularly important in medical imaging tasks, where models can memorize small dataset variations rather than learning generalizable patterns. In this project, the following two values of weight decay were investigated:

- 1e-5 - lower weight decay which retains more model flexibility, potentially improving performance on complex datasets
- 1e-4 - higher weight decay encouraging sparsity, helping to prevent overfitting but potentially limiting model expressivity.

A balance between these three hyperparameters is crucial to obtain training stability and generalization.

5.5.1.4 Hyperparameter configurations tested

The following four configurations were evaluated using 5-fold cross-validation to determine a robust and consistent setup for each model in the ensemble (see Table 5.8).

Table 5.8: Hyperparameter configurations of learning rate, optimizer, and weight decay.

Learning rate	**Optimizer**	**Weight decay**
1e-4	AdamW	1e-5
1e-4	AdamW	1e-4
1e-3	SGD	1e-5
1e-3	SGD	1e-4

These configurations were carefully chosen to cover:

- Adaptive vs. momentum-based optimization (AdamW vs. SGD with momentum).
- Stable vs. more aggressive learning rates (1e-4 vs. 1e-3).
- Weaker vs. stronger regularization (1e-5 vs. 1e-4).

The choice to test only four hyperparameter configurations was dictated by the need to balance computational feasibility and methodological rigor. Each configuration was evaluated using 5-fold cross-validation, requiring multiple full training cycles for each model. Since a single 5-fold cross-validation run per model took over three days, performing an exhaustive hyperparameter search was not practical.

By selecting a small yet representative set of configurations, it was possible to systematically examine the influence of key hyperparameters while keeping training time reasonable. After identifying the best-performing configuration, it was applied consistently across all models to ensure that any performance differences were primarily due to architectural variations rather than discrepancies in hyperparameters.

5.5.2 Cross-validation

Cross-validation is a model evaluation technique where the dataset is split repeatedly into training and validation subsets. This method allows the model to be evaluated on different validation subsets while being trained on diverse portions of the data consequently, ensuring that the bias from class imbalances is reduced and providing a more robust evaluation of the generalization performance. Cross-validation was applied to obtain multiple independent estimates of

model performance for each hyperparameter configuration to address the class imbalance present in the dataset. The following is the step-by-step description of this process:

1. The dataset was divided into five stratified folds, maintaining the same distribution of region combinations in each fold.
2. In each iteration, one fold (20% of the data) was used for validation, while the remaining four folds (80% of the data) served as the training set.
3. This process was repeated five times, ensuring that each case was used for validation **exactly once** while being part of the training set in the remaining four iterations. The final data splits were saved as a JSON file, preserving fold assignments and patient-specific paths.

Figure 5.15 shows the distribution of training and validation samples per fold, grouped by region combination. Each row represents a cross-validation fold, while the bottom row highlights the dataset's stratification based on region combination.

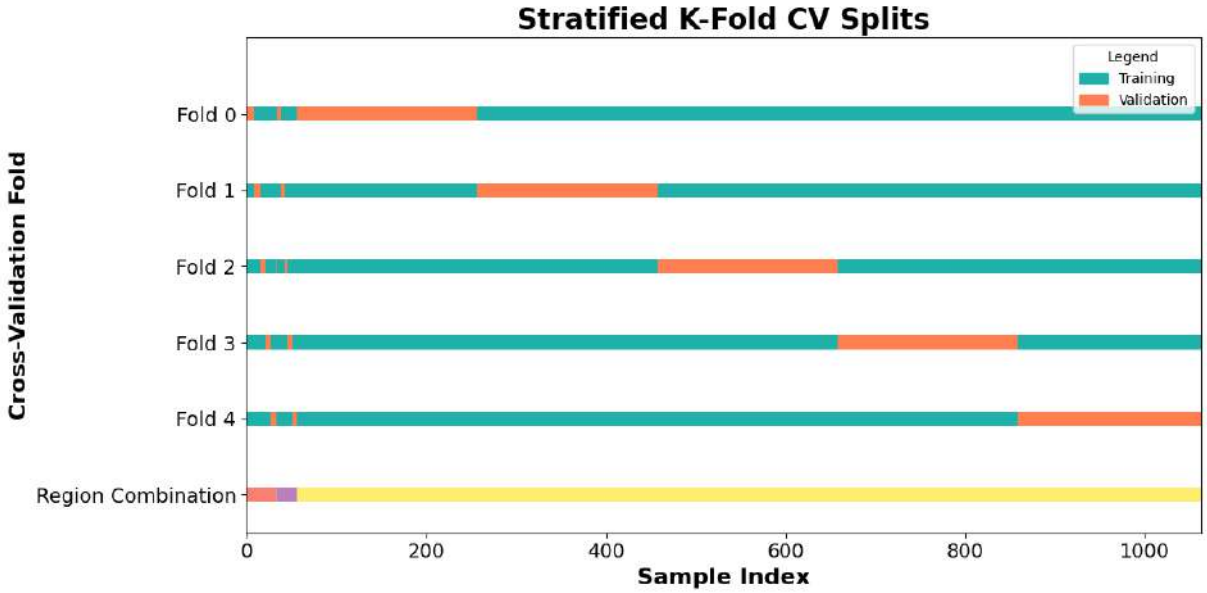


Figure 5.15: Visualization of the stratified 5-fold cross-validation splits. Each fold preserves the percentage of samples for each sub-region combination.

A sliding window inference strategy was used to process large 3D volumes of the validation set within the available GPU memory. Logging was performed using TensorBoard for each fold,

recording the tested hyperparameters as well as loss and performance metrics for each tumor sub-region. Moreover, early stopping and checkpointing mechanisms were utilized to reduce overfitting and retain the best-performing model for each fold. Finally, performance metrics (as described in Section 5.12) were aggregated across folds to provide reliable performance estimates.

Once the best hyperparameters were selected, the models were retrained on the full training data and subsequently evaluated before being incorporated into the ensemble. Figure 5.16 provides an overview of the entire **individual** model training pipeline - from hyperparameter tuning using cross-validation to final model integration into the ensemble.

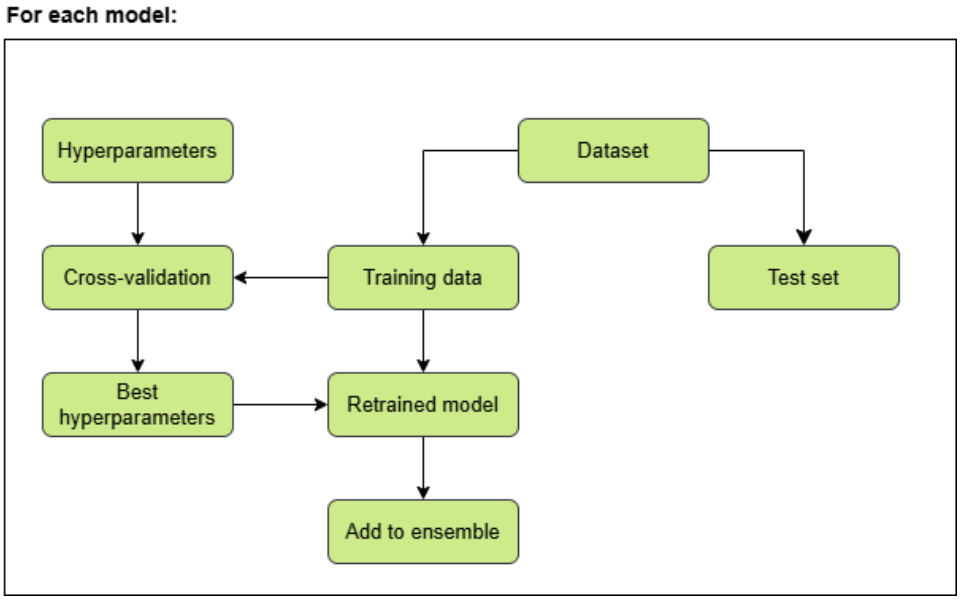


Figure 5.16: Overview of the individual model training pipeline. Each model undergoes hyperparameter tuning through cross-validation to determine the best configuration. The selected parameters are then used to train the model on the training data. Once trained, the model is added to the ensemble for the final prediction. The test set is reserved for the evaluation of both individual models and the final ensemble.

5.6 Uncertainty estimation

5.6.1 Overview of uncertainty estimation in this study

Uncertainty estimation techniques aim to provide an estimation of a deep learning model's confidence in its own predictions, allowing for a safer deployment of neural networks in safety-critical tasks such as medical image analysis [83]. In this work, uncertainty estimation is utilized

to improve model reliability and guide ensemble predictions in a brain tumor segmentation task.

The focus of this work is on epistemic uncertainty (model-dependent uncertainty due to, e.g., limited training data) and aleatoric uncertainty (data-dependent uncertainty resulting, e.g., from noise and data artifacts), which will be estimated using test-time dropout and test-time augmentation, respectively. Uncertainty is computed separately for each tumor sub-region: NCR, ED, and ET, resulting in individual uncertainty maps for each sub-region.

5.6.2 Uncertainty Estimation Methods

5.6.2.1 Test-time Dropout (TTD)

To estimate epistemic uncertainty, this thesis utilizes test-time dropout (TTD), following the Bayesian approximation method introduced by Gal and Ghahramani [88]. Unlike deterministic predictions, TTD enables dropout layers during inference, allowing the model to produce multiple stochastic outputs from a single input. By aggregating these outputs, voxel-wise uncertainty is estimated based on the variability in predictions.

Two key parameters in TTD are the dropout rate and the number of stochastic forward passes. Many works on uncertainty estimation in image segmentation have utilized dropout rates between 0.2 and 0.5 [88, 89, 118–120]. This thesis employs a dropout rate of 0.2 for all models, following Kendall and Gal [118], to balance capturing meaningful uncertainty with maintaining segmentation accuracy.

Previous research has used varying numbers of stochastic forward passes in TTD (ranging from 5 [89] to 50 [118]). Fewer passes reduce inference time but may provide noisy uncertainty estimates, whereas more passes yield more stable uncertainty maps at the cost of increased computation. Chlebus et al. [121] found that 20 passes offer a reasonable tradeoff. Hence, 20 stochastic forward passes are used in this study.

For each tumor sub-region, the uncertainty is modeled as the variance (V) of the predicted probability for each class over the stochastic forward passes:

$$V_c^i = \frac{1}{N} \sum_{n=1}^{N} \left(y_n^{i,c} - y_\mu^{i,c} \right)^2 \tag{5.5}$$

where:

- N is the number of stochastic forward passes;
- $y_n^{i,c}$ is the predicted probability for class c at voxel i during the n-th dropout-enabled forward pass;
- $y_\mu^{i,c}$ is the mean predicted probability for class c at voxel i, averaged over all passes.

Analyzing V produces an uncertainty map that highlights regions where the model is less confident in its segmentation.

5.6.2.2 Test-time Augmentation (TTA)

To model aleatoric uncertainty, test-time augmentation (TTA) was implemented by applying a set of data augmentations to the input MRI volumes during inference. By applying these augmentations, the model's sensitivity to inherent data noise (such as scanner variability or motion artifacts) can be assessed.

TTA reprocesses the same MRI scan multiple times with different augmentations, and the predictions are aggregated to quantify uncertainty. To ensure consistency with the training pipeline, all training augmentations were reused during TTA (with the exception of random cropping, as full MRI volumes are processed during inference). The augmentations applied in TTA and their role are summarized in Table 5.9.

Table 5.9: Augmentations used in TTA.

Augmentation	Purpose
Flipping	Ensures symmetry invariance.
Intensity Scaling	Simulates variability in scanner contrast settings.
Intensity Shifting	Simulates variability in overall brightness.

For each voxel i and for each class c, aleatoric uncertainty is computed as the variance of the predicted probability over multiple augmented inferences:

$$V_c^i = \frac{1}{N} \sum_{n=1}^{N} \left(y_n^{i,c} - y_\mu^{i,c} \right)^2 \tag{5.6}$$

where:

- N is the number of augmented inferences;
- $y_n^{i,c}$ is the predicted probability for class c at voxel i during the n-th augmentation;
- $y_\mu^{i,c}$ is the mean predicted probability for class c at voxel i, averaged over all augmented inferences.

Higher variance values indicate regions where the model is less confident. As with TTD, 20 augmented inferences are utilized.

5.7 Ensemble Strategy

Brain-tumor segmentation is challenging because of the heterogeneity of tumor sub-regions. A single model may not optimally segment all sub-regions due to differences in feature representation, tissue contrast, and patient variability. To address these challenges, three ensemble strategies were implemented:

1. **Simple Averaging Ensemble**: equal-weight averaging of model logits.
2. **Performance-Weighted Ensemble**: weighting by each model's validation performance per sub-region.
3. **Performance- and Uncertainty-Weigthed Ensemble**
    - **TTD-Weighted Ensemble**: weighting by inverse Test-Time Dropout (epistemic) uncertainty.
    - **TTA-Weighted Ensemble**: weighting by inverse Test-Time Augmentation (aleatoric) uncertainty.

- **Hybrid Uncertainty- and Performance-Weighted Ensemble**: combining inverse TTD + TTA weighting with performance weights.

The overall goal is to evaluate the impact of incorporating uncertainty estimates on segmentation quality, compared with simpler averaging and performance-based weightings. The hypothesis is that incorporating uncertainty will result in a more refined (and potentially more accurate) ensemble prediction, especially in the more challenging sub-regions like NCR.

5.7.1 Simple Averaging Ensemble

In this baseline strategy, each model outputs a logit volume $l_{k,s}^{i}$ for voxel i and tumor sub-region $s \in \mathsf{BG}, \mathsf{NCR}, \mathsf{ED}, \mathsf{ET}$. The logits are then fused by an element-wise arithmetic mean

$$\bar{l}_{s}^{i} = \frac{1}{K} \sum_{k=1}^{K} l_{k,s}^{i}. \tag{5.7}$$

Unlike averaging post-soft-max scores, averaging logits preserves the relative confidence gaps between classes and as a result [122]. The predicted label is obtained with a single soft-max followed by *argmax*:

$$\hat{y}^{i} = \underset{s}{\operatorname{argmax}} \big(\operatorname{softmax}(\bar{l}^{\cdot i})\big)_{s}, \quad \bar{l}^{\cdot i} = \big[\bar{l}_{\mathsf{BG}}^{i}, \bar{l}_{\mathsf{NCR}}^{i}, \bar{l}_{\mathsf{ED}}^{i}, \bar{l}_{\mathsf{ET}}^{i}\big]. \tag{5.8}$$

5.7.2 Performance-Weighted Ensemble

Validation performance for each model k and sub-region s is summarised by a composite score

$$C_{k}^{s} = w_{\text{Dice}} D_{k}^{s} + w_{\text{HD95}} \frac{1}{1 + \text{HD95}_{k}^{s}} + w_{\text{Sens}} \text{Sens}_{k}^{s} + w_{\text{Spec}} \text{Spec}_{k}^{s}, \tag{5.9}$$

with $(w_{\text{Dice}}, w_{\text{HD95}}, w_{\text{Sens}}, w_{\text{Spec}}) = (0.45, 0.15, 0.30, 0.10)$. The four weights reflect the clinical emphasis on overlap and sensitivity, while HD95 measures boundary accuracy and specificity penalises false positives.

Scores are normalised per region to give

$$\tilde{w}_{k,s} = \frac{C_k^s}{\sum_{j=1}^{K} C_j^s}, \qquad \sum_{k=1}^{K} \tilde{w}_{k,s} = 1; \forall s. \tag{5.10}$$

Fused logits and the final label are then

$$\bar{l}_s^i = \sum_{k=1}^{K} \tilde{w}_{k,s} l_{k,s}^i, \qquad \hat{y}^i = \underset{s}{\operatorname{argmax}} \big(\operatorname{softmax}(\bar{l}^{\,i})\big)_s. \tag{5.11}$$

The same region-wise weights $\tilde{w}_{k,s}$ are reused by the uncertainty-aware ensembles below.

5.7.3 Performance- and Uncertainty-Weighted Ensemble

Region-wise performance weights (Section 5.7.2) are combined with per-voxel uncertainty estimates obtained via Monte Carlo sampling. At test time, N stochastic forward passes under dropout and M stochastic augmentations yield, for each model k, voxel i, and sub-region s:

$$V_{k,\mathrm{TTD}}^{i,s} = \mathrm{Var}\{\, l_{k,s}^{(n,i)}\}_{n=1}^{N}, \qquad V_{k,\mathrm{TTA}}^{i,s} = \mathrm{Var}\{\, l_{k,s}^{[m,i]}\}_{m=1}^{M},$$

$$\mu_k^{i,s} = \frac{1}{N+M}\Big(\sum_{n=1}^{N} l_{k,s}^{(n,i)} + \sum_{m=1}^{M} l_{k,s}^{[m,i]}\Big).$$

Here $\{l_{k,s}^{(n,i)}\}$ are logits under Test-Time Dropout (TTD) and $\{l_{k,s}^{[m,i]}\}$ are logits under Test-Time Augmentation (TTA). The common mean $\mu_k^{i,s}$ represents the averaged prediction.

5.7.3.1 Voxel-wise Uncertainty Adjustment

Each mean logit is down-weighted by the inverse of its variance(s):

$$\tilde{l}_k^{i,s} = \mu_k^{i,s} \big(V_{k,\mathrm{TTD}}^{i,s} + \varepsilon\big)^{-\mathbf{1}_{\mathrm{TTD}}} \big(V_{k,\mathrm{TTA}}^{i,s} + \varepsilon\big)^{-\mathbf{1}_{\mathrm{TTA}}}, \quad \varepsilon = 10^{-6}. \tag{5.12}$$

The indicator exponents $(\mathbf{1}_{\mathrm{TTD}}, \mathbf{1}_{\mathrm{TTA}})$ specify which uncertainty source is active. When

equal to 1 they include the corresponding inverse-variance term, and when 0 they omit it.

Table 5.10: Indicator settings for each ensemble variant.

Variant	$\mathbf{1}_{\mathrm{TTD}}$	$\mathbf{1}_{\mathrm{TTA}}$
TTD-Only	1	0
TTA-Only	0	1
Hybrid	1	1

5.7.3.2 Final Fusion

Adjusted logits are fused using region-wise weights $\tilde{w}_{k,s}$ (Section 5.7.2):

$$\bar{l}_s^{\,i} = \sum_{k=1}^{K} \tilde{w}_{k,s}\, \tilde{l}_k^{\,i,s}, \qquad \hat{y}^{\,i} = \arg\max_s \big(\mathrm{softmax}(\bar{\mathbf{l}}^{\,i})\big)_s.$$

A single soft-max across the fused logit vector ensures a valid probability distribution.

5.7.3.3 Uncertainty Maps

Hybrid (TTD+TTA).

Combined uncertainty at voxel i, class s is computed as

$$U_k^{i,s} = \tfrac{1}{2}\big(V_{k,\mathrm{TTD}}^{\,i,s} + V_{k,\mathrm{TTA}}^{\,i,s}\big),$$

then fused by

$$\bar{U}^{\,i,s} = \sum_{k=1}^{K} \tilde{w}_{k,s}\, U_k^{i,s},$$

and finally min–max normalised to $[0, 1]$.

TTD-Only and TTA-Only.

For single-source variants the unused term is omitted:

$$U_k^{i,s} = \begin{cases} V_{k,\mathrm{TTD}}^{i,s}, & \text{TTD-Only}, \\ V_{k,\mathrm{TTA}}^{i,s}, & \text{TTA-Only}, \end{cases}$$

followed by the same weighted fusion and normalisation.

5.8 Calibration of ensemble probabilities

Preliminary experiments showed that raw ensemble softmax scores are often over-confident, i.e. pushed towards 0 or 1. Temperature scaling [123] is therefore applied to all probability maps.

A dedicated calibration split containing 20 volumes (drawn from the validation set and thus, not used for subsequent performance evaluation) is used to fit T. The optimisation minimises the negative log-likelihood with the L-BFGS algorithm (learning-rate 0.1, `max_iter` = 30), starting from $T = 1.0$. One temperature is learned per ensemble and afterwards kept fixed for all test volumes. This procedure preserves the arg-max label map but produces probability values that better match empirical correctness frequencies.

Ultimately, the final calibrated probability map can be defined as follows:

$$\hat{p}_{\mathrm{calib}}(x)_s \;=\; \frac{\exp\big(l_s(x)/T\big)}{\sum_j \exp\big(l_j(x)/T\big)},$$

where $l_s(x)$ denotes the ensemble logit for class s and $T > 0$ is a scalar temperature. Dividing by $T > 1$ softens the distribution and lowers over-confidence; $T < 1$ would have the opposite effect.

5.8.1 Expected calibration error

Calibration quality is measured with the *Expected Calibration Error* (ECE)[123, 124]. A model is perfectly confidence-calibrated if:

$$\mathbb{P}\big(Y = \arg\max \hat{p}(X) \;\big|\; \max \hat{p}(X) = c\big) = c, \quad \forall\, c \in [0,1],$$

where X is an image, Y its true label and $\hat{p}(X) \in \Delta^K$ the predicted class distribution, and $c \in [0, 1]$ is a confidence level.

In practice, the interval $[0, 1]$ is partitioned into M equally sized confidence bins $\{B_m\}_{m=1}^{M}$. Let n be the total number of samples (voxels). Then:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|,$$

with

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}\{\hat{y}_i = y_i\}, \qquad \mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \max \hat{p}(x_i).$$

where:

- M is the number of confidence bins;
- n is the total number of samples (voxels) evaluated;
- B_m is the set of indices whose maximum predicted confidence $\max \hat{p}(x_i)$ falls in the m-th bin;
- $|B_m|$ is the number of samples in bin B_m;
- $\max \hat{p}(x_i)$ is the maximum predicted probability for sample i;
- $\hat{y}_i$ is the predicted label for sample i;
- $\mathrm{acc}(B_m)$ is the empirical accuracy in bin m;
- $\mathrm{conf}(B_m)$ is the average predicted confidence in bin m.

ECE is reported for each tumor sub-region (NCR, ED, ET) because the background class outnumbers tumor voxels by roughly two orders of magnitude and would otherwise dominate the statistic.

Uncertainty maps produced by test-time dropout (TTD) or test-time augmentation (TTA) represent predictive variance rather than class probabilities and therefore do not meet the assumptions of ECE. These maps are analysed qualitatively in Section 5.9.

5.8.2 Reliability Diagrams

Calibrated probabilities $\hat{p}_{\text{calib}}(x_i)$ are partitioned into M equal-width bins B_m based on $\max_s \hat{p}_{\text{calib}}(x_i)$. For each bin:

$$\text{acc}_m = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}\{\arg\max_s \hat{p}_{\text{calib}}(x_i) = y_i\}, \quad \text{conf}_m = \frac{1}{|B_m|} \sum_{i \in B_m} \max_s \hat{p}_{\text{calib}}(x_i).$$

Plotting acc_m versus conf_m (with the diagonal $y = x$ as reference) yields the reliability curve. Bin counts are optionally annotated at each point.

In this work, reliability diagrams are generated separately for each sub-region (NCR, ED, ET). The calibration curve and the diagonal line provide a visual complement to the numeric ECE metric (Section 5.8.1).

5.9 Evaluation of uncertainty estimates

5.9.1 Uncertainty–error correlation analysis

Inspired by Mehrtash et al. [125], who demonstrated a strong Pearson correlation between segment-level average entropy and Dice score using binned-error plots, this thesis instead computes a voxel-wise Spearman correlation between NLL-based error and variance-based uncertainty maps for each tumor subregion. Here, error was measured using negative log-likelihood (NLL) derived from softmax probabilities, and uncertainty was extracted from variance-based maps for each subregion (NCR, ED, ET). Percentile-based uncertainty bins are then used to track how median error varies across uncertainty levels.

For each voxel in the tumor mask:

- Uncertainty was extracted from the respective TTA, TTD, or Hybrid uncertainty map.
- Error was computed as the negative logarithm of the predicted probability for the ground truth label:

$$\text{Error} = -\log(\hat{p}_{\text{true}}(x_i) + \varepsilon) \tag{5.13}$$

where $\varepsilon = 10^{-8}$ ensures numerical stability.

The resulting correlation coefficients indicate the degree to which higher uncertainty is associated with greater prediction error. A positive and statistically significant correlation suggests that the uncertainty estimate is informative for identifying unreliable predictions.

5.9.2 Risk–coverage analysis

To evaluate the practical utility of uncertainty maps for selective prediction, risk–coverage curves were computed as described in [126, 127]. These curves plot the average prediction error over subsets of voxels retained after excluding those with the highest uncertainty, providing insight into the usefulness of uncertainty for risk-aware segmentation.

Formally, let $C \in [0, 1]$ denote the coverage fraction (percentage of voxels with the lowest uncertainty retained), and let $R(C)$ be the average NLL of those voxels. A desirable curve is one where:

$$\frac{dR(C)}{dC} > 0 \tag{5.14}$$

i.e., average error increases with more uncertain voxels [126]. This suggests that the uncertainty metric can be used for risk-aware voxel filtering or triage.

Risk–coverage curves were computed for each sub-region using the same uncertainty and error maps as in the correlation analysis. Methods with steeper upward curves are considered more suitable for clinical decision support, where high-certainty predictions are prioritized.

5.10 Overview of the Ensemble Fusion Pipeline

Figure 5.17 provides an overview of the ensemble fusion pipeline employed in this thesis. This pipeline illustrates how multiple deep learning models contribute to the final segmentation through three distinct fusion strategies.

The pipeline proceeds as follows:

1. **Model Training and Inference:**

   - Four deep learning architectures (SegResNet, SwinUNETR, Attention UNet, and VNet) are tuned and trained individually (see Section 5.5 and Figure 5.16) on multi-modal MRI scans.
   - Each model outputs logits corresponding to the background and three tumor sub-regions: NCR, ED, and ET.

2. **Ensemble Fusion Strategies:** The pipeline supports three fusion strategies that differ only in the method used to aggregate the model predictions:

   - **Simple Averaging Ensemble:** The logit predictions from all models are averaged equally.
   - **Performance-Weighted Ensemble:** Each model's logits are scaled by a global performance weight derived from validation metrics (Dice, HD95, Sensitivity, and Specificity) for each tumor sub-region prior to averaging.
   - **Performance and Uncertainty-Weighted Ensemble:** In this approach, the fusion is extended by first adjusting each model's base prediction using voxel-level uncertainty estimates. Specifically, the TTD mean prediction is modulated by inverse uncertainty weights computed from TTD and/or TTA estimates (with three variants: TTD-Only, TTA-Only, and a hybrid TTD+TTA combination). The uncertainty-adjusted predictions are then fused using global performance weights.

3. **Probability Map Generation and Final Segmentation:**

   - Following fusion, the ensembled logits undergo a softmax transformation to generate voxel-wise probability maps.
   - Finally, the segmentation for each voxel is determined by applying an argmax operation over the probability maps.

4. **Ensemble Uncertainty Estimation:**

   - For the performance and uncertainty-weighted ensemble, the uncertainty maps from

individual models are also aggregated—using a weighted average based on the individual model contributions—to produce a final ensemble uncertainty map.

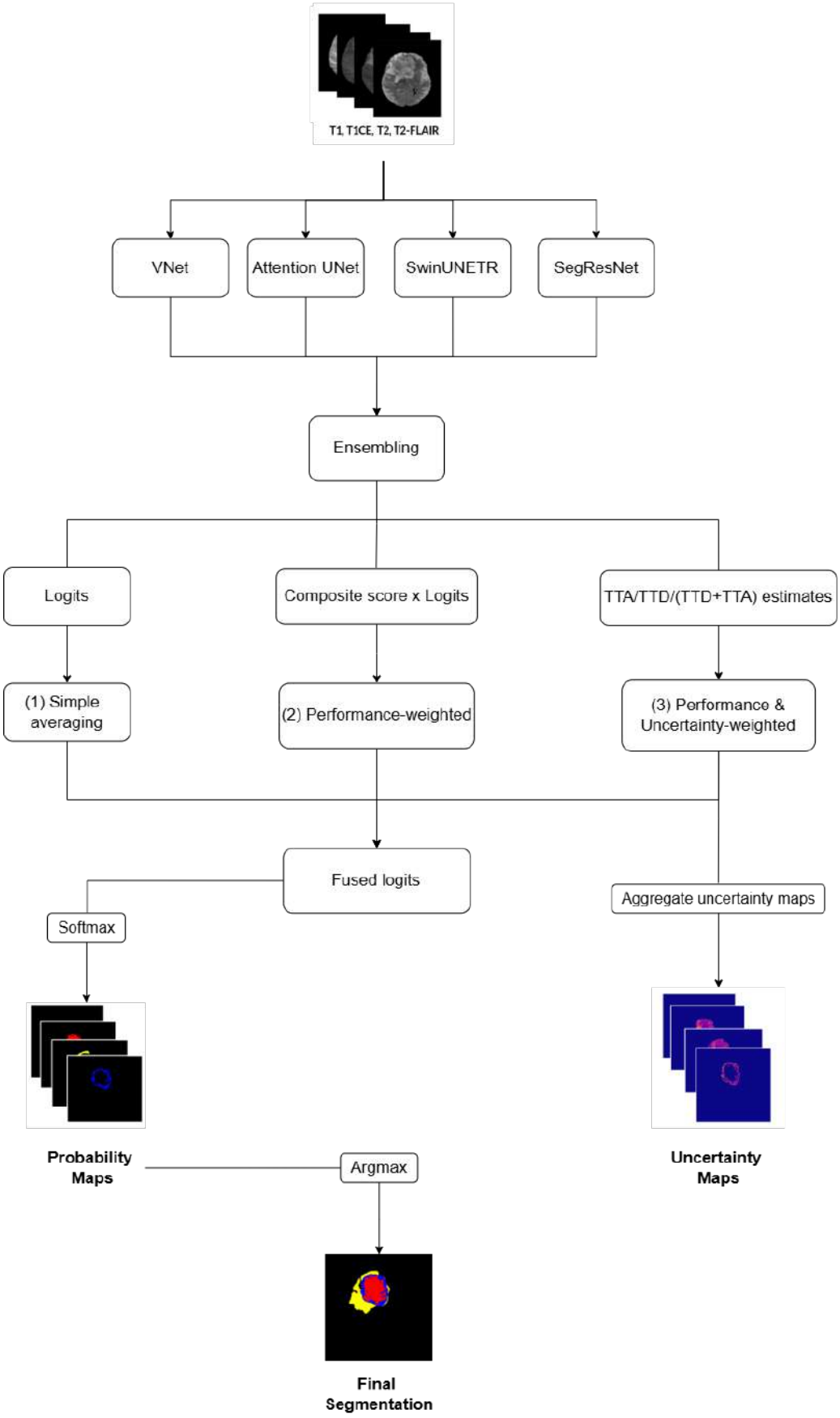


Figure 5.17: Overview of the ensemble fusion pipeline. Individual deep learning models generate segmentation logits, which are then fused using one of three strategies: simple averaging, performance-weighted averaging, or performance and uncertainty-weighted averaging.

5.11 Computational resources

The entire project was implemented in Python 3.10. The models were implemented using a `MONAI` [106] library. Training and testing were implemented using the `PyTorch` [128] library.

All training and evaluation experiments were conducted on a high-performance GPU cluster equipped with two NVIDIA A100 GPUs (each with 24GB VRAM). The models were trained in parallel, with each model assigned to a separate GPU. This setup ensured efficient resource utilization, reducing overall training time.

The key GPU specifications are:

- GPU Model: NVIDIA A100
- Number of GPUs Used: 2 (each model trained on a separate GPU)
- VRAM per GPU: 24GB
- CPU Allocation: 2 CPU cores per job
- Driver Version: 525.116.04
- CUDA Version: 12.0

Two models were trained simultaneously on separate GPUs to maximize throughput. Mixed precision training was enabled via Torch's AMP (Automatic Mixed Precision) to optimize memory usage and speed up training. To optimize performance on NVIDIA GPUs, cuDNN benchmarking was enabled. This setting allows `PyTorch` to select the most efficient convolution algorithms, leading to faster training times.

The training jobs were submitted using the Slurm workload manager. A detailed example of a Slurm job script used for hyperparameter tuning is provided in Appendix A.1.

5.12 Evaluation metrics

Performance evaluation in the BraTS 2021 Adult Glioma challenge was primarily carried out using the Dice Similarity Coefficient and Hausdorff distance at the 95th percentile [37]. Additionally, sensitivity and specificity metrics are provided for participant guidance to determine potential over- or under-segmentations of the tumor sub-regions [37]. However, they did not impact the ranking in the challenge [37]. Similarly, and for consistency, all four metrics were used in the reported experiments to assess model performance. The following sections provide a detailed explanation of each metric and its relevance to brain tumor segmentation.

5.12.1 Dice Similarity Coefficient

The **Dice Similarity Coefficient (DSC)** serves as the primary metric for assessing spatial overlap accuracy [129]. It is commonly used in evaluating the performance of segmentation algorithms, particularly in medical image analysis. It quantifies the spatial overlap between a predicted segmentation and a ground truth annotation, making it highly relevant for assessing the accuracy of brain tumor segmentation models. The DSC ranges from 0 to 1, where 1 signifies a perfect segmentation, while 0 indicates no overlap at all, meaning that the segmentation completely failed.

$$DSC = \frac{2|A \cap B|}{|A| + |B|} \tag{5.15}$$

where:

- A represents the **ground truth** segmentation mask;
- B represents the **predicted segmentation** mask;
- $|A \cap B|$ denotes **the number of overlapping voxels** between A and B;
- $|A|$ and $|B|$ are **the total number of voxels** in the ground truth and predicted masks, respectively.

In brain tumor segmentation, the Dice coefficient is typically computed for different tumor sub-regions (NCR, ED, ET). Each sub-region is evaluated separately to ensure that the segmentation model accurately detects different tumor tissues.

5.12.2 Hausdorff distance 95%

5.12.2.1 Definition of Hausdorff Distance

The **Hausdorff Distance (HD)** is a commonly used metric in medical image segmentation that measures the worst-case boundary discrepancy between two segmentations [130]. It quantifies the largest distance from a point in one segmentation to the closest point in the other segmentation, reflecting boundary accuracy.

The standard HD between two sets A (ground truth) and B (predicted segmentation) is defined as [130]:

$$d_H(A, B) := \max \left\{ \sup_{a \in A} d(a, B), \sup_{b \in B} d(A, b) \right\}, \tag{5.16}$$

where:

- $d(a, B)$: The **distance** from a point $a \in A$ to the closest point in set B;
- $d(A, b)$: The **distance** from a point $b \in B$ to the closest point in set A;
- $\sup_{a \in A} d(a, B)$: The **supremum** of all minimum distances from points in A to B;
- $\sup_{b \in B} d(A, b)$: The **supremum** of all minimum distances from points in B to A;
- max: The maximum operator ensures that the Hausdorff Distance captures the **worst-case boundary mismatch** between A and B.

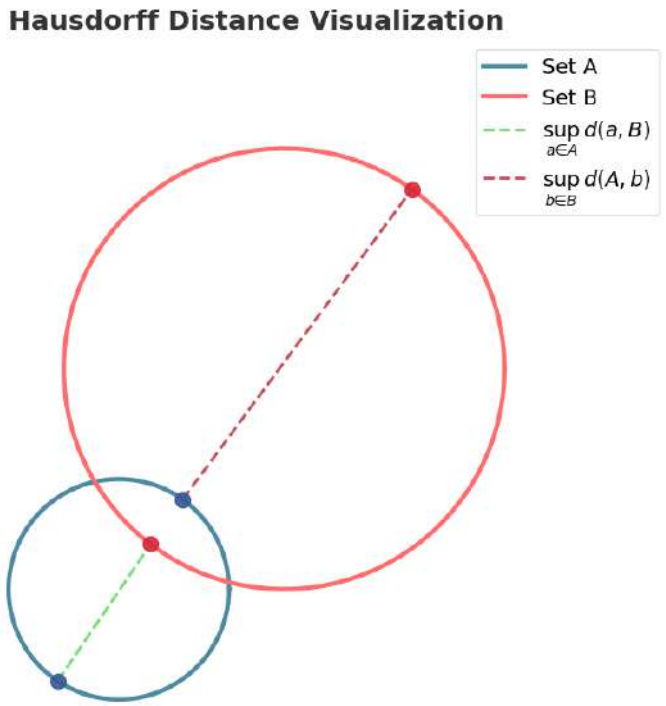


Figure 5.18: Graphical representation of the Hausdorff distance, where the maximum distances from one region to the other are highlighted.

5.12.2.2 95^{th} percentile Hausdorff Distance

Since the HD can be sensitive to outliers, the **95^{th} percentile Hausdorff Distance (HD95)** is often used instead. HD95 is defined as:

$$HD_{95}(A, B) = P_{95}\left(d(A, B)\right), \tag{5.17}$$

where P_{95} represents the 95^{th} percentile of all computed distances. By focusing on the distances that encompass the majority of the data (95% of the distances) the impact of extreme outliers is reduced.

The HD95 is always non-negative and ranges from 0 (indicating perfect boundary alignment) to a value proportional to the image size or the extent of misalignment. In medical image

segmentation, smaller HD95 values indicate better boundary consistency between the ground truth and prediction.

5.12.2.3 Limitations of HD95 in tumor segmentation

One limitation of HD95 arises when a tumor sub-region (e.g., NCR) is either absent in the ground truth but predicted by the model or present in the ground truth but not predicted by the model. In the first case (false positives), the HD becomes undefined or infinite, making it difficult to compute meaningful averages. In the second case (false negatives), the model fails to generate any segmentation, leading to the opposite issue: a distance that cannot be computed due to the absence of a prediction.

Some prior studies have addressed this issue by replacing infinity with an arbitrarily large constant [93]. However, this approach fails to account for the spatial extent of the missing regions, treating all misclassified regions equally, regardless of their size.

5.12.2.4 Proposed correction for undefined HD95

To ensure a geometrically meaningful HD95 value in these cases, I propose computing the HD95 between the available region (either predicted or ground truth) and its own center of mass (CoM). This method effectively creates a virtual reference point inside the segmented region, ensuring that the computed HD95 reflects the spatial extent of the missing structure rather than defaulting to an arbitrarily large threshold.

This method is applied differently depending on whether the missing structure is in the ground truth (false negative) or the prediction (false positive):

Case 1: False Positives (ground truth absent, prediction present) If a tumor sub-region is not present in the ground truth but is predicted, the correction is applied as follows:

1. Compute the CoM of the predicted segmentation mask P:

$$\mathbf{CoM}(P) = \frac{\sum_{a \in P} \mathbf{r_a}}{|P|} \tag{5.18}$$

where:

- $|P|$ denotes the number of voxels in the predicted mask
- $\mathbf{r_a} = (x_a, y_a, y_c)$ is the position vector of point a.

2. Construct a reference mask R containing a single voxel at the computed CoM location:

$$R(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} = \mathbf{CoM}(P) \\ 0, & \text{otherwise} \end{cases} \tag{5.19}$$

3. Compute the corrected HD95 as the HD between the predicted mask P and the reference mask R:

$$HD_{95}^{\text{CoM}}(P) = HD_{95}(P, R) \tag{5.20}$$

Case 2: False Negatives (ground truth present, prediction absent) If a tumor subregion exists in the ground truth but is not predicted, the correction is applied similarly, but with respect to the ground truth:

1. Compute the CoM of the predicted segmentation mask P:

$$\mathbf{CoM}(G) = \frac{\sum\limits_{a \in G} \mathbf{r_a}}{|G|} \tag{5.21}$$

where:

- $|G|$ denotes the number of voxels in the ground truth mask
- $\mathbf{r_a} = (x_a, y_a, y_c)$ is the position vector of point a.

2. Construct a reference mask R containing a single voxel at the computed CoM location:

$$R(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} = \mathbf{CoM}(G) \\ 0, & \text{otherwise} \end{cases} \tag{5.22}$$

3. Compute the corrected HD95 as the HD between the predicted mask P and the reference mask R:

$$HD_{95}^{\text{CoM}}(G) = HD_{95}(G, R) \tag{5.23}$$

This approach ensures that HD95 remains well-defined in both cases, reflecting the spatial extent of the missing structure rather than defaulting to an arbitrary value. In the case of false positives, it penalizes large over-segmentations more than small ones. In the case of false negatives, it ensures that completely missing structures are penalized based on their expected size and location.

5.12.3 Sensitivity

Sensitivity, also known as the true positive rate (TPR) or recall, is a crucial evaluation metric in medical image segmentation. It measures the model's ability to correctly identify positive cases (i.e., tumor regions) while minimizing false negatives. Mathematically, sensitivity is defined as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{5.24}$$

where:

- TP (True Positives): The number of correctly segmented tumor voxels;
- FN (False Negatives): The number of tumor voxels incorrectly labeled as background.

Sensitivity ranges from 0 to 1, where high sensitivity values indicate that the segmentation model effectively detects most tumor regions, reducing the risk of missing critical structures. Once again, in brain tumor segmentation, sensitivity is evaluated across the different tumor sub-regions.

