TU WIEN Informatics

Evaluating the Impact of Parameter Tuning on Glioblastoma Segmentation using Deep Learning

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Medizinische Informatik

eingereicht von

Dr.med.univ. Lukas Machegger, PhD, MBA, Bakk.techn.
Matrikelnummer 09900294

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dipl.-Ing. Dr.techn. Eduard Gröller
Mitwirkung: Assistant Prof. Renata Raidou, MSc, PhD

Wien, 2. März 2025

Lukas Machegger	Eduard Gröller

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Evaluating the Impact of Parameter Tuning on Glioblastoma Segmentation using Deep Learning

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Medical Informatics

by

Dr.med.univ. Lukas Machegger, PhD, MBA, Bakk.techn.
Registration Number 09900294

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dipl.-Ing. Dr.techn. Eduard Gröller
Assistance: Assistant Prof. Renata Raidou, MSc, PhD

Vienna, March 2, 2025

Lukas Machegger	Eduard Gröller

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Dr.med.univ. Lukas Machegger, PhD, MBA, Bakk.techn.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, haben ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT-Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 2. März 2025

Lukas Machegger

Danksagung

Zunächst möchte ich meine Dankbarkeit gegenüber Univ.Prof. Dipl.-Ing. Dr.techn. Eduard Gröller für seine hervorragende Betreuung und seine unschätzbare Unterstützung während des gesamten Entstehungsprozesses dieser Arbeit ausdrücken. Seine Anleitung war entscheidend für die Ausrichtung meiner Forschungsarbeit.

Ebenso möchte ich mich bei Assistant Prof. Renata Raidou, MSc, PhD für ihre Unterstützung und ihr Engagement während meiner Forschung bedanken. Ihre wertvollen Einblicke und ihr Feedback waren von großem Nutzen, und ich bin sehr dankbar für ihre Hingabe.

Ein besonderer Dank gilt meinem Kollegen und Freund Dipl.-Ing. Dr.techn. Jürgen Steinbacher, dessen ständige Bereitschaft, sich in Diskussionen einzubringen und Ratschläge zu geben, mir enorm geholfen hat. Seine praktischen Tipps und sein professioneller Input waren während des gesamten Projekts unverzichtbar.

Besonders dankbar bin ich meiner Frau Melanie, meinem Sohn Georg und meiner Tochter Marlene, die in den letzten zwölf Monaten so viel geopfert haben, um mir zu ermöglichen, mich ganz auf die Fertigstellung dieser Arbeit zu konzentrieren. Ohne ihr Verständnis und ihre unerschütterliche Unterstützung wäre dies nicht möglich gewesen.

Acknowledgements

First and foremost, I would like to express my gratitude to Univ.Prof. Dipl.-Ing. Dr.techn. Eduard Gröller for his excellent supervision and invaluable assistance throughout the process of this thesis. His guidance was fundamental in shaping the direction of my work.

I would also like to sincerely thank Assistant Prof. Renata Raidou, MSc, PhD for her support and involvement during this research. Her insights and feedback were of great value, and I am grateful for her dedication.

A special thanks go to my colleague and friend Dipl.-Ing. Dr.techn. Jürgen Steinbacher, whose constant willingness to engage in discussions and provide advice helped me tremendously. His practical tips and professional input were indispensable throughout the entire project.

I am especially thankful to my wife Melanie, my son Georg, and my daughter Marlene, who sacrificed so much over the last 12 months to allow me to focus on completing this thesis. Without their understanding and unwavering support, this would not have been possible.

Kurzfassung

Hintergrund: Glioblastoma multiforme (GBM) ist die aggressivste Form der Hirntumoren, gekennzeichnet durch schnelles Wachstum und Infiltration in umliegendes Hirngewebe. Eine präzise Segmentierung von GBM, insbesondere der pathologischen Kontrastmittelaufladung und des nekrotischen Kerns, ist entscheidend für die chirurgische Planung und Behandlung. Manuelle Segmentierung ist zeitaufwändig und unterliegt hoher Interrater-Variabilität, dies erfordert automatisierte Ansätze.

Ziel: Diese Arbeit optimiert wichtige Parameter in der Deep-Learning-Segmentierung von Glioblastomen. Der Fokus liegt auf Batch-Größe, Datenaugmentierung und der Anzahl der Trainingsfälle sowie der Abstimmung des fokalen Gewichtsfaktors in der kombinierten Verlustfunktion. Ziel ist es, die Genauigkeit der Segmentierung klinisch relevanter Tumorregionen zu verbessern.

Methoden: 3D U-Net-Modelle wurden mit dem BraTS Challenge-Datensatz trainiert, der multimodale MRT-Scans (T1 post-Kontrast, FLAIR, T2) enthält, die von einem Neuroradiologen überarbeitet wurden, um Interrater-Variabilität zu eliminieren. Die Modelle wurden an 108 Patienten des Universitätsklinikums Salzburg getestet, um Generalisierungsfähigkeit und Leistung zu bewerten. Die Genauigkeit wurde mit Intersection-over-Union (IoU) und einem benutzerdefinierten gewichteten Dice-Score gemessen, wobei der Fokus auf den Dice-Koeffizienten der Kontrastmittelaufladung und des nekrotischen Kerns lag. Vier Fallgruppen (80, 160, 240, 314 Fälle) wurden untersucht, um den Einfluss der Fallzahl auf die Leistung zu analysieren.

Ergebnisse: Modelle mit Batch-Größe 4 gehörten konsistent zu den besten, 80% davon unter den Top 10. Größere Batch-Größen führten zu besserer Generalisierung und Stabilität bei steigender Fallzahl. Augmentierungen führten in der Regel zu schlechteren Ergebnissen, außer beim besten Modell, das mit einem 1:1-Verhältnis von Augmentierungen zu Originalen, Fallgruppe 314 und einer Batch-Größe von 1 trainiert wurde und außergewöhnlich gut abschnitt.

Fazit: Augmentierungen mit einem 1:3-Verhältnis schnitten schlecht ab, besonders, wenn drei Varianten eines Originals in einer Batch von 4 waren, was zu Overfitting führte. Ein Mangel an Diversität innerhalb der Batches verursachte Overfitting, während eine Strategie, die verschiedene Augmentierungen mischte, besser generalisierte. Modelle der Fallgruppe 314 erzielten die besten Ergebnisse, was die Bedeutung größerer Datenmengen für eine verbesserte Leistung zeigt.

Abstract

Background: Glioblastoma multiforme (GBM) is the most aggressive form of brain cancer, characterized by rapid growth and infiltration into surrounding brain tissue. Precise segmentation of GBM, particularly the contrast-enhancing region and necrotic (non-contrast-enhaning) core, is critical for surgical planning and treatment. Manual segmentation methods are time-consuming and subject to high interrater variability, necessitating automated approaches for greater consistency.

Objective: This thesis aims to optimize key parameters in deep learning-based segmentation of glioblastomas, focusing on the impact of Batch size, data augmentation strategies, and the number of training cases on model performance, along with tuning the Focal Weight Factor in the Combined Loss Function. The goal is to improve the accuracy of segmenting clinically relevant tumor regions.

Methods: In this study, 3D U-Net models were trained using the BraTS Challenge dataset, which includes multimodal MRI scans (T1 post-contrast, FLAIR, and T2) with expert-labeled segmentations reviewed by a neuroradiologist to eliminate interrater variability. The models were evaluated on 108 unseen clinical cases from patients at the University Hospital Salzburg to assess their generalization capability and performance. Segmentation accuracy was measured using Intersection over Union (IoU) and a Custom Weighted Dice Score, focusing on Dice coefficients for the contrast-enhancing and non-contrast-enhancing tumor. Four Case Groups (80, 160, 240, and 314) were used to examine the effect of Case Group size on performance.

Results: Models trained with Batch size of four consistently ranked among the top performers, with 80% making it into the top 10, suggesting that larger Batch sizes contribute to better generalization and stability as number of training cases increase. However, augmentations generally resulted in worse performance, except for one outlier—the best performing model—trained with a 1:1 ratio of augmentations to originals, Case Group 314, and a Batch size of one, which performed exceptionally well.

Conclusion: Augmentations with a ratio of 1:3 performed poorly, particularly when three variants of one original were included in a Batch size of four, leading to overfitting. This suggests a lack of diversity within the batches caused the model to overfit, whereas a strategy mixing different augmentations within each batch led to better generalization. Case Group 314 models performed best, highlighting the importance of more training data for improved performance.

Contents

CHAPTER 1

Introduction

1.1 Motivation and Problem Definition

Glioblastoma multiforme (GBM) is the most frequent and most lethal tumor of the central nervous system. It is characterized by its aggressive growth, including rapid proliferation and extensive infiltration into surrounding brain tissue, which complicates treatment [DLCG16]. Even with modern treatments, including surgery, radiotherapy, and chemotherapy, the prognosis remains poor [BV09]. The median survival for patients undergoing optimal treatment, including maximal resection and adjuvant chemoradiotherapy, is around 14 months [BV09, DLCG16].

Accurate segmentation of glioblastoma, especially the contrast-enhancing regions and necrotic (non-contrast-enhancing) core, is crucial for surgical planning. These regions represent the primary targets for resection because they contain the most aggressive tumor cells [FTPM20]. Incomplete removal of these regions significantly increases the risk of tumor recurrence, especially at the resection margins, due to GBM's infiltrative nature [CIF+23]. Precise delineation is essential to reduce recurrence risks and improve patient outcomes [FTPM20].

Manual segmentation of glioblastoma in Magnetic Resonance Imaging (MRI) scans is time-consuming and depends heavily on the expertise of neuroradiologists. Interrater variability, a well-known issue, leads to inconsistencies in tumor volume assessments [CMB+22]. This variability stems from the subjective nature of manual delineation, particularly in difficult cases where tumor boundaries are unclear, impacting both clinical decisions and treatment outcomes [WRP20]. Manual segmentation is especially error-prone if tumor margins are difficult to define or infiltrate surrounding tissue. The challenge of distinguishing tumor tissue from surrounding edema further complicates achieving accurate and consistent segmentation [VMV+18].

Automated segmentation with advanced machine learning techniques has proven effective in reducing the variability of manual methods, offering more consistent and reliable results [LQX+24]. This approach enhances the accuracy of organ and tumor segmentation compared to manual methods and helps reduce discrepancies in treatment planning, which can impact therapies like radiation treatment [LQX+24].

In recent years, deep learning has emerged as a powerful tool for automating medical image analysis, particularly for segmentation tasks [LJZZ23]. Convolutional neural networks (CNNs), especially U-Net architectures [HJHK19], have demonstrated superior performance in biomedical image segmentation tasks, enabling accurate delineation of complex anatomical structures. For GBM, deep learning models hold the potential to standardize the segmentation process, reduce interrater variability, and provide fast, reproducible results [CAMS+23]. These methods are especially effective in capturing critical tumor subregions, such as contrast-enhancing and necrotic (non-contrast-enhancing) areas, which are essential for surgical resection and radiation therapy planning [HJHK19, LQX+24].

The heterogeneity of GBMs poses significant challenges for automated segmentation. Each tumor subregion, including contrast-enhancing areas, the necrotic (non-contrast-enhancing) core, and surrounding edema (as shown in Figure 2.1), has distinct radiological and morphological features that complicate the process [FLG+24]. While contrast-enhancing regions are usually clear in MRI scans, the necrotic (non-contrast-enhancing) core lacks contrast, making it harder to differentiate [WC18]. Additionally, edema often has diffuse, poorly defined borders that overlap with healthy brain tissue, further complicating segmentation [FLG+24].

Since the tumor volume is significantly smaller compared to healthy tissue, this leads to class imbalance, causing models to often favor the majority class. As a result, smaller but clinically relevant tumor areas may be segmented suboptimally [WC18]. To address this issue, advanced loss functions such as the Focal Dice Loss have been developed [WC18]. These integrate class weights to prioritize accurate segmentation of smaller, critical regions such as the necrotic (non-contrast-enhancing) core and contrast-enhancing areas [YSSR22]. This method reduces bias toward the majority class and enhances segmentation accuracy for minority classes [YSSR22].

Optimizing deep learning models for medical image segmentation involves fine-tuning key hyperparameters, including Batch size, data augmentation strategies, and the loss function [MD22]. Batch size is especially important for model convergence and segmentation accuracy, with smaller sizes often preferred in imbalanced datasets as they introduce gradient noise, preventing the model from getting trapped in local minima and improving generalization [MD22]. Well-tuned data augmentation strategies are also crucial for enhancing model robustness, allowing better generalization to new, unseen data, and helping to prevent overfitting, particularly when training on small datasets [OdANC23].

Data augmentation is vital for improving the generalizability of deep learning models, especially when training data is limited. Transformations like rotation, flipping and scaling artificially increase the size and diversity of the dataset, helping to reduce overfitting

[YXZ+22]. This allows the model to learn more robust features and perform better on unseen data [SK19]. In medical imaging, where large datasets are often hard to obtain, data augmentation has proven effective in enhancing model performance by introducing variations that improve generalization [SK19, YXZ+22].

1.2 Aim of the Work

The main goal of this work is to create an automated multisequence volumetric segmentation pipeline for glioblastoma, using deep learning to optimize performance through parameter tuning. The focus is on understanding how modifiable parameters—such as Batch size, data augmentation, and the number of training cases—impact the accuracy and efficiency of segmentation. Additionally, the work examines optimizing the Focal Weight Factor within the Combined Loss Function. Through hyperparameter tuning, the optimal Focal Weight Factor is determined to achieve the best segmentation performance. This is assessed using evaluation metrics such as Intersection over Union (IoU) and Dice Score shown in Section 4.1. The ultimate aim is to improve segmentation accuracy, especially in the challenging peritumoral regions, by maximizing these key metrics.

Main Research Question:
How do different parameters impact the performance of glioblastoma segmentation using deep learning, as measured by Intersection over Union (IoU) and Dice Score? Additionally, how does optimizing the Focal Weight Factor in the Combined Loss Function affect segmentation accuracy? This research focuses on how modifiable parameters—Batch size, data augmentation, the number of training cases, and the Focal Weight Factor—affect segmentation performance, particularly in clinically relevant tumor subregions.

Sub-Questions:

Q1. How does the choice of Batch size affect the training time and segmentation accuracy, as measured by the IoU and Dice Score?

Q2. Effect of augmentations:

- **a.** What is the impact of augmentation on segmentation performance?
- **b.** How does the number of augmentation cases affect the outcome?
- **c.** What is the effect of the ratio of augmentation cases to original training cases on the segmentation results?
- **d.** How do different augmentation strategies (e.g., using variants of a single original case within a batch vs. employing random variants) influence segmentation performance?

Q3. Effect of the number of training cases:

a. How does segmentation performance (IoU and Dice Score) change with an increasing number of training cases?

b. What is the effect of the number of training cases and augmentations on training time, especially in relation to Batch size?

Q4. How does tuning the Focal Weight Factor (within the range of 0.0 to 5.0, in 0.1 increments) affect metrics like the IoU and Dice Score?

Q5. Is it possible to detect signs of overfitting during the training process? What indicators suggest model saturation, and how can these be addressed?

1.3 Methodological Approach

The methodological approach of this work involves developing and evaluating a deep learning-based segmentation pipeline for glioblastoma, using three MRI sequences: T1-weighted post-contrast, FLAIR, and T2. These sequences were chosen for their ability to capture different aspects of glioblastoma tissue, essential for accurately segmenting the contrast-enhancing tumor, necrotic (non-contrast-enhancing) core, and surrounding edema. The pipeline is illustrated in Figure 1.1.

The model is trained using the RSNA-ASNR-MICCAI Brain Tumor Segmentation Challenge (BraTS) dataset [BGM+23], detailed in Section 5.1. This dataset includes multimodal MRI scans of glioblastoma patients with expert-labeled ground truth segmentations, enabling robust training with diverse data. Specifically, the BraTS dataset contains T1-weighted post-contrast, FLAIR, and T2 sequences, all critical for comprehensive tumor segmentation.

For evaluation, a separate set of 108 patients with newly diagnosed glioblastoma, treated at the University Hospital Salzburg between February 2009 and August 2022, is used. These patients met the inclusion criteria outlined in Section 5.1. This unseen dataset allows the model to be tested on real clinical data, providing insights into its generalizability beyond the training dataset.

Segmentation is performed using a 3D U-Net architecture [RFB15], which is particularly effective for capturing spatial information in volumetric images. Ground truth segmentations from an experienced neuroradiologist serve as the benchmark for evaluating model performance. The main metrics for assessing segmentation accuracy are the Intersection over Union (IoU) and a custom-developed Dice Score. This Dice Score is calculated using weighted Dice coefficients for the three tumor classes: non-contrast-enhancing, contrast-enhancing, and edema described in Section 4.1.6. Weighted coefficients reflect the varying clinical importance of each subregion, as outlined by Taha and Hanbury [TH15], who reviewed metrics for 3D medical image segmentation.

The strength of this approach lies in the weighted Dice Score, which focuses on clinically relevant tumor subregions, particularly those critical for surgical intervention. By assigning higher weights to the contrast-enhancing tumor, necrotic (non-contrast-enhancing) core, and edema in descending order, the model prioritizes accuracy in areas that directly affect treatment decisions. This ensures that segmentation emphasizes regions critical for improving surgical outcomes, such as contrast-enhancing tumor margins, while accounting for other subregions like necrosis and edema. Furthermore, the automated segmentation pipeline significantly reduces interrater variability, a common issue in manual segmentation. This consistency leads to more reliable treatment planning and predictions of outcomes. Final results are rigorously compared with expert-labeled ground truth segmentations, reviewed by an experienced neuroradiologist, ensuring the model's clinical relevance and robustness.

1.4 Requirements

This thesis outlines several key requirements for developing and evaluating a successful deep learning-based segmentation model for glioblastoma:

High-quality MRI data: The research relies on multimodal MRI data, specifically T1-weighted post-contrast, FLAIR, and T2 sequences, to fully capture the characteristics of glioblastoma.

Segmentation Accuracy: The model's accuracy is assessed using metrics like Intersection over Union (IoU) and a customized Dice Score, focusing on the main tumor classes (contrast-enhancing, necrotic (non-contrast-enhancing), and edema). Weighted Dice coefficients reflect the clinical importance of each tumor class.

Computational Resources: Given the large dataset and the complexity of the 3D U-Net architecture, high-performance computational resources are essential. A multi-GPU setup or cloud-based resources, as discussed in Section 6.10, are recommended to efficiently manage the training process and prevent computational bottlenecks.

Reproducibility: Ensuring consistency in results requires reproducibility across experiments. This is achieved by using a fixed random seed for data shuffling and augmentation, allowing the exact reproduction of training conditions in different runs.

Parameter Optimization: Hyperparameter tuning—such as adjusting Batch size, augmentation ratios, and the Focal Weight Factor—is essential for improving segmentation performance, particularly in clinically important tumor subregions. This optimization is crucial for ensuring that the model generalizes well to new data and avoids overfitting.

1.5 Contribution

In existing studies on reproducibility, such as those by Leventi-Peetz et al. [LPO22] and Chen et al. [CWS+22], the reproducibility of deep learning models is often achieved

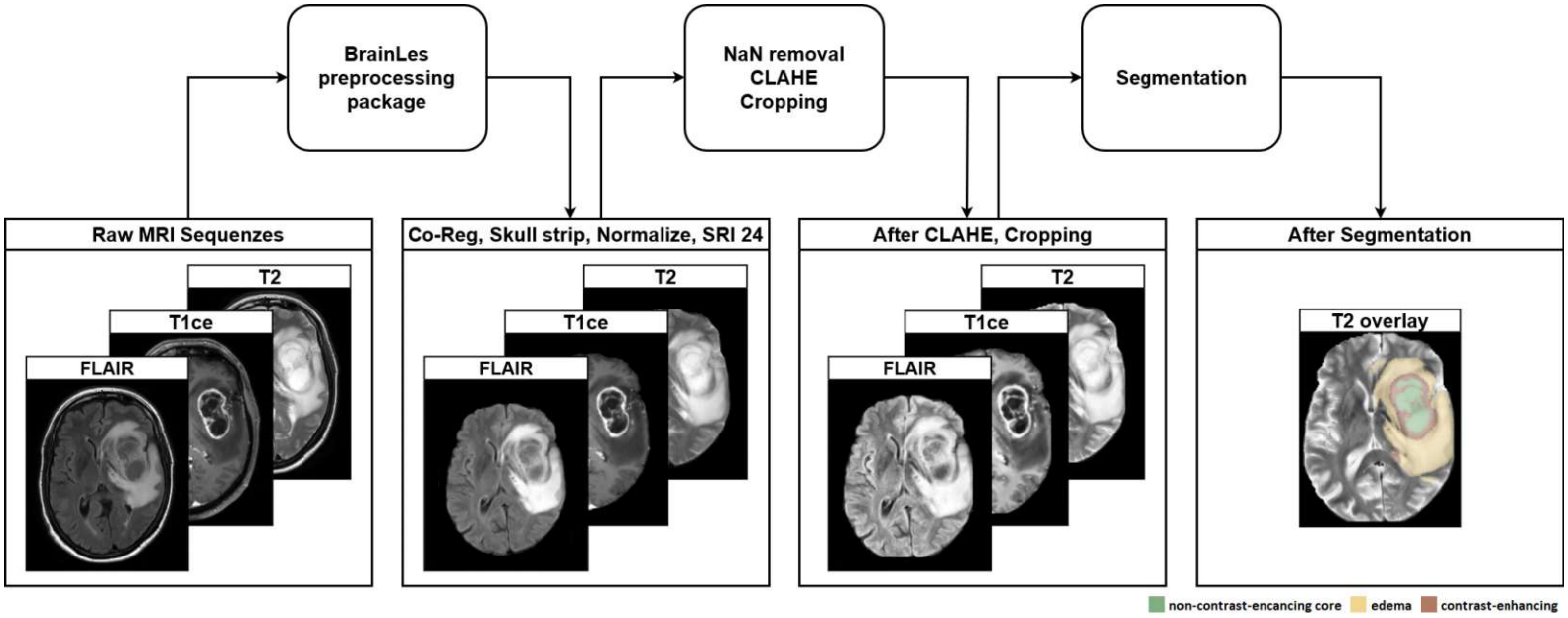


Figure 1.1: This diagram illustrates the glioblastoma segmentation pipeline. Raw MRI sequences (FLAIR, T1ce, and T2) are preprocessed using the BrainLes Preprocessing Package [KBW+20], including co-registration, skull stripping, normalization, and conversion to the SRI-24 [RZSP10] space. Next, NaN values are removed, and CLAHE [Zui94] (Contrast Limited Adaptive Histogram Equalization) is applied to enhance contrast. The images are then cropped to exclude non-relevant areas, focusing on the brain and tumor tissue. The trained 3D U-Net is applied to the unseen evaluation dataset, resulting in tumor segmentation, which is exemplarily overlaid on the T2 sequence. Segmentation labels showing necrotic/cystic (non-contrast-enhancing) core (green), edema (yellow) and contrast-enhancing tumor (brown).

by setting random seeds for weight initialization and data shuffling. These approaches primarily focus on minimizing random factors within the training pipeline by systematically standardizing environmental conditions and hardware configurations. Chen et al. [CWS+22], in particular, employ a systematic approach in which random software operations are managed through a record-and-replay system, while hardware-related non-determinisms are controlled. However, these methods are often either complex or challenging to integrate into existing systems.

The approach chosen in this study goes beyond these methods by not only setting seeds for consistent weight initialization and shuffling, but also managing the exact sequence in which training cases are processed across multiple runs. This is achieved through the use of a pre-defined seed list that enables the exact reproduction of the training case order. By ensuring that the model encounters the same training data in the same order, this approach provides finer control over learning steps. This precise management surpasses the methods commonly used in prior studies, significantly contributing to the reduction of variations and non-determinisms during training.

In existing literature, approaches to hyperparameter optimization are often focused on broad, generalizable methods across machine learning tasks. For example, Yang and Shami [YS20] discuss techniques such as grid search, random search, and Bayesian optimization for hyperparameter tuning. These methods are designed to explore a wide parameter space efficiently, providing high-level optimization for various machine learning models. However, they do not explore domain-specific configurations in depth, such as those required for complex medical image segmentation tasks. In contrast, the approach taken in this work emphasizes a more targeted, granular optimization of parameters. Specifically, by systematically tuning the Focal Weight Factor in small increments, this study achieves a level of control and sensitivity necessary for accurate segmentation of glioblastoma tumor regions—a critical area where general hyperparameter tuning methods may fall short.

Similarly, Wistuba et al. [WRP19] explore neural architecture search (NAS) using reinforcement learning and evolutionary algorithms, which are primarily intended for finding optimal architectures through structural experimentation. These methods address architecture-level adjustments and rely on broad parameter search strategies, without focusing on incremental dataset expansion or specific medical imaging configurations. The approach used in this study applies an incremental expansion strategy, where training datasets are gradually expanded from smaller to larger subsets (e.g., from 80 to 314 cases). This structured expansion allows for refined control of both model complexity and dataset size, optimizing performance progressively across stages rather than relying on general architecture search methods.

Another notable difference lies in reproducibility practices. While Yang and Shami [YS20] and Wistuba et al. [WRP19] emphasize reproducibility through deterministic processes, they do not address the exact loading order of training cases or augmentations across iterations. In contrast, this work achieves reproducibility not only by setting random seeds but also by preserving the precise sequence of training cases and augmentations.

This consistency ensures that the model encounters data in the same order across runs, enhancing reproducibility and minimizing variability—a critical factor in medical deep learning where reliable and replicable results are essential.

1.6 Structure of the Work

The diploma thesis is organized into eight chapters, as outlined in the following:

Chapter 1 ***Introduction*** outlines the motivation, problem definition, objectives of the thesis, and provides an overview of the methodology and key requirements.

Chapter 2 ***Clinical Background*** gives an overview of glioblastoma multiforme, focusing on its invasive nature and the imaging methods used for diagnosis and treatment planning.

Chapter 3 ***Tumor Segmentation on Medical Images*** describes how convolutional neural networks (CNNs), particularly U-Net, are applied for tumor segmentation. Relevant studies and methods are also discussed.

Chapter 4 ***Evaluation Metrics and Loss Functions*** explains the performance metrics such as Intersection over Union (IoU), Dice Score, and Hausdorff Distance, along with the Loss functions used during training.

Chapter 5 ***Methodology*** presents the design and training of the 3D U-Net model, covering the data sources, preprocessing steps, augmentation strategies, and the training process, including the development of the Custom Weighted Dice Score.

Chapter 6 ***Implementation*** details the preprocessing steps for training and evaluation datasets and the methods used to assess model performance on unseen data. Techniques such as Contrast Limited Adaptive Histogram Equalization (CLAHE) and data cropping are detailed.

Chapter 7 ***Results*** examines the segmentation performance of different models using metrics such as IoU, Dice Score, and Custom Weighted Dice Score. Additionally, it evaluates the models' generalization on the unseen dataset and highlights the top-performing results.

Chapter 8 ***Discussion, Outlook and Conclusion*** provides a summary of the findings, addresses the study's limitations, and gives recommendations for future work.

CHAPTER 2

Clinical Background

2.1 Glioblastoma Multiforme

Glioblastoma multiforme (GBM) is classified as a grade IV astrocytoma according to the World Health Organization (WHO) and is the most aggressive and malignant form of brain cancer. GBM has an incidence rate of approximately 3 per 100,000 adults annually [LPW+21]. This tumor entity is the most common and aggressive form of brain cancer in adults and presents significant clinical challenges due to its highly invasive nature and poor prognosis [MOK+21, OCW+21]. Originating from the white matter of the brain, GBMs diffusely infiltrate surrounding healthy brain tissue, rendering complete surgical removal of all tumor cells virtually impossible [DPSS19]. This infiltration not only disrupts normal brain function but also accelerates the decline in key neurological functions such as motor skills, speech, vision, and cognitive abilities [Org24].

Additionally, GBMs are notorious for rapid growth and resistance to standard therapies, including chemotherapy and radiotherapy. Microscopic tumor cells, which often remain undetected in imaging, increase the risk of recurrence following surgery. These deeply embedded cell nests further complicate treatment and contribute to the high recurrence rate even after maximal resection [SPPDSBO22].

2.2 Medical Imaging Methods

Several medical imaging techniques are used to diagnose and monitor glioblastoma. Each provides unique insights into the tumor's characteristics and growth patterns. These imaging modalities include:

2.2.1 Computed Tomography (CT)

Cranial Computed Tomography (cCT) is frequently used in emergency settings for the initial detection of a brain tumor and to quickly assess the presence of a mass effect. On CT images, GBMs can appear as irregular, hyperdense (in case of acute hemorrhage), or hypodense (reflecting necrosis or cystic components) areas [RT12], which fits the appearance shown in the CT scan in Figure 2.1, A. Often with surrounding edema, that appears hypodense in peritumoral regions. However, cCT lacks the specificity to differentiate glioblastoma from other types of brain tumors or non-tumor lesions based on imaging characteristics alone. Therefore, Magnetic Resonance Imaging (MRI) is the gold standard for the diagnosis and evaluation of glioblastoma due to its superior soft tissue contrast [Dhe14].

2.2.2 Magnetic Resonance Imaging (MRI)

Glioblastoma multiforme usually shows a typical garland-shaped marginal enhancement in the T1-weighted post-contrast sequence (Figure 2.1, C), with the surrounding edema spreading finger-like over a large area, which is best visualized in the FLAIR sequence (Figure 2.1, B). The contrast-enhancement alone does not provide information about the malignancy of the tumor itself, e.g., astrocytomas WHO III occasionally do not show a contrast-enhancement but are highly malignant [GKD+11].

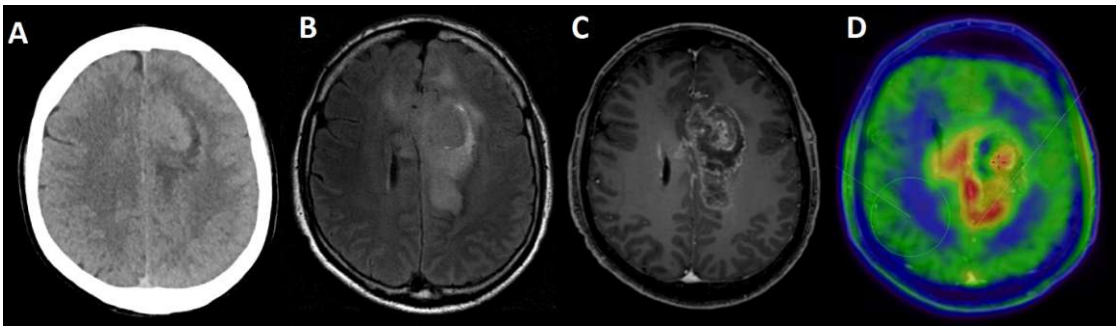


Figure 2.1: A: cCT showing tumor mass left frontal, B: Fluid Attenuated Inversion Recovery (FLAIR) magnetic resonance imaging (MRI) suppressing fluid signals to highlight edema and gliotic changes, C: Post-contrast T1-weighted sequence demonstrating a garland-shaped, contrast-enhancing lesion parasagittal left frontal, D: FET PET-CT with average SUV 1.88 in tumor mass (histologically confirmed glioblastoma multiforme, IDH-wildtype).

2.2.3 Diffusion-Weighted MRI

Magnetic Resonance Imaging (MRI), particularly when employing Diffusion Tensor Imaging (DTI) [BML94], is a sophisticated neuroimaging technique that provides detailed insights into the brain's microstructural environment at a molecular level. DTI provides a unique ability to map the diffusion of water molecules within brain tissue, which

reveals critical information about the brain's microarchitecture. This methodology allows for the visualization of neuronal fiber tracts, offering a detailed representation of the directional pathways and connectivity within the brain's white matter [EOPB+22]. DTI is instrumental in understanding and diagnosing a range of neurological conditions, as it provides an unparalleled view of the intricate network of neural pathways and their integrity or disruption in various disease states.

Diffusion Tensor Imaging (DTI) is a potent MRI modality capable of delineating neuronal structural abnormalities that result in loss of function and are undetectable through conventional MRI sequences [MBC+08]. In the context of glioblastoma, where edematous alterations can obscure critical details, DTI provides a significant advantage. It achieves this by mapping the diffusion patterns of water molecules in brain tissue, which are altered in the presence of disrupted neural architecture [SMAS13]. Direction-dependent diffusion can be well described with three-dimensional ellipsoids by introducing a tensor into the Stejskal and Tanner equation [ST65].

$$S = S_0 e^{-b\hat{g}^T D\hat{g}} \tag{2.1}$$

Where S is the signal intensity measured after the application of the diffusion gradient, and S_0 is the baseline signal intensity without any diffusion gradient. The parameter b represents the b-value, which quantifies the strength, duration, and temporal spacing of the diffusion gradients in MRI. In modern MRI scanners, b-values typically range from 0 to 3000 sec/mm² [BDEY01], with values around 1000 sec/mm² being common for diffusion-weighted imaging [KM04], and higher values, up to 1500 to 2000 sec/mm², used for specific clinical applications like stroke imaging to enhance contrast between normal and ischemic tissues [KM04, MBC+08]. The vector $\hat{g}$ denotes the direction of the applied diffusion gradient, while D is the diffusion tensor, a 3x3 matrix that characterizes the diffusion properties, specifically the rate and direction of water diffusion in the tissue [MBC+08]. The term $\hat{g}^T D\hat{g}$ represents the projection of the diffusion tensor along the direction of the gradient vector $\hat{g}$ [SMAS13].

The formula that characterizes the b-value is given by:

$$b = \gamma^2 G^2 \delta^2 \left(\Delta - \frac{\delta}{3}\right) \tag{2.2}$$

Where γ is the gyromagnetic ratio, a constant specific to the nucleus being imaged, G represents the amplitude of the diffusion gradient, which controls the strength of diffusion weighting, δ refers to the duration of the applied gradient pulses, and Δ is the time between the onset of the two diffusion gradients. These variables together influence the degree of diffusion weighting and thus the sensitivity of the MRI scan to water molecule movement within the tissue [MBC+08].

Fractional anisotropy FA is a measure of the tensor and can be calculated using its eigenvectors [BP96]:

$$FA = \frac{1}{\sqrt{2}}\sqrt{\frac{(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_1 - \lambda_3)^2}{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}} \tag{2.3}$$

Here λ_1, λ_2, and λ_3 are the eigenvalues of the diffusion tensor D and represent the magnitude of diffusion along the principal axes of the tensor [MBC+08]. The tensor is defined as:

$$D = \begin{pmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{pmatrix} \tag{2.4}$$

By following the primary diffusion direction, typically the eigenvector corresponding to λ_1, the structure of white matter pathways can be traced [MBC+08]. This method maps and visualizes the connectivity between different brain regions and can reveal changes in neural integrity, which may remain hidden in conventional imaging, especially in cases of glioblastomas [EOPB+22].

2.2.4 Dynamic Susceptibility Contrast (DSC) Perfusion MRI

Dynamic Susceptibility Contrast DSC Perfusion MRI refers to a magnetic resonance imaging technique used to evaluate blood flow through tissues and organs [BRK+90]. This method involves the rapid injection of a contrast agent, usually gadolinium-based, into a vein. As the contrast agent passes through the brain (or another organ being imaged), it causes a temporary decrease (susceptibility effect) in the signal intensity on T2*-weighted Perfusion MRI images. By measuring these changes in signal intensity over time, it is possible to generate various parameters related to blood flow, such as cerebral blood volume (CBV), cerebral blood flow (CBF), and mean transit time (MTT) [JLOC14, SOM+20]. DSC Perfusion MRI is particularly useful for brain tumors because it can differentiate between tumor types based on their blood flow, blood volume, and leakage characteristics [TMC+11]. For instance, higher-grade tumors such as glioblastomas typically show increased perfusion and relative cerebral blood volume (rCBV) compared to lower-grade tumors, due to their higher vascularity and angiogenic activity [SOM+20]. This information can aid in determining tumor grade, assessing tumor aggressiveness, and distinguishing between tumor recurrence and treatment-related changes such as radiation necrosis [WRM+14]. The radiation-induced necrosis tends to result in reduced blood flow [WRM+14].

2.2.5 Dynamic Contrast Enhanced (DCE) Perfusion MRI

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is a technique that involves the injection of a contrast agent to enhance the visualization of tissues

and blood vessels in Perfusion MRI scans [LWS+84]. Glioblastomas are known for their aggressive and highly vascular nature. They often exhibit areas of high perfusion and vascular permeability because of the presence of abnormal and leaky blood vessels, a phenomenon known as neoangiogenesis [FHP+21]. The parameter K^{trans} measures the transfer constant of the contrast agent from blood plasma into the extravascular extracellular space (EES). Contrast agent accumulates within the tissue extracellular space at a rate determined by perfusion, capillary permeability, and surface area [Tof97]. It is a reflection of both blood flow and the permeability of the tissue to the contrast agent. This parameter can be particularly useful in differentiating between glioblastomas and other types of tumors due to the unique vascular characteristics of glioblastomas [CB13]. K^{trans} values in glioblastomas tend to be higher compared to other tumors and normal brain tissue, reflecting these tumors' higher blood flow and vascular permeability [FHP+21].

2.2.6 Positron Emission Tomography (PET)

PET-CT, which stands for Positron Emission Tomography combined with Computed Tomography, is an advanced imaging technique that provides both functional and anatomical information. In the context of glioblastoma diagnosis, PET-CT can play a significant role by offering insights into the metabolic activity of the brain tissues, alongside detailed structural images. An increased metabolism is reflected in a higher uptake of certain radiotracers. Therefore, high-grade tumors like glioblastomas typically exhibit higher uptake levels compared to lower-grade gliomas, this has been observed specifically with 18F-FDG [AWS+16, SDW+19].

18F-Fluoroethyltyrosine (FET) is a radiolabeled amino acid, which is specifically used as a tracer in PET imaging for brain tumors originating from glial cells. It is taken up by active tumor cells due to their increased amino acid transport, making it a valuable tool for assessing the metabolic activity of brain tumors. FET has several advantages over the more traditional FDG (Fluorodeoxyglucose), especially in the brain, where FDG's effectiveness is limited by the high background glucose metabolism of normal brain tissue [HNP+13]. As shown in Figure 2.1, D, FET PET-CT provides better contrast between tumor and normal brain tissue, improving the ability to delineate tumor boundaries, assess tumor grade, and monitor treatment response or disease progression [HNP+13].

Despite its diagnostic advantages, PET-CT also presents several significant drawbacks, particularly related to the exposure to ionizing radiation from both the CT scan and the radioactive tracers used in PET imaging. This is especially concerning for patients requiring multiple scans over time for treatment monitoring, as cumulative radiation exposure can heighten the risk of radiation-induced damage or secondary malignancies [MKF+20]. Additionally, logistical challenges arise due to the production of radiotracers like 18F-Fluoroethyltyrosine (FET), which have a relatively short half-life (approximately 110 minutes [WLSC22]), requiring them to be produced in specialized facilities close to the imaging site. This not only limits the availability of PET imaging but also increases the complexity and cost of the procedure [TRX23, BGPB14].

2.2.7 Magnetic Resonance Spectroscopy

Magnetic Resonance Spectroscopy (MRS) is a non-invasive diagnostic tool that supplements traditional MRI by providing biochemical information about tissues. It measures the concentration of specific molecules, offering insights into the metabolic changes in brain tumors, including glioblastoma. For glioblastoma diagnosis, MRS can help by identifying unique metabolic patterns, such as elevated choline, reduced N-acetyl aspartate (NAA) (as shown in Figure 2.2), and the presence of lactate and lipids, which are indicative of high tumor cellularity, proliferation, and necrosis [WKSM21, PKF+11]. These metabolic signatures aid in distinguishing glioblastoma from other types of brain lesions, thereby enhancing diagnostic accuracy and treatment planning. Furthermore, the ratios of the metabolites mentioned, especially choline/NAA and choline/creatine, can be used to differentiate between low and high-grade astrocytomas [PKF+11].

The presence of a 2-Hydroxyglutarate (2-HG) peak at 2.25 ppm provides evidence of the presence of an IDH mutation [CGD+12]. According to the currently valid WHO classification from 2021, the presence of 2-HG is not compatible with glioblastoma multiforme, as this designation is reserved exclusively for the wildtype [LPW+21].

2.3 Histopathological Diagnosis (Biopsy & Resection)

The histopathological diagnosis, which is made by biopsy and resection, is considered the gold standard for the diagnosis of glioblastoma, as it reveals the characteristic pathological features of this aggressive brain tumor with unparalleled accuracy [LPW+21]. With this method, the tumor tissue can be examined directly under the microscope, allowing pathologists to observe the specific cellular abnormalities and mitotic activity that characterize glioblastoma. This level of detailed examination is critical for a definitive diagnosis, distinguishing glioblastoma from other types of brain tumors and guiding the selection of the most appropriate therapeutic strategies. Histopathological analysis not only provides insights into the tumor's grade and aggressiveness but also identifies molecular markers that can influence treatment decisions, such as O6-methylguanine-DNA methyltransferase (MGMT) promoter methylation status and isocitrate dehydrogenase (IDH) mutations, which have been linked to response to certain chemotherapies and targeted therapies [MVL+14].Therefore, the accuracy and depth of information provided by histopathological diagnosis are essential for optimizing patient outcomes through tailored therapy plans [LPW+21].

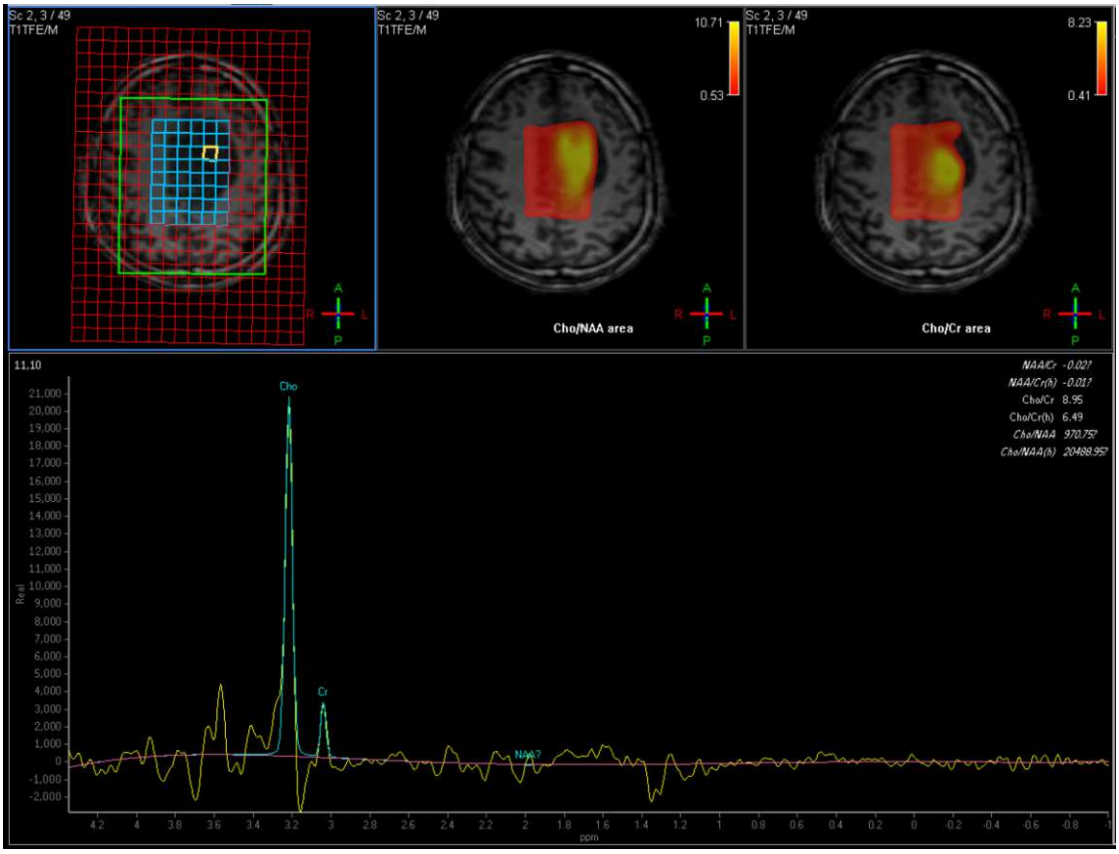


Figure 2.2: Multi-voxel spectroscopy showing an increased choline (Cho) peak, a significantly reduced creatine (Cr) peak, and a nearly non-existent N-acetyl aspartate (NAA) peak. Additionally, a distinct M-shaped lactate peak at 1.3 ppm is present, although it was not labeled, and there is no evidence of a lipid peak (same patient as in Figure 2.1).

CHAPTER 3

Tumor Segmentation on Medical Images

In tumor segmentation, a critical aspect of medical image processing, various machine learning, and deep learning approaches are employed to automatically identify and delineate tumors in imaging techniques such as MRI, CT, or PET. Below are some of the most common approaches.

3.1 Convolutional Neural Networks (CNNs)

CNNs are a class of deep learning models that are particularly well-suited for image recognition and image processing [KSH17]. In tumor segmentation, they are used to automatically extract features from images and learn to distinguish tumor tissue from healthy tissue.

3.2 U-Net

U-Net, a special architecture of a Convolutional Neural Network (CNN), was first introduced by Ronneberger et al. [RFB15], for image segmentation tasks. Its distinctive U-shaped structure, as shown in Figure 3.1, allows the combination of contextual information across different scales. The network uses a contracting path to capture context and a symmetric expanding path for precise localization [RFB15]. U-Net is designed to work efficiently with limited images, performing segmentation through convolutional operations, downsampling with pooling, and upsampling. This architecture allows U-Net to learn from small training datasets with extensive data augmentation, particularly elastic deformations, making it highly effective for biomedical image segmentation [CAL+16].

U-Net's design is highly effective for segmenting tumors like glioblastomas, which often have blurry edges and heterogeneous regions [DYL+17]. Its precise ability to distinguish between tumor and non-tumor tissue is crucial for the complex task of glioblastoma segmentation [DYL+17]. Typically, glioblastomas are imaged using MRI with sequences such as T1, T1 post-contrast, T2, and FLAIR, and U-Net's versatility allows for the processing of these multisequence inputs [MJB+15, DYL+17]. This flexibility makes U-Net well-suited for recognizing glioblastoma subtypes and different growth patterns, making it a valuable tool for complex medical imaging analysis [DYL+17].

3.2.1 U-Net Fundamentals

The U-Net architecture is shown in Figure 3.1, which details its contracting and expansive paths and demonstrates how each part contributes to the network's functionality. Understanding these components helps clarify how U-Net achieves accurate segmentation in biomedical images.

Contracting Path
The contracting path follows the standard convolutional network architecture. It involves repeated applications of convolutional layers (with ReLU activation) followed by max-pooling for downsampling, capturing high-level features and contextual information from the input image.

- **Convolutional Layers (Conv 3x3, ReLU):** Each blue arrow in the diagram represents a convolution operation with 3x3 filters followed by a ReLU activation. The numbers indicate the dimensions of the feature maps at each stage.
- **Max Pooling Layers (Max Pool 2x2):** The red arrows represent 2x2 max-pooling operations, which down-sample the feature maps, reducing their spatial dimensions while increasing the depth.

Bottleneck
At the U-Net's base, two convolutional layers with a high number of feature channels capture the most abstract features from the input image.

Expansive Path
The expansive path involves upsampling operations to increase the resolution of the output. Each upsampling step is followed by a convolutional layer that reduces the number of feature channels. Skip connections are also used to concatenate feature maps from the contracting path, allowing for precise localization.

- **Up-Convolutional Layers (Up-Conv 2x2):** The green arrows indicate up-convolution or transposed convolution operations, which up-sample the feature maps, increasing their spatial dimensions.

- **Concatenation (Copy and Crop)**: The grey arrows show where feature maps from the contracting path are concatenated with the up-sampled feature maps, allowing the network to combine low-level and high-level features.

- **Convolutional Layers (Conv 1x1):** The final layer in the expansive path uses a 1x1 convolution to map the feature vector to the desired number of classes (e.g., 2 for binary segmentation).

Output Segmentation Map
The output is a segmentation map of dimensions 388 x 388, representing the predicted segmentation of the input image [RFB15].

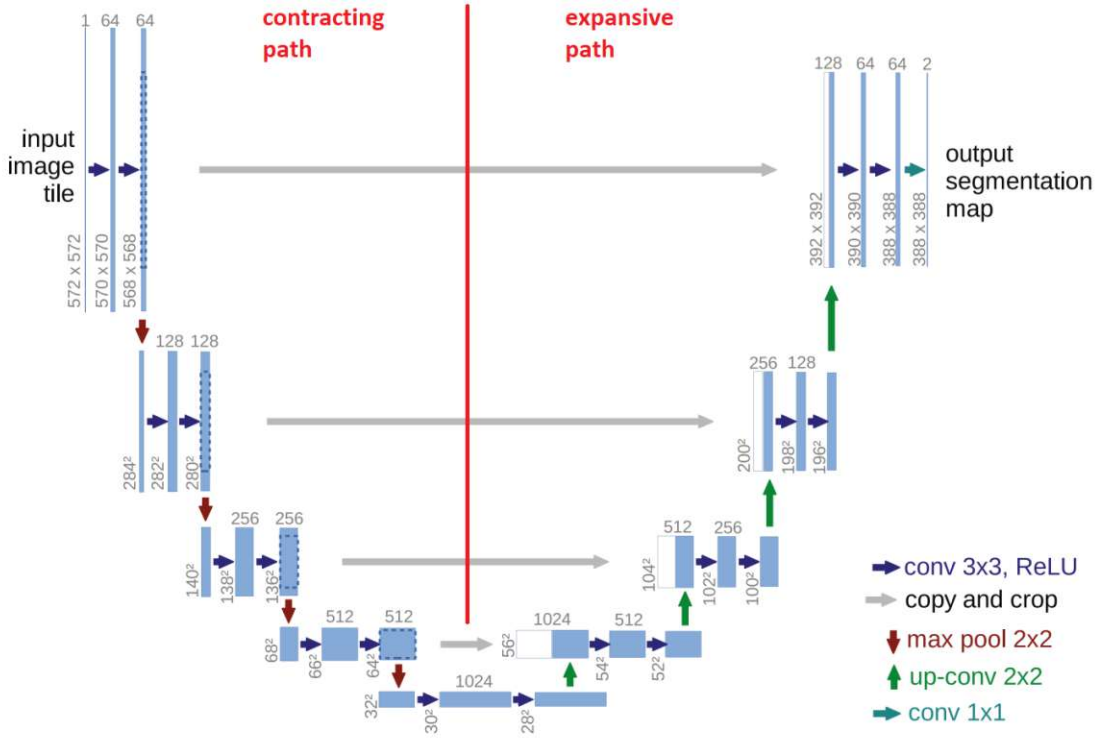


Figure 3.1: U-Net Architecture for Biomedical Image Segmentation This figure illustrates the U-Net architecture designed for biomedical image segmentation, highlighting its distinctive U-shaped structure that enables both precise localization and contextual understanding. The architecture consists of a contracting path (left side) and an expansive path (right side), with a vertical red line and labels added to enhance the visual separation and identification of these paths. Adapted from: [RFB15].

3.2.2 Key Components and Principles

Non-Linear Activation Functions

According to Goodfellow et al. [GBC16], non-linear activation functions are essential in neural networks as they introduce the non-linearity needed to capture complex relationships within data. Among these, the Rectified Linear Unit (ReLU) is particularly favored and is defined as:

$$f(x) = \max(0, x) \tag{3.1}$$

Where $f(x)$ is the output function, returning zero for any negative input x and retaining positive values as they are. ReLU offers distinct advantages over hyperbolic tangent and logistic sigmoid functions, particularly for medical image segmentation. Its non-saturating, linear structure helps avoid the vanishing gradient problem, which allows deeper networks to learn more effectively [NPN+21]. By setting negative values to zero, ReLU promotes sparsity in the network, reducing computational demands and enhancing efficiency [NPN+21]. This sparse activation also improves feature focus, helping to emphasize relevant structures in medical slice images and thereby contributing to more accurate segmentation outcomes [VCS24].

Convolutional Layers

Convolutional layers form the core of the U-Net architecture. They apply convolution operations to the input, followed by an activation function (typically ReLU), to extract features such as edges, textures, and patterns from the input image.

In the U-Net diagram in Figure 3.1, which is adapted from Ronneberger et al. [RFB15], each blue arrow represents a convolution operation. For example, the initial input image tile (572 x 572) undergoes multiple convolutions, reducing its dimensions while increasing the number of feature channels (e.g., from 1 to 64). This process continues as the image is downsampled in the contracting path. A convolution is defined as:

$$(f * g)(t) = \sum_{a=-\infty}^{\infty} f(a)g(t-a) \tag{3.2}$$

Where $f(a)$ represents the input image data, where a refers to the position of a specific pixel in the image. The filter or kernel applied during the convolution is denoted by $g(t-a)$. This filter g is shifted across the image, with t representing the position where the filter is currently centered. At each position t, the filter is applied to a local neighborhood of pixels in the input image. The summation runs over all possible values of a, which represent the pixel positions within this local neighborhood [DV16].

Overall, this convolution describes the process in which the filter g "slides" over the input image f, and at each position t, it calculates a weighted sum of the neighboring pixel values. This operation, described by Ian Goodfellow et al. [GBC16] in "Deep Learning", forms the core of convolutional neural networks (CNNs), which are widely used in

image processing. As explained by Goodfellow, convolution allows for the extraction of important features such as edges and patterns by applying small kernels across an image in this systematic manner.

Pooling Layers

Max-pooling layers are used in the contracting path to down-sample the feature maps as described by Ian Goodfellow et al. [GBC16]. This reduces their spatial dimensions while increasing the depth, helping to capture contextual information.

In Figure 3.1, red arrows indicate max-pooling operations. For example, after the initial convolutions, the feature map (568 x 568) undergoes max-pooling to reduce its size to 284 x 284, while increasing the depth to 128 channels. This process helps to retain the essential features while discarding irrelevant details.

For a max-pooling layer with a 2x2 filter:

$$y_{i,j} = \max(x_{2i,2j}, x_{2i,2j+1}, x_{2i+1,2j}, x_{2i+1,2j+1}) \tag{3.3}$$

This formula describes the max-pooling operation, where a 2x2 region of the input feature map, starting at position $(2i, 2j)$, is reduced to a single value. The function selects the maximum value from the four elements in the 2x2 window: $x_{2i,2j}, x_{2i,2j+1}, x_{2i+1,2j}, x_{2i+1,2j+1}$. The output value $y_{i,j}$ is the maximum of these four values, effectively downsampling the input while preserving the most prominent feature in each 2x2 region [DV16].

Max-pooling reduces the spatial dimensions of the input by a factor of 2 along both axes, decreasing computational load and introducing translation invariance. This invariance arises because the precise location of features within each 2x2 window becomes less significant [SMB10].

Up-sampling Layers

Up-sampling layers in the expansive path increase the resolution of the feature maps, a crucial step for pixel-wise classification in segmentation tasks. These operations are essential for restoring the spatial dimensions of the feature maps while retaining the learned representations.

In Figure 3.1, green arrows represent up-sampling operations. For instance, after reaching the bottleneck, the feature map (32 x 32) is up-sampled to higher resolutions, ultimately reaching 388 x 388 in the final layer. This up-sampling is accomplished through operations like transposed convolutions, which allow for precise reconstruction of spatial details.

For an up-convolution (transposed convolution):

$$y_{i,j} = \sum_{k,l} x_{i+k,j+l} w_{k,l} \tag{3.4}$$

Where $y_{i,j}$ represents the output value at position (i, j), $x_{i+k,j+l}$ denotes the input values within the receptive field centered at this position, and $w_{k,l}$ corresponds to the weights of the filter at position (k, l). The expression $\sum_{k,l}$ signifies a summation over all positions within the filter dimensions [DV16].

Skip Connections

Skip connections link corresponding layers of the contracting and expansive paths, enabling the network to use fine-grained features from earlier layers, which aids in precise localization [RFB15]. These connections play a critical role in mitigating the loss of spatial information during down-sampling operations in the contracting path, as they reintroduce high-resolution features into the expansive path.

The grey arrows in Figure 3.1 represent these skip connections. For example, the feature map from the contracting path (284 x 284) is concatenated with the up-sampled feature map of the same resolution in the expansive path. This preserves high-resolution features, improving segmentation accuracy.

Loss Function

Commonly used loss functions for medical segmentation tasks are detailed in Section 4.2.

3.3 3D U-Net

The 3D U-Net is an advanced extension of the original U-Net architecture, developed to process 3D volumetric data, making it highly suitable for medical imaging tasks such as MRI and CT scan analysis. Unlike the original U-Net, which operates with 2D convolutions, the 3D U-Net utilizes 3D convolutions, allowing it to capture spatial relationships across all three dimensions—depth, width, and height. This design is especially beneficial in the medical field, where accurate segmentation of complex structures, such as brain tumors like glioblastomas, is critical. By using 3D convolutions, the network captures complex spatial features, resulting in improved segmentation accuracy over 2D models [CAL+16].

In brain tumor segmentation, the 3D U-Net's ability to process entire volumes ensures that critical spatial information across slices is retained, making it particularly effective for tasks like glioblastoma segmentation, where the tumor's shape and boundaries are complex and extend through multiple MRI slices. This volumetric processing improves the network's capability to handle intricate anatomical structures, delivering more accurate results than traditional 2D U-Nets [CAL+16, ZZG22].

3.3.1 U-Net Model Development

Since its introduction in 2015 by Ronneberger et al. [RFB15], the U-Net architecture has undergone significant evolution, with each successive model addressing specific challenges

encountered in medical image segmentation. dResU-Net, first introduced by Raza et al. [RIBM+23], represents an early advancement addressing the vanishing gradient problem. By incorporating residual connections, this architecture enabled the development of deeper and more efficient networks.

Later, Attention U-Net [GSD24] improved upon this by introducing attention mechanisms, which enhanced the model's ability to focus on the most relevant image features. nnU-Net [IJK+21] further expanded the flexibility of the U-Net by automating the model's configuration to adapt to different datasets, reducing the need for manual intervention. Most recently, Swin UNETR [HNT+21] leveraged transformer-based architectures, capturing long-range dependencies in medical images and delivering improved performance in highly complex segmentation tasks.

3.3.2 State-of-the-Art Models

The development of advanced U-Net-based models has introduced various improvements to address specific challenges in medical image segmentation. The dResU-Net, developed by Raza et al. [RIBM+23], enhances the traditional U-Net by incorporating Residual Blocks from ResNet [HZRS16], effectively addressing the vanishing gradient problem. This adaptation allows for the training of deeper networks that can capture more intricate features, which is particularly valuable for segmenting detailed or irregular structures such as brain tumors by facilitating gradient propagation and improving convergence during training.

Building on this, the Attention U-Net, introduced by Oktay et al. [OSF+18], integrates attention mechanisms that enable the network to focus on relevant areas of the image, filtering out less important regions. This is especially useful in medical imaging, where small but critical structures, like tumors, may be surrounded by complex anatomy. Attention gates enhance important features, leading to more precise segmentations, particularly in challenging cases such as glioblastoma.

The nnU-Net, developed by Isensee et al. [IJK+21], introduced a fully self-adapting framework, eliminating the need for manual adjustments. It automatically configures preprocessing, architecture, and hyperparameters to fit the dataset at hand, adapting to the specific nuances of each dataset. This flexibility allows nnU-Net to perform well across diverse segmentation tasks, often surpassing more complex models.

Most recently, Swin UNETR, proposed by Hatamizadeh et al. [HNT+21], combines the U-Net structure with Swin Transformer blocks, leveraging the transformer's ability to capture long-range dependencies within images. This makes Swin UNETR particularly effective for segmenting large and detailed anatomical structures, providing a state-of-the-art solution in complex medical imaging tasks.

3.4 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) were first introduced by Ian Goodfellow and his colleagues in 2014 [GPAM+14], they brought a new approach to machine learning, particularly for generating synthetic data. The key innovation of GANs lies in their architecture, which includes two neural networks: the Generator (G) and the Discriminator (D). These networks are trained together through adversarial learning, where they compete to improve their performance. While U-Net and its variants focus on segmentation, GANs provide a complementary approach, especially for generating synthetic data and improving training through augmentation [FADK+18].

The Generator's goal is to create data that closely resembles real data by taking random noise or a latent vector as input and producing synthetic samples. Meanwhile, the Discriminator's task is to differentiate between real data from the training set and the synthetic data created by the Generator. It receives a sample and outputs a probability indicating whether the sample is real or generated [FT18]. The training process of GANs can be described as a minimax game, where the Generator tries to maximize the Discriminator's error rate, and the Discriminator aims to minimize its classification error [FT18].

This adversarial process can be formalized using the following objective functions:

For the Discriminator:

$$L_D = -\mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] - \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))] \tag{3.5}$$

For the Generator:

$$L_G = -\mathbb{E}_{z\sim p_z}[\log D(G(z))] \tag{3.6}$$

In these equations, L_D and L_G represent the loss functions for the Discriminator and Generator, respectively. The term p_{data} denotes the distribution of the real data, while p_z represents the distribution of the noise input fed into the Generator. For the Discriminator's loss L_D, $\mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)]$ calculates the expected log probability of correctly identifying real data samples, and $\mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]$ represents the expected log probability of correctly identifying generated samples as fake. For the Generator's loss L_G, $\mathbb{E}_{z\sim p_z}[\log D(G(z))]$ is the expected log probability that the Discriminator mistakenly identifies the generated samples as real, which the Generator tries to maximize to improve its output quality [FT18].

GANs have been widely applied across many fields, especially for generating realistic data. In medical imaging, they are particularly useful for augmenting training datasets, improving image resolution, and performing style transfer. These features help tackle challenges like data scarcity and variability in medical datasets [MMAC23, MGM+24].

Application of GANs in Brain Tumor Segmentation

In the BraTS 2023 Challenge [BBF+23], Ferreira et al. [FSL+24] used a GAN with a Swin UNETR (Swin Transformer U-Net) architecture [HNT+21] as the Generator and a Convolutional Neural Network (CNN) as the Discriminator to generate synthetic brain tumor data for augmenting the training dataset.

The Generator used the Swin UNETR framework [HNT+21], which incorporates Swin Transformer blocks to capture long-range dependencies in the input data. This architecture was selected for its efficiency in handling high-resolution images. The Generator received input patches of size 96×96×96, with four input channels and one output channel, and a feature size set to 48. These parameter values were part of the approach to balance computational efficiency and detailed spatial representation but were not discussed in depth regarding specific design choices by Ferreira et al. [FSL+24].

The Discriminator was a CNN with additional layers to enhance its ability to differentiate between real and synthetic data. It included an extra 3D convolutional layer at the end, configured with a stride of 1 and a kernel size of 3, without spectral normalization. A sigmoid activation function was applied before the output to produce a probability score indicating whether the input data was real. According to Ferreira et al. [FSL+24], this configuration aimed to improve the Discriminator's effectiveness but was not further elaborated on in terms of the specific parameter values chosen.

Following the methodology of Ferreira et al. [FSL+24], the training of the GAN was conducted in two distinct stages to optimize both the realism of the synthetic tumors and the surrounding brain tissue. In the initial phase, the GAN underwent 200,000 training iterations, with the adversarial loss weight (λ_1) set to 1 and the mean absolute error (MAE) loss weight (λ_2) set to 5. These specific values were chosen to prioritize the generation of realistic tumor structures during this phase, balancing the adversarial and MAE losses to enhance the distinctiveness of tumor regions while maintaining computational stability.

In the subsequent refinement phase, the training process shifted to improve the visual quality of the surrounding tissue, ensuring a cohesive appearance between tumors and adjacent brain structures. To achieve this, the weight of the MAE loss was gradually increased, while the adversarial loss weight was proportionally reduced. This adjustment allowed the model to focus more on refining the non-tumor regions, creating a more seamless integration of synthetic tumors within realistic brain tissue. Through this two-stage approach, as outlined by Ferreira et al. [FSL+24], the GAN achieved high-quality, realistic synthetic images, effectively enhancing the dataset for training purposes.

The input data for the Generator, following the approach by Ferreira et al. [FSL+24], was preprocessed by carefully cropping around the tumor region, normalizing voxel values, and introducing Gaussian noise to simulate realistic variations. This preprocessing approach was designed to focus the Generator's learning on the tumor and its immediate surroundings, ensuring that the generated samples retained essential anatomical and pathological features seen in real brain tumors. The cropping around the tumor area

allowed the model to concentrate computational resources on relevant structures, while normalization standardized voxel intensity values, making it easier for the model to learn consistent patterns. Adding Gaussian noise introduced variability in the data, which simulated natural variations in tumor appearance, thereby enhancing the robustness and generalizability of the synthetic samples. This preprocessing was crucial to ensure that the generated data closely mimicked real brain tumors in both structure and texture.

The use of GANs), especially with advanced architectures like Swin UNETR [HNT+21], has proven highly effective in medical imaging tasks, such as brain tumor segmentation. GANs generate synthetic medical images that closely replicate real MRI data, offering a solution to the problem of limited data, which is common in medical imaging. This synthetic data can improve the segmentation quality of machine learning models by providing diverse training examples, thereby enhancing model robustness and accuracy [AAHK24].

3.5 Transfer Learning & Fine-Tuning

Transfer Learning, as defined in *A Comprehensive Survey on Transfer Learning* by Zhuang et al. [ZQD+20], "aims at improving the performance of target learners on target domains by transferring the knowledge contained in different but related source domains". It has become a key method in medical imaging to enhance model performance, especially in scenarios with limited annotated data [YBL+24, ASM+21, TWZ+22]. It allows models to leverage knowledge from pre-trained networks on extensive datasets, retaining generalizable features that are useful for new tasks with smaller, specialized datasets [YBL+24, ASM+21]. This approach can significantly reduce training time and improve model convergence, as shown by its successful application in various medical imaging tasks, where large, labeled datasets are often scarce [YBL+24]. Fine-tuning these pre-trained models provides a performance boost by adapting to task-specific features, enhancing the accuracy and efficiency of segmentation and classification in medical contexts [ASM+21, TWZ+22].

While Transfer Learning offers potential advantages, it is important to acknowledge specific drawbacks as well. A model trained without Transfer Learning offers methodological advantages by allowing precise control over parameterization and architecture. This is particularly important if specific adaptations are required to optimize the model for glioblastoma segmentation tasks [RZKB19]. A model trained directly on the target data can be finely tuned, providing more valid insights into the performance and robustness of various hyperparameters [RZKB19, KEB+21]. Additionally, the quality and specificity of data used for pre-training in Transfer Learning may differ significantly from MRI data specific to glioblastomas. Pre-trained models are often based on general imaging datasets or other medical modalities and may not capture the intricate details found in glioblastoma-specific imaging sequences, such as T1- or T2-weighted MRI, which can impair transferability and accuracy in segmenting these complex structures [RZKB19, KEB+21].

3.6 Radiomic Feature Extraction

The main distinction between 3D U-Net and radiomic feature extraction lies in their focus. While 3D U-Net models are designed for segmenting medical images, radiomic analysis provides valuable insights into tumor characteristics—such as intensity, shape, and texture—which can be used to predict clinical outcomes like survival [SKA+20, HHAH+23]. Precise segmentation of tumors can thus provide the basis for further analysis, including radiomics [KPG+24]. Several studies have demonstrated the effectiveness of radiomic features in predicting survival in GBM patients [SKA+20, HHAH+23, KPG+24].

Overview of Radiomic Feature Extraction

Chaddad et al. [CDT16] used texture features from multi-contrast MRI, including T1-weighted post-contrast and FLAIR images, to predict survival times. Their study found that features like Energy, Correlation, and Variance from contrast-enhanced regions were significant survival predictors.

Baid et al. [BRT+20] combined intensity, volume, shape, and texture features from FLAIR and T1ce MRI data. They used a stationary wavelet transform to capture directional information, which improved overall survival (OS) prediction. Their approach achieved strong results in the BraTS 2018 challenge.

Suter et al. [SKA+20] identified robust radiomic features from pre-operative MRI to classify survival in GBM patients. They performed over 16 million perturbation tests to simulate multi-center data variability, emphasizing the importance of robust feature selection when applying models across different datasets.

Kaur et al. [KRA23] improved survival predictions in GBM patients using machine learning models built on radiomic features from segmented MRI scans, focusing on texture, shape, and intensity features.

Key differences between radiomic feature extraction and 3D U-Net segmentation

Radiomic feature extraction focuses on deriving features for prediction models, whereas the 3D U-Net is an end-to-end deep learning model specifically designed for segmenting medical images [ZGJ23]. Radiomic approaches often require manual or semi-automated segmentation to define regions of interest (ROIs) from which features are extracted. This step can introduce variability and potential biases, depending on the accuracy of the initial segmentation [WZZ+24]. The primary purpose of radiomic feature extraction is to predict clinical outcomes, such as survival, treatment response, and recurrence. In contrast, the 3D U-Net is designed to provide precise and automated segmentation, which can be used as input for further analysis or treatment planning [WZZ+24, ZGJ23].

An alternative approach to segmentation within the context of radiomics

The software Brain Tumor Image Analysis (BraTumIA) is based on machine learning techniques and specifically developed for the automatic segmentation of glioblastomas

[OKA+19, MKL+16]. BraTumIA utilizes a combination of classical image processing algorithms and predefined radiomic features that are manually selected and crafted based on prior knowledge of tumor characteristics in MRI images [IOK+16, PBP+14]. These features include specific metrics such as intensity distributions, texture patterns, and shape descriptors that are chosen to represent the different tumor compartments accurately [KKR+18, PHM+16]. Once these features are defined, BraTumIA applies machine learning algorithms to analyze these engineered features across the tumor regions and classify different tissue types (e.g., necrotic (non-contrast-enhancing) tissue, edema, and contrast-enhancing tumor) [DMFAMJRV19, PHM+16]. This method contrasts with the 3D U-Net approach, where the model learns relevant features directly from ground truth annotations during training, without requiring predefined feature selection [IOK+16, MKL+16]. In BraTumIA, however, the reliance on predefined features offers the advantage of interpretability and control, as each feature's contribution to the segmentation can be assessed, which can be valuable in clinical contexts focused on radiomic analysis [OKA+19, PBP+14].

CHAPTER 4

Evaluation Metrics and Loss Functions

4.1 Evaluation Metrics

A comprehensive assessment of medical image segmentation models requires multiple metrics to cover all aspects of performance. Each metric provides a unique perspective on segmentation accuracy and reliability. Using metrics such as the Dice Similarity Coefficient (DSC) [Sø48] and Intersection over Union (IoU) [Jac01] offers a detailed view of model performance. DSC measures the general agreement of segmentation volumes, while IoU provides insight into the overlap relative to the union of predicted and ground truth segments. Cross entropy and accuracy are essential for understanding probability distributions and overall classification performance. Although the 95% Hausdorff Distance (HD95) [VNL20] is not used in this work, it is mentioned for completeness, as it is commonly used to assess boundary accuracy in segmentation tasks. In summary, the combination of different metrics provides a comprehensive evaluation of segmentation models, ensuring both overall agreement and detailed accuracy are considered.

4.1.1 Overlap Metrics

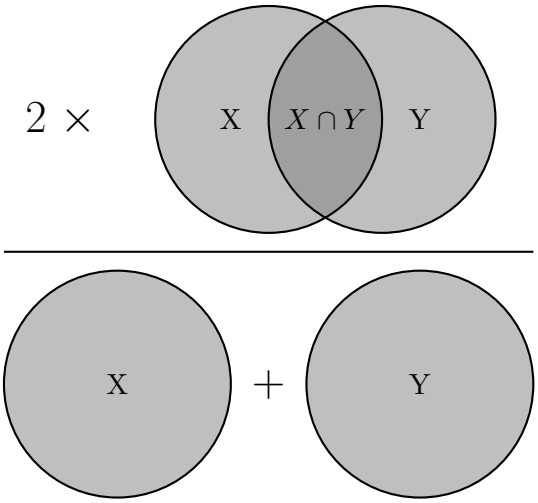
Overlap-based metrics focus on the degree to which the predicted and ground truth segmentations overlap [MSRK22]. These metrics are widely used in segmentation tasks as they provide a clear measure of similarity between two segmentations, with the Dice Similarity Coefficient (DSC) [Sø48] and Intersection over Union (IoU) [Jac01] being among the most commonly applied.

Dice Similarity Coefficient

The Dice Similarity Coefficient (DSC) [Sø48], also known as the Dice coefficient, is a statistical measure of similarity between two datasets. In tumor segmentation, it evaluates the overlap between two segmentation results. Commonly used in medical image processing, the Dice coefficient assesses the accuracy of segmentation models by comparing automatically generated results with manually created segmentations, which are considered the ground truth [Sø48].

The Dice coefficient is calculated as follows:

$$DSC = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{4.1}$$



This relationship is visually represented in the accompanying illustration, where X is the set of pixels or voxels belonging to the first segmentation (for example, the prediction of a segmentation model), and Y is the set of pixels or voxels belonging to the second segmentation (for example, the ground truth). $|X \cap Y|$ is the number of pixels or voxels that match in both segmentations, while $|X|$ and $|Y|$ are the counts of pixels or voxels in the respective segmentations.

The Dice coefficient ranges from 0 to 1, with 1 representing a perfect match between two segmentations and 0 indicating no overlap. A higher Dice coefficient reflects greater segmentation accuracy, meaning the model's output closely aligns with the ground truth. This metric is particularly useful for evaluating segmentation models as it balances both sensitivity and specificity, offering a comprehensive measure of accuracy [MRV03].

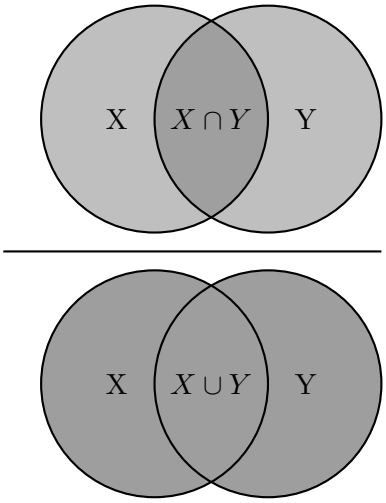
Intersection over Union (IoU)

The Intersection over Union (IoU) [Jac01], also called the Jaccard Index or Jaccard similarity coefficient, is a key metric for image segmentation tasks. It measures the

overlap between the predicted segmentation and the ground truth, defined as the ratio of the intersection of the predicted and actual segmentation masks to their union.

IoU is defined as:

$$\text{IoU} = \frac{|X \cap Y|}{|X \cup Y|} \tag{4.2}$$



This relationship is visually represented in the accompanying illustration, where X is the predicted segmentation, Y is the ground truth segmentation, $|X \cap Y|$ is the area of overlap between the two segmentations, and $|X \cup Y|$ is the total area covered by both the prediction and the ground truth.

IoU Scores range from 0 to 1, with 1 indicating perfect alignment between predicted segmentation and ground truth, and 0 indicating no overlap.

This metric is widely used in medical imaging segmentation tasks due to its clear and interpretable results. It provides a direct measure of segmentation accuracy by penalizing both false positives (FP)—where the model predicts a tumor that doesn't exist—and false negatives (FN)—where the model misses parts of the tumor. This makes IoU a robust choice for evaluating models where precise anatomical delineation is crucial, such as in tumor segmentation and organ boundary detection [MSRK22].

One key advantage of IoU is its ability to reduce both over-segmentation and under-segmentation, offering a more stringent evaluation compared to metrics like the Dice Similarity Coefficient (DSC). This is particularly valuable when assessing smaller tumor regions where precision is critical, as false positives or negatives can have significant clinical consequences [MSRK22].

Since IoU penalizes incorrect predictions more heavily, it is especially useful for fine-tuning model performance in challenging tumor regions. For example, if a model overestimates

the size of edema or misses parts of the necrotic (non-contrast-enhancing) core, IoU provides a more accurate measure of error than metrics like DSC, which may be more tolerant in cases of class imbalance [MSRK22].

By using IoU alongside other metrics like the Dice Similarity Coefficient, researchers and clinicians can achieve a more comprehensive evaluation of the model's performance. This combination ensures accurate and reliable predictions across various types of medical images and conditions [HJHK19].

Combined Dice-IoU Metric

In this work, Dice and IoU are combined to better address the small and heterogeneous tumor regions. While Dice offers a balanced perspective, IoU is particularly useful for identifying cases where false positives heavily affect segmentation. The model's performance will be evaluated by using both metrics.

4.1.2 Boundary Metrics

Boundary accuracy is critical, especially if precise delineation of tumor margins is required for treatment planning.

Hausdorff Distance

The Hausdorff Distance is a boundary-based metric that measures the greatest distance between predicted and ground truth segmentations. It is especially useful for assessing boundary accuracy, where small errors can have significant clinical consequences [HKR93]. The formula is defined as:

$$HD(X,Y) = \max\left\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y)\right\} \tag{4.3}$$

Where $d(x,y)$ is the Euclidean distance between points x and y in the predicted and ground truth segmentations. The infimum (inf) refers to the greatest lower bound of a set. In this context, for each point $x \in X$, it represents the smallest distance to any point $y \in Y$, providing the minimum distance from that point x to the other set. The supremum (sup), or least upper bound, takes the largest of these minimal distances, effectively selecting the worst-case scenario.

This metric is sensitive to outliers and evaluates extreme cases where the model might misinterpret tumor boundaries. In practical terms, the minimum distance from each point in the first segmentation (e.g., the algorithm's output) to any point in the second segmentation (the ground truth) is calculated. Then, the same calculation is performed in reverse, measuring the minimum distance from each point in the ground truth to the automated segmentation. The Hausdorff Distance is the maximum of these minimum distances, capturing the greatest possible discrepancy between the two segmentations [HKR93].

A low Hausdorff Distance indicates that the two segmentations are highly similar, not only in shape and volume but also in the precise location of their boundaries. In contrast, a high Hausdorff Distance suggests significant discrepancies, highlighting areas where the algorithm may have failed to accurately capture the tumor boundary [KS20].

Due to its sensitivity to maximum error, the Hausdorff Distance is often used alongside other metrics, such as the Dice Similarity Coefficient, to provide a more comprehensive evaluation of segmentation quality. It offers insights into spatial accuracy that other metrics may overlook [MSRK22].

95% Hausdorff Distance (HD95)

The 95% Hausdorff Distance (95th percentile of the Hausdorff Distance) is an important metric in evaluating segmentation models, especially in medical imaging [ATH+21]. While the Hausdorff Distance measures how close two subsets of a metric space are, HD95 represents the maximum distance between two sets after excluding the most extreme 5% of values. This makes HD95 a more robust metric, less sensitive to outliers than the standard Hausdorff Distance [CRF23].
The HD95 is defined as:

$$HD_{95}(X,Y) = \max\left\{\text{quantile}_{95}\min_{y\in Y} d(x,y),\ \text{quantile}_{95}\min_{x\in X} d(y,x)\right\} \tag{4.4}$$

Where X and Y are the sets of points in the predicted segmentation and ground truth images, respectively. $d(x,y)$ is the Euclidean distance between a point $x \in X$ and a point $y \in Y$. The term quantile_{95} refers to the 95th percentile of the minimum distances, which is used to reduce the influence of outliers. By focusing on the 95th percentile, this measure captures the boundary mismatch between the predicted and actual segmentations without being overly affected by extreme deviations.

Medical datasets often include noise and outliers, which can skew evaluations when using the standard Hausdorff Distance. HD95 strikes a balance by penalizing boundary errors while ignoring extreme outliers that may not be clinically relevant. This helps capture the worst-case error while remaining robust to outliers. Its robustness makes HD95 ideal for evaluating segmentation models in detecting anatomical structures, where accurate and reliable boundary delineation is crucial, such as in tumor and organ segmentation [PCP+22b].

4.1.3 Probability Metrics

Cross Entropy

Cross entropy is a widely used loss function in deep learning, especially for classification and segmentation tasks in medical imaging. It measures the difference between two probability distributions: the true distribution of the labels and the predicted distribution [RZA22]. The cross entropy loss is defined as:

$$L_{CE} = -\sum_{i} Y_i \log(p_i) \tag{4.5}$$

Where Y_i represents the true label (ground truth) for pixel i, and p_i represents the predicted probability for pixel i belonging to the target class. In the context of medical image segmentation, cross entropy is used to evaluate how well the predicted segmentation probabilities match the ground truth labels.

Cross entropy is effective in handling multi-class problems and is often used due to its simplicity and efficiency. It helps in training models to output probabilities that are as close as possible to the true distribution, thus improving the accuracy of segmentation. However, it may not always handle class imbalances well, which is common in medical imaging where the regions of interest (e.g., tumors) are often much smaller than the background [WC18].

Binary Cross-Entropy

The Binary Cross-Entropy (or Log Loss) is a loss function commonly used in binary classification tasks, where the goal is to classify inputs into one of two classes, such as distinguishing between tumor and non-tumor tissue. The function measures how well the predicted probabilities for each class match the actual labels.

The binary cross-entropy loss function is expressed as:

$$H(p, q) = -(y \log(p) + (1 - y) \log(1 - p)) \tag{4.6}$$

Where y represents the true label (0 or 1), and p represents the predicted probability for the positive class. Thus, binary cross-entropy is not applied here due to the multi-class nature of the problem, where a more fitting alternative, such as categorical cross-entropy, should be used.

4.1.4 Pixel Metrics

Accuracy

Accuracy is a straightforward metric used to evaluate the performance of segmentation models. It measures the proportion of correctly classified pixels over the total number of pixels. Mathematically, it is the ratio of correctly predicted pixels to the total number of pixels. In the context of image segmentation, accuracy can be misleading if the dataset is imbalanced [WWZ20, TH15]. For instance, in medical images where the background vastly outnumbers the region of interest, a model that predicts mostly background can achieve high accuracy despite poor performance in identifying the actual region of interest [MSRK22].

Sensitivity (True Positive Rate)

Sensitivity, or the True Positive Rate (TPR), measures the proportion of actual positives that are correctly identified by the model. In medical imaging, this is crucial as it reflects the model's ability to detect diseased or abnormal regions, such as tumors [TH15].

The formula for Sensitivity is:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4.7}$$

Where TP (True Positives) are the correctly predicted positive instances, and FN (False Negatives) are the positive instances that the model failed to predict. A high sensitivity means the model is good at identifying positive cases (e.g., detecting tumors), which is essential in healthcare to minimize missed diagnoses [Yer47].

Specificity (True Negative Rate)

Specificity, or the True Negative Rate (TNR), measures the proportion of actual negatives that are correctly identified. It is equally important to ensure that healthy regions or normal anatomy are not misclassified as abnormal [TH15].

The formula for Specificity is:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4.8}$$

Where TN (True Negatives) are the correctly predicted negative instances, and FP (False Positives) are the negative instances that were incorrectly predicted as positive. Specificity is crucial in medical imaging, especially in cases where false positives might lead to unnecessary further testing or treatments [Yer47].

4.1.5 Generalized Dice Score

The Generalized Dice Score (GDS) is an extension of the Dice Similarity Coefficient (DSC) [Sø48], designed to handle multi-class segmentation tasks by incorporating class-specific weights [SLV+17]. This score is particularly useful in medical image segmentation, where imbalanced class distributions are common [SLV+17]. By weighting each class according to its relevance or occurrence in the dataset, the GDS can provide a more balanced evaluation of segmentation performance across all classes [SLV+17].

The Generalized Dice Score is defined as:

$$\text{GDS} = \frac{2 \sum_{c=1}^{C} w_c \cdot |X_c \cap Y_c|}{\sum_{c=1}^{C} w_c \cdot (|X_c| + |Y_c|)} \tag{4.9}$$

In this formula, C represents the total number of classes, and X_c and Y_c are the sets of pixels for each class c in the predicted segmentation and the ground truth segmentation,

respectively. The term $|X_c \cap Y_c|$ indicates the overlapping pixels between the predicted and ground truth segmentations for a given class c, while $|X_c|$ and $|Y_c|$ denote the total number of pixels for class c in the prediction and ground truth, respectively. The weight w_c is assigned to each class c and is often calculated as the inverse of the class frequency, giving more significance to underrepresented classes. This weighting approach enables the GDS to provide a balanced evaluation across all classes, which is especially useful in imbalanced datasets commonly found in medical image segmentation.

The weighting factor w_c allows the GDS to balance contributions from each class, particularly in datasets where some classes are underrepresented [SLV+17]. This makes the GDS a robust choice for evaluating segmentation performance across diverse and imbalanced classes, such as those often found in medical imaging applications.

4.1.6 Custom Weighted Dice Score

The Custom Weighted Dice Score is a tailored version of the Dice Similarity Coefficient (DSC), designed to emphasize the most clinically significant regions in glioma segmentation, namely the contrast-enhancing tumor, necrotic (non-contrast-enhancing) core, and surrounding edema. Although inspired by the Generalized Dice Score (GDS) [SLV+17], which also uses weighted contributions for different classes, the Custom Weighted Dice Score differs in key ways to serve specific clinical priorities.

Unlike the GDS, which typically derives class weights based on class frequency (giving less frequent classes more weight), the Custom Weighted Dice Score assigns weights $w_1, w_2,$ and w_3 based on the clinical importance of each tumor region rather than its prevalence. The formula for this score is:

$$\text{Custom Dice Score} = w_1 \times \text{Dice}_{\text{contrast-enhancing}} + w_2 \times \text{Dice}_{\text{necrotic}} + w_3 \times \text{Dice}_{\text{edema}} \tag{4.10}$$

Where $w_1, w_2,$ and w_3 are weights summing to 1, ensuring a balanced contribution of each clinically relevant region. Here, $\text{Dice}_{\text{contrast-enhancing}}, \text{Dice}_{\text{necrotic}}, \text{Dice}_{\text{edema}}$ represent the Dice scores for each tumor class.

This approach allows the model to prioritize tumor regions with the greatest impact on treatment planning and prognosis, enhancing its clinical applicability. However, unlike the GDS, which provides a more standardized weighting by class frequency, the Custom Weighted Dice Score introduces subjectivity into weight selection, which may lead to inconsistencies across studies or clinical applications. Furthermore, by heavily weighting certain regions, it risks reducing accuracy in lower-weighted areas, potentially overlooking comprehensive information necessary for treatment. This highlights the importance of carefully balancing weights to ensure that the metric remains clinically relevant without sacrificing a holistic view of tumor segmentation.

4.1.7 Metric Comparison

In conclusion, the use of multiple evaluation metrics provides a comprehensive assessment of glioblastoma segmentation models. The Dice Similarity Coefficient (DSC) is advantageous for its ability to measure general overlap, making it effective for evaluating the overall segmentation volume [MSRK22]. However, it can be less sensitive to boundary details. Intersection over Union (IoU) offers a similar advantage with the added benefit of penalizing discrepancies in the overlap more harshly, but it also shares the sensitivity issues of DSC [BEB+19].

The Cross-Entropy loss is particularly useful for its probabilistic interpretation, as it measures the difference between predicted and true probability distributions, helping to fine-tune model predictions during training. However, it does not fully capture the spatial accuracy of segmentations, as it focuses primarily on pixel-level classification [Jad20]. Accuracy, while straightforward and easy to interpret, often fails to provide meaningful insights in the context of imbalanced datasets. This is because it may overemphasize the correct classification of dominant classes, such as the background in medical images, at the expense of smaller, critical regions like tumor compartments [Jad20].

The greatest advantage of the 95% Hausdorff Distance (HD95) lies in its sensitivity to significant boundary deviations, making it particularly effective for identifying large errors at segmentation edges [CRF23, KS20]. This focus on extreme discrepancies is valuable in applications where minimizing critical errors is essential, such as multimodal medical image registration or surgical planning. However, HD95's sensitivity to isolated large deviations can also be a disadvantage. Such deviations, often caused by noise or artifacts, may disproportionately impact the metric, leading to an overemphasis on localized errors and providing a less representative view of overall segmentation accuracy [KS20].

By leveraging these metrics together, a comprehensive evaluation can be achieved that accounts for both volume overlap and boundary precision, ensuring robust glioblastoma segmentation models suitable for clinical applications. This multi-metric approach is particularly important for capturing the complex and heterogeneous characteristics of glioblastomas, ultimately enhancing diagnostic accuracy and improving treatment planning [RB24].

4.2 Loss Functions

The use of Combined Loss Functions, such as Dice Loss paired with cross entropy loss, has proven to be highly effective in medical image segmentation, particularly for complex tasks like glioblastoma segmentation, where challenges such as class imbalance and precise boundary delineation are critical [RKF+22, YSSR22]. By combining complementary loss functions, such as Dice and Focal Loss (as defined in Section 4.2.2), it is possible to address multiple challenges simultaneously, including accurate segmentation and the management of class imbalance inherent in medical imaging datasets [WC18, LGG+20]. This approach leverages the strengths of different loss functions, such as Dice Loss's

ability to measure overlap and Focal Loss's capability to address class imbalance, thereby enhancing the model's overall performance [YSSR22, LGG+20].

In glioblastoma segmentation, several Combined Loss Functions have been widely adopted. A common combination is Dice Loss paired with Cross-Entropy Loss, which balances global segmentation accuracy with pixel-wise prediction reliability [RKF+22, YSSR22]. Tversky Loss is particularly effective for fine-tuning sensitivity and precision in underrepresented tumor regions, and its potential combination with Cross-Entropy Loss could provide additional flexibility in balancing accuracy and robustness [YSSR22]. Another approach involves combining Dice Loss with boundary-aware losses, such as Hausdorff Distance Loss, to enhance accuracy along tumor margins, a critical requirement for clinical applications [RKF+22, YSSR22].

Among these combinations, the integration of Dice Loss and Focal Loss offers distinct advantages. Dice Loss ensures high overlap between predicted and actual segmentations, making it ideal for capturing larger tumor regions, while Focal Loss complements this by addressing class imbalances and focusing on smaller, harder-to-segment areas [YSSR22, LGG+20]. Together, these functions enable precise segmentation of clinically important regions, such as the contrast-enhancing tumor, while maintaining robust overall performance. This combination is particularly well-suited for glioblastoma segmentation, where accurate delineation of tumor regions is crucial for diagnosis and treatment [RKF+22, YSSR22, LGG+20].

4.2.1 Dice Loss

Dice Loss focuses on maximizing the overlap between predicted and true segmentation masks and is particularly effective in addressing significant class imbalance, such as when the tumor region occupies a much smaller volume than the background [ZLLW21]. It is directly derived from the Dice Similarity Coefficient (DSC) and is defined as Dice Loss $= 1 - \text{DSC}$. By minimizing the Dice Loss, the network aims to maximize the overlap between predicted and ground truth segmentations.

The Dice Loss formula is:

$$\text{Dice Loss}(y, \hat{p}) = 1 - \frac{2 \times |y \cap \hat{p}| + \epsilon}{|y| + |\hat{p}| + \epsilon} \tag{4.11}$$

Here y represents the set of true segments, also known as the ground truth segmentation. It refers to the actual segmentation mask of the tumor in the imaging data. $\hat{p}$ represents the set of predicted segments. It refers to the segmentation mask predicted by the model. $y \cap \hat{p}$ denotes the intersection between the true and predicted segments, representing the common area between the ground truth and predicted segmentations. $|y|$ represents the number of voxels in the true segmentation. $|\hat{p}|$ represents the number of voxels in the predicted segmentation. ϵ is a small constant value added to avoid division by zero, typically set to a very small number like 10^{-6} [RKF+22, WC18, YSSR22].

In summary, Dice Loss evaluates the degree of overlap between predicted and ground truth segmentations. A higher overlap results in a smaller Dice Loss, reflecting improved segmentation performance [ZLLW21].

The notation y and $\hat{p}$ in the Dice Loss formula refer specifically to the ground truth labels and predicted values, respectively. This differs from the general notation X and Y used in Equation 4.1, which represents a more abstract view of two sets being compared. Despite this difference in notation, the underlying calculation for Dice Loss and the Dice Similarity Coefficient remains identical.

4.2.2 Focal Dice Loss

The standard Focal Loss was introduced to tackle the class imbalance problem commonly found in tasks like dense object detection and medical image segmentation [NPSA21]. This loss function is particularly useful for datasets where the background class is overrepresented, and a small number of positive or minority examples (e.g., tumors in medical images) require accurate detection [LGG+20].

The formula for the standard Focal Loss is:

$$FL_{\text{standard}}(p_t) = -(1 - p_t)^{\gamma} \log(p_t) \tag{4.12}$$

Where p_t represents the model's predicted probability for the true class. If the true label is 1, p_t is the probability assigned to that class; otherwise, it is the probability assigned to the negative class. γ is the focusing parameter. A higher γ value reduces the relative loss for well-classified examples, focusing the model more on misclassified or difficult examples [LGG+20].

In this form, Focal Loss modulates the standard cross-entropy loss by the factor $(1-p_t)^{\gamma}$. This factor scales the contribution of each example to the loss based on how easy it is to classify. For easy examples (i.e., those with high p_t), this factor approaches zero, thus reducing their influence on the model's training. Conversely, for hard examples (i.e., those with low p_t), the factor is close to one, allowing these examples to contribute more to the loss. This mechanism makes Focal Loss especially effective in scenarios with high class imbalance, as it reduces the overwhelming influence of well-classified background examples, enabling the model to focus on learning from harder, often underrepresented instances [ZLLW21, Jad20, LGG+20].

Focal Loss with Class Weights

In medical image segmentation, particularly for brain tumor segmentation, certain regions of interest (e.g., contrast-enhancing tumor, necrotic (non-contrast-enhancing) core, edema) are much smaller compared to the background, leading to extreme class imbalance. To address this, an extension of the Focal Loss incorporates class weights into the formula,

giving more importance to these smaller, clinically relevant regions. This weighted version of Focal Loss is beneficial as it allows the model to learn the nuances of smaller classes without being overshadowed by the background [WC18, YSSR22].

The modified Focal Loss with class weights can be expressed as:

$$FL_{\text{weighted}}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{4.13}$$

Where α_t is a class-specific weight that adjusts the importance of each class. In brain tumor segmentation, for instance, higher weights might be assigned to the contrast-enhancing tumor, necrotic (non-contrast-enhancing) core, and edema classes to ensure they contribute more significantly to the loss calculation [LGG+20]. The focusing parameter γ continues to control the contribution of hard examples, as in the standard Focal Loss.