5.12.4 Specificity

Specificity, also known as the true negative rate (TNR), is another important evaluation metric in medical image segmentation. It measures the model's ability to correctly identify negative cases (i.e., non-tumor regions) while minimizing false positives. Mathematically, specificity is defined as:

$$Specificity = \frac{TN}{TN + FP} \tag{5.25}$$

where:

- TN (True Negatives): The number of correctly classified non-tumor voxels (i.e., healthy brain tissue);
- FP (False Positives): The number of non-tumor voxels that were incorrectly labeled as tumor.

A high specificity value indicates that the model effectively avoids over-segmenting tumor regions, reducing the risk of misclassifying normal brain tissue as a tumor.

5.13 Interactive segmentation application implementation

To bridge the gap between advanced segmentation algorithms and clinical usability, a web-based interactive application was developed using Streamlit [131]. The application supports a full segmentation pipeline — from uploading and preprocessing MRI scans to executing ensemble segmentation and visualizing results including uncertainty maps.

5.13.1 Design Objectives

The main objectives of the application were:

1. Provide an **intuitive workflow** for uploading and preprocessing MRI scans in `DICOM` or `NIfTI` format.
2. Enable **execution of an ensemble segmentation model** with softmax- and uncertainty-based outputs.
3. Offer **interactive visualizations**, including slice-by-slice overlays for segmentation, probability, and uncertainty.
4. Compute and display **tumor volume estimates**.

5.13.2 Requirements

The application was designed to meet specific functional and non-functional requirements, which are detailed in Tables 5.11 and 5.12.

5.13.2.1 Functional requirements

Functional requirements express the functions that a (software) system or system element must be able to perform, where a function is defined as a summary of behaviour between inputs and outputs [132]. In short, functional requirements describe *what* the system must be able to do. Table 5.11 showcases 10 functional requirements defined for the clinical toolbox prototype.

Table 5.11: Functional Requirements of the Application

ID	Requirement Description
FR1	The system shall accept up to 4 MRI scans in `DICOM` or `NIfTI` format.
FR2	The system shall convert `DICOM` files to `NIfTI` format automatically.
FR3	The system shall allow optional preprocessing steps including skull stripping and image registration.
FR4	The system shall reorder modalities to match the expected input order (Flair, T1ce, T1, T2).
FR5	The system shall execute a deep learning ensemble model for tumor segmentation.
FR6	The system shall output segmentation masks, softmax probability maps, and voxel-wise uncertainty maps.
FR7	The system shall allow interactive slice-by-slice visualization with adjustable overlays.
FR8	The system shall compute and display tumor subregion volumes in cubic centimeters.
FR9	The system shall allow users to download thresholded uncertainty maps and segmentation figures.
FR10	The system shall output performance metrics if the user provides the ground truth segmentation to compare the predicted segmentation.

5.13.2.2 Non-functional requirements

Non-functional requirements, on the other hand, are specifications used to judge a quality attribute of the system such as performance, security, or scalability [133]. Table 5.12 shows the

non-functional requirements defined for the system developed in this thesis.

Table 5.12: Non-Functional Requirements of the Application

ID	Requirement Description
NFR1	The interface must be intuitive and usable by clinicians or researchers without technical expertise.
NFR2	The application requires access to a GPU for model inference and must be run on hardware with CUDA-compatible support.
NFR3	The system should return segmentation results within five minutes for typical scan sizes.
NFR4	The application must support modular integration of alternative models and postprocessing steps.
NFR5	The system must process data locally to ensure privacy and compliance with medical data regulations.
NFR6	The codebase must be structured in a modular and maintainable way.
NFR7	The system must provide clear error messages for file format or runtime issues.
NFR8	The design must support future extensions such as batch processing or clinical data integration.

5.13.3 Sequence diagram

A UML sequence diagram is a behavioral diagram that models the interactions among system components (objects or actors) over time. It presents a set of lifelines (the vertical lines) representing participants, and message arrows between them to show the order and content of exchanges. In essence, it lays out the temporal sequence of operations and the lifelines of each component, making it ideal for capturing end-to-end workflows and control flow across modules [134].

Figure 5.19 presents the complete clinical toolbox pipeline as a sequence diagram. The vertical lifelines represent the User, the Streamlit App, the File Upload Module, the Preprocessing Pipeline, the Ensemble Inference engine, the Output Metrics calculator, and the Visualization Layer. Numbered arrows trace the key steps: from uploading DICOM/NIfTI scans, through format detection and skull-stripping, to sliding-window segmentation, uncertainty estimation, metric computation (e.g., Dice, HD95), and finally interactive slice overlays and downloadable reports.

The diagram aims to clarify the data flow and the interactions between the different components of the app, illustrating how user inputs propagate through the system .

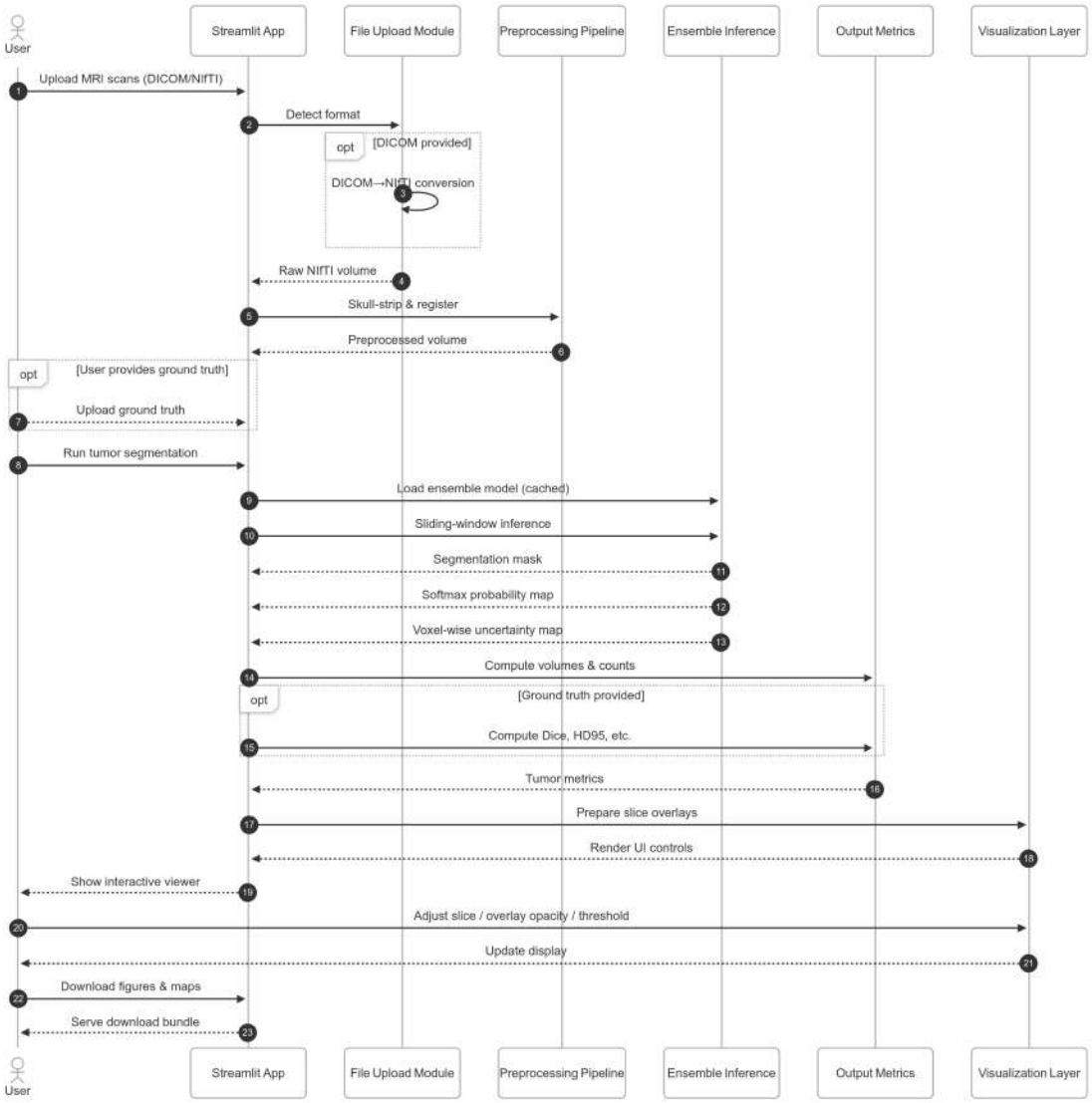


Figure 5.19: Sequence diagram

5.13.4 Implementation details

5.13.4.1 File upload and preprocessing

NiBabel library handles reading of `DICOM` or `NIfTI` inputs and conversion of `DICOM` series into a single `NIfTI` volume. Uploaded files are matched to modalities (FLAIR, T1CE, T1, T2) via filename patterns and any ambiguous cases trigger a user warning.

In the preprocessing pipeline, skull-stripping uses HD-BET [135], and spatial registration/reslicing is implemented using SimpleITK [136]. Both steps are cached in Streamlit via '@st.cache' to eliminate redundant computation within a session.

5.13.4.2 Segmentation execution

The MONAI ensemble model is loaded once per session (cached), then inference is performed using MONAI's sliding-window engine. Outputs (logits, probability maps, TTD/TTA uncertainty maps) are stored in session state for downstream visualization and analysis.

5.13.4.3 Visualization and output analysis

Streamlit sliders and checkbox widgets enable:

- **Slice navigation:** Users can traverse 3D volumes via sliders that display 2D anatomical slices.
- **Overlay controls:** Users can toggle and adjust opacity for segmentation masks, probability maps, and uncertainty overlays.
- **Uncertainty thresholding:** Users can filter the uncertainty maps based on voxel-wise thresholds and download the resulting maps.

Volumetric measurements (in cm^3) for each sub-region and performance metrics (Dice, HD95) are computed on demand and displayed in the Results tab if the ground truth is computed.

5.13.5 System Specifications

The web application was developed and tested on a local machine with the following configuration:

- **GPU:** NVIDIA GeForce RTX 4060 Laptop GPU (8 GB VRAM)

- **Driver Version:** 561.17
- **CUDA Version:** 12.6
- **Python:** 3.10
- **Operating System:** Ubuntu 22.04
- **Deep Learning Libraries:**
  - PyTorch (GPU build)
  - MONAI
  - HD-BET
  - SimpleITK
  - Streamlit

Model inference requires a CUDA-compatible GPU. The segmentation pipeline, particularly when using 3D models, is **not** optimized for CPU-only environments due to memory and performance constraints.

Chapter 6

Experiments and Results

6.1 Model training and cross-validation performance

In order to ensure that the differences in performance across the models are primarily due to their architectures and not sub-optimal hyperparameters, tuning of hyperparameters was performed. Three hyperparameters - learning rate (LR), optimizer, and weight decay (WD) - were tested, as they play a crucial role in model convergence, stability, and generalization. Four different configurations of these hyperparameters were tested using a 5-fold cross-validation approach, which helps mitigate biases due to dataset variability and ensures robust performance evaluation. Each fold consisted of 30 epochs and a batch size of one.

The following sections analyze the performance of each model per each configuration tested and determine the best configuration for later re-training and final evaluation of the models. Four evaluation metrics were analyzed - Dice score, HD95, sensitivity, and specificity. These metrics were evaluated at both the sub-region level (individual tumor sub-regions) and the global level (entire tumor). The global values of the metrics were computed by aggregating the scores over all predicted tumor regions and recalculating them globally. This approach prevents overly optimistic performance estimations, as errors in one region cannot be fully compensated by higher scores in another, leading to a more realistic and reliable evaluation.

6.1.1 Hyperparameter tuning results

Across all models, configurations using the AdamW optimizer with a learning rate of 0.0001 consistently provided the most balanced and robust segmentation. These setups achieved the highest Dice scores and sensitivity, particularly in the enhancing tumor (ET) sub-region, while also maintaining competitive boundary delineation as indicated by HD95.

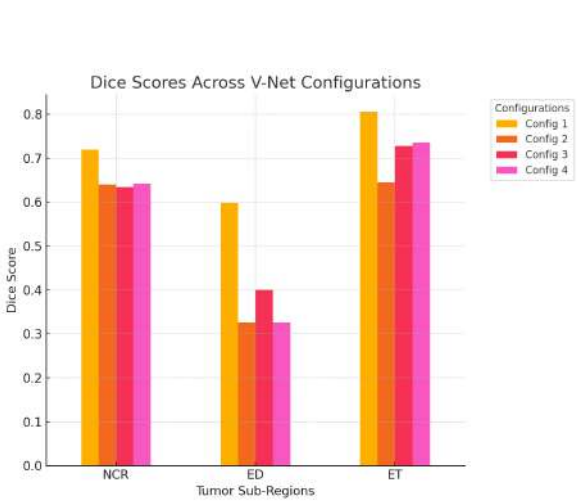
- For V-Net, Configuration 1 (AdamW, LR=0.0001, WD=1e-5) yielded the best performance, especially in ET segmentation and overall sensitivity. Although Configuration 4 (SGD-based) achieved lower HD95 in some regions, it lagged in volumetric accuracy. The full results are summarized in Appendix A.2.1.

- SegResNet showed similarly strong results with AdamW. Configuration 1 had the highest Dice and sensitivity, while Configuration 2 offered slightly improved HD95. ET remained the easiest region to segment, while NCR showed the most variability.

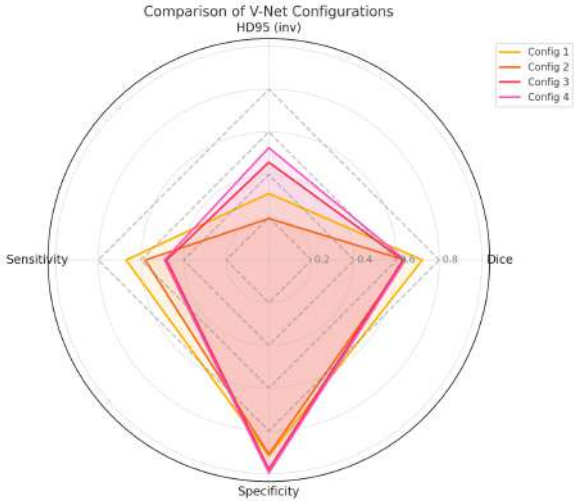
- For Attention UNet, Configuration 2 achieved the best average Dice score, while Configuration 1 had slightly better HD95. Both AdamW-based setups substantially outperformed the SGD ones across all metrics.

- SwinUNETR followed the same trend. Configuration 2 achieved the highest Dice and lowest HD95 across most sub-regions, confirming the strength of the AdamW optimizer and lower LR.

Figures 6.1 through 6.4 illustrate the cross-validation results for each model, with Dice scores and radar charts shown side by side for visual comparison. Complete quantitative tables are available in Appendix A.2.

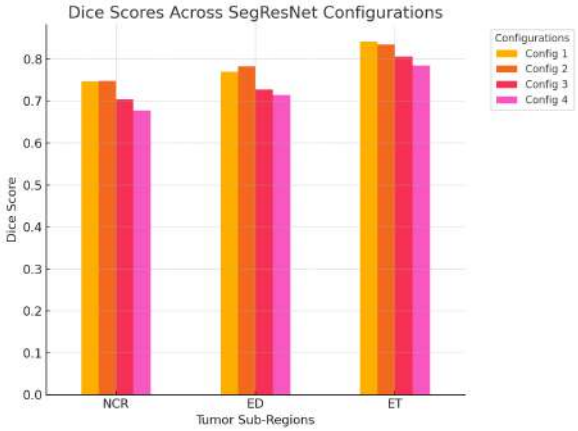


(a) Dice scores by sub-region

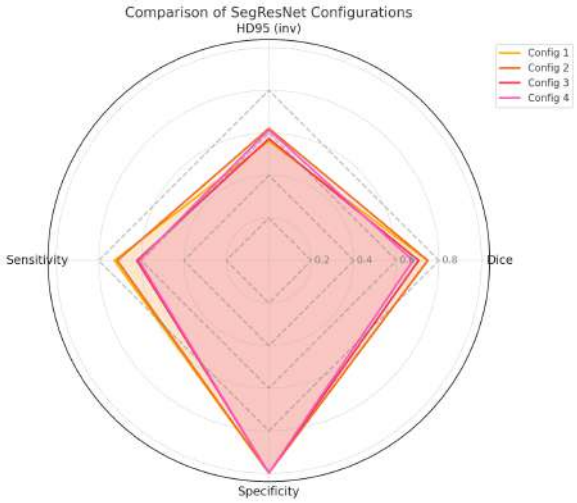


(b) Radar chart across metrics

Figure 6.1: Cross-validation performance of V-Net across four configurations.

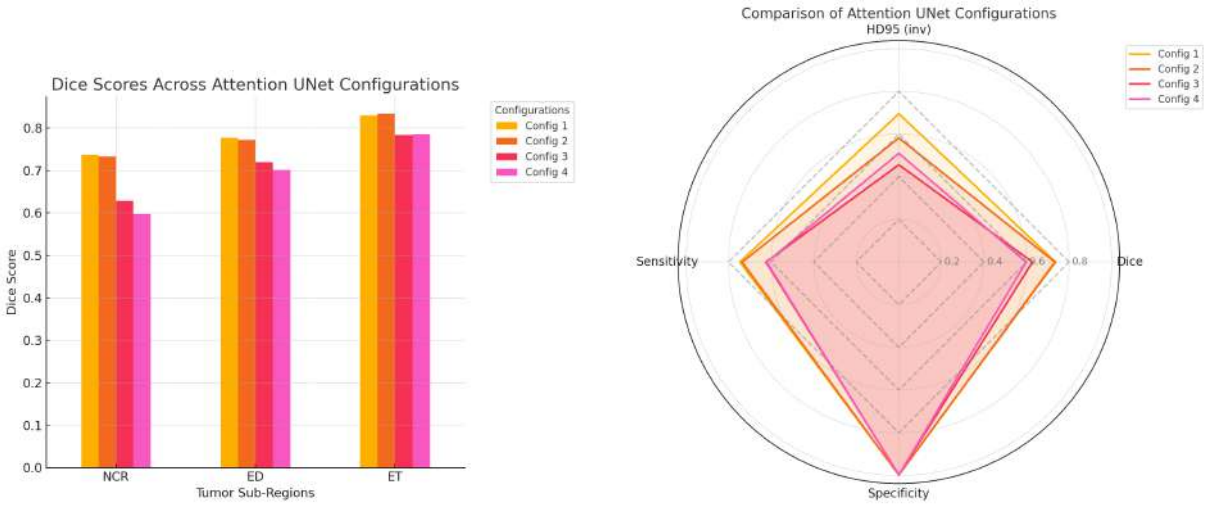


(a) Dice scores by sub-region



(b) Radar chart across metrics

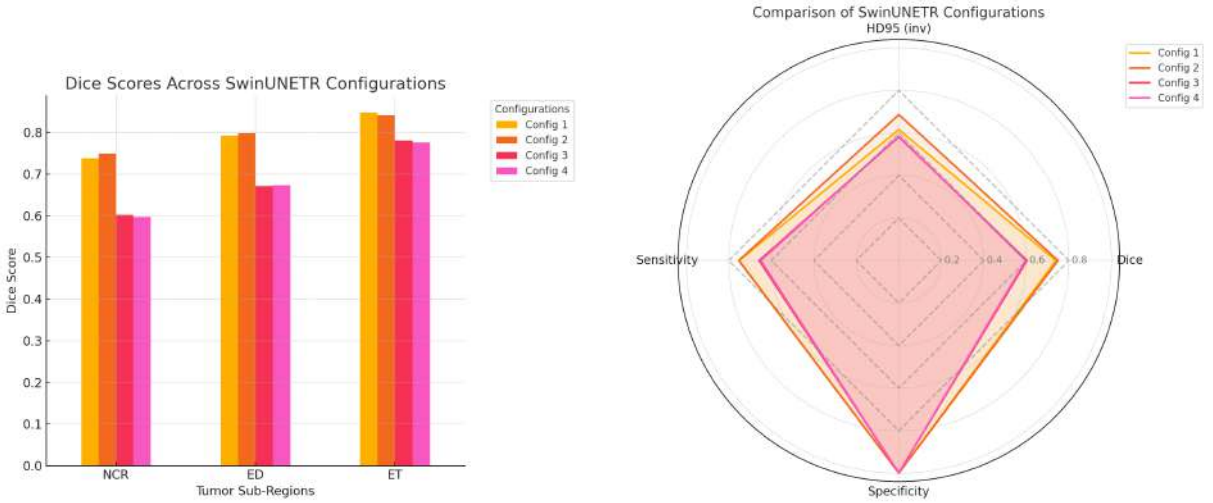
Figure 6.2: Cross-validation performance of SegResNet across four configurations.



(a) Dice scores by sub-region

(b) Radar chart across metrics

Figure 6.3: Cross-validation performance of Attention UNet across four configurations.



(a) Dice scores by sub-region

(b) Radar chart across metrics

Figure 6.4: Cross-validation performance of SwinUNETR across four configurations.

6.1.2 Final hyperparameter configurations

Based on cross-validation results, the optimal hyperparameters were chosen for each model. Table 6.1 presents the final hyperparameter configurations selected for re-training each of them. In all cases, configurations using the AdamW optimizer and a lower learning rate (0.0001) led to superior performance, particularly in terms of volumetric accuracy (Dice score) and boundary precision (HD95).

Each model was subsequently retrained using its selected configuration on 90% of the data for training and 10% for validation.

Table 6.1: Selected hyperparameter configurations for retraining each model.

Model	Configuration	Learning rate	Optimizer	Weight decay
V-Net	1	0.0001	AdamW	1e-5
SegResNet	1	0.0001	AdamW	1e-5
Attention UNet	2	0.0001	AdamW	0.0001
SwinUNETR	2	0.0001	AdamW	0.0001

6.2 Model performance on test set

6.2.1 Performance across individual models

6.2.1.1 Performance metrics overview

After re-training the individual models using the optimal hyperparameters, their performance was compared using Dice scores as well as sensitivity, HD95, and specificity. Table 6.2 provides a detailed numerical comparison of the tumor segmentation performance for each model. The data reveals that:

- Attention UNet achieves the best performance in the NCR sub-region, obtaining the highest Dice score (0.7244), a low HD95 (7.81), and the highest sensitivity (0.7319) in this sub-region. This suggests that Attention UNet is particularly good at segmenting the necrotic core sub-region of the tumor.

- SegResNet demonstrates its strengths primarily in the ED and ET sub-regions. It leads in the ET region with the highest Dice score (0.8226) and the lowest HD95 value (5.47), and it also achieves superior sensitivity (0.8183) in this sub-region. In the case of ED, SegResNet also achieves the best performance, obtaining a slightly higher Dice score than Attention UNet (0.7639) and the lowest HD95 (13.7480).

Table 6.2: Tumor segmentation performance metrics for each model with standard deviations

Metric	Attention UNet	SegResNet	SwinUNETR	V-Net
Dice Scores				
Dice NCR	0.7244 ± 0.3114	0.7144 ± 0.3091	0.7043 ± 0.3277	0.6365 ± 0.3258
Dice ED	0.7604 ± 0.1832	0.7639 ± 0.1926	0.7502 ± 0.1921	0.7038 ± 0.2024
Dice ET	0.8024 ± 0.2059	0.8226 ± 0.2091	0.8200 ± 0.2094	0.5895 ± 0.3012
Dice Overall	0.7624 ± 0.1862	0.7670 ± 0.1957	0.7581 ± 0.2001	0.6432 ± 0.2250
HD95				
HD95 NCR	7.8143 ± 10.3893	8.0385 ± 10.4800	8.3355 ± 10.4441	44.4316 ± 40.6994
HD95 ED	14.2434 ± 19.0894	13.7480 ± 19.4973	15.8083 ± 22.2843	32.6853 ± 30.6074
HD95 ET	5.8104 ± 10.8230	5.4743 ± 11.3163	7.1022 ± 16.9847	63.9401 ± 36.2688
HD95 Overall	7.0690 ± 6.9459	6.9421 ± 6.9130	7.9350 ± 8.8697	38.5781 ± 26.3360
Sensitivity				
Sensitivity NCR	0.7319 ± 0.3368	0.6943 ± 0.3246	0.6685 ± 0.3384	0.6961 ± 0.3180
Sensitivity ED	0.7139 ± 0.1905	0.7134 ± 0.1982	0.7054 ± 0.1948	0.6658 ± 0.2085
Sensitivity ET	0.7374 ± 0.2199	0.8183 ± 0.2182	0.7805 ± 0.2238	0.7669 ± 0.2057
Sensitivity Overall	0.7277 ± 0.1900	0.7420 ± 0.1967	0.7181 ± 0.1992	0.7096 ± 0.1736
Specificity				
Specificity NCR	0.9999 ± 0.0002	0.9999 ± 0.0002	0.9999 ± 0.0002	0.9998 ± 0.0004
Specificity ED	0.9993 ± 0.0010	0.9994 ± 0.0010	0.9993 ± 0.0012	0.9990 ± 0.0015
Specificity ET	0.9999 ± 0.0002	0.9998 ± 0.0003	0.9999 ± 0.0002	0.9981 ± 0.0026
Specificity Overall	0.9997 ± 0.0004	0.9997 ± 0.0004	0.9997 ± 0.0004	0.9990 ± 0.0011

Figure 6.5 shows box-and-whisker plots of the Dice scores for each model and tumor sub-region, and Figure 6.6 shows the same data as violin-density plots. In the box plots, each box spans the interquartile range (IQR) and the whiskers mark the furthest non-outlier values. In the violin plots, the width at each value indicates the empirical score density. Notably, V-Nets box plot shows the widest whiskers, indicating that this model has the biggest variability in performance. This is further confirmed by the violin plots, where the V-Net shows the lest skewness towards higher Dice scores, suggesting more varied (and often worse) performance on more cases.

Both representations indicate that the distributions are clearly asymmetric, with the bulk of scores being present mainly near the top end and the long tails extending downward. That skew toward higher Dice values was statistically confirmed by Shapiro–Wilk tests (all models and sub-regions $p < 0.0001$), so the assumption of normality is violated in every case.

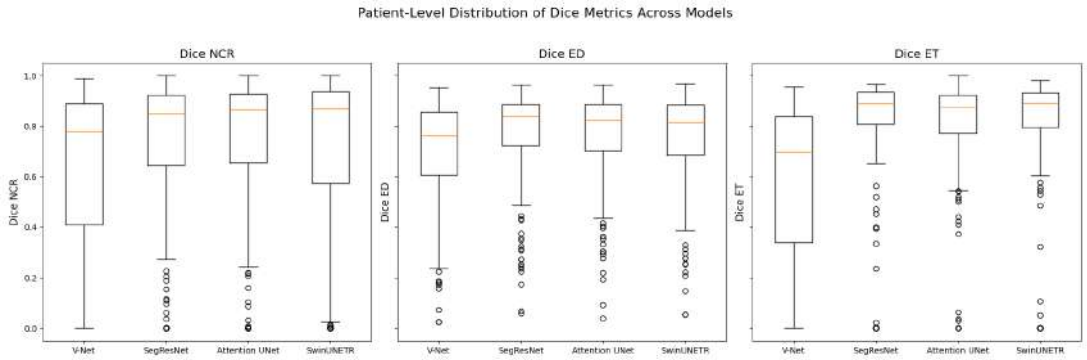


Figure 6.5: Box plots of Dice scores for the tumor sub-regions (NCR, ED, ET) across all models.

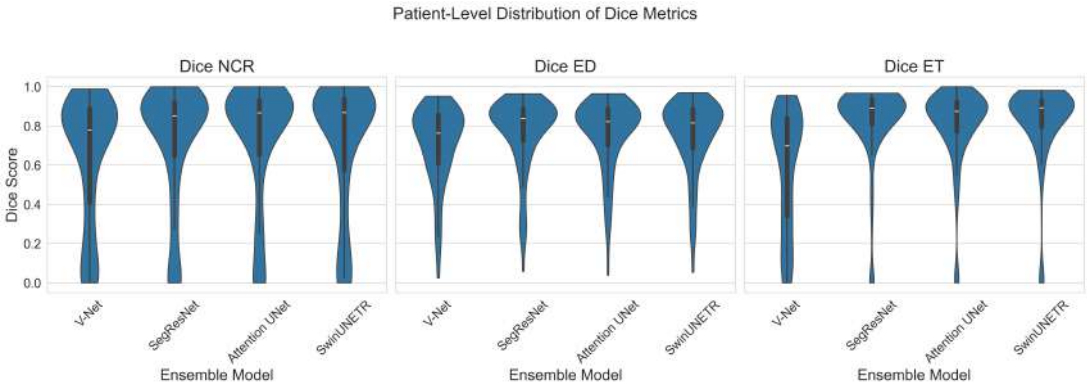


Figure 6.6: Violin plots of Dice scores for the tumor sub-regions (NCR, ED, ET) across all models.

Across all four models, the choice of architecture had a highly significant impact on each of the four performance metrics (all Kruskal–Wallis tests $p < 0.05$). Below the most important pairwise differences are highlighted. The full set of significant post-hoc results is presented in Appendix A.2.5 (Table 5).

Dice scores

Figure 6.7 illustrates the differences in mean Dice scores for the three tumor sub-regions (NCR, ED, ET) across the models. Model differences were significant in every tumor sub-region (NCR:$H(3) = 19.49$, $p = 2.2 \times 10^{-4}$; ED: $H(3) = 18.60$, $p = 3.3 \times 10^{-4}$; ET: $H(3) = 127.38$, $p < 10^{-26}$). In all three regions, V-Net performed substantially worse than the other networks. For example, V-Net's median Dice in ET was lower than SegResNet (U=7561, $p < 10^{-20}$) and Attention UNet (U=8975, $p < 10^{-15}$).

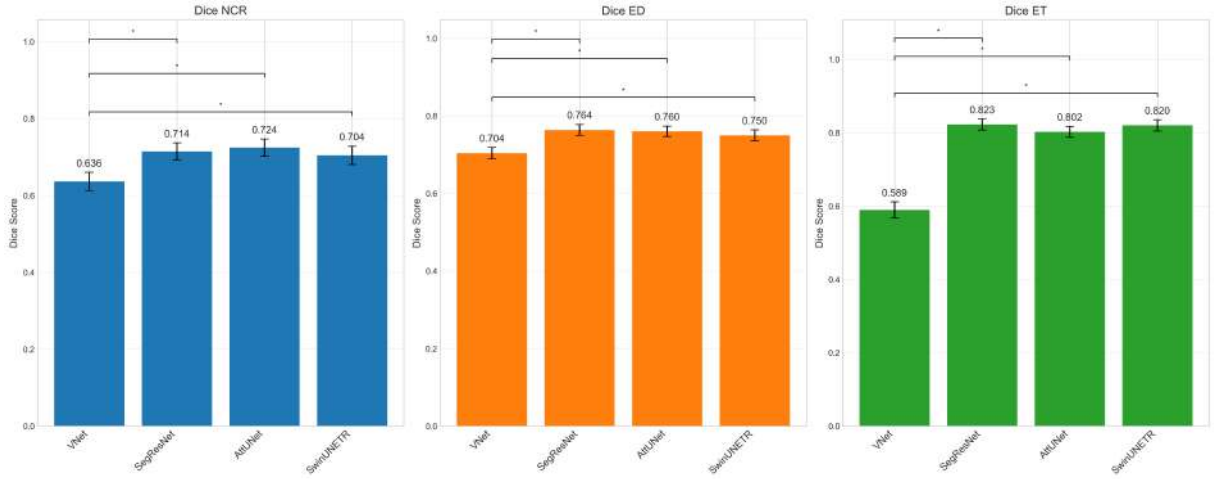


Figure 6.7: Dice scores for each sub-region (NCR, ED, ET) across all models, including standard deviation bars and significance bars.

HD95 distances

Significant effects were also observed for the HD95 distance in NCR ($H(3) = 112.99$, $p < 10^{-23}$), ED ($H(3) = 56.46$, $p < 10^{-11}$), and ET ($H(3) = 252.78$, $p < 10^{-53}$). Again, V-Net yielded markedly worse boundary agreement: in ET, V-Net's HD95 exceeded SegResNet's by nearly an order of magnitude (U=31496, $p < 10^{-38}$).

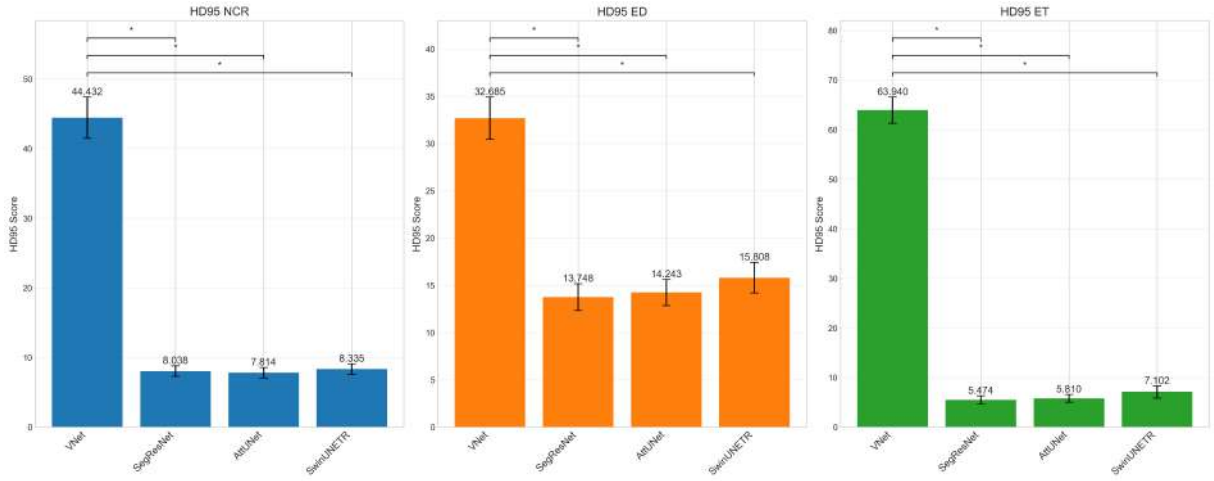


Figure 6.8: HD95 distances for each sub-region (NCR, ED, ET) across all models, including standard deviation bars and significance bars.

Sensitivity and specificity

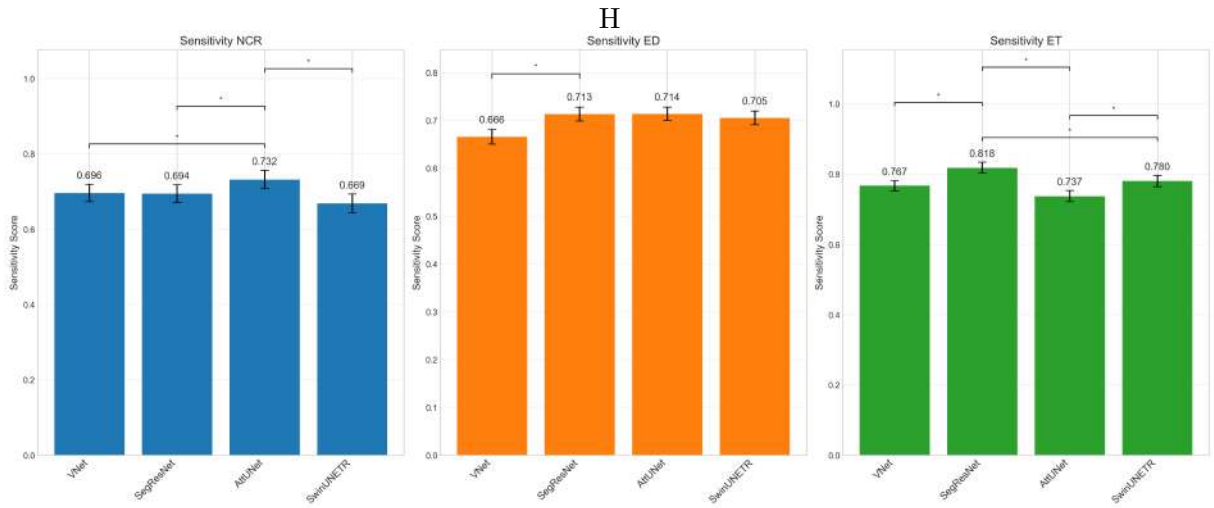


Figure 6.9: Sensitivity for each sub-region (NCR, ED, ET) across all models, including standard deviation bars and significance bars.

Sensitivity differed by model in NCR ($H(3) = 12.29$, $p = 0.0065$), ED ($H(3) = 9.19$, $p = 0.0269$), and ET ($H(3) = 41.57$, $p < 10^{-8}$). Attention UNet showed the highest detection rate in the necrotic core (NCR; V-Net < Attention UNet, U=14885, $p = 0.049$), whereas SegResNet led in ET sensitivity (SegResNet > V-Net, U=12742, $p = 1.7 \times 10^{-5}$).

Although specificity differences were statistically significant, absolute values exceeded 0.99 for all models and so are omitted here. See Appendix A.2.5, Table 5, for the complete post-hoc comparisons, including specificity results.

Absent sub-regions

Additionally, the robustness of each model in cases where specific tumor sub-regions were absent was assesed. As previously mentioned, certain cases lacked sub-regions such as NCR or ET. Since most cases contained all sub-regions, the absence of a particular tissue could be challenging for the models. Table 6.3 presents the average Dice scores for each sub-region in cases where that tissue was absent. A higher average Dice score indicates a better ability of the model to correctly predict the absence of the respective tissue.

Table 6.3: Average Dice scores for cases where certain tissue was absent.

	NCR absent (n=6)	**ED absent (n=1)**	**ET absent (n=6)**
V-Net	0	0	0
SegResNet	0.167	0	0
Attention UNet	0.5	0	0.5
SwinUNETR	0.5	0	0

The results show that the Attention UNet was the most robust, correctly predicting the absence of NCR and ET in 50% of cases. The SwinUNETR also achieved a 50% success rate for predicting the absence of NCR. SegResNet correctly identified the absence of NCR in one out of six cases, while V-Net consistently failed to predict the absence of tissues. None of the models successfully predicted the absence of the ED region, likely because only one such case was available, providing insufficient examples for the models to learn this scenario.

6.2.1.2 Confusion matrices

To gain better insights into the types of errors made by each model at the voxel level, confusion matrices were generated. Figure 6.10 displays these matrices, normalized to percentages, for each model. In these matrices, the rows represent the true voxel labels (ground truth), while the columns represent the labels predicted by the model. The diagonal cells thus indicate the percentage of voxels correctly classified for each label, while off-diagonal cells reveal the percentage of voxels misclassified as a different label.

Notably, all models demonstrated near-perfect accuracy in identifying the background, approaching 100%. This explains the consistently high specificity scores observed across all models. The reason for such superior performance is likely the fact that the background takes up the majority of the MRI scan.