By introducing α_t, the Focal Loss becomes more versatile for imbalanced medical datasets. This customization allows the model to prioritize clinically critical but underrepresented regions, like small tumor parts, by increasing their impact on the total loss. Consequently, the model's sensitivity to these regions improves, reducing bias towards larger, less significant areas such as the background. This weighted approach thus mitigates the imbalance issue more effectively than the standard Focal Loss, promoting better segmentation performance on small but essential tumor structures [LGG+20, YSSR22, WC18].

Focal Loss with Class Weights for Multi-Class Segmentation

The weighted Focal Loss, originally designed for binary classification, has been adapted for multi-class segmentation tasks where each pixel can belong to one of several classes. In the context of the BraTS Challenge, there are four classes: background, necrotic (non-contrast-enhancing) core, edema, and contrast-enhancing tumor. The loss function calculates the error across all classes for each sample, then averages it over the dataset to provide a balanced learning signal across different structures [ZLLW21].

The formula for the multi-class Focal Loss with class weights is:

$$FL_{\text{multi-class}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{4}\alpha_c\, y_{true,i,c}\,(1-p_{i,c})^{\gamma}\log(p_{i,c}) \tag{4.14}$$

Where N represents the number of samples (or pixels, in image segmentation), and the summation over c (from 1 to 4) accounts for the four classes in a multi-class segmentation setting: background, necrotic (non-contrast-enhancing) core, edema, and contrast-enhancing tumor. The term α_c denotes a class-specific weight, allowing the loss function to assign a different importance to each class, which is especially useful in

cases where certain classes, like necrotic core or edema, have fewer pixels compared to the background. The binary indicator $y_{true,i,c}$ is set to 1 if class c is the correct class for sample i, otherwise it is 0. The predicted probability for class c for sample i is represented by $p_{i,c}$, which is raised to the power of $(1 - p_{i,c})^{\gamma}$, where γ is the focusing parameter that reduces the impact of well-classified examples, thus concentrating more on difficult-to-classify cases [ZLLW21].

This extended Focal Loss formula with class weights directly addresses the challenge of class imbalance in multi-class segmentation by integrating both class-specific importance and a mechanism to focus on harder-to-classify examples, ensuring a more nuanced learning process across diverse tumor regions [YSSR22, ZLLW21].

4.2.3 Combined Loss

The Combined Loss Function integrates both Dice Loss and Focal Loss to ensure that the model performs well on both class-imbalanced data and tumor segmentation accuracy.

The combined loss formula is:

$$\text{Combined Loss} = \text{Dice Loss} + \beta \times \text{Focal Loss} \tag{4.15}$$

Where β is a Focal Weight Factor, a weighting parameter that controls the balance between Dice Loss and Focal Loss. This parameter is tuned to optimize the model's performance across different tumor regions.

The focusing parameter γ, introduced by Lin et al. [LGG+20] in Focal Loss for Dense Object Detection in 2017, is used within the Focal Loss to reduce the impact of easily classified examples and emphasize harder cases. Typically, γ is set between 1 and 3. Values in this range have proven effective for addressing class imbalance by reducing the influence of easily classified examples and emphasizing harder cases. A value of 1 adds moderate weighting to difficult examples without compromising model stability, while 2 is often seen as optimal, striking a good balance between reducing focus on easy examples and prioritizing challenging ones. Higher values, like 3, further enhance the focus on difficult cases, which can be beneficial in highly imbalanced datasets, although very high values risk model instability by overly weighting a small set of challenging samples.

The combined use of Dice Loss and Focal Loss provides several advantages. Dice Loss encourages the model to maximize the overlap between predictions and ground truth, which is crucial for achieving high segmentation accuracy [YSSR22, SLV+17]. Focal Loss complements this by helping the model concentrate on difficult-to-segment areas, effectively addressing class imbalance by assigning higher importance to challenging examples [Jad20, WC18]. Additionally, incorporating class weights into the Focal Loss further enhances class balance; by assigning specific weights to each tumor class, the model can place greater emphasis on underrepresented or clinically significant regions, such as the contrast-enhancing tumor [WC18, ZLLW21].

In additive manufacturing (AM), a process where three-dimensional objects are built layer by layer, a Combined Loss Function using Dice Loss and Focal Loss has been applied to improve segmentation by addressing class imbalances and refining boundary detection of complex structures, such as melt pools in laser-powder bed fusion [SKS+21].

These method can be effectively adapted to medical image segmentation, including glioblastoma, where irregular contrast-enhancing tumor regions are often difficult to differentiate from solid and cystic non-contrast-enhancing regions. By applying this combined loss approach, segmentation accuracy, particularly in the delineation of these clinically significant subregions, can be enhanced.

This is illustrated in Figure 4.1, adapted from Schmid et al. [SKS+21], which highlights the challenges of segmenting complex structures. The figure demonstrates how the Combined Loss Function improves boundary detection, making it particularly useful for segmenting small, irregular regions that demand precise delineation.

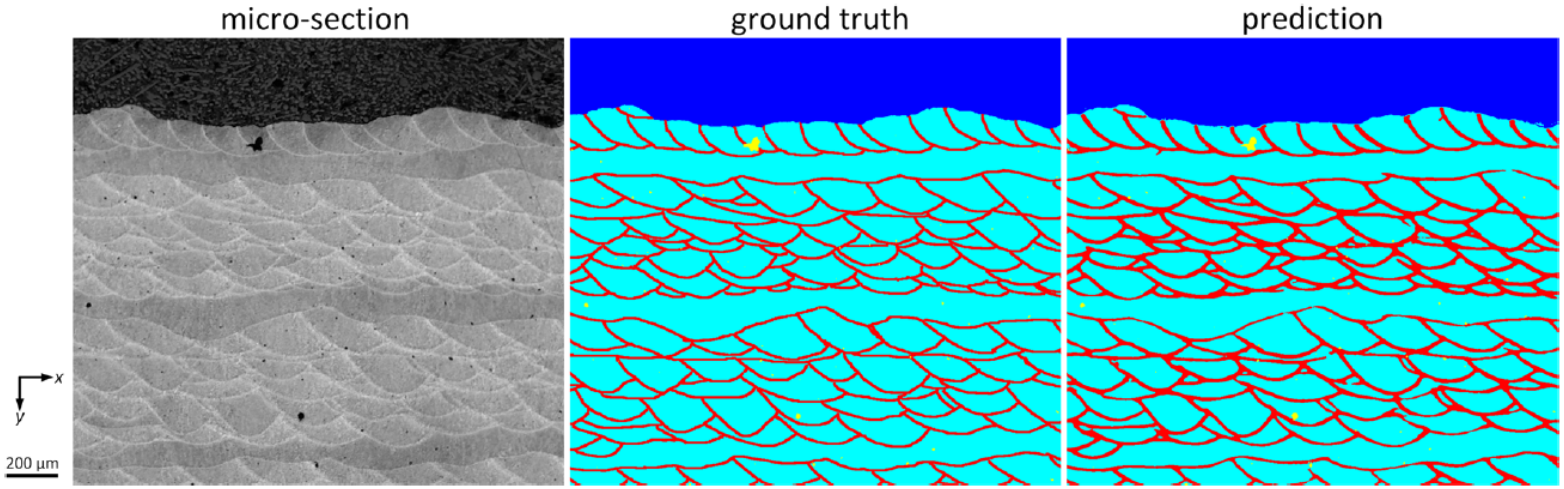


Figure 4.1: Adapted from Schmid et al. [SKS+21], the left panel shows a micro-section of a laser-powder bed fusion specimen with visible melt pools. The middle panel represents the ground truth segmentation (red boundaries), and the right panel shows the prediction from the model using a Combined Loss Function (Dice Loss and Focal Loss), highlighting improved boundary detection.

CHAPTER 5

Methodology

5.1 Patient Selection

Training Dataset

The training and validation datasets used in this study were derived from two primary sources, both provided by the same organization as part of their RSNA-ASNR-MICCAI Brain Tumor Segmentation Challenge (BraTS) [BGM+23].

The first dataset contains 369 glioma cases with histologically confirmed diagnoses, including both low-grade (LGG) and high-grade gliomas (HGG). From this dataset, 293 HGG cases were selected, as this is the target tumor type chosen for segmentation (Source 1). The second dataset consists of 125 cases for which no ground truth segmentation or histological diagnosis was available (Source 2). To expand the training pool, the thesis author as an experienced neuroradiologist evaluated these cases based on morphological criteria indicative of high-grade gliomas (HGG), aligning with the diagnostic emphasis on integrating morphological and molecular features outlined in the WHO Classification of Tumors of the Central Nervous System (CNS) by Louis et al. [LPW+21] in 2021. During this process, 25 cases were excluded due to morphological characteristics inconsistent with HGG, ensuring that only cases with a high likelihood of being HGG were included in the training dataset.

The selection of HGG cases across both datasets was motivated by their morphological and histopathological similarity to the evaluation dataset, which exclusively contains glioblastomas (GBM, WHO Grade IV). Glioblastomas and HGG share key features such as diffuse infiltration, high mitotic activity, and extensive vascular proliferation [WKTLRR22]. Including LGG cases, which lack these critical characteristics, in the training process would likely impair the segmentation performance of the 3D U-Net. Rebsamen et al. [RKR+19] demonstrated that stratifying training data by tumor grade improves segmentation performance, particularly for HGG, by reducing heterogeneity

within training data. They emphasize that glioblastomas (WHO Grade IV) exhibit distinct imaging features, such as necrosis, peritumoral edema, and contrast enhancement, which are not observed in LGG. Stratified training allows models to focus on consistent tumor phenotypes, optimizing segmentation accuracy for high-grade tumors.

After combining both datasets, a total of 393 cases are available for training. As described in Section 5.2.2, ground truth segmentations are created or revised for all selected cases to ensure consistent and accurate training data.

Cross-Validation Split

To evaluate the model during training, the 393 cases in the training dataset were split into 80% training and 20% validation subsets. This approach aligns with practices commonly adopted in the BraTS challenges, where standardized cross-validation strategies are employed to optimize model performance and ensure robust internal validation while maintaining a sufficient amount of data for training [FSL$^+$24, ZKMUB21]. For example, the BraTS 2021 challenge emphasized that such methodologies are critical for the development of accurate segmentation models, enabling thorough performance evaluation on independent subsets [ZKMUB21, ZKBMU22].

Zeineldin et al. [ZKBMU22] further highlight that cross-validation and similar techniques have become integral to glioblastoma segmentation within the BraTS framework. These methods facilitate iterative parameter tuning and reliable performance assessments, thereby ensuring model generalizability to unseen data—a requirement essential for clinical applicability and broader validation of segmentation algorithms [FSL$^+$24, ZKBMU22].

Case Group Formation and Dataset Splitting

To further analyze the effect of different Case Group sizes on model performance, four distinct Case Groups were created. These Case Groups were designed to maintain the proportional source distribution between Source 1 and Source 2 across both training and validation subsets. The cases from Source 1 and Source 2 were randomly assigned to the training and validation subsets, ensuring proportional representation but without preserving the original order:

- Cases group 80 (80 training plus 20 validation cases)
- Cases group 160 (160 training plus 40 validation cases)
- Cases group 240 (240 training plus 60 validation cases)
- Cases group 314 (314 training plus 79 validation cases)

Both the training and validation Case Groups build incrementally on the smaller ones. For example, Case Group 160 includes Case Group 80, and Case Group 240 includes Case Group 160. The largest group, Case Group 314, contains all available cases. The source ratio (293 cases from Source 1 and 100 cases from Source 2) was consistently

preserved across both training and validation sets. The models were trained using the training cases exclusively, while the validation cases were used to monitor and evaluate the training process.

Evaluation Dataset (External Unseen Data)

The evaluation dataset consists of 108 patients, retrospectively selected, who were treated for newly diagnosed glioblastoma (GBM) at the University Hospital Salzburg between February 2009 and August 2022. These patients were not involved in the training process and serve as completely unseen data, providing a basis for the external evaluation of the model's performance.

The Inclusion Criteria for this dataset are as follows:

- Adult patients (aged 18 years or older).
- Newly diagnosed GBM with treatment initiated at the University Hospital Salzburg.
- Preoperative MRI scans are available for segmentation, with imaging performed prior to any treatment (surgical, radiological, or chemotherapy).
- Histopathological grading is based on the latest World Health Organization (WHO) classification available at the time of diagnosis [LPW+21].
- No study-specific MRI protocols, surgical procedures, or treatments were performed. All imaging and treatment protocols followed standard clinical practice.
- Clinical data were extracted from electronic medical records, and all patients or their legal guardians provided consent for imaging and treatment as part of their clinical care.

This evaluation dataset is used to assess the generalizability and robustness of the trained model on real-world, unseen data after the training phase has been completed. The external evaluation helps determine the clinical applicability of the segmentation model.

5.2 Imaging Data Source

Training Dataset (BraTS Dataset)

The BraTS dataset [BGM+23] used for training the segmentation model includes multi-modal MRI scans from patients with gliomas. These scans include the following sequences: native T1, post-contrast T1-weighted, T2-weighted, and FLAIR. The data were acquired using different clinical protocols on MRI scanners from 19 different institutions, ensuring a diverse dataset that reflects a wide range of clinical settings. The dataset is provided in NIfTI format [CAB+04], allowing for easy integration into segmentation algorithms.

The acquisition of these scans was performed in different orientations, with the axial plane being the most commonly used. However, some sequences, particularly those acquired in the coronal plane, exhibit thicker slices and, as a result, lower resolution in the axial direction. Additionally, there is variation in the native T1-weighted sequences, which include both Fast Field Echo (FFE) and Turbo Spin Echo (TSE) acquisitions. The TSE sequences were used more frequently in older scans, suggesting that the data span is approximately two decades.

Due to the inconsistency in the acquisition protocols of the native T1 sequences and the differing image impressions caused by the use of both Fast Field Echo (FFE) and Turbo Spin Echo (TSE), the decision was made to exclude the native T1 sequence from the training process. This exclusion not only ensures consistency in the data but also reduces the overall dataset size by approximately one-quarter, making the training process more efficient.

Evaluation Dataset

The evaluation dataset consists of 108 patients treated at the University Hospital Salzburg between February 2009 and August 2022. The MR imaging for these patients was acquired either in-house or from external sources. A total of 83 cases were performed using in-house protocols, 4 cases were performed externally, and 21 cases had external imaging, which was subsequently complemented by additional in-house sequences.

For in-house MR imaging, a 3T MRI machine (Achieva dStream, Philips Medical Systems, Best, Netherlands) with a 32-channel head coil was used. The following imaging parameters were applied:

- **T2-weighted imaging (TSE):** Axial orientation with 28 slices, echo time (TE) of 80 ms, repetition time (TR) of 3000 ms, field of view (FOV) of 560 mm × 560 mm, voxel size of 0.65 × 1.13 mm, with a slice thickness of 4 mm and a gap of 5 mm between slices.
- **FLAIR imaging (TSE):** Axial orientation with 28 slices, echo time (TE) of 125 ms, repetition time (TR) of 10000 ms, inversion time (TI) of 2800 ms, field of view (FOV) of 560 mm × 560 mm, voxel size of 0.65 × 1.13 mm, with a slice thickness of 4 mm and a gap of 5 mm between slices.
- **T1-weighted imaging (FFE):** Sagittal alignment, echo time (TE) of 4.04 ms, repetition time (TR) of 8.66 ms, field of view (FOV) of 320 mm × 320 mm, voxel size of 1 × 1 × 1 mm, performed before and after the administration of gadolinium, following the acquisition of T2 and FLAIR sequences.

For patients with external imaging, variability in the imaging protocols was present. External MRI examinations were performed outside the neuroradiology department, prior to admission to the neurosurgery department of the Christian Doppler Clinic.

Consequently, the protocols used in these external scans varied in terms of parameters such as Gradient Echo (FFE) vs. Turbo Spin Echo (TSE) sequences and spatial orientation (sagittal, coronal, and axial planes).

Specifically, for the T2-weighted sequences, eight cases were acquired in coronal orientation and one case in sagittal orientation, while the remaining cases were acquired in axial orientation. For the FLAIR sequences, eight cases were also acquired in coronal orientation and one case in sagittal orientation, with the remaining cases being acquired in axial orientation.

This variability in the external scans posed a challenge for standardizing the evaluation dataset. However, the inclusion of complementary in-house imaging for 21 cases ensured that essential sequences were available for evaluation in most cases.

5.2.1 Tumor Annotation

Gliomas can be substructured according to their appearance. Menze et al. [MJB+15] defined four types of intra-tumoral structures, namely “edema,” “non-enhancing (solid) core,” “necrotic (or fluid-filled) core,” and “contrast-enhancing core”. Strictly speaking, the term "intra-tumoral" does not pathognomonically apply to the edema, as it is not part of the tumor itself, but is caused by it. Nevertheless, the segmentation of the edema is of great relevance subsequently. The mentioned anatomical substructure can essentially be found in any glioblastoma, which leads to deriving an annotation protocol from this fact.

Menzel et al. [MJB+15], for example, outline their process in five sequential steps, which should be completed in the following order:

1. The “edema” was segmented primarily from T2-weighted images. FLAIR was used to cross-check the extension of the edema and discriminate against ventricles and other fluid-filled structures. The initial “edema” segmentation in T2 and FLAIR contained the core structures that were then relabeled in subsequent steps, as shown in Figure 5.1 (A).

2. As an aid to the segmentation of the other three tumor substructures, the so-called gross tumor core—including both enhancing and non-enhancing structures—was first segmented by evaluating hyperintensities in T1 post-contrast (for high-grade cases) together with the inhomogeneous component of the hyperintense lesion visible in T1 and the hypointense regions visible in T1, shown in Figure 5.1 (B).

3. The “contrast-enhancing core” of the tumor was subsequently segmented by thresholding T1 post-contrast intensities within the resulting gross tumor core, including the Gadolinium enhancing tumor rim and excluding the necrotic center and vessels. The appropriate intensity threshold was determined visually on a case-by-case basis, as shown in Figure 5.1 (C).

4. The "necrotic (or fluid-filled) core" was defined as the tortuous, low-intensity necrotic structures within the enhancing rim visible in T1 post-contrast. The same label was also used for the very rare instances of hemorrhages in the BraTS data, as shown in Figure 5.1 (C).

5. Finally, the "non-enhancing (solid) core" structures were defined as the remaining part of the gross tumor core, i.e., after subtraction of the "contrast-enhancing core" and the "necrotic (or fluid-filled) core" structures, as shown in Figure 5.1 (D).

With regard to the approach described by Menze et al. [MJB+15], it is important to emphasize that experienced neuroradiologists may not necessarily require native T2-weighted and T1-weighted sequences for the sub-segmentation of glioblastoma. The edema can be accurately identified using the FLAIR sequence alone, while the contrast-enhancing regions of the tumor are best visualized using the post-contrast T1-weighted sequence. These two sequences provide the essential information needed for segmentation without relying on native T2- or T1-weighted images. If the tumor has non-contrast-enhancing margins, the information about its margins is obtained from the FLAIR sequence. Since the patient population consists primarily of newly diagnosed, non-previously operated glioblastoma patients (as described in Section 5.1), it is possible to deviate from the procedure described above.

Müller et al. [MKE+24] emphasize that the most important differential diagnosis of glioblastoma (GBM) is cerebral metastasis, as both entities can mimic each other on anatomical imaging due to similar or overlapping features. In their comparison of multiple imaging characteristics, they identified that glioblastomas and brain metastases may present similar imaging patterns on T2/FLAIR sequences, complicating differentiation.

Buchner et al. [BPE+23] conducted a detailed evaluation of MRI sequences to optimize automated segmentation for brain metastases. They concluded that the T1 post-contrast sequence alone was sufficient for segmenting brain metastases, achieving a median Dice similarity coefficient (DSC) of 0.96. For edema segmentation, however, the combined use of T1 post-contrast and FLAIR sequences was critical, with the best-performing models achieving a median DSC of 0.93. Their study highlights that optimizing MRI protocols by excluding unnecessary sequences can streamline clinical workflows and enhance segmentation accuracy for neural network-based target delineation.

These findings, derived from independent studies, suggest that it is not always necessary to include all four conventional MRI sequences for tumor segmentation. Instead, sequence selection should be guided by the specific clinical or research objectives, as demonstrated in the segmentation of brain metastases.

5.2.2 Ground Truth Segmentation

Glioblastoma segmentation was performed using the 3D Slicer version 5.6.2 [FBKC+12], an open-source platform widely used for medical image segmentation. This software enables comprehensive segmentation, visualization, and analysis of medical images. In

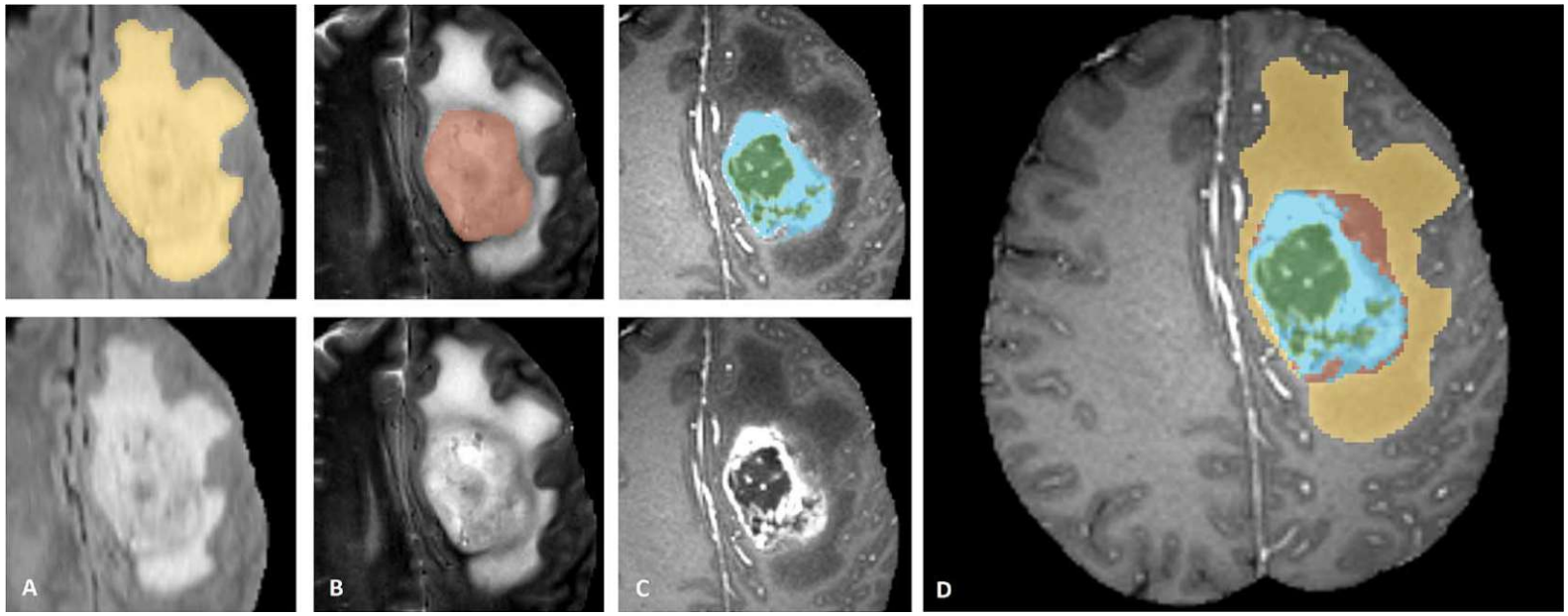


Figure 5.1: Manual annotations by expert raters, adapted from Menze et al. [MJB+15]. Left: Tumor components in different modalities—(A) whole tumor in FLAIR, (B) tumor core in T2, and (C) contrast-enhancing tumor (blue) and necrotic core (green) in T1ce. Right: Final segmentation labels showing edema (yellow), non-enhancing solid core (red), necrotic/cystic core (green), and contrast-enhancing tumor (blue).

this study, we focused on segmenting glioblastomas into three specific classes: non-contrast-enhancing tumor (Class 1), edema (Class 2), and contrast-enhancing tumor (Class 3).

Image Registration and Preprocessing

The process begins with the registration of the various imaging sequences, specifically FLAIR and FFE T1-weighted post-contrast (T1ce) images, using the General Registration (BRAINS) module within 3D Slicer. The T1ce sequence, being the highest-resolution sequence, serves as the reference onto which the other sequences are registered. This step ensures that all images are aligned accurately, facilitating precise segmentation, especially considering the differing resolutions of the FLAIR and FFE T1-weighted post-contrast images. Following the registration, the sequences were normalized to standardize the intensity values, which significantly eased the semi-automatic segmentation using the "Grow from Seeds" and "Sphere Brush" tools.

Segmentation Process

The "Grow from Seeds" algorithm in 3D Slicer's Segment Editor module [PLF19] utilizes a region-growing technique for semi-automatic segmentation of structures within medical images. This method involves placing seed points inside and outside the target region; the algorithm then iteratively expands these seeds based on image intensity and spatial information until the regions converge, effectively delineating the structure's boundaries. Such region-growing techniques are widely used in medical image segmentation due to their simplicity and effectiveness in capturing complex anatomical structures [POC14].

The "Sphere Brush" tool in 3D Slicer's Segment Editor module [PLF19] is designed to facilitate precise manual segmentation by allowing users to paint spherical regions directly onto medical images. This tool is particularly effective if combined with the "Editable Intensity Range" setting, which restricts modifications to voxels within a specified intensity range. This combination is especially useful for delineating contrast-enhancing tumor regions that exhibit distinct intensity differences from surrounding tissues. By setting appropriate intensity thresholds, users can accurately target and segment these regions, enhancing the precision of the segmentation process.

Classification

- **Class 1 (non-contrast-enhancing tumor)** and **Class 3 (contrast-enhancing tumor):** These classes are primarily identified using the FFE T1ce sequence. The distinct contrast differences between these regions and the surrounding parenchyma allow for reliable differentiation between contrast-enhancing and non-contrast-enhancing tumor regions.

- **Class 2 (edema):** Edema is predominantly segmented using the FLAIR sequence. This modality is particularly effective in highlighting hyperintense areas associated with tumor-related edema, providing clear boundaries for segmentation.

- **Class 0 (no tumor):** This class comprises all voxels that do not belong to Classes 1, 2, or 3. These represent regions free of any tumor-associated signal abnormalities, encompassing normal brain parenchyma and other non-pathological areas.

Manual Adjustment and Validation

To ensure consistent segmentation quality, the initial segmentations were iteratively reviewed and refined. This process involved the neuroradiologist repeatedly revisiting and adjusting the segmentation boundaries to enhance accuracy and achieve finer delineation. The improvement in segmentation precision through repetition and practice is well-supported by the findings of Liu et al. [LQX+24], who emphasize that iterative interaction between human expertise and segmentation tools is essential for achieving high-quality results in medical image analysis. The entire adjustment and validation process was time-intensive, ranging from 30 to 60 minutes per case, depending on the complexity of the tumor structures and the clarity of the imaging data. This iterative refinement process, supported by the neuroradiologist's growing familiarity with the dataset and segmentation nuances, aligns with the best practices highlighted in the systematic review. Once the segmentation was performed according to the defined classes, the final result was obtained, as demonstrated in Figure 5.2.

Revision of expert segmentation

To ensure consistency across the dataset and optimize the quality of segmentations for downstream analysis, the expert-provided segmentations were systematically reviewed and refined. This process was essential to mitigate interrater variability, ensuring that all cases in the training, cross-validation, and evaluation datasets were segmented by a single expert. As highlighted by Conze et al. [CAMS+23], manual segmentation by multiple raters often introduces variability due to differences in individual expertise and interpretation, which can adversely affect the consistency of the dataset and the performance of downstream models. It is important to emphasize that the revisions did not aim to critique the work of the experts, whose segmentations provided a robust foundation for the analysis. Instead, the revisions served to further standardize the dataset and ensure alignment with the specific requirements of the study. The refinement process also enabled a sharper focus on achieving precise sub-segmentation of the contrast-enhancing tumor regions, which are critical for clinical decision-making and model evaluation. The effort required for this process was significant, with each case requiring an average of 15 to 25 minutes for review and refinement, depending on the complexity of the tumor structure and the quality of the initial segmentation. Given the dataset of 293 cases, this resulted in a total estimated time investment of approximately 100 hours. Figure 5.3 illustrates two

examples from the training dataset, comparing the initial expert segmentation with the refined segmentation.

Statistics

For the analysis of all data, SPSS version 29.0.2.0 [Cor24] was used. Measurement data with normal distribution and uniform variance are expressed as mean ± standard deviation, while non-normally distributed data are stated as median and interquartile range (IQR). Comparisons of patient age in the different datasets were performed using the Mann-Whitney U test [MW47]. Additionally, Matplotlib [Mat24], a comprehensive library for creating static, animated, and interactive visualizations in Python [VR09], was employed exclusively for the graphical representation of numerical data through static charts and graphs.

5.3 Data Preprocessing

The preprocessing pipeline ensures consistency and comparability between the evaluation and training datasets, aligning them to a standardized format suitable for machine learning applications. Unlike the training dataset, which is already provided in the BraTS-compatible format, the evaluation dataset requires additional preprocessing steps to achieve compatibility. Therefore, the BrainLes Preprocessing Package of the BraTS Toolkit [KBW+20] is used. This Toolkit is specifically designed to facilitate this process, ensuring that the evaluation dataset adheres to the same standards as the training data. Once the evaluation dataset has been processed to match the BraTS standards, additional preprocessing steps are applied to both datasets. The complete preprocessing pipeline for the evaluation dataset is visualized in Figure 1.1, outlining all steps required.

Preprocessing for the Evaluation Dataset

For the evaluation dataset, the following preprocessing steps are applied to transform the data into a BraTS-compatible format:

- **DICOM to NIfTI Conversion:** All MR images are converted from DICOM format to NIfTI format (Neuroimaging Informatics Technology Initiative), which is the standard format used in the BraTS dataset.
- **Co-Registration:** Co-registration ensures that all imaging modalities of a patient (native T1, post-contrast T1-weighted, T2-weighted, and FLAIR) are aligned within the same spatial coordinates. This alignment ensures that anatomical structures appear in the same location across all modalities, which is essential for accurate segmentation in machine learning models.
- **Normalization:** This process adjusts the intensity values of MRI images across all modalities to a common scale. By reducing variability between patients and imaging

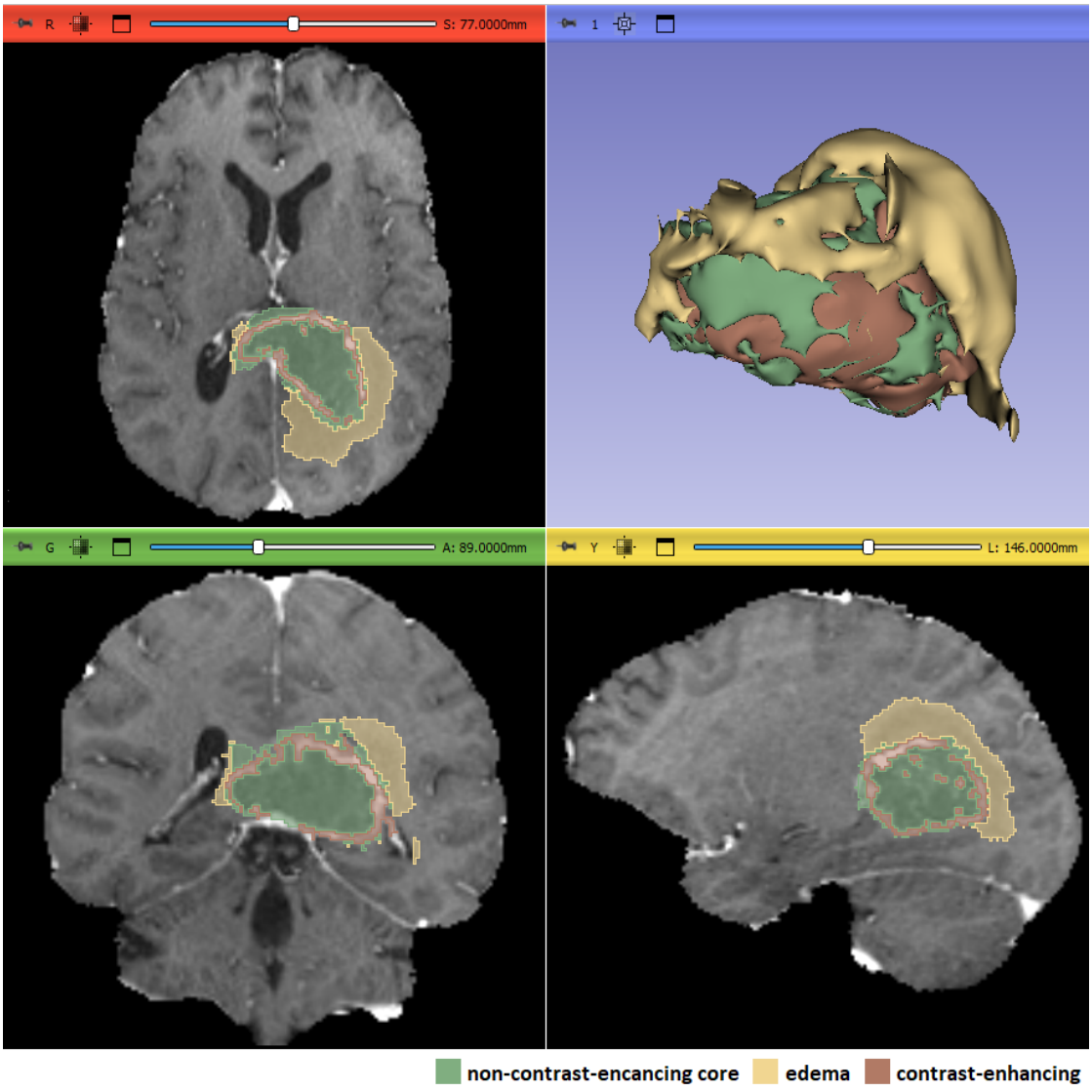


Figure 5.2: The underlying image sequence is the FFE T1ce sequence, displaying the segmentation of three glioblastoma classes: non-contrast-enhancing tumor (Class 1, green), edema (Class 2, yellow) and contrast-enhancing tumor (Class 3, brown). The segmentation of the edema was based on the FLAIR sequence, facilitated by the hyperintense difference in intensity between the tumor edema and the surrounding parenchyma. The upper-right image shows a 3D visualization of the glioblastoma, illustrating the spatial arrangement of the tumor components.

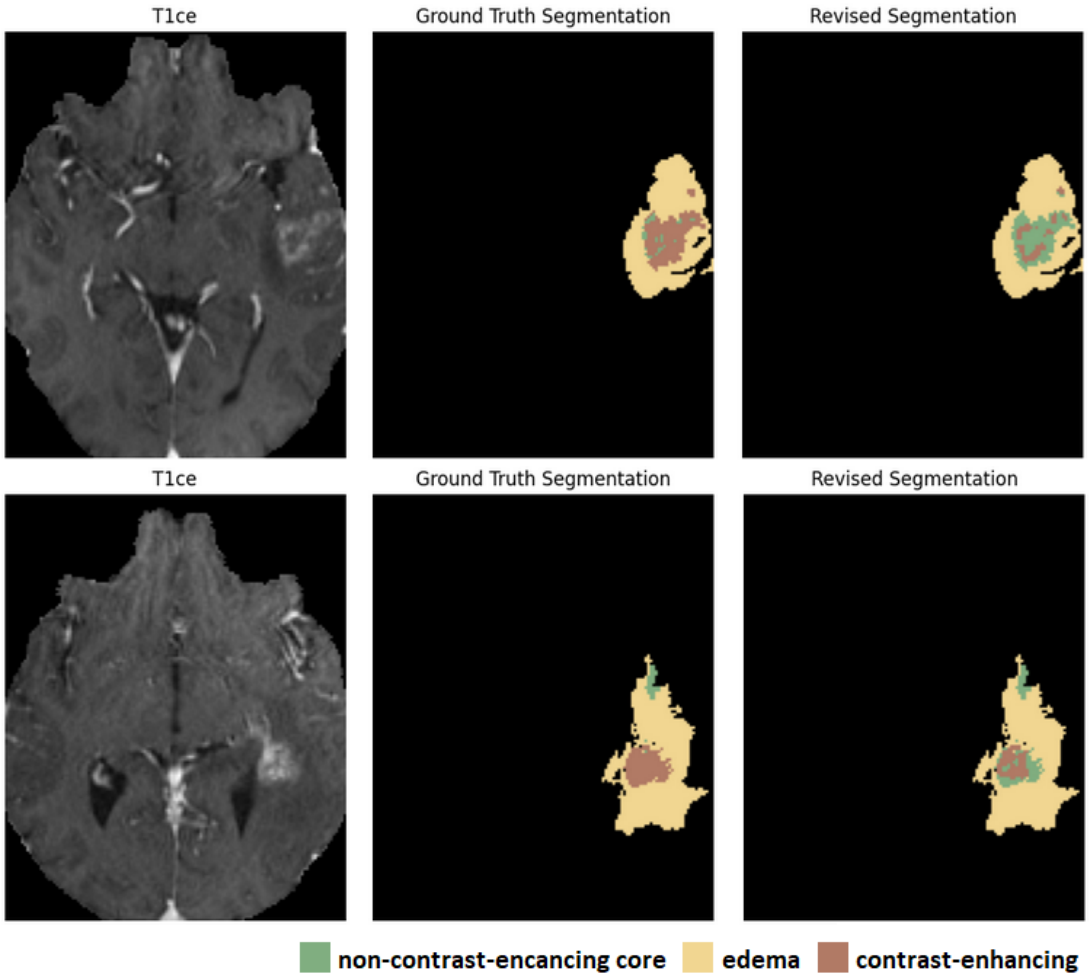


Figure 5.3: Comparison of Expert and Revised Segmentations. Two examples from the training dataset are displayed, with the first example in the upper row and the second in the lower row. On the left of each triplet is the T1 post-contrast sequence (T1ce), in the center is the ground truth segmentation provided by the expert panel, and on the right is the revised segmentation generated by our method. The segmentation highlights the non-contrast-enhancing tumor (green), tumor edema (yellow), and contrast-enhancing tumor (brown). The expert segmentation appears coarser, while the revised segmentation demonstrates a finer delineation.

sessions, normalization ensures more consistent image analysis and segmentation results, regardless of differences in scanner settings or patient positioning. A percentile normalizer is used to adjust the intensity values of the scans, reducing the impact of outliers and standardizing the intensity distribution.

- **Skull-Stripping:** Skull-stripping removes non-brain structures, particularly the skull, from the images. This step ensures that machine learning algorithms focus exclusively on the brain and tumor regions, without interference from surrounding structures. The High-Definition Brain Extraction Tool (HD-BET) [ISP+19], a deep learning-based tool, is used to automate this process, ensuring precise and efficient skull removal.

- **Conversion into SRI-24 Space:** The SRI-24 space [RZSP10] is a standardized brain reference space based on a set of 24 normal brain scans developed by the Stanford Research Institute (SRI). Converting the MRI data to this standardized space ensures consistency across patients and imaging protocols, which is crucial for comparability in machine learning applications. NiftyReg [ORS+01] is used for registration and alignment to the SRI-24 anatomical atlas, ensuring the data is in the expected format for BraTS challenge algorithms.

Common Preprocessing Steps for Both Datasets

After the initial preprocessing of the evaluation dataset to ensure consistency with the BraTS dataset, the following additional steps are applied to both the evaluation and training datasets. While these steps are briefly introduced here for the sake of completeness, their implementation is described in greater detail in the implementation Sections 6.3 and 6.4.

- **Contrast Limited Adaptive Histogram Equalization (CLAHE):** CLAHE [Zui94] is applied to enhance the contrast of the MRI images, making it easier for the segmentation algorithms to distinguish between different brain structures. This contrast adjustment is crucial for improving the visibility of tumor boundaries.

- **Cropping:** The images are cropped to reduce their dimensions by removing areas outside the brain, such as the skull and surrounding tissues. Cropping reduces the size of the dataset, improving computational efficiency during training by focusing only on the relevant anatomical structures.

Data Augmentation (Training Dataset Only)

Data augmentation is applied exclusively to the training dataset to artificially increase the number of training samples and improve model generalization. Specifically, only rotations and flips are used as augmentation techniques, which are explained in detail in the following Section 5.4. No augmentations are applied to the evaluation dataset, as it is strictly reserved for testing the model after training. The implementation details

of the augmentation strategy, including how and in which scenarios augmentations are applied, are discussed comprehensively in the implementation Section 6.5. To ensure reproducibility and significantly reduce training time, the augmentations are pre-generated and stored as NumPy arrays, so they do not need to be generated repeatedly during runtime.

Efficient Data Storage as NumPy Arrays

After applying the preprocessing steps, the training, validation, and augmentation datasets are stored as NumPy arrays. This approach ensures faster data loading during training and validation, which improves computational efficiency. By storing the augmentations as NumPy arrays, reproducibility is also guaranteed, as the augmentations do not need to be regenerated at runtime.

5.4 Data Augmentation

Geometric transformations are an essential aspect of data augmentation in medical image analysis, as they help simulate variability in datasets without altering the anatomical integrity of the structures. As highlighted by Paschali et al. [PSR+19] and Goceri [Cos23], transformations such as rotations, translations, and scalings are crucial for increasing dataset diversity and improving the robustness of models against variations in image orientation and positioning.

The selection of augmentation techniques was guided by the need to preserve the clinical relevance of the images. Methods such as color adjustments, intensity transformations, or noise injection were intentionally excluded. These techniques, though commonly applied in data augmentation, risk introducing artifacts or distortions that could degrade model performance. Kumar et al. [KBMB24] emphasize that while non-geometric augmentations such as color space manipulation and noise injection can improve model robustness, they may also introduce noise or distortions, potentially leading to degraded performance in specific contexts. Similarly, Nanni et al. [NPBL22] highlight the necessity of carefully choosing augmentation techniques to avoid generating samples that do not align with the original data distribution.

Extensive preprocessing was conducted to optimize image quality and enhance detail and clarity. To preserve the diagnostic integrity of the medical images, noise and photometric alterations were avoided, as these could counteract these improvements and compromise image reliability. Instead, augmentation focused exclusively on geometric transformations, such as rotations and flips, to balance dataset enhancement with the preservation of image clarity.

5.4.1 Rotations and Flips

To introduce variability while preserving anatomical integrity, rotations and flips were employed as geometric transformations. Rotations were applied by randomly rotating

each image within a range of ±10° along the x, y, and z axes. This approach, commonly used to introduce subtle variations while maintaining anatomical fidelity, has been demonstrated in studies by Goceri [Goc23] and Paschali et al. [PSR+19]. The restricted range ensured that critical brain structures remained visible, avoiding the exclusion of important diagnostic information. Limiting the rotation angle was necessary because, in the final step of preprocessing, the surrounding regions of the brain in the images were cropped to reduce the data size, as detailed in Section 6.4. Consequently, larger rotation angles, such as the fixed rotations of 90° and 270° employed by Nanni et al. [NPBL22], could result in anatomical details being rotated out of the image, compromising the integrity of the data. By restricting the angle to ±10°, sufficient variability was introduced while ensuring that essential anatomical content remained within the field of view, enhancing model robustness without risking the loss of critical information.

Flips were applied by randomly mirroring the images along the x, y, or z axes. This straightforward yet effective transformation increased the dataset size by introducing positional variance without altering the anatomical structures or associated labels. By consistently applying these transformations across all modalities—including FLAIR, T1ce, T2, and segmentation images—the augmentation process ensured uniformity and maintained the diagnostic relevance of the dataset.

5.4.2 Optimal Augmentation Ratio

The determination of the optimal ratio between original and augmented data is crucial for maximizing the performance of deep learning models in medical image segmentation tasks, such as Glioblastoma segmentation. Literature suggests that the appropriate mix of original and augmented data can significantly impact model accuracy and generalization capabilities.

The determination of the optimal ratio between original and augmented data is a critical factor in maximizing the performance of deep learning models in medical image segmentation tasks. Abdalla et al. [AMS23] demonstrate that the balance between original and augmented data directly impacts model accuracy, with improperly balanced ratios leading to either underfitting or overfitting, thereby reducing generalization capabilities. Similarly, Pattilachan et al. [PDK+22] emphasize that while augmentation enhances dataset diversity, an excessive reliance on augmented data may distort the representation of the original dataset, highlighting the need for a carefully calibrated ratio. The literature suggests that this ratio must be optimized based on the specific dataset and task requirements, as an imbalance could compromise the clinical relevance and robustness of the trained model [Cos23].

Studies, such as the comprehensive survey by Mumuni et al. [MM22], have shown that training with a higher proportion of augmented images generally yields better results compared to training with a smaller proportion of augmented images. This survey explored various augmentation ratios for medical imaging tasks and highlighted that using three times as many augmented images as original images led to higher accuracy compared

to using twice as many or an equal number of original and augmented images. These findings underscore the importance of carefully optimizing the ratio between original and augmented data, especially if dealing with small datasets, which are common in medical image segmentation tasks. The survey's conclusions emphasize that an increased quantity of augmented data, when appropriately balanced, can significantly improve model performance and generalization capabilities.

Based on the survey by Mumuni et al. [MM22], a recommended starting point for training is a 1:3 ratio, with augmented images outnumbering original images by three to one. This baseline can be fine-tuned to account for the specific characteristics of the dataset. Ratios beyond 1:3, such as 1:4 or 1:5, were not tested in this study due to concerns about potential overfitting to augmented data, which could reduce the model's ability to generalize effectively to unseen cases. Previous research emphasizes that excessive augmentation may introduce augmentation-specific features or Out-of-Distribution (OOD) data, thereby degrading model performance and limiting generalization [PS20, ZYQ+23]. Consequently, an over-reliance on augmented data might cause the model to focus on features derived from augmentation techniques rather than the inherent characteristics of the original data. Comparing performance across other ratios, such as 1:1 and 1:2, provides a balanced exploration of augmentation effects, helping to identify the optimal mix for Glioblastoma segmentation while maintaining robust model performance.

5.5 Model Architecture

The model architecture is based on the 3D U-Net, an extension of the U-Net tailored for volumetric data processing. This architecture was chosen for its ability to capture spatial relationships across all axes (x, y, z) in medical imaging data, making it particularly effective for MRI-based segmentation tasks [HHY+19]. The 3D U-Net is designed to segment glioblastomas by processing multisequence inputs, such as T1ce, T2, and FLAIR.

The core structure consists of a contracting path and an expanding path connected by skip connections. The following modifications were made to adapt the architecture for glioblastoma segmentation. Dropout rates were progressively increased in deeper layers to mitigate overfitting, allowing the model to generalize better to unseen data. Additionally, the number of filters was expanded in these layers to enable the network to capture more complex and abstract features, essential for identifying intricate tumor structures. Furthermore, a Combined Loss Function was employed, integrating Dice Loss and Focal Loss, to optimize both overlap accuracy and the model's ability to handle class imbalances effectively. Details regarding these specific adjustments are discussed in Section 6.7.

5.5.1 Input & Output

The input to the 3D U-Net consists of a three-channel volumetric image composed of the post-contrast T1-weighted, T2-weighted, and FLAIR sequences. Each input volume

is preprocessed, including skull-stripping, normalization, and alignment to the SRI-24 space [RZSP10], as described in Section 5.3 and Section 6.1.

The output is a segmentation mask with the same spatial dimensions as the input volume. The mask provides voxel-wise predictions for four classes: background (Class 0, no tumor), non-contrast-enhancing tumor (Class 1), edema (Class 2), and contrast-enhancing tumor (Class 3). Each voxel is assigned to one of these classes based on the model's predictions, ensuring a comprehensive delineation of both tumor regions and non-tumor areas.

5.5.2 Loss Function & Optimization

The Combined Loss Function integrates both Dice Loss and Focal Loss to ensure the model performs well on class-imbalanced data while achieving accurate tumor segmentation.

Dice Loss: Derived from the Dice Similarity Coefficient (DSC), the Dice Loss ensures that the overlap between the predicted and true tumor regions is maximized. The Dice Loss is weighted according to the class distribution to prioritize the accurate segmentation of smaller tumor regions. The class weights are calculated based on the ground truth segmentations in the training dataset, ensuring that each class (necrotic (non-contrast-enhancing) core, edema, contrast-enhancing tumor, and background) is properly balanced.

Focal Loss: The Focal Loss helps the model focus on hard-to-classify voxels by applying a modulating factor to down-weight easy examples. The class distribution from the training dataset is used to calculate the alpha values in the Focal Loss function, which helps address the imbalance between different tumor classes.

The **class weights** used in the Focal Loss function are derived from the ground truth segmentation of the training dataset, which is discussed in detail in Section 5.2.2. They ensure that the model pays more attention to smaller and more clinically relevant classes, such as the necrotic (non-contrast-enhancing) core and contrast-enhancing tumor, which are critical for accurate segmentation of glioblastomas.

Combined Loss: This loss function is the sum of the Dice Loss and the Focal Loss, weighted by a Focal Weight Factor. This factor is tuned using Keras [Cho15] Tuner to find the optimal balance between the two loss components during training. The theoretical foundations and rationale for using this particular loss function in the project are described in Section 4.2.3.

Optimizer: The model is trained using the Adam optimizer with an initial learning rate of 0.001. The learning rate is dynamically adjusted during training using a ReduceLROnPlateau callback to prevent overfitting and ensure smooth convergence.

5.5.3 Hyperparameter Tuning

To optimize the Focal Weight Factor, Keras [Cho15] Tuner is used to explore a range of values between 0.0 and 5.0, with a step size of 0.1. The optimal Focal Weight Factor is

determined based on the validation Dice Score, and the results of the tuning process are discussed in Chapter 7. The MyHyperModel class is used to systematically test different values and fine-tune the model's performance to find the optimal balance between the two losses, allowing the model to better handle challenging cases and smaller tumor regions while maintaining overall segmentation accuracy.

5.5.4 Optimization & Reproducibility

To ensure both reproducibility and flexibility in the loading order of training cases, a seed list with 100 entries is employed to control the random generator for shuffling the dataset. For each Focal Weight Factor, the training process begins with the first entry in the seed list, ensuring a consistent, deterministic order. The random generator's behavior is standardized across NumPy [HMvdW+20], TensorFlow [Dev16], and Python's random module using an initial seed, which is set to 7070 for all random operations to ensure that the training process is reproducible. This guarantees comparable results across different configurations while allowing controlled randomness.

During training, early stopping and model checkpoint callbacks are applied to manage the process efficiently. The random seed is reset for each Focal Weight Factor, ensuring consistent behavior even with the inclusion of data augmentation and shuffling.

Additionally, Mixed Precision Training with the *mixed float16* policy is utilized to optimize performance and reduce memory consumption. This approach leverages 16-bit floating-point precision on supported hardware, enabling faster computations on GPUs with Tensor Cores [Ten24], significantly improving training speed without sacrificing accuracy—an essential advantage if working with large 3D medical datasets.

5.5.5 Training Process

The model is trained with up to 393 cases in 4 different Case Groups, with an 80/20 split for training and validation. The Batch size for each epoch is set to 1, 2, or 4, with pre-loaded augmentations either used or skipped, depending on the scenario. Training continues up to 100 epochs for each Focal Weight Factor, with early stopping based on validation Loss. The best model, as determined by the validation IoU Score, is saved for evaluation on the unseen evaluation dataset.

The training process of the 3D U-Net model involves a combination of data generators, Batch size variations, and advanced techniques like early stopping, and dynamic learning rate adjustments to optimize the model's performance and ensure reproducibility.

Data Generators

The training and validation data generators handle data loading in real time, while augmentations are pre-generated and applied during the loading process. Depending on the scenario, different ratios of original to augmented cases are used, where the first

number represents the count of original cases and the second number indicates the count of augmentations:

The explored ratios include 2:1, 1:1, 1:2, and 1:3, as these are found to impact the model's generalization capabilities and are further discussed in Chapter 7.

Both generators are implemented with a seed management system to shuffle the data in a reproducible manner, ensuring consistent results across experiments with different configurations of the Focal Weight Factor (explained in Section 6.8).

Batch Sizes

Different Batch sizes are employed throughout the experiments, depending on the scenario: Batch sizes 1, 2, and 4 are used to explore the impact of varying Batch sizes on model performance, training time, and memory usage. These Batch sizes are chosen to achieve a balance between training efficiency and model generalization, with results depending on the experimental design and research questions.

Early Stopping and Model Checkpoints

To prevent overfitting, early stopping is applied using a validation loss-based approach, where training halts if no improvement is observed over a specified number of epochs. In this case, training is stopped after three consecutive epochs without validation loss improvement. The best model, identified by the highest validation IoU Score, is restored for further use.

Additionally, model checkpoints are saved at the end of each epoch. The best-performing checkpoint, determined by the lowest validation loss or highest validation IoU Score, is retained for evaluation. This approach aligns with practices discussed in Wang et al. [WHSX23], where early stopping and checkpoint selection based on validation loss are highlighted as effective methods to improve generalization while minimizing overfitting.

Dynamic Learning Rate Adjustment

The ReduceLROnPlateau callback dynamically adjusts the learning rate during training. If the validation Loss stagnates for two consecutive epochs, the learning rate is reduced by a factor of 0.2, with a minimum limit of 0.0001. This mechanism ensures a more stable and efficient convergence, as highlighted in Al-Kababji et al. [AKBD22], where the ReduceLROnPlateau scheduler demonstrated faster convergence compared to other methods, achieving competitive results with fewer epochs and better model generalization.

Reproducibility through Seed Management

The entire training process is designed to be fully reproducible. A seed list containing 100 predefined values is used to shuffle the training data in a deterministic manner. The initial seed is set to 7070, ensuring consistency across all random operations. The seed

is reset at the start of each experiment to maintain consistent data sequences across different runs.

Model Output and Evaluation Metrics

The model is evaluated using internal validation metrics, with the validation IoU Score serving as the primary criterion for model selection. Additionally, metrics such as the Dice coefficient of the necrotic (non-contrast-enhancing) core, edema, and contrast-enhancing tumor are computed to provide a detailed breakdown of the model's performance across different tumor subregions.

The model undergoes internal validation during the training phase, where its performance is monitored on a separate validation dataset to ensure generalization and prevent overfitting. This process optimizes hyperparameters, such as the Focal Weight Factor, and utilizes techniques like early stopping to refine the model. After training is completed, the model is tested on the unseen evaluation dataset, which comprises 108 cases (as described in Section 5.2). This final evaluation assesses the model's generalization capabilities on entirely new and therefore unseen data, providing a robust measure of its performance beyond the training and validation datasets.

5.6 Custom Weighted Dice Score for Evaluation

From Human to Predicted Dice Coefficient Weights: Methodology and Justification

This section outlines the methodology used to calculate and justify the weights applied to Dice coefficients for the individual tumor classes in the evaluation of segmentation models. By systematically optimizing these weights, the Custom Weighted Dice Score ensures that model evaluation aligns with both statistical performance and clinical relevance. The following subsections explain the step-by-step process, from statistical metric calculation to the selection and justification of the final Dice coefficient weights.

Calculation of Statistical Metrics

The process began with calculating statistical metrics for various segmentation models, grouped by parameters such as augmentation strategy, Case Group size, Batch size, model version, and Focal Weight Factor. Median values and interquartile ranges (IQR) were computed for the Intersection over Union (IoU) and Dice coefficients of the contrast-enhancing tumor, necrotic (non-contrast-enhancing) core, and edema regions (in descending order of clinical importance). Notably, the Dice coefficient for the background class (no tumor) was not calculated, as it does not provide meaningful insights into the segmentation of tumor regions. These metrics served as robust indicators of model performance across different parameter settings, ensuring reliable comparisons between configurations.

Grid Search for Optimal Weights of Dice Coefficients

A grid search was performed to determine the optimal weights for the Dice coefficients corresponding to the three glioblastoma subregions: contrast-enhancing tumor, necrotic (non-contrast-enhancing) core, and edema. The weights for each region were systematically varied within the range [0,1] in increments of 0.05.

The problem of finding all possible combinations of weights for the three tumor regions (contrast-enhancing tumor w_1, necrotic (non-contrast-enhancing) core w_2, edema w_3), where each weight is within the rage of [0,1], their sum equals 1, and the increments for weights are 0.005, can be modeled using the stars and bars theorem from combinatorics [Sta11]:

$$\text{Total Combinations for } (w_1, w_2, w_3) = \binom{n+k-1}{k-1} \tag{5.1}$$

Here, n represents the number of discrete steps for the total sum ($n = 21$, as the total sum of 1.00 is divided into increments of 0.05) and k is the number of weights (tumor regions, $k = 3$).

$$\binom{21+3-1}{3-1} = \binom{23}{2} = \frac{23 \cdot 22}{2} = 231$$

Thus, there are 231 possible combinations of weights before applying additional constraints. This calculation includes all combinations of weights that could theoretically sum to 1 using the given range and step size. However, after filtering out invalid combinations where the weights do not sum exactly to 1, 194 valid combinations remained for further testing. This systematic exploration of possible weight distributions helps identify, in the next step, the weight combination that maximizes model performance while addressing clinical priorities.

Selection of the Best Dice Coefficient Weights

For each valid weight combination, the top three models with the highest median IoU Scores were selected. A composite score was calculated by combining the Dice coefficients for the contrast-enhancing tumor, necrotic (non-contrast-enhancing) core, and edema regions, weighted according to each tested parameter combination. The optimal Dice coefficient weight configuration was defined as the one that maximized this composite score. This step ensured that the final configuration balanced clinical importance and statistical performance. As discussed in Section 4.1.6, the Custom Weighted Dice Score is employed to evaluate model performance by applying different weights to the tumor subregions, reflecting their clinical importance. The optimal Dice coefficient combination, determined through the grid search process described above, was compared to the class

distribution weights in the ground truth segmentation, as described in Section 5.2.2. This comparison was conducted to select a weight configuration that most closely aligned with the actual class distributions while maintaining robust performance.

Justification for the Final Dice Coefficient Weights

The final Dice coefficient weights—0.45 for the contrast-enhancing tumor, 0.15 for the non-contrast-enhancing tumor, and 0.05 for the edema—were selected to balance clinical relevance and statistical performance. The contrast-enhancing tumor, critical for assessing tumor progression and treatment response, was assigned the highest weight due to its clinical significance. The non-contrast-enhancing tumor, while important, has a secondary role compared to the contrast-enhancing tumor, justifying a moderate weight. The edema, with the least immediate clinical impact, received the lowest weight.

From a performance perspective, the grid search across 194 valid combinations identified this configuration as optimal, achieving a 93.75% match (30 out of 32 configurations) with the ground truth tumor class distribution. This ensures that the model emphasizes the clinically most significant tumor classes while maintaining adequate accuracy for the non-contrast-enhancing tumor and edema, providing a balanced solution aligned with both clinical and statistical priorities.

CHAPTER 6

Implementation

6.1 Pipeline Overview

The glioblastoma segmentation pipeline follows a structured and modular approach, beginning with comprehensive data preprocessing, followed by data augmentation, and concluding with training and testing a deep learning model on multimodal MRI data. The pipeline operates across different environments, ensuring both privacy and efficiency in handling sensitive medical imaging data. The pipeline is illustrated in Figure 1.1.

For the evaluation dataset, the first step of the pipeline is required, which involves the application of the BrainLes pre-processing package [KBW+20]. However, this step is not necessary for the training/validation dataset, as it is already provided in a format compatible with BraTS. In the pipeline diagram (Figure 1.1), this distinction applies only to the initial step; All subsequent steps, including CLAHE and cropping, are identical for all datasets. The specific preprocessing steps required to achieve BraTS compatibility are detailed in Section 6.2.

1. **Data Preprocessing:**

   - **Training/Validation Dataset:** The training and validation datasets are sourced from the BraTS Challenge and are already preprocessed to a standard comparable to BrainLes [KBW+20] output. Thus, no further preprocessing steps like skull-stripping or registration are necessary for these datasets.
   - **Unseen Evaluation Dataset:** The unseen evaluation dataset is processed using the BrainLes Preprocessing Package [KBW+20]. This includes co-registration, skull-stripping, normalization, and conversion to SRI-24 space [RZSP10]. These preprocessing steps align the different imaging modalities (T1, T1 post-contrast, T2, FLAIR), remove irrelevant structures such as the

skull, and standardize intensity values across patients and scans. Preprocessing is conducted locally for privacy reasons, with only skull-removed and non-identifiable data further processed in the cloud.

- **Normalization and NaN Removal:** Before applying CLAHE [Zui94], all datasets (training, validation, and evaluation) are normalized to the [0, 1] range. As part of this process, any NaN values present in the sequences are converted to zeroes. This ensures consistency in intensity values and prepares the data for contrast enhancement.
- **CLAHE (Contrast Limited Adaptive Histogram Equalization):** CLAHE is applied to all datasets, including the training, validation, and evaluation datasets. This step enhances the contrast of the MRI images, improving the visibility of tumor boundaries. CLAHE is applied to each 2D slice of the 3D image volumes, with voxel intensities normalized to the [0, 1] range. Optimal parameters for CLAHE, such as Clip Limit and Tile Grid Size, are selected to balance contrast enhancement and image quality preservation.
- **Cropping:** After CLAHE, the MRI images are cropped to remove irrelevant areas outside the brain. This reduces the overall dataset size and optimizes computational efficiency, focusing on the brain and tumor regions.

2. **Data Augmentation:**

   Augmentations are applied to the preprocessed and cropped training data to increase the dataset size and improve model generalization. These augmentations include transformations such as rotations and flips. To ensure reproducibility and minimize runtime overhead, the augmented images are pre-generated and stored as NumPy arrays.

3. **Storage of Processed Data:**

   Before the 3D U-Net is trained, both the training/validation datasets and the augmentations are saved as NumPy arrays. This improves the efficiency of data loading during the training process and ensures consistency across experiments.

4. **Model Training:**

   A 3D U-Net architecture, implemented using Python [VR09] with TensorFlow [Dev16] and Keras [Cho15], is trained on the preprocessed, augmented, and stored MRI data. The model uses ground truth segmentations provided by a trained neuroradiologist, and class imbalances are addressed using a combination of Dice Loss and Focal Loss. Hyperparameter tuning is conducted to optimize the focal weight, and mixed precision training is used to improve computational efficiency.

5. **Model Testing:**

   After training, the model is evaluated using key metrics: Intersection over Union (IoU), Dice coefficients for the three segmentation classes (necrotic/non-contrast-enhancing tumor, edema, contrast-enhancing tumor), accuracy, and the Custom Weighted Dice Score. These metrics provide a comprehensive evaluation of the model's performance in segmenting the tumor regions.

6. **Experimental Testing Environment:**

   The pipeline is executed in a hybrid environment. Preprocessing steps using the BrainLes Preprocessing Package [KBW+20] are performed locally to maintain data privacy. Only skull-removed, non-identifiable MRI data are uploaded to Google Colab [Col23], where the model is trained and tested using an NVIDIA A100 GPU. This setup ensures compliance with data protection regulations while providing the necessary computational resources for the deep learning tasks.

6.2 Evaluation-Dataset Preprocessing

The preparation of medical image data for the evaluation dataset follows the preprocessing protocol of the BraTS datasets, with the objective of ensuring an optimal transfer of the image characteristics learned from the training to the evaluation dataset in the context of machine learning. This ensures that the evaluation dataset is processed in a manner consistent with the training data, facilitating a fair comparison between training and evaluation results.

Initially, the DICOM data is converted into the NIfTI format (Neuroimaging Informatics Technology Initiative). Following this, crucial preprocessing steps are applied, including co-registration, skull-stripping, normalization, and conversion to the SRI-24 space [RZSP10]. These steps are essential for maintaining data consistency and enhancing the accuracy of the segmentation model.

- **Co-registration:** All imaging modalities of a patient (e.g., T1, T1 post-contrast, T2, FLAIR) are co-registered to ensure that they are aligned within the same spatial coordinates. This alignment ensures that anatomical structures appear in the same location across all modalities, which is essential for accurate segmentation.
- **Normalization:** Intensity normalization is applied to adjust the intensity values across all modalities to a common scale. This reduces variability between patients and across imaging sessions caused by differences in scanner settings or patient positioning. Standardizing intensity values enhances the segmentation model's ability to consistently analyze brain tissues and improves the accuracy of the segmentation results.

- **Skull-stripping:** Skull-stripping removes non-brain structures, particularly the skull, from the MRI images. This is an important step to ensure that the segmentation algorithm focuses exclusively on brain tissues and tumor regions, without interference from surrounding structures.
- **Conversion into SRI-24 space:** The MRI images are converted into the SRI-24 space [RZSP10], a standardized anatomical template based on 24 normal brain scans developed by the Stanford Research Institute (SRI). This space provides a reference framework that facilitates spatial normalization of brain images, aligning different datasets to a common coordinate system. By using the SRI-24 space, variations introduced by differences in scanners, imaging protocols, or patient anatomy are minimized, ensuring comparability of segmentation results across datasets. This consistency is especially important for reproducibility in multi-center studies and the evaluation of segmentation algorithms in challenges like BraTS.

These preprocessing steps ensure that the evaluation dataset is aligned, normalized, and homogenized in terms of intensity, enabling more consistent analysis across the entire dataset. The BrainLes preprocessing package from the BraTS Toolkit [KBW+20] is used to perform these tasks, offering the advantage of automatically converting the image data into the SRI-24 space [RZSP10], which is expected by the BraTS challenge algorithms.

To achieve this, the following preprocessing steps are performed:

- **Registration and correction** are performed using NiftyReg [ORS+01], a software package that aligns the scans with a reference anatomical atlas (SRI-24) using rigid and non-rigid transformations.
- **Normalization** is done using a percentile normalizer, adjusting the intensity values of MRI scans based on percentile thresholds to reduce the impact of outliers and ensure the dataset is on a consistent scale.
- **Skull-stripping** is automated using the High-definition Brain Extraction Tool (HD-BET) [ISP+19], a deep learning-based tool designed specifically for removing non-brain tissues from MRI images, streamlining the skull-stripping process.

Although the BrainLes Preprocessing Package [KBW+20] performs foundational steps such as intensity standardization and alignment, additional refinements are necessary to ensure optimal data consistency and reliability for segmentation. These refinements include further normalization and handling of NaN values, which prepare the datasets for downstream processing and enhance their compatibility with subsequent steps like CLAHE.

Normalization and NaN Removal

Before further processing, including the application of CLAHE, all datasets (training, validation, and evaluation) are normalized to the range [0, 1]. This normalization step

is crucial to ensure that the intensity values of the MRI images are consistent across different modalities and patient datasets. Intensity normalization is a widely established preprocessing step in medical image analysis, particularly for segmentation tasks involving MRI images. Since MRI image intensities are not inherently tissue-specific, they vary significantly between scanners, protocols, and even scans of the same patient. Normalizing the intensity range facilitates consistency across datasets and reduces scanner-induced intensity variability, thus improving model robustness and segmentation performance [PNA18].

To ensure data consistency, any NaN values that may appear during preprocessing are converted to zeros. Although a detailed review of the literature did not explicitly identify the NaN issue encountered in this study, related works provide a plausible explanation for its occurrence. For instance, interpolation and resampling during the co-registration of multimodal MRI datasets can introduce artifacts, particularly at the boundaries of imaging regions, due to phenomena such as the Gibbs effect. These artifacts may result in voxel intensities becoming inconsistent or undefined in specific areas [PCCDM21, MCV+97].

To mitigate this issue, converting NaN values to zeros was incorporated as a preprocessing step. This approach ensures dataset integrity and prevents unpredictable behavior during the training process. Unresolved NaN values can propagate errors that affect gradient computations, leading to unstable model convergence. By replacing NaN values with zeros, the risk of unstable gradients is reduced, facilitating a more stable training process and ensuring that each voxel consistently contributes to model learning [PNA18].

Normalization is essential not only for preparing the data for downstream tasks, but also plays a significant role in improving the stability and performance of the segmentation model. By standardizing the intensity ranges, the model can better differentiate between different brain structures and tumor regions, minimizing the impact of scanner variability or image noise [Jan15, PNA18]. Furthermore, intensity normalization allows the model to generalize more effectively across datasets from different sources. Studies have shown that intensity normalization enhances segmentation performance by reducing variability in intensity distributions, which can otherwise hinder segmentation accuracy [Jan15]. Once the data is normalized and NaN-free, it is ready for the next step in the pipeline.

6.3 CLAHE (Contrast Limited Adaptive Histogram Equalization)

(CLAHE) represents a significant advancement in the field of image processing, addressing limitations found in traditional histogram equalization methods. The algorithm was refined and popularized by Karel Zuiderveld [Zui94], who detailed it in "Graphics Gems IV". However, the foundational concepts of adaptive histogram equalization were explored earlier by researchers such as Pizer et al. [PAA+87].

Initially, CLAHE [Zui94] was developed and used for medical imaging purposes. It played a decisive role in enhancing the visibility of anatomical structures in CT scans and MRI images. This improved visualization aids in the more accurate diagnosis and analysis of medical professionals [YMI+24]. The method's ability to enhance local contrast while maintaining the overall image quality makes it particularly suitable for medical applications where detail and clarity are paramount. The method works by dividing the image into small tiles, applying histogram equalization to each tile, and then combining the tiles using bilinear interpolation to remove artificial boundaries.

The implemented algorithm operates by generating multiple histograms for different sections of an image and using these histograms to redistribute the brightness values. Two primary parameters must be set for CLAHE: Clip-Limit (CL) and Tile Grid Size or Block size (BS).

Parameters for CLAHE

The Clip Limit (CL) defines the maximum threshold for contrast adjustments, preventing excessive enhancement that could amplify noise. Higher clip limit values increase contrast, with typical values ranging from 2.0 to 4.0. The Tile Grid Size or Block Size (BS) specifies how the image is divided into smaller regions for localized contrast adjustment. Each tile undergoes individual enhancement, with typical sizes set to (8, 8) or (16, 16). Together, these parameters ensure effective contrast optimization while preserving image quality.

Determining Optimal Parameters

To determine the optimal parameters for CLAHE, various combinations of the clip limit and tile grid size can be tested and evaluated visually and statistically. By evaluating the histograms and images generated with various parameter combinations, the parameters that yield a more uniform and consistent voxel intensity distribution can be identified. For example, with clipLimits = [2.0, 3.0, 4.0] and tileGridSizes = [(8, 8), (16, 16)], six different combinations result.

Programmatic Implementation

The function `apply_clahe_3D` applies CLAHE to each 2D slice of a 3D image, scaling voxel intensities before and after processing to ensure that they remain within the range [0, 1]. After applying CLAHE, the data are rescaled to this range and converted to float32, making the normalized data ready for further processing in TensorFlow [Dev16].

However, the provided function does not fully take advantage of the 3D structure of the data, which a true 3D CLAHE implementation would do. True 3D CLAHE could provide more consistent and uniform contrast enhancement across the entire volume by considering the spatial context in all three dimensions, which was first described by Amorim et al. [AFdMSP18].

Despite the promising approach of 3D CLAHE, the decision was made to use 2D CLAHE instead. This decision is driven by the need for a more manageable and computationally efficient solution. The 2D CLAHE provides a good balance between enhancing image

contrast and maintaining a practical level of complexity and resource usage. Additionally, the implementation of 3D CLAHE introduced stability issues, making it difficult to achieve consistent results. Therefore, 2D CLAHE was chosen as the more suitable option for the current implementation.

Optimal Parameter Selection Based on Histogram Metrics

To select the optimal parameter combination based on histogram metrics, various characteristics of the histogram are analyzed. The three primary metrics considered are entropy, kurtosis, and skewness. Each of these measures provides distinct insights into the nature of voxel intensity distributions and their impact on image quality and interpretability.

Entropy serves as a measure of the complexity of the histogram, reflecting the degree of unpredictability or randomness in voxel intensities [NWS14]. A higher entropy value signifies a more uniform distribution of intensities, indicating better contrast distribution across the intensity range. This is particularly useful in medical imaging, where a higher entropy can enhance the visibility of subtle differences between tissue types. In practical applications, maximizing entropy has been shown to enhance image details and improve diagnostic precision.

Kurtosis quantifies the peakedness or flatness of the histogram relative to a normal distribution [HSF24]. In medical imaging, a flatter distribution is desirable as it promotes a more even spread of voxel intensities, reducing the likelihood of over-saturation or underexposure in specific intensity ranges. This balance ensures that subtle anatomical details are preserved, enhancing the diagnostic utility of the images. A lower kurtosis value indicates a flatter distribution, where voxel intensities are more evenly spread across the range, reducing the occurrence of overly bright or dark regions. This contributes to a more natural and realistic appearance of the image, improving its interpretability. In the context of medical imaging, controlling kurtosis helps to avoid over-saturation or underexposure of specific intensity ranges, thereby preserving the visual quality of the image.

Skewness measures the asymmetry of the histogram, reflecting how voxel intensities are distributed around the mean [VARS24]. A skewness value close to zero indicates a balanced distribution, while positive or negative skewness suggests a bias towards higher or lower intensity values, respectively. In medical imaging, a skewness value near zero ensures a balanced intensity distribution, maintaining an equilibrium between light and dark regions in the image. This balance is essential for accurate interpretation and analysis, as it avoids dominance of certain intensity ranges.

To combine these three metrics into a single quantitative measure, a score is calculated that aims to maximize entropy while simultaneously minimizing kurtosis and the absolute value of skewness. The score is defined as follows:

$$\text{Score} = \text{Entropy} - \text{Kurtosis} - |\text{Skewness}| \tag{6.1}$$

This formula reflects a logical approach to optimizing image quality. By maximizing entropy, the goal is to achieve a diverse range of voxel intensities, ensuring that subtle

features are not lost. Simultaneously, reducing kurtosis prevents intensity values from clustering too tightly, and minimizing the absolute value of skewness ensures a balanced intensity distribution. The combination of these three objectives promotes an optimal balance between contrast, brightness, and symmetry in the resulting images.

The rationale behind this formulation lies in the physical significance of each metric in image processing [HSF24]. Entropy is widely recognized as a measure of information content, and its maximization aligns with the goal of preserving as much detail as possible. Kurtosis and skewness, on the other hand, control the shape and symmetry of the intensity distribution. By minimizing kurtosis, the aim is to distribute voxel intensities more evenly, while the minimization of skewness ensures that no specific intensity range dominates the image. The combined effect of these adjustments results in images that are visually balanced, have enhanced contrast, and preserve crucial anatomical details.

To identify the optimal parameter combination, the score is calculated for each of the six parameter configurations. The combination yielding the highest score is selected as the optimal choice. This approach ensures that the selected parameters produce an image with the most favorable balance of contrast, brightness, and symmetry, as shown in Figure 6.1.

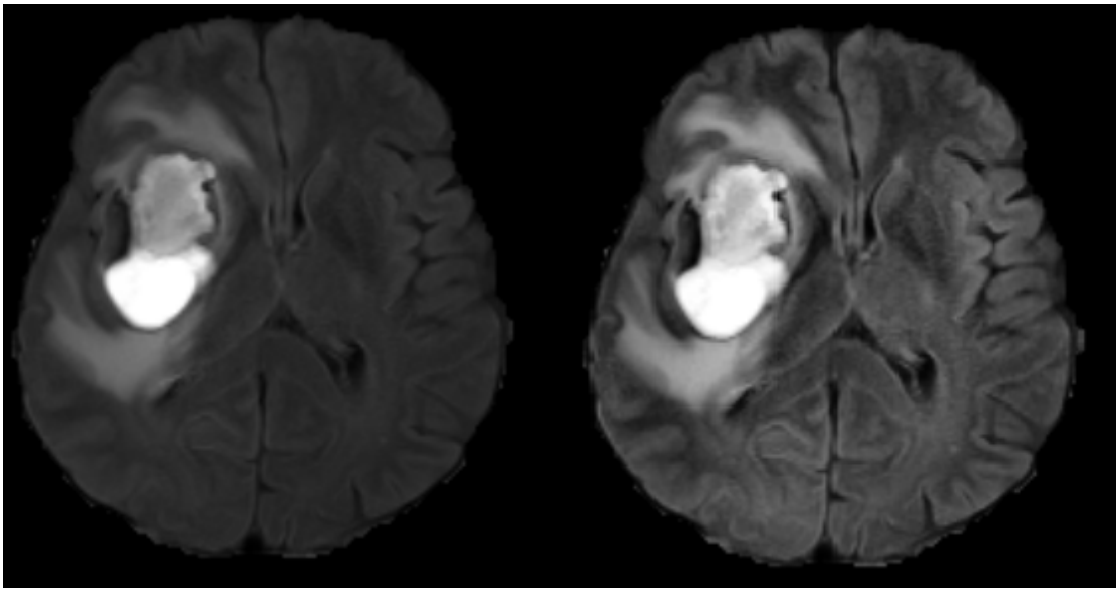


Figure 6.1: Comparison of FLAIR MRI images before (left) and after (right) applying Contrast Limited Adaptive Histogram Equalization (CLAHE). The CLAHE-enhanced image shows significantly improved local contrast, enhancing the visibility of the tumor and surrounding structures. The histogram transformation of the FLAIR sequence is illustrated in the plots of Figure 6.2.

The stripes observed in the right histogram in Figure 6.2 after the CLAHE transformation occur due to the local contrast enhancement applied to each tile. CLAHE divides the

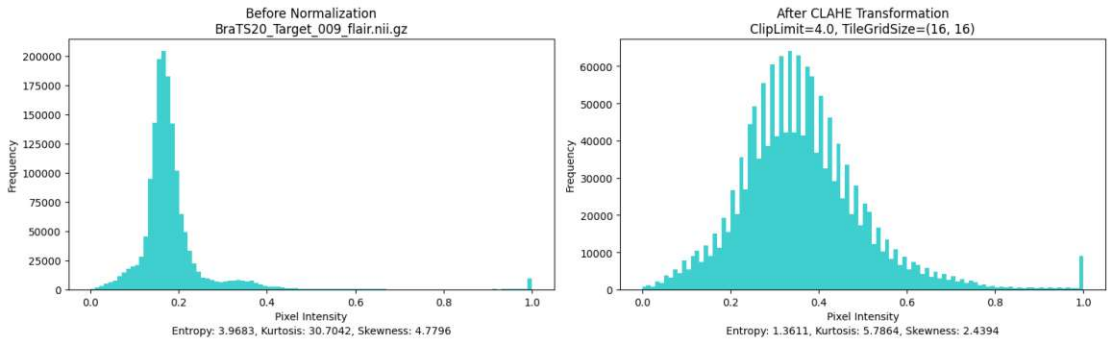


Figure 6.2: Histograms of voxel intensity before (left) and after (right) CLAHE transformation for the FLAIR sequence in Figure 6.1. The CLAHE-enhanced histogram shows improved contrast distribution, with reduced entropy and skewness, indicating better equalization and visibility of image details.

image into smaller regions (tiles), applies histogram equalization to each, and then combines them. This process can introduce slight discontinuities at the borders of the tiles, resulting in a striped appearance in the histogram. These stripes reflect the boundaries between tiles with varying contrast levels, as each tile is adjusted independently before being merged back into the overall image. However, these discontinuities usually do not affect the overall quality and utility of the image for diagnostic purposes. If necessary, additional post-processing steps could be applied to smooth out these transitions and minimize the visual impact of the stripes.

The study by Yoshimi et al. [YMI+24] demonstrates that preprocessing with CLAHE significantly enhances the performance of deep learning models for segmenting MRI images. Specifically, the application of CLAHE resulted in higher values for metrics such as the Dice similarity coefficient, sensitivity, and positive predictive value when compared to models trained on non-preprocessed images. This improvement underlines the effectiveness of CLAHE in dealing with low contrast images and varying brightness levels, which are common in medical imaging. The results suggest that CLAHE is a robust preprocessing method to boost the performance of deep learning models in medical image analysis.

Alternatives to the CLAHE Method

While CLAHE is a widely used approach for local contrast enhancement, other image processing techniques, such as the Wavelet Transform, have been considered in the field. However, the Wavelet Transform exhibits several limitations in this context, as it does not directly enhance contrast. Instead, it performs a multi-scale decomposition of the image into frequency components, separating low-frequency (coarse) structures from high-frequency (fine-detail) information [HMAZ23]. This decomposition allows for operations like noise suppression and feature extraction, which are valuable in certain image processing tasks but not directly applicable for localized contrast enhancement [VKG+22, HMAZ23].

The indirect influence of the Wavelet Transform on contrast is achieved through the manipulation of high-frequency components, often enhancing edges or emphasizing fine details [Pyk17]. However, this process does not involve a direct adjustment of voxel intensity distributions, which is essential for localized contrast enhancement. Unlike methods that directly redistribute intensity values within image regions, the Wavelet Transform focuses on frequency-based separation, making it better suited for noise reduction and feature analysis [HMAZ23, VKG+22].

Given the objective of this work — to achieve local contrast enhancement rather than frequency decomposition or noise suppression — the Wavelet Transform was not a suitable option. Its reliance on frequency-based operations does not meet the requirement of dynamically adjusting voxel intensities in small image regions. Consequently, the CLAHE method was selected as it is specifically designed to achieve this localized enhancement, ensuring better visibility of subtle image details [GBGC18].

6.4 Data Cropping & Resource Utilization

After skull-stripping, a significant portion of the images, particularly in areas near the vertex (high frontal) and the nose (ventro-frontal), consists of non-informative data. This is especially evident after removing the skull, leaving large empty regions in the images. These black regions, visible in the images, result from the removal of the skull and soft tissue using the BrainLes Preprocessing Framework, as shown in Figure 6.3. By focusing on these non-informative regions and reducing the overall data size through cropping, we can significantly improve resource utilization. This optimization not only decreases memory usage but also enhances processing speeds, reduces storage requirements, and streamlines data handling and analysis.

The BrainLes Preprocessing Package [KBW+20] used earlier in the pipeline outputs images with a resolution of 240x240x155, equating to 8,928,000 voxels per image. By applying cropping techniques, we reduce the dimensions to 128x160x128, resulting in only 2,621,440 voxels. This adjustment leads to a 70% reduction in data size, a significant improvement that directly impacts computational efficiency, as illustrated in Figure 6.3.

The specific cropping dimensions were carefully selected to retain the most relevant brain structures while eliminating non-informative regions. In this process, some parts of the brain surface, particularly in the bifrontal region, lose a few voxels due to the tightly selected cropping window. Despite this minimal loss, the method preserves the anatomically relevant structures, ensuring the accuracy of downstream analyses while significantly improving computational efficiency. The images were cropped from slice 55 to slice 183 along the frontal (coronal) axis, from slice 47 to slice 207 along the sagittal axis, and from slice 10 to slice 138 along the cranio-caudal axis. This asymmetric cropping reduces the data size while preserving crucial anatomical features, ensuring that no important information is lost.

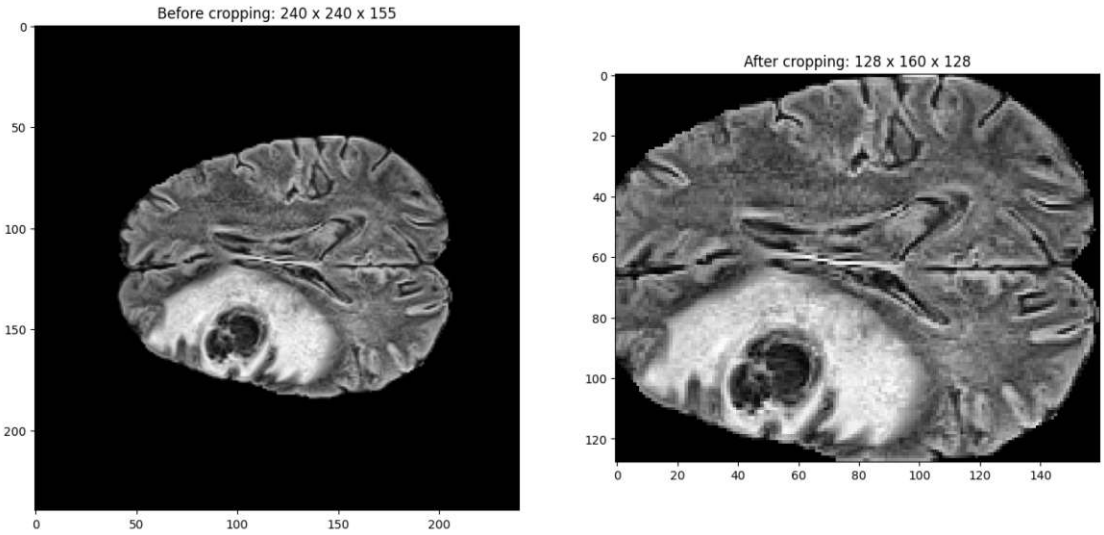


Figure 6.3: FLAIR MRI sequence before (left) and after (right) applying cropping. The initial resolution is 240x240x155 voxels, reduced to 128x160x128 voxels after cropping, significantly decreasing data size and improving computational efficiency.

Asymmetric Volume Alignment

The asymmetric alignment of the preprocessed volume is due to the need to avoid cutting off the nose during image acquisition. After the skull is removed from the image, the resulting empty space can be cropped out to focus on the brain. Since the human head is typically longer than it is wide, the cropping dimensions of 128x160x128 are chosen to reflect this natural asymmetry. Although this approach may result in the loss of some peripheral information, the trade-off is deemed reasonable given the substantial reduction in data size and the increased computational efficiency.

Dimensional Requirements for Deep Learning Models

Another critical factor to consider when cropping images for deep learning models, such as the 3D U-Net architecture, is that the image dimensions should be powers of two. This is essential because, in the contraction path of U-Net, the image resolution is halved at each downsampling layer. Consequently, the image dimensions must be divisible by two to allow smooth downsampling. The specific dimensions chosen—128 (which can be expressed as 7 x 2^7, allowing for 7 layers) and 160 (which can be expressed as 5 x 2^5, allowing for 5 layers)—reflect the depth to which the network can operate efficiently. The higher the factor of two, the deeper the network can become, enabling more layers without requiring complex architectural adjustments.

The reduction in data size has a direct impact on the utilization of computing resources. By minimizing the amount of data processed in each training iteration, memory usage is reduced, which lowers the risk of memory overflow during training. Smaller data sizes also accelerate processing times, improving overall computational efficiency. Storage

requirements are also minimized, which is a critical factor when managing large medical imaging datasets. Additionally, the reduced data size facilitates faster data transmission and analysis, enabling quicker results and more efficient project execution.

6.5 Augmentation Implementation

Geometric transformations, such as rotations and flips, are applied to expand the dataset, introducing variability that enhances the model's generalization capabilities. These transformations are particularly effective for medical image segmentation, as they simulate plausible variations while preserving anatomical structures.

Using the `skimage.transform` package, each FLAIR, T1ce, T2 image, and the corresponding segmentation masks can be rotated randomly between -10° and 10° along the x, y, or z axis. The rotation angle (between -10° and 10°) and axis (x, y, or z) are selected at random to introduce variability while preserving anatomical structures. Segmentation masks are rotated using order 0 interpolation to maintain the integrity of labeled regions. Additionally, these images and segmentation masks are flipped along a randomly chosen axis (x, y, or z) using `numpy.flip`. This simple yet effective augmentation technique duplicates the dataset, effectively increasing its size.

The function `augment_data` randomly selects one of three augmentation operations —rotation only, flip only, or a combination of both—and applies it to the FLAIR, T1ce, T2 images, and the segmentation masks. In the first scenario, the images are rotated by a random angle between -10° and 10° along a randomly selected axis (x, y, or z). In the second option, the images are flipped along a randomly chosen axis (x, y, or z). Finally, the third option combines both transformations, where a random rotation is followed by a random flip. This random selection strategy guarantees an equal probability for each augmentation type, achieving balanced dataset enhancement and preventing overfitting to specific transformations.

Figure 6.4 illustrates the application of these augmentation techniques. The first row displays the original images with the corresponding segmentation mask, the subsequent rows display augmented versions of the original images. The second row demonstrates flipping along the x-axis. The third row shows an example of rotation by -4.02 degrees along the z-axis combined with flipping along the y-axis, where the effect of rotation is visible in the altered position of the green area (non-contrast-enhancing tumor). In the fourth row, a flip along the z-axis is applied, combined with a rotation of 3.28 degrees along the y-axis, resulting in notable changes in the segmentation mask. These augmentations enhance the dataset's variability while preserving essential anatomical structures, thereby improving the robustness and generalizability of the deep-learning model.

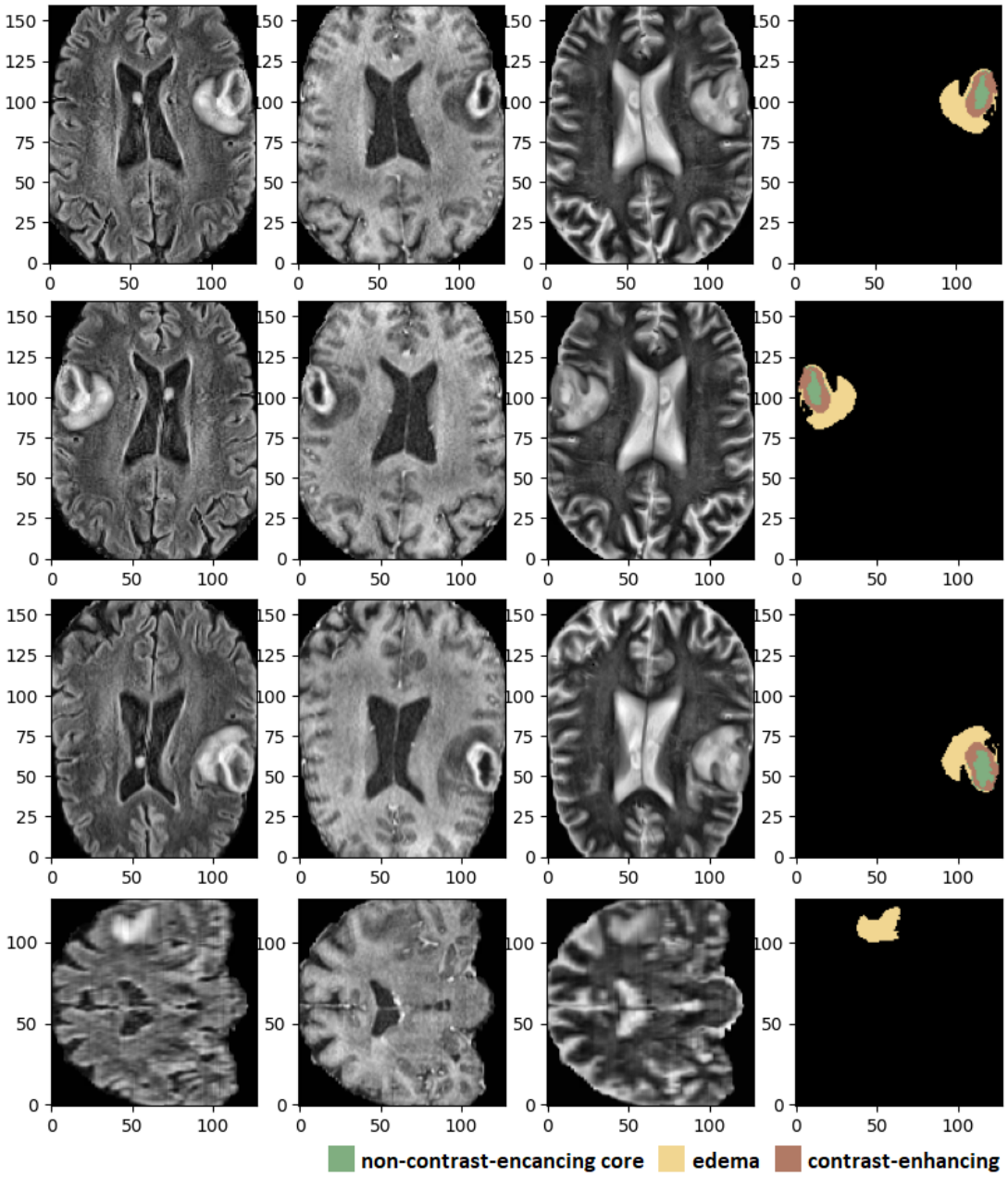


Figure 6.4: Example of Original and Augmented MRI Images with Corresponding Segmentations. The first row displays the original images, with columns from left to right showing FLAIR, T1ce, T2, and the segmentation, which includes non-contrast-enhancing tumor (green), tumor edema (yellow), and contrast-enhancing tumor (brown). Subsequent rows present augmented images with flipping, rotation, or a combination of both, enhancing dataset variability while preserving anatomical structures.

6.6 Optimized Data Storage & Sequence Selection

To further optimize data loading and reduce latency, three of the four imaging sequences—T2, FLAIR, and T1ce—are stacked into a single NumPy array. This approach minimizes the time required to load the data into memory during training and ensures that the images are readily available for processing. The native T1 sequence is omitted from the stack because it does not provide significant additional information beyond what is already captured by the other sequences. Menze et al. [MJB+15] and Bakas et al. [BRJ+19] explicitly emphasize that the primary focus for segmentation models lies on T1ce, T2, and FLAIR, as they provide the most relevant information for delineating tumor subregions. The native T1, in contrast, does not reveal additional features for identifying non-contrast-enhancing tumor, contrast-enhancing tumor, or peritumoral edema, as these structures are better visualized with the existing sequences. This redundancy supports the decision to exclude the native T1 sequence, as its contribution to segmentation accuracy is marginal. By excluding redundant data, the preprocessing pipeline is streamlined without compromising the quality or completeness of the diagnostic information.

An additional reason for excluding the native T1 sequence is its inconsistency in appearing as either a Turbo Spin Echo (TSE) or Fast Field Echo (FFE) sequence, which can introduce variability into the dataset. This inconsistency could potentially affect the model's ability to converge during training, as the network may struggle to generalize across differing sequence types. As noted in Section 5.2, this variability in the native T1 sequence can lead to unwanted impacts on the convergence of the network, making it less reliable for the task at hand. Similar issues have been observed in multi-center MRI datasets, where variations in acquisition protocols, such as pulse sequences and scanner parameters, cause significant changes in image contrast. This variability hinders the generalization ability of convolutional neural networks, as they tend to overfit to specific contrast distributions from the training data [JHG+19]. Therefore, its exclusion is justified to maintain the overall consistency and robustness of the training data.