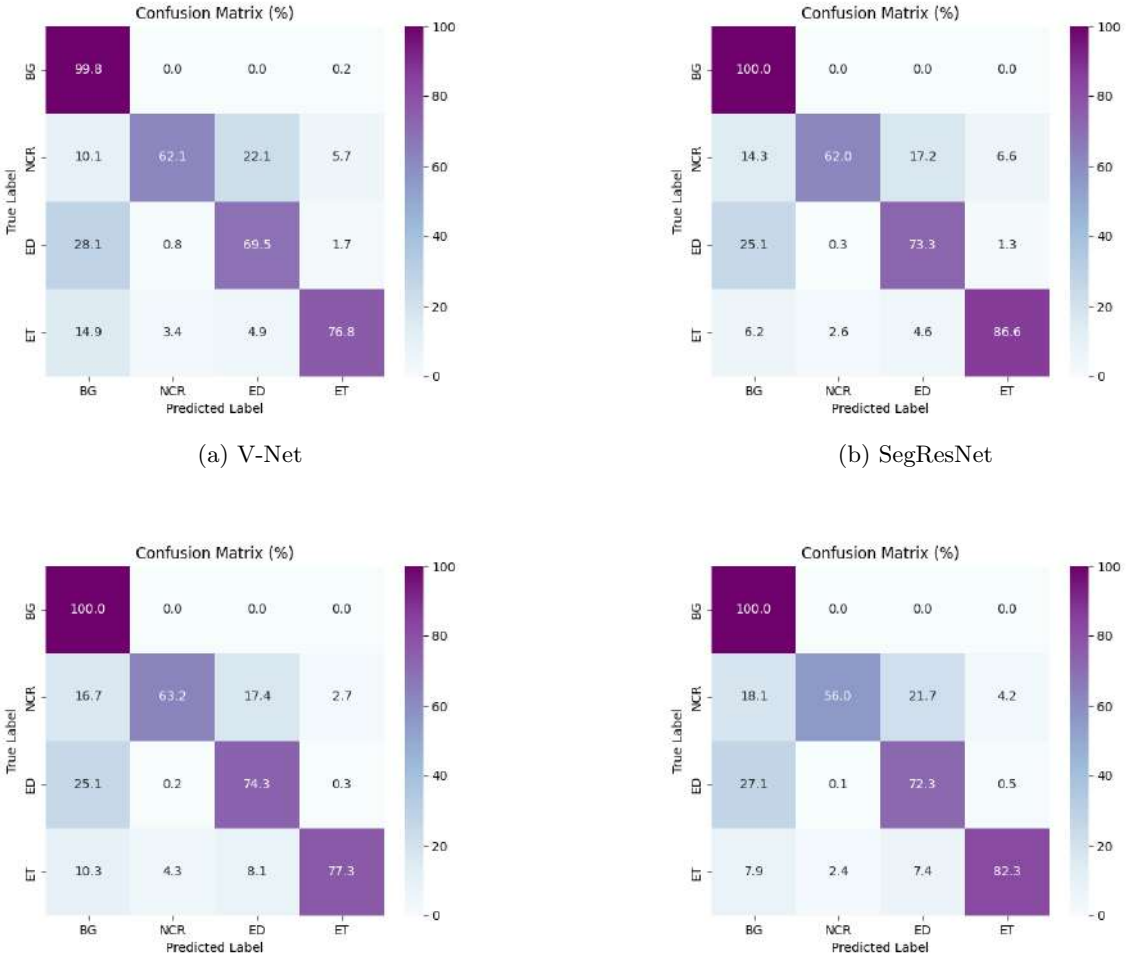
The NCR sub-region proved to be the most difficult to segment accurately. Consequently, models frequently misclassified this sub-region as either ED or background. Notably, misclassification of the NCR as ET was less prevalent. Several factors likely contribute to this pattern. ET exhibits hyper-intensity on T1-weighted contrast-enhanced (T1CE) scans compared to T1,

creating a stark visual contrast [37]. Conversely, the NCR typically appears hypo-intense on T1CE compared to T1, with a signal intensity similar to cerebrospinal fluid (CSF) and healthy white matter, making it distinct from the bright ET [37]. On FLAIR, the NCR is only mildly bright and sits inside the highly hyper-intense ED which is typically depicted by the abnormal hyper-intense signal in the FLAIR modality [37].

Furthermore, the NCR's small volume makes it particularly vulnerable to being overwhelmed by the signals from surrounding tissues. As a result, the models may learn to disregard the subtle NCR signal, often mislabeling it as background or incorporating it into the hyper-intense FLAIR signal characteristic of ED.

The ED tissue was usually mislabeled as background. ED often has infiltrative boundaries, gradually transitioning into normal brain tissue. This can lead to blurring at the edges in the MRI scans and cause the mislabeling of ED as background.

All models showed the most robust performance on the ET. The likely cause for it is that this area shows high contrast, particularly in the T1CE modality, and tends to have more distinct boundaries compared to ED and NCR.



(a) V-Net

(b) SegResNet

(c) Attention UNet

(d) SwinUNETR

Figure 6.10: Confusion matrices for individual models.

6.2.1.3 Correlation analysis with MRI features

A correlation analysis was performed between various MRI-derived features and the Dice scores obtained by each model for the NCR, ED, and ET regions to better understand the reasons for differences in segmentation performance across tumor sub-regions. For each patient, extracted several types of features were extracted from each MRI modality, including tumor volume (in mm^3), voxel intensity statistics (mean, minimum, maximum, standard deviation), as well as radiomics features such as first order uncertainty, GLCM contrast, GLCM homogeneity, and GLDM dependence uncertainty. Figures 6.11-6.13 demonstrate significant correlations (p-value < 0.05) between the extracted MRI features and Dice scores on the specific sub-region.

NCR

For the NCR sub-region, V-Net demonstrated moderate positive correlations with tumor volume ($r \approx 0.24$) and several texture-related features (e.g., FLAIR and T1CE first-order uncertainty, GLCM contrast, and GLDM dependence uncertainty with $r \approx 0.17$–0.25). This indicates that V-Net benefits from larger tumors and greater textural heterogeneity, likely because such lesions present more distinct boundaries and richer local information.

Similarly, SegResNet showed strong positive correlations with FLAIR-derived radiomic features (with r values reaching even 0.28), suggesting that subtle textural variations in FLAIR images are particularly important for its performance in segmenting the NCR.

In contrast, Attention UNet and SwinUNETR exhibited fewer correlations in the NCR region, implying that these models are less dependent on local intensity and textural properties for accurate segmentation. Similarly, the relatively modest correlations of SwinUNETR's Dice scores and MRI features suggest that it might be less affected by local intensity variations or textural properties captured by these features. This could be due to SwinUNETR's transformer-based architecture, which tends to capture more global context instead of focusing mainly on local features. Therefore, SwinUNETR's segmentation performance may be more consistent across a range of tumor characteristics.

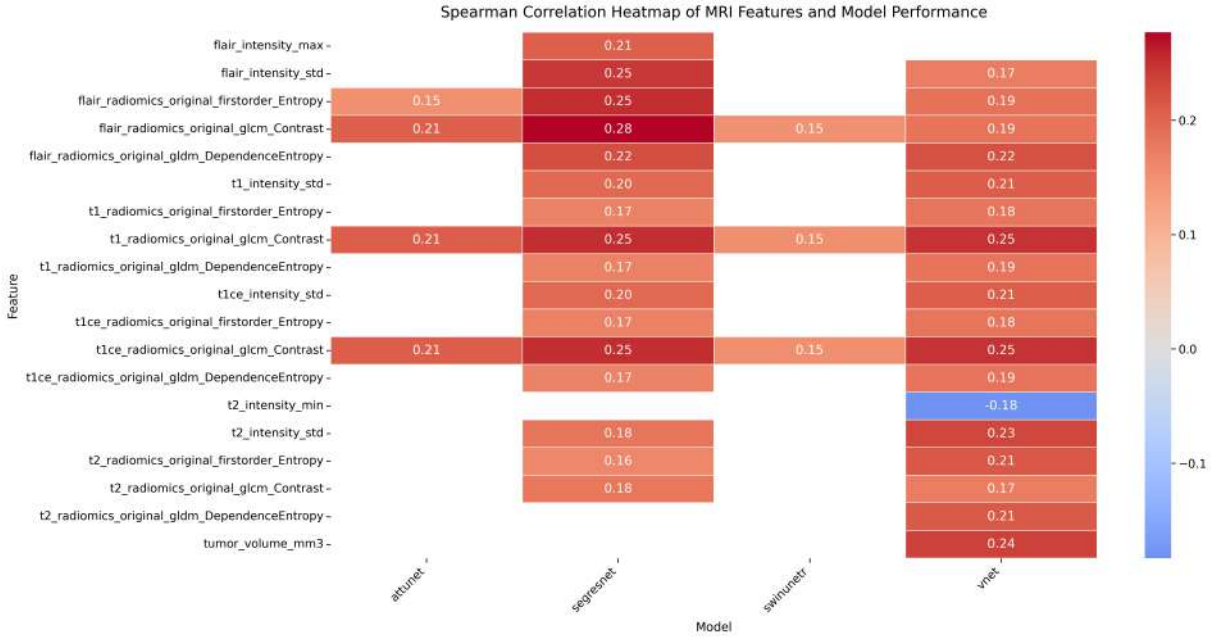


Figure 6.11: Correlations between model Dice scores on NCR sub-regions and MRI features.

ED

In the ED sub-region, the correlations were generally more modest. Both V-Net and SegResNet presented moderate positive correlations with tumor volume ($r \approx 0.15$ to 0.18), along with negative correlations for T1CE and T1 intensity means ($r \approx -0.18$ to -0.20). These results suggest that lower average intensities (potentially leading to higher contrast) are associated with better segmentation of edema. Attention UNet also showed a similar pattern, while SwinUNETR revealed modest correlations (e.g., negative correlations with T1CE intensities around $r \approx -0.19$ and a positive correlation with T2 intensity minimum around $r \approx 0.17$). This indicates that, compared to the NCR region, radiomic features have a less pronounced impact on performance in ED.

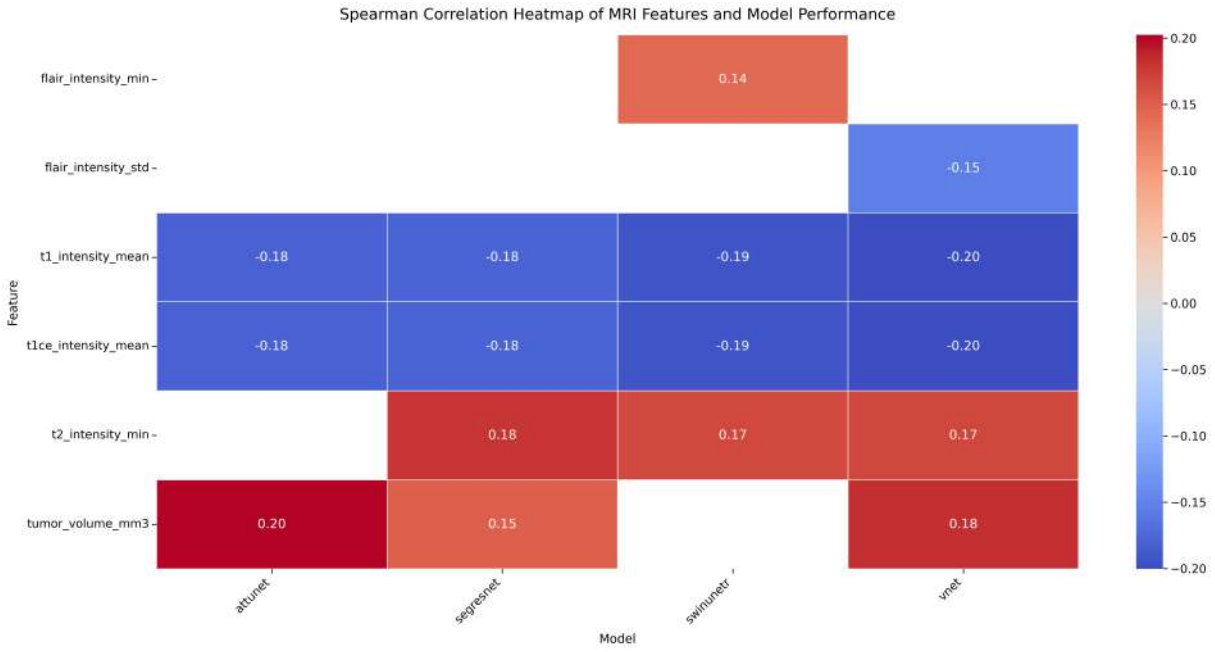


Figure 6.12: Correlations between model Dice scores on ED sub-region and MRI features.

ET

For the ET region, V-Net again displayed strong positive correlations. In particular, tumor volume correlated very strongly ($r \approx 0.46$) with its Dice scores. In addition, several textural metrics - such as T1CE intensity minimum (with $r \approx -0.30$), and T2 radiomics features (with $r \approx 0.20$ to 0.25) - were significantly correlated. These findings suggest that V-Net's performance in segmenting the enhancing tumor is highly influenced by both the size of the lesion and local textural heterogeneity. SegResNet showed moderate correlations with T2-based features (e.g., first-order uncertainty with $r \approx 0.17$), while both Attention UNet and SwinUNETR had only one significant correlation each (with T2 intensity minimum at $r \approx 0.20$ and $r \approx 0.19$, respectively), implying that these models are less affected by local intensity variations in ET.

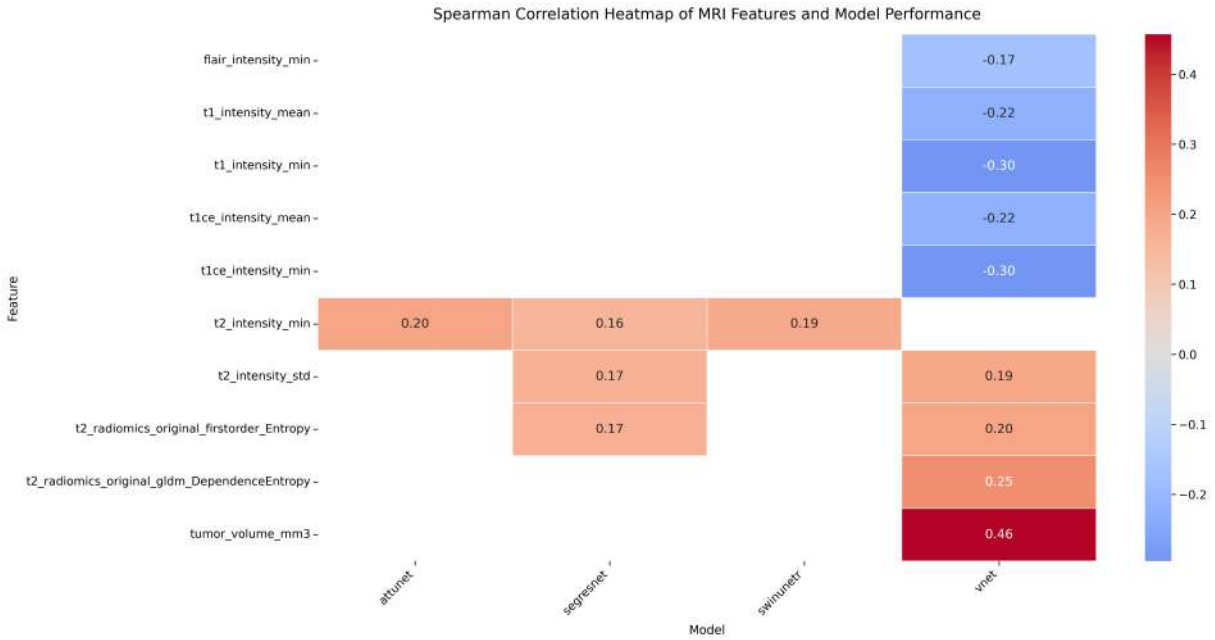


Figure 6.13: Correlations between model Dice scores on ET sub-region and MRI features.

Overall, the NCR sub-region exhibits a higher number of significant correlations with MRI features compared to ED and ET. This suggests that accurate segmentation of the necrotic core is particularly sensitive to variations in texture and voxel intensity statistics.

The differences across models indicate that V-Net and SegResNet are more strongly influenced by these local image characteristics, while Attention UNet and SwinUNETR appear more robust to such variations - likely due to their differing architectural strategies (e.g., global contextual modeling in SwinUNETR). These insights not only enhance the understanding of model-specific performance but also provide another argument for employing an ensemble method: by combining models with complementary sensitivities, it might be possible to achieve a more robust overall segmentation performance across diverse tumor presentations.

6.2.1.4 Visual examples

To supplement the quantitative analysis, Figure 6.14 presents qualitative examples for three representative patient cases. These cases were selected to demonstrate the varying performance of the models and further highlight the need for an ensemble approach.

The leftmost column displays the segmentation overlays for a patient where Attention UNet demonstrated the best performance in delineating the NCR sub-region, obtaining the Dice score of 0.78, and outperforming the second best model, SegResNet, by a difference of 0.51. The predictions show that, indeed, the NCR sub-region delineated by Attention UNet resembles the NCR sub-region in the ground truth most closely, while the SwinUNETR model did not predict the NCR region at all for this slice of the scan.

The second column showcases the strengths of the SegResNet model. The segmentation overlays reveal that SegResNet not only achieves higher ovelap in the NCR sub-region (quantiatively, Dice score of approximately 0.61 compared to the second-best value of 0.16, achieved by Attention UNet) but also exhibits superior performance across the ED and ET sub-regions (obtaining 0.68 and 0.80 Dice scores, respectively).

The third column presents the case of patient 01405 which proved to be challenging to all models, however, SwinUNETR achieved significantly better performance on segmenting this tumor than other models (average Dice of ≈ 0.62 compared to ≈ 0.50 (SegResNet), ≈ 0.42 (Attention UNet), and ≈ 0.39 (V-Net)). While most models miss the small ET and NCR sub-region completely, SwinUNETR does not miss it and also predicts small parts of the disconnected ED sub-region. One possible reason for such a difference could be the fact that SwinUNETR utilizes more global context and can thus better spot differences that are less prominent and more distant from each other, rather than predicting them as the background label.

Lastly, the rightmost column illustrates a scenario in which all models struggle. The overall Dice scores for this patient are lower and exhibit high variability (ranging from roughly 0.31 to 0.58 across models), indicating a challenging case where tumor heterogeneity or atypical morphology limits segmentation accuracy. A likely cause for the models' mistake could be the size of the NCR sub-region in this patient which, although usually very small, in this patient takes a significant portion of the tumor. This unexpected feature could be misleading for the models which safely predict that this must be an ED tissue. This example highlights that, despite robust average performance, individual cases may still present significant challenges. This motivates the development of ensemble methods that can potentially compensate for such variability.

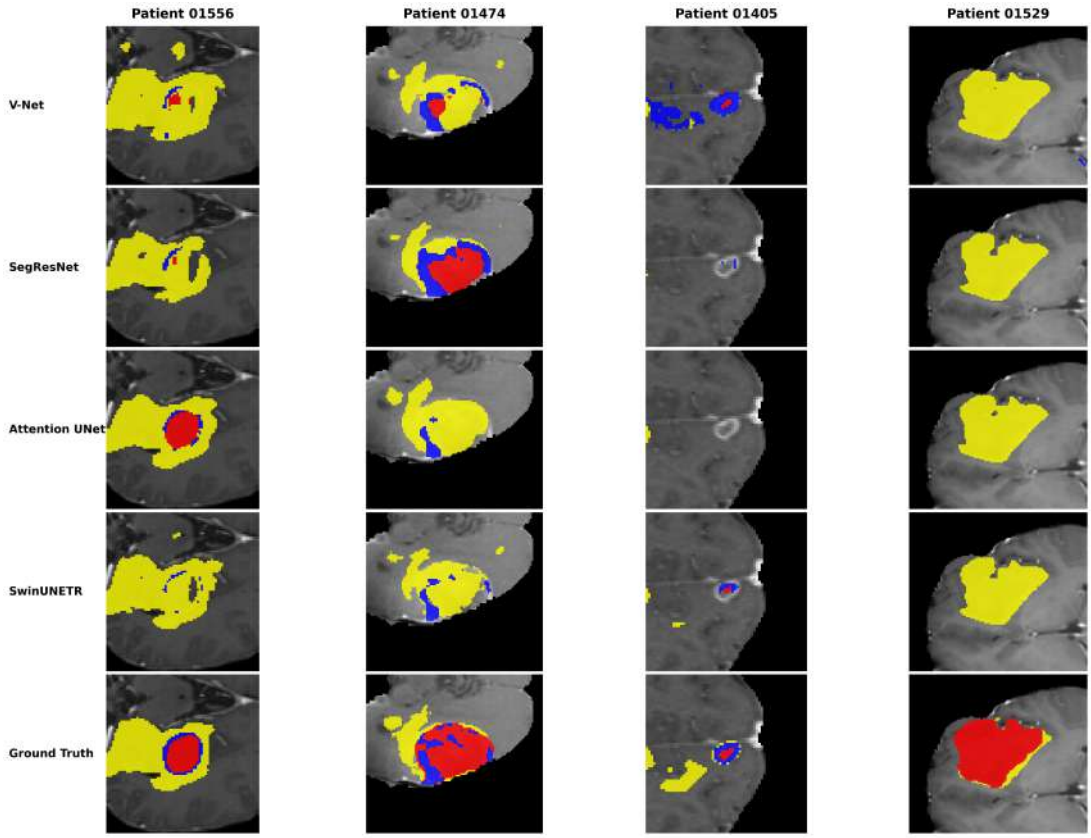


Figure 6.14: Segmentation overlays for three representative patient cases demonstrating strengths and weaknesses of the different models.

Summary

In summary, the statistical and qualitative analyses indicate that V-Net consistently underperforms compared to SegResNet, Attention UNet, and SwinUNETR across all tumor subregions, as evidenced by its significantly lower Dice scores and greater performance variability. Since the ultimate goal is to construct an ensemble that leverages the strengths of individual models for robust segmentation, the considerably worse performance of V-Net suggests that including it in the ensemble might degrade overall accuracy. Therefore, excluding V-Net could enhance ensemble performance by relying on the more consistently performing models.

6.2.2 Performance of ensemble models

The main goal of the ensemble experiments was to investigate whether combining predictions using various weighting schemes (simple averaging, performance weighting, and performance plus

uncertainty weighting with TTD and TTA) can lead to improved segmentation performance.

1. **Simple Averaging Ensemble**: This method combines the predictions of the individual models by calculating their arithmetic mean. Its performance is intended to serve as a baseline for evaluating more sophisticated weighting strategies.

2. **Performance-Weighted Ensemble**: In this approach, the contribution of each model is weighted globally based on its individual performance on a **validation set**. Consequently, each model is assigned a specific weight for each sub-region, reflecting its performance on that particular region.

3. **Performance and Uncertainty-Weighted Ensemble**: This method determines model weights by considering both their performance and associated uncertainty measures. Similar to the Performance-Weighted ensemble, global performance weights are assigned per sub-region. However, uncertainty weights are applied at the voxel level using voxel-wise maps for each voxel within the sub-region. Uncertainty was assessed using:

   (a) Test Time Dropout (TTD) uncertainty alone.

   (b) Test Time Augmentation (TTA) uncertainty alone.

   (c) A combination of TTA and TTD uncertainties.

6.2.2.1 Composite scores in Performance-Weighted ensemble

Firstly, composite scores for each model were computed on the validation set to avoid data leakage into the test evaluation. A description and justification for the composite score can be found in the Methodology chapter Section 5.7.2.

Table 6.4 presents the raw composite scores and the normalized scores (which sum to one) for each tumor sub-region. For example, in the NCR sub-region, Attention UNet achieved the highest raw score of 0.672. In the ED sub-region, Attention UNet and SegResNet obtained similar scores of 0.665 and 0.667, respectively. For the ET sub-region, SegResNet performed best with a raw score of 0.742, followed by SwinUNETR with 0.731. The normalized scores

were then applied in the Performance-Weighted and Performance and Uncertainty-Weighted ensemble.

Table 6.4: Composite scores for each model and tumor sub-region

Metric	Attention UNet	SegResNet	SwinUNETR
Composite Scores			
NCR	0.672	0.659	0.641
ED	0.665	0.667	0.658
ET	0.703	0.741	0.731
Normalized Composite Scores (final weights)			
NCR	0.341	0.334	0.325
ED	0.334	0.335	0.331
ET	0.323	0.341	0.336

6.2.2.2 Performance metrics overview

Table 6.5 summarizes the four core metrics—Dice, HD95, sensitivity, specificity—along with their standard deviations, for all five ensemble strategies.

The TTA-Only ensemble yields the highest overall Dice overlap (0.7816 ± 0.1986), closely followed by the Hybrid (TTD+TTA) at 0.7766 ± 0.2018. Breaking this down by sub-region, TTA-Only also leads on NCR (0.7331 ± 0.3173) and ED (0.7827 ± 0.1853), while the Performance-Weighted model slightly edges out on ET overlap (0.8311 ± 0.1984). This shift suggests that test-time augmentations contribute more to overlap accuracy than the approach that combines TTD and TTA uncertainty estimation, but does not improve the overlap performance on the ET sub-region.

In terms of boundary accuracy, TTD-Only again excels with the lowest overall HD95 (6.8317 ± 7.6474 mm), and the best ED boundary performance at 8.6778 ± 13.5215 mm. Both TTA-Only and Hybrid ensembles, by contrast, shows a higher overall HD95 of 7.0580 ± 7.9239 mm and 8.6647 ± 10.9392 mm, respecitvely, indicating that the smoothing effect of the Hybrid ensemble broadens some tumor boundaries.

For sensitivity, the TTA ensemble remains strongest overall (0.7522 ± 0.1998), highlighting its ability to detect positive regions with fewer false negatives. The Hybrid approach records

the highest sensitivity on ED (0.7579 ± 0.1956), while TTA-Only leads on ET detection (0.7957 ± 0.2225). All methods maintain very high specificity($\geq$ 0.9993 ± 0.0012 across all models), underscoring their robustness in correctly excluding non-tumor tissue.

Finally, it is worth noting the large variability in NCR segmentation (standard deviations > 0.31 for all ensembles), reflecting the inherent difficulty of this small sub-region. These findings point to a potential need for region-specific strategies—such as finer input resolution or targeted loss weighting—to stabilize performance on the NCR class.

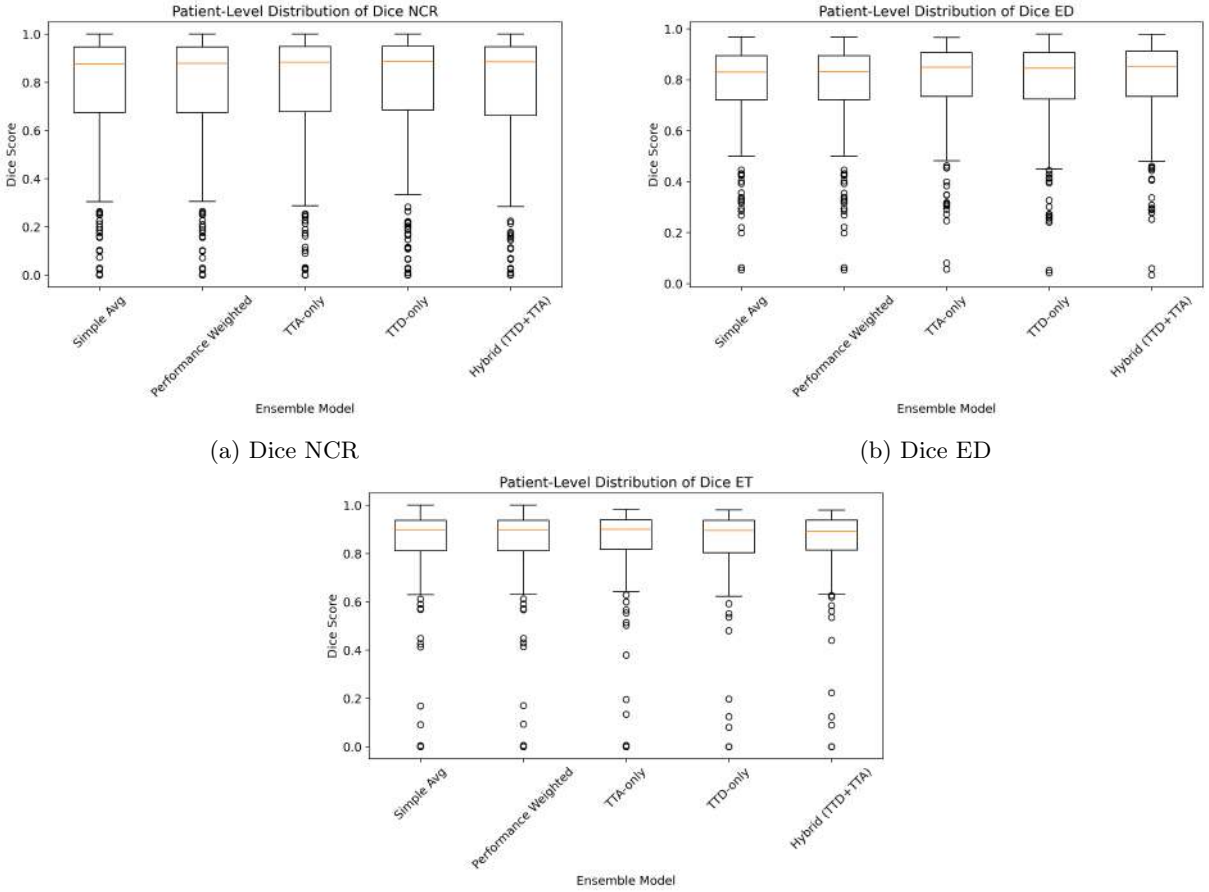
Table 6.5: Performance metrics for the Simple Averaging, Performance-Weighted, TTD-Only, TTA-Only, and Hybrid (TTD+TTA) ensembles with standard deviations.

Metric	Simple Averaging	Perf.-Weighted	TTD-Only	TTA-Only	TTD+TTA
Dice Scores					
Dice NCR	0.7296 ± 0.3202	0.7297 ± 0.3202	0.7314 ± 0.3143	0.7331 ± 0.3173	0.7193 ± 0.3299
Dice ED	0.7677 ± 0.1891	0.7678 ± 0.1891	0.7762 ± 0.1907	0.7827 ± 0.1853	0.7849 ± 0.1875
Dice ET	0.8309 ± 0.1985	0.8311 ± 0.1984	0.8275 ± 0.2030	0.8291 ± 0.2136	0.8258 ± 0.2093
Dice Overall	0.7761 ± 0.1911	0.7762 ± 0.1910	0.7784 ± 0.1924	0.7816 ± 0.1986	0.7766 ± 0.2018
HD95					
HD95 NCR	7.4401 ± 9.5124	7.4426 ± 9.5220	6.9740 ± 8.9093	7.0369 ± 9.0372	7.8722 ± 9.9323
HD95 ED	12.0162 ± 17.3564	11.9972 ± 17.3592	8.6778 ± 13.5215	9.5395 ± 15.3121	12.2299 ± 20.3462
HD95 ET	4.9050 ± 9.9415	4.8963 ± 9.9358	4.8432 ± 10.2844	4.5975 ± 10.0191	5.8920 ± 15.3080
HD95 Overall	8.1204 ± 8.4524	8.1120 ± 8.4544	6.8317 ± 7.6474	7.0580 ± 7.9239	8.6647 ± 10.9392
Sensitivity					
Sensitivity NCR	0.7210 ± 0.3362	0.7211 ± 0.3361	0.7191 ± 0.3330	0.7238 ± 0.3336	0.7022 ± 0.3394
Sensitivity ED	0.7152 ± 0.1954	0.7152 ± 0.1954	0.7241 ± 0.1999	0.7372 ± 0.1920	0.7579 ± 0.1956
Sensitivity ET	0.7872 ± 0.2125	0.7878 ± 0.2123	0.7909 ± 0.2134	0.7957 ± 0.2225	0.7892 ± 0.2211
Sensitivity Overall	0.7411 ± 0.1929	0.7414 ± 0.1928	0.7447 ± 0.1952	0.7522 ± 0.1998	0.7498 ± 0.2026
Specificity					
Specificity NCR	0.9999 ± 0.0002	0.9999 ± 0.0002	0.9999 ± 0.0002	0.9999 ± 0.0002	0.9999 ± 0.0002
Specificity ED	0.9994 ± 0.0011	0.9994 ± 0.0011	0.9995 ± 0.0010	0.9994 ± 0.0011	0.9993 ± 0.0012
Specificity ET	0.9999 ± 0.0002	0.9999 ± 0.0002	0.9999 ± 0.0002	0.9999 ± 0.0002	0.9999 ± 0.0002
Specificity Overall	0.9997 ± 0.0004	0.9997 ± 0.0004	0.9997 ± 0.0004	0.9997 ± 0.0004	0.9997 ± 0.0004

Figures 6.15 and 6.16 depict the patient-level distributions of Dice scores for the NCR, ED and ET sub-regions across all five ensemble strategies. In the boxplots (Figure 6.15), the overall overlap performance on ET remains uniformly high for every ensemble, with medians clustered around 0.90 and relatively narrow IQR, indicating consistently strong segmentation of the enhancing tumor.

By contrast, the ED sub-region exhibits a clear progression in median Dice — from roughly 0.83 for Simple Averaging and performance weighting, to about 0.85 for TTD-only, 0.87 for

TTA-only and 0.86 for the Hybrid approach—alongside the largest spread of values and the greatest number of low-score outliers. The NCR region falls in between, with medians near 0.88–0.89 for Hybrid and TTA-only ensembles but wider whiskers and more variability than ET, reflecting a moderate level of patient-to-patient inconsistency.



(a) Dice NCR

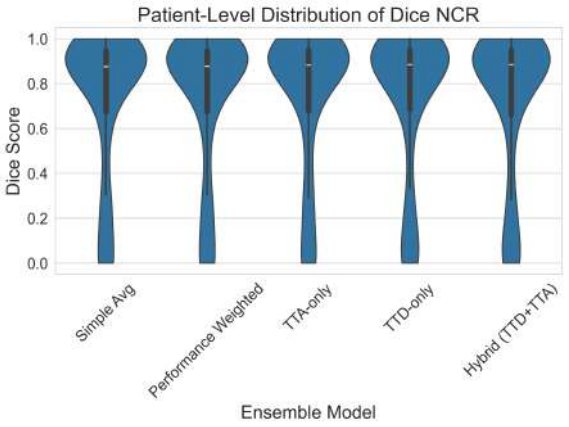
(b) Dice ED

(c) Dice ET

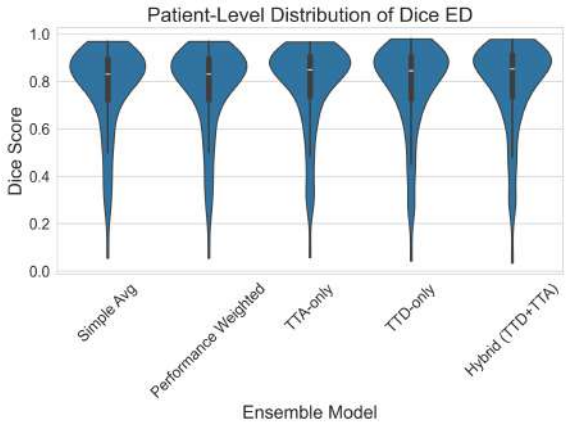
Figure 6.15: Boxplots of Dice scores for the tumor sub-regions (NCR, ED, and ET) across ensemble strategies.

The violin plots in Figure 6.16 reinforce these findings by showing the underlying density of scores. Each distribution is heavily skewed toward high Dice values (approximately 0.8–1.0), yet the width and shape of the violins differ by sub-region. The ED violins are the bulkiest toward the lower end, underscoring the heavier tail of poorly segmented cases. In contrast, the ET violins are tall and narrow, confirming that most patients achieve near-perfect overlap on ET. The NCR violins lie somewhere between these extremes, with a modest leftward extension indicating occasional failures on the necrotic core. Together, these plots illustrate that test-time

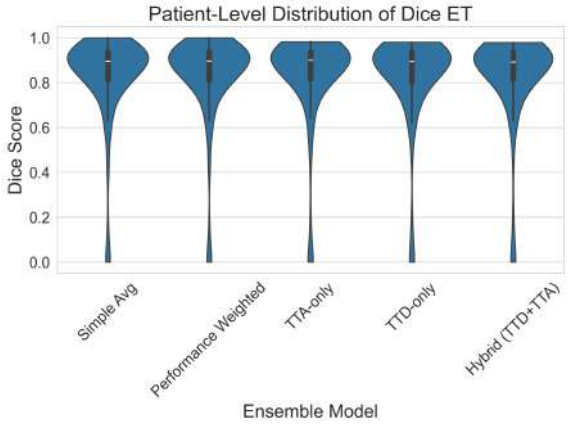
augmentation and the Hybrid ensemble offer the most consistent gains for the more challenging NCR and ED tissues, while all methods perform similarly well on ET.



(a) Dice NCR



(b) Dice ED



(c) Dice ET

Figure 6.16: Distribution of Dice scores for the tumor sub-regions (NCR, ED, and ET) across ensembles.

6.2.2.3 Statistical analysis

All ensemble strategies were compared pairwise across the full set of metrics to assess whether any method outperformed the others in a statistically significant way. First, each metric's distribution across the five ensembles (Simple-Avg, Performance-Weighted, TTD-Only, TTA-Only, Hybrid) was tested for normality using the Shapiro–Wilk test. In every case at least one group violated normality (all $p < 0.05$), so the non-parametric Kruskal–Wallis test was performed for the overall comparison.

For the Dice scores, the Kruskal–Wallis statistic for NCR was $H = 0.1639$ ($p = 0.9968$), for

ED $H = 3.7662$ ($p = 0.4386$), for ET $H = 0.6504$ ($p = 0.9573$), and for the overall Dice score $H = 23.1797$ ($p = 0.0001$).

The HD95 distances yielded $H = 0.6815$ ($p = 0.9536$) for NCR, $H = 6.3533$ ($p = 0.1743$) for ED, $H = 2.1182$ ($p = 0.7140$) for ET, and $H = 6.5281$ ($p = 0.1630$) overall.

Sensitivity produced $H = 1.8180$ ($p = 0.7692$) for NCR, $H = 11.6009$ ($p = 0.0206$) for ED, $H = 1.4834$ ($p = 0.8296$) for ET, and $H = 21.0180$ ($p = 0.0003$) overall. While specificity showed $H = 3.8855$ ($p = 0.4217$) for NCR, $H = 13.3462$ ($p = 0.0097$) for ED, $H = 6.7429$ ($p = 0.1501$) for ET, and $H = 678.0968$ ($p < 0.0001$) overall.

After applying a Bonferroni correction to control for multiple comparisons, none of the adjusted p-values fell below the significance threshold. Consequently, no post-hoc pairwise tests were performed. Thus, all five ensemble strategies demonstrate statistically equivalent performance across Dice, HD95, sensitivity, and specificity.

6.2.2.4 Confusion matrices

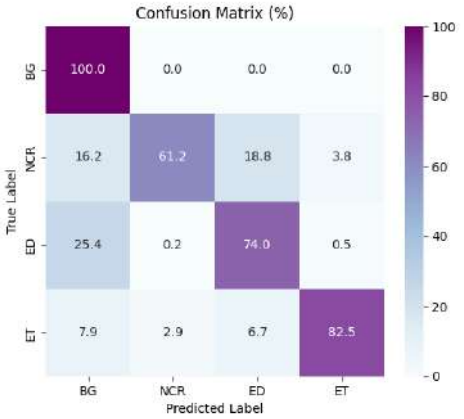
All ensembles achieve perfect segmentation of BG (100%). Recall for the NCR sub-region is very consistent: both Simple Averaging and Performance-Weighted ensembles obtain 61.2%, TTA-Only and Hybrid also 61.2%, while TTD-Only attains 63.8%, equivalent to the leading single-model performance.

ED recall is enhanced by ensembling the single models. Simple Averaging and Performance-Weighted reach 74.0%, TTD-Only 74.6%, and both Hybrid and TTA-Only achieve 77.8%, surpassing the strongest individual model.

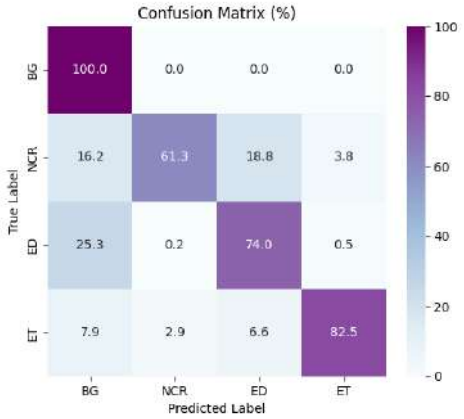
For ET, the highest recall of 83.1% is observed under Hybrid and TTA-Only, followed by 82.5% for Simple Averagingg/Performance-Weighted and 76.9% for TTD-Only. Despite good performance, none of the ensembles exceeded SegResNet's recall of 86.6%.

The almost identical results of Simple Averaging and Performance-Weighted ensemble strategies imply that validation-score weighting affords minimal benefit when base models are of similar quality. TTD-Only excels in NCR at the slight expense of ED recall, indicating a trade-off be-

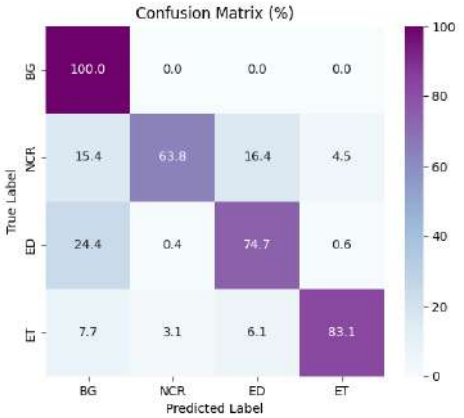
tween boundary precision (at which TTD-Only model was superior) and core lesion sensitivity. TTA-Only approach demonstrates boundary robustness particularly for ED and ET on par with the Hybrid approach.