6.7 3D U-Net Design

This implementation of the 3D U-Net for glioblastoma segmentation combines key principles of the U-Net architecture with tailored modifications for handling volumetric MRI data. The model utilizes increased filter depths, dropout regularization, and skip connections to preserve critical spatial information, making it highly effective for segmenting complex tumor structures in 3D.

Contracting Path (Encoder)
The contracting path is designed to capture the context of the input MRI scans by progressively downsampling the input images while increasing the depth of the feature maps. Key characteristics of this implementation are:

- **Convolutional Layers:** At each stage, two 3D convolutional layers are applied with ReLU activations, followed by a dropout layer to prevent overfitting. The filters at each layer follow an increasing pattern (16, 32, 64, 128, and 256), allowing the network to progressively capture more complex and abstract features.
- **Dropout:** Dropout layers are employed after each convolutional block, with dropout rates increasing from 0.1 to 0.3 as the depth of the network increases. This strategy helps reduce overfitting by randomly dropping units during training [SHK+14].
- **Max Pooling:** The spatial dimensions of the input image are reduced using 3D max pooling with a pool size of (2, 2, 2). This reduces computational complexity while retaining essential spatial information.

Expanding Path (Decoder)
The expanding path reconstructs the high-resolution segmentation map by progressively upsampling the feature maps. Key characteristics include:

- **Transposed Convolutions:** Conv3DTranspose layers are used to upsample the feature maps, doubling the spatial resolution at each step. This is essential for reconstructing the segmentation map.
- **Skip Connections:** Skip connections between corresponding layers in the contracting and expanding paths are critical for preserving high-resolution information. These connections ensure that fine-grained spatial details are preserved, even after the aggressive downsampling in the encoder.

Skip Connections
Skip connections are a hallmark of the U-Net architecture and are employed between layers of corresponding spatial dimensions in the encoder and decoder. These connections are crucial for combining the abstracted high-level features from the contracting path with the fine-grained spatial information from earlier layers. By concatenating feature maps from the encoder with those in the decoder, the model retains detailed spatial information critical for accurate tumor segmentation [SZL+24, TB24].

Output Layer
The final layer of the network is a Conv3D layer with a kernel size of (1, 1, 1) and a softmax activation. This outputs voxel-wise class probabilities across the four defined segmentation classes: non-contrast-enhancing tumor, contrast-enhancing tumor, edema, and background.

Model Input and Output

- **Input Shape:** The input to the network consists of 3D volumes with a shape of (128, 160, 128), representing the downsampled MRI images. The input includes three MRI modalities (T1ce, FLAIR, and T2), providing the necessary information for the segmentation task.

- **Number of Classes:** The model outputs probabilities for four classes, including non-contrast-enhancing tumor, contrast-enhancing tumor, edema, and background.

Optimization for Efficient Training

- **Dropout Regularization**: Dropout layers are strategically placed in both the contracting and expanding paths to reduce overfitting. As the network becomes deeper, the dropout rates increase to introduce more regularization.

- **Batch Normalization:** While not included in this specific implementation, Batch Normalization could be introduced between convolutional layers to normalize the activations and potentially speed up training convergence.

Batch Normalization was deliberately not included in this 3D U-Net implementation due to its limited effectiveness in 3D convolutional networks, particularly if small Batch sizes are used. As highlighted by Kolarik et al. [KBR20], Batch Normalization requires sufficiently large Batch sizes to compute reliable statistics for mean and variance. If the Batch size is reduced to 1, the computed variance becomes zero, leading to numerical instability and unreliable parameter updates. This issue is particularly relevant for 3D segmentation networks, where the larger memory demands of 3D images often necessitate smaller Batch sizes [KBR20]. Consequently, using Batch Normalization in this setting can degrade the convergence behavior and hinder the stability of the optimization process.

Furthermore, Batch Normalization introduces additional memory overhead, which is particularly costly in 3D networks due to their high-dimensional input. Qin et al. [QGT+19] note that smaller Batch sizes increase the overhead for computing global statistics across batches, resulting in longer training times and increased memory consumption. Given the memory-intensive nature of 3D image segmentation, avoiding Batch Normalization helped to maintain efficient GPU usage. Additionally, the benefits of Batch Normalization, such as smoothing the optimization landscape, can be achieved through alternative approaches such as data augmentation and dropout, both of which are less dependent on Batch size [STIM19]. These alternative strategies were used to stabilize the training without introducing the computational overhead associated with Batch Normalization.

6.8 Reproducibility via Seed Management

The initial seed is set to 7070 and is consistently applied using TensorFlow's [Dev16] random seed mechanism to ensure reproducibility in operations such as weight initialization and data shuffling. Additionally, a Seed List with 100 predefined values is used to control randomization during each epoch, ensuring that the order in which training and validation cases are loaded remains consistent across experiments.

At the beginning of each epoch, a new seed is selected from the Seed List, which determines the specific order in which the original, validation, and optionally augmented data are loaded. Each seed ensures a consistent and reproducible sequence of data loading for that particular seed. For example, Seed 860 results in a different data order than Seed 5390, but for any given seed, the same order is always used. This mechanism guarantees that the data is shuffled in a specific and repeatable way for each seed, making the training process reproducible across experiments. Since augmentations are pre-generated and stored separately, seed management only affects the order of data loading, not the generation of augmentations.

Limitations of Deterministic Data Shuffling and Justification

One potential limitation of using a deterministic shuffling approach through predefined seeds is the risk of overfitting to specific data patterns that occur early in the training sequence. Since the training and validation data are always presented in the same order, the model might learn to rely on these early patterns, which can negatively impact generalization performance. Such effects are observed in deep reinforcement learning (DRL), where overfitting occurs "robustly", even in the presence of added stochasticity [ZVMB18]. Moreover, the fixed sequence of data presentation reduces the exposure to random variability, which can further hinder generalization to unseen data [ZVMB18].

Despite these potential drawbacks, the decision to use deterministic shuffling was made to ensure reproducibility and fair model comparison. In deep learning research, reproducibility is a critical factor, especially if training models with stochastic processes. Deterministic implementations help eliminate the variance caused by nondeterminism, which is a known issue in DRL [NWS18]. By fixing the order of data presentation, it becomes possible to consistently compare the performance of models with different hyperparameters under identical conditions [NWS18]. This consistency is crucial if evaluating multiple models (e.g., variations in Batch size, augmentation strategies, or Focal Weight Factors) to ensure that observed differences in performance are solely attributable to model configurations and not to random fluctuations in data presentation [NWS18].

The use of deterministic shuffling does not necessarily limit generalization. The variability of the training data is still ensured through data augmentations, which generate unique transformations for each epoch. This is supported by findings on data augmentation, where random transformations (e.g., rotation, flipping, and intensity shifts) introduce diversity into the training data, thereby mitigating the risk of overfitting [SK19]. In this context, the fixed shuffling order does not reduce variability, as the image content changes

with each epoch due to augmentations, even if the image indices remain constant [SK19].

To address the potential risk of overfitting to specific patterns, the training set was designed to be diverse and representative of the underlying data distribution. Additionally, augmentations such as rotation, flips, and intensity shifts were applied to increase variability during training [SK19]. This approach ensures that the model does not rely on static patterns in the data order but instead learns from a wide range of transformations [SK19].

In summary, while deterministic shuffling has potential limitations regarding generalization, its use is justified by the need for reproducibility and fair model comparison. Deterministic implementations play a crucial role in achieving reproducibility, as they eliminate nuisance noise and ensure consistent experimental conditions, which is particularly relevant in fields like deep reinforcement learning [NWS18]. The presence of data augmentations further mitigates the risk of overfitting to a fixed data order, as the model is exposed to variable image transformations across epochs [SK19]. This balance between reproducibility and generalization aligns with best practices for deep learning experimentation, especially if evaluating numerous model configurations.

DataGenerator: Seed Management for Reproducible Data Shuffling

The DataGenerator is essential for loading both original and pre-generated augmented data during training. To ensure reproducibility, the Seed List is passed to the DataGenerator, which controls the shuffling of data at the start of each epoch by selecting a seed. This guarantees consistent data ordering (including training, validation, and augmented images) across different runs of the experiment.

Since augmentations are pre-generated, seed management strictly governs the data shuffling, ensuring deterministic loading sequences while allowing variability in the order across epochs, and maintaining reproducibility throughout the process.

Callback Management for Seed Control

In addition to the DataGenerator, callbacks like ModelCheckpoint and EarlyStopping are employed to manage checkpoints and optimize training while ensuring that the random seeds are reset correctly at each epoch. The CustomSeedCallback ensures that the seed is reset at the beginning of each epoch, maintaining consistency in the shuffling of the training and validation data, and keeping the overall process deterministic.

The callback is triggered at the beginning of each epoch, ensuring that the training data—including both the original and pre-generated augmented images—are reshuffled based on the selected seed. This consistency ensures that every experiment remains fully reproducible while still introducing some variability by changing the order in which the data is presented across epochs.

Pre-Generated Augmentations for Reproducibility

As explained in Section 5.4, augmentations are created with random rotations and flips to increase variability in the training process and stored as NumPy arrays. These pre-

generated augmentations, along with the training and validation datasets, are loaded using the seed management system. This ensures that the data is consistently shuffled and loaded in the same deterministic order across training epochs, allowing for precise comparisons between different training runs. By controlling the random seed, the augmented data is shuffled and loaded in the same reproducible (deterministic) order as the training and validation data for each experiment.

Impact on Model Training

By carefully managing seed values and utilizing pre-generated augmentations, the model training process is made fully reproducible. This ensures that each experiment can be replicated under identical conditions, allowing for consistent comparisons across different Focal Weight Factor configurations. Seed management controls data shuffling and augmentation, ensuring that randomness remains controlled, leading to reliable and comparable results across experiments. This reproducibility is crucial for validating the model's performance in scientific research.

6.9 Model Training & Validation

Data Loading and Augmentation Customized data generators manage the loading and augmentation of large 3D medical images. The **DataGenerator class** handles loading original and augmented data from predefined directories. The dimensions of the images are set to 128x160x128 (depth, width, height) with three channels representing T1ce, FLAIR, and T2 MRI sequences. Augmentations, including rotations and flips, are pre-generated to reduce computational load during training, and seed management ensures consistent shuffling.

Key aspects:

- **Batch Size Flexibility:** Depending on the experiment, Batch sizes of 1, 2, or 4 are employed to explore their impact on segmentation accuracy and training efficiency. The data generator dynamically adapts to these variations.
- **Shuffling and Seed Management:** A seed list with 100 predefined values ensures consistent shuffling of training and validation data across experiments, using a seed value (e.g., 7070) applied through TensorFlow's [Dev16] seed management functions. At the start of each epoch, the seed is reset to maintain consistency.

Performance Optimization Techniques To prevent overfitting and improve model performance, several optimization techniques are employed throughout the training process:

- **Early Stopping:** Early stopping is applied when validation Loss shows no improvement after three consecutive epochs, preventing overfitting and saving computational resources.

- **Dynamic Learning Rate Adjustment:** The learning rate starts at 0.001 and is dynamically adjusted using a ReduceLROnPlateau callback. If validation Loss stagnates for two epochs, the learning rate is reduced by a factor of 0.2, with a minimum learning rate of 0.0001, allowing the model to fine-tune weights more effectively.
- **Multi-GPU Support:** TensorFlow's [Dev16] MirroredStrategy is used to distribute the workload across multiple GPUs, if multi-GPU support is unavailable, the model defaults to single-GPU or CPU training.

Monitoring and Saving the Best Model
During training, performance metrics such as the Dice coefficient and Intersection over Union (IoU) are logged after each epoch. These metrics guide model adjustments and help in selecting the best-performing model for evaluation on unseen data.

Key aspects:

- **Model Checkpoints:** At the end of each epoch, the model is saved based on its validation IoU Score, ensuring that the best configuration is retained for later evaluation.
- **Performance Logging:** A CSV logger tracks key metrics, including training/validation Loss and Dice Scores, to provide a detailed analysis of model performance across multiple runs.

Reproducibility
Reproducibility is ensured through consistent seed management. A seed list with 100 predefined values guarantees that data shuffling and augmentation are applied in a consistent, reproducible manner, even across different experiments and variations in the Focal Weight Factor.

Representative Slices and Visualizations
For each test scenario, a selection of representative slices from the FLAIR, T1ce, and T2 modalities is visualized, focusing on key tumor regions. These slices were chosen based on the presence of distinct tumor features, such as contrast enhancement, non-contrast enhancing core, or peritumoral edema. The visualizations include both the predicted and ground truth segmentation masks, providing qualitative insights into the model's segmentation accuracy. Particularly, overlaying the predicted masks onto T2 images, as shown in Figure 7.22, allows for a clearer assessment of segmentation performance.

Saving and Comparison of Results
Performance metrics, including the Custom Weighted Dice Score, are saved in a CSV file, enabling comparison and further evaluation s disscussed in Chapter 7. This allows for a comprehensive analysis of the model's performance across the entire validation dataset and supports the selection of the best-performing models based on the saved metrics.

6.10 Testing Environment

Efficient model training for deep learning-based glioblastoma segmentation requires a computational environment that balances high-performance hardware, scalability, and compliance with data privacy regulations. This section outlines the hardware setup, development tools, and cloud-based infrastructure used in this study.

Hardware Specifications

The experiments were conducted using the NVIDIA A100 GPU, which features 6912 CUDA cores and 40 GB of high-bandwidth memory (HBM2). This GPU is optimized for high-performance computing and large-scale AI workloads, making it ideal for the computationally intensive task of training 3D U-Net models for glioblastoma segmentation [NVI21]. Its advanced architecture provides substantial computational power, enabling faster model convergence and efficient experimentation.

Cloud-Based Environment

To leverage high-performance hardware without the constraints of local resources, all experiments were carried out within the Google Colab environment [Col23], a cloud-based Jupyter notebook platform that provides access to various GPUs, including the NVIDIA A100. Google Colab [Col23] offers seamless integration with Google Drive, which was essential for handling the large datasets used in this project. Additionally, the platform comes pre-installed with machine learning libraries, simplifying setup and reducing the overhead associated with configuring a local environment. Although lower-cost GPUs such as the L4 and T4 were available, the A100 was selected for its superior performance in computationally intensive tasks.

Development Tools

The implementation was carried out in Python 3.11 [VR09], using TensorFlow [Dev16] version 2.15.0 and Keras [Cho15] version 3.3.3 as the primary deep learning frameworks. TensorFlow was chosen over alternatives like PyTorch [PGM+19] and MONAI [Con20] due to specific compatibility and performance considerations in this project. These tools provided a robust and stable platform for developing, training, and evaluating 3D U-Net models tailored for medical image segmentation.

Motivation for Cloud Usage

A key motivation for selecting Google Colab [Col23] was the need for computational flexibility due to software incompatibilities on local hardware. The A100's advanced architecture significantly accelerated deep learning tasks, reducing training times that might otherwise take over 24 hours on a standard GPU such as the NVIDIA RTX 4070 mobile to approximately 6–8 hours per model. This fourfold reduction in training time enabled more extensive experimentation within the project's timeframe. Additionally, Google Colab's pricing model, which charges approximately €1.2 per hour for the A100 GPU, made it a cost-effective choice, despite total costs exceeding €1500 due to the extensive experimental runs conducted during the project. The pre-configured

environment further streamlined the workflow, enabling rapid experimentation without the need for additional setup.

Data Privacy and Security

To comply with data privacy regulations, preprocessing steps such as skull-stripping were performed locally using an NVIDIA RTX 4070 mobile GPU (4608 graphics cores, 8 GB GDDR6X memory), an AMD Ryzen 9 7940HS CPU (8 cores/16 threads, clock speed: 4.00–5.20 GHz), and 64 GB of DDR5-4800 RAM. This ensured that patient faces could not be reconstructed from the anonymized image data. Only NIfTI files, stripped of header information traceable to patient identities, were uploaded to the cloud. Patient names were replaced with anonymized identifiers, ensuring no personally identifiable information was stored on Google servers.

In summary, the experimental testing environment was meticulously designed to balance computational demands with privacy and cost-efficiency. The NVIDIA A100 GPU, integrated within the Google Colab [Col23] platform, provided robust performance for deep learning tasks, while TensorFlow [Dev16] and Keras [Cho15] ensured a reliable framework for model development. This setup enabled efficient and secure experimentation, facilitating the successful implementation of gliobastoma segmentation using 3D U-Net models.

CHAPTER 7

Results

Introduction to Chapter 7

This chapter systematically evaluates the segmentation results obtained from the trained deep learning models. The objective is to analyze the impact of different hyperparameters and dataset characteristics on model performance and to address the research questions formulated in Section 1.2. The chapter is structured in a logical sequence that first establishes the dataset properties, then examines segmentation performance across various conditions, and finally explores model optimization strategies and computational efficiency. Each section contributes to answering specific research questions regarding dataset influence, augmentation strategies, hyperparameter tuning, and computational constraints.

Section 7.1 provides an overview of the training and evaluation datasets, ensuring transparency regarding their composition. This is essential for assessing the generalizability of the models and evaluating whether differences between datasets may have influenced the results. Since dataset composition directly affects model learning, Section 7.2 further investigates the distribution of ground truth segmentation classes, which is relevant for understanding potential class imbalances and their impact on performance. These sections contribute to the broader research question concerning the representativeness of the dataset and its implications for training deep learning models.

Building on this foundation, Section 7.3 presents the segmentation results for the best-performing model in each of the four Case Groups during training and validation. Understanding how dataset size and augmentation strategies influence segmentation accuracy is a key aspect of this analysis. However, this does not yet include performance evaluation on the unseen evaluation dataset, which is addressed in Section 7.9. Section 7.4 then systematically examines the effect of the number of training cases on evaluation metrics, addressing the research question of how dataset size influences model generalization during training.

Section 7.5 and Section 7.6 provide a more detailed analysis of hyperparameter tuning and computational considerations. Section 7.5 focuses on the optimization of the Focal Weight Factor, a key component in the Combined Loss Function, and its influence on segmentation accuracy. This section directly addresses the research question regarding the impact of Focal Loss adjustments on model performance. Section 7.6 evaluates the role of the Batch size in training efficiency and model performance, including the inflection point where computational demand shifts from linear to accelerated growth. This analysis is particularly relevant for optimizing resource allocation and model training strategies.

The next sections provide further insights into augmentation strategies and overfitting detection. Section 7.7 investigates the effect of data augmentation on model performance, addressing the research question of how different augmentation strategies influence segmentation quality. Section 7.8 then examines the detection of overfitting, particularly in relation to Batch size and augmentation ratio, shedding light on the risk of performance degradation when excessive augmentations are used.

Sections 7.9 to 7.11 focus on model evaluation and comparative analyses. Section 7.9 provides a comparative assessment of models trained with different augmentation strategies, while Section 7.10 compares the performance of models trained with varying Batch sizes. These sections contribute to a deeper understanding of how the Batch size and augmentation interact in model training. Section 7.11 explores the impact of the Custom Weighted Dice Score, ensuring that the selected weighting scheme aligns with tumor class distributions and clinical relevance.

The final Sections 7.12 to 7.14 summarize key findings and provide broader insights. Section 7.12 presents a comparison of segmentation accuracy across different parameter configurations, offering a comprehensive overview of the best-performing setups. Section 7.13 discusses the clinical implications of the findings, linking the segmentation performance to potential real-world applications. Finally, Section 7.14 offers a critical reflection on the limitations of the study and identifies potential directions for future research.

By following this structured approach, Chapter 7 systematically addresses the key research questions outlined in Section 1.2. It provides a clear connection between dataset characteristics, hyperparameter choices, computational constraints, and segmentation performance, ultimately guiding the interpretation of results and their implications for deep learning-based glioblastoma segmentation.

7.1 Dataset Demographics

The age of patients was provided for the training datasets from the RSNA-ASNR-MICCAI as part of the Brain Tumor Segmentation Challenge BraTS [BGM+23], with a median of 61 years, IQR 53 – 69, compared to the evaluation dataset with a median of 64 years, IQR 54 – 72. There is no significant difference in age between the two datasets according to the

Mann-Whitney U test, with a p-value of 0.069. Gender distribution is only available for the evaluation dataset, with 67 males out of a total of 108 patients. The age distribution, with a predominance of male patients, is consistent with findings in the literature, which report that glioblastomas are more common in males, with the average age at diagnosis typically ranging between 55 and 65 years [OCW+21]. A comparison of the two datasets can be seen in Figure 7.1.

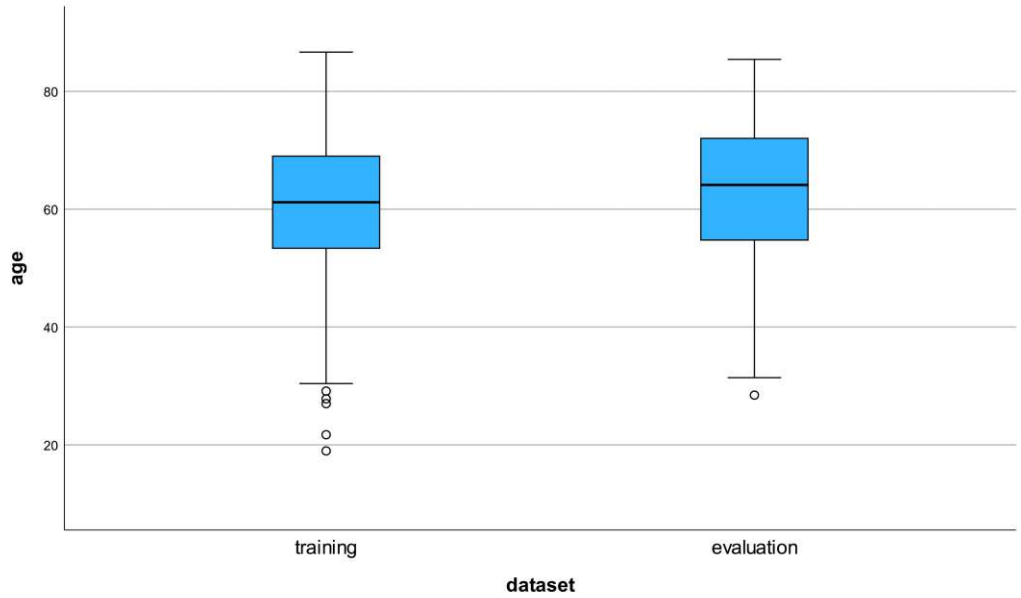


Figure 7.1: Boxplots showing the age distribution of patients. The left boxplot represents the training dataset (provided by BraTS as described in detail in Section 5.1), while the right boxplot represents the evaluation dataset.

Training Dataset
The training dataset included up to 393 cases, divided into the 80% training and 20% validation subsets, with four distinct Case Groups created to ensure a proportional representation of Source 1 and Source 2, as described in detail in Section 5.1.

- Cases group 80 (80 training plus 20 validation cases)
- Cases group 160 (160 training plus 40 validation cases)
- Cases group 240 (240 training plus 60 validation cases)
- Cases group 314 (314 training plus 79 validation cases)

Each Case Group builds upon the previous one: Case Group 160 includes the initial Case Group 80, Case Group 240 includes Case Group 160, and the final Case Group 314 contains all available cases. This incremental expansion ensures that the larger

groups are made up of the smaller subsets. The models were trained exclusively on the 80/160/240/314 training cases, while the validation cases were used solely to guide the optimization process and monitor model performance.

Evaluation dataset
We retrospectively identified 108 Caucasian patients (67 men, median age 64 years, IQR 54–72) with preoperative imaging. All patients underwent neurosurgical treatment for a histopathologically confirmed de novo GBM, classified according to the WHO guidelines valid at the time of diagnosis. According to the current classification, these 108 patients would have been diagnosed with glioblastoma WHO grade 4, IDH-wildtype [LPW+21].

7.2 Ground Truth Segmentation Distribution

The frequencies of the tumor segmentation classes in the training and evaluation datasets are illustrated in the diagrams in Figure 7.2. The top diagram represents the training dataset, while the bottom diagram shows the evaluation dataset. Each class is represented by a different color: green for non-contrast-enhancing tumor (Class 1), yellow for edema (Class 2), and brown for contrast-enhancing tumor (Class 3).

The horizontal axis of the diagrams in Figure 7.2 corresponds to the cranio-caudal direction (z-axis) of the MRI images, ranging from slice 0 to 127. The vertical axis indicates the frequency of the respective voxels for each class. In the diagrams, the lowest peak represents the contrast-enhancing tumor (Class 3), the next higher peak corresponds to the non-contrast-enhancing tumor (Class 1), and the largest peak is associated with the edema (Class 2). The background (Class 0) was not included in the analysis, as it not only represents tumor-free brain tissue but also the regions outside the brain. These external regions can vary significantly depending on the size and shape of the patient's head, introducing potential confounders and inaccuracies into the analysis. Therefore, excluding the background ensures a more robust and meaningful comparison of the tumor-related classes.

From these diagrams, we can observe that the overall distribution pattern of the segmentation classes along the the cranio-caudal axis of the MRI images is consistent between the training and evaluation datasets. However, there are differences in the exact frequencies of the voxels for each class in the respective slices. These differences might impact the performance of the 3D U-Net model, as variations in voxel frequencies between the training and evaluation dataset can affect the model's ability to accurately generalize and segment the tumor regions.

The cumulative percentage distribution of all segmentation classes, including the background (Class 0), is shown in Figure 7.3. The diagrams illustrate that the background class dominates the overall distribution, as indicated by the significantly higher percentage of the blue bars. To ensure the smaller proportions of tumor-related classes are visible, a secondary vertical axis is used for Classes 1–3. This highlights the pronounced imbalance

between the background and the tumor-related classes, which is addressed during training by applying class weights.

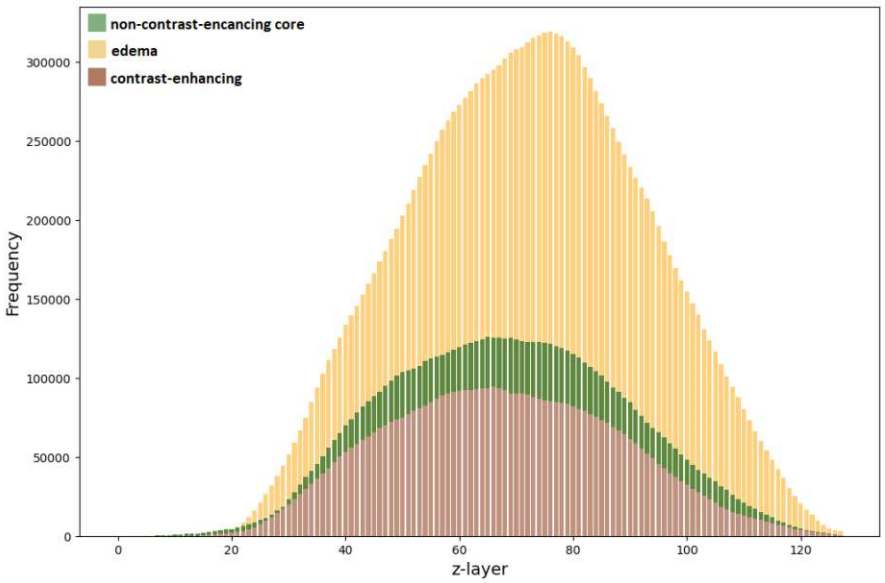
The anatomical locations of the segmented regions, if viewed as cumulative frequencies along the the cranio-caudal axis, are depicted in Figure 7.2 and appear relatively consistent between the two datasets. This does not imply that tumors are always located in the exact same position but rather that their distribution along the cranio-caudal axis is similar across both datasets. This observation aligns with findings that glioblastomas predominantly occur in the cerebral hemispheres, as 62% of gliomas are located in the supratentorial compartment, including the frontal, temporal, parietal, and occipital lobes [PCP+22a]. The majority of the tumors are observed in slices 50–80, which correspond to the coverage of these four lobes. This consistency in the spatial distribution of tumors across both cohorts is essential for the model's ability to generalize effectively and accurately capture the spatial characteristics of the tumors.

From the relative frequencies of the classes, the class weights can be calculated. The principle behind calculating class weights is to adjust for class imbalance by assigning a higher weight to less frequent classes and a lower weight to more frequent classes. This ensures that the model does not become biased towards the more common classes.

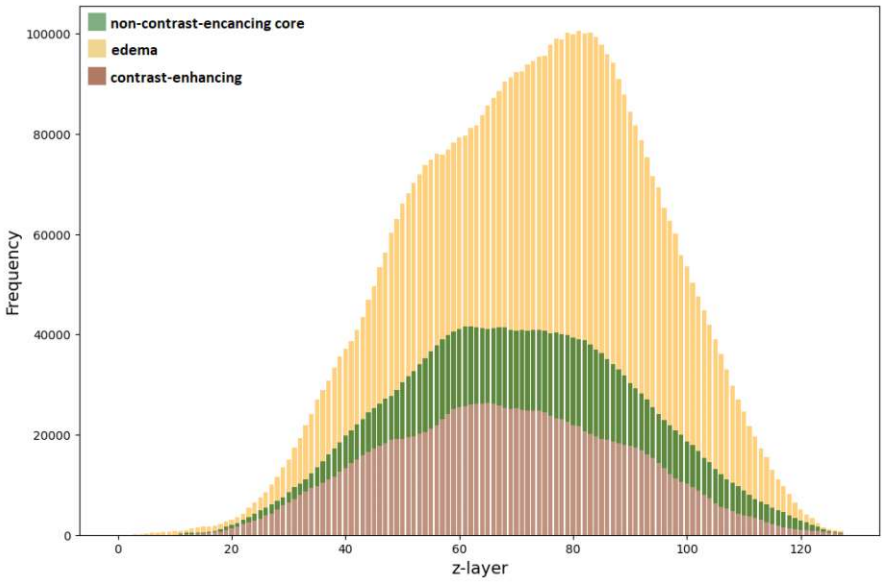
For training a 3D-UNet model, it is advisable to calculate the class weights based solely on the training dataset, as the validation data should be used separately for model validation. Consequently, 314 out of the 393 cases are used for training. This approach ensures that the validation data do not influence the training process in any way.

The resulting class weights for the training dataset (314 cases) and the evaluation dataset (108 cases) are presented in Table 7.1. These weights address class imbalances by assigning higher importance to less frequent classes, such as the contrast-enhancing tumor, ensuring the model does not become biased toward more common classes. They are integral to the Combined Loss Function (Section 4.2.3) and the Custom Weighted Dice Score (Section 4.1.6 and Section 5.6), balancing the loss during training and improving segmentation performance.

The tumor characteristics represented by these weights show only minor differences between the training and evaluation datasets, with variations of approximately 5%, 10%, and 11% for the three tumor classes. This consistency indicates that glioblastomas in both datasets are relatively comparable in their characteristics.

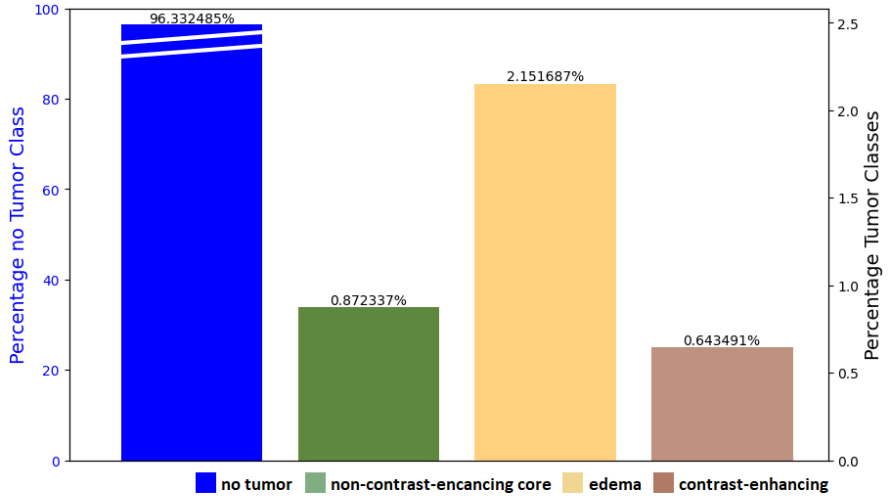


(a) Frequency distribution of tumor segmentation classes in the training dataset.

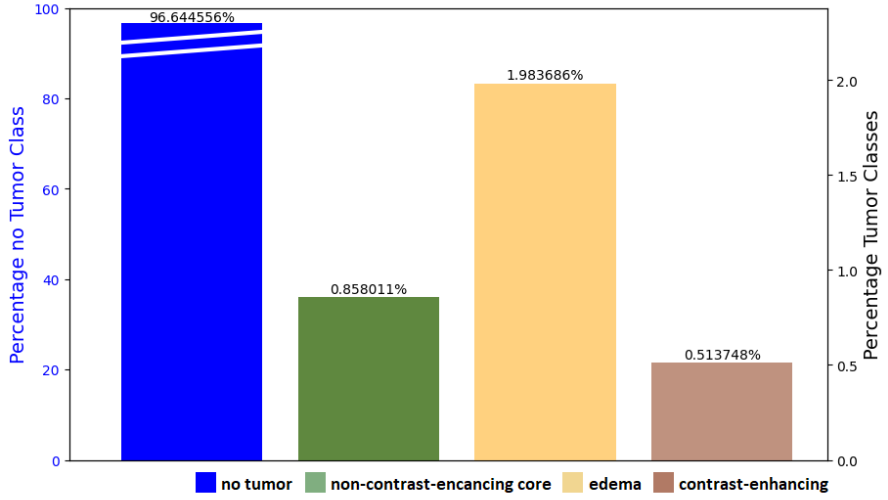


(b) Frequency distribution of tumor segmentation classes in the evaluation dataset.

Figure 7.2: The diagrams illustrate the frequency distribution of the tumor segmentation classes, with the training dataset shown in (a) and the evaluation dataset in (b). The colors represent the classes as follows: non-contrast-enhancing tumor (green), edema (yellow), and contrast-enhancing tumor (brown). The horizontal axis represents the cranio-caudal direction of the MRI images, ranging from slice 0 to 127, and the vertical axis represents the frequency of the corresponding voxels for each tumor class.



(a) Percentage distribution of the four classes in the training dataset.



(b) Percentage distribution of the four classes in the evaluation dataset.

Figure 7.3: The diagrams illustrate the percentage distribution of the four classes in the training (a) and evaluation (b) datasets: no tumor (blue), non-contrast-enhancing tumor (green), edema (yellow), and contrast-enhancing tumor (brown). To account for the significantly higher percentage of the no tumor class, a breakline is introduced in the blue bars, indicating that the actual bar height exceeds the displayed scale compared to the tumor classes. The tumor classes are represented on a secondary vertical axis (right) for enhanced visibility of their smaller proportions.

Class	Class Description	Training Dataset Weight	Evaluation Dataset Weight	Percentage Difference	Absolute Difference (in % points)
0	no tumor	0.0033	0.0029	-12.12%	0.0004
1	non-contrast-enhancing tumor	0.3610	0.3214	-10.97%	0.0396
2	edema	0.1464	0.1390	-5.05%	0.0074
3	contrast-enhancing tumor	0.4894	0.5367	+9.68%	0.0473

Table 7.1: Comparison of class weights in the training cases (314 cases) and evaluation (108 cases) datasets. Absolute differences and percentage differences highlight variations in class weights between the datasets. Note that Class 0 (background) includes not only tumor-free brain tissue but also regions outside the brain, which can vary significantly between cases.

7.3 Model Overview

To investigate the effect of various parameters on segmentation performance, a total of 1632 models were computed. The models were generated based on the following combinations:

- **Case Groups:** Four Case Groups were used: 80, 160, 240, and 314.
- **Batch sizes and augmentation variations:**
  - For each Batch size (1, 2, and 4), models were calculated with and without augmentations.
  - **Batch size 1:**
    - **Version 1:** With a 1:1 ratio of original to augmented cases.
    - **Version 2:** With a 2:1 ratio of original to augmented cases.
  - **Batch size 2:** With a 1:1 ratio, consisting of one original case and one augmentation of a random case.
  - **Batch size 4:**
    - **Version 1:** With a 1:3 ratio, consisting of one original case and three augmentations of the same case.
    - **Version 2:** With a 1:3 ratio, consisting of one original case and three augmentations of random cases.
- **Focal Weight Factor:** For each of the 32 combinations (8 variations across 3 Batch sizes and 4 Case Groups), the Focal Weight Factor was varied from 0.0 to 5.0 in increments of 0.1, leading to 51 models per combination.
- **Epochs:** Each model was trained for 3 to 100 epochs, with an average of 30 epochs.

This resulted in a total of 1632 models, calculated as 32 combinations multiplied by 51 models per combination. To reduce this number, a selection process was implemented. For each of the 32 combinations, the models with the top 10 Focal Weight Factors that yielded the highest validation IoU Score during training were selected, resulting in a total of 320 models.

These 320 models were then applied to the unseen 108 evaluation cases to calculate metrics such as the IoU score, the Dice coefficient for the three tumor classes (non-contrast-enhancing tumor, edema, and contrast-enhancing tumor), accuracy, and, based on these, the Custom Weighted Dice Score (as described in Section 4.1.6).

7.4 Focal Weight Factor Tuning

The Focal Weight Factor was hyperparameterized to optimize the Combined Loss Function and improve segmentation performance. The models were trained with Focal Weight Factors ranging from 0.0 to 5.0 in increments of 0.1, and the validation IoU Score was measured for each setting across the four Case Groups: 80, 160, 240, and 314 training cases. This analysis was performed for Batch size 4 without augmentations to isolate the effect of the Focal Weight Factor on model performance.

Comparison of validation IoU Score for different Case Groups
In Figure 7.4, the highest IoU Scores achieved on the validation cases for each Focal Weight Factor are plotted for the four Case Groups. There is a significant difference between the lower Case Groups (80 and 160) and the higher Case Groups (240 and 314), with the Case Group 314 consistently achieving the best performance. The colors in the plot transition from low saturation (80) to high saturation (314), visually encoding the increase in training data. The gap in IoU Score performance is most pronounced between the smallest Case Group (80) and the largest Case Group (314). This demonstrates the clear advantage of increasing the number of training cases for optimizing model accuracy, particularly when fine-tuning the Focal Weight Factor.

Heatmap Visualization
The heatmap in Figure 7.4b provides an alternative visualization of the data presented in Figure 7.4. It illustrates how the validation IoU Score varies as a function of the Focal Weight Factor across all Case Groups. The darker red regions encode higher IoU Scores, independent of the specific Case Group, and highlight the range of Focal Weight Factors (approximately 0.5 to 3.5) where the highest IoU Scores are achieved, particularly for the larger Case Groups. Unlike the line plot, this visualization does not differentiate between the individual Case Groups but instead focuses on overall performance trends. The Colorbar on the right represents the density of IoU Scores, with a maximum value of 0.61 observed in this highlighted range, indicating the best segmentation performance during validation.

However, it is important to note that the density shown in the Colorbar is dimensionless and represents the distribution of the IoU Scores. This allows for a relative comparison

of regions with higher and lower IoU Scores, it does not provide an absolute measure of performance. Additionally, the heatmap emphasizes the IoU intensity distribution rather than absolute performance metrics per Case Group. This limitation means that the Colorbar serves primarily as a visual aid to identify trends and patterns, rather than a precise quantitative metric. Despite this, the heatmap complements the line plot in Figure 7.4 by offering a more intuitive understanding of the data trends and the effectiveness of different Focal Weight Factors.

Statistical Analysis
To assess the statistical significance of these differences, a series of tests were performed:

- **Shapiro-Wilk Normality Test:** Indicated that none of the Case Groups followed a normal distribution (all p-values < 0.05).
- **Levene's Test:** Suggested no significant differences in variance between the Case groups (p-value $= 0.1223$).
- **ANOVA Test:** Demonstrated a significant effect of the size of the Case Group on the highest validation IoU Scores observed during training (p-value < 0.001).
- **Tukey HSD Post-Hoc Test:** Confirmed significant differences between most Case Groups, with the largest difference between 80 and 314.
- **Boxplot Comparison** of results for the maximum IoU Scores.

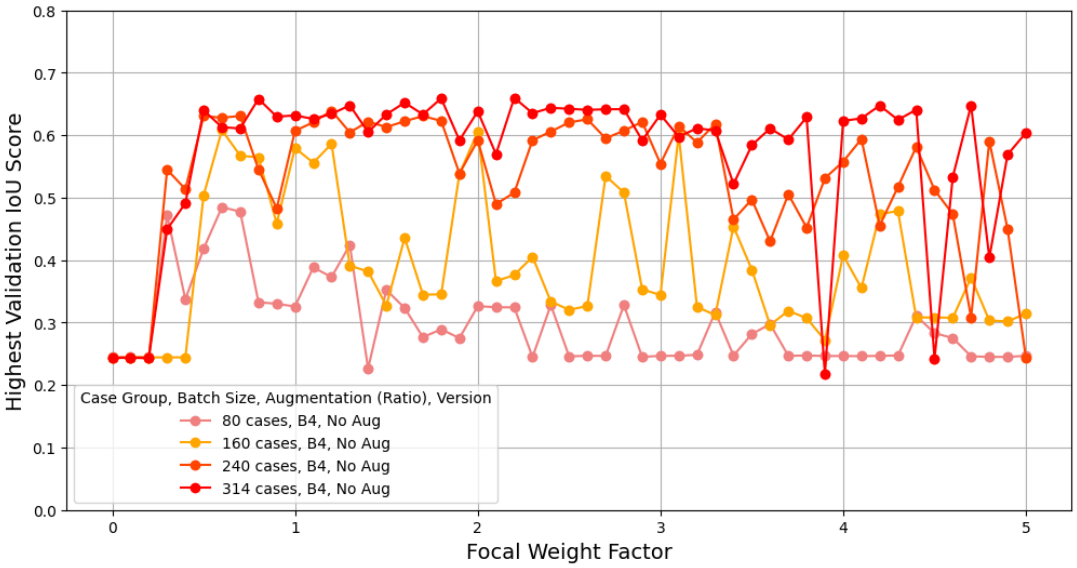
Finally, Figure 7.5 shows boxplots comparing the maximum IoU Scores achieved across the four Case Groups. The distribution of IoU Scores highlights the clear advantage of larger Case Groups, with 240 and 314 outperforming smaller groups (80 and 160) in terms of maximum IoU Score. This underlines the benefit of increasing the number of training cases to improve segmentation performance.

7.5 Training Process

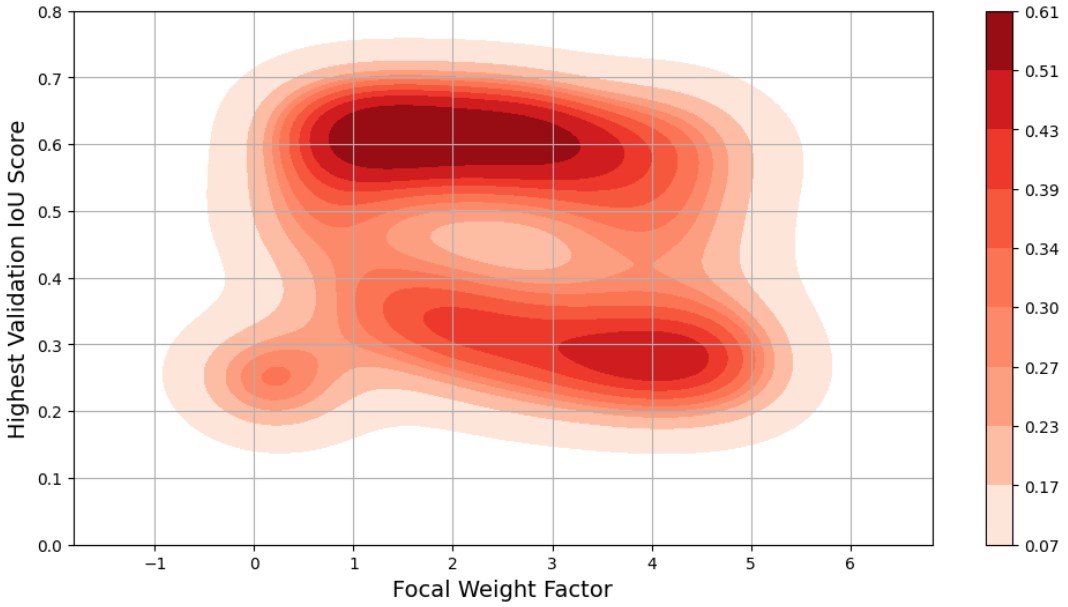
The training process for the 3D U-Net model was monitored using several key metrics, including Loss, accuracy, IoU (Intersection over Union), and the Dice coefficients for the three tumor classes (non-contrast-enhancing tumor, edema, and contrast-enhancing tumor). The training process was governed by the training and validation IoU Score, to select the models that demonstrated the best validation performance.

Loss function and Accuracy development

As shown in Figure 7.6, the top left plot depicts the progression of the Combined Loss Function for both the training and validation datasets. The training Loss (blue line) steadily decreases throughout the epochs, while the validation Loss (red line) reaches a plateau after approximately 30 epochs. This indicates that while the model continues to improve on the training data, the validation performance stabilizes, suggesting that



(a) Highest Validation IoU Scores for different Focal Weight Factors.



(b) Heatmap representation of the IoU Scores shown in (a).

Figure 7.4: (a) Highest validation IoU Scores for different Focal Weight Factors across the four Case Groups (80, 160, 240, and 314) with Batch Size 4 and no augmentations. The results show a clear advantage of increasing the number of training cases, with Case Group 314 consistently achieving the highest scores. (b) Heatmap representation of the same data, highlighting the range of Focal Weight Factors (0.5 to 3.5) where the highest IoU Scores are observed.

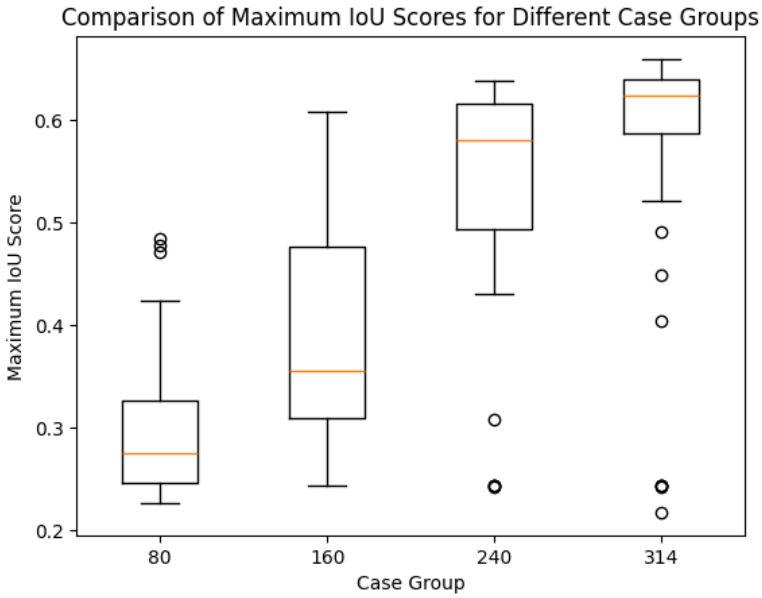


Figure 7.5: Boxplot comparing the maximum IoU Scores for the four Case Groups, showing that the larger groups (240 and 314) consistently outperformed the smaller groups in terms of maximum IoU Score.

further training beyond this point does not yield significant improvement in generalization. The accuracy of the model, presented in the top middle plot, mirrors this trend. After a sharp initial increase in both training and validation accuracy, the validation accuracy also plateaus after about 30 epochs, reflecting the stabilization of model performance. Interestingly, the validation accuracy outperforms the training accuracy at this stage. This phenomenon can be attributed to regularization techniques such as dropout and early stopping, which introduce stochasticity into the training process by randomly deactivating neurons or connections [WZZ+13, LWL+21]. These methods prevent overfitting by penalizing the co-adaptation of neurons, enabling better generalization. As a result, the model achieves higher validation accuracy compared to the training accuracy, despite the induced randomness slightly lowering the training performance [WZZ+13, LWL+21].