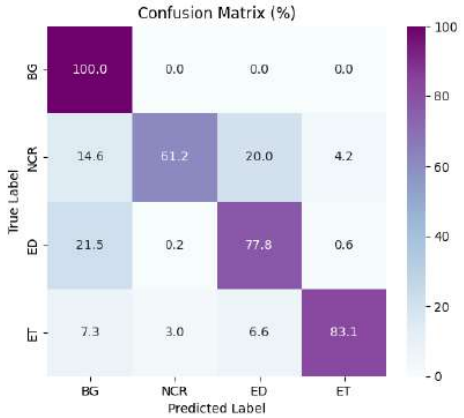
(a) Simple Averaging ensemble



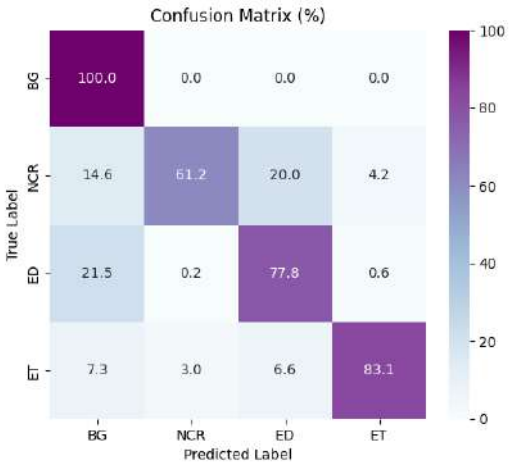
(b) Performance-Weighted ensemble



(c) TTD-Only ensemble



(d) TTA-Only ensemble



(e) Hybrid ensemble

Figure 6.17: Confusion matrices for ensemble models.

6.2.2.5 Correlation analysis with MRI features

To further understand the differences across the ensemble models, a correlation analysis was performed between key MRI features and the Dice scores obtained on each sub-region (NCR, ED, and ET). Only correlations with $p < 0.005$ are reported. Three heatmaps (Figures 6.18–6.20) summarize these associations for the NCR, ED, and ET sub-regions, respectively.

Necrotic Core

For the NCR sub-region, 11 of the 33 inspected descriptors reached the significance threshold of $p < 0.005$. The largest coefficients were obtained for the grey–level co-occurrence matrix (GLCM) contrast. Figure 6.11 presents all significant correlations for the NCR sub-region. In the FLAIR sequence, this texture feature correlated with Dice score at $\rho = 0.21$ for Simple Averaging and Performance-Weighted ensembles and reached $\rho = 0.25$ for TTD, TTA, and Hybrid approaches. An almost identical progression was observed for the T1 and T1CE counterparts, whose values rose from $\rho = 0.19$ to $\rho = 0.24$.

First-order statistics showed more moderate but still reliable links. FLAIR uncertainty increased from $\rho = 0.16$ (simple ensemble methods) to $\rho = 0.21$ (uncertainty-weighted methods), and the standard deviation of FLAIR intensities followed a similar pattern, moving from $\rho = 0.15$ to $\rho = 0.20$.

The maximum FLAIR intensity became significant only for the three most Performance and Uncertainty-Weighted ensembles, yielding $\rho = 0.17$–0.18, while GLDM dependence-uncertainty reached $\rho = 0.16$–0.17 under the same conditions. For T1 and T1CE, intensity spread showed small yet consistent effects ($\rho = 0.17$–0.18), whereas their first-order uncertainty was significant solely for the Hybrid configuration ($\rho = 0.14$).

A nearly monotonic rise in absolute ρ values can be traced from the Simple Avgeraging baseline to the Hybrid ensemble. This trend indicates that models benefiting from performance and unertainty weighting are better able to exploit the fine-grained textural details present in FLAIR and T1-weighted MRI scans when delineating the necrotic core. No T2-based feature exceeded $\rho = 0.19$ in any ensemble, suggesting limited relevance of pure T2 information for this

sub-region.

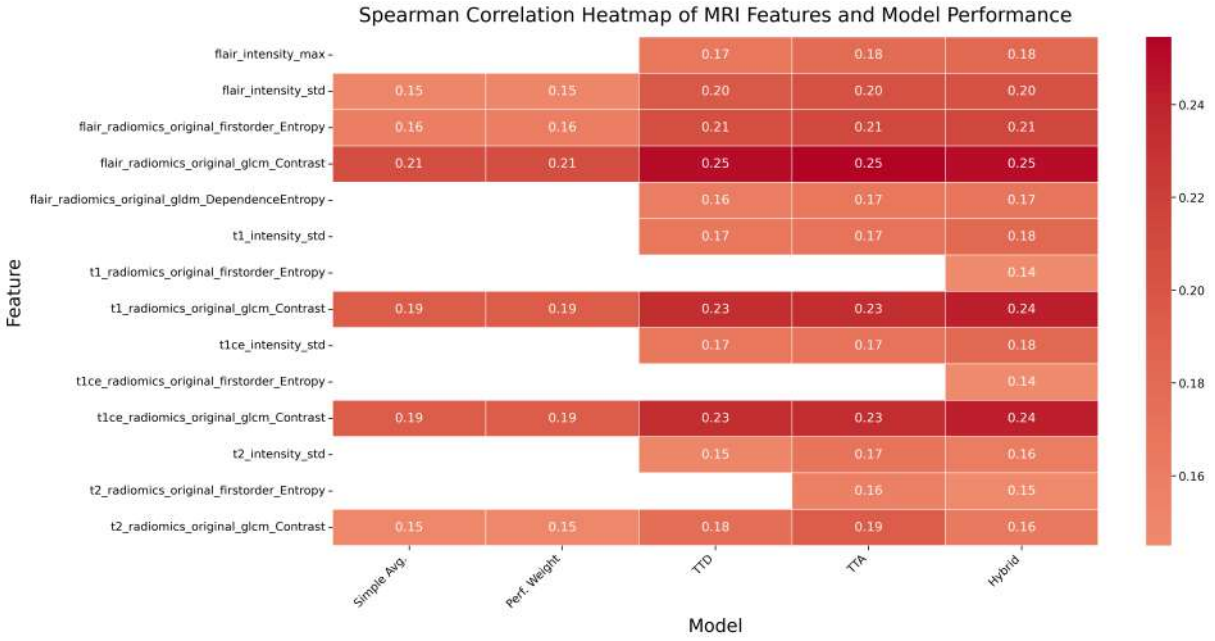


Figure 6.18: Correlations between model Dice scores on NCR sub-region and MRI features for the ensemble models.

Edema

The correlation pattern obtained for the ED sub-region differs from that of the NCR and points to a stronger dependence on coarse intensity cues. Only seven feature–ensemble pairs reached the $p < 0.005$ significance level and none of the absolute coefficients exceeded $|\rho| = 0.21$.

The most stable positive association was produced by the minimum T2 intensity, which correlated at $\rho = 0.17$ for both the Simple Averaging and Performance-Weighted ensembles, dipped slightly to $\rho = 0.16$ in TTD, and rose again to $\rho = 0.18$ under TTA and Hybrid. Because ED is typically hyper-intense on T2, a low voxel minimum may indicate that the surrounding white matter is well-suppressed, thereby increasing the local contrast and facilitating segmentation.

The minimum FLAIR intensity also showed significant correlations, although significance was limited to Simple Averaging and Performance-Weighted ensembles ($\rho = 0.15$ and $\rho = 0.14$, respectively). The loss of significance for Performance and Uncertainty-Weighted ensembles suggests that this feature becomes redundant when additional synthetic views are available, possibly because FLAIR already performs CSF suppression and yields a cleaner ED signal.

In contrast to these positive links, the mean intensity in both T1 and T1CE images showed a consistent negative relationship with performance. Across the first four ensembles, the coefficient remained fixed at $\rho = -0.18$, while the Hybrid scheme intensified the effect to $\rho = -0.21$. Lower mean values in T1-weighted modalities are expected when ED appears hypointense, implying that clearer signal depression aids the models.

For the Hybrid ensemble, first-order uncertainty from T1 and T1CE also entered the significant range ($\rho = -0.15$), hinting that a more heterogeneous intensity distribution within the ED sub-region may degrade its performance on the ED sub-region.

Finally, overall tumor volume correlated positively ($\rho = 0.15$) only in the Hybrid ensemble. Given that larger lesions often occupy more voxels with unambiguous T2/FLAIR hyperintensity, this mild effect is expected and may reflect a basic signal-to-noise advantage.

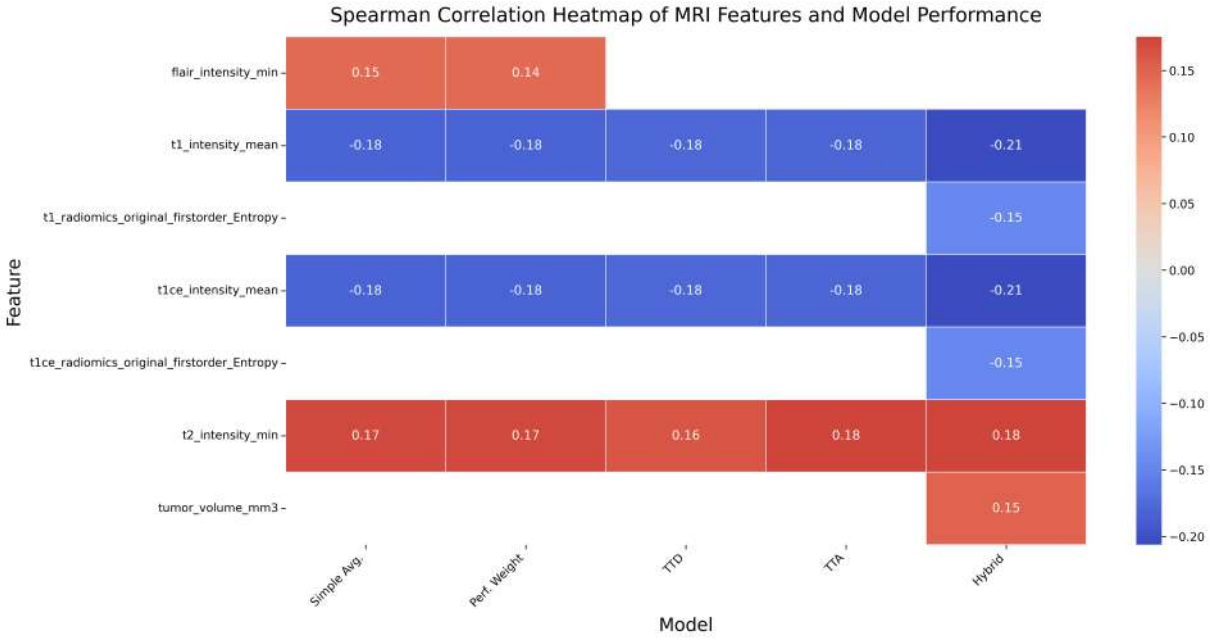


Figure 6.19: Correlations between model Dice scores on ED sub-region and MRI features for the ensemble models.

Enhancing tumor (ET).

Finally, Figure 6.20 displays the correlation results for the ET sub-region. The correlation map for the enhancing component is the sparsest of the three sub-regions and is driven exclusively by features extracted from the T2-weighted sequence. The minimum T2 intensity exhibits

the largest and most persistent association with Dice, reaching $\rho = 0.23$ for both the Simple Averaging and Performance-Weighted ensembles, falling slightly to $\rho = 0.20$ under TTD, and ending at $\rho = 0.17$ and $\rho = 0.19$ for TTA and Hybrid, respectively. Lower signal minima on T2 likely emphasise the compact, relatively hypo-intense core of the enhancing tumor, thereby improving the contrast against ED and normal tissue.

No correlations are detected for any FLAIR, T1, or T1CE feature, which were present in the case of NCR and ED sub-regions. A plausible reason is that the very bright, gadolinium-enhanced ET in T1CE images is already an clear and robust cue for the segmentation models. As a result, additional T1/T1CE statistics do not improve performance further. Instead, subtle T2 signal reductions might offer the only supplementary information that still differentiates better- and worse-performing cases.

In summary, compared with the other two sub-regions, ET shows the smallest set of significant predictors, which are only present in the T2 modality, and a clear decrease of correlations once augmentation or hybrid weighting is deployed. These findings suggest that, once the network captures the primary T1CE signal, very little additional information (beyond a simple T2 contrast cue) is required to achieve better delineation of the enhancing component.

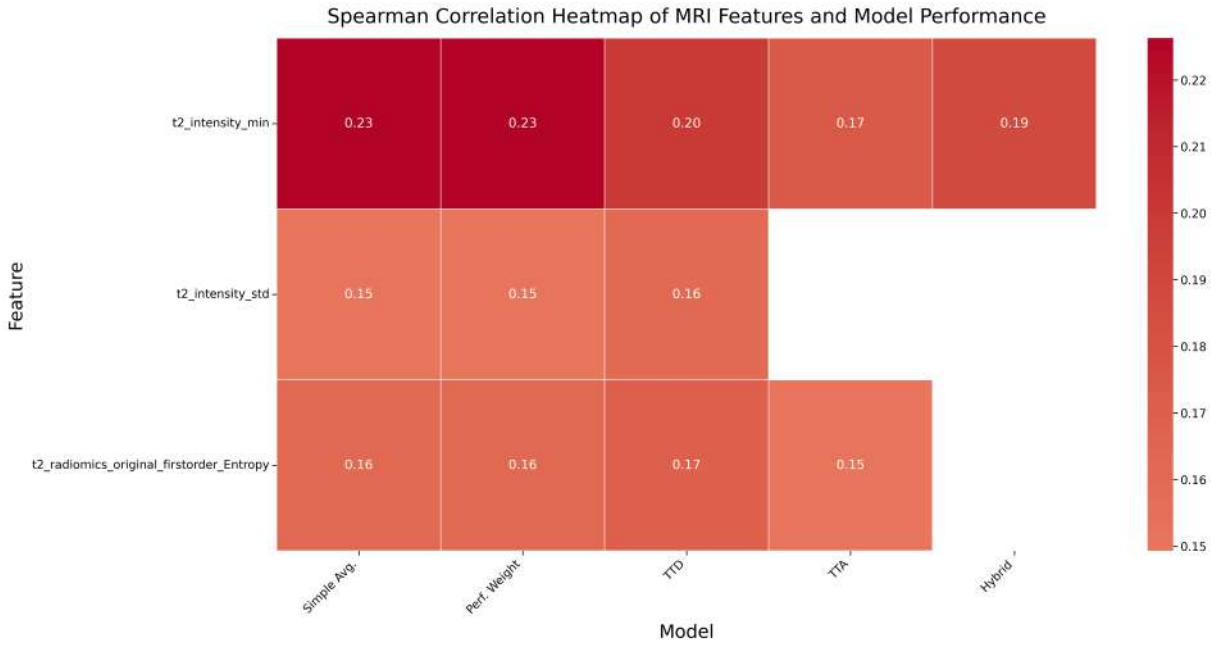


Figure 6.20: Correlations between model Dice scores on ET sub-region and MRI features for the ensemble models.

In conclusion, the correlation patterns are highly sub-region specific. NCR segmentation benefits most from local texture and intensity variability, ED segmentation is sensitive to mean signal levels (particularly in T1 and T1CE), and ET segmentation relies primarily on minimum T2 intensities, with supplementary gains from intensity heterogeneity in some ensemble methods.

These findings indicate that segmentation could possibly be improved by employing tailored pre-processing steps or developing model architectures that specifically exploit the unique features of each imaging modality for different tumor sub-regions.

6.2.2.6 Visual examples

Figure 6.21 presents four representative patients that highlight the strengths and limitations of the different ensemble strategies.

In Patient 00332 (leftmost column), the performance and uncertainty-weighted ensembles clearly outperform the averaging-based methods. The Hybrid ensemble achieves an overall Dice of 0.82, compared to 0.71 for TTD-only, 0.72 for TTA-Only and approximately 0.75 for both Simple Averaging and performance weighting. This gain is particularly evident in the NCR, where the Hybrid model scores 0.56, TTA-Only ensemble scores 0.46, TTD-Only obtains 0.45 Dice score, while the Performance-Weighted ensemble strategy obtained a Dice score of 0.37. ED and ET delineation also benefit, with Dice ED of 0.87 (Hybrid) versus 0.81 (Performance-Weighted), and Dice ET of 0.87 (Hybrid) versus 0.83 (Performance-Weighted). The fragmentation of the necrotic regions and the low contrast at the ED and ET boundaries make overlap-based segmentation especially challenging in this case, highlighting the value of uncertainty guidance.

In Patient 01032 (second column), the simple-average and Performance-Weighted ensembles both achieve moderate overall Dice scores of approximately 0.71. The TTD-Only and Hybrid methods degrade to around 0.66 and 0.65, respectively, while the TTA-Only ensemble performs poorest, with an overall Dice of 0.58. A consistent error across all strategies is the misclassification of the bulky ET sub-region as NCR, indicating that when intensity and textural signatures of ET and NCR overlap substantially, none of the ensembling approaches can fully resolve the ambiguity, but the Simple Averaging approach does this better than the the ensemble strategies

utilizing uncertainty.

The third case (Patient 01147) yields uniformly strong results across all ensembles (overall Dice $\approx$ 0.92–0.95), with the Hybrid and TTD-only approaches each reaching 0.95. The clear separation in intensity and texture between NCR, ED, and ET sub-regions seems to enable all ensemble strategies to perform accurately.

The last column demonstrates a case where none of the ensembles correctly identify the extensive necrotic core — instead, they label it as ED, resulting in uniformly low NCR Dice scores. This failure is most likely caused by the minimal intensity heterogeneity in the MRI scans. As the earlier correlation analysis demonstrated, low variance and weak textural contrast impede accurate NCR segmentation which might be the cause of the hindered performance of the models in this case. Additionally, given a usually small size of the NCR, the ensembles might haved repeated the mistakes of the single models and predict it with more confidence as the ED sub-region.

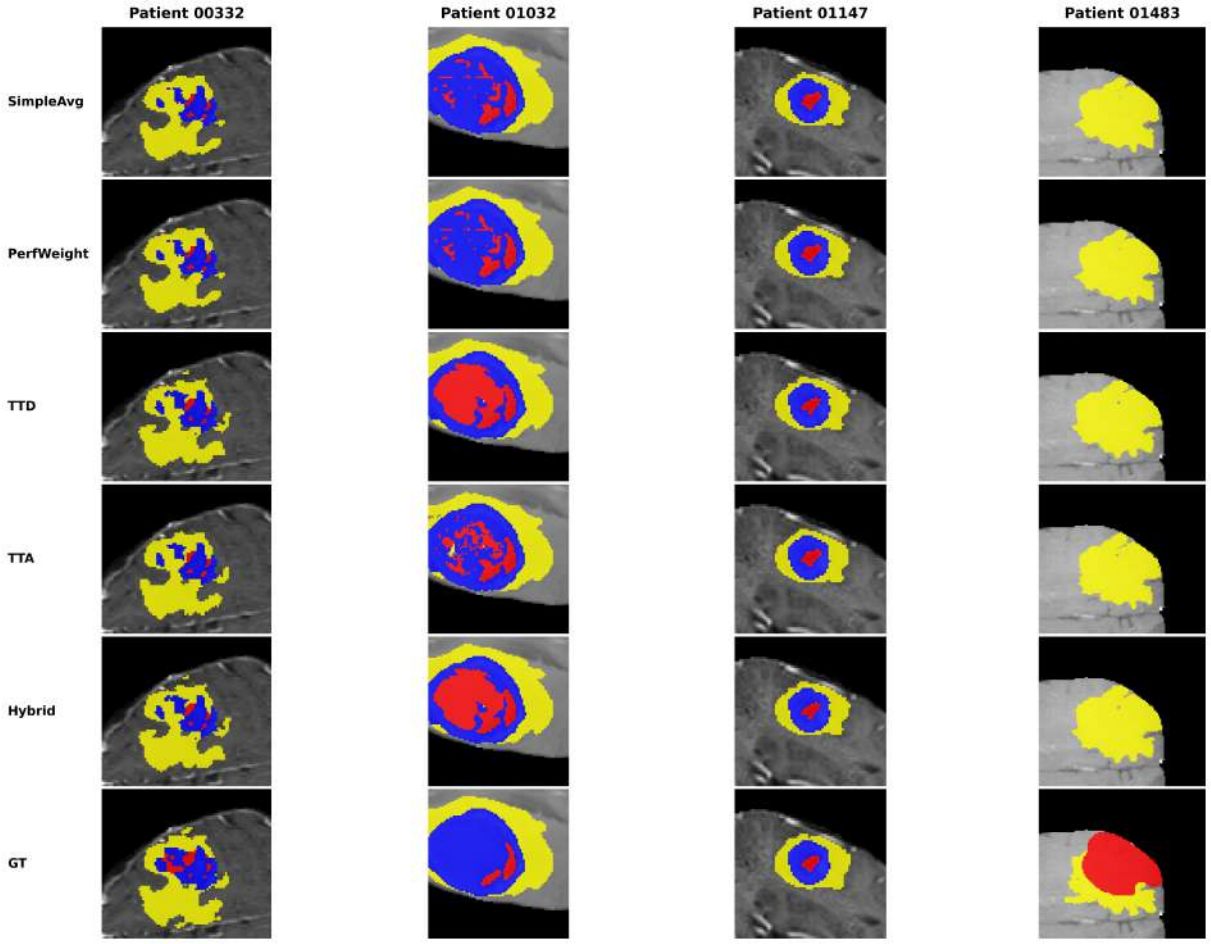


Figure 6.21: Segmentation overlays for four representative patient cases demonstrating strengths and weaknesses of the different ensemble methods.

6.2.3 Comparison Between Individual and Ensemble Results

The performance of individual tumor segmentation models (Attention UNet, SegResNet, SwinUNETR, and V-Net) and their ensembles (Simple Averaging, Performance-Weighted, TTD-Only, TTA-Only, and Hybrid) was evaluated on Dice score, HD95, sensitivity, and specificity. Tables 6.2 and 6.5 give the full results with standard deviations.

6.2.3.1 Dice Scores

Across the three tumor sub-regions, ensembles and single models did not differ significantly in case-wise Dice distributions: NCR ($H = 6.53$, $p > 0.05$), ED ($H = 13.46$, $p > 0.05$), ET ($H = 11.88$, $p > 0.05$) (Fig. 6.22) However, the omnibus Kruskal–Wallis test on the overall

(per-patient) Dice scores was significant ($H = 25.8588$, $p = 0.00053$), indicating at least one pair of methods differs in their case-wise distributions—even though the voxel-aggregate means in Table 6.5 appear very close (Simple Averaging 0.7761, Performance-Weighted 0.7762, TTD-Only 0.7784, TTA-Only 0.7816, Hybrid 0.7766).

Post-hoc Mann–Whitney U-tests reveal that:

- Simple Averaging > TTD-Only ($U = 21223$, $p = 0.0211$)
- Performance-Weighted>TTD-Only ($U = 21227$, $p = 0.0208$)

This arises likely because the voxel-aggregated mean for TTD-Only is slightly inflated by a few very high-Dice cases, whereas in the majority of patients its Dice is marginally lower than Simple Averaging and Performance-Weighted. Thus, although the grand-mean Dice differences are small, the nonparametric pairwise tests capture consistent case-level shifts.

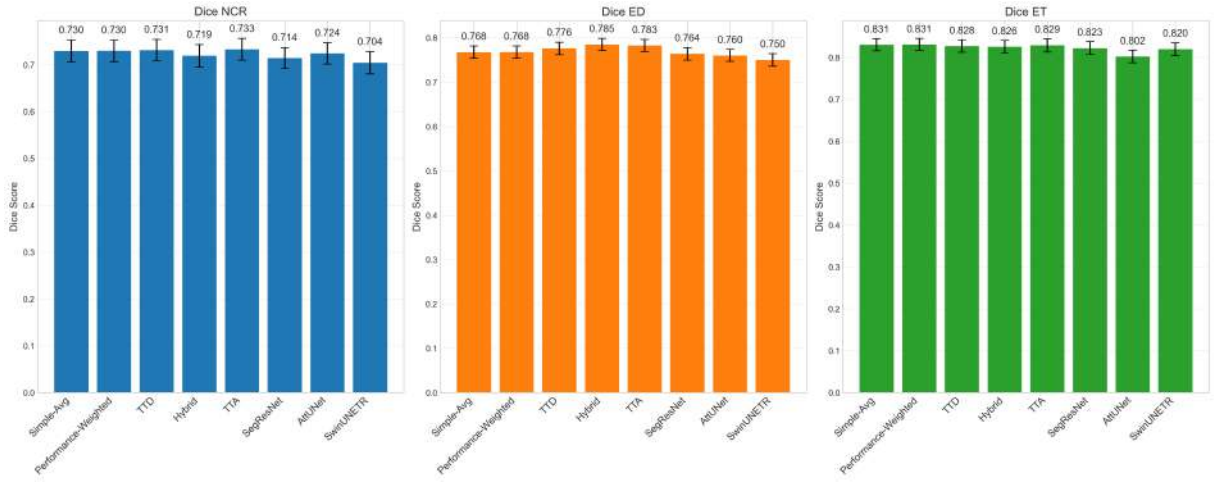


Figure 6.22: Dice scores with error bars for each model across three tumor sub-regions.

6.2.3.2 Hausdorff Distance (HD95) analysis

A notable gain appeared for boundary accuracy in the ED sub-region. The omnibus test indicated a significant effect for HD95 ED ($H = 30.21$, $p = 8.7 \times 10^{-5}$). Post-hoc analysis showed that TTD-Only (8.68 mm) and TTA-Only (9.54 mm) achieved significantlyshorter distances than Attention UNet (14.24 mm; $p = 0.0033$ and $p = 0.0086$, respectively) and SwinUNETR

(15.81 mm; $p = 0.0048$ and $p = 0.0132$, respectively). No significant differences emerged for NCR, ET or for the overall HD95 (Fig. 6.23).

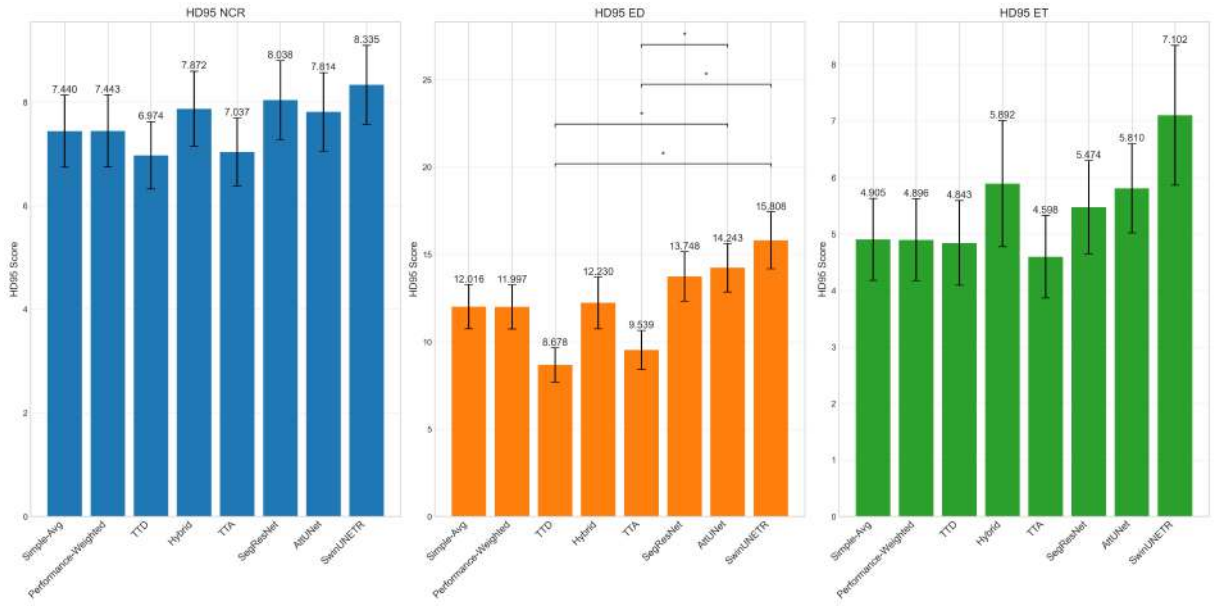


Figure 6.23: HD95 values with error bars for each model across three tumor sub-regions.

6.2.3.3 Sensitivity and Specificity

For the ED sub-region, the Hybrid ensemble increased sensitivity to 0.7579 and significantly surpassed SwinUNETR (0.7054; $U = 21\,282$, $p = 0.0172$). In the ET sub-region, every ensemble exceeded Attention UNet by at least 0.059, with TTA-Only reaching the highest sensitivity (0.7957; Kruskal–Wallis $H = 40.80$, $p = 8.8 \times 10^{-7}$).

Across the entire test set, mean overall sensitivity values ranged from 0.7411 for Simple Averaging to 0.7522 for TTA-Only (Performance-Weighted: 0.7414; TTD-Only: 0.7447; Hybrid: 0.7498). A Kruskal–Wallis test on these per-patient overall-sensitivity scores was significant ($H = 27.27$, $p = 0.00030$). Post-hoc Mann–Whitney U-tests revealed that both Simple Averaging and Performance-Weighted ensembles had significantly higher sensitivity than TTD-Only (Simple Averaging > TTD-Only: $U = 21\,289$, $p = 0.0168$; Performance-Weighted > TTD-Only: $U = 21\,304$, $p = 0.0159$), and that TTD-Only itself was marginally more sensitive than the best single model, SegResNet (0.7447 vs. 0.7420; $U = 13\,958$, $p = 0.0119$). This apparent paradox occurs again because a handful of cases in the TTD-Only ensemble exhibit very high sensitivity

and lift its overall mean, while the majority of patients experience slightly lower sensitivity under TTD-Only—a shift that the rank-based tests detect.

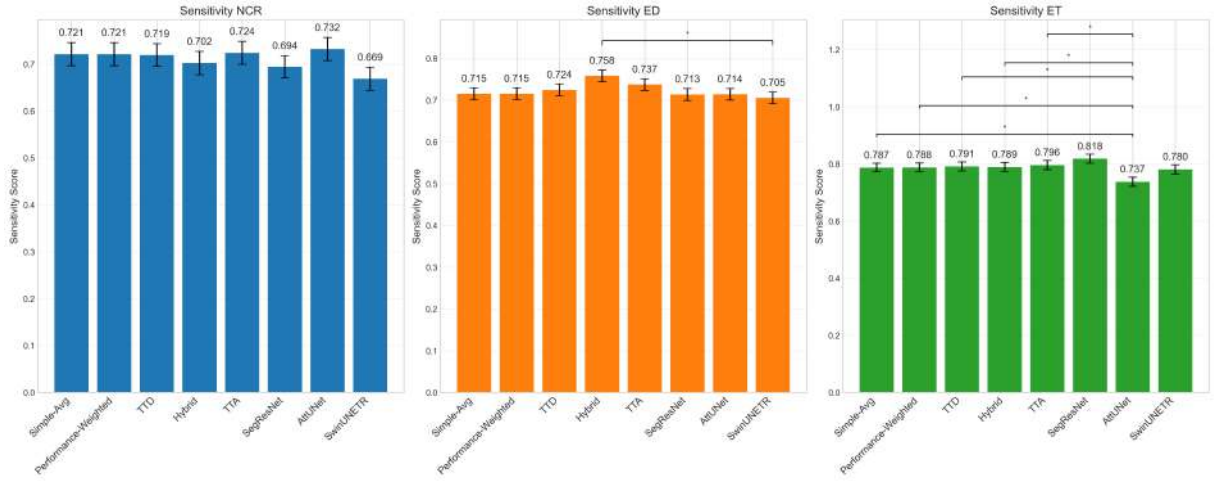


Figure 6.24: Sensitivity scores with error bars and significance lines for each model across three tumor sub-regions.

Specificity differed significantly in several settings (e.g. NCR: $H = 34.73$, $p = 1.3 \times 10^{-5}$). However, the absolute gaps were very small (on the order of 10^{-5}), reflecting the high power of the test rather than a clinically relevant change.

6.2.3.4 Summary of the comparison

Ensembling keeps Dice scores on par with those of the strongest single architectures while offering targeted improvements. For instance, TTD-Only and TTA-Only ensembles sharpen ED boundaries; Hybrid raises recall in ED and maintains a favourable balance between sensitivity and specificity; TTD-Only delivers the highest overall specificity but sacrifices a small amount of overall sensitivity. Compared with the conservative Attention UNet, all ensemble methods trade a slight rise in false positives for a larger increase in true positives, most clearly seen in the ET.

6.3 Calibration of probability maps

Probability maps were first calibrated with temperature scaling (see Section 6.3 for details). The optimisation yielded virtually identical temperatures for all three ensembles with the value of $T = 4.12$.

6.3.1 Expected Calibration Error (ECE)

In the context of trustworthy brain tumor segmentation, it is desired that the models do not only correctly predict the location of the tumor, but also that they provide reliable estimates of their own uncertainty. To evaluate this reliability, the ensemble models' calibration is assessed - that is, how well the model's predicted probability of the specific labels matches the actual likelihood of that segmentation being correct. In this thesis, ECE is computed on the softmax probabilities produced by the ensemble corresponding to each uncertainty estimation strategy. While ECE does not measure uncertainty quality directly, it is used to assess how well the ensemble's confidence aligns with correctness, providing a complementary perspective to voxel-wise uncertainty analyses.

The Expected Calibration Error (ECE) is a quantitative measure to assess the calibration of models. It measures the difference between the model's predicted confidence and the actual accuracy of those predictions, averaged across different confidence levels. A model with lower ECE provides more reliable uncertainty estimates.

In this thesis, the ECE was calculated for three different uncertainty estimation methods (TTD, TTA, and Hybrid) and for three different tumor sub-regions (NCR, ED, ET). Calculating ECE for each sub-region allows to determine whether the model's probabilities are better calibrated for some sub-regions than others, e.g. due to differences in segmentation difficulty or the characteristics of the specific sub-region.

Table 6.6 summarises the mean and standard deviation of ECE across the BraTS 2021 subjects present in the test set.

Table 6.6: Average Expected Calibration Error (ECE) for the Performance and Uncertainty-Weighted ensembles on each tumor sub-region (NCR, ED, ET).

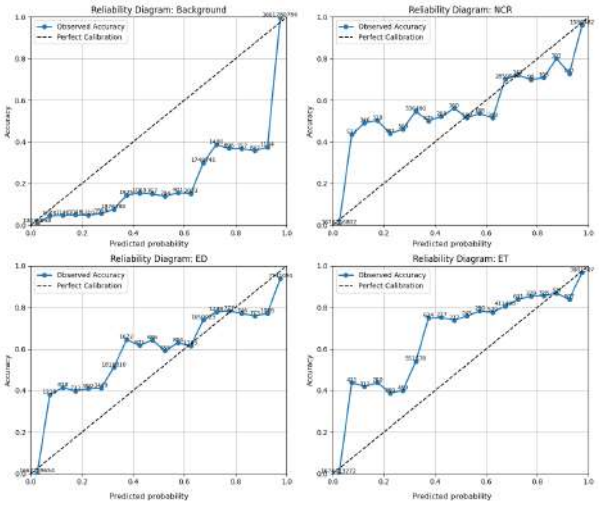
	NCR	**ED**	**ET**
TTD	0.127± 0.210	0.117 ± 0.131	0.086 ± 0.124
TTA	0.181 ± 0.215	0.179 ± 0.160	0.130 ± 0.148
Hybrid	0.129 ± 0.211	0.105 ± 0.124	0.087 ± 0.127

The Hybrid ensemble attains the lowest average ECE on two of the three regions (ED, ET) and is a close second on NCR. TTD follows closely, while TTA is consistently the least well calibrated. Wide standard deviations — often larger than the means — reveal strong case-to-case variation, especially for the small and heterogeneous NCR.

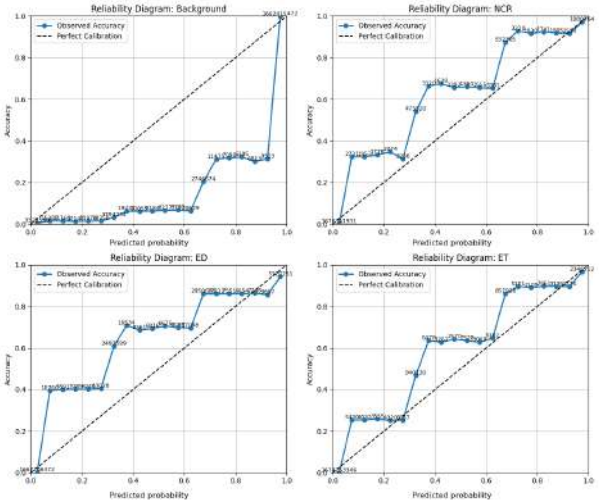
6.3.1.1 Reliability diagrams.

A reliability diagram bins predictions by their stated confidence and, for each bin, plots *empirical accuracy* (y-axis) against *predicted probability* (x-axis). Perfect calibration lies on the diagonal. Curves below the diagonal indicate over-confidence and those above under-confidence. Figure 6.25 compares the three ensembles.

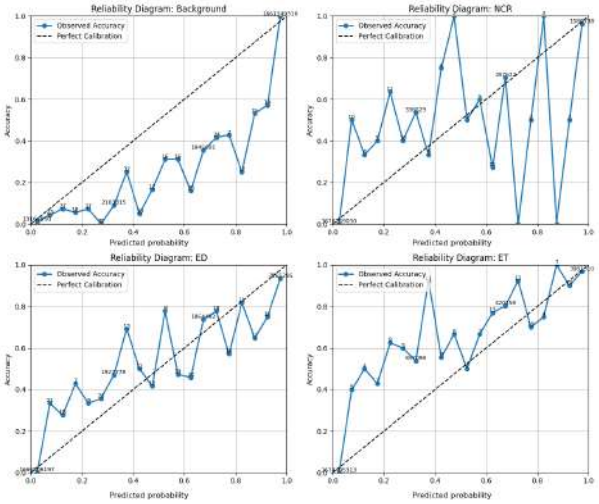
- **TTD** (Fig. 6.25a) produces curves that track the diagonal above a confidence of$\sim$0.4 for all regions, indicating mild but acceptable over-confidence in the mid-range and good calibration elsewhere.
- **TTA** (Fig. 6.25b) shows a staircase pattern: plateaus in the middle probabilities and sharp jumps at high confidence. This behaviour explains the higher ECE numbers.
- **Hybrid** (Fig. 6.25c) inherits the smoothness of TTD while slightly reducing the residual bias in ED, which leads to its best average ECE.



(a) TTD ensemble



(b) TTA ensemble



(c) Hybrid ensemble

Figure 6.25: Reliability diagrams for Performance and Uncertainty-Weighted ensembles (TTD-Only, TTA-Only, and Hybrid).

6.3.1.2 Summary

Calibration is most challenging for NCR. The reason for that is likely to be its small volume and diverse appearance that result in the highest ECE and the largest spread. Meanwhile, ET sub-region is the best calibrated one. All three curves approach the diagonal once the confidence exceeds$\sim$ 0.5. ED sits between the two extremes: the ensembles are moderately over-confident but remain within 10–18 % of perfect calibration across the full confidence range.

Overall, temperature scaling with a single global temperature already yields useful probability maps. Among the three ensembling strategies, TTD-Only and Hybrid ensembles provide better calibrated scores than the TTA-Only ensemble strategy.

6.4 Uncertainty estimation analysis

6.4.1 Uncertainty vs. error correlation analysis

To further assess the trustworthiness of the estimated uncertainty, a correlation analysis was performed between voxel-wise uncertainty and prediction error using Spearman's rank correlation. Table 6.7 shows the Spearman correlation values between voxel-wise uncertainty and prediction error across tumor sub-regions (NCR, ED, ET) for TTD-Only, TTA-Only, and Hybrid methods. In general, a well-performing uncertainty estimation method should exhibit a positive trend, where higher uncertainty is associated with higher prediction error.

Table 6.7: Spearman correlation (ρ) between voxel-wise uncertainty and NLL error. All p-values are $<10^{-3}$ because of the large voxel count.

Method	**NCR**	**ED**	**ET**
TTD	0.221	−0.042	0.070
TTA	0.282	0.268	0.131
Hybrid	0.310	0.015	0.085

6.4.1.1 TTA

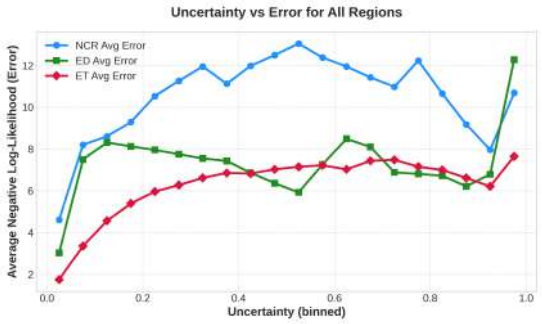
TTA showed the most consistent and monotonic increase in error across all sub-regions (Figure 6.26a), especially for NCR and ED. The corresponding Spearman correlation coefficients support this trend, with $\rho = 0.2815$ for NCR, $\rho = 0.2678$ for ED, and $\rho = 0.1307$ for ET (all $p < 0.001$). The moderate magnitude of the correlation suggests that while higher uncertainty is generally associated with higher error, the relationship is not extremely strong.