IoU Score and Dice coefficients

The Intersection over Union (IoU), shown in Figure 7.6 (top right plot), is a critical metric for evaluating the overlap between the predicted and ground truth tumor regions. Similar to the accuracy, the validation IoU Score plateaus around 30 epochs, confirming the model's generalization performance on the validation data. In the bottom row, the Dice coefficients for the three tumor classes (non-contrast-enhancing tumor, edema, and contrast-enhancing tumor) are illustrated. These coefficients show the overlap between the predicted segmentations and the ground truth for each tumor class. While the training Dice coefficients improve continuously, the validation Dice coefficients show a

similar plateau effect after 30 epochs, particularly for the non-contrast-enhancing and contrast-enhancing tumor classes.

- The Dice coefficient for the **non-contrast-enhancing tumor**, as shown in Figure 7.6 (bottom left), shows steady improvement in both training and validation, with the validation coefficient reaching a plateau.
- The **edema** Dice coefficient, depicted in Figure 7.6 (bottom middle), follows a similar pattern but shows slightly more variation, suggesting that edema segmentation might be more challenging for the model.
- The **contrast-enhancing tumor**, illustrated in Figure 7.6 (bottom right), shows the best performance among the three classes, with the validation Dice coefficient reaching a high value and stabilizing after 30 epochs.

Model Selection Based on Validation IoU Score

These trends, including the plateau in validation IoU Score and Dice coefficients after approximately 30 epochs, provide important insights into the model's performance. The stabilization of these metrics on the validation data suggests that further training beyond this point does not significantly improve the model's generalization ability. Based on these observations, models with the highest validation IoU Score during the training process were selected for further evaluation. The IoU Score serves as the primary criterion for identifying the best-performing models, which were then applied to the unseen evaluation dataset for final testing.

7.6 Model Application to Unseen Data

After selecting the top 10 Focal Weight Factors based on the highest validation IoU Score for each of the 32 parameter combinations (Batch size, augmentation, and Case Group), these 320 models were applied to the 108 unseen evaluation cases. For this analysis, we focus on the scenario of Batch size 4 without augmentations in the Case Group 314. Figure 7.7 shows the results of applying the selected models to the unseen data. The IoU Score, Dice coefficients for the three tumor classes non-contrast-enhancing tumor, edema, and contrast-enhancing tumor, and overall Accuracy are plotted against the Focal Weight Factor. Each boxplot represents the distribution of scores across the unseen evaluation dataset for different Focal Weight Factors.

It is important to note that the Focal Weight Factors on the horizontal axis correspond to those that yielded the highest validation IoU Scores during training. Consequently, the specific Focal Weight Factors displayed in each diagram may differ across experiments. For instance, in Figure 7.16, which depicts a different model configuration, the Focal Weight Factors shown on the horizontal axis are not necessarily the same as those in Figure 7.7.

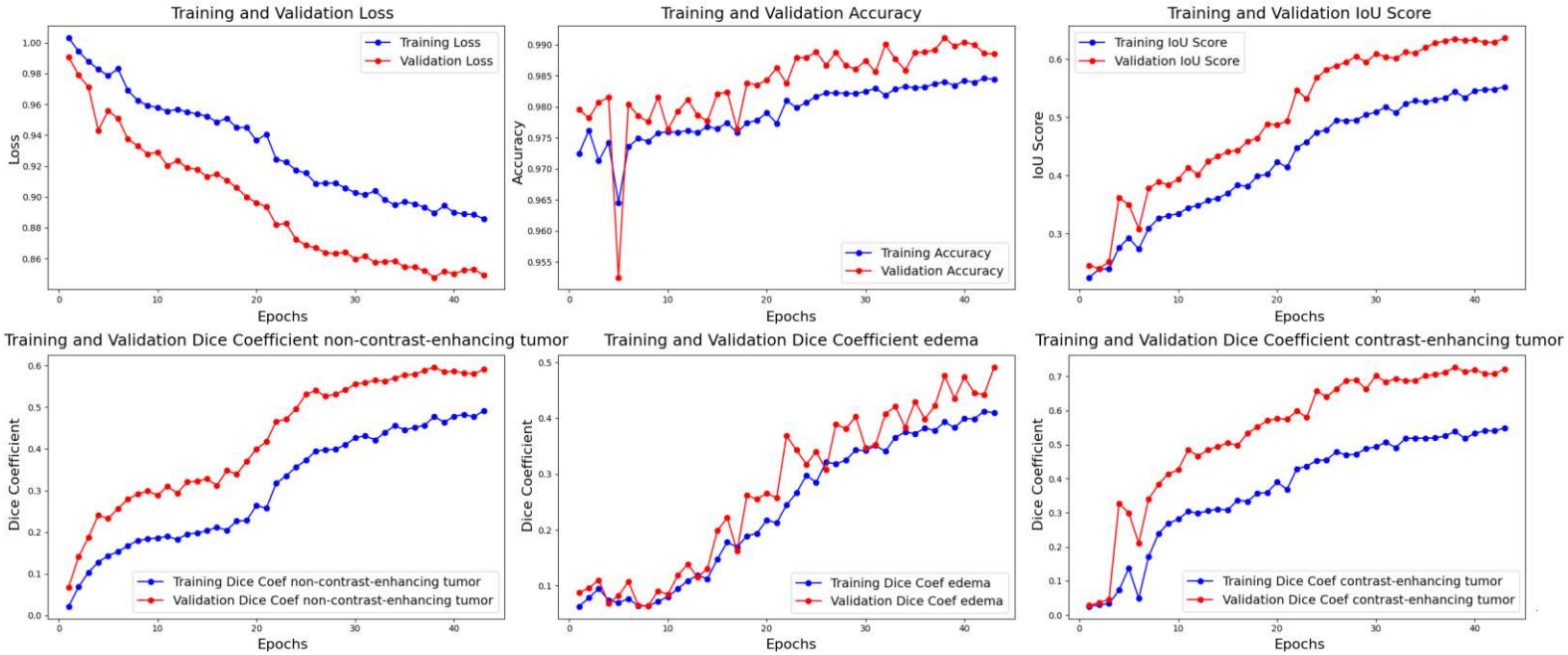


Figure 7.6: The top left plot shows the Combined Loss Function for training and validation. The top middle plot depicts accuracy, and the top right plot illustrates the IoU (Intersection over Union) Score for training and validation. The bottom row presents the Dice coefficients for the three tumor segmentation classes: non-contrast-enhancing tumor, edema, and contrast-enhancing tumor. In all plots, the blue lines represent the training dataset, and the red lines represent the validation dataset.

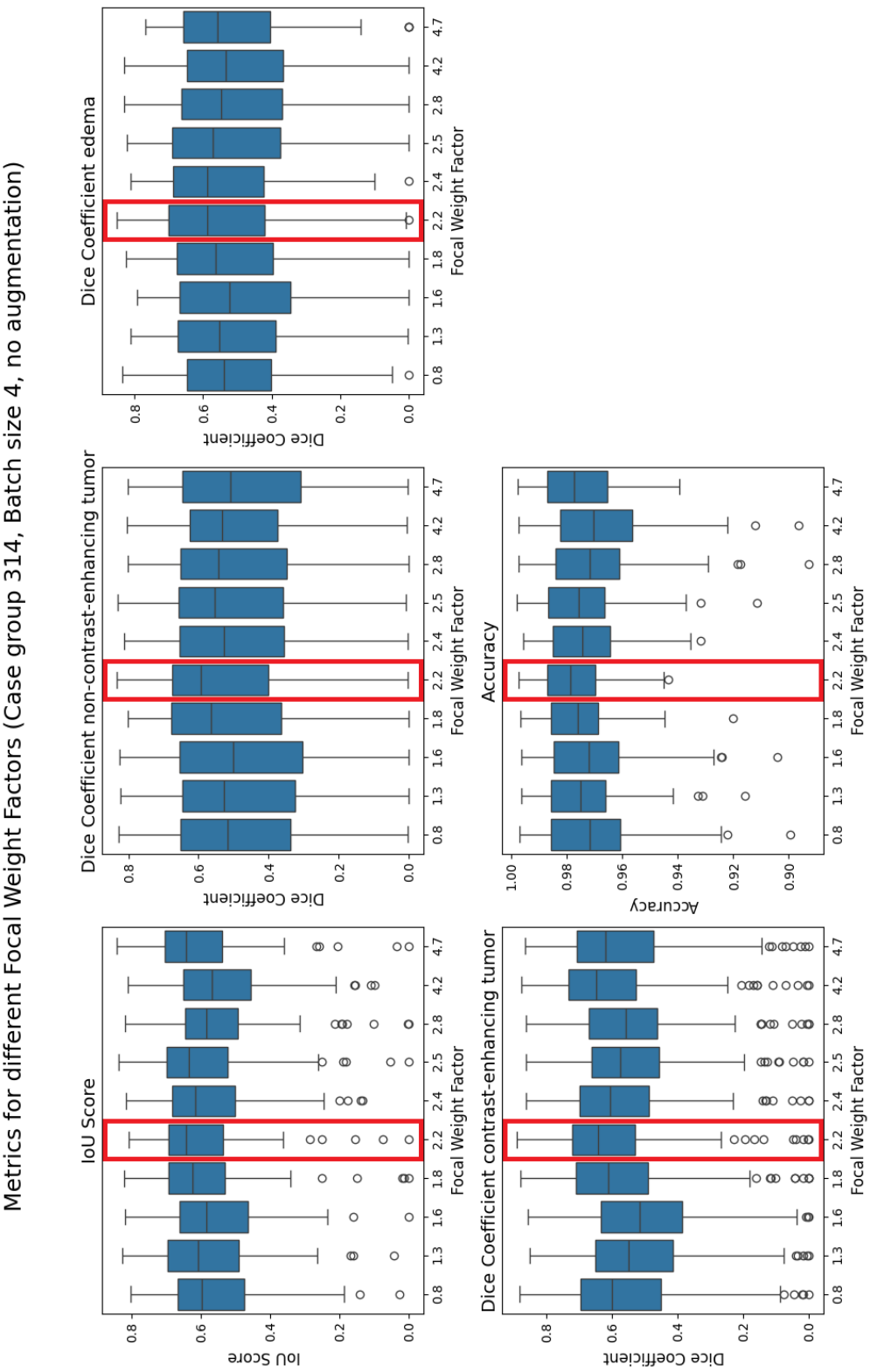


Figure 7.7: Metrics for different Focal Weight Factors in the Case Group 314 with Batch size 4 and no augmentations. Boxplots represent the distribution of IoU Score, Dice coefficients for non-contrast-enhancing tumor, edema, and contrast-enhancing tumor, and accuracy across 108 unseen evaluation cases. Only the 10 models with the highest IoU Score (out of all Focal Weight Factors tested) are shown on the horizontal axis. The model with the best performance, based on the Custom Weighted Dice Score, is highlighted in red.

The model with the best overall performance, as determined by the Custom Weighted Dice Score (described in Section 4.1.6), is marked in red in each plot shown in Figure 7.7. This model achieved the highest balance between IoU Score and Dice coefficients across all tumor classes, making it the optimal model for this specific scenario.

7.7 Training Time Analysis

The computational training time for the 3D U-Net was systematically analyzed across different Case Groups and models with Batch sizes of 1, 2, and 4, both with and without data augmentation. For each scenario, 51 different Focal Weight Factors were tested, leading to a total of 32 model configurations derived from the 8 combinations of Batch sizes, augmentation strategy, and augmentation version, applied to 4 Case Groups. Figure 7.8 presents the cumulative training time across these variations, highlighting a significant increase for models with Batch size 1 combined with augmentations, particularly at higher Case Group sizes. This increase was further investigated to identify its underlying causes.

To maintain a structured notation, the figure legend follows a consistent format: 'B1, No Aug' represents Batch size 1 without augmentation, while 'B1, Aug (1:1)' and 'B1, Aug (2:1)' indicate Batch size 1 with augmentation at a 1:1 or 2:1 ratio, respectively. Similarly, 'B4, Aug (1:3), V1' and 'B4, Aug (1:3), V2' correspond to Batch size 4 with an augmentation ratio of 1:3, where V1 and V2 denote different augmentation strategies. The results illustrate the computational impact of Batch size and augmentation on training duration.

For improved visual clarity, colors were consistently assigned across all figures following the same legend structure: red for Batch Size 1 (B1), blue for Batch Size 2 (B2), and green for Batch Size 4 (B4). Models without augmentation are represented by fully saturated colors, whereas models with augmentation appear in lighter shades of the corresponding Batch size color. This differentiation allows for a clear distinction between augmentation strategies while maintaining Batch size consistency across visualizations.

7.7.1 Impact of Augmentations on Training Time

For Batch size 1, the addition of augmentations significantly increased the computation time, resulting in a rapid increase in training duration as the number of cases grew. While the observed trend suggests a more-than-linear growth, the limited number of data points does not allow for a definitive determination of an exponential relationship. In contrast, for Batch sizes 2 and 4, the growth in computation time remained linear, even when augmentations were applied, as shown in Figure 7.8.

7.7.2 Transition to Rapid Growth ("Elbow Point")

To explore the transition from linear to more rapid growth in computation time, we analyze the critical points where the rate of increase in training time becomes significantly

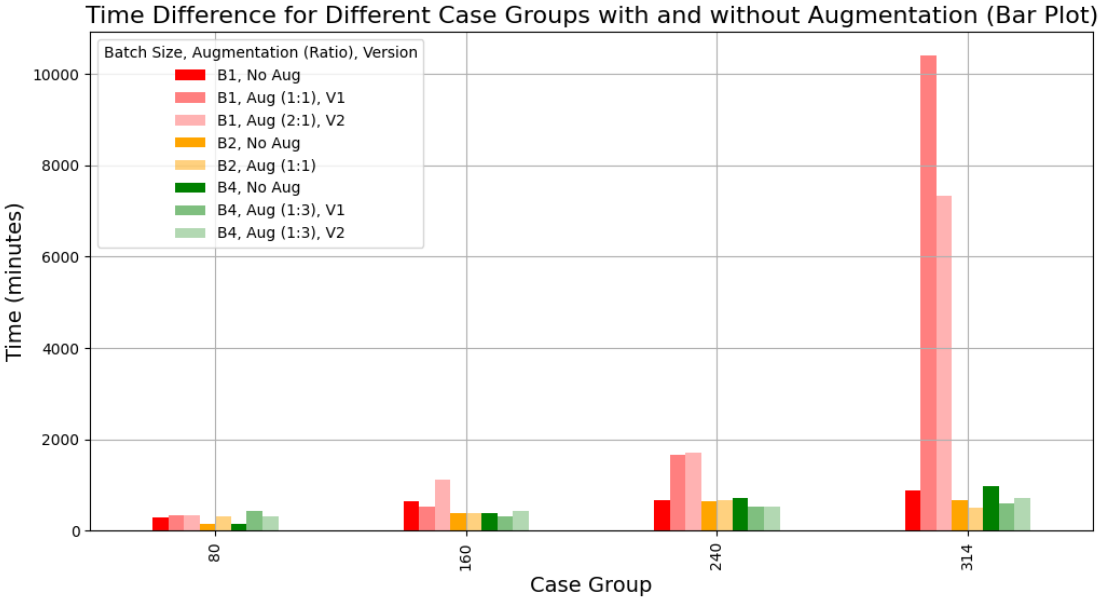


Figure 7.8: Bar plot illustrating the cumulative training time across different Batch sizes and augmentation strategies. Each model was trained with 51 different Focal Weight Factors per scenario, leading to a total of 32 model configurations across the four Case Groups.

higher. While the term "inflection point" in mathematics refers to a point where the second derivative changes sign, our focus lies on identifying the "elbow point", where the slope of the computation time curve increases sharply. This transition is observed exclusively for models with Batch size 1 and data augmentation. In contrast, for Batch sizes 2 and 4, computation time grows linearly, even when augmentations are applied, as illustrated in Figure 7.11.

Analyzed Configurations Two key configurations were examined: Version 1, which used a Batch size of 1 with a 1:1 ratio of original to augmented data, and Version 2, which used a Batch size of 1 with a 2:1 ratio of original to augmented data.

Quadratic Fit and Derivatives To model the observed computation time growth, quadratic equations were fitted to the data. Here, x represents the number of cases, while y denotes the corresponding computation time in minutes. Figure 7.10 shows the progression of the training time with quadratic fits. The dashed lines represent the fitted quadratic functions for Batch size 1 with augmentation strategies (Version 1 and 2). The corresponding equations are:

For Version 1: $$y = 0.000015x^2 - 0.004897x + 6.244658 \tag{7.1}$$

For Version 2: $$y = 0.000006x^2 + 0.004926x + 5.241635 \tag{7.2}$$

Although a quadratic model was applied, the small values of the coefficient of x^2 (denoted as a) suggest that computation time initially increases in an almost linear fashion. Given that only four Case Groups were tested, the robustness of this fit is limited, and additional data points would be needed to confirm a truly quadratic trend.

Defining the Transition Point To quantify the transition to accelerated growth, we analyze the derivative of the quadratic equation:

$$\frac{d}{dx}(ax^2 + bx + c) = 2ax + b \tag{7.3}$$

This derivative represents the rate of change in computation time at any given number of cases x. Instead of using an arbitrary threshold, the transition point is defined as the first Case Group where the rate of change exceeds a data-driven threshold, set to 50% of the maximum observed rate of change. Based on this criterion, the transition to rapid growth occurs at Case Group 314 for both augmentation strategies, as seen in Figure 7.9. The threshold values for the transition point are approximately 0.007 for Version 1 and 0.006 for Version 2, which closely align with observed trends.

The mathematically computed transition points, determined by setting the first derivative equal to the computed threshold, suggest a transition at approximately:

For Version 1: $$x = 484.18 \tag{7.4}$$

For Version 2: $$x = 456.57 \tag{7.5}$$

To enhance the visualization of the transition from linear to accelerated growth, a logarithmic scale was applied to the vertical axis in Figure 7.10, in contrast to Figure 7.9, which uses a linear scale. The transition points, where computation time starts increasing more rapidly, are marked with black dots. While these computed values indicate a theoretical threshold for rapid growth, empirical evidence suggests that practical computational constraints already become significant at Case Group 314. This discrepancy may be due to the limited number of data points in the quadratic fit and the influence of augmentation on memory consumption.

When examining the growth of computation times for models with Batch sizes 2 and 4, as shown in Figure 7.11, a clear linear growth pattern is evident. Notably, models with a Batch size of 4 processed a substantially higher number of cases, reaching up to 1256 cases when considering the augmentation ratio of 1:3. However, it is important to note that Figure 7.11 displays only the number of original cases on the horizontal axis, meaning that the augmented cases are not explicitly visualized. For comparison, the model with a Batch size of 1 and an augmentation ratio of 1:1 processed a total of 628 cases (comprised of 314 original and 314 augmented cases), but these values are not represented in Figure 7.11.

To avoid confusion, it is important to note that the total number of processed cases for Batch size 4 is derived from the multiplication of the Batch size by the number of original

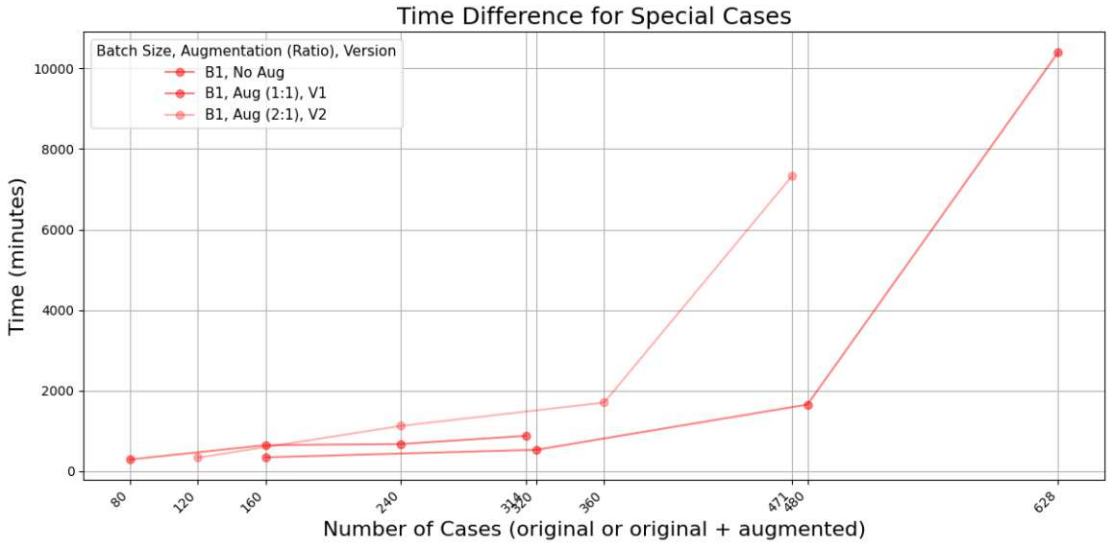


Figure 7.9: Computation time for Batch size 1 with and without augmentation. The plot highlights the rapid increase in training time if augmentations are applied, particularly at higher Case Group sizes. Each Case Group contains only four data points, reflecting the limited number of tested models per configuration.

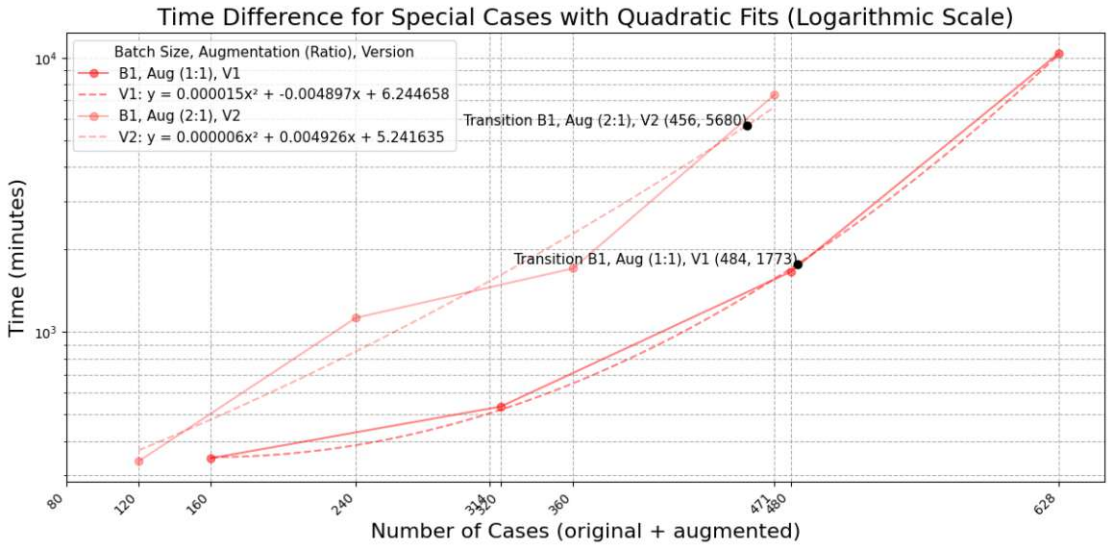


Figure 7.10: Training time differences with quadratic fits on a logarithmic scale. This figure presents the same data as Figure 7.9 but with a logarithmic vertical axis to highlight the quadratic fit. Dashed lines represent the fitted quadratic functions, while black markers indicate the transition points where computation time increases more rapidly.

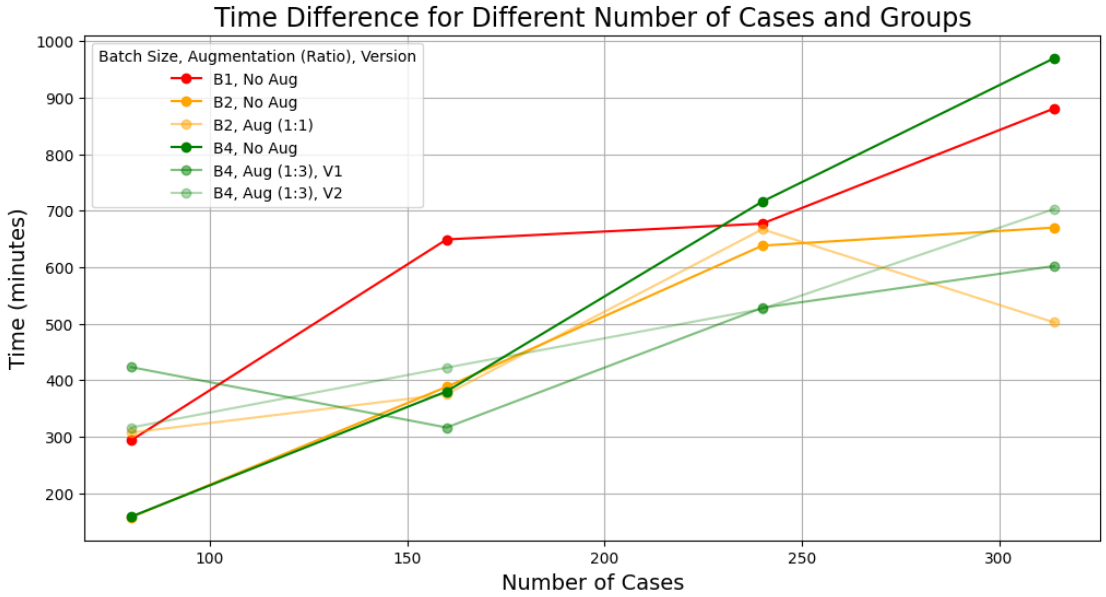


Figure 7.11: This plot illustrates the linear growth in training time for Batch sizes 2 and 4 with augmentations. To improve readability, the rapidly growing computation time for Batch size 1 with augmentations is not shown in this figure. Note: Case Group 314 with Batch size 1 (both Version 1 and Version 2) is excluded.

cases. For the largest Case Group, which consists of 314 original cases and follows an augmentation ratio of 1:3, this results in a total of 1256 processed cases. Since Batch sizes 2 and 4 exhibit linear time growth, the number of processed cases does not negatively impact training duration, as long as Batch size is adjusted accordingly.

7.8 Augmentation Impact on Performance

To evaluate the effect of different augmentation strategies on model performance, two approaches were tested using Batch Size 4. In the first strategy (Version 1), each batch consisted of one original case combined with three augmentations derived from the same original case. In contrast, the second strategy (Version 2) introduced greater variability by including one original case along with three augmentations sourced from different original cases within the batch.

Version 1: Augmentations of the Same Case

In Version 1, illustrated in Figure 7.12a, it is evident that the choice of Focal Weight Factor has little impact on the performance. Across all Case Groups (80, 160, 240, and 314), the models achieve a consistently low IoU Score, with minimal variation between different Focal Weight Factors. This suggests that using multiple augmentations of the same original case within a batch does not introduce sufficient diversity, leading to limited

generalization. The heatmap in Figure 7.12b reinforces this observation, showing that the IoU Scores are concentrated at the lower range, regardless of the Focal Weight Factor or the Case Group.

Version 2: Random Augmentations from Different Cases

In contrast, Figure 7.13a illustrates Version 2, which exhibits a different pattern. While the average IoU Score remains relatively low, there are instances where models achieve significantly higher IoU Scores. This suggests that incorporating augmentations from different original cases introduces beneficial variability, which helps the model generalize better and occasionally leads to higher segmentation accuracy. The heatmap in Figure 7.13b further illustrates this increased variability. Unlike Version 1, where performance remains consistently low, Version 2 shows a wider distribution of IoU Scores, with several peaks—especially in the larger Case Groups (240 and 314). This indicates that a greater number of training cases, in combination with diverse augmentations, can enhance model robustness.

Reproducibility of Random Augmentations

While the selection of augmentations from different cases in Version 2 follows a randomized approach, it is fully reproducible using the predefined seed list, as described in Section 6.8. This ensures that the same training pipeline can be re-executed with identical augmentation distributions, enabling consistent comparisons between different models and training setups.

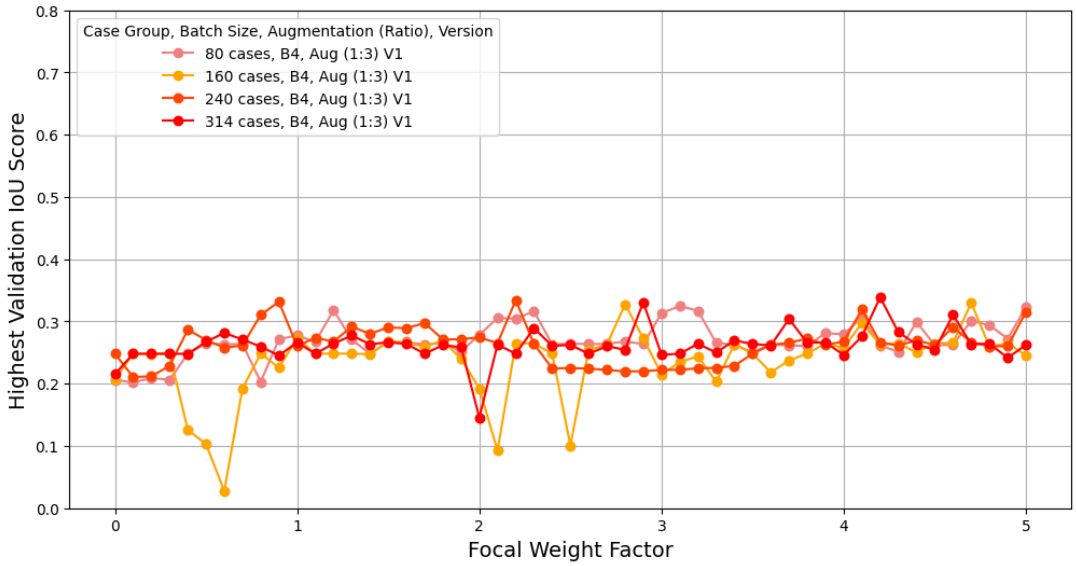
These findings highlight the importance of selecting an appropriate augmentation strategy. Augmenting the same case (Version 1) does not provide meaningful improvements, whereas introducing variability through augmentations from different original cases (Version 2) leads to better generalization. This effect becomes more pronounced as the number of training cases increases.

7.8.1 Generalization of Augmentation Strategy Version 1

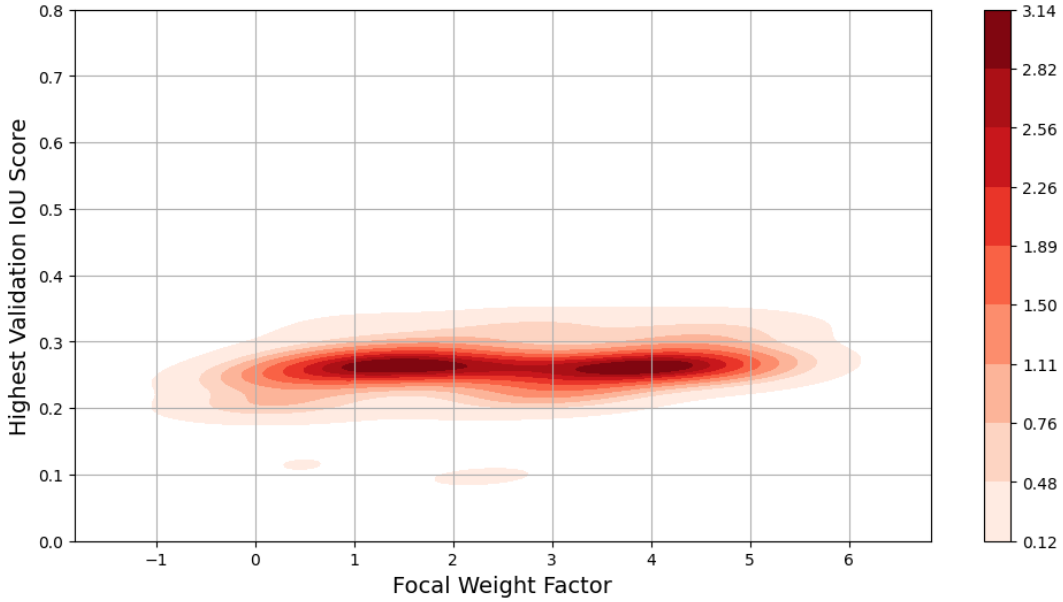
The 10 best models, trained on Case Group 314 using Batch size 4 with augmentation (Version 1), were applied to the evaluation dataset of 108 unseen cases. The results, illustrated in Figure 7.12, show a clear reduction in generalization performance compared to models trained without augmentations, as seen in Figures 7.4.

As expected, models trained using this augmentation strategy performed significantly worse if applied to unseen data. The IoU Scores, Dice coefficients for the tumor classes non-contrast-enhancing tumor, edema, and contrast-enhancing tumor, and overall accuracy are notably lower than those of models trained without augmentations. This suggests that augmenting the same case within the batch, as in Version 1, does not provide the necessary diversity for the model to generalize well to new, unseen cases.

This comparison highlights the limitations of augmentation strategy Version 1 in terms of generalization. The lack of variability between the augmented cases within the batch results in overfitting to the training data and poor performance on the evaluation dataset.

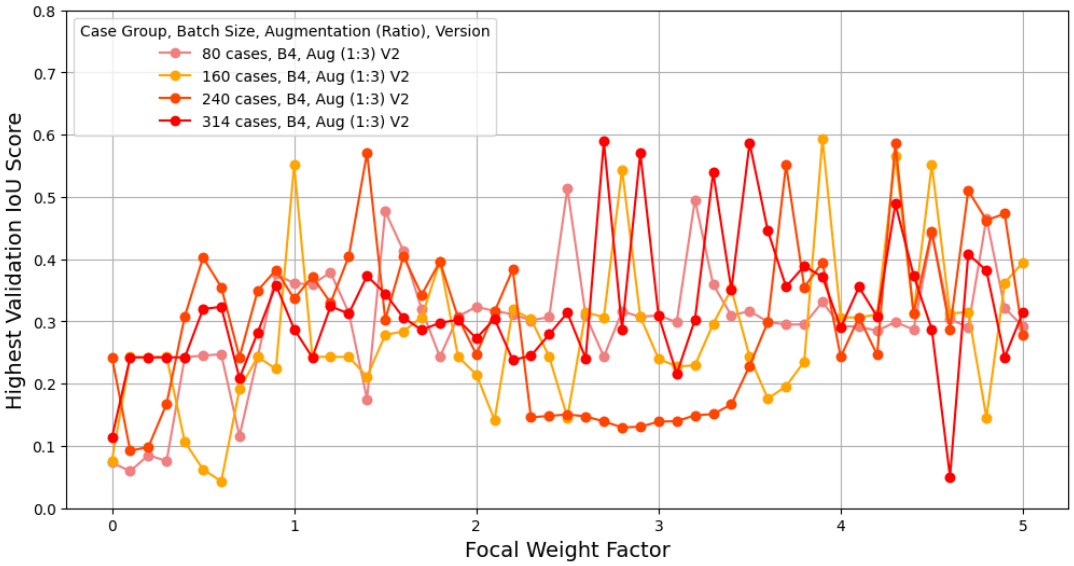


(a) Highest Validation IoU Scores for different Focal Weight Factors.

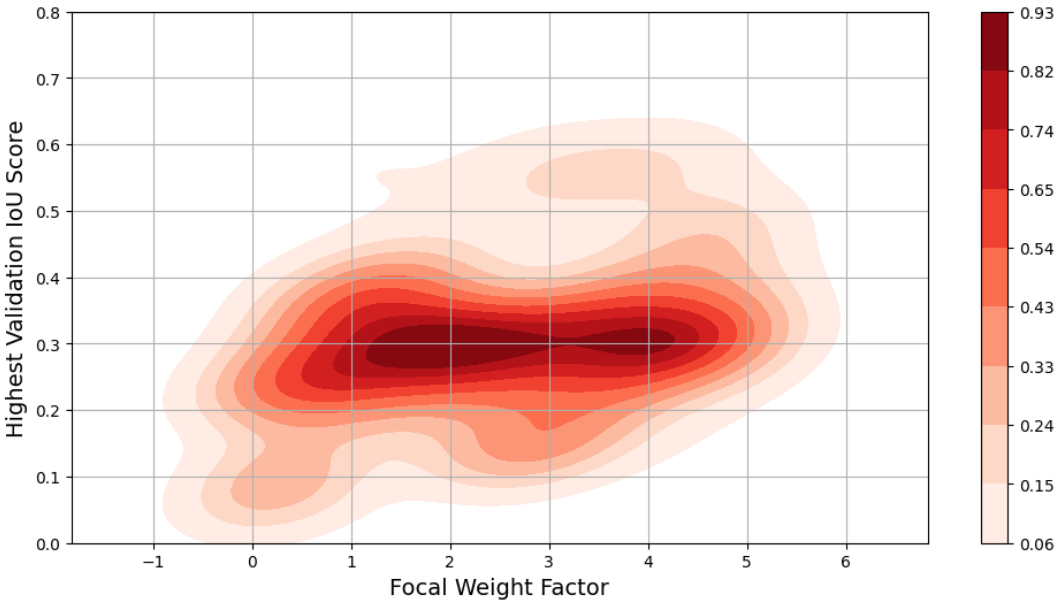


(b) Heatmap representation of the IoU Scores shown in (a).

Figure 7.12: (a) Highest validation IoU Scores for different Focal Weight Factors for Batch Size 4 with augmentation (Version 1), separated by Case Groups (80, 160, 240, and 314). The augmentation strategy generates three variations of the same case within a batch. (b) Heatmap visualization of the same data, showing the concentration of IoU Scores at a lower range, confirming that the Focal Weight Factor has little impact on performance.



(a) Highest Validation IoU Scores for different Focal Weight Factors.



(b) Heatmap representation of the IoU Scores shown in (a).

Figure 7.13: (a) Highest validation IoU Scores for different Focal Weight Factors for Batch Size 4 with augmentation (Version 2), separated by Case Groups (80, 160, 240, and 314). The augmentation strategy combines an original case with three random augmentations from other original cases. (b) Heatmap visualization of the same data, highlighting a broader distribution of IoU Scores and several peaks, particularly in larger Case Groups, indicating that random augmentations introduce beneficial variability into the training process.

Detection of Overfitting in the Training Process

The poor generalization of augmentation strategy Version 1 suggests that the models fail to learn robust features and instead memorize the training data. To further investigate this issue, the training dynamics were analyzed to determine whether overfitting played a role in the observed performance drop.

One of the primary goals was to develop a method for detecting overfitting during the training process. This phenomenon can be observed in models trained with augmentation strategy Version 1 on the Case Group 160, using Batch size 4. Figure 7.14 illustrates the training process, showing that the Dice coefficients for the non-contrast-enhancing and contrast-enhancing tumor exhibit a sigmoidal jump, reaching values above 0.8 within a few epochs, typically around 4 epochs. After this sudden increase, the curves saturate, indicating that the model has overfitted to the training data.

A key observation is that the training and validation Dice coefficients for Version 1 behave similarly, following the same trend closely, which reinforces the overfitting effect. In contrast, Figure 7.15 shows the training process for Version 2, where the training and validation Dice coefficients diverge. This reduction in overlap between training and validation performance indicates that Version 2 introduces more variability, which could potentially help the model generalize better to unseen data. However, it does not necessarily guarantee improved generalization, as the validation Dice coefficients may still reflect suboptimal recognition of validation data.

To systematically detect overfitting, a method was developed using a grid search to identify optimal thresholds that can distinguish models trained with augmentation strategy Version 1 from all other models (including Version 2). The following parameters were used for the thresholding algorithm:

- **Minimum validation threshold:** The lowest validation Dice coefficient score detected before the sigmoidal jump is 0.242. This indicates the baseline performance before the model begins to improve significantly.
- **Maximum validation threshold:** The highest validation Dice coefficient score after the sigmoidal jump is 0.66. This value represents the model's upper limit in terms of generalization performance for the non-contrast-enhancing and contrast-enhancing tumor.
- **Jump threshold:** The difference between the minimum and maximum validation Dice coefficients during the jump is 0.55. This threshold marks the size of the performance improvement during the sigmoidal transition.
- **Epoch difference threshold:** The number of epochs over which the sigmoidal jump occurs is set at 4 epochs. This threshold ensures that rapid changes in Dice coefficients, indicating overfitting, can be detected.

- **Steepness threshold:** The transition from low to high Dice coefficients during the sigmoidal jump is characterized by a steep increase, measured both in terms of slope (0.8) and duration (4 epochs). These parameters help differentiate between models with a sudden performance surge, indicative of overfitting, and those with a more gradual learning curve.

A detection algorithm was developed based on sigmoidal curve fitting and predefined thresholds (validation threshold, slope, jump size, and duration), enabling the identification of overfitting models. Using this approach, 31 out of 40 models trained with augmentation strategy Version 1 and Batch size 4 were correctly classified as overfitting. Importantly, there were no false positives. The remaining 9 models were trained for 4 epochs or fewer. Since both the epoch difference threshold and steepness duration threshold were set at 4, models with such a short training duration could not be detected by the algorithm.

Performance Evaluation of Models Trained with Augmentation Strategy Version 1 Figure 7.16 presents the evaluation results of the 10 best models trained with Batch size 4 and augmentation Version 1 on Case Group 314. The boxplots show key metrics, including the IoU Score, Dice coefficients for non-contrast-enhancing tumor, edema, and contrast-enhancing tumor, as well as accuracy. These results indicate significantly poorer generalization compared to models trained without augmentations, which performed better in Figure 7.7.

Sigmoidal Curve Fitting for Overfitting Detection To improve the detection of overfitting, a sigmoidal curve was fitted to the Dice coefficient score during training, as illustrated in Figure 7.17. The dashed red lines represent the fitted sigmoidal curves, while the green vertical lines indicate the turning points, determined at Epoch 9 for the non-contrast-enhancing tumor and Epoch 7 for the contrast-enhancing tumor. This sigmoidal fitting process was applied to the model from Figure 7.14 (Batch size 4, augmentation strategy Version 1, Case Group 160) and was used to calculate the threshold values for the overfitting detection algorithm.

7.9 Best Models (IoU Score & Custom Weighted Dice Score)

In line with the research questions, one of the primary goals was to identify the best-performing models based on the IoU metric as applied to the evaluation dataset of 108 unseen cases. Initially, models were selected purely based on their IoU Scores, which directly measure the overlap between predicted and ground truth segmentations of tumor classes. This selection process led to the results shown in Figure 7.18, where the top 10 models with the highest IoU Scores are displayed. However, further analysis revealed that relying solely on the IoU Score does not fully capture the clinical relevance of tumor segmentation, particularly in distinguishing critical tumor regions such as the non-contrast-enhancing and the contrast-enhancing tumor. The variation in the Interquartile

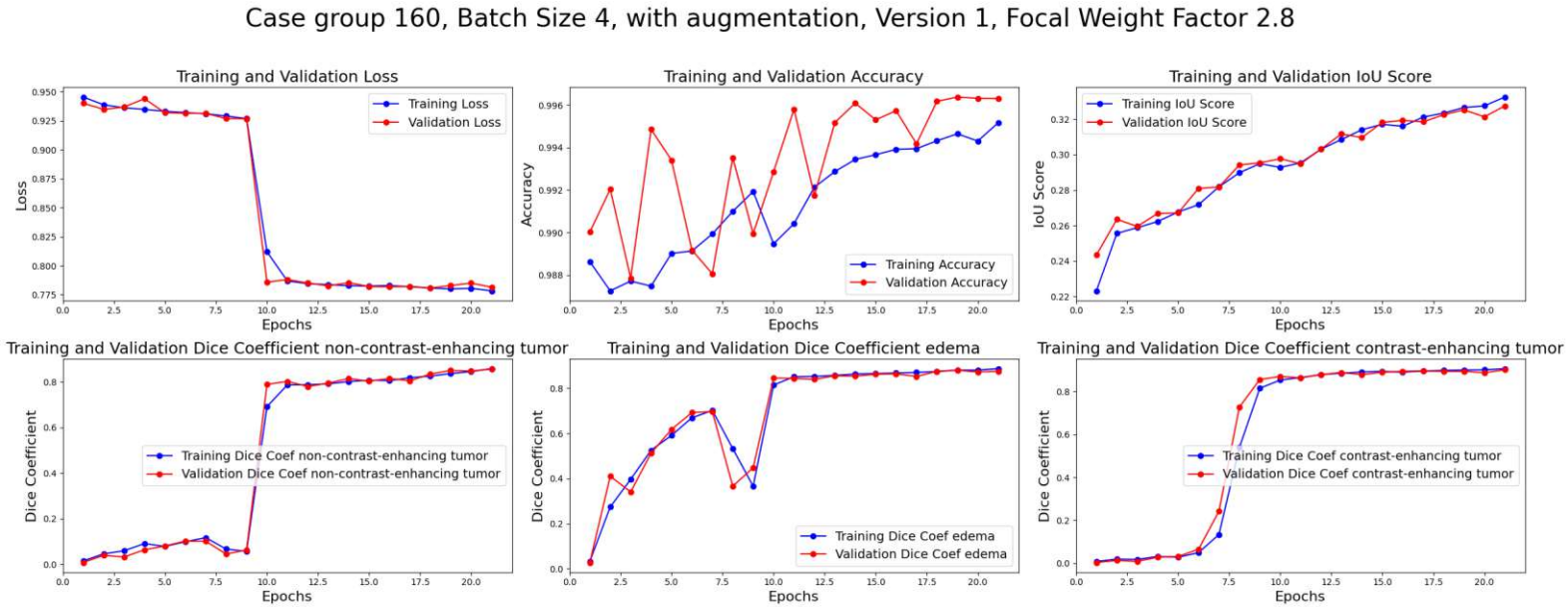


Figure 7.14: Training process for a model with Batch size 4, augmentation strategy Version 1, and Case Group 160. The Dice coefficients for the non-contrast-enhancing tumor and contrast-enhancing tumor show a sigmoidal jump to values above 0.8 within a few epochs, after which the curves saturate, indicating overfitting.

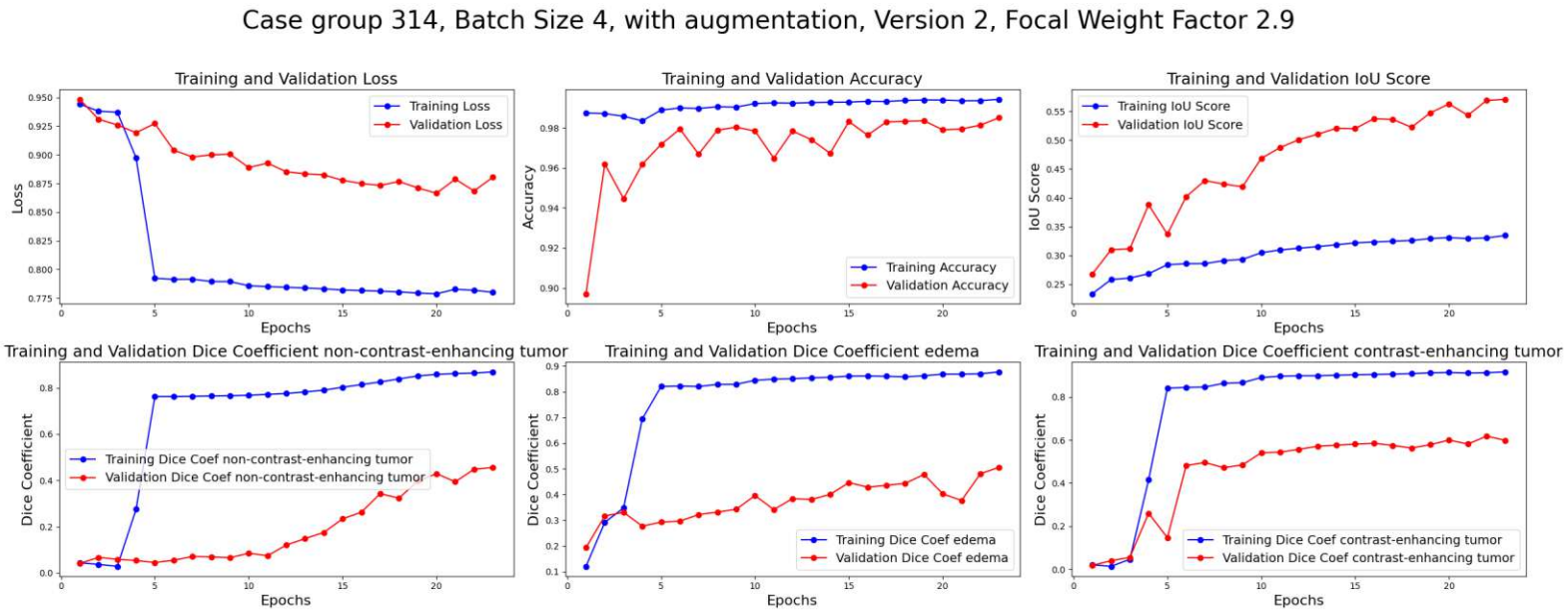


Figure 7.15: Training process for a model with Batch size 4, augmentation strategy Version 2, and Case Group 314. In contrast to Version 1, the training and validation Dice coefficients diverge, reducing overfitting and improving generalization.

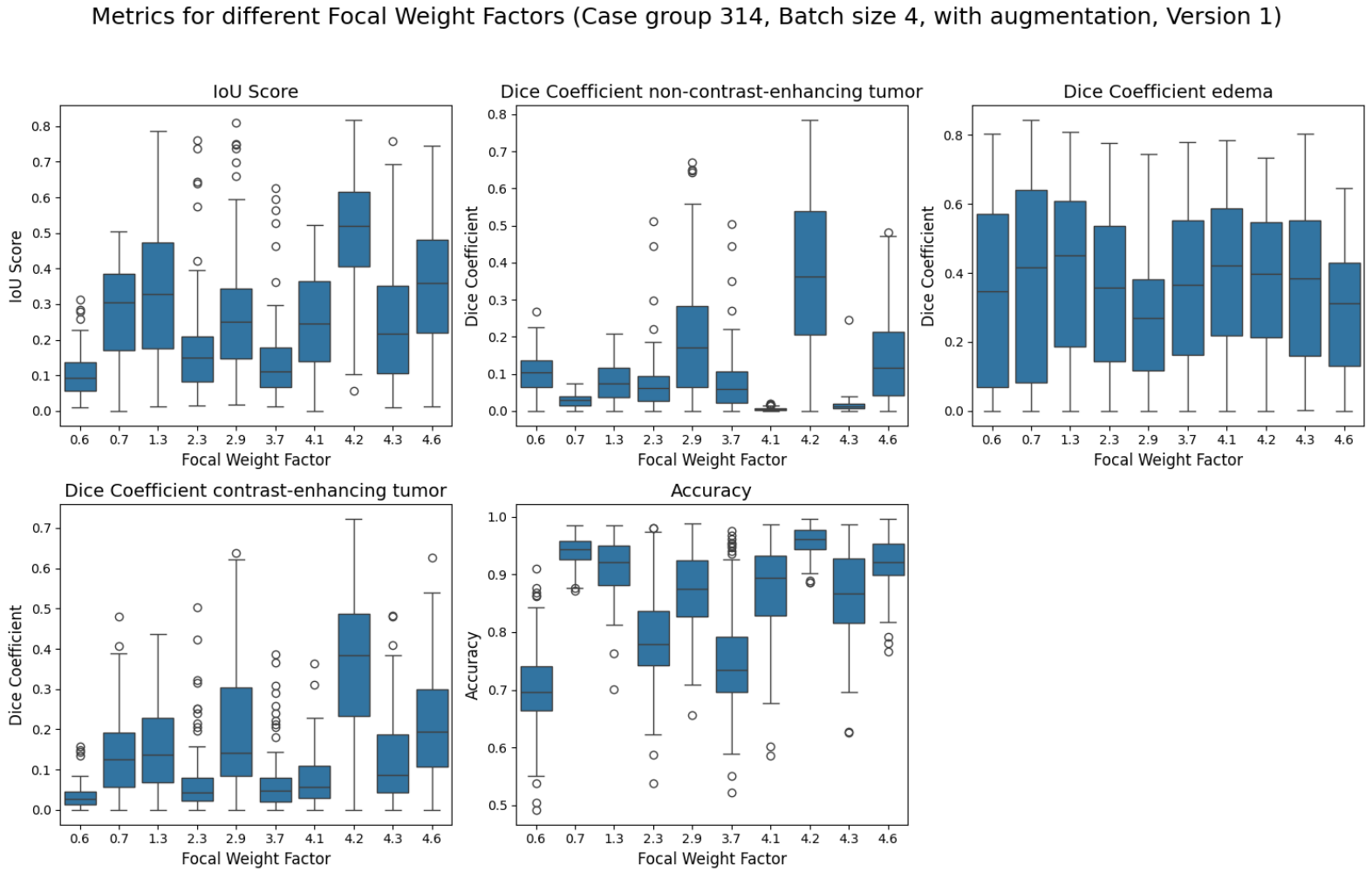


Figure 7.16: Boxplots showing key evaluation metrics for the 10 best models trained with Batch size 4 and augmentation Version 1 for Case Group 314, applied to the 108 unseen evaluation cases. The results demonstrate significantly poorer generalization compared to models trained without augmentations (as seen in Figure 7.7).

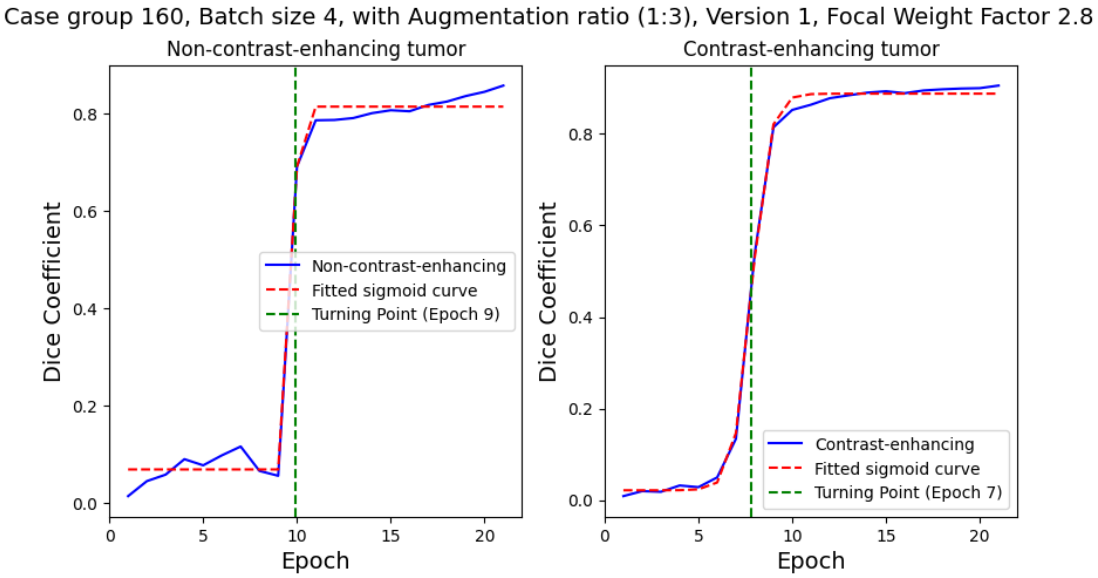


Figure 7.17: Sigmoidal curve fitting for the Dice coefficients of non-contrast-enhancing and contrast-enhancing tumor. The green vertical lines mark the turning points (Epoch 9 and Epoch 7), helping to define threshold values for the overfitting detection algorithm. This example is based on the model from Figure 7.14 (Batch size 4, augmentation strategy Version 1, Case Group 160).

Range (IQR) observed in Figure 7.18 indicates that models achieving high IoU Scores can still exhibit significant performance inconsistencies. This suggests that a more refined metric is needed to ensure robust segmentation of clinically relevant tumor classes. To address these limitations, the Custom Weighted Dice Score was introduced (Section 4.1.6), allowing for a more nuanced evaluation that prioritizes segmentation accuracy in the most clinically significant tumor regions. This metric was subsequently applied to the top-performing models ranked by IoU, with the aim of selecting models that achieve a more balanced segmentation quality across all tumor subregions.

Figures 7.19 and 7.20 demonstrate the impact of this refined evaluation strategy. Figure 7.19 presents the results after applying the Custom Weighted Dice Score to the top 20 models based on the IoU Score, while Figure 7.20 extends this analysis to the top 30 models. The key observations from these results are:

1. **Top-performing models:** Across both analyzed sets (top 20 and top 30), the best-performing model consistently includes Batch Size 1, augmentation ratio of 1:1, and Case Group 314. This model consistently outperforms others based on both IoU Score and the Custom Weighted Dice Score.

2. **Case Group dominance:** Models trained with Case Group 314 constitute the majority (90%) of the top-ranked models. Among these, models trained with

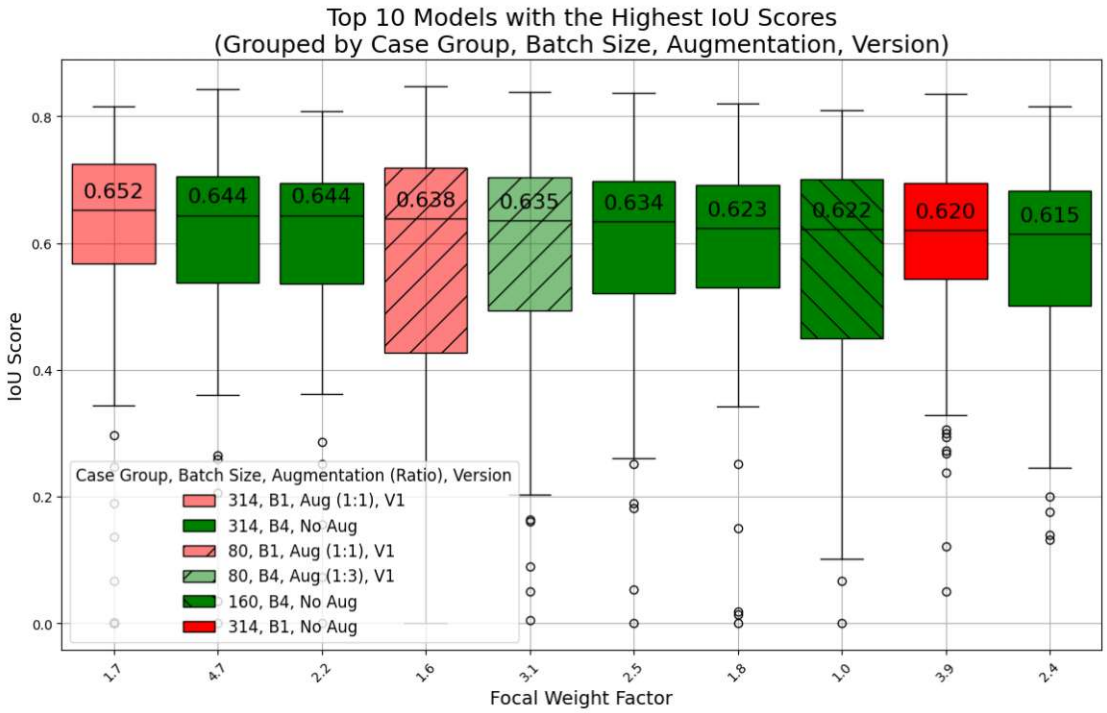


Figure 7.18: Boxplots of the 10 models with the highest IoU Scores, evaluated on the 108 unseen cases. While these models achieve high IoU Scores, the IQR variations indicate inconsistent segmentation quality across different cases, emphasizing the need for a more refined evaluation metric.

Batch Size 4 account for 80% of the highest-ranked models, further reinforcing the advantage of larger training datasets.

3. **Minimal augmentation:** Although models with augmentations appear in the top ranks, they only represent 10–20% of the best-performing models. This suggests that, while augmentation can be beneficial, its impact is less pronounced when applied to the largest Case Groups.

4. **Robustness of the score:** The Custom Weighted Dice Score results in a more stable and clinically meaningful ranking of models, as evidenced by the reduced IQR variations in Figures 7.19 and 7.20. This underscores the score's ability to reduce inconsistencies observed when models were ranked solely based on their IoU Scores.

It is important to note that the horizontal axis in Figures 7.18, 7.19, and 7.20 represents different Focal Weight Factors corresponding to the best-performing models evaluated on unseen data. The presence of multiple models with similar IoU Scores but different Focal Weight Factors suggests that no single Focal Weight Factor consistently outperforms

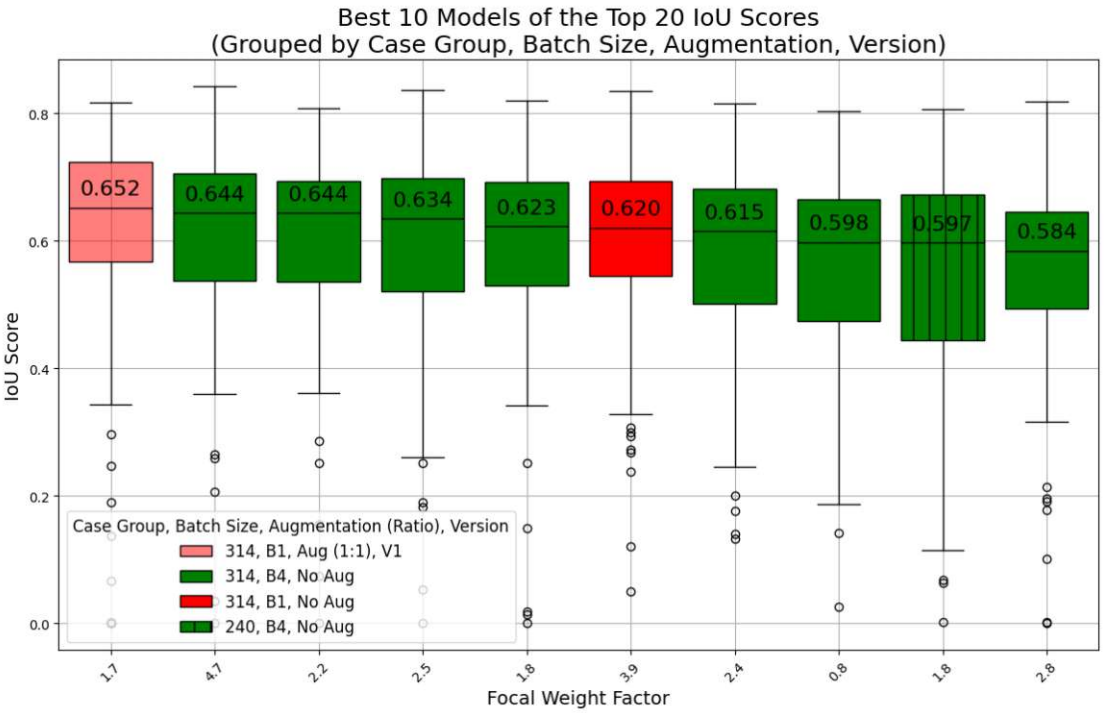


Figure 7.19: Boxplots of the best models from the top 20 IoU Scores, after applying the Custom Weighted Dice Score. The best-performing model is characterized by Batch Size 1, augmentation ratio of 1:1, and Case Group 314. The Custom Weighted Dice Score stabilizes model ranking by reducing performance variability, as reflected in the reduced IQR variations.

others. Instead, multiple Focal Weight Factor values lead to comparable segmentation results, indicating that the model is not highly sensitive to a specific Focal Weight Factor setting. This observation highlights that while Focal Weight Factor tuning influences segmentation performance, its effect remains within a certain range without exhibiting a clear, consistent trend.

7.10 Training Case Group Size & Performance

A key aspect of evaluating the glioblastoma segmentation models was examining how the number of training cases affected model performance. This investigation focused on understanding how increasing the amount of data impacted segmentation accuracy, particularly in the clinically relevant tumor classes: non-contrast-enhancing tumor, edema, and contrast-enhancing tumor.

The comparison was made across the four Case Groups: 80, 160, 240, and 314. These Case Groups were carefully designed to maintain proportional class distributions across both the training and validation datasets, as discussed in Section 5.1. The aim was to

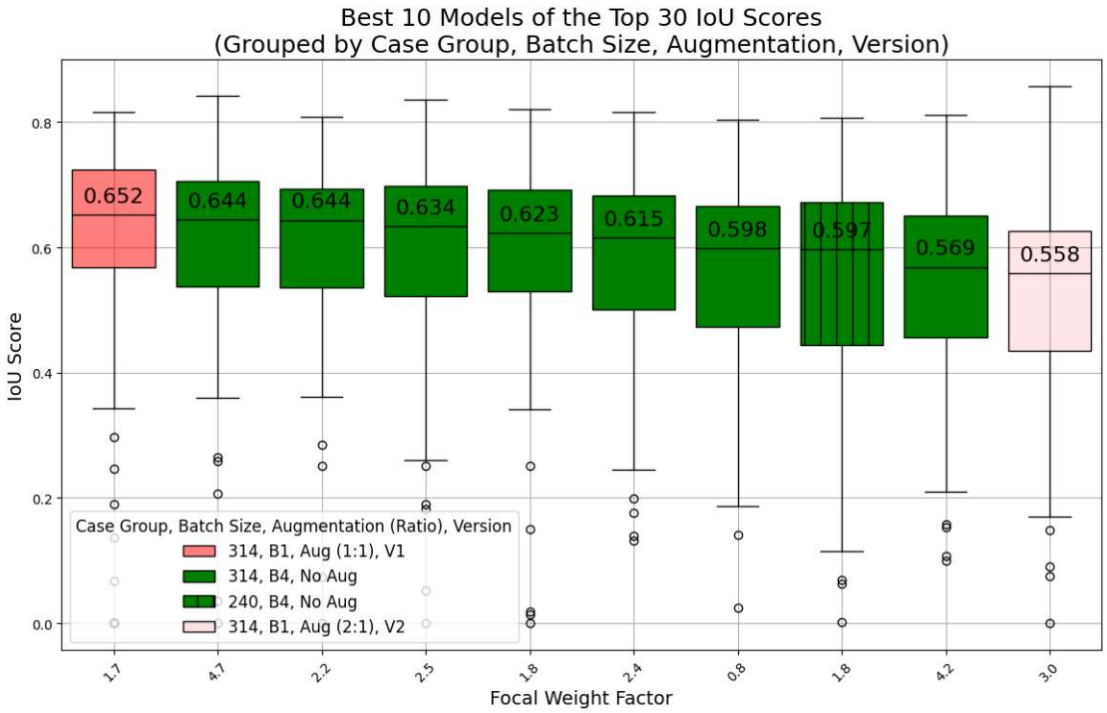


Figure 7.20: Boxplots of the best models from the top 30 IoU Scores, after applying the Custom Weighted Dice Score. The observed trends remain consistent, with Batch Size 4 and Case Group 314 dominating the top ranks. Models without augmentation generally outperform those with augmentations, except for the top model. The Custom Weighted Dice Score further validates these observations by providing a more consistent ranking.

assess how the models' performance improves with more training data and whether larger datasets provide significant benefits in segmenting these critical tumor classes.

The evaluation of model performance was based on several key metrics: the IoU Score, Dice coefficients for the three tumor classes non-contrast-enhancing tumor, edema, and contrast-enhancing tumor, and the overall accuracy. For each Case Group, the best-performing model was selected, and its results were compared across these metrics.

As shown in Figure 7.21, the comparison between the different Case Groups reveals several trends regarding the impact of increasing training data. The analysis below highlights how each metric behaved as the number of cases increased:

IoU Score:
The IoU Score generally improves as the number of cases increases, with the model trained on Case Group 314 achieving the highest median IoU (0.652). However, the model trained on Case Group 240 exhibits a slight dip in performance with a lower IoU (0.601), indicating a non-linear trend where larger training sets mostly improve generalization and segmentation accuracy, but with an unexpected drop at Case Group

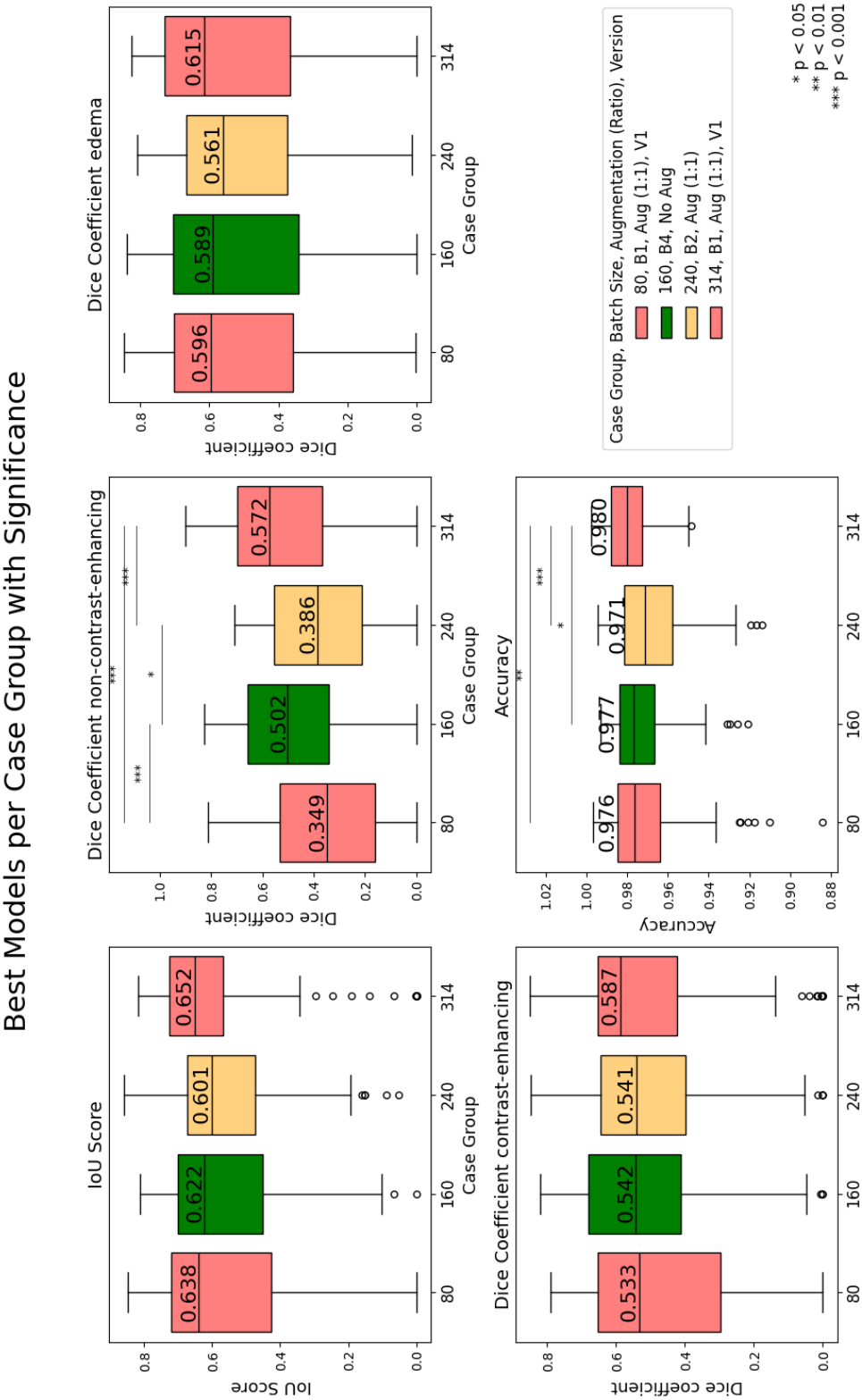


Figure 7.21: Comparison of best models per Case Group with statistical significance. The figure presents the best models for each Case Group (80, 160, 240, 314) based on their IoU Scores and Dice coefficients for the three tumor classes. Significant differences between Case Groups are indicated by p-values (*p $<$ 0.05, **p $<$ 0.01, **p $<$ 0.001), particularly in the non-contrast-enhancing tumor class, where increasing the number of cases led to notable improvements in performance.

240.

Dice Coefficient – non-contrast-enhancing tumor:
The necrotic tumor Dice coefficient exhibits the most substantial improvement, particularly between Case Group 80 (0.349) and Case Group 314 (0.572). The differences here are statistically significant ($p < 0.001$), indicating that more cases provide the model with a better representation of this critical tumor class.

Dice Coefficient – edema:
While the Dice coefficient for edema also improves with increasing case numbers, the improvement is more gradual, rising from 0.596 with Case Group 80 to 0.615 with Case Group 314. The variation across models is less pronounced, but the trend still favors larger datasets.

Dice Coefficient – contrast-enhancing tumor:
The contrast-enhancing tumor Dice coefficient shows a steady improvement as well, from 0.533 with Case Group 80 to 0.587 with Case Group 314. This is particularly important given the clinical significance of accurately segmenting the contrast-enhancing tumor region.

Accuracy:
The accuracy metric also increases as more cases are used, with a modest improvement from 0.976 (Case Group 80) to 0.980 (Case Group 314). Although the differences in accuracy are relatively small, the statistical tests reveal significant differences between the Case Groups, particularly between 80 and 314 ($p < 0.001$).

The analysis reveals that increasing the number of training cases generally improves performance, particularly in the segmentation of the three tumor classes. Significant differences are especially notable in the non-contrast-enhancing tumor Dice coefficient and accuracy between smaller and larger Case Groups, indicating the importance of a larger dataset for accurate segmentation.

7.11 Segmentation Evaluation: Best Models per Case Group

To assess the impact of different dataset sizes on segmentation performance, the best-performing models from each Case Group (80, 160, 240, and 314) were applied to the same unseen test case. The selected models achieved the highest IoU Score and Dice coefficients for the three tumor classes: non-contrast-enhancing tumor, edema, and contrast-enhancing tumor. The corresponding numerical values can be found in Table 7.2.

The segmentation results are visualized in Figure 7.22, which provides a structured comparison of the segmentation performance of the four different Case Groups. The first row displays the input MRI sequences, including FLAIR, T1ce, and T2, along with the ground truth segmentation. The second row presents the predicted segmentations for the same patient using the best models from each Case Group. The third row overlays

these segmentations onto the corresponding T2-weighted image for anatomical reference, highlighting how well each model aligns with the actual tumor anatomy. The predicted segmentations use a consistent color scheme, where the non-contrast-enhancing tumor is represented in green, tumor edema in yellow, and the contrast-enhancing tumor in brown.

7.11.1 Performance Across Case Groups

Each model exhibited different segmentation accuracies, primarily influenced by the number of training cases and the presence of augmentations. The model from Case Group 80 (Batch Size 1, augmentation ratio of 1:1, Version 1) achieved the highest IoU Score (0.8151) among all Case Groups. The Dice coefficients were 0.7178 for the non-contrast-enhancing tumor, 0.8442 for edema, and 0.6685 for the contrast-enhancing tumor. Despite the high overall performance, the model showed slightly lower accuracy in segmenting the non-contrast-enhancing tumor, potentially indicating challenges in accurately delineating smaller non-enhancing regions when trained on a limited dataset.

The model from Case Group 160 (Batch Size 4, no augmentation) showed a slight decrease in IoU Score (0.7606). However, the Dice coefficients remained stable, with 0.7182 for the non-contrast-enhancing tumor, 0.7945 for edema, and 0.7069 for contrast-enhancing tumor. The model maintained a high accuracy (0.9737), suggesting that training with a larger dataset without augmentation still leads to robust segmentation performance.

For Case Group 240 (Batch Size 2, augmentation ratio of 1:1, Version 1), a further decline in performance was observed, with an IoU Score of 0.6901. The Dice coefficients were 0.4870 for non-contrast-enhancing tumor, 0.7528 for edema, and 0.5813 for contrast-enhancing tumor. The drop in non-contrast-enhancing and contrast-enhancing tumor segmentation performance suggests that augmentation introduced additional variability, which may have negatively impacted the model's ability to generalize effectively to unseen data.

In contrast, the model from Case Group 314 (Batch Size 1, augmentation ratio of 1:1, Version 1) demonstrated strong overall performance, with an IoU Score of 0.7966 and high Dice coefficients for non-contrast-enhancing tumor (0.7478), edema (0.8255), and contrast-enhancing tumor (0.6653). The model's accuracy (0.9773) remained consistently high, reinforcing the observation that larger training datasets, combined with augmentations, enhance model generalization across all tumor subregions.

7.11.2 Key Observations and Interpretation

Figure 7.22 illustrates the impact of Case Group size on segmentation quality. Interestingly, the best segmentation performance, based on visual inspection, is achieved by the models from Case Group 80 and Case Group 314, despite the considerable difference in the number of training cases. Both of these models were trained with Batch Size 1, which suggests that this configuration may contribute to improved segmentation accuracy. The strong performance of Case Group 80 further indicates that a smaller dataset does not

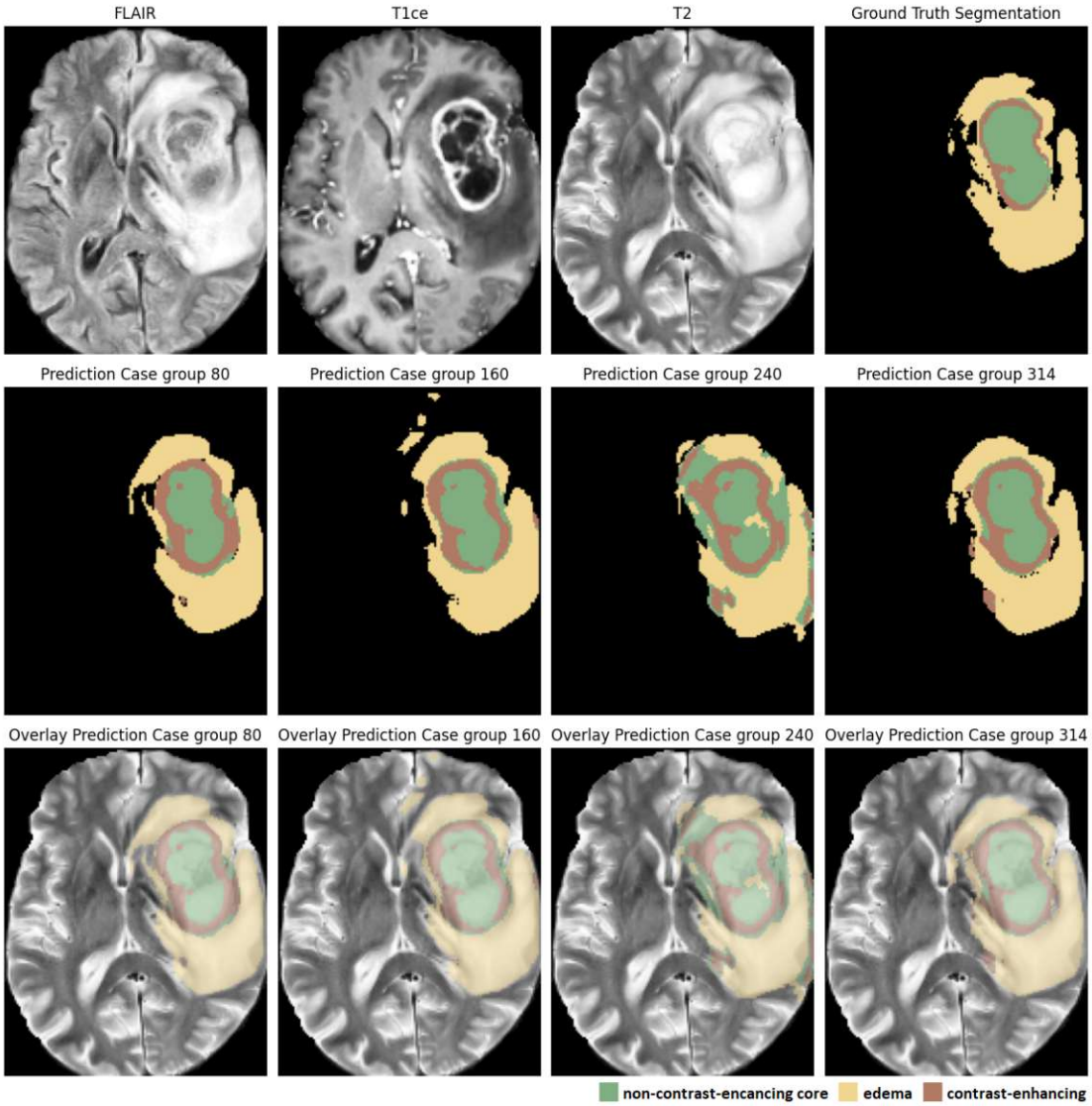


Figure 7.22: Visualization of predicted glioblastoma segmentation across the four Case Groups. The first row shows the input MRI sequences, including FLAIR, T1ce, T2, and the ground truth segmentation mask. The second row presents the predicted segmentation results from the four best-performing models for each Case Group (80, 160, 240, and 314). The third row displays the predicted segmentations overlaid on the corresponding T2-weighted image for anatomical reference. The segmentations highlight the non-contrast-enhancing tumor (green), tumor edema (yellow), and contrast-enhancing tumor (brown), illustrating the influence of Case Group size on segmentation performance.

				Dice Coefficients			
Case Group	**Batch Size**	**Aug.**	**IoU Score**	**non-contrast enhanc.**	**edema**	**contrast enhanc.**	**Accu.**
80	1	Aug (V1)	0.8151	0.7178	0.8442	0.6685	0.9789
160	4	no Aug	0.7606	0.7182	0.7945	0.7069	0.9737
240	2	Aug (V1)	0.6901	0.4870	0.7528	0.5813	0.9602
314	1	Aug (V1)	0.7966	0.7478	0.8255	0.6653	0.9773

Table 7.2: Segmentation performance of best models per Case Group. This table compares the IoU Score, Dice coefficients of the three tumor classes, and accuracy across the best-performing Case Groups 80, 160, 240, and 314. augmentation strategies (with and without, including Version) and Batch size are included to show their impact on model performance.

necessarily lead to inferior segmentation quality. Despite its limited number of training cases, this model performs on par with Case Group 314, highlighting the importance of model optimization over absolute dataset size.

While augmentation plays a role in improving segmentation, its impact is not uniform across all Case Groups. In Case Group 314, augmentation appears to enhance generalization, whereas in Case Group 240, it introduces greater variability in segmentation performance, particularly in certain tumor casses. This suggests that augmentation can be beneficial, but its effectiveness depends on other factors such as Batch size and dataset composition.

The anatomical overlays in Figure 7.22 provide further insight into how well each model's predictions align with the actual tumor structures. The comparison across Case Groups highlights that both larger training datasets and well-optimized smaller datasets can yield robust segmentation results. These findings emphasize that fine-tuned model configurations, particularly Batch size selection, may be just as crucial as dataset size in achieving optimal segmentation performance.

7.12 Comparison with State-of-the-Art Segmentation Benchmarks

The performance of the best-performing model in this study (Case Group 314, Batch Size 1, with an augmentation ratio of 1:1) was compared to state-of-the-art results reported in challenges such as BraTS 2018. According to Braid et al. [BTR+20], Dice coefficients for glioblastoma segmentation typically range from 0.75 to 0.93 for edema, 0.77 to 0.91 for the non-contrast-enhancing tumor, and 0.67 to 0.83 for the contrast-enhancing tumor. The Dice coefficients achieved by the best model in this study were 0.825 for edema, 0.748 for the non-contrast-enhancing tumor, and 0.665 for the contrast-enhancing tumor.

These results indicate that the model performs competitively with state-of-the-art meth-

ods, particularly for edema segmentation, where the Dice coefficient of 0.825 lies within the upper range of reported values. The performance for the non-contrast-enhancing tumor is slightly below the benchmark range, while the contrast-enhancing tumor segmentation approaches the lower bound of state-of-the-art results.

The slight discrepancies in performance can be attributed to the simplified 3D U-Net architecture employed in this study, which was intentionally chosen to balance computational efficiency and model interpretability (Section 6.7). Additionally, differences in preprocessing strategies, augmentation techniques, and the specific composition of the training dataset likely influenced the observed outcomes.

7.13 Data Augmentation Impact on Evaluation

The comparison is made between models trained on augmented data with fewer original cases (shown in lighter shades on the left) and models trained on non-augmented data with a larger number of original cases (shown in darker shades on the right). The models represent the best 10 ones based on the highest IoU Scores from the training process and are applied to unseen data for evaluation. The numbers used for further interpretation can be found in Table 7.3.

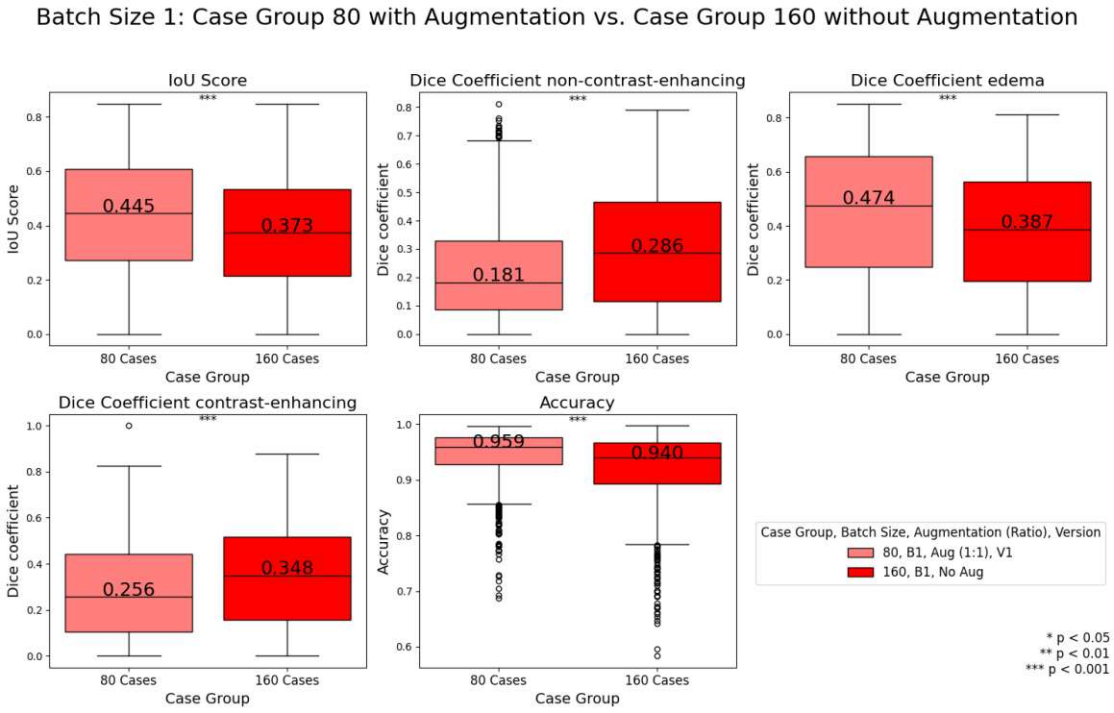


Figure 7.23: Comparison of Batch size 1: Case Group 80 with augmentation vs. Case Group 160 without augmentation. The augmented models perform better in IoU Score, edema segmentation, as well as in accuracy, while the non-augmented models perform better in non-contrast-enhancing and contrast-enhancing tumor segmentation.

Batch size 1: Case Group 80 with augmentation vs. Case Group 160 without augmentation (Figure 7.23)

- **IoU Score:** Interestingly, Case Group 80 with 80 additional augmented cases performs better with a higher median IoU Score (0.445) than the non-augmented Case Group 160 (0.373). This result is counterintuitive, as one might expect more original cases to lead to better performance. However, the diversity introduced by data augmentation may have led to improved generalization in the Case Group 80 models.
- **Dice coefficients of non-contrast-enhancing tumor, edema and contrast-enhancing tumor:** The augmented Case Group 80 models outperform the non-augmented Case Group 160 models in segmenting non-contrast-enhancing tumor and edema, while the non-augmented models perform better for the contrast-enhancing tumor.
- **Accuracy:** The augmented models also achieve higher accuracy (0.959 vs. 0.940), further supporting the idea that augmentation enhances generalization if the number of original cases is limited.

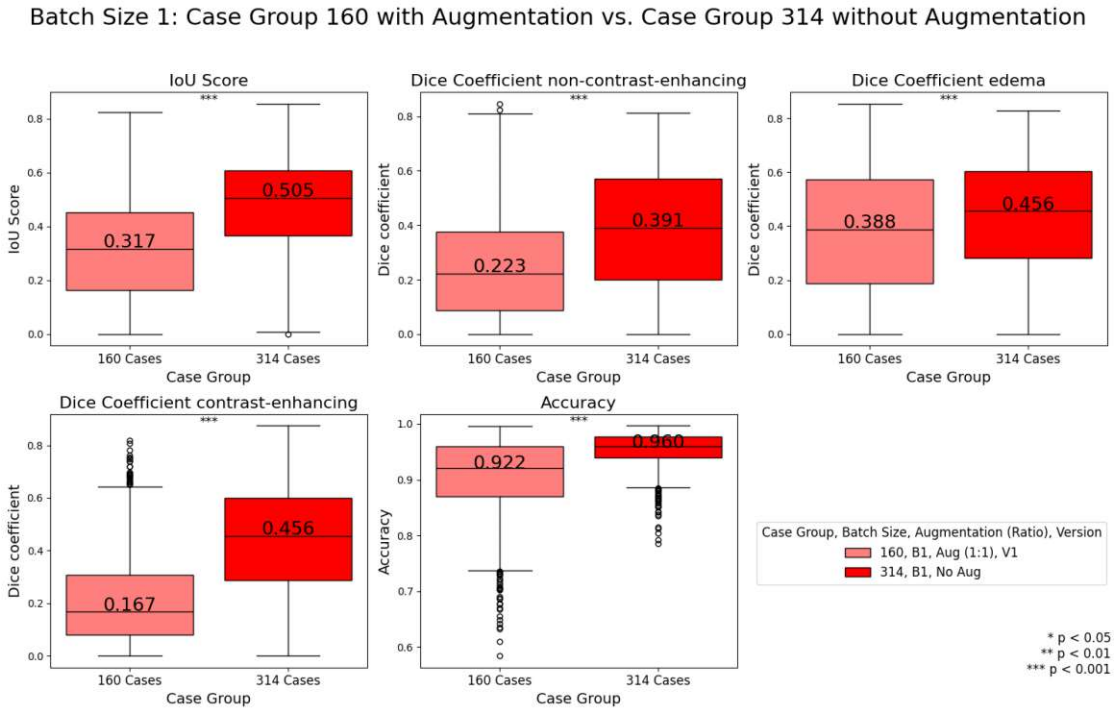


Figure 7.24: Comparison of Batch size 1: Case Group 160 with augmentation vs. Case Group 314 without augmentation. The non-augmented models outperform the augmented models across all metrics, showing that augmentation is less beneficial if the number of original cases is larger.

Batch size 1: Case Group 160 with augmentation vs. Case Group 314 without augmentation (Figure 7.24)

- **IoU Score:** The non-augmented Case Group 314 models outperform the augmented Case Group 160 models, achieving a significantly higher IoU Score (0.505 vs. 0.317). This suggests that with a sufficient number of original cases, augmentation becomes less necessary and may even reduce performance.
- **Dice coefficients:** Segmentation of non-contrast-enhancing and contrast-enhancing tumor are significantly better in the non-augmented Case Group 314 models. The larger number of original cases provides more robust training for the model.
- **Accuracy:** The non-augmented Case Group 314 models also achieve higher accuracy (0.960 vs. 0.922), indicating better generalization.

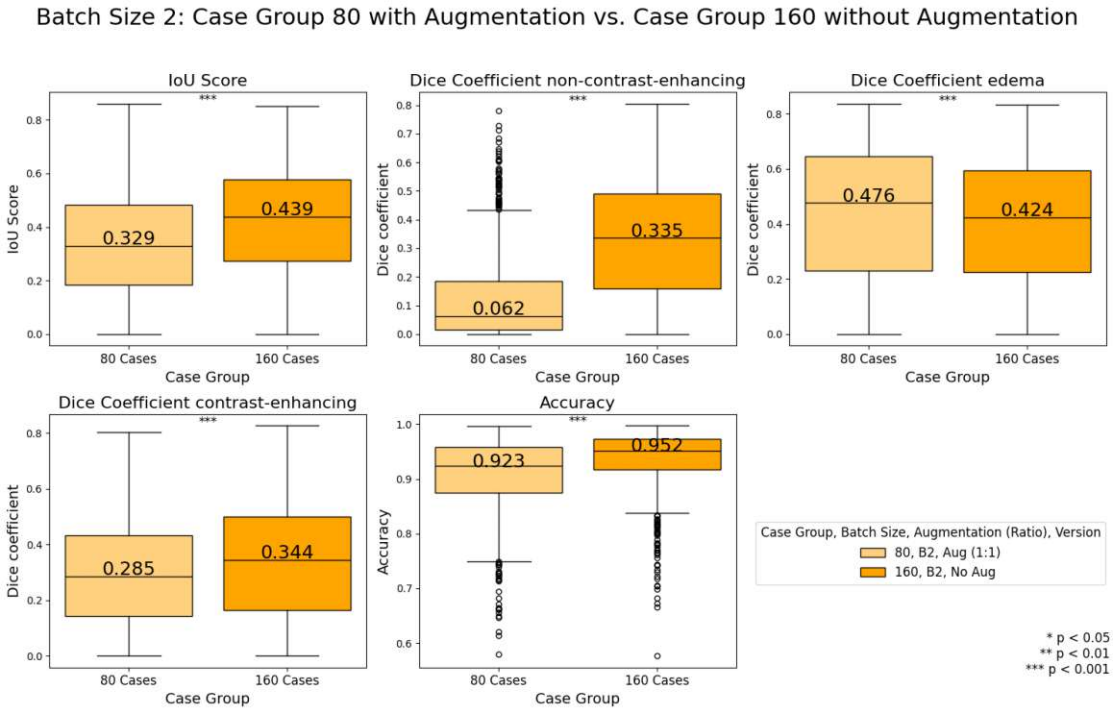


Figure 7.25: Comparison of Batch size 2: Case Group 80 with augmentation vs. Case Group 160 without augmentation. The non-augmented models perform better across IoU Score, non-contrast-enhancing and contrast-enhancing tumor segmentation, as well as accuracy, suggesting that Batch size 2 benefits more from a larger number of original cases than from augmentation.