6.4.1.2 TTD

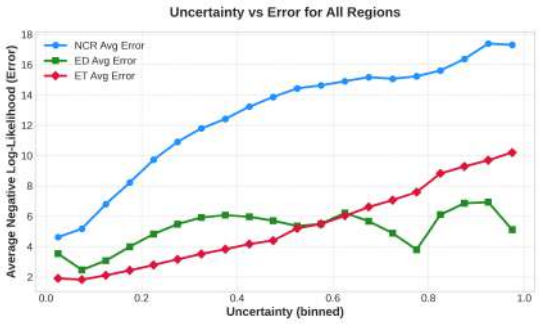
In contrast to TTA, TTD (Fig. 6.26b) shows a positive but weaker correlation for NCR (ρ=0.22) and an almost flat—or even slightly negative—relationship for ED $\rho = -0.04$). The ET curve rises late but only modestly (ρ=0.07). These patterns suggest that dropout-based epistemic uncertainty is informative in the compact, difficult NCR core but fails to flag errors in the highly ambiguous edema region, where model variance is apparently small even when the likelihood is wrong.

6.4.1.3 Hybrid

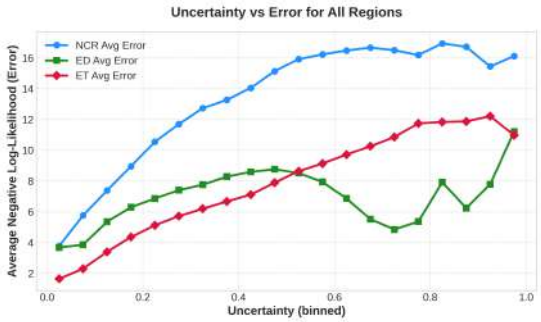
Combining epistemic and aleatoric terms (Fig. 6.26c) boosts performance on NCR to the strongest correlation in the entire table (ρ=0.31). This suggests that when both model variance and data noise are taken into account, uncertainty estimates become most informative precisely where the tissue appearance is most heterogeneous and error-prone. However, that gain does not carry over to the edema (ED) or enhancing tumor (ET) regions. In ED, Hybrid's correlation ($\rho = 0.015$) collapses almost to zero, far below TTA's 0.27 and only marginally better than TTD's slight negative association. And in ET, Hybrid (0.085) sits between TTD (0.07) and TTA (0.13) but still lags well behind TTA alone. In practice this means that blending both uncertainty sources helps most in the compact, high-variability core but offers little additional insight where the predicted boundaries are smoother or where noise dominates.



(a) TTA uncertainty versus error.



(b) TTD uncertainty versus error.



(c) Hybrid uncertainty versus error.

Figure 6.26: Combined sub-figures showing uncertainty vs error correlation.

6.4.1.4 Uncertainty quality vs. calibration

Interestingly, the best error correlation (TTA) coincides with the worst probability calibration (highest ECE in Section 6.3). Conversely, TTD exhibits superior calibration but a poorer uncertainty–error link. These findings show that a method can be well calibrated in probabilities while still failing to mark where it will be wrong, and vice versa. For uncertainty-aware clinical applications, both perspectives are important and should be considered together.

6.4.1.5 Summary

Among the three sub-regions, the highest correlation between uncertainty and error for all models was observed in NCR. This is particularly noteworthy given that NCR is widely recognized as one of the most challenging regions to segment, likely due to its small size. The strong correlation suggests that although the model struggles with accurate predictions in NCR, the uncertainty estimate is effective at identifying where those errors occur. In other words, higher uncertainty aligns well with higher error in this difficult region, demonstrating the ensemble capacity to flag

unreliable predictions in anatomically ambiguous areas.

TTA seems to offer the most informative voxel-wise uncertainty maps despite calibration shortcomings (according to the quantitative evaluation), while the Hybrid approach yielded the single best correlation in the particularly error-prone NCR region. These results provide useful insights for future explorations of uncertainty estimation in deep learning models, further exploring and utilizing the strengths of the different uncertainty estimation method for improving performance on segmentation tasks.

6.4.2 Risk coverage curves

While ECE provides a global measure of calibration by averaging the difference between probabilities and accuracy, and the correlation analysis assesses the relationship between voxel-wise uncertainty and error, risk coverage curves offer a complementary perspective by evaluating the model's ability to selectively exclude uncertain predictions. A model with well-calibrated uncertainty estimates should exhibit a risk coverage curve that starts with low risk at low coverage (i.e., when only the most confident predictions are included) and gradually increases in risk as coverage increases (i.e., as more uncertain predictions are added).

6.4.2.1 TTA

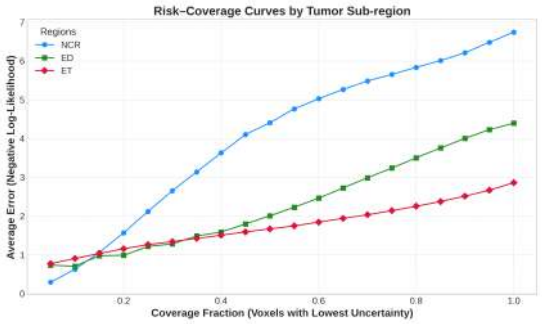
Figure 6.27 compares risk–coverage curves for each uncertainty estimation method across all tumor sub-regions. The TTA curves exhibit the steepest and most monotonic increase in risk. In NCR the average error drops by $\approx 4.5\,\mathrm{NLL}$ units when the model is allowed to abstain on the most uncertain $20\,\%$ of voxels; ED and ET show a similar but slightly less pronounced trend. The area under the RC curve (AURC) is consequently the lowest of the three methods, confirming that TTA uncertainty is highly useful for selective-prediction scenarios—even though its softmax probabilities were the least calibrated.

6.4.2.2 TTD

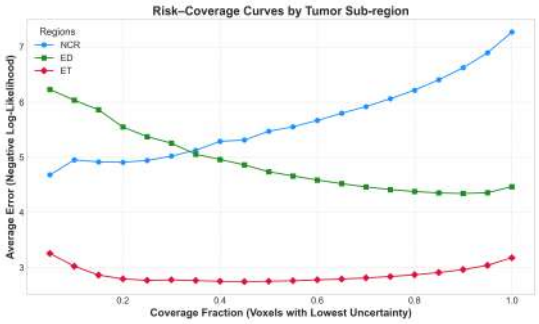
In contrast, TTD generates non-monotonic RC curves. For ED and ET the risk decreases as coverage approaches 100 %, sometimes even dipping below the risk at 80 % coverage. This discordant behaviour is consistent with the near-zero (or negative) ρ values reported earlier and indicates that voxel-wise TTD uncertainty is only weakly, and sometimes inversely, related to the true segmentation error. In practice, using TTD to reject uncertain voxels would therefore remove many correct predictions while retaining erroneous ones.

6.4.2.3 Hybrid

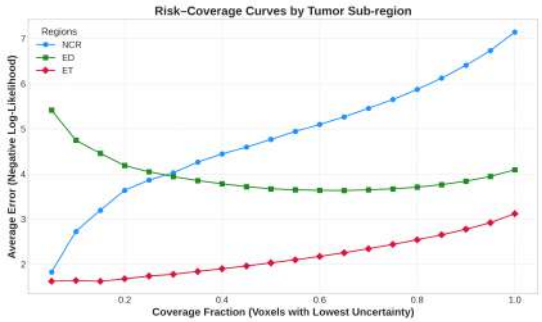
The Hybrid curves shown in Figure 6.27c strike a balance between the two extremes. They are increasing over the full range of coverages (although ED shows a slight dip at the beginning), thus avoiding the pathological inversions seen with TTD, yet remain less steep than TTA, indicating a more conservative uncertainty estimate. In the NCR, error rises smoothly from an NLL of ≈ 1.8 at 1% coverage to over 7.0 at full coverage, indicating a clean, monotonic ranking. In ED, uncertainty first filters out the very worst voxels—dropping from ≈ 5.4 down to ≈ 3.7 by about 60% coverage—then gently admits the remainder, climbing back toward ≈ 4.1. In the ET, the curve is the flattest, increasing steadily from ≈ 1.6 to ≈ 3.1. Compared to TTA, all three Hybrid curves avoid steep early rises, yielding a more calibrated uncertainty ordering—at the cost of a slight non-monotonic dip in ED before the final ascent.



(a) Risk coverage for TTA-Only ensemble.



(b) Risk coverage for TTD-Only ensemble.



(c) Risk coverage for Hybrid ensemble.

Figure 6.27: Combined subfigures showing risk coverage curves.

6.4.2.4 Summary

RC curves reinforce the key message that a good softmax calibration does not guarantee useful spatial uncertainty. TTA, despite its poor ECE, allows the clearest risk stratification. Meanwhile, TTD, although better calibrated on average, is not as suitable for selective prediction. The Hybrid strategy seems to combine the strengths of both and yields relatively monotonic RC behaviour across all tumor sub-regions and well-calibrated softmax probabilities.

6.4.3 Visual examples

To qualitatively assess the differences between the uncertainty estimation methods, the voxel-wise uncertainty maps were overlaid on MRI slices for two representative patients (Figure 6.28 and Figure 6.29). Each row corresponds to a different method — TTA, TTD, or Hybrid — and each column illustrates the uncertainty for one tumor sub-region (NCR, ED, ET), alongside the model's predicted segmentation.

6.4.3.1 Patient 00332

For Patient 00332 (Figure 6.28), all methods achieved strong performance across sub-regions. Hybrid achieved the highest overall Dice (0.819) and strong sub-region scores (e.g. Dice ED = 0.871, ET = 0.863), closely followed by TTA (overall Dice = 0.736) and TTD (0.758). Nevertheless, there is a notable variation in how uncertainty was expressed by each method:

- TTA produces a crisp yellow–red rim that tightly hugs the tumor border in all three sub-regions. Interior voxels stay mostly dark, indicating high confidence where the prediction is correct.
- TTD displays many bright, dispersed speckles all over the slice, both inside and outside the tumor. The boundary itself is only weakly emphasised. This noisy behaviour mirrors the non-monotonic risk–coverage curve reported earlier (Section 6.4.2).
- Hybrid retains the clear rim of TTA and overlays fine-grain speckle patterns like TTD—yielding a heatmap that is both boundary-focused and sensitive to isolated mis-segmented voxels.

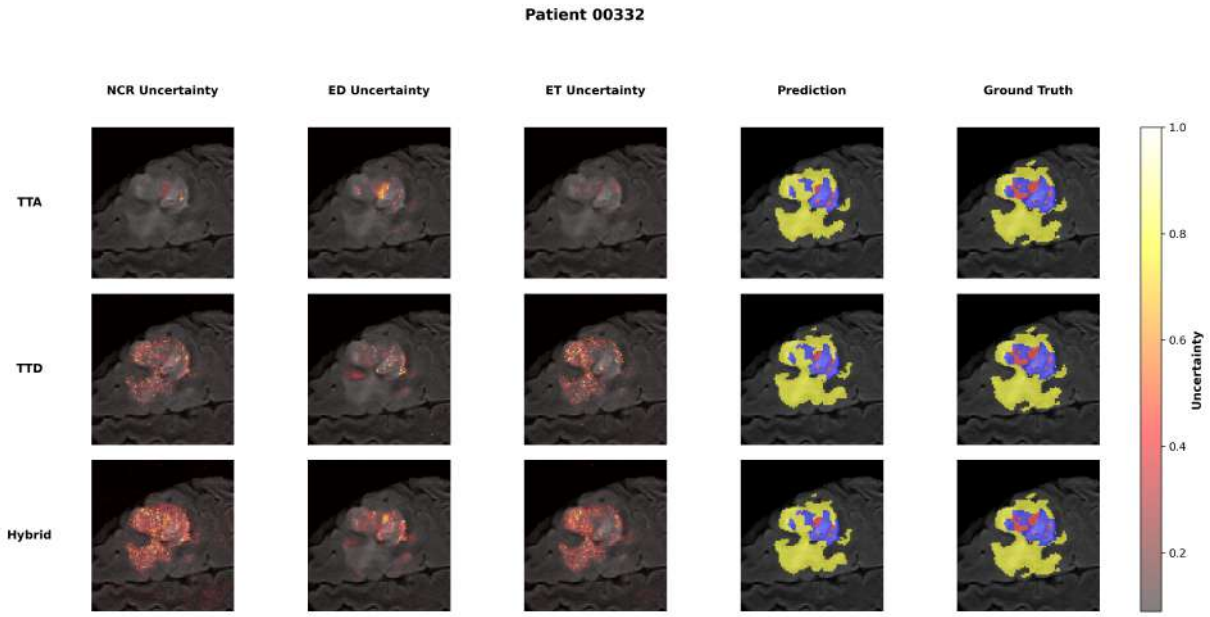


Figure 6.28: Voxel-wise uncertainty maps and predicted segmentations for Patient 00332.

6.4.3.2 Patient 01502

For the more difficult Patient 01502 (Figure 6.29), segmentation quality deteriorated across all methods. Hybrid achieved the best overall Dice (0.353), outperforming TTD (0.324) and TTA (0.103), although all three struggled to segment ED (Dice = 0.0). Importantly, TTA correctly predicted the absence of the ET sub-region, yielding Dice ET = 1.0, while both Hybrid and TTD incorrectly predicted ET (Dice = 0.071 and 0.0, respectively). Each method illustrates its uncertainty in the following ways:

- TTA correctly omits the ET class (Dice ET=1.0) and assigns high uncertainty to the difficult NCR/ED interface, faithfully reflecting its own errors.
- TTD again shows salt-and-pepper noise. Crucially, it remains confident (dark) inside a false-positive ET island, echoing the inverse trend seen in its risk coverage curve.
- Hybrid highlights both the missed NCR tongue and the spurious ET region, but without the shot-noise artefacts of TTD. Its response in ET is slightly muted compared with TTA, consistent with the more conservative slope of the Hybrid risk-coverage curve.

Quantitatively, Hybrid method reaches the highest overall Dice = 0.353 by penalising fewer false positives, while TTA sacrifices Dice but signals its uncertainty more reliably. TTD underperforms on both counts.

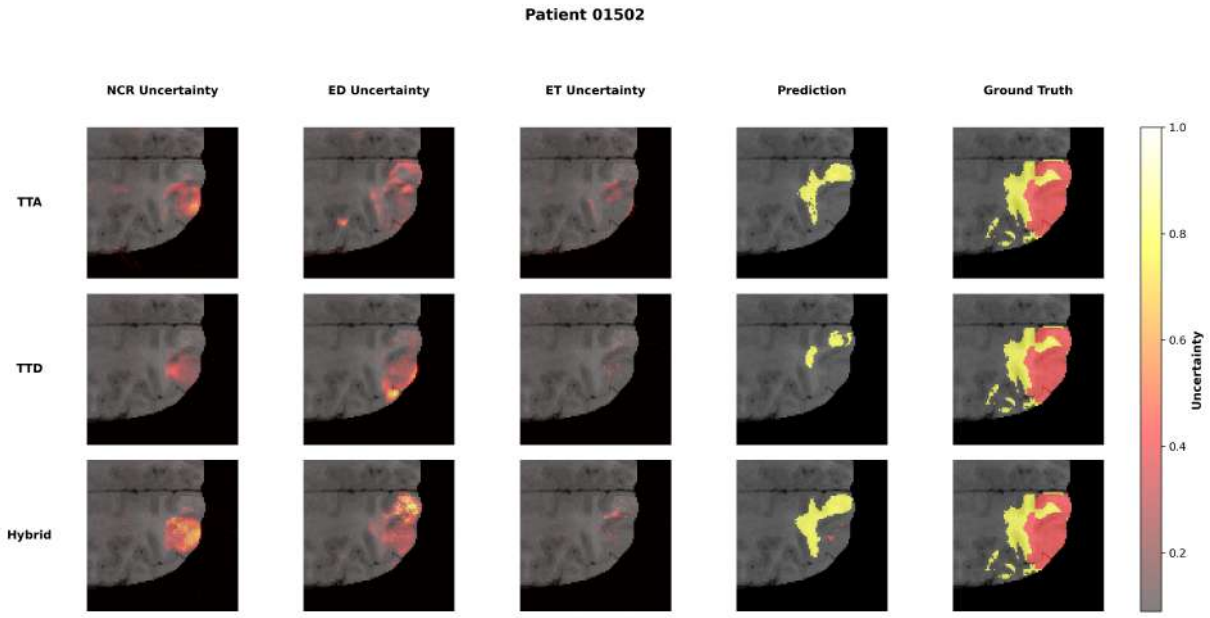


Figure 6.29: Voxel-wise uncertainty maps and predicted segmentations for Patient 01502.

6.4.4 Probability maps vs. uncertainty maps

Figures 6.30 and 6.31 present voxel-wise probability outputs and epistemic uncertainty estimates for the ED sub-region (yellow in the prediction and ground truth) using three ensembling strategies: test-time augmentation (TTA), test-time dropout (TTD) and their hybrid.

For both Patient 00332 and Patient 01502, high-probability areas align with low uncertainty, marking the lesion core as reliable. Intermediate probabilities coincide with high uncertainty, highlighting voxels where the model hesitates. In Patient 00332, the Hybrid's probability map matches TTA's smooth, slightly oversized ED region, but uncertainty forms a tight halo along the true boundary. That halo pinpoints exactly where errors are most likely.

TTD shows more variation in probabilities and a broader spread of uncertainty. Its uncertainty map fills low- and mid-probability patches—particularly at irregular edges and thin protrusions. In Patient 01502, this appears as high uncertainty in the under-segmented tail and scattered interior voxels with sub-threshold probabilities. The overlap of a probability dip and an uncertainty peak clearly marks challenging areas.

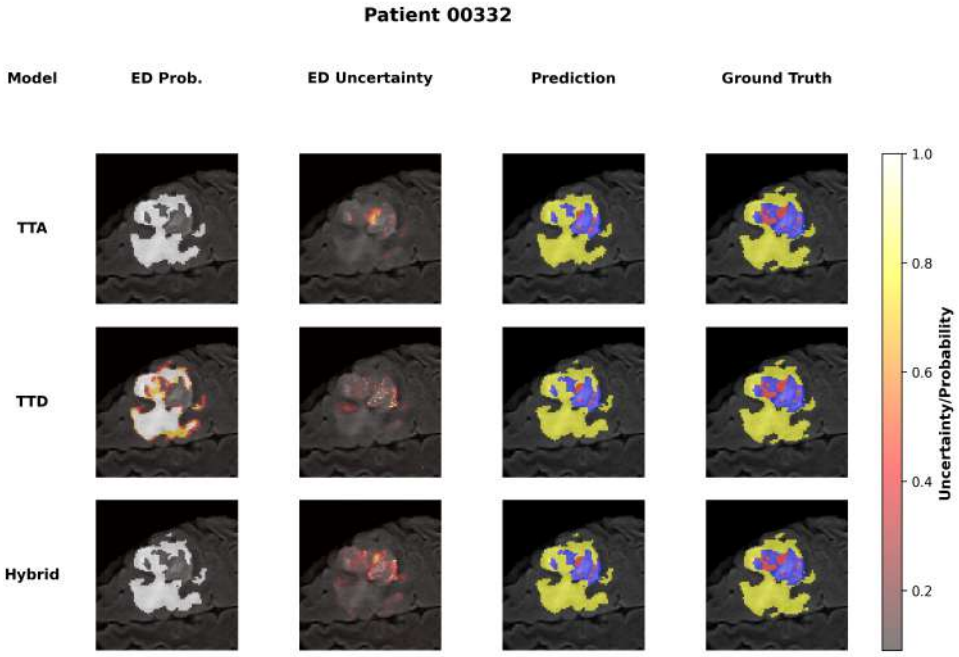


Figure 6.30: Probability and uncertainty maps for ED sub-region of Patient 00332.

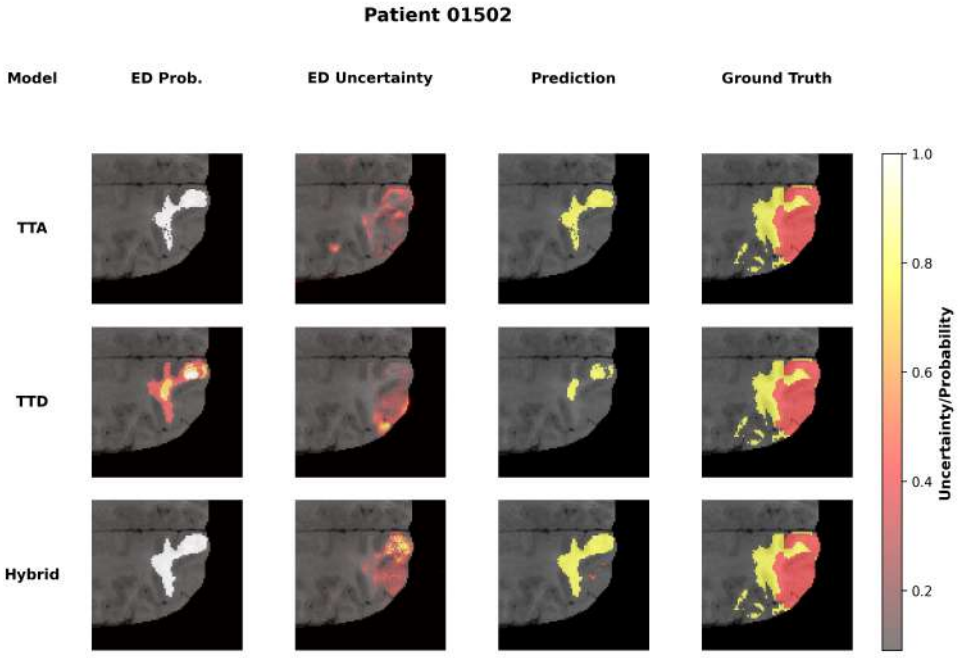


Figure 6.31: Probability and uncertainty maps for ED sub-region of Patient 01502.

6.4.4.1 Summary

These findings demonstrate the complementary roles of probability and uncertainty maps: the former quantifies the model's confidence, while the latter identifies regions susceptible to error. Consequently, voxels exhibiting high predictive probability and low uncertainty may be automatically accepted, whereas those with moderate-to-low probability combined with elevated uncertainty — particularly along the difficult-to-segment sub-regions — should be referred for expert evaluation. Out of the three ensembles, the Hybrid ensemble effectively integrates the smooth regional delineation characteristic of TTA with the boundary-sensitive uncertainty estimation of TTD.

6.5 Performance on out-of-distribution samples

To further evaluate the generalisability of the developed models, their performance was on two out-of-distribution (OOD) samples acquired from Hospital de Bellvitge in Barcelona. These cases differ from the BraTS 2021 dataset used during training and therefore serve as a benchmark to assess robustness in clinical deployment. All eight methods (Simple Averaging, Performance-Weighted, TTA, TTD, Hybrid, Attention UNet, SwinUNETR, SegResNet) were tested, and their quantitative performance is reported in Tables 6.8 and 6.9. Figures 6.32 and 6.33 illustrate qualitative segmentation examples.

6.5.1 Quantitative analysis

The quantitative performance of the brain tumor segmentation models on the two OOD cases, VIGO_01 and VIGO_03, is presented in this section. Tables 6.8 and 6.9 summarize the Dice score, Hausdorff Distance (HD95), sensitivity, and specificity achieved by each of the eight methods (Simple Averaging, Performance-Weighted, TTA, TTD, Hybrid, Attention UNet, SwinUNETR, SegResNet) for the NCR, ED, and ET sub-regions. The following subsections provide a detailed analysis of these metrics, highlighting the strengths and limitations of the different segmentation methods in the context of these unseen data. Overall, ensemble approaches (Hybrid and TTA) yield the most balanced trade-off between Dice score, HD95, and sensitivity on the ED sub-

region, but all methods fail to recover the ET region under domain shift.

Table 6.8: Segmentation metrics for patient VIGO_01

Model	Dice			HD95			Sensitivity			Specificity		
	NCR	ED	ET	NCR	ED	ET	NCR	ED	ET	NCR	ED	ET
AttUNet	1.000	0.820	0.000	0.000	15.264	114.952	1.000	0.755	0.000	1.000	0.999	1.000
SegResNet	1.000	0.794	0.001	0.000	18.466	121.65	1.000	0.677	0.000	1.000	0.999	1.000
SwinUNETR	1.000	0.834	0.000	0.000	18.708	114.952	1.000	0.742	0.000	1.000	0.999	1.000
Simple Averaging	1.000	0.817	0.000	0.000	32.155	114.952	1.000	0.709	0.000	1.000	0.999	0.999
Perf.-weight.	1.000	0.817	0.000	0.000	32.155	114.952	1.000	0.709	0.000	1.000	0.999	0.999
TTA	1.000	0.829	0.000	0.000	32.295	114.952	1.000	0.733	0.000	1.000	0.999	1.000
TTD	1.000	0.828	0.000	0.000	31.416	114.952	1.000	0.734	0.000	1.000	0.999	1.000
Hybrid	1.000	0.845	0.000	0.000	32.710	114.952	1.000	0.770	0.000	1.000	0.999	1.000

Table 6.9: Segmentation metrics for patient VIGO_03

Model	Dice			HD95			Sensitivity			Specificity		
	NCR	ED	ET	NCR	ED	ET	NCR	ED	ET	NCR	ED	ET
AttUNet	0.000	0.610	0.000	24.370	11.705	71.694	0.000	0.526	0.000	1.000	0.998	1.000
SegResNet	0.000	0.615	0.000	1.682	13.342	45.351	0.000	0.544	0.000	1.000	0.998	1.000
SwinUNETR	1.000	0.645	0.000	0.000	14.629	74.469	1.000	0.580	0.000	1.000	0.998	1.000
Simple Averaging	1.000	0.642	0.000	0.000	11.045	71.686	0.000	0.544	0.439	1.000	0.999	1.000
Perf.-weight.	1.000	0.642	0.000	0.000	11.045	71.686	0.000	0.544	0.439	1.000	0.999	1.000
TTA	1.000	0.644	0.000	0.000	10.677	88.196	1.000	0.583	0.000	1.000	0.997	1.000
TTD	1.000	0.637	0.000	0.000	11.180	71.686	1.000	0.552	0.000	1.000	0.998	1.000
Hybrid	1.000	0.652	0.000	0.000	11.747	71.686	1.000	0.621	0.000	1.000	0.997	1.000

6.5.1.1 Necrotic core

Neither OOD case contains a necrotic core in the ground truth. In VIGO_01, all methods correctly predicted its absence (Dice = 1.000, HD95 = 0 mm). In VIGO_03, Attention UNet and SegResNet produced false positives (Dice = 0.000; HD95 = 24.37 mm and 1.68 mm, respectively), whereas SwinUNETR and all five ensemble methods (Simple Averaging, Performance-Weighted, TTA-Only, TTD-Only, Hybrid) again omitted the NCR region perfectly (Dice = 1.000, HD95 = 0 mm).

6.5.1.2 Edema

On VIGO_01, the Hybrid ensemble achieved the highest edema Dice (0.845) and sensitivity (0.770), although its boundary error (HD95 = 32.71 mm) was larger than that of Attention

UNet (HD95 = 15.26 mm, Dice = 0.820, sensitivity = 0.755). SwinUNETR delivered the next lowest HD95 (18.71 mm) with Dice = 0.834 and sensitivity = 0.742. On VIGO_03, edema Dice scores ranged from 0.610 (Attention UNet) to 0.652 (Hybrid), and ensemble methods again outperformed single models: TTA-Only produced the lowest HD95 (10.68 mm), while Hybrid retained the highest sensitivity (0.621). Specificity for edema remained above 0.997 for every method.

6.5.1.3 Enhancing tumor

No method achieved any Dice overlap for ET in either case (all Dice = 0.000). In VIGO_01, HD95 values were extremely high (114.95–121.65 mm) and sensitivity was zero across the board. In VIGO_03, boundary errors improved (45.35 mm for SegResNet up to 88.20 mm for TTA-Only). Notably, Simple Averaging and Performance-Weighted ensembles did predict some ET voxels (sensitivity = 0.439) but these did not coincide with the true ET region, yielding zero Dice. Specificity for ET remained at 1.000 for all methods.

6.5.1.4 Summary

Ensembles—particularly the Hybrid method—consistently outperformed or matched single models on OOD cases (see Tables 6.8–6.9). For edema, Hybrid had the highest Dice on both VIGO_01 and VIGO_03 and kept HD95 within the range of other ensemble methods. Single networks sometimes edged out ensembles on boundary precision in VIGO_01 but not in VIGO_03. All methods perfectly omitted NCR in VIGO_01; only Attention UNet and SegResNet produced NCR false positives in VIGO_03. No method detected ET under domain shift.

Overall, these results underscore the value of ensembling for robust ED segmentation across sites and the urgent need for domain-adaptation or uncertainty-based QC to catch the uniformly missed ET component.

6.5.2 Visual examples

6.5.2.1 Predicted segmentations

Figures 6.32 and 6.33 show example FLAIR slices and overlaid predictions for VIGO_01 (top row) and VIGO_03 (bottom row). Yellow indicates predicted ED and blue indicates ET in the ground truth.

VIGO_01 All methods successfully localize the large ED sub-region in both hemispheres. Individual models (Attention UNet, SegResNet, SwinUNETR) sometimes produce small holes or spurious islands within the main lesion (e.g. the grey patch in Attention UNet), whereas ensemble masks (Simple Averaging, Performance-Weighted, TTA, TTD, Hybrid) are more spatially coherent and fill interior gaps. None of the methods recover the small enhancing-tumor island (blue) anterior to the ventricle, reflecting the zero-Dice ET performance reported in Table 6.8. Attention UNet and SegResNet also generate minor false positives around the cortex in VIGO_01, consistent with their non-zero NCR sensitivity on this case.

VIGO_03 In this case, the dominant ED tissue is well delineated by all ensembles, with Hybrid producing the smoothest, most contiguous boundary. Individual methods again show fragmented predictions: SegResNet under-segments the inferior lobe, and SwinUNETR yields a patchy inferior extension. No model captures the enhancing region (blue), mirroring the uniform ET failure (Dice=0) in Table 6.9. Furthermore, Simple Averaging and Performance-Weighted introduce a few isolated false positives near the cortex, whereas TTA, TTD, and Hybrid keep false positives minimal.

Overall, the visual examples confirm that ensemble approaches enhance spatial consistency and reduce fragmentation, but all methods still fail to detect small enhancing-tumor regions under domain shift.

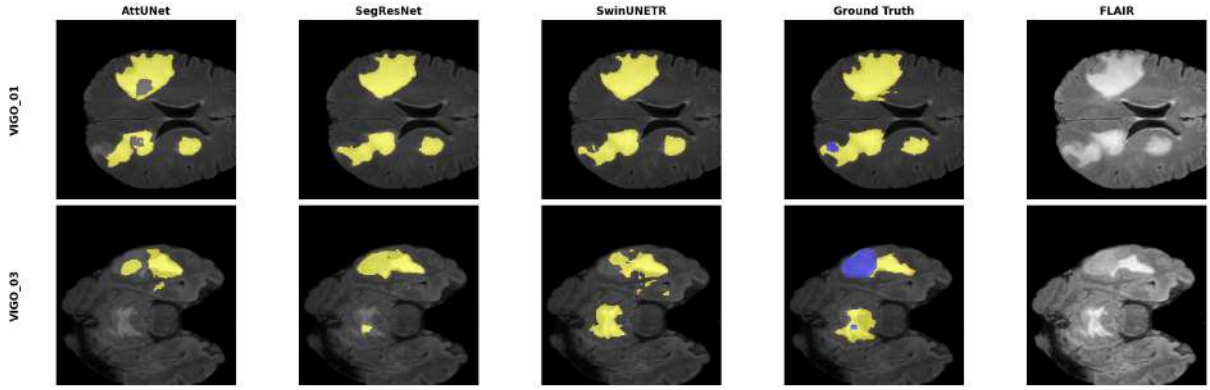


Figure 6.32: Qualitative comparison of individual model predictions on two out-of-distribution cases (VIGO_01 and VIGO_03). Red color indicates NCR sub-region, yellow signifies the ED tissue, and blue the ET tissue.

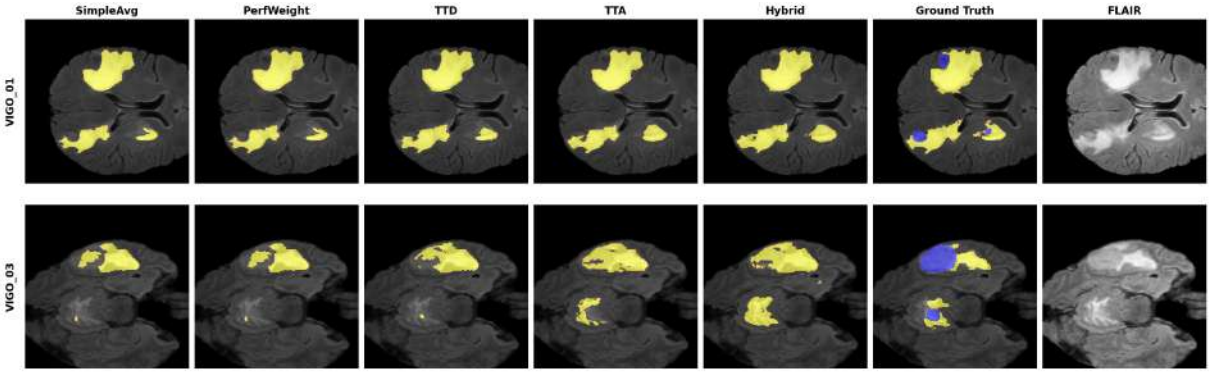


Figure 6.33: Qualitative comparison of ensemble model predictions on two out-of-distribution cases (VIGO_01 and VIGO_03). Red color indicates NCR sub-region, yellow signifies the ED tissue, and blue the ET tissue.

6.5.2.2 Uncertainty maps

Figures 6.34 and 6.35 display voxel-wise uncertainty overlaid on FLAIR for the three ensemble methods (TTD, TTA, Hybrid) and the three sub-regions (NCR, ED, ET).

VIGO_01

Across all three uncertainty-weighted ensemble strategies, uncertainty concentrates along class boundaries, with the most prominent hot-spots located in the ET sub-region that every method fails to segment (cf. Fig. 6.33). TTD exhibits the sparsest pattern: a thin red rim surrounds the ED and NCR, while the ET area is highlighted only weakly—reflecting dropout's

tendency to under-estimate overall uncertainty. TTA behaves in the opposite way: it produces a diffuse "speckle" of low-to-moderate uncertainty throughout the slice and a broad yellow patch covering the ET, indicating higher recall but lower precision in its uncertainty signal. The Hybrid map inherits the TTA hot-spot over ET yet suppresses most of the background speckle, yielding a sharper, reviewer-friendly cue. Inside the large bilateral ED mass, all methods show uniformly low uncertainty, consistent with the near-perfect ED Dice reported in Table 6.8.

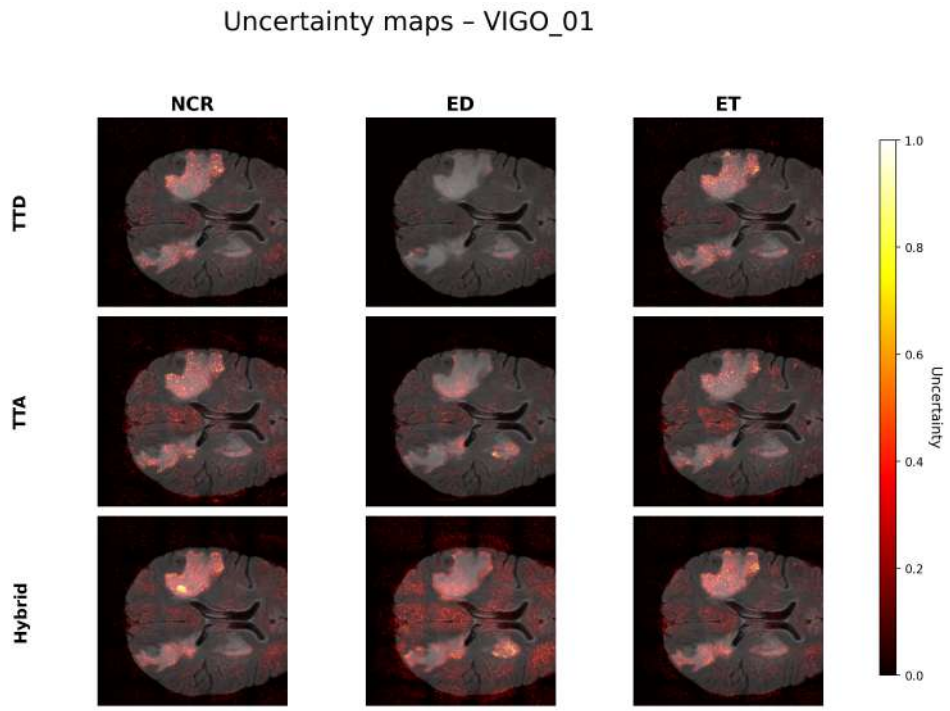


Figure 6.34: Uncertainty maps for patient VIGO_01.

VIGO_03

The sagittal slice of VIGO_03 reveals the same ordering of methods. TTD again confines uncertainty to a narrow rim, providing the cleanest but least sensitive map. TTA displays a lattice-like pattern of red voxels in normal cortex and marks almost the entire superior tumor surface yellow, mirroring its fragmented segmentation mask. The Hybrid ensemble reduces the cortical noise while preserving the high-uncertainty plume over the missed ET lobule. Notably, the interior of the ED remains low-uncertainty for all methods, whereas voxels on the tumor–brainstem interface show elevated uncertainty, in line with the small false-positive clusters produced by the individual models (Fig. 6.32).

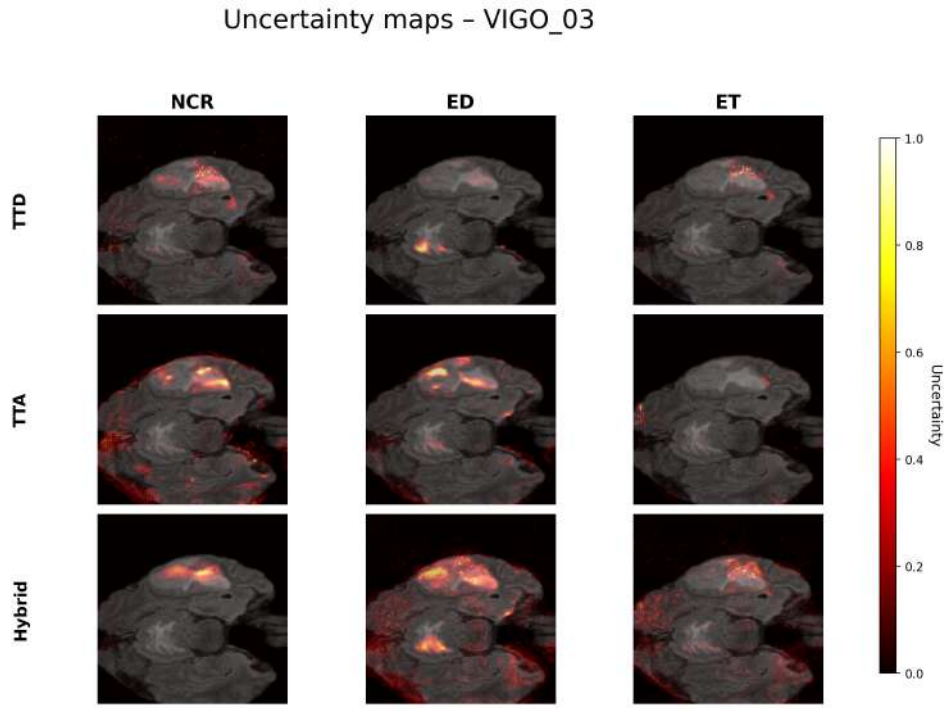


Figure 6.35: Uncertainty maps for patient VIGO_03.

Summary

In the two out-of-distribution cases considered, voxels exhibiting higher uncertainty generally corresponded to mis-segmented regions, most notably the undetected ET tissue. TTD tended to produce conservative, rim-shaped uncertainty maps with relatively few false positives, whereas TTA offered broader coverage at the expense of increased background noise. The Hybrid approach appeared to strike a compromise, retaining much of TTA's sensitivity while reducing spurious signals toward the specificity levels of TTD. Notably, the Hybrid uncertainty maps' high-contrast voxels occupied less than 5% of the slice yet encompassed a substantial fraction of the observed segmentation errors, suggesting that manual reviewers could focus their attention on this limited subset of voxels.

Collectively, these visualisations align with the quantitative findings and indicate that ensemble-based epistemic measures may serve as useful indicators of model failures under domain shift. While additional validation across larger and more diverse cohorts is definitely required, the Hybrid strategy emerges as a promising candidate for enhancing downstream quality-control

workflows.

6.6 Web application

6.6.1 Segmentation workflow

When a user uploads four co-registered volumes (FLAIR, T1, T1CE, T2), the application immediately displays them in synchronized slice views (Figs. 6.37–6.39). Preprocessing (skull-stripping, reslicing, intra-modal registration) is started with a single click; progress bars and subtle dimming on each panel convey status (Fig. 6.38).

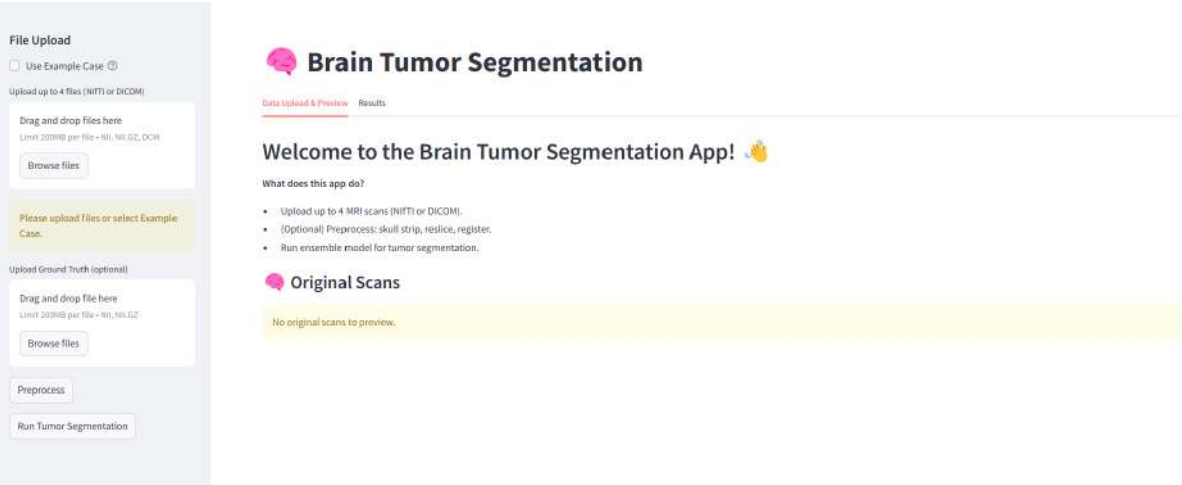


Figure 6.36: Starting Page

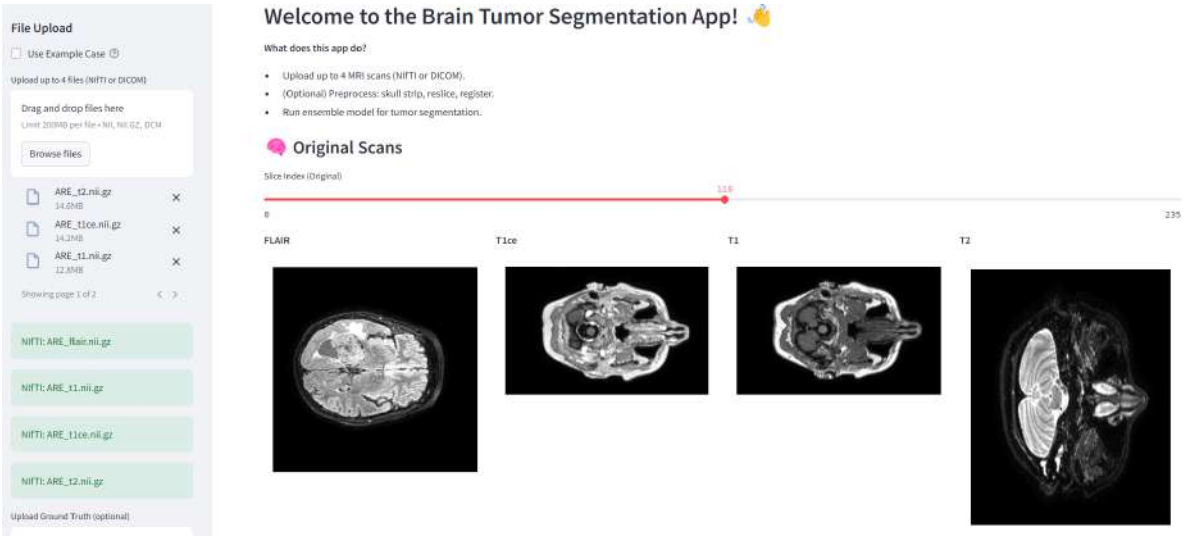


Figure 6.37: Uploading original, raw MRI scans.

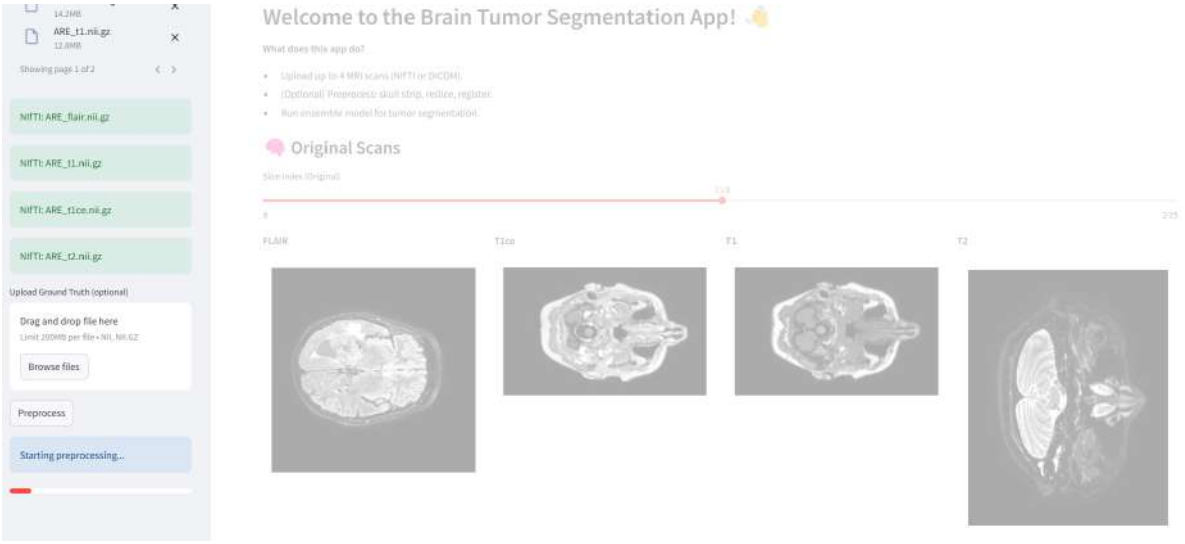


Figure 6.38: Preprocessing in progress.

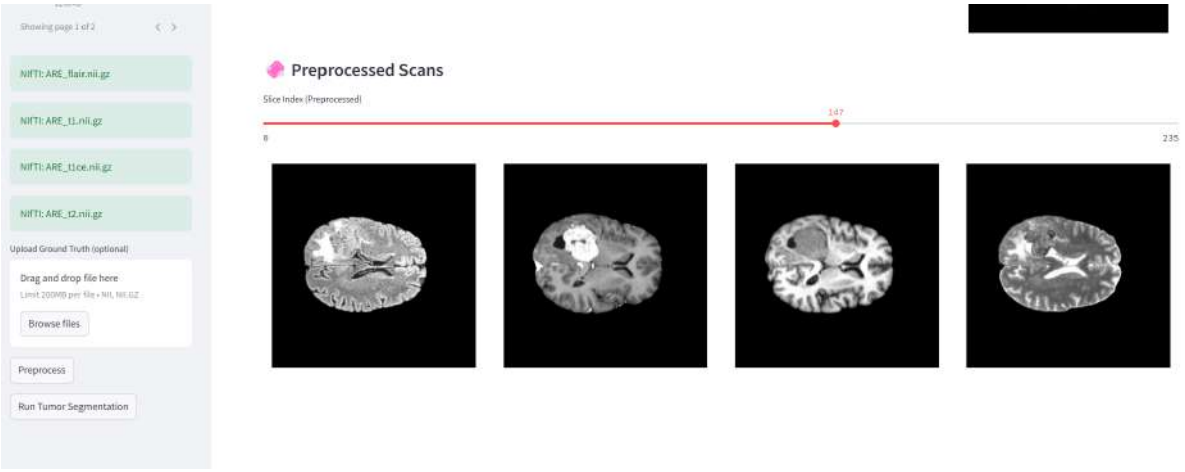


Figure 6.39: Preprocessed MRI scans displayed on the website.

Once preprocessing completes (≈ 144.20 s on RTX 4060 GPU), clicking Run Tumor Segmentation launches the ensemble model, with real-time feedback (Fig. 6.40).

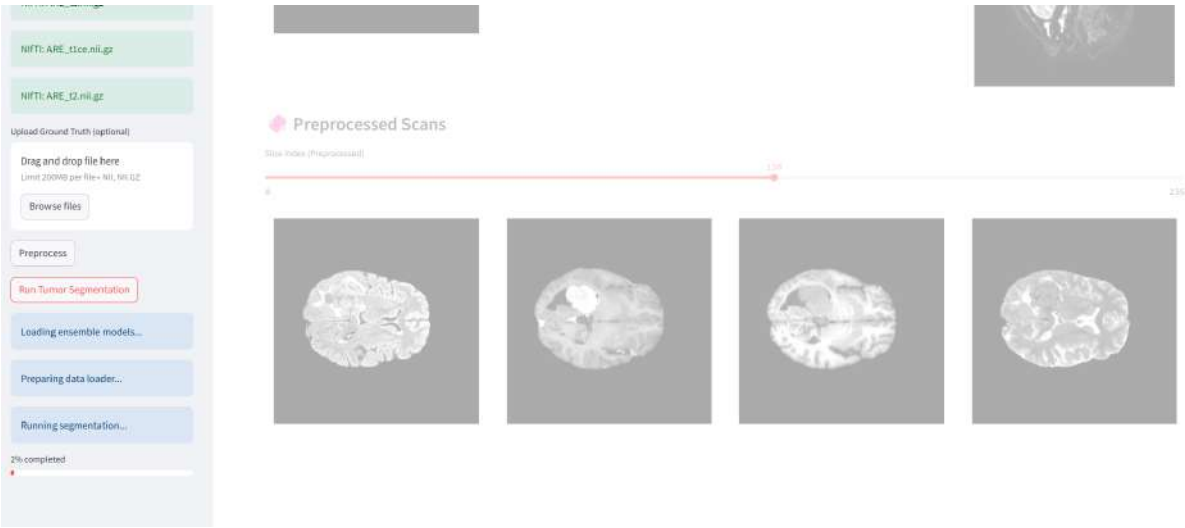


Figure 6.40: Segmentation in progress

Once the segmentation finishes (≈ 300 s, depending on file size) In the Results & Visualization tab, users can:

- Toggle between predicted segmentation and ground truth (Fig. 6.41 and 6.42),
- Show or hide individual sub-regions (NCR, ED, ET) (Fig. 6.41),
- Overlay softmax probability heatmaps or voxel-wise uncertainty maps (Fig. 6.43),

- Adjust threshold sliders to highlight voxels above (or below) a chosen confidence level (Fig. 6.41).

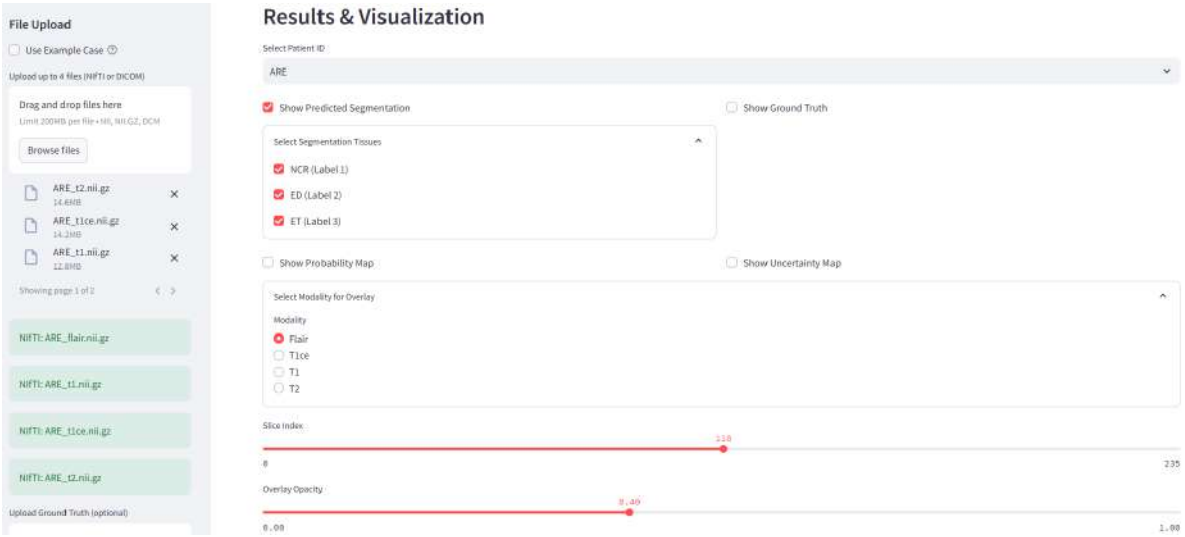


Figure 6.41: Options to select on the results tab.

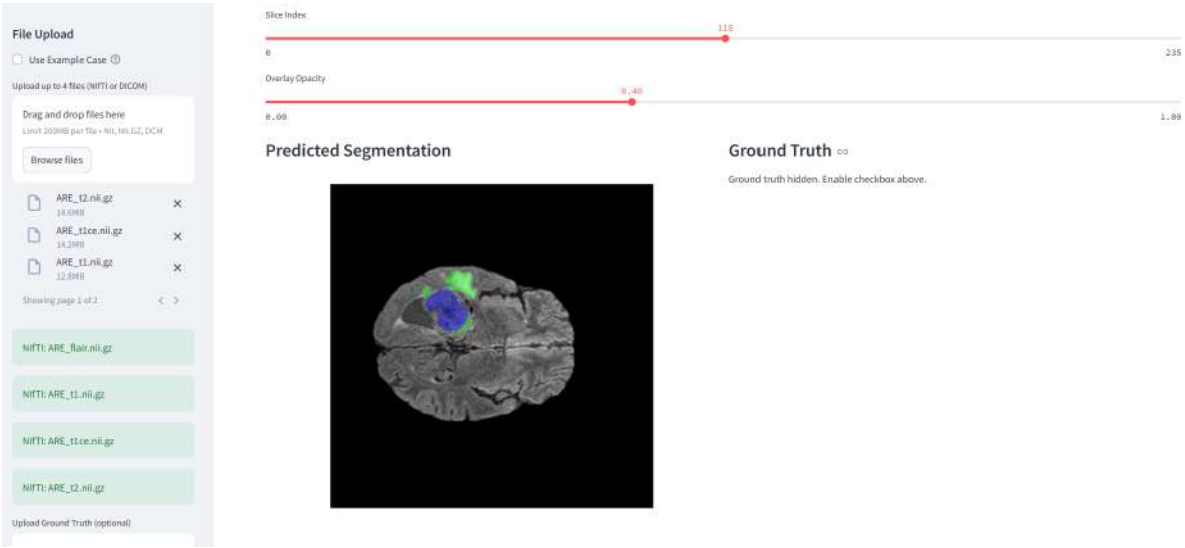


Figure 6.42: Predicted segmentation.

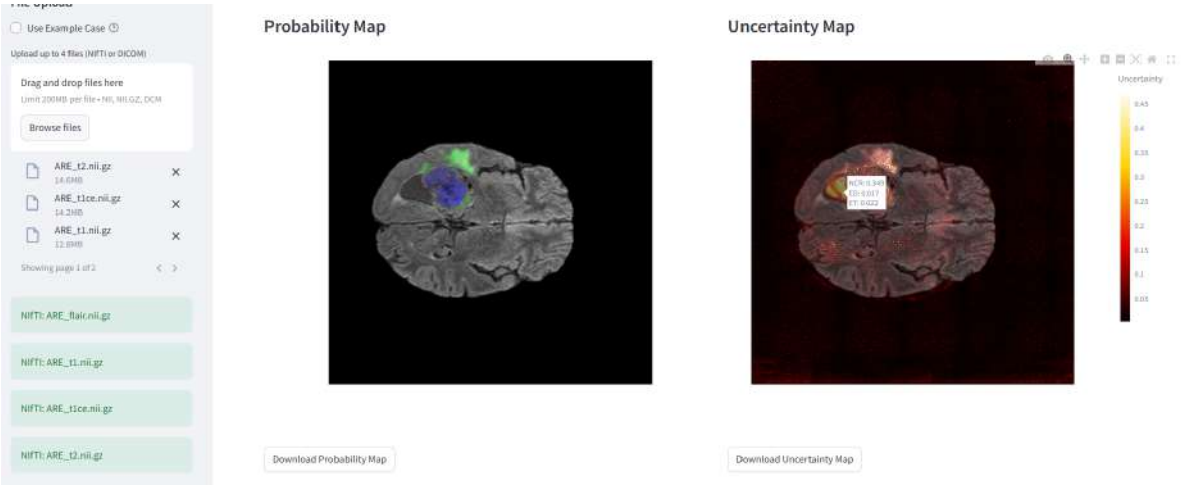


Figure 6.43: Probability map and uncertainty map.

Finally, quantitative volumetric outputs are computed in real time: the app reports volumes (in cm^3) for each tumor sub-region and the total tumor burden (Fig. 6.44), and provides buttons to download the overlaid segmentation figure, raw probability map, and uncertainty map.

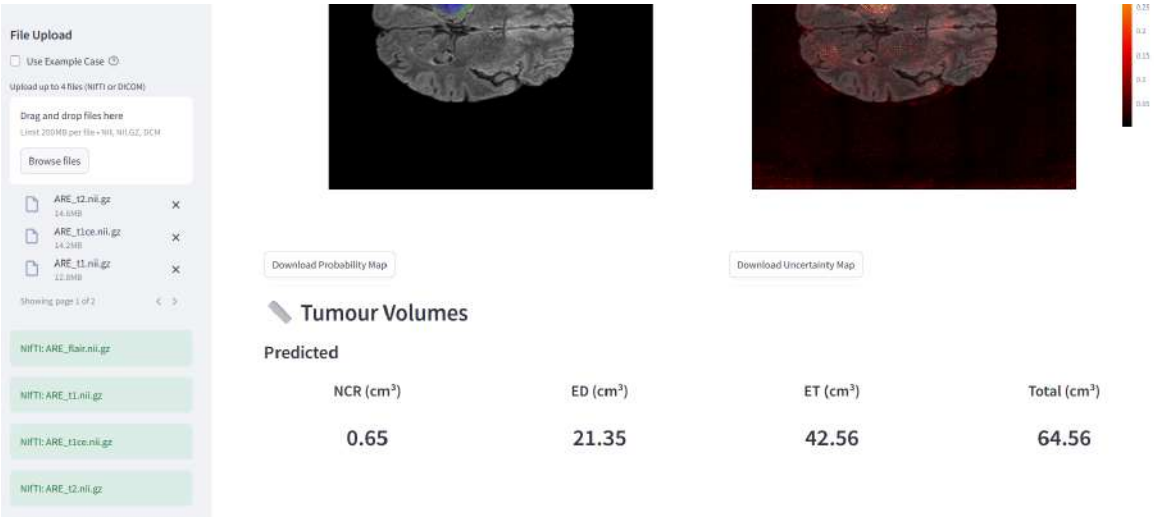


Figure 6.44: Tumor volumes are displayed below the visualizations.

When provided with ground truth masks (BraTS 2021 validation set), the application reports per-slice Dice scores and Hausdorff distances alongside the images (Fig. 6.46).

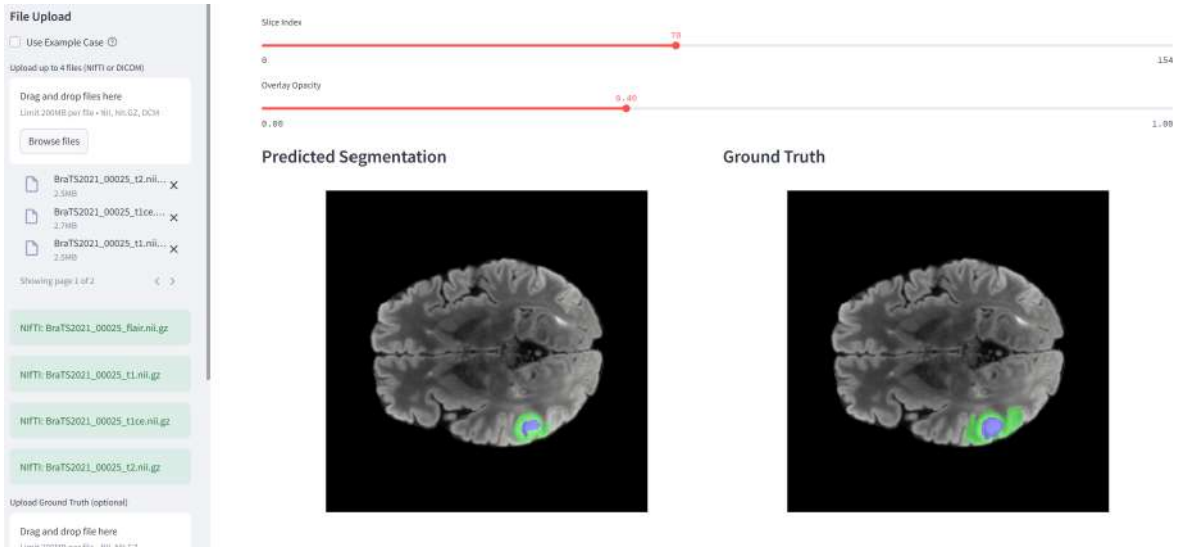


Figure 6.45: Prediction and ground truth shown for patient 00025 from the BraTS dataset.

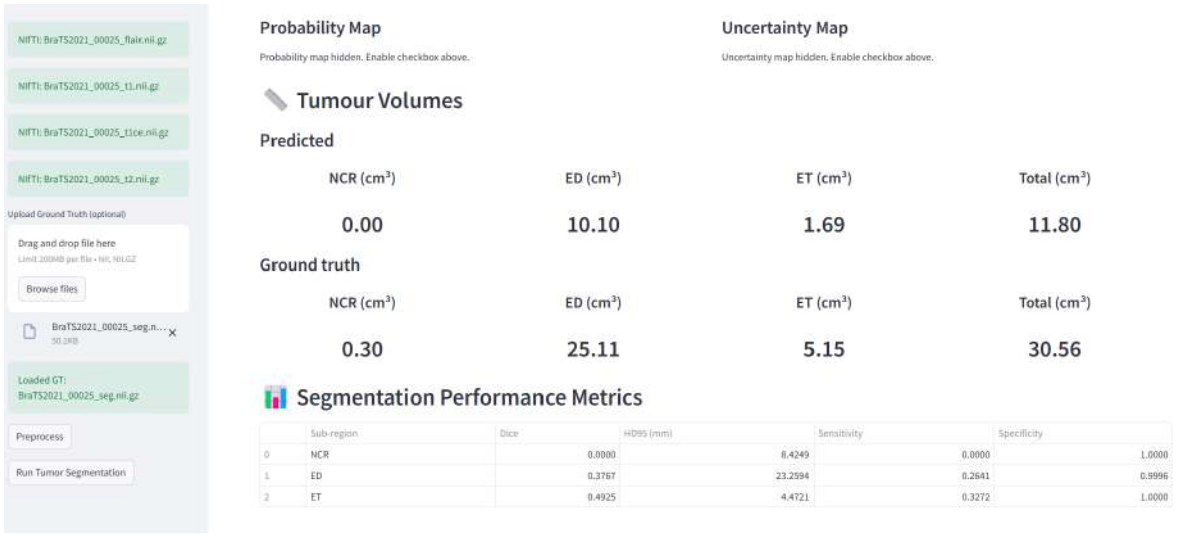


Figure 6.46: Performance metrics

6.6.2 Status of requirements

Table 6.10 confirms that all ten functional requirements (FR1–FR10) have been met. In particular, the real-time interactive overlays (FR7) and on-demand performance metrics (FR10) demonstrate that the application not only implements the core segmentation pipeline but also delivers the user controls and feedback essential for clinical usability.

Table 6.10: Compliance with Functional Requirement

ID	Requirement Description	Met?
FR1	The system shall accept up to 4 MRI scans in `DICOM` or `NIfTI` format.	✓
FR2	The system shall convert `DICOM` files to `NIfTI` format automatically.	✓
FR3	The system shall allow optional preprocessing steps including skull stripping and image registration.	✓
FR4	The system shall reorder modalities to match the expected input order (FLAIR, T1CE, T1, T2).	✓
FR5	The system shall execute a deep learning ensemble model for tumor segmentation.	✓
FR6	The system shall output segmentation masks, softmax probability maps, and voxel-wise uncertainty maps.	✓
FR7	The system shall allow interactive slice-by-slice visualization with adjustable overlays.	✓
FR8	The system shall compute and display tumor sub-region volumes in cubic centimeters.	✓
FR9	The system shall allow users to download thresholded uncertainty maps and segmentation figures.	✓
FR10	The system shall output performance metrics if the user provides the ground truth segmentation to compare the predicted segmentation.	✓

Table 6.11 shows full compliance with non-functional requirements. Fast turnaround (NFR3) and local data processing (NFR5) ensure both efficiency and patient privacy, while modular code organization (NFR4, NFR6) lays the groundwork for future extensions.

Table 6.11: Compliance with Non-Functional Requirements

ID	Requirement Description	Met?
NFR1	The interface must be intuitive and usable by clinicians or researchers without technical expertise.	✓
NFR2	The application requires access to a GPU for model inference and must be run on hardware with CUDA-compatible support.	✓
NFR3	The system should return segmentation results within five minutes for typical scan sizes.	✓
NFR4	The application must support modular integration of alternative models and postprocessing steps.	✓
NFR5	The system must process data locally to ensure privacy and compliance with medical data regulations.	✓
NFR6	The codebase must be structured in a modular and maintainable way.	✓
NFR7	The system must provide clear error messages for file format or runtime issues.	✓
NFR8	The design must support future extensions such as batch processing or clinical data integration.	✓

Chapter 7

Discussion

7.1 Summary of key findings and contributions

This thesis addressed the crucial challenge of enhancing the clinical utility of automated brain tumor segmentation by explicitly integrating uncertainty estimation within ensemble methods. The investigation involved a comprehensive evaluation of four distinct 3D architectures and five ensemble strategies on the BraTS 2021 dataset, followed by a rigorous analysis of the resulting segmentations and associated uncertainty estimates. The key findings and contributions of this work span several critical areas, each of which will be discussed in detail in the subsequent sections of this chapter:

- **Performance of individual state-of-the-art segmentation models**: The study established the strengths, weaknesses, and inherent biases of V-Net, Attention UNet, SegResNet, and SwinUNETR, providing a baseline for ensemble evaluation.
- **Effectiveness of ensemble strategies**: The investigation demonstrated the improvements in segmentation accuracy and robustness achieved by various fusion techniques, including the uncertainty-aware methods.
- **Added value of uncertainty estimation**: A comprehensive evaluation of uncertainty maps through calibration, error correlation, and risk-coverage analysis highlighted their potential for clinical decision support.

- **Contextualization and implications**: The thesis contextualizes the findings by comparing the best ensembles to the state-of-the-art, discussing study limitations and future directions, and outlining potential clinical implications.

The following sections will delve deeper into each of these key areas, providing a detailed analysis of the findings and their significance for the advancement of trustworthy and clinically meaningful brain tumor segmentation.

7.2 Performance of individual segmentation models

Understanding the baseline performance and inherent biases of the individual segmentation models – V-Net, Attention UNet, SegResNet, and SwinUNETR – is a crucial first step in evaluating the potential benefits of ensemble strategies. While a comprehensive analysis involving feature map and attention heatmap visualization lies beyond the scope of this discussion, examining their overall performance, architectural nuances, and correlations with radiomic features provides valuable insights into their strengths and limitations, ultimately informing the ensemble design implemented in this thesis.

7.2.1 Overall performance comparison

Overall, the three models that were ultimately included in the ensemble models, performed on a similar level, showing no statistically significant differences in performance. However, the V-Net model significantly underperformed compared to the other three models. It achieved the lowest Dice scores across all sub-regions (e.g. average Dice = 0.6432 ± 0.2250), and its HD95 distances were substantially higher, indicating poorer boundary delineation and more outlier errors.

Despite the differences in architectures, most models performed similarly on the brain tumor segmentation task. This aligns with a recent review found that U-Net variants, attention-augmented CNNs, and transformer-based architectures typically achieve comparable Dice scores (differences within 1–2%) on multiple medical segmentation benchmarks, with no one family consistently dominating across tasks [137]. Moreover, on well-studied datasets like BraTS, performance nears annotation variability limits, so architectural gains might yield diminishing

returns. Furthermore, all models employ extensive augmentation, regularization, and optimization strategies that may narrow performance gaps. The significant underperformance of V-Net provides the first key insight: not all models are equally suitable for inclusion in an ensemble, and its exclusion was a necessary first step to maximize ensemble potential.

7.2.2 Why V-Net struggles

A post-hoc Mann–Whitney U test confirmed that V-Net's performance was significantly worse than each of the other three models ($p \leq 0.0068$). One key factor that could be the cause of the model's underperformance is its reliance on batch normalization, which generally requires larger batch sizes to estimate reliable mean and variance statistics [138]. Due to GPU memory limits, the batch size in this study was limited to only 1 case per batch. This likely caused unstable normalization and degraded convergence in V-Net's deep, volumetric convolutions. This highlights the importance of considering the interplay between model architecture and training constraints. In contrast, SegResNet and Attention UNet models explicitly use GroupNorm, which is independent of batch sizes, and its accuracy is stable in a wide range of batch sizes [139]. SwinUNETR uses LayerNorm in its Transformer layers, which, like GroupNorm, does not depend on batch statistics.

7.2.3 Complementary inductive biases

Despite a lack of significant differences across the single models, individual model's inductive bias gave them a slight edge in different sub-regions, e.g. SegResNet on ET (sensitivity of ≈ 0.82) and Attention U-Net on NCR (sensitivity of ≈ 0.73). A possible reason why SegResNet had the best sensitivity on the NCR tissue could be found in its residual connections. SegResNet is built from stacked "ResBlock" modules that add identity-skip connections across convolutions [45]. Those skip connections allow for preserving low-frequency information from earlier layers with high-frequency information in subsequent layers [140]. As a result, the SegResNet model is capable of delineating high-contrast boundaries like those of the ET sub-region, but also preserving boundary delineation as can be seen in its low HD95 distance (average HD95 = 6.9421 ± 6.9130). Meanwhile, in the Attention UNet, attention gates allow the network to

focus on subtle, low-contrast necrotic regions. In one BraTS-21 study, a 3D Attention U-Net outperformed a plain U-Net on the tumor core (ET+NCR) task—improving core accuracy by $\approx 2\%$—implying better sensitivity on necrotic parts as well.

The strength of the SwinUNETR model lies in its transformer block, which allows for computing attention across non-overlapping windows that shift between layers and therefore, each voxel can "see" and integrate information from distant parts of the volume. This can be seen in the example of patient 01405 presented in Figure 6.14 in the Results chapter. While SegResNet and Attention UNet completely miss the NCR and ET sub-regions of the tumor, possibly treating it as noise (average Dice score of ≈ 0.50 and ≈ 0.42, respecitvely), and V-Net delineates most of the tumor as the ET sub-region (average Dice ≈ 0.39), SwinUNETR predicts the tumor boundaries most reliably, although still, with a moderate performance of average Dice ≈ 0.62. SwinUNETR's global attention plus edge-preserving decoder together let it see the ET ring and correctly predict that it is a tumor boundary rather than patchy noise.

These distinct sub-region strengths provide a clear rationale for combining these models, aiming to create a more robust and accurate overall segmentation by capitalizing on their complementary "ways of perceiving" the tumor.

7.2.4 Radiomic feature correlations

Another interesting exploration in this thesis was the correlation between MRI radiomic features (e.g. first-order entropy, minimum intensity) and each model's sub-region Dice scores. Notably, SegResNet and V-Net exhibited far more significant correlations with necrotic-core (NCR) Dice—16 and 18 features, respectively—than Attention U-Net (3 features) and SwinUNETR (4 features) (see Fig. 6.11). A likely reason for such a difference is the architecture of those models:

- V-Net and SegResNet, built on purely convolutional and residual blocks, rely heavily on first- and second-order intensity/textural cues (e.g. first-order entropy in all modalities) to localize and delineate the NCR region.
- Attention UNet's attention gates and SwinUNETR's transformer self-attention learn to

integrate richer context and consequently show less dependence on individual radiomic metrics.

7.2.5 Summary

In summary, the analysis of individual model performance reveals a landscape of comparable overall accuracy among Attention UNet, SegResNet, and SwinUNETR, yet with distinct subregion strengths stemming from their unique architectural inductive biases. The clear underperformance of V-Net, likely attributable to batch normalization limitations in this setup, justified its exclusion from subsequent ensemble experiments. The key takeaway from this section is that the observed complementary strengths of the remaining three models provide a strong foundation for investigating the potential of ensemble methods to surpass the performance of any single model alone, which will be the focus of the next section.

7.3 Effectiveness and insights from ensemble methods

Combining multiple segmentation models is a well-established strategy to boost performance in brain tumor MRI segmentation. In the 2019 BraTS challenge, the winning team fused 12 different 3D U-Net variants, pushing ET Dice to ∼0.83 (versus ∼0.79 with single models) [141]. Recent winners of the BraTS 2023 challenge leveraged the heterogeneity of segmentation models, mixing a CNN-based nnU-Net with a Transformer-based Swin UNETR to exploit complementary strengths, achieving higher Dice scores and lower boundary errors than any single model [72]. Here, SegResNet, Attention U-Net, and SwinUNETR were ensembled (V-Net was excluded for its poor standalone performance), leveraging their complementary biases—local detail vs. global context—via three strategies:

- Simple averaging
- Performance-weighted averaging
- Performance & uncertainty-weighted fusion, where voxel-level weights derive from (a) TTA-Only, (b) TTD-Only, and (c) a Hybrid of TTA+TTD uncertainties.

The following subsections present the main findings regarding the performance of the different ensemble strategies.

7.3.1 Marginal Dice gains reflect high model homogeneity

The ensemble approaches developed in this research demonstrated robust segmentation performance, particularly for the ET sub-region. However, they did not reach the peak accuracy reported by the top BraTS challenge winners from 2021 to 2023. These leading solutions often employ large ensembles of highly optimized 3D U-Nets, such as nnU-Net 2.0 [142], frequently incorporating mechanisms like axial attention [142] or STAPLE fusion [143] to achieve impressive Dice scores on the hidden test set (Whole Tumor (WT=NCR+ED+ET) ≈ 0.92, Tumor Core (TC=NCR+ET) ≈ 0.88, and ET ≈ 0.88). In contrast, the presented Hybrid test-time augmentation (TTA-Only) ensemble achieved Dice scores of 0.733 (NCR), 0.783 (ED), and 0.829 (ET) on the same BraTS 2021 split, placing the implemented methods approximately 5–10 percentage points below the challenge winners for each sub-region. This difference is likely attributable to factors such as the scale of the ensemble and the absence of external data or extensive test-time augmentation techniques commonly employed by the top-performing teams.

All three ensemble variants — simple averaging (Dice ≈ 0.7761), performance-weighted ($\approx$ 0.7762), and uncertainty-aware (TTA-only ≈ 0.7816, TTD-only ≈ 0.7784, Hybrid ≈ 0.7766) — outperformed the best single model (Dice ≈ 0.7674) but did not reach statistical significance ($p > 0.05$). This aligns with previous research which pinpointed model diversity as key driver of ensemble benefit. When base models perform similarly and share biases, averaging yields only marginal 1–3% Dice improvements. For example, Henry et al. (2020) noted that increasingly complex architectures yielded diminishing returns on BraTS – after a certain point, "*more complicated training schemes and neural network architectures*" gave no significant improvement [144].

The impact of ensembling on each tumor sub-region was generally small, and no one fusion strategy outperformed the others by a statistically significant margin. However, when compared to the single-model baselines, all ensembles did yield measurable improvements in certain sub-regions:

For the NCR, Dice and boundary error gains were minimal—less than 0.005 in overlap and roughly 0.4 mm in HD95—reflecting the inherently heterogeneous nature of necrosis. In such a variable tissue, no single uncertainty strategy proved dominant, and the ensembles provided only marginal smoothing of false positives and jagged edges.

In ED, the benefits were much clearer. Both the TTD-Only ensemble (HD95 $\approx$ 8.68 mm) and the TTA-Only ensemble ($\approx$ 9.54 mm) significantly outperformed every individual model (for example, Attention U-Net's 12.01 mm, $p < 0.01$), demonstrating that incorporating either epistemic or aleatoric uncertainty at inference time can effectively smooth spurious blobs and irregular boundaries in diffuse ED. Interestingly, the Hybrid (TTA+TTD) fusion saw its ED HD95 rebound to roughly 12 mm—worse than either uncertainty source alone—suggesting that in some cases blending both types of uncertainty can introduce conflicting boundary cues that degrade performance.

Finally, in the ET sub-region, Attention U-Net performed worst on its own (sensitivity $\approx$ 0.695), frequently missing the thin, high-contrast rim. By contrast, every ensemble improved ET sensitivity, with the TTA-Only fusion reaching the highest value ($\approx$ 0.796). This highlights how consensus across multiple models can rescue small, sharply defined structures that individual networks tend to under-segment.

Regarding boundary precision, top BraTS submissions report single-digit HD95 values: for example, strong single-model methods achieved HD95 $\approx$ 5.5 mm for ET and $\approx$ 7.9 mm for TC on validation [142], while the 2022 winning ensemble attained ET HD95 $\approx$ 11–12 mm (due to a few outliers) but an exceptionally low WT HD95 of $\approx$ 3–4 mm [143]. By comparison, the TTA-only ensemble yielded an average HD95 of $\approx$ 7.06 mm, and the TTD-only ensemble further reduced this to $\approx$ 6.83 mm by smoothing spurious boundary fragments. These results demonstrate that lightweight uncertainty-based sampling can rival more expensive multi-model fusion schemes in boundary delineation.

7.3.2 Performance-based weights offer little above uniform fusion

Weighting models by their validation Dice produced virtually the same result as uniform averaging. Because SegResNet, Attention U-Net, and SwinUNETR differed by only 1–2% in single-model Dice, performance-based weights offered no additional discrimination.

7.3.3 Aleatoric vs. epistemic ensembling: Dice vs. boundary precision

- TTA-only ensemble achieved the highest Dice and markedly improved calibration of probabilities, in line with reports that geometric test-time augmentations often outperform MC dropout for uncertainty-guided segmentation, yielding both higher overlap and better uncertainty-error correlation [145].
- TTD-only produced the lowest overall HD95 ($\approx$ 6.83 mm), demonstrating dropout's strength in smoothing extreme boundary outliers by averaging model-weight uncertainty.
- The Hybrid (TTA+TTD) ensemble lies between the two, suggesting that dropout adds complementary epistemic uncertainty but that most performance gains stem from aleatoric TTA diversity. This results mimics the findings of Wang et al. (2019) who observed that generally uncertainty maps produced by the Hybrid (TTA+TTD) approach looks similar to the aleatoric uncertainty maps [146].

7.3.4 Radiomic feature correlations

Ensembling reshapes how radiomic markers relate to segmentation accuracy by tempering each model's unique biases and amplifying their shared cues.

In the NCR sub-region, single models diverged sharply: SegResNet's Dice correlated with GLCM contrast as high as $\rho \approx 0.28$ and first-order variability at $\rho \approx 0.17$–0.22, while Attention U-Net and SwinUNETR showed few significant links. After fusion—especially with test-time augmentations or hybrid weighting—the strongest contrast correlation falls to about $\rho \approx 0.25$, weaker T1/T1CE signals rise to match it, and first-order metrics (e.g. FLAIR SD, GLDM entropy) become consistently significant ($\rho \approx 0.15$–0.21).

For ED, individual architectures leaned on different predictors—T2 minima for SegResNet ($\rho \approx 0.18$), volume for Attention U-Net ($\rho \approx 0.20$), T1CE mean for SwinUNETR ($\rho \approx -0.19$)—but every ensemble converges on T2 minimum ($\rho \approx 0.16$–0.18). Hybrid fusion further strengthens the negative T1/T1CE mean link ($\rho \approx -0.21$) and adds a mild volume effect ($\rho \approx 0.15$), refocusing the feature set onto robust contrast cues.

In the ET sub-region, both single models and ensembles depend almost exclusively on T2 minimum intensity ($\rho \approx 0.19$–0.23), with no other modality contributing significant predictors once this key contrast is captured.

Overall, ensembling trims 30–70% of significant feature–performance correlations, concentrating on the most reliable, sub-region-specific markers. By dampening extreme biases and elevating consensus signals, it yields smoother boundaries and clearer, clinically interpretable mappings between MRI characteristics and model performance.

7.4 Added value of uncertainty estimation

In clinical practice, knowing where the model is likely to be wrong is as important as the segmentation accuracy. Uncertainty estimation brings not only numbers but also actionable insights which keep human in the loop. In this project, three complementary views on uncertainty - probability calibration, the correlation between uncertainty and true error, and risk–coverage behavior - have been employed to assess the reliability of the implemented uncertainty estimation methods.

7.4.1 Probability calibration

A crucial aspect of trustworthy AI in clinical settings is the reliability of the model's confidence in its predictions. Calibration measures how closely a model's stated confidence (softmax probability) matches its empirical accuracy. Temperature scaling was applied to all ensembles (T=4.12), incorporating uncertainty-weighting, and the Expected Calibration Error (ECE) per sub-region was computed (Table 6.6). The Hybrid ensemble achieved the lowest average ECE in two of three regions (ED, ET) and a close second on NCR, indicating it produces the most

reliably calibrated softmax probabilities overall. TTD followed closely, delivering only slightly higher ECE. TTA was the least well calibrated across all sub-regions (highest ECE), often producing a "staircase" pattern in its reliability diagram (Figure 6.25b), with sharp jumps and mid-range plateaus. This aligns with the finding of Wang et al. (2018) [145] who found that TTA does not necessarily yield well-calibrated probabilities without additional post-processing. All performance & uncertainty-weighted ensembles showed wide standard deviations — especially in the small, heterogeneous NCR — revealing that calibration quality varies strongly from case to case.

Overall, the superior calibration of the Hybrid ensemble suggests greater trustworthiness in its probability outputs, which could aid clinicians in interpreting the model's confidence.

7.4.2 Uncertainty vs. error

Additionally, beyond mere confidence scores, a valuable characteristic of uncertainty estimates is their ability to predict where the model is likely to fail. Thus, the correlation between the uncertainties and error (negative log-likelihood) was analyzed. Spearman correlations revealed that TTA's uncertainty maps track true segmentation error most faithfully—especially in necrotic core ($\rho \approx 0.28$) and edema ($\rho \approx 0.27$) — while TTD's epistemic signals only modestly flag necrotic errors ($\rho \approx 0.22$) and even inversely correlate with edema mistakes. This finding aligns with the Wang et al. (2019) analysis which showed that has fewer overconfident incorrect predictions compared to TTD, outperforming TTD at localizing erroneous voxels [146].

7.4.3 Risk coverage

For practical clinical adoption, uncertainty estimates should enable informed decision-making. Risk-coverage analysis explores how effectively the uncertainty measures can guide selective review. When allowing the model to "abstain" on its least confident voxels, TTA produced the steepest, most monotonic risk–coverage trade-off, confirming that its uncertainty map can effectively guide clinicians to review only the voxels most likely to be wrong. The Hybrid ensemble, by blending TTA's aleatoric strengths with TTD's calibration, yields smooth, monotonic RC curves that balance conservatism with reliable risk reduction. TTD's non-monotonic

risk–coverage curves—where dropping voxels sometimes increases error—underscore the practical limitations of relying solely on epistemic signals for abstention, a phenomenon also noted in previous research [146].

7.4.4 Summary

In summary, the Hybrid fusion strategy offers a compelling approach by combining the reliability of well-calibrated probabilities with the practical utility of robust error localization and a predictable risk-coverage trade-off, enhancing its potential for implementation in clinical practice.

7.5 Limitations of the study and directions for future work

7.5.1 Model diversity and ensemble gains

7.5.1.1 Limitations

Despite combining three state-of-the-art architectures, ensemble gains—especially for the necrotic core (NCR)—were modest. When base models share similar strengths, their errors overlap, leaving little room for fusion to improve Dice or boundary metrics.

7.5.1.2 Future directions

Sub-region specialist networks: The limited improvements — especially in NCR Dice and overall accuracy - provided by the ensemble models stem primarily from the high performance of standalone models on BraTS, leaving little room for overlap gains which resulted in low error diversity among three state-of-the-art models. Future studies, rather than training all models to segment every tumor sub-region, could train each network to focus on a single task - e.g., one model dedicated solely to NCR, another to ED, a third to ET. Such expert specialists would develop highly tailored inductive biases (e.g. an NCR expert might use loss functions or augmentations tuned for low-contrast regions), increasing error diversity across the ensemble and potentially boosting both Dice and boundary accuracy when their outputs are fused.

Radiomics-driven weighting: The understanding of how each model performance correlates with the radiomic features of the MRI data can inform the development of more sophisticated, case-adaptive ensemble strategies in future work, potentially weighting models based on the characteristics of the input scan. Rather than fixed or solely uncertainty-based fusion, one could train a small meta-learner on validation data that ingests the scan's radiomic feature vector and predicts optimal weights for each base model. For instance, high T2 entropy (a cue that SegResNet and V-Net handle well) would increase their ensemble weight, whereas a very thin ET sub-region (low T1CE mean) would shift weight toward SwinUNETR's global attention. A similar idea was demonstrated by Chen et al. (2023), who used voxel-level radiomics maps alongside deep features to inform ensemble fusion in glioma segmentation [68].

7.5.2 Uncertainty Estimation Framework

7.5.2.1 Limitations

All uncertainty quantification relied on Bayesian approximations—test-time augmentation (TTA) for aleatoric and test-time dropout (TTD) for epistemic uncertainty—which carries two main drawbacks:

- **Computational overhead:** Both TTA and TTD incur tens of forwards per volume, straining GPU resources and potentially limiting real-time clinical use.
- **Calibration variability & OOD generalization:** Although Hybrid ensembles improved in-distribution calibration, the Expected Calibration Error (ECE) varied widely—especially for NCR—and improvements may not hold on out-of-distribution (OOD) scans [96].
- **Simple averaging of TTA and TTD:** Current Hybrid fusion simply averages aleatoric (TTA) and epistemic (TTD) uncertainty signals, giving each equal weight. In practice, some regions may benefit more from one type of uncertainty than the other.

7.5.2.2 Future directions

Exploring alternative uncertainty estimation methods There is a wide variety of uncertainty estimation methods - e.g., deep ensembles, evidential networks, or the recently proposed spectral-normalized Gaussian processes (SNGP) - which might capture richer uncertainty, especially under distributional shift.

Evaluating uncertainty on external datasets Rigorously evaluate calibration and uncertainty reliability on external datasets—different institutions, scanner vendors, tumor types (e.g. metastases)— to quantify how well in-distribution gains transfer.

Learned uncertainty weighting A more flexible approach than weighting both uncertainty estimation methods equally would be to learn per-voxel or per-region weighting factors — e.g. via a small meta-network — that would tune how much each uncertainty source influences the final fusion. For instance, regions with highly variable augmentations but stable dropout might lean more on TTA, while the opposite would favor TTD. Optimizing these weights could sharpen both overlap (Dice) and boundary precision (HD95) beyond what a fixed Hybrid can achieve.

7.5.3 Data scope and clinical validation

7.5.3.1 Limitations

This work used only the BraTS adult glioma dataset and retrospective testing. It lacks broad tumor diversity such as other tumor types (e.g. meningiomas, metastases), pediatric cases, or multi-institutional protocols. Moreover, no user studies were performed to show whether uncertainty maps actually improve clinician confidence, reduce review time, or influence treatment planning.

7.5.3.2 Future directions

Expand to varied cohorts Apply the ensemble and uncertainty pipeline to scans derived from pediatric patients, metastases, and meningioma cases.

Reader studies & workflow integration Embed uncertainty maps into a clinical workstation prototype and run prospective trials, measuring clinician's time savings and segmentation accuracy when guided by uncertainty, and assess trust and usability via structured feedback.

7.6 Clinical implications

This thesis introduces a system designed to enhance the monitoring of glioma patients. By generating detailed segmentation, probability, and uncertainty maps, this methodology facilitates the identification of regions exhibiting temporal variations in model confidence. This capability is crucial for clinicians in differentiating genuine tumor progression from potential confounding factors such as treatment-related effects or imaging artifacts. The primary objective of this research was to establish a comprehensive framework for researchers and clinicians, enabling efficient preprocessing of MRI datasets, precise tumor segmentation, quantitative volumetric analysis, and informative visualization of clinical trial outcomes.

The integration of quantitative uncertainty assessment within the developed toolbox offers a significant advantage in clinical practice. It allows expert clinicians to prioritize the review of regions characterized by high model uncertainty, thereby focusing diagnostic attention on the most ambiguous areas. Furthermore, these uncertainty metrics can be translated into standardized, readily interpretable reports (e.g., "The enhancing margin demonstrated a segmentation confidence exceeding 90% across 95% of its extent"). Such reports enhance interdisciplinary communication, aiding in the evaluation of treatment risks, the delineation of expected resection boundaries, and the facilitation of transparent discussions regarding prognosis and follow-up strategies with patients and their families.

Beyond immediate clinical applications, this work also represents a substantive advancement towards the development of transparent and accountable artificial intelligence models within healthcare. By explicitly quantifying and presenting model limitations through uncertainty estimations, and by maintaining the critical role of human oversight, this system aligns with the evolving requirements of regulatory bodies, such as the EU AI Act, which increasingly mandate the provision of transparent uncertainty information for high-risk medical AI applications.

Chapter 8

Sustainability Analysis and Ethical Implications

8.1 Introduction

Beyond evaluating the technical performance of the project, it is also crucial to assess its broader impacts, which aimed to develop an uncertainty-aware brain tumor segmentation tool on the environment, economy, and society. Machine learning research, and often technology in general, is commonly seen as value-neutral, and scientists put more emphasis on its potential positive applications, omitting the environmental, economic, and societal dimensions of their work [147]. The focus on novelty and performance overshadows a deeper analysis of the societal impact of research, and reduces the impact statements merely to an administrative formality.

Many analyses show that AI research often downplays (or omits entirely) the economic, social, and environmental implications of its work. A meta-analysis of highly-cited NeurIPS/ICML papers between 2008 and 2019 revealed that out of the 100 analyzed papers, none of them mentions user rights and ethical principles and 98% do not mention any potential negative impacts of their work [147]. At NeurIPS 2021, only 32.9% of submitted papers contained any "societal impact" or "broader impact" statement, while the remainder omitted it entirely [148].

To move towards a more responsible AI research paradigm, a fundamental shift in perspective is necessary. The notion of value-neutrality must be challenged, and researchers need to

recognize that their work is inherently embedded within a complex web of social, economic, and environmental factors. Thus, recognizing this critical gap in the AI research, this chapter aims to address these often-neglected dimensions specifically within the context of the uncertainty-aware brain tumor segmentation tool developed in this project. The aim is to provide a holistic evaluation of the work and foster a more responsible approach to AI in medical imaging.

8.2 Sustainability matrix

8.2.1 Environmental perspective

Training four large 3D deep learning models (V-Net, SwinUNETR, SegResNet, and Attention UNet) on a high-performance cluster incurs non-negligible energy costs. The project used two NVIDIA A100 GPUs (each ∼400W TDP) in parallel to perform hyperparameter tuning using cross-validation. The entire 5-fold cross-validation took ≈ 72h for each model. The electricity consumption would total on the order of tens of kWh, translating to several kilograms of CO_2 emissions. Despite being small compared to large language models such as GPT-4, whose training consumed 1,750 MWh, which is equivalent to the annual consumption of approximately 160 average American homes [149], it still has a direct environmental footprint of the research. Some impact was mitigated by using PyTorch's Automatic Mixed Precision, which accelerates training on modern GPUs [150], thereby cutting runtime and energy use.

The ensemble inference is also resource-intensive. Running the segmentation pipeline in the Streamlit prototype takes about 5-7 minutes (depending on the size of the MRI scans) on a laptop GPU (NVIDIA RTX 4060) per patient. Translating this to the realities of a busy hospital, it could become significant. Local deployment of the application (preferred for privacy reasons) means each hospital would run its own hardware for inference instead of a shared cloud server. This avoids the energy overhead of data transfer and large cloud data centers, but it also means duplicated hardware at each site. Avoiding the cloud also means no additional data center footprint. Edge computing can sometimes be up to 30% more energy-efficient than cloud computing by eliminating constant data transfer and idle server time [151].

One could also argue that intangible outputs of this project (models, code) support a circular

economy. The trained models and codebase could be reused or fine-tuned for other segmentation tasks, maximizing the value extracted from the initial resource investment.

Moving forward, there are several ways to reduce the environmental impact of this project. Firstly, the knowledge of the ensemble could be distilled into a single compact model to reduce inference costs. This way, the ensemble performance could be retained, while avoiding triple compute each run. A distilled model would require less memory and could even run on lower-power hardware, reducing energy per prediction.

Secondly, scheduling the training and inference runtimes during periods of low grid carbon intensity can also reduce the carbon footprint. Additionally, using infrastructure that is powered by renewable energy, when possible, can further minimize the project's environmental impact.

In summary, while this thesis shows some awareness of efficiency (e.g., by using mixed-precision), however, it will need further optimizations like knowledge distillation to minimize its environmental impact, especially as it scales.

8.2.2 Economic perspective

In terms of the development costs, this project benefited from the computational resources of the Technical University of Catalonia, but it still generated costs. If the training was conducted on the cloud, like Azure, utilizing two A100 GPUs could cost on the order of tens of dollars per hour. Using open software also reduced development costs, as did leveraging the publicly available BraTS 2021 dataset.

For deployment in a hospital or a research facility, hardware acquisition poses a significant cost. A suitable local server or high-end PC with a CUDA-capable GPU is required. NVIDIA RTX 40 GPU series costs are in the order of $300-$2000 [152]. Local deployment is a capital expense - hospitals would have to buy and maintain the equipment themselves. This can be cost-effective if utilization is high, but if the tool is used infrequently, the hardware investment might sit idle. On the other hand, avoiding cloud means no ongoing subscription or API fees, which can be quite expensive for medical AI software.

Furthermore, the choice of an ensemble over a single model also touches upon economic effi-

ciency. The marginal accuracy gain from the ensemble was modest, while an ensemble utilizing TTA and TTD raises computational costs significantly compared to a single model. This raises a cost-benefit question: is the small performance boost and uncertainty quantification worth the extra complexity and the costs coming with it? Uncertainty quantification is important and might improve the adoption of such a tool. Here, knowledge distillation could be also the solution - one could train the big ensemble for maximum performance and then compress it into a single model for deployment, delivering better performance per dollar during inference.

If this prototype were to become a broadly adopted clinical tool, its software engineering and regulatory expenses must be included. Converting the Streamlit proof-of-concept into a production-grade application demands substantial developer effort, testing, and maintenance. In addition, conformity with the EU Medical Device Regulation (MDR) — including CE-marking — adds both time and cost.

Under MDR Rule 11, any software that "provides information used to make decisions for diagnosis or therapeutic purposes" is automatically at least Class IIa. Its exact wording is:

> *Software intended to provide information which is used to take decisions with diagnosis or therapeutic purposes is classified as class IIa, except if such decisions have an impact that may cause:*
>
> - *death or an irreversible deterioration of a person's state of health, in which case it is in class III;*
> - *or — a serious deterioration of a person's state of health or a surgical intervention, in which case it is classified as class IIb [153].*

While errors in brain tumor segmentation could have severe consequences, they typically do not have the same immediate life-threatening potential as a malfunctioning pacemaker. Moreover, the intention is always for a qualified clinician to review and approve the AI's output.

CE-marking a Class IIa device typically incurs on the order of €200 000–€600 000 in initial certification costs—covering Notified-Body fees, technical documentation, clinical evaluation, software verification, and consulting—followed by €50 000–€100 000 per year in surveillance and maintenance fees [154].

On the positive side, the project could results in economic gains for hospitals and research facilities from faster and more accurate tumor delineations. Neuroradiologists would spend less time manually contouring tumors (a labor-intensive task), reallocating time to other duties and increasing overall productivity. There is also reuse potential in the developed models and pipeline: the models trained on BraTS could be fine-tuned with relatively low effort to new tumor segmentation datasets, rather than training from scratch each time, saving future development costs.

In conclusion, while this work utilized university compute and open-source assets to create the prototype, the transition to a clinically deployed system introduces substantial additional expenditures. Hospitals or research facilities would have to weigh the one-time costs of specialized hardware and regulatory certification against the ongoing savings in clinician time and the value of enhanced diagnostic consistency. The modest performance gains of an uncertainty-aware ensemble must be balanced with its computational burden, a trade-off that techniques such as knowledge distillation may help resolve.

Ultimately, a detailed financial model - incorporating hardware depreciation, maintenance, regulatory fees, and projected efficiency improvements - would be crucial to determine whether the long-term clinical and research benefits justify the up-front and recurring investments required to integrate this brain tumor segmentation tool into clinical practice.

8.2.3 Social perspective

The biggest social benefit of this project would lie in reducing clinician workload. Segmenting a brain tumor by hand can take substantial time for a radiologist. An AI toolbox that preprocesses MRI scans and pre-segments the images could free up time for the clinician to focus on more complex interpretative tasks or on patient interaction. This can in turn result in higher job satisfaction across clinicians as they would be able to spend more time on what requires their expertise and less on routine labour.

The inclusion of uncertainty quantification is especially noteworthy in a social context, as it allows for risk-aware decision making. For example, if the model flags certain tumor regions

as high-uncertainty, a clinician knows they should examine those areas more closely or order additional tests instead of relying blindly on the AI. This could prevent harm – knowing where the model is likely wrong is as important as its accuracy. Such transparency can lead to increased trust in a tool and keeps human in the loop - AI is treated as a decision support tool, not an autonomous decision-maker. Nevertheless, there is still a potential negative: if the clinicians become too reliant on a tool they might, over time, lose some sharpness in their own segmentation skills.

A key point in the analysis of social implications is the consideration of *who might be excluded or harmed* by the model. The model was trained on the BraTS 2021 Adult glioma dataset which means its knowledge is rooted in adult brain tumor characteristics. Pediatric brain tumors often have very different characteristics and challenges compared to the adult brains [155]. Thus, if the tool were applied to a child's MRI without caution, it might perform poorly, potentially missing tumors or mis-segmenting due to those differences. This points to a need for further training on diverse data or a clear usage disclaimer that it's for adult data only.

Furthermore, there is an inherent bias in the BraTS dataset itself. BraTS data comes from 15 institutions, but is not globally representative of all demographic groups or MRI machine types – if the model encounters an image from a substantially different distribution (e.g., different MRI vendor or demographic group), its accuracy would suffer. Tackling this requires both technical improvements (continuous learning, more diverse training data) and social awareness (clinicians recognizing when the tool might be inappropriate to use).

An issue of accountability is also important here. If a segmentation error leads to harm, there must be clarity on whether it was a user mistake or an AI failure and how to prevent it in the future. The uncertainty feature contributes significantly to this by documenting instances of the AI's uncertainty, signaling situations where human expertise should have been the primary driver. Maintaining meaningful human control and fostering genuine understanding prevents over-reliance and ensures the technology remains a tool that enhances human capabilities while respecting dignity and agency.

8.3 Ethical implications

This research addressed the issue of trust and transparency in AI-powered applications in medical settings. Normally, clinicians do not know to what extent they can trust an automated segmentation, as deep learning models do not provide uncertainty estimates to their own predictions. This thesis addresses this issue by incorporating TTA and TTD in ensemble models to create voxel-level uncertainty maps, giving insights into tumor regions the model is less certain about. Consequently, researchers obtain a robust, reproducible image analysis, while clinicians are provided with the decision support that they can trust.

Addressing accountability is paramount when integrating AI into clinical workflows. To counter the potential for automation bias, where users might over-rely on AI, clear documentation and comprehensive user training are crucial. These should equip users to understand the tool's limitations and interpret its output appropriately, including identifying situations requiring expert review, such as high uncertainty regions. Ultimately, maintaining human oversight in final decisions is a necessary safeguard, ensuring the human element remains central.

Furthermore, handling medical imaging data entails strict privacy obligations. A local deployment of the application is important to ensure that patient scans can be processed on-site (e.g., in a hospital) without needing to upload data to an external server. Furthermore, the system must be compliant with data protection regulations (such as GDPR) and institutional ethics guidelines when using real patient MRI data. This includes anonymizing the data (like in BraTS dataset) and obtaining patient consent to use an automatic segmentation tool on their data. Following the UPC's ethical principles regarding respect for individuals, it is essential to protect personal health information. For this reason, ethically, it is crucial to have robust security and privacy controls in place to prevent the misuse of sensitive patient data.

Another issue is potential misuse of the application. Although developed for beneficial purposes in research and future clinical practice, the tool could be applied in contexts beyond its original intent. For instance, it could be the case that someone without medical expertise uses the tool on personal MRI data and makes health decisions without consulting a doctor, or integrates the algorithm into a clinical workflow without regulatory approval. Additionally,

the segmentation tool might be used to draw conclusions that it was not validated for (e.g., estimating tumor grade or patient prognosis). Thus, it is important to communicate to the end users the purpose of the application and possibly release a license limiting clinical use.

Overall, this thesis project strived to align with the ethical principles outlined by the UPC. The UPC Code of Ethics highlights values such as honesty, integrity, respect for individuals, transparency, equity, and sustainability [156].

The work respects the dignity and rights of patients by using anonymized data and focusing on improving patient care outcomes. There is an implicit respect for the autonomy of clinicians and patients – the tool is not meant to override human decision-making, but to support it with additional information. Moreover, in line with "respect for individuals" in the UPC code, the inclusion of uncertainty estimates can be seen as respecting the end-user's right to know the limits of the AI's knowledge.

Furthermore, this thesis closely follows the principles of integrity and honesty. The tool's capabilities are not overstated and the project is clearly presented as a prototype for research use, not a finalized clinical solution. The detailed documentation of the methods as well as the prototype's availability for local testing contribute to the reproducibility of the research, an important aspect of scientific integrity.

This thesis also advances United Nations Sustainable Development Goal 3 ("Good Health and Well-Being") by enhancing the accuracy, consistency, and transparency of brain-tumor segmentation—thereby contributing to target 3.4 ("reduce premature mortality from non-communicable diseases through early diagnosis and treatment") and target 3.b ("support research and development of vaccines and medicines for the communicable and non-communicable diseases that primarily affect developing countries"). Additionally, by providing open-source code and uncertainty-aware methods, the project also fosters capacity-building in low-resource research settings, aligning with Goal 17 ("Partnerships for the Goals") through knowledge sharing and collaborative science. In doing so, it supports equitable access to advanced AI tools, promoting sustainability in both health outcomes and research infrastructure.

In conclusion, the ethical analysis confirms that this uncertainty-aware brain tumor segmen-

tation tool was developed with a strong ethical awareness. By addressing a significant need and integrating safeguards, it responsibly navigates potential risks like misuse and bias. Its alignment with the UPC's ethical code and research integrity principles underscores its role as an example of accountable AI innovation, promising benefits for research and future clinical applications.

Chapter 9

Conclusions

9.1 The overall aim of the thesis

The thesis set out to determine whether an uncertainty-aware ensemble of state-of-the-art 3-D neural networks can provide brain-tumour segmentations that are both quantitatively competitive and accompanied by voxel-level confidence information that clinicians can interpret. A complementar aim was to deliver a lightweight Streamlit prototype that demonstrates the full workflow - from scan upload to uncertainty visualisation - and could form a basis for the future iterations of clinical toolboxes incorporating uncertainty estimation.

9.2 How the sub-objectives were met

To achieve these aims, the thesis pursued the following sub-objectives:

Sub-objective 1: Investigate the performance of state-of-the-art models across tumour sub-regions This objective was met by benchmarking four architectures (SwinUNETR, SegResNet, Attention UNet, and V-Net) on necrotic core (NCR), edema (ED), and enhancing tumor (ET) in Section 6.2.1.1. V-Net performed significantly worse than the other three models and was excluded from subsequent analysis. Among the remaining models, performance differences were minor and not statistically significant, justifying their ensemble.

Sub-objective 2: Develop an uncertainty-aware ensemble-based segmentation model This objective was addressed by proposing and evaluating three fusion strategies described in Chapter 5, incorporating test-time dropout (TTD), test-time augmentation (TTA), or a hybrid of both for uncertainty estimation. The TTA-Only ensemble achieved the highest overall Dice, while the TTD-Only variant produced the sharpest edema boundaries. However, none of the ensembles significantly outperformed the best individual model in Dice, suggesting limited model diversity. Notably, uncertainty estimates proved informative, with TTA yielding strong risk-coverage curves and uncertainty correlating with prediction error.

Sub-objective 3: Create a prototype of the clinical toolbox A Streamlit application was developed, as described in Section 5.13, to execute preprocessing, ensemble inference, and visualization (segmentation, probabilities, uncertainty) within approximately 5-7 minutes on a laptop GPU. This prototype demonstrates the technical feasibility of a clinician-facing tool, supporting DICOM/NIfTI input and providing output of predicted masks and uncertainty volumes.

9.3 Looking Forward

While this thesis demonstrates the potential of uncertainty-aware ensembles for brain tumor segmentation, several avenues for future research could further enhance its clinical utility. Primarily, increasing model diversity within the ensemble, perhaps by incorporating sub-region specialists or foundation models, could yield more substantial and statistically significant improvements. Streamlining uncertainty estimation techniques to reduce computational cost, such as exploring lightweight evidential or spectral-normalized approaches, is another important direction. Ultimately, prospective clinical validation and conducting multi-reader studies, will be essential to fully assess the impact of uncertainty information on clinical decision-making. Finally, while the prototype incorporates elements that align with the EU AI Act's human-oversight principle, a formal conformity-assessment study is a necessary step towards real-world deployment.

9.4 Final remarks

This work demonstrates that combining complementary segmentation networks and weighting their predictions with explicit voxel-level uncertainty can enhance the interpretability – and occasionally the boundary accuracy – of automated glioma segmentation. The accompanying open-source toolbox translates these research findings into an end-to-end pipeline that clinicians can explore, laying a concrete foundation for future, larger-scale studies aimed at integrating uncertainty-aware AI into routine neuro-oncological practice.

Appendix

A.1 Slurm job script for hyperparameter tuning

Listing 1: Example Slurm job script used for hyperparameter tuning of SegResNet.

```
#!/bin/bash
#SBATCH -J brain_tumor_segmentation_segresnet # Job name
#SBATCH -o output_segresnet%j.log # Standard output log
#SBATCH -e error_segresnet%j.log # Standard error log
#SBATCH -c 2 # CPU cores per job
#SBATCH --mem=32000 # RAM per job (32GB)
#SBATCH --gpus=1 # Number of GPUs allocated (1)
#SBATCH -p gpu # Specify GPU partition

# Select the GPU
export CUDA_VISIBLE_DEVICES=1
# Activate the Python environment
source ~/.thesis_env/bin/activate
# Navigate to the directory with the script
cd ~/src/train/
# Execute the script
python hyperparameter_tuning.py --model_name "segresnet"
# Deactivate the Python environment
deactivate
```

A.2 Hyperparameter tuning results

A.2.1 V-Net

Table 1: V-Net cross-validation results for four hyperparameter configurations.

Metric	Config 1	Config 2	Config 3	Config 4
Dice Scores				
Dice NCR	0.7191 ± 0.0085	0.6395 ± 0.1611	0.6341 ± 0.0515	0.6413 ± 0.0227
Dice ED	0.5976 ± 0.2997	0.3251 ± 0.3590	0.4009 ± 0.3045	0.3253 ± 0.3112
Dice ET	0.8047 ± 0.0135	0.6447 ± 0.3189	0.7272 ± 0.0412	0.7348 ± 0.0415
Dice Overall	0.4853 ± 0.1930	0.3903 ± 0.3284	0.3915 ± 0.2262	0.4052 ± 0.2524
HD95 (mm)				
HD95 NCR	32.8416 ± 48.4487	57.5419 ± 28.8051	31.8601 ± 19.6046	22.7782 ± 14.9702
HD95 ED	22.0405 ± 19.1293	17.7562 ± 14.4112	27.1971 ± 8.7715	18.2340 ± 7.4368
HD95 ET	22.9703 ± 7.6755	58.1370 ± 54.4720	27.8920 ± 16.0916	19.8119 ± 4.2749
HD95 Overall	68.8090 ± 34.9144	80.5465 ± 38.7520	54.2084 ± 20.2886	47.4053 ± 13.2021
Sensitivity				
Sensitivity NCR	0.7136 ± 0.0296	0.6848 ± 0.0321	0.6164 ± 0.0589	0.6345 ± 0.0489
Sensitivity ED	0.6191 ± 0.3183	0.3185 ± 0.3482	0.3809 ± 0.3026	0.2903 ± 0.2977
Sensitivity ET	0.8444 ± 0.0241	0.8614 ± 0.0181	0.7076 ± 0.0406	0.7413 ± 0.0886
Sensitivity Overall	0.6700 ± 0.2068	0.5792 ± 0.3033	0.4800 ± 0.2262	0.4949 ± 0.2646
Specificity				
Specificity NCR	0.9998 ± 0.0001	0.9989 ± 0.0019	0.9998 ± 0.0001	0.9998 ± 0.0001
Specificity ED	0.9989 ± 0.0008	0.9994 ± 0.0005	0.9993 ± 0.0006	0.9996 ± 0.0004
Specificity ET	0.9996 ± 0.0001	0.8667 ± 0.2655	0.9996 ± 0.0002	0.9996 ± 0.0002
Specificity Overall	0.9175 ± 0.0006	0.9096 ± 0.1655	0.9799 ± 0.0004	0.9968 ± 0.0003

A.2.2 SegResNet

Table 2: SegResNet cross-validation results for four hyperparameter configurations.

Metric	Config 1	Config 2	Config 3	Config 4
Dice Scores				
Dice NCR	0.7468 ± 0.0077	0.7478 ± 0.0047	0.7042 ± 0.0099	0.6777 ± 0.0286
Dice ED	0.7699 ± 0.0123	0.7828 ± 0.0143	0.7280 ± 0.0155	0.7149 ± 0.0379
Dice ET	0.8410 ± 0.0073	0.8355 ± 0.0145	0.8064 ± 0.0079	0.7847 ± 0.0248
Dice Overall	0.6864 ± 0.0411	0.6685 ± 0.0380	0.5967 ± 0.0452	0.5821 ± 0.0541
HD95 (mm)				
HD95 NCR	81.6802 ± 13.5496	19.8825 ± 13.8670	16.4750 ± 23.8776	62.5112 ± 4.5718
HD95 ED	25.6414 ± 6.7586	17.4683 ± 3.3954	17.5292 ± 2.2409	25.3683 ± 8.9381
HD95 ET	12.9155 ± 30.9467	13.5375 ± 4.6398	12.5232 ± 5.4083	16.7710 ± 6.4104
HD95 Overall	43.9034 ± 21.1887	37.8846 ± 11.0973	42.7405 ± 16.8410	38.5737 ± 6.9394
Sensitivity				
Sensitivity NCR	0.7214 ± 0.0211	0.7347 ± 0.0175	0.6863 ± 0.0215	0.6505 ± 0.0496
Sensitivity ED	0.7805 ± 0.0343	0.7742 ± 0.0436	0.6792 ± 0.0211	0.7215 ± 0.0876
Sensitivity ET	0.8357 ± 0.0200	0.8387 ± 0.0224	0.7996 ± 0.0238	0.7501 ± 0.0518
Sensitivity Overall	0.7265 ± 0.0534	0.7128 ± 0.0524	0.6207 ± 0.0604	0.6109 ± 0.0776
Specificity				
Specificity NCR	0.9999 ± 0.0000	0.9999 ± 0.0001	0.9999 ± 0.0000	0.9999 ± 0.0000
Specificity ED	0.9990 ± 0.0003	0.9992 ± 0.0003	0.9994 ± 0.0001	0.9989 ± 0.0007
Specificity ET	0.9998 ± 0.0001	0.9997 ± 0.0001	0.9997 ± 0.0001	0.9998 ± 0.0001
Specificity Overall	0.9988 ± 0.0004	0.9986 ± 0.0003	0.9988 ± 0.0002	0.9990 ± 0.0006

A.2.3 Attention UNet

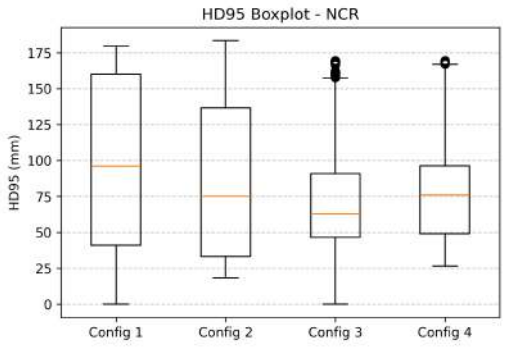
Table 3: Attention UNet cross-validation results for four hyperparameter configurations.

Metric	Config 1	Config 2	Config 3	Config 4
Dice Scores				
Dice NCR	0.7369 ± 0.0175	0.7333 ± 0.0178	0.6281 ± 0.0235	0.5982 ± 0.0211
Dice ED	0.7777 ± 0.0187	0.7721 ± 0.0174	0.7202 ± 0.0064	0.7013 ± 0.0344
Dice ET	0.8295 ± 0.0087	0.8330 ± 0.0091	0.7838 ± 0.0095	0.7851 ± 0.0076
Dice Overall	**0.6875 ± 0.0410**	**0.7009 ± 0.0438**	**0.5417 ± 0.0657**	**0.5484 ± 0.0801**
HD95 (mm)				
HD95 NCR	28.0799 ± 9.7868	55.9355 ± 40.8262	24.2125 ± 6.6369	19.3021 ± 21.9625
HD95 ED	15.9326 ± 5.3329	14.3248 ± 1.7203	28.2644 ± 5.0522	28.9433 ± 5.7738
HD95 ET	11.7476 ± 3.9359	10.8893 ± 1.2395	23.4984 ± 4.4669	22.4647 ± 3.2019
HD95 Overall	**46.7259 ± 9.1335**	**50.7672 ± 31.1677**	**54.4128 ± 5.8030**	**49.0182 ± 14.1030**
Sensitivity				
Sensitivity NCR	0.7181 ± 0.0250	0.6985 ± 0.0239	0.6863 ± 0.0215	0.6682 ± 0.0590
Sensitivity ED	0.7691 ± 0.0478	0.7361 ± 0.0376	0.7211 ± 0.0356	0.7061 ± 0.0591
Sensitivity ET	0.8120 ± 0.0285	0.8021 ± 0.0205	0.8196 ± 0.0228	0.8361 ± 0.0130
Sensitivity Overall	**0.7474 ± 0.0521**	**0.7364 ± 0.0513**	**0.6245 ± 0.0627**	**0.6206 ± 0.0869**
Specificity				
Specificity NCR	0.9999 ± 0.0000	0.9999 ± 0.0000	0.9996 ± 0.0001	0.9996 ± 0.0001
Specificity ED	0.9992 ± 0.0003	0.9994 ± 0.0002	0.9990 ± 0.0003	0.9989 ± 0.0003
Specificity ET	0.9998 ± 0.0001	0.9998 ± 0.0001	0.9996 ± 0.0001	0.9995 ± 0.0001
Specificity Overall	**0.9970 ± 0.0004**	**0.9982 ± 0.0002**	**0.9973 ± 0.0003**	**0.9981 ± 0.0004**

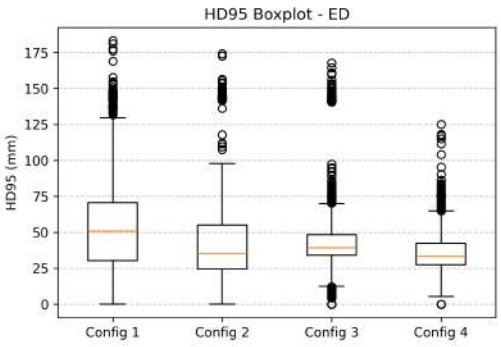
A.2.4 SwinUNETR

Table 4: SwinUNETR cross-validation results for four hyperparameter configurations.

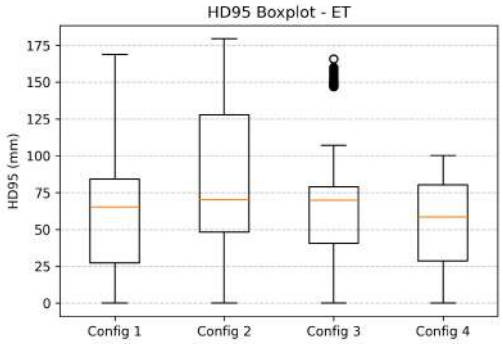
Metric	Config 1	Config 2	Config 3	Config 4
Dice Scores				
Dice NCR	0.7380 ± 0.0083	0.7486 ± 0.0131	0.6024 ± 0.0189	0.5970 ± 0.0220
Dice ED	0.7921 ± 0.0143	0.7984 ± 0.0086	0.6720 ± 0.0202	0.6737 ± 0.0202
Dice ET	0.8469 ± 0.0079	0.8418 ± 0.0054	0.7808 ± 0.0059	0.7766 ± 0.0087
Dice Overall	0.7358 ± 0.0457	0.7382 ± 0.0392	0.5981 ± 0.0752	0.6004 ± 0.0757
HD95 (mm)				
HD95 NCR	14.9706 ± 25.7039	16.2190 ± 39.0418	22.6297 ± 5.5712	23.4653 ± 4.1021
HD95 ED	13.9314 ± 2.7663	15.0363 ± 4.6691	29.7044 ± 8.0778	29.8281 ± 6.9895
HD95 ET	10.1485 ± 3.6440	9.9956 ± 2.5351	22.4407 ± 5.3503	21.5397 ± 3.5179
HD95 Overall	38.4182 ± 27.2265	31.3730 ± 23.0622	41.8183 ± 7.1029	41.4929 ± 6.1860
Sensitivity				
Sensitivity NCR	0.7070 ± 0.0220	0.7342 ± 0.0285	0.6154 ± 0.0472	0.6244 ± 0.0841
Sensitivity ED	0.7709 ± 0.0222	0.8004 ± 0.0264	0.6416 ± 0.0544	0.6375 ± 0.0332
Sensitivity ET	0.8391 ± 0.0209	0.8460 ± 0.0201	0.7734 ± 0.0187	0.7659 ± 0.0296
Sensitivity Overall	0.7509 ± 0.0582	0.7529 ± 0.0524	0.6566 ± 0.0814	0.6496 ± 0.0842
Specificity				
Specificity NCR	0.9999 ± 0.0000	0.9999 ± 0.0000	0.9998 ± 0.0001	0.9998 ± 0.0001
Specificity ED	0.9993 ± 0.0001	0.9992 ± 0.0002	0.9990 ± 0.0003	0.9991 ± 0.0002
Specificity ET	0.9998 ± 0.0001	0.9997 ± 0.0001	0.9997 ± 0.0001	0.9997 ± 0.0001
Specificity Overall	0.9992 ± 0.0003	0.9993 ± 0.0003	0.9984 ± 0.0004	0.9986 ± 0.0003



(a) HD95 Boxplot for NCR

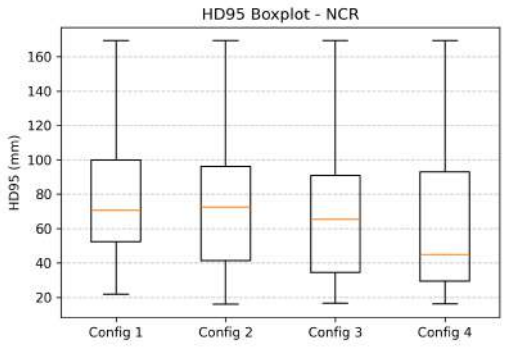


(b) HD95 Boxplot for ED

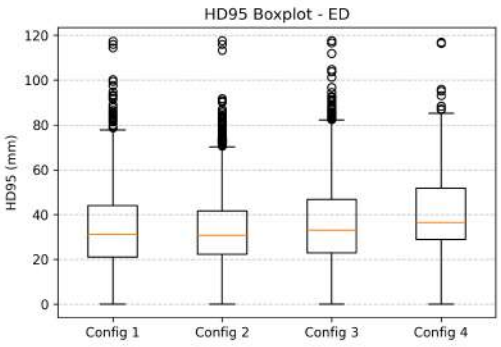


(c) HD95 Boxplot for ET

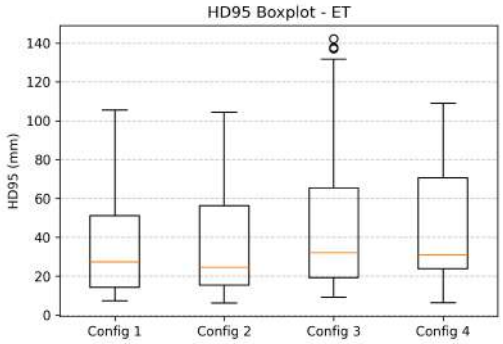
Figure 1: Boxplots showing the distribution of HD95 values for the NCR, ED, and ET tumor subregions across four V-Net hyperparameter configurations. Lower HD95 values indicate better boundary delineation and reduced boundary error.



(a) HD95 Boxplot for NCR

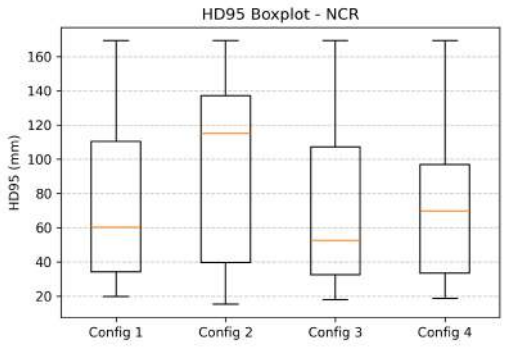


(b) HD95 Boxplot for ED

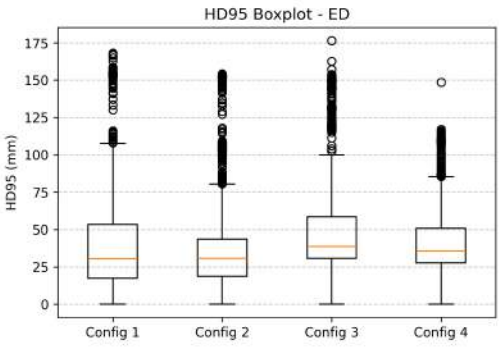


(c) HD95 Boxplot for ET

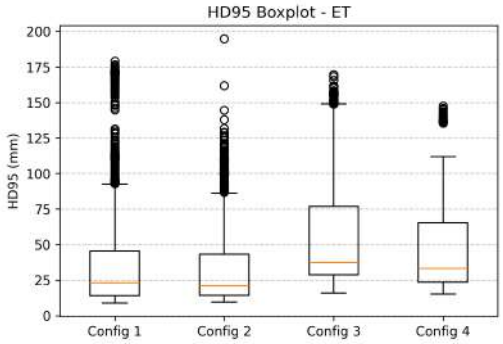
Figure 2: Comparison of HD95 boxplots for NCR, ED, and ET across different SegResNet configurations.



(a) HD95 Boxplot for NCR

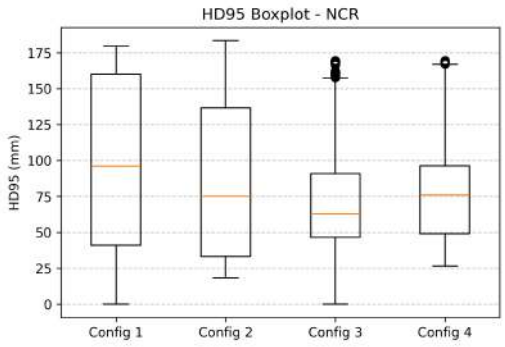


(b) HD95 Boxplot for ED

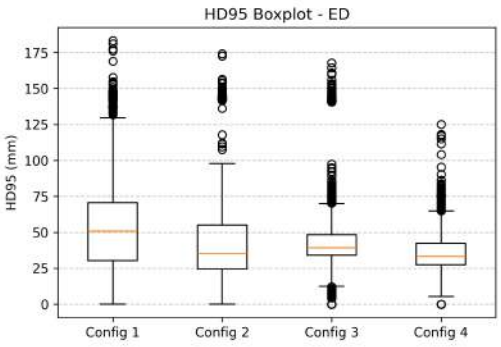


(c) HD95 Boxplot for ET

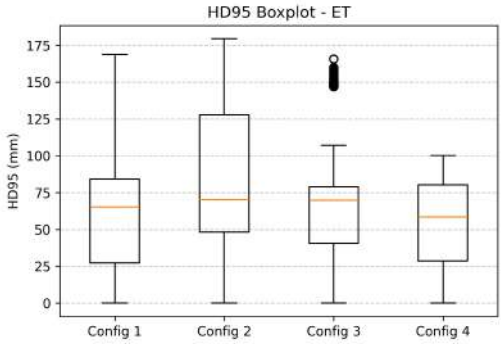
Figure 3: Comparison of HD95 boxplots for NCR, ED, and ET across different Attention UNet configurations.



(a) HD95 Boxplot for NCR



(b) HD95 Boxplot for ED



(c) HD95 Boxplot for ET

Figure 4: Comparison of HD95 boxplots for NCR, ED, and ET across different SwinUNETR configurations.

A.2.5 Significant differences between single models after pos-hoc Mann-Whitney pairwise test

Table 5: Pairwise post-hoc Mann–Whitney U–test results for all significant model differences across metrics.

Metric	**Comparison**	U	p
Dice NCR	V-Net < SegResNet	14242	0.0068
Dice NCR	V-Net < AttUNet	13601.5	0.0007
Dice NCR	V-Net < SwinUNETR	13944.5	0.0024
Dice ED	V-Net < SegResNet	13495	0.0004

Continued on next page

Table 5 – *continued from previous page*

Metric	**Comparison**	U	p
Dice ED	V-Net < AttUNet	14124	0.0046
Dice ED	V-Net < SwinUNETR	14646	0.0245
Dice ET	V-Net < SegResNet	7561	5.0e-21
Dice ET	V-Net < AttUNet	8975	9.3e-16
Dice ET	V-Net < SwinUNETR	7838	6.2e-20
Dice overall	V-Net < SegResNet	10544	8.1e-11
Dice overall	V-Net < AttUNet	11087	2.5e-09
Dice overall	V-Net < SwinUNETR	11223	5.6e-09
HD95 NCR	V-Net > SegResNet	26896	1.1e-17
HD95 NCR	V-Net > AttUNet	26967.5	6.1e-18
HD95 NCR	V-Net > SwinUNETR	26406.5	6.3e-16
HD95 ED	V-Net > SegResNet	24602.5	2.9e-10
HD95 ED	V-Net > AttUNet	23959	1.5e-08
HD95 ED	V-Net > SwinUNETR	23594	1.1e-07
HD95 ET	V-Net > SegResNet	31496	1.1e-38
HD95 ET	V-Net > AttUNet	31270	2.0e-37
HD95 ET	V-Net > SwinUNETR	31022.5	3.9e-36
HD95 overall	V-Net > SegResNet	30804.5	7.2e-35
HD95 overall	V-Net > AttUNet	30592.5	8.8e-34
HD95 overall	V-Net > SwinUNETR	30257	4.3e-32

Continued on next page

Table 5 – *continued from previous page*

Metric	**Comparison**	U	p
Sensitivity NCR	V-Net < AttUNet	14885	0.0490
Sensitivity NCR	SegResNet < AttUNet	14611	0.0220
Sensitivity NCR	AttUNet > SwinUNETR	20801	0.0179
Sensitivity ED	V-Net < SegResNet	14780	0.0364
Sensitivity ET	V-Net < SegResNet	12742	1.7e-05
Sensitivity ET	SegResNet > AttUNet	24083	7.1e-09
Sensitivity ET	SegResNet > SwinUNETR	20632.5	0.0298
Sensitivity ET	AttUNet < SwinUNETR	14233.5	0.0066
Sensitivity overall	V-Net < SegResNet	14231	0.0066
Specificity NCR	V-Net < SegResNet	9312.5	1.3e-14
Specificity NCR	V-Net < AttUNet	12303.5	2.1e-06
Specificity NCR	V-Net < SwinUNETR	6755.5	2.3e-24
Specificity NCR	SegResNet < SwinUNETR	13868.5	0.0018
Specificity NCR	AttUNet < SwinUNETR	12241	1.5e-06
Specificity ED	V-Net < SegResNet	13336.5	0.0002
Specificity ET	V-Net < SegResNet	7418.5	1.3e-21
Specificity ET	V-Net < AttUNet	3193	3.5e-42
Specificity ET	V-Net < SwinUNETR	4428	1.9e-35
Specificity ET	SegResNet < AttUNet	8308	3.8e-18
Specificity ET	SegResNet < SwinUNETR	11954.5	3.5e-07

Continued on next page

Table 5 – *continued from previous page*

Metric	**Comparison**	U	p
Specificity ET	AttUNet > SwinUNETR	22836.5	5.7e-06
Specificity overall	V-Net < SegResNet	14435	0.0128

Bibliography

[1] National Brain Tumor Society. Brain Tumor Facts - National Brain Tumor Society, 2 2024.

[2] Quinn T Ostrom, Mackenzie Price, Corey Neff, Gino Cioffi, Kristin A Waite, Carol Kruchko, and Jill S Barnholtz-Sloan. CBTRUS Statistical Report: Primary brain and other central nervous system tumors diagnosed in the United States in 2015–2019. *Neuro-Oncology*, 24(Supplement_5):v1–v95, 10 2022.

[3] Faizan Ullah, Muhammad Nadeem, Mohammad Abrar, Muna Al-Razgan, Taha Alfakih, Farhan Amin, and Abdu Salam. Brain Tumor Segmentation from MRI Images Using Handcrafted Convolutional Neural Network. *Diagnostics*, 13(16):2650, 8 2023.

[4] Tirivangani Magadza and Serestina Viriri. Deep Learning for Brain Tumor Segmentation: A Survey of State-of-the-Art. *Journal of Imaging*, 7(2):19, 1 2021.

[5] Chandrakant M. Umarani, S.G. Gollagi, Shridhar Allagi, Kuldeep Sambrekar, and Sanjay B. Ankali. Advancements in deep learning techniques for brain tumor segmentation: A survey. *Informatics in Medicine Unlocked*, 50:101576, 1 2024.

[6] Muhammad Ansab Butt and Absaar Ul Jabbar. Hybrid Multihead Attentive UNET-3D for brain tumor segmentation. *arXiv (Cornell University)*, 5 2024.

[7] Rui Hua, Quan Huo, Yaozong Gao, He Sui, Bing Zhang, Yu Sun, Zhanhao Mo, and Feng Shi. Segmenting brain tumor using cascaded V-Nets in multimodal MR images. *Frontiers in Computational Neuroscience*, 14, 2 2020.

[8] Tao Lei, Wenzheng Zhou, Yuxiao Zhang, Risheng Wang, Hongying Meng, and Asoke K. Nandi. Lightweight v-net for liver segmentation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1379–1383, 2020.

[9] Ali Hatamizadeh, Dong Yang, Holger Roth, and Daguang Xu. UNETR: Transformers for 3D Medical Image Segmentation. *arXiv (Cornell University)*, 1 2021.

[10] Jeppe Thagaard, Søren Hauberg, Bert Van Der Vegt, Thomas Ebstrup, Johan D. Hansen, and Anders B. Dahl. *Can you trust predictive uncertainty under real dataset shifts in digital pathology?* 1 2020.

[11] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 3 2021.

[12] Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act) (text with eea relevance). Official Journal of the European Union L, 2024/1689, 12 July 2024, 2024. Legal status: In force.

[13] Gianluca Quaglio. Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts, 2022. Available at: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2022)729512.

[14] World Health Organization: WHO. Cancer, 7 2019.

[15] Harry S. Greenberg, William F. Chandler, and Howard M. Sandler. *Brain tumors*. Contemporary Neurology, 1 1999.

[16] Fabio Maria Triulzi. *Neuroradiology of brain tumors*. Springer, 11 2023.

[17] Nicola J. Allen and David A. Lyons. Glia as architects of central nervous system formation and function. *Science*, 362(6411):181–185, 10 2018.

[18] Farina Hanif, Kanza Muzaffar, Kahkashan Perveen, Saima M Malhi, and Shabana U Simjee. Glioblastoma Multiforme: A Review of its Epidemiology and Pathogenesis through Clinical Presentation and Treatment. *PubMed*, 18(1):3–9, 1 2017.

[19] Spyridon Bakas, Mauricio Reyes, András Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Shinohara, Christoph Berger, Sung Ha, Martin Rozycki, Marcel Prastawa, Esther Alberts, Jana Lipkova, John Freymann, Justin Kirby, Michel Bilello, Hassan Fathallah-Shaykh, Roland Wiest, Jan Kirschke, and Jakub Nalepa. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge, 04 2019.

[20] Elton Diêgo Bonifácio, Cleudmar Amaral Araújo, Marcília Valéria Guimarães, Márcio Peres de Souza, Thiago Parente Lima, Bethânia Alves de Avelar Freitas, and Libardo Andrés González-Torres. Computational model of the cancer necrotic core formation in a tumor-on-a-chip device. *Journal of Theoretical Biology*, 592:111893, 2024.

[21] Evert C.A Kaal and Charles J Vecht. The management of brain edema in brain tumors. *Current Opinion in Oncology*, 16(6):593–600, 10 2004.

[22] Sina Mohammadi and Mohamed Allali. Advancing brain tumor segmentation with spectral–spatial graph neural networks. *Applied Sciences*, 14:3424, 04 2024.

[23] Esther Alberts. Multi-modal multi-temporal brain tumor segmentation, growth analysis and texture-based classification, 2019.

[24] Martin J. Van Den Bent, Michael Weller, Patrick Y. Wen, Johan M. Kros, Ken Aldape, and Susan Chang. A clinical perspective on the 2016 WHO brain tumor classification and routine molecular diagnostics. *Neuro-Oncology*, 19(5):614–624, 2 2017.

[25] Daniel G. Eichberg, Long Di, Alexis A. Morell, Ashish H. Shah, Alexa M. Semonche, Christopher N. Chin, Rita G. Bhatia, Aria M. Jamshidi, Evan M. Luther, Ricardo J. Komotar, and Michael E. Ivan. Incidence of high grade gliomas presenting as radiographically

non-enhancing lesions: experience in 111 surgically treated non-enhancing gliomas with tissue diagnosis. *Journal of Neuro-Oncology*, 147(3):671–679, 3 2020.

[26] A. Berger. How does it work?: Magnetic resonance imaging. *BMJ*, 324(7328):35, 1 2002.

[27] Kevin Coyne. MRI: A guided tour. `http://www.magnet.fsu.edu/education/tutorials/magnetacademy/mri/fullarticle.html`, 2012. Accessed: 2025-01-27.

[28] Marc C. Mabray, Ramon F. Barajas, and Soonmee Cha. Modern Brain Tumor Imaging. *Brain Tumor Research and Treatment*, 3(1):8, 1 2015.

[29] Makoto Hosono, Mamoru Takenaka, Hajime Monzen, Mikoto Tamura, Masatoshi Kudo, and Yasumasa Nishimura. Cumulative radiation doses from recurrent PET–CT examinations. *British Journal of Radiology*, 94(1126), 6 2021.

[30] Javier E. Villanueva-Meyer, Marc C. Mabray, and Soonmee Cha. Current clinical brain tumor imaging. *Neurosurgery*, 81(3):397–415, 2 2017.

[31] W. R. Nitz and P. Reimer. Contrast mechanisms in MR imaging. *European Radiology*, 9(6):1032–1046, 7 1999.

[32] University of Wisconsin. Magnetic resonance imaging. Online resource, 2016. Archived from the original on 2017-05-10. Accessed: 2025-01-28.

[33] Gaurav Shukla, Gregory S. Alexander, Spyridon Bakas, Rahul Nikam, Kiran Talekar, Joshua D. Palmer, and Wenyin Shi. Advanced magnetic resonance imaging in glioblastoma: a review. *Chinese Clinical Oncology*, 6(4), 2017.

[34] Rishabh Dhabalia, Shivali V Kashikar, Pratap S Parihar, and Gaurav V Mishra. Unveiling the intricacies: A Comprehensive review of magnetic resonance Imaging (MRI) assessment of T2-Weighted hyperintensities in the neuroimaging landscape. *Cureus*, 2 2024.

[35] R Kates, D Atkinson, and M Brant-Zawadzki. Fluid-attenuated inversion recovery (FLAIR): clinical prospectus of current and future applications. *PubMed*, 8(6):389–96, 12 1996.

[36] Zhihua Liu, Lei Tong, Long Chen, Zheheng Jiang, Feixiang Zhou, Qianni Zhang, Xiangrong Zhang, Yaochu Jin, and Huiyu Zhou. Deep learning based brain tumor segmentation: a survey. *Complex Intelligent Systems*, 9(1):1001–1026, 7 2022.

[37] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C. Kitamura, Sarthak Pati, Luciano M. Prevedello, Jeffrey D. Rudie, Chiharu Sako, Russell T. Shinohara, Timothy Bergquist, Rong Chai, James Eddy, Julia Elliott, Walter Reade, Thomas Schaffter, Thomas Yu, Jiaxin Zheng, Ahmed W. Moawad, Luiz Otavio Coelho, Olivia McDonnell, Elka Miller, Fanny E. Moron, Mark C. Oswood, Robert Y. Shih, Loizos Siakallis, Yulia Bronstein, James R. Mason, Anthony F. Miller, Gagandeep Choudhary, Aanchal Agarwal, Cristina H. Besada, Jamal J. Derakhshan, Mariana C. Diogo, Daniel D. Do-Dai, Luciano Farage, John L. Go, Mohiuddin Hadi, Virginia B. Hill, Michael Iv, David Joyner, Christie Lincoln, Eyal Lotan, Asako Miyakoshi, Mariana Sanchez-Montano, Jaya Nath, Xuan V. Nguyen, Manal Nicolas-Jilwan, Johanna Ortiz Jimenez, Kerem Ozturk, Bojan D. Petrovic, Chintan Shah, Lubdha M. Shah, Manas Sharma, Onur Simsek, Achint K. Singh, Salil Soman, Volodymyr Statsevych, Brent D. Weinberg, Robert J. Young, Ichiro Ikuta, Amit K.

Agarwal, Sword C. Cambron, Richard Silbergleit, Alexandru Dusoi, Alida A. Postma, Laurent Letourneau-Guillon, Gloria J. Guzman Perez-Carrillo, Atin Saha, Neetu Soni, Greg Zaharchuk, Vahe M. Zohrabian, Yingming Chen, Milos M. Cekic, Akm Rahman, Juan E. Small, Varun Sethi, Christos Davatzikos, John Mongan, Christopher Hess, Soonmee Cha, Javier Villanueva-Meyer, John B. Freymann, Justin S. Kirby, Benedikt Wiestler, Priscila Crivellaro, Rivka R. Colen, Aikaterini Kotrotsou, Daniel Marcus, Mikhail Milchenko, Arash Nazeri, Hassan Fathallah-Shaykh, Roland Wiest, Andras Jakab, Marc-Andre Weber, Abhishek Mahajan, Bjoern Menze, Adam E. Flanders, and Spyridon Bakas. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification, 2021.

[38] Wenyin Zhang, Yong Wu, Bo Yang, Shunbo Hu, Liang Wu, and Sahraoui Dhelim. Overview of Multi-Modal Brain Tumor MR Image Segmentation. *Healthcare*, 9(8):1051, 8 2021.

[39] Amit Verma, Shiv Naresh Shivhare, Shailendra P. Singh, Naween Kumar, and Anand Nayyar. Comprehensive Review on MRI-Based Brain Tumor Segmentation: A Comparative Study from 2017 Onwards. *Archives of Computational Methods in Engineering*, 5 2024.

[40] Md. Faysal Ahamed, Md. Munawar Hossain, Md. Nahiduzzaman, Md. Rabiul Islam, Md. Robiul Islam, Mominul Ahsan, and Julfikar Haider. A review on brain tumor segmentation based on deep learning methods with federated learning techniques. *Computerized Medical Imaging and Graphics*, 110:102313, 11 2023.

[41] Supun Nakandala and Arun Kumar. Materialization trade-offs for feature transfer from deep cnns for multimodal data analytics. 2018.

[42] Satoshi Takahashi, Yusuke Sakaguchi, Nobuji Kouno, Ken Takasawa, Kenichi Ishizu, Yu Akagi, Rina Aoyama, Naoki Teraya, Amina Bolatkan, Norio Shinkai, Hidenori Machino, Kazuma Kobayashi, Ken Asada, Masaaki Komatsu, Syuzo Kaneko, Masashi Sugiyama, and Ryuji Hamamoto. Comparison of vision transformers and convolutional neural networks in Medical Image Analysis: a systematic review. *Journal of Medical Systems*, 48(1), 9 2024.

[43] Darko Zikic, Yani Ioannou, Matthew Brown, and Antonio Criminisi. Segmentation of brain tumor tissues with convolutional neural networks. 09 2014.

[44] G. Urban, M. Bendszus, F.A. Hamprecht, and Jens Kleesiek. Multi-modal brain tumor segmentation using deep convolutional neural networks. *Miccai-Bratss*, pages 31–35, 01 2014.

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[46] Andrew Beers, Ken Chang, James Brown, Emmett Sartor, CP Mammen, Elizabeth Gerstner, Bruce Rosen, and Jayashree Kalpathy-Cramer. Sequential 3d u-nets for biologically-informed brain tumor segmentation, 2017.

[47] Kamlesh Pawar, Zhaolin Chen, N. Shah, and Gary Egan. Residual encoder and convolutional decoder neural network for glioma segmentation. 02 2018.

[48] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2015.

[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[50] Sarahi Rosas-Gonzalez, Taibou Birgui-Sekou, Moncef Hidane, Ilyess Zemmoura, and Clovis Tauber. Asymmetric ensemble of asymmetric U-Net models for brain tumor segmentation with uncertainty estimation. *Frontiers in Neurology*, 12, 9 2021.

[51] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation, 2016.

[52] Canxuan Gang. A novel convolutional-free method for 3d medical imaging segmentation, 2025.

[53] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[54] Yutong Xie, Bing Yang, Qingbiao Guan, Jianpeng Zhang, Qi Wu, and Yong Xia. Attention mechanisms in medical image segmentation: A survey, 2023.

[55] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019.

[56] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas, 2018.

[57] Wenguang Yuan, Jia Wei, Jiabing Wang, Qianli Ma, and Tolga Tasdizen. Unified attentional generative adversarial network for brain tumor segmentation from multimodal unpaired images, 2019.

[58] Mehrdad Noori, Ali Bahri, and Karim Mohammadi. Attention-guided version of 2d unet for automatic brain tumor segmentation. In *2019 9th International Conference on Computer and Knowledge Engineering (ICCKE)*. IEEE, October 2019.

[59] Hai Xu, Hongtao Xie, Yizhi Liu, Chuandong Cheng, Chaoshi Niu, and Yongdong Zhang. Deep cascaded attention network for multi-task brain tumor segmentation. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 420–428, Cham, 2019. Springer International Publishing.

[60] Chenhong Zhou, Changxing Ding, Xinchao Wang, Zhentai Lu, and Dacheng Tao. One-pass multi-task networks with cross-task guided attention for brain tumor segmentation. *IEEE Transactions on Image Processing*, 29:4516–4529, 2020.

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[62] Yue Zhou, Xue Jiang, Guozheng Xu, Xue Yang, Xingzhao Liu, and Zhou Li. Pvt-sar: An arbitrarily oriented sar ship detector with pyramid vision transformer. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP:1–15, 01 2022.

[63] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation, 2022.

[64] Cheng Liu and Hisanor Kiryu. 3d medical axial transformer: A lightweight transformer model for 3d brain tumor segmentation. In Ipek Oguz, Jack Noble, Xiaoxiao Li, Martin Styner, Christian Baumgartner, Mirabela Rusu, Tobias Heinmann, Despina Kontos, Bennett Landman, and Benoit Dawant, editors, *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pages 799–813. PMLR, 10–12 Jul 2024.

[65] Wenxuan Wang, Chen Chen, Meng Ding, Jiangyun Li, Hong Yu, and Sen Zha. Transbts: Multimodal brain tumor segmentation using transformer, 2021.

[66] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, 2022.

[67] Ranadeep Bhuyan and Gypsy Nandi. *Brain Tumour—Augmentation, Segmentation and Classification Using Deep Learning—A Review*, pages 209–229. 11 2023.

[68] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew Lungren, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou. 3d transunet: Advancing medical image segmentation through vision transformers, 2023.

[69] Fuxin Fan, Jingna Qiu, YiXing Huang, and Andreas Maier. Enhancing cross-modality synthesis: Subvolume merging for mri-to-ct conversion, 09 2024.

[70] Jeremiah Fadugba, Isabel Lieberman, Olabode Ajayi, Mansour Osman, Solomon Oluwole Akinola, Tinashe Mustvangwa, Dong Zhang, Udunna C Anazondo, and Raymond Confidence. Deep ensemble approach for enhancing brain tumor segmentation in resource-limited settings, 2025.

[71] Konstantinos Kamnitsas, Wenjia Bai, Enzo Ferrante, Steven McDonagh, Matthew Sinclair, Nick Pawlowski, Martin Rajchl, Matthew Lee, Bernhard Kainz, Daniel Rueckert, and Ben Glocker. Ensembles of multiple models and architectures for robust brain tumour segmentation, 2017.

[72] André Ferreira, Naida Solak, Jianning Li, Philipp Dammann, Jens Kleesiek, Victor Alves, and Jan Egger. How we won brats 2023 adult glioma challenge? just faking it! enhanced synthetic data augmentation and model ensemble for brain tumour segmentation, 2024.

[73] Shiv Naresh Shivhare and Nitin Kumar. Tumor bagging: a novel framework for brain tumor segmentation using metaheuristic optimization algorithms. *Multimedia Tools and Applications*, 80(17):26969–26995, 5 2021.

[74] Sheik Imran and Pradeep N. A review on ensemble machine and deep learning techniques used in the classification of computed tomography medical images. *International Journal of Health Sciences and Research*, 14(1):201–213, 1 2024.

[75] Asadullah Shaikh, Samina Amin, Muhammad Ali Zeb, Adel Sulaiman, Mana Saleh Al Reshan, and Hani Alshahrani. Enhanced brain tumor detection and segmentation using densely connected convolutional networks with stacking ensemble learning. *Computers in Biology and Medicine*, 186:109703, 2025.

[76] S.K. Warfield, K.H. Zou, and W.M. Wells. Simultaneous Truth and Performance Level Estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 7 2004.

[77] Alain Jungo, Fabian Balsiger, and Mauricio Reyes. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in Neuroscience*, 14, 4 2020.

[78] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, and Christian Wachinger. Bayesian quicknat: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage*, 195:11–22, 2019.

[79] Ke Zou, Zhihao Chen, Xuedong Yuan, Xiaojing Shen, Meng Wang, and Huazhu Fu. A Review of Uncertainty Estimation and its Application in Medical Imaging. *arXiv (Cornell University)*, 1 2023.

[80] Shahriar Faghani, Mana Moassefi, Pouria Rouzrokh, Bardia Khosravi, Francis I. Baffour, Michael D. Ringler, and Bradley J. Erickson. Quantifying uncertainty in deep learning of radiologic images. *Radiology*, 308(2), 8 2023.

[81] Martijn J. Mulder, Max C. Keuken, Pierre-Louis Bazin, Anneke Alkemade, and Birte U. Forstmann. Size and shape matter: The impact of voxel geometry on the identification of small nuclei. *PLoS ONE*, 14(4):e0215382, 4 2019.

[82] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(S1):1513–1589, 7 2023.

[83] Janis Postels, Mattia Segu, Tao Sun, Luca Sieber, Luc Van Gool, Fisher Yu, and Federico Tombari. On the practicality of deterministic epistemic uncertainty, 2022.

[84] Thierry Judge, Olivier Bernard, Mihaela Porumb, Agis Chartsias, Arian Beqiri, and Pierre-Marc Jodoin. Crisp - reliable uncertainty estimation for medical image segmentation, 2022.

[85] Ke Zou, Yidi Chen, Ling Huang, Xuedong Yuan, Xiaojing Shen, Meng Wang, Rick Siow Mong Goh, Yong Liu, and Huazhu Fu. Towards reliable medical image segmentation by utilizing evidential calibrated uncertainty, 2024.

[86] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, December 2021.

[87] Theodore Papamarkou, Jacob Hinkle, M. Todd Young, and David Womble. Challenges in Markov chain Monte Carlo for Bayesian neural networks. *Statistical Science*, 37(3), 6 2022.

[88] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016.

[89] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics Data Analysis*, 142:106816, 2020.

[90] Qingqiao Hu, Hao Wang, Jing Luo, Yunhao Luo, Zhiheng Zhangg, Jan S. Kirschke, Benedikt Wiestler, Bjoern Menze, Jianguo Zhang, and Hongwei Bran Li. Inter-rater uncertainty quantification in medical image segmentation via rater-specific bayesian neural networks, 2023.

[91] Lohith Konathala. Bayesian neural networks for 2d mri segmentation, 2024.

[92] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[93] Laura Mora Ballestar and Veronica Vilaplana. Mri brain tumor segmentation and uncertainty estimation using 3d-unet architectures. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 376–390, Cham, 2021. Springer International Publishing.

[94] Nikita Moshkov, Botond Mathe, Attila Kertesz-Farkas, Reka Hollandi, and Peter Horvath. Test-time augmentation for deep learning-based cell segmentation on microscopy images. *Scientific Reports*, 10(1), 3 2020.

[95] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017.

[96] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, 2019.

[97] Raghav Mehta, Angelos Filos, Ujjwal Baid, Chiharu Sako, Richard McKinley, Michael Rebsamen, Katrin Dätwyler, Raphael Meier, Piotr Radojewski, Gowtham Krishnan Murugesan, Sahil Nalawade, Chandan Ganesh, Ben Wagner, Fang F. Yu, Baowei Fei, Ananth J. Madhuranthakam, Joseph A. Maldjian, Laura Daza, Catalina Gómez, Pablo Arbeláez, Chengliang Dai, Shuo Wang, Hadrien Reynaud, Yuanhan Mo, Elsa Angelini, Yike Guo, Wenjia Bai, Subhashis Banerjee, Linmin Pei, Murat Ak, Sarahi Rosas-González, Ilyess Zemmoura, Clovis Tauber, Minh H. Vu, Tufve Nyholm, Tommy Löfstedt, Laura Mora Ballestar, Veronica Vilaplana, Hugh McHugh, Gonzalo Maso Talou, Alan Wang, Jay Patel, Ken Chang, Katharina Hoebel, Mishka Gidwani, Nishanth Arun, Sharut Gupta, Mehak Aggarwal, Praveer Singh, Elizabeth R. Gerstner, Jayashree Kalpathy-Cramer, Nicolas Boutry, Alexis Huard, Lasitha Vidyaratne, Md Monibor Rahman, Khan M. Iftekharuddin, Joseph Chazalon, Elodie Puybareau, Guillaume Tochon, Jun Ma, Mariano Cabezas, Xavier Llado, Arnau Oliver, Liliana Valencia, Sergi Valverde, Mehdi Amian, Mohammadreza Soltaninejad, Andriy Myronenko, Ali Hatamizadeh, Xue Feng, Quan Dou, Nicholas Tustison, Craig Meyer, Nisarg A. Shah, Sanjay Talbar, Marc-André Weber, Abhishek Mahajan, Andras Jakab, Roland Wiest, Hassan M. Fathallah-Shaykh, Arash Nazeri, Mikhail Milchenko, Daniel Marcus, Aikaterini Kotrotsou, Rivka Colen, John Freymann, Justin Kirby, Christos Davatzikos, Bjoern Menze, Spyridon Bakas, Yarin Gal, and Tal Arbel. QU-BRATS: MICCAI BRATS 2020 Challenge on Quantifying Uncertainty in Brain Tumor segmentation – Analysis of ranking scores and benchmarking

results. *The Journal of Machine Learning for Biomedical Imaging*, 1(August 2022):1–54, 8 2022.

[98] Hao Li, Yang Nan, Javier Del Ser, and Guang Yang. Region-based evidential deep learning to quantify uncertainty and improve robustness of brain tumor segmentation, 2022.

[99] Joohyun Lee, Dongmyung Shin, Se-Hong Oh, and Haejin Kim. Method to minimize the errors of AI: Quantifying and exploiting uncertainty of deep learning in brain tumor segmentation. *Sensors*, 22(6):2406, 3 2022.

[100] Zain Ul Abidin, Rizwan Ali Naqvi, Amir Haider, Hyung Seok Kim, Daesik Jeong, and Seung Won Lee. Recent deep learning-based brain tumor segmentation models using multi-modality magnetic resonance imaging: a prospective survey. *Frontiers in Bioengineering and Biotechnology*, 12, 7 2024.

[101] David Bouget, Demah Alsinan, Valeria Gaitan, Ragnhild Holden Helland, André Pedersen, Ole Solheim, and Ingerid Reinertsen. Raidionics: an open software for pre- and postoperative central nervous system tumor segmentation and standardized reporting. *Scientific Reports*, 13(1), 9 2023.

[102] Evgenii Belykh, Kurt V. Shaffer, Chaoqun Lin, Vadim A. Byvaltsev, Mark C. Preul, and Lukui Chen. Blood-Brain barrier, Blood-Brain tumor barrier, and Fluorescence-Guided Neurosurgical Oncology: delivering optical labels to brain tumors. *Frontiers in Oncology*, 10, 6 2020.

[103] Costas D. Arvanitis, Gino B. Ferraro, and Rakesh K. Jain. The blood–brain barrier and blood–tumour barrier in brain tumours and metastases. *Nature reviews. Cancer*, 20(1):26–41, 10 2019.

[104] Peter Jagd Sørensen, Claes Nøhr Ladefoged, Vibeke Andrée Larsen, Flemming Littrup Andersen, Michael Bachmann Nielsen, Hans Skovgaard Poulsen, Jonathan Frederik Carlsen, and Adam Espe Hansen. Repurposing the public BRATS dataset for Postoperative Brain Tumour Treatment Response monitoring. *Tomography*, 10(9):1397–1410, 9 2024.

[105] Ramin Ranjbarzadeh, Abbas Bagherian Kasgari, Saeid Jafarzadeh Ghoushchi, Shokofeh Anari, Maryam Naseri, and Malika Bendechache. Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images. *Scientific Reports*, 11(1), 5 2021.

[106] MONAI Consortium. MONAI: Medical Open Network for AI (1.4.0), 2024.

[107] Andriy Myronenko. *3D MRI brain tumor segmentation using Autoencoder regularization.* 1 2019.

[108] Zhifan Jiang, Daniel Capellán-Martín, Abhijeet Parida, Xinyang Liu, María J. Ledesma-Carbayo, Syed Muhammad Anwar, and Marius George Linguraru. Enhancing generalizability in brain tumor segmentation: Model ensemble with adaptive post-processing. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–4, 2024.

[109] Ebtihal J. Alwadee, Xianfang Sun, Yipeng Qin, and Frank C. Langbein. Latup-net: A lightweight 3d attention u-net with parallel convolutions for brain tumor segmentation. *Computers in Biology and Medicine*, 184:109353, January 2025.

[110] Md Alamin Talukder, Md Abu Layek, Md Aslam Hossain, Md Aminul Islam, Mohammad Nur e Alam, and Mohsin Kazi. Acu-net: Attention-based convolutional u-net model for segmenting brain tumors in fmri images. *DIGITAL HEALTH*, 11:20552076251320288, 2025.

[111] Suorong Yang, Weikang Xiao, Mengchen Zhang, Suhan Guo, Jian Zhao, and Furao Shen. Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*, 2022.

[112] Fabio Garcea, Alessio Serra, Fabrizio Lamberti, and Lia Morra. Data augmentation for medical imaging: A systematic literature review. *Computers in Biology and Medicine*, 152:106391, 2023.

[113] W.R. Crum, O. Camara, and D.L.G. Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, 25(11):1451–1461, 2006.

[114] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. *Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations*, page 240–248. Springer International Publishing, 2017.

[115] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.

[116] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[117] Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nathan Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning, 2018.

[118] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[119] Terrance DeVries and Graham W. Taylor. Learning confidence for out-of-distribution detection in neural networks, 2018.

[120] Aryan Mobiny, Pengyu Yuan, Supratik K. Moulik, Naveen Garg, Carol C. Wu, and Hien Van Nguyen. DropConnect is effective in modeling uncertainty of Bayesian deep networks. *Scientific Reports*, 11(1), 3 2021.

[121] Grzegorz Chlebus, Andrea Schenk, Horst K. Hahn, Bram Van Ginneken, and Hans Meine. Robust segmentation models using an uncertainty slice sampling-based annotation workflow. *IEEE Access*, 10:4728–4738, 2022.

[122] Cedrique Tassi, Jakob Gawlikowski, Auliya Fitri, and Rudolph Triebel. The impact of averaging logits over probabilities on ensembles of neural networks. 01 2022.

[123] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017.

[124] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015.

[125] Alireza Mehrtash, William M Wells III, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation. *arXiv preprint arXiv:1911.13273*, 2020.

[126] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[127] Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2151–2159. PMLR, Jun 2019.

[128] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[129] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 7 1945.

[130] M.-P. Dubuisson and A.K. Jain. A modified hausdorff distance for object matching. In *Proceedings of 12th International Conference on Pattern Recognition*, volume 1, pages 566–568 vol.1, 1994.

[131] Streamlit • A faster way to build and share data apps.

[132] Randall Fulton and Roy Vandermolen. *Airborne Electronic Hardware Design Assurance: A Practitioner's Guide to RTCA/DO-254*. CRC Press, Inc., USA, 2014.

[133] Martin Glinz. On non-functional requirements. In *Proceedings of the 15th IEEE International Requirements Engineering Conference (RE'07)*, pages 21–26, New Delhi, India, October 2007. IEEE.

[134] Martin Fowler. *UML Distilled: A brief guide to the standard object modeling language*. 1 1997.

[135] Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, Martin Bendszus, Klaus H. Maier-Hein, and Philipp Kickingereder. Automated brain extraction of multisequence mri using artificial neural networks. *Human Brain Mapping*, 40(17):4952–4964, August 2019.

[136] Richard Beare, Bradley Lowekamp, and Ziv Yaniv. Image segmentation, registration and characterization in r with simpleitk. *Journal of Statistical Software*, 86(8):1–35, 2018.

[137] Wenjian Yao, Jiajun Bai, Wei Liao, Yuheng Chen, Mengjuan Liu, and Yao Xie. From CNN to Transformer: A review of Medical Image Segmentation models. *Deleted Journal*, 37(4):1529–1547, 3 2024.

[138] Martin Kolarik, Radim Burget, and Kamil Riha. Comparing normalization methods for limited batch size segmentation neural networks. In *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*, page 677–680. IEEE, July 2020.

[139] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[140] Shayan Shekarforoush, David B. Lindell, David J. Fleet, and Marcus A. Brubaker. Residual multiplicative filter networks for multiscale reconstruction, 2022.

[141] Zeyu Jiang, Changxing Ding, Minfeng Liu, and Dacheng Tao. Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I 5*, pages 231–241. Springer, 2020.

[142] Huan Minh Luu and Sung-Hong Park. Extending nn-unet for brain tumor segmentation. In *International MICCAI brainlesion workshop*, pages 173–186. Springer, 2021.

[143] Ramy A. Zeineldin, Mohamed E. Karar, Oliver Burgert, and Franziska Mathis-Ullrich. Multimodal cnn networks for brain tumor segmentation in mri: A brats 2022 challenge solution, 2022.

[144] Theophraste Henry, Alexandre Carre, Marvin Lerousseau, Theo Estienne, Charlotte Robert, Nikos Paragios, and Eric Deutsch. Brain tumor segmentation with self-ensembled, deeply-supervised 3d u-net neural networks: a brats 2020 challenge solution, 2020.

[145] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sebastien Ourselin, and Tom Vercauteren. Test-time augmentation with uncertainty estimation for deep learning-based medical image segmentation, 07 2018.

[146] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, April 2019.

[147] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The values encoded in machine learning research, 2022.

[148] David Liu, Priyanka Nanayakkara, Sarah Ariyan Sakha, Grace Abuhamad, Su Lin Blodgett, Nicholas Diakopoulos, Jessica R. Hullman, and Tina Eliassi-Rad. Examining responsibility and deliberation in ai impact statements and ethics reviews. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 424–435, New York, NY, USA, 2022. Association for Computing Machinery.

[149] Clyde Morgan. ChatGPT's Energy Consumption: A Closer look. 2 2025.

[150] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training, 2018.

[151] Rusty Flint. Edge computing Benefits: Edge vs cloud computing. 9 2024.

[152] John Loeffler. GPU prices in 2025: current prices on all the latest graphics cards on the market, 5 2025.

[153] Regulation - 2017/745 - EN - Medical Device Regulation - EUR-LEX.

[154] F. Windisch, N. Zimmermann, K. Habimana, V. Knoll, S. Fischer, and S. Vogler. One-pager (long version) on "study supporting the monitoring of the availability of medical devices on the eu market (md availability)". Technical report, Gesundheit Österreich GmbH, Vienna, 2023.

[155] Max Bengtsson, Elif Keles, Gorkem Durak, Syed Anwar, Yuri S. Velichko, Marius G. Linguraru, Angela J. Waanders, and Ulas Bagci. A new logic for pediatric brain tumor segmentation, 2025.

[156] Code of ethics and good practice. https://canviaelmon.upc.edu/en/social-responsibility-at-the-upc-1/policies-institutional-commitment/code-ethics-good-practice. Accessed: 2025-05-05.