Batch size 2: Case Group 80 with augmentation vs. Case Group 160 without augmentation (Figure 7.25)

- **IoU Score:** With Batch size 2, the non-augmented Case Group 160 models outperform the augmented Case Group 80 models (0.439 vs. 0.329). This suggests that the benefits of augmentation diminish as Batch size increases, making original data more beneficial.
- **Dice coefficients:** The non-contrast-enhancing tumor Dice coefficient is significantly higher for the non-augmented Case Group 160, further supporting the advantage of using more original data in this scenario.
- **Accuracy:** Accuracy also follows this trend, with the non-augmented models achieving higher accuracy (0.952 vs. 0.923).

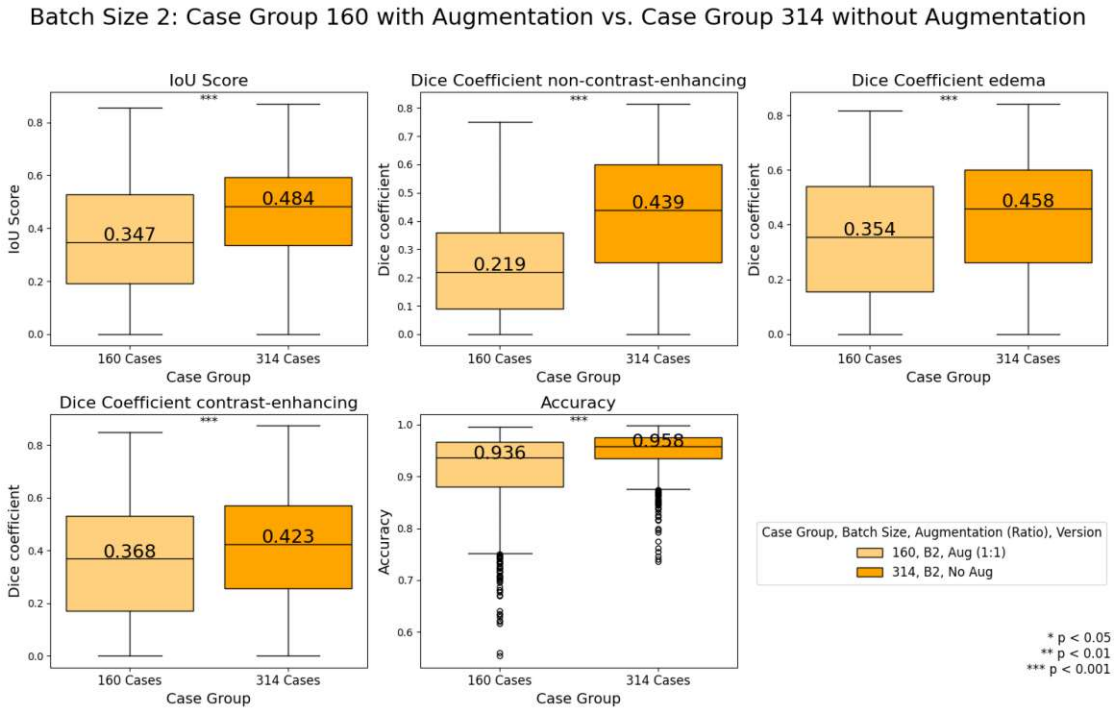


Figure 7.26: Comparison of Batch size 2: Case Group 160 with augmentation vs. Case Group 314 without augmentation. The non-augmented Case Group 314 consistently outperform the augmented Case Group 160 across all metrics, emphasizing the diminishing returns of augmentation with increasing original case numbers.

Batch size 2: Case Group 160 with augmented cases vs. Case Group 314 without augmented cases (Figure 7.26)

- **IoU Score:** The non-augmented models trained on Case Group 314 show a substantial improvement in IoU Score (0.484 vs. 0.347 for Case Group 160 with augmented cases). This supports the notion that larger datasets without augmentation outperform smaller augmented datasets if Batch size 2 is used.
- **Dice coefficients:** Segmentation of non-contrast-enhancing and contrast-enhancing tumor is more accurate in the non-augmented models, reflecting the importance of original data if available in higher volumes.
- **Accuracy:** As expected, accuracy is higher for the non-augmented Case Group 314 models (0.958 vs. 0.936).

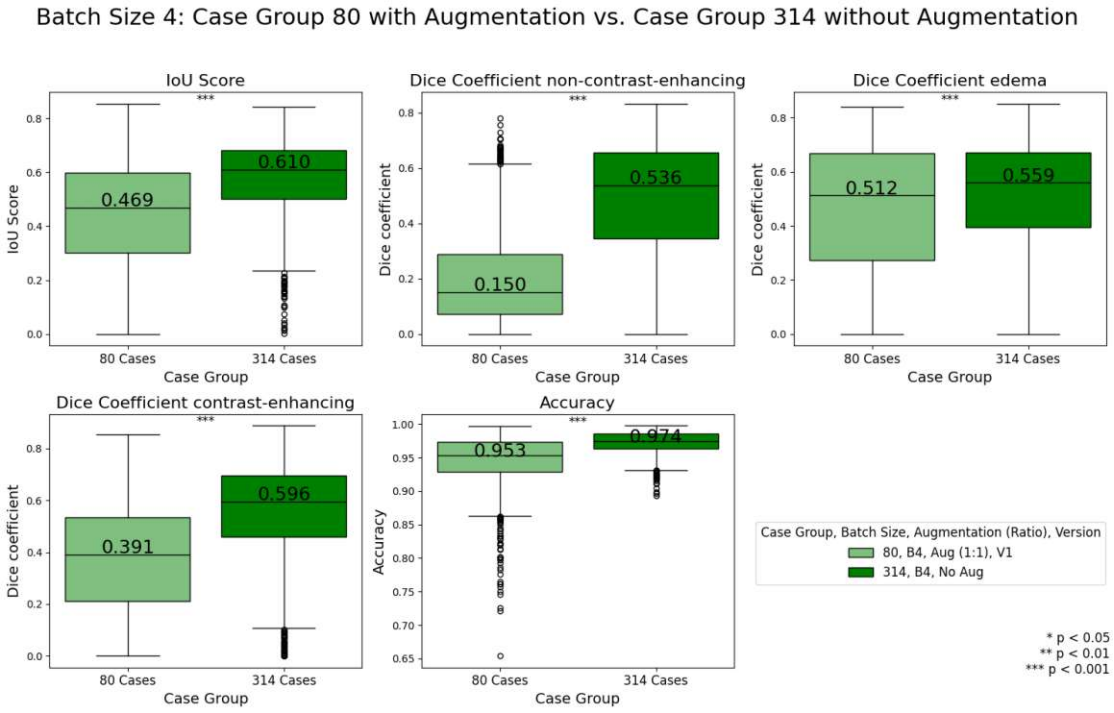


Figure 7.27: Comparison of Batch size 4: Case Group 80 with augmentation vs. Case Group 314 without augmentation. The 314 non-augmented models perform better in all metrics, particularly for non-contrast-enhancing and contrast-enhancing tumor segmentation, and accuracy.

Batch size 4: Case Group 80 with augmentation vs. Case Group 314 without augmentation (Figure 7.27)

- **IoU Score:** The non-augmented Case Group 314 models consistently outperform their augmented counterparts, achieving a significantly higher IoU Score (0.610 vs. 0.469 for the 80 augmented cases). This suggests that augmentation is less beneficial if a large Batch size is used, as the model can leverage the high volume of original cases for better generalization.
- **Dice coefficients:** The non-augmented models show superior performance in segmenting non-contrast-enhancing tumor, contrast-enhancing tumor, and edema.
- **Accuracy:** The non-augmented Case Group 314 models also achieve the highest accuracy (0.974 vs. 0.953).

General Insights

These findings suggest that data augmentation can significantly improve model performance if the number of original cases is limited, as observed in the Case Group 80 vs. 160 comparison with Batch size 1. However, as the number of original cases increases (e.g., to 160 or 314), augmentation becomes less beneficial, and non-augmented models tend to outperform the augmented ones, especially if lager Batch sizes are used. This supports the idea that augmentation serves as a substitute for limited data but loses its advantage as more original cases become available.

7.14 Case Number Impact on Metrics by Augmentation

The central research question here is: Does an increase in the number of cases result in improved model performance? Intuitively, one would hypothesize that training on more cases should enhance the model's ability to generalize. However, this hypothesis needs to be tested separately for augmented and non-augmented models, as the effect of case number may vary depending on whether augmentation is applied. The numbers used for further interpretation can be found in Table 7.4.

Batch	Case Group		IoU		Dice Coefficients		Accuracy		Ref.
	Aug	no Aug	Aug	no Aug	Aug	no Aug	Aug	no Aug	
Size 1	80	160	0.445	0.373	Better non-con.enh. & edema	Better con.enh.	0.959	0.940	Fig. 7.23
Size 1	160	314	0.317	0.505	Worse non-con.enh. & con.enh.	Better non-con.enh. & con.enh.	0.922	0.960	Fig. 7.24
Size 2	80	160	0.329	0.439	Worse non-con.enh.	Better non-con.enh.	0.923	0.952	Fig. 7.25
Size 2	160	314	0.347	0.484	Worse non-con.enh. & con.enh.	Better non-con.enh. & con.enh.	0.936	0.958	Fig. 7.26
Size 4	80	314	0.469	0.610	Worse non-con.enh., edema & con.enh.	Better non-con.enh., edema & con.enh.	0.953	0.974	Fig. 7.27

Table 7.3: This table compares the performance of models across Batch sizes 1, 2, and 4. For each Batch size, the Case Groups with augmentation contain half the number of original cases compared to the non-augmented models. Metrics include IoU Score, Dice coefficients of non-contrast-enhancing tumor, edema, and contrast-enhancing tumor, as well as accuracy, with columns for augmentation "Aug" and "no Aug" showing the respective performance.

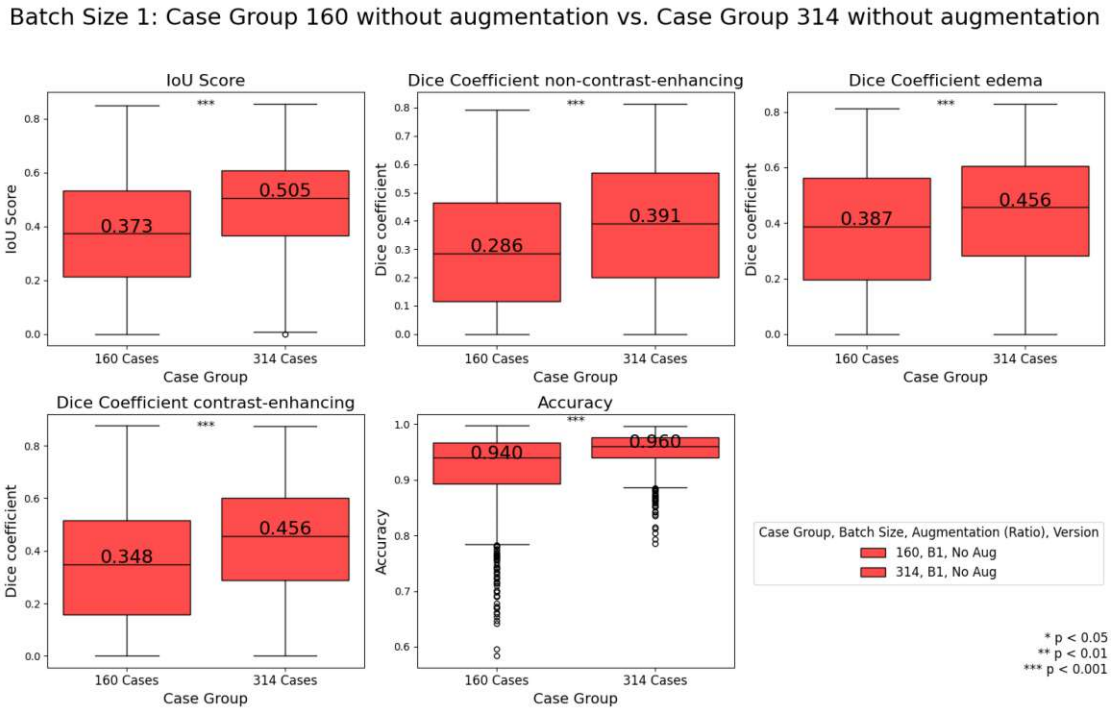


Figure 7.28: Comparison of Batch size 1 without augmentation for Case Group (160 vs. 314). The boxplots compare the performance metrics of models trained with Case Group 160 and 314 without augmentation. Adding more cases significantly improves all metrics, with the most notable increase in IoU Score, Dice coefficient for non-contrast-enhancing and contrast-enhancing tumor, as well as overall accuracy (***$p < 0.001$).

Batch size 1, without augmentation, Case Group 160 vs. 314 (Figure 7.28)

- The **IoU Score** increases from 0.373 for Case Group 160 to 0.505 for Case Group 314. This significant improvement ($p < 0.001$) suggests that adding more cases without augmentation contributes positively to the overall segmentation performance.
- **Dice coefficient – non-contrast-enhancing tumor** similarly shows a marked improvement, from 0.286 to 0.391 ($p < 0.001$). This is critical as the non-contrast-enhancing tumor is a challenging class to segment, and increasing the number of cases appears to improve the model's ability to learn its features.
- **Dice coefficient – edema** shows improvement from 0.387 to 0.456, and **Dice coefficient – contrast-enhancing tumor** improves from 0.348 to 0.456 (both with $p < 0.001$), further solidifying that higher case numbers provide better results across all tumor classes.
- **Accuracy** improves from 0.940 to 0.960 ($p < 0.001$), showing a trend of increased reliability with more data.

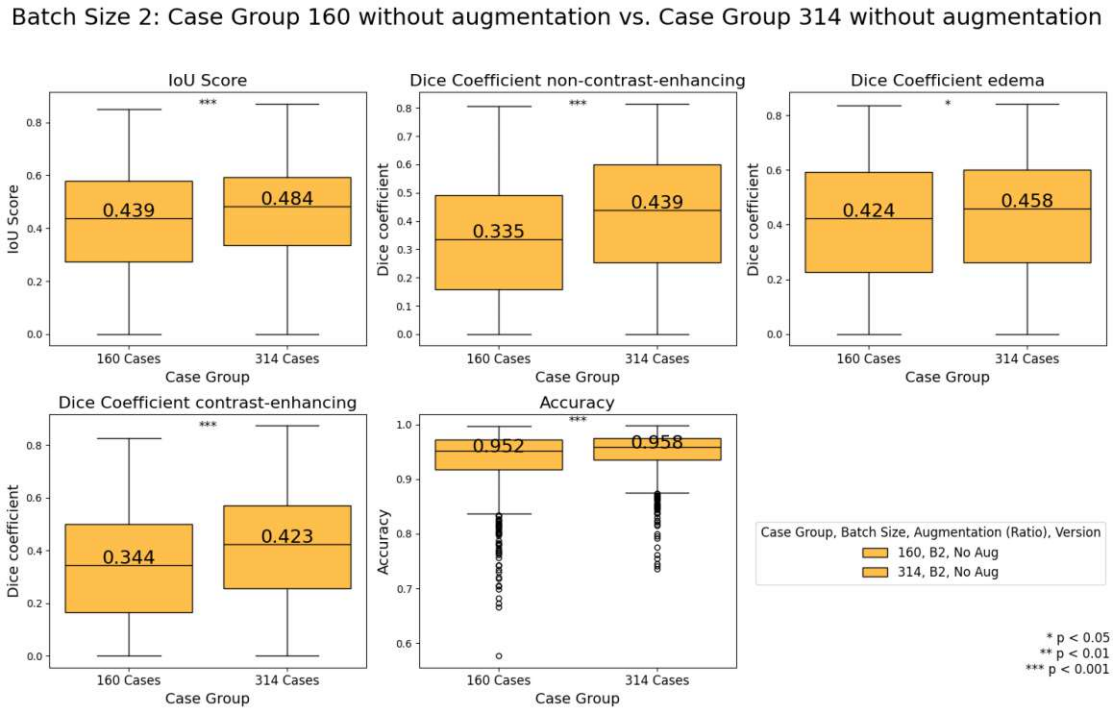


Figure 7.29: Comparison of Batch Size 2 without augmentation Case Group (160 vs. 314). This figure compares models trained with Case Group 160 and 314 without augmentation, holding Batch size 2 constant. All metrics show improvements with more cases, confirming the advantage of more training data in non-augmented datasets.

Batch size 2, without augmentation, Case Group 160 vs. 314 (Figure 7.29)

- The **IoU Score** shows a similar upward trend, from 0.439 to 0.484, reinforcing the positive impact of adding more cases.
- **Dice coefficients** also show improvements across the board. **non-contrast-enhancing tumor** improves from 0.335 to 0.439 ($p < 0.001$), **edema** from 0.424 to 0.458, and **contrast-enhancing tumor** from 0.344 to 0.423 ($p < 0.001$). These trends are consistent with Batch size 1 results.
- **Accuracy** increases from 0.952 to 0.958 ($p < 0.001$), again validating that higher case numbers lead to more accurate models, even without augmentation.

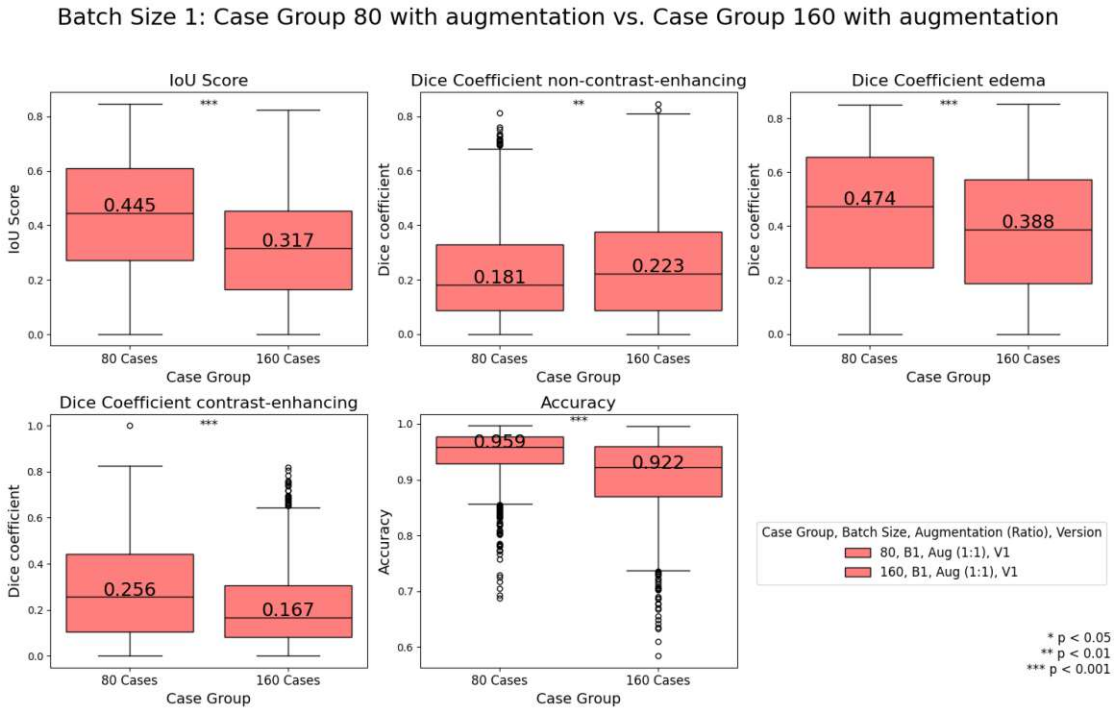


Figure 7.30: Comparison of Batch size 1 with augmentation Case Group (80 vs. 160). Augmented models trained on Case Group 80 outperformed those trained on Case Group 160, counterintuitively. The IoU Score and Dice coefficients for all tumor classes dropped with Case Group 160, with the exception of edema. This suggests that the applied augmentation strategy may not generalize well as the Case Group size increases.

Batch size 1, with augmentation, Case Group 80 vs. 160 (Figure 7.30)

- Surprisingly, the **IoU Score** drops from 0.445 for Case Group 80 to 0.317 for Case Group 160 ($p < 0.001$), which is counterintuitive. Typically, one would expect the IoU Score to improve with more cases, but this drop suggests that adding more augmented cases may introduce noise or redundant information.
- Similarly, **Dice coefficients** for **non-contrast-enhancing tumor** increase from 0.181 to 0.223, and **edema** from 0.474 to 0.388. These drops, despite increasing case numbers, could indicate that the augmentation strategy used here might not be as beneficial when applied to larger datasets.
- **Contrast-enhancing tumor** also drops from 0.256 to 0.167, reflecting poorer performance.
- **Accuracy** follows the same trend, dropping from 0.959 to 0.922 ($p < 0.001$), suggesting that this augmentation strategy may not generalize well if applied to more cases.

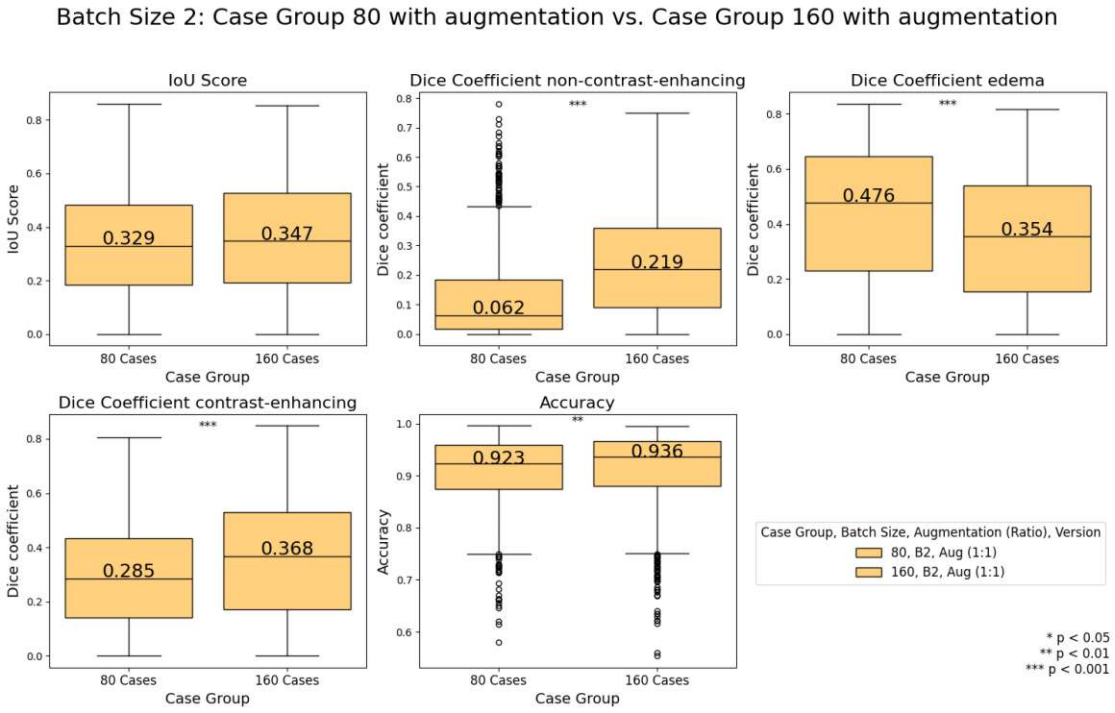


Figure 7.31: Comparison of Batch size 2 with augmentation Case Group (80 vs. 160). If a Batch size 2 is used, the IoU Score and Dice coefficients for most tumor classes show improvements with increasing Case Group size. Unlike Batch size 1, this augmentation strategy seems to perform better with a larger Case Group size.

Batch size 2, with augmentation, Case Group 80 vs. 160 (Figure 7.31)

- The **IoU Score** shows a similar behavior as the other comparisons, increasing slightly from 0.329 to 0.347.
- Dice coefficient for the **non-contrast-enhancing tumor** increases from 0.062 to 0.219, indicating a measurable but still insufficient improvement. Despite this increase, segmentation performance for this tumor class remains low and far from clinically useful. **edema** decreases from 0.476 to 0.354 ($p < 0.001$), highlighting a notable decline in segmentation accuracy.
- **Contrast-enhancing tumor** shows a better improvement from 0.285 to 0.368, giving a less volatile behavior compared to Batch size 1.
- **Accuracy** shows a slight increase from 0.923 to 0.936 ($p < 0.001$), again emphasizing that Batch size 2 shows better stability.

General Insights

Effect of augmentation: Interestingly, if augmentation is used, particularly with **Batch size 1**, models trained with **fewer cases (80 cases)** outperform models trained with **more cases (160 cases)**. This trend is most clearly observed in the **IoU Score** and **Dice coefficients for the non-contrast-enhancing and edema class**. The augmentation strategy appears to be more effective with fewer original cases, leading to better generalization. However, as the number of cases increases, the performance begins to degrade, likely due to the introduction of redundant or noisy information during the augmentation process. This suggests that the chosen augmentation strategy may not scale well if applied to larger datasets.

Batch size consideration: For **Batch size 1**, the effect of augmentation on fewer cases is unexpectedly positive, but performance diminishes if the case number increases. In contrast, **Batch size 2** demonstrates more stability, with the models benefiting from both augmentation and an increased number of cases. This implies that Batch size plays a crucial role in the interaction between augmentation and dataset size. Smaller Batch sizes seem more sensitive to augmentation, and careful tuning may be needed to avoid overfitting or poor generalization.

Non-contrast-enhancing and contrast-enhancing tumor: In augmented models with **Batch size 1**, the **non-contrast enhancing** and **contrast-enhancing tumor** classes are particularly sensitive to the number of cases. If fewer cases (80) are augmented, the model captures these challenging tumor classes more effectively compared to using 160 cases, where performance drops significantly. This finding highlights the potential of augmentation to artificially enhance limited datasets, but it also underscores the risk of diminishing returns as more data is introduced.

Batch	Aug	Case Group	IoU	Dice Coefficients			Accuracy	Ref.
				non-contrast enhancing	edema	contrast enhancing		
Size 1	no Aug	160	0.373	0.286	0.387	0.348	0.940	Fig. 7.28
		314	0.505	0.391	0.456	0.456	0.960	
Size 2	no Aug	160	0.439	0.335	0.424	0.344	0.952	Fig. 7.29
		314	0.484	0.439	0.458	0.423	0.958	
Size 1	Aug	80	0.445	0.181	0.474	0.256	0.959	Fig. 7.30
		160	0.317	0.223	0.388	0.167	0.922	
Size 2	Aug	80	0.329	0.062	0.476	0.285	0.923	Fig. 7.31
		160	0.347	0.219	0.354	0.368	0.936	

Table 7.4: This table compares model performance across Batch sizes 1 and 2, with and without data augmentation. For each combination of Batch size and augmentation, the larger Case Group contains about twice as many cases as the smaller group. Metrics include IoU Score, Dice coefficients of non-contrast-enhancing tumor, edema, contrast-enhancing tumor, as well as accuracy.

CHAPTER 8

Discussion, Outlook & Conclusion

8.1 Discussion

The results showed that models trained with a Batch size of 4 consistently ranked among the top-performing models, with 80% of these models ranking among the top 10 performers. This indicates that larger Batch sizes contribute to better generalization and stability, particularly as the number of training cases increases. Augmentations generally performed worse, except the best model, which used a 1:1 ratio of augmentations to originals and a Batch size of 1. This setup proved to be an outlier, performing exceptionally well compared to other models.

A deeper investigation into the augmentation strategies revealed that performance degradation occurred even at relatively low augmentation ratios, such as 1:3 (original: augmentation). This contrasts with findings in other studies, which suggest that excessive augmentation (typically with ratios higher than 1:3) confuses the model. In our case, the performance issues emerged earlier, indicating a sensitivity to augmentation in this specific setup. One notable problem was observed with the augmentation strategy that placed one original and three augmentations of that same original into a Batch of size 4. This configuration resulted in clear overfitting, likely due to the lack of variability between the augmentations and the original data. It became evident that batch diversity was critical, as a second strategy, which shuffled different augmentations within each batch, generalized better. Although both strategies used the same number of augmentations, the distribution within the batches had a significant impact on model performance.

Interestingly, the Case Group 80 with a 1:1 ratio of augmentations to originals, trained with Batch size 1, performed unexpectedly well in several metrics. Although Batch size 1 typically results in longer training times and poorer generalization, this particular setup produced results that deviated from the expected trends. This suggests that Batch size and augmentation strategies may interact in complex ways that warrant further investigation.

Despite considerable efforts to minimize the influence of randomness in the training process through seed management, it cannot be entirely eliminated. Notable fluctuations in model performance were observed across different Focal Weight Factors, with no clear trend towards a specific optimal value. This variability is likely attributable to the heterogeneity of the MRI sequences used in the different Case Groups. The recommendations derived from these findings are discussed in Section 8.4.

8.2 Limitations

Despite the success in achieving high-performing models with certain configurations, this study encountered several limitations that should be addressed in future work. One significant limitation was the long runtime of the training process, even on high-performance hardware such as the Nvidia A100. Although efforts were made to reduce data processing times—such as cropping, excluding one sequence (T1 native), shifting the normalization process outside of the data generator, and saving data as NumPy arrays to accelerate loading—the training times remained considerable. This issue was particularly pronounced when using Batch size 1, where the computational time grew drastically with an increasing number of cases, especially beyond 480 cases.

Another limitation relates to the performance of the augmentation strategies. While geometric transformations were chosen for their simplicity and safety (to avoid altering anatomical context), the overall performance of augmented models was suboptimal compared to models trained without augmentations. Interestingly, the degradation in performance occurred even at relatively low original-to-augmentation ratios, such as 1:3, which contrasts with previous findings that suggest performance degradation typically occurs at higher ratios. This highlights the sensitivity of the chosen model to even moderate levels of augmentation.

A technical limitation was encountered with memory overflow on smaller GPUs, such as the Nvidia 3070 Mobile, which has 8 GB of VRAM. This limitation made it impossible to run Batch sizes higher than 1 on this hardware. Consequently, the project had to be run on cloud-based services like Colab, which significantly increased costs, exceeding €1500. The cost factor is a critical consideration for future projects and should be taken into account during planning.

8.3 Lessons Learned

Several key lessons emerged throughout this project, offering valuable insights for future work in glioblastoma segmentation using deep learning techniques. These insights span three main areas: the impact of augmentation strategies, the role of data preprocessing, and the limitations imposed by hardware constraints.

First, the sensitivity of augmentation strategies was more pronounced than anticipated. Although the use of augmentations is a common technique to improve model generalization,

this study found that even relatively modest augmentation ratios (e.g., 1:3) led to performance degradation. This emphasizes the importance of carefully balancing the ratio of augmentations to originals and selecting augmentation strategies that maintain sufficient variability within batches. The finding that overfitting occurred if a batch contained multiple augmentations of the same original case highlights the need for greater diversity in the augmentation process.

Second, the handling of data preprocessing had a significant impact on both model performance and computational efficiency. Early issues, such as the presence of NaN values in the data, underscored the importance of thorough data cleaning before model training. Converting NaN values to zeros, along with other preprocessing steps like skull removal and normalization, were crucial in avoiding complications during training. Additionally, tight cropping of the images, while necessary to manage large data volumes, led to a limitation of the rotation applied during augmentations to +/- 10 degrees. Although other studies have used rotations of up to 20 degrees, larger rotations were avoided to prevent parts of the brain surface from being rotated out of the field of view, which could result in data loss. It is also assumed that smaller rotations may have been too similar to the original images, potentially negatively affecting the training process.

Third, another significant challenge encountered was related to hardware limitations, particularly with the NVIDIA RTX 4070 Laptop GPU, which, despite its computational power, could not be fully utilized for deep learning tasks. The laptop featured an AMD Ryzen 9 7940HS CPU with 64GB of DDR5 RAM, providing ample resources for preprocessing tasks. However, due to the inability to disable the onboard graphics card under Linux, direct access to the NVIDIA GPU via CUDA was not possible. Attempts to utilize the GPU through WSL2 (Windows Subsystem for Linux) on Windows 11 also proved unstable, leading to frequent issues during model training. While the CPU was sufficient for preprocessing tasks using tools like the BrainLes Preprocessing Package [KBW+20], it was inadequate for the computational demands of model training.

These limitations became a major bottleneck, forcing the project to rely on cloud-based solutions. This experience highlighted the importance of factoring in both hardware constraints and potential costs when planning large-scale deep-learning projects. Efficient memory management became particularly important when using Batch size 1, which led to a steep increase in runtime as the number of cases grew. The inability to fully utilize the high-performance GPU, combined with the memory limitations of the laptop, demonstrated that gaming laptops running Windows 11 might not be well suited for deep learning projects that require stable and intensive GPU usage.

Finally, the overall structure and execution of the training pipeline highlighted the importance of ensuring a reproducible order in which cases are loaded during training. This was achieved by using a seed list with 100 predefined entries, ensuring a fixed loading order for up to 100 training epochs. This control mechanism guaranteed that if a new training run was restarted from Epoch 0—after, for instance, an initial run had been interrupted at Epoch 30—the case-loading sequence for the first 30 epochs would remain identical to the previous run. More specifically, if training was resumed from

Epoch 15, the order of cases for these first 15 epochs would be exactly the same as in the prior execution. Maintaining this level of consistency was crucial for reproducibility, particularly in deep learning experiments where even small variations in data loading could lead to significant differences in model performance. Ensuring such reproducibility is especially critical in medical applications, where reliable and consistent results are essential.

8.4 Recommendations

Based on the findings of this study, several key recommendations can be made for future work in glioblastoma segmentation using deep learning. To ensure clarity and structure, these recommendations follow the chronological workflow used in the project. Each recommendation consists of two elements: first, a brief rationale explaining its necessity based on the challenges and findings of this study, and second, an actionable guideline derived from this rationale to improve future implementations. This structured approach ensures that each recommendation is well-founded while providing clear, practical guidance for future applications.

1. **Dataset Similarity:**
    - Differences in population demographics, such as age and gender distribution, can introduce biases in model performance. To minimize such effects, the training and evaluation datasets should be as demographically similar as possible.
    - Heterogeneous pathologies within the dataset can introduce noise and mislead the segmentation model. To minimize misclassification, patients with additional tumor entities, such as meningiomas, should be excluded, as these could be mistaken for glioblastomas. Likewise, patients with severe vascular leukoencephalopathy should be removed, as the associated white matter changes may be erroneously interpreted as tumor edema.
2. **Ground Truth Segmentation:**
    - Interrater variability can introduce inconsistencies that negatively impact model training. To ensure consistency, a single experienced neuroradiologist should perform the annotations rather than multiple raters.
    - The segmentation process is subject to a learning curve, meaning that initial segmentations may be less refined. To address this, an iterative approach should be used, allowing early segmentations to be reviewed and refined as expertise increases.
    - Standardized annotation protocols have been shown to reduce variability in segmentation quality, as demonstrated in later iterations of the BraTS Challenge [CdVSG+24]. To maintain consistency, adherence to a clinically approved annotation protocol is essential.

3. **Data Acquisition:**

   - Variations in slice thickness and orientation can lead to inconsistencies in model training. To prevent this, uniform slice thickness and orientation should be ensured in sequences like T2 and FLAIR (e.g., 4 mm axial).
   - T1 contrast-enhanced imaging using Fast Field Echo (FFE) provides more consistent contrast than Turbo Spin Echo (TSE). For this reason, FFE should be used instead of TSE, and isotropic voxels with a 1 mm edge length should be preferred.
   - Thick-slice T2/FLAIR sequences are often used in clinical practice to reduce scan time, despite the advantages of isotropic voxels. If isotropic voxels cannot be used, dataset consistency should be ensured through standardized acquisition settings.
   - Variability in T1 imaging protocols can introduce inconsistencies in contrast-enhanced versus native sequences. If native T1 is used, it should match the parameters of T1ce.
   - Heterogeneous datasets with varying acquisition parameters may negatively impact model convergence. To ensure optimal performance, datasets that deviate from standardized acquisition parameters should be excluded, even if this reduces the number of cases.

4. **Data Preprocessing:**

   - Inconsistent preprocessing can introduce systematic errors in model training. To avoid this, a standardized framework like the BrainLes Preprocessing Package [KBW+20] should be used, including steps like skull removal, atlas mapping (e.g., SRI-24 space [RZSP10]), and normalization.
   - The presence of NaN values in the dataset can cause issues in later processing steps. To prevent this, all NaN values should be converted to zeroes.
   - Unnecessary regions in MRI scans, such as the skull and soft tissues (e.g., nasal structures), do not contribute to tumor segmentation and may introduce noise; their removal creates empty spaces in the scan, which should then be eliminated through cropping to ensure that the model focuses only on the regions of interest. This also reduces the data size by removing areas without relevant information, leading to more efficient processing.
   - Cropping can limit augmentation flexibility, as tight cropping reduces the available rotation range. Due to this limitation, augmentation rotations were restricted to ±10 degrees, whereas literature suggests that 20 degrees may be optimal.
   - Contrast enhancement techniques, such as Contrast Limited Adaptive Histogram Equalization (CLAHE), can improve image quality and segmentation performance. To enhance contrast, CLAHE should be applied, followed by

renormalization to the range of 0 to 1. While this study used a 2D implementation of CLAHE, an extension to the third dimension, as in True 3D CLAHE [AFdMSP18], could further improve contrast uniformity across slices and enhance segmentation consistency.

- Advanced denoising techniques, such as the Boosted Anisotropic Diffusion Filter (BADF), have been shown to improve image quality in other medical imaging applications [NMS23]. Future work should explore the combination of CLAHE and BADF to further enhance segmentation accuracy.
- Repeated normalization steps during training can increase computational overhead. To optimize efficiency, 3D image data should be stored as pre-normalized NumPy arrays for faster loading in Python [VR09].

5. **Data Augmentation:**

   - Geometric transformations such as rotation and flipping are widely used for data augmentation, as they preserve anatomical context. To maintain structural integrity, these transformations should be prioritized, while the effects of different augmentation-to-original ratios should be further investigated.
   - Generative adversarial networks (GANs) have been proposed as a method to create more complex and realistic augmentations [FSL+24]. Unlike traditional augmentations, GAN-generated samples could introduce synthetic but anatomically plausible variations, potentially improving model robustness. Future research should consider GAN-based augmentations to enhance data variability.

6. **3D U-Net Parameters:**

   - The risk of overfitting increases when training progresses beyond the optimal number of epochs. To prevent overfitting, callbacks such as Early Stopping and Reduce Learning Rate on Plateau should be used.
   - Batch size influences both training stability and memory requirements. Batch sizes should be adjusted based on the available hardware, with larger Batch sizes generally providing better performance.
   - Variability in case loading order can affect model training. To ensure reproducibility, the data generator should be modified to maintain a fixed random order for case loading throughout the training process.

7. **Loss Function:**

   - Medical tumor segmentation tasks are often affected by class imbalance, requiring specialized loss functions. A Combined Loss Function, such as the one used in this project (Dice Loss and Focal Loss), should be applied, with class weights adjusted accordingly.

- Class imbalance can disproportionately affect segmentation performance across different tumor classes. To address this, class weights should be systematically optimized, ensuring that non-contrast-enhancing tumor, contrast-enhancing tumor, and edema are appropriately balanced in the loss function to prevent under-segmentation of less prevalent regions.

8. **Computational Considerations:**

   - Limited memory capacity can lead to training instability or necessitate the use of cloud-based services, increasing project costs. To avoid unexpected expenses, memory capacity should be considered when selecting hardware, and cloud service costs should be factored into the budget.
   - Mixed precision training can reduce memory usage and speed up computations without significant loss of model accuracy. To improve efficiency, mixed precision training should be implemented, particularly for large-scale models, as it allows for reduced memory consumption while maintaining numerical stability.

8.5 Future Work

This study has highlighted several areas where future research could build upon the findings and address the limitations encountered during the project. The following suggestions outline potential directions for further exploration:

1. **Exploration of Advanced Augmentation Techniques:**

   Future research could explore the use of more sophisticated augmentation techniques, such as GAN-based augmentations, which have the potential to generate more realistic and diverse synthetic data. This could help overcome the limitations observed with geometric transformations, which performed poorly in this study.

2. **Further Investigation of Augmentation Ratios**:

   Given the unexpected performance degradation at relatively low augmentation-to-original ratios (e.g., 1:3), future work should focus on systematically evaluating the effects of different augmentation ratios. This would help clarify whether the results observed in this project are specific to the dataset or if they represent a broader issue in medical image segmentation.

3. **Optimization of CLAHE for 3D Images:**

   Although 2D CLAHE improved image quality in this project, the use of True 3D CLAHE could be explored in future work [AFdMSP18]. This would likely provide better contrast enhancement for volumetric data, potentially improving segmentation accuracy in 3D U-Net models.

4. **Evaluation of larger Batch sizes:**

   While Batch sizes of 4 performed well in this study, future research could investigate the impact of even larger Batch sizes, provided that the hardware allows for it. This could lead to further improvements in model stability and generalization, especially as datasets grow in size.

5. **Investigation of More Complex Model Architectures:**

   This project focused on a standard 3D U-Net architecture. Future studies could explore more complex architectures, such as attention-based models or hybrid models combining 3D U-Net with other approaches like transformers. These advanced architectures could further improve segmentation accuracy, particularly in more challenging tumor classes.

6. **Refinement of Preprocessing Techniques:**

   Additional preprocessing techniques, such as more advanced denoising methods or adaptive cropping strategies, could be explored. The goal would be to further reduce noise in the data without sacrificing important anatomical information, particularly in patients with unusual anatomical features.

7. **Handling of Diverse Pathologies in Datasets:**

   Future work should also focus on refining the inclusion and exclusion criteria for datasets. This project identified the need to exclude patients with pronounced neurodegenerative changes (e.g., Fazekas scale grade III [FCA+87] vascular leukoencephalopathy) or other tumor types (e.g., meningiomas) to avoid confusion in the segmentation process. A more rigorous approach to dataset curation could lead to more homogeneous data and better model performance.

8. **Cost-Effective Computational Solutions:**

   The high costs associated with cloud-based training environments in this project highlight the need for more cost-effective computational solutions. Future work could explore alternative hardware configurations or optimized cloud usage strategies to reduce costs without compromising performance.

8.6 Conclusion

This thesis focused on the optimization of parameter tuning for the deep learning-based segmentation of glioblastoma multiforme (GBM). By analyzing key parameters such as Batch size, data augmentation strategies, the effect of Case Group size, and the Focal

Weight Factor in the Combined Loss Function, several important findings emerged from the conducted experiments.

First, models trained with a Batch size of 4, particularly those without augmentations, generally performed better. This trend highlights the stability and generalization benefits of larger Batch sizes. The exception to this trend was the best-performing model, which utilized a Batch size of 1 and a 1:1 ratio of original to augmented images, showing that there are circumstances where smaller Batch sizes combined with specific augmentation strategies can outperform other models.

Second, the augmentation strategy played a crucial role in model performance, particularly with Batch size 4, where an augmentation ratio of 1:3 (original to augmented images) led to overfitting, especially when three variants of one original were included in a batch. This emphasizes the importance of carefully balanced augmentation strategies to prevent overfitting, particularly with larger Batch sizes.

Third, the number of training cases had a clear impact on segmentation accuracy. Models trained with the full set of cases (Case Group 314) outperformed those with fewer cases, highlighting the importance of larger datasets in improving model generalization and accuracy.

While the tuning of the Focal Weight Factor was explored, its influence on segmentation performance was relatively minor compared to other parameters. Although it contributed to managing class imbalance, it was less impactful than Batch size, Case Group size, or the use of augmentations.

Additionally, the computation time for models trained with a Batch size of 1 displayed a linear growth up to a certain number of training cases. However, as the number of cases increased, there was a clear transition to drastically growth in training time. This elbow point, marked by a sharp increase in the slope of the computation time curve, became evident when augmentations were used, with up to 628 cases included for Batch size 1. The transition was observed at approximately 480 cases. This highlights the need for optimized resource management when working with small Batch sizes on hardware with limited memory. Less powerful GPUs can lead to memory overflow, making costly cloud-based solutions necessary.

The best-performing model in this study (Case Group 314, Batch Size 1, augmentation ratio 1:1) was compared to state-of-the-art results from segmentation challenges such as BraTS 2018. The achieved Dice coefficient for edema segmentation (0.825) was within the upper range of reported values, whereas the performance for the non-contrast-enhancing tumor (0.748) and the contrast-enhancing tumor (0.665) approached the lower bound of state-of-the-art results. These deviations can likely be attributed to the use of a simplified 3D U-Net architecture, which was selected to balance computational efficiency and interpretability. Additionally, differences in preprocessing techniques, augmentation strategies, and dataset composition may have contributed to the observed variations. Nonetheless, these findings demonstrate that with careful parameter tuning, even relatively simple model architectures can achieve competitive results.

In conclusion, this work demonstrated that optimizing modifiable parameters significantly improves the accuracy and robustness of deep learning models for GBM segmentation. These results provide useful guidelines for developing models that can be applied more effectively in clinical settings. Future work could focus on refining augmentation strategies and exploring the broader applicability of these findings to other medical image segmentation tasks.

Overview of Generative AI Tools Used

In this thesis, I have utilized generative AI tools as supplementary aids during the writing, revision, and structuring process. The overall intellectual and creative contribution remains predominantly my own, with all AI-suggested outputs critically reviewed and revised by myself to ensure accuracy, coherence, and alignment with the research objectives.

AI assistance was specifically employed in areas such as refining text, improving clarity, restructuring content, and enhancing the scientific writing style. The generative AI tools used are as follows:

- **Tool:** ChatGPT (OpenAI)
  - **Version:** GPT-4 (September 2024)
  - **Usage:** ChatGPT served as a dialog-based assistant to iteratively refine and revise various sections of the thesis. Through continuous exchanges, I clarified ideas, reorganized content, and ensured consistency in tone and style. Specific contributions include:
    - Restructuring complex sections for improved logical flow.
    - Simplifying overly complex sentences while maintaining scientific precision.
    - Suggesting concise and descriptive chapter headings.
    - Refining explanations of methods, results, and visualizations.
    - Providing alternative phrasing for repetitive or unclear passages.
  - **Critical Review:** All AI-generated suggestions were carefully reviewed and edited to preserve the scientific accuracy, integrity, and originality of the thesis.
- **Tool:** Grammarly (Grammarly, Inc.)
  - **Usage:** Grammarly was used for grammar and style checks to ensure clarity, coherence, and consistency. It helped identify minor errors in spelling, punctuation, and word choice, enabling adherence to academic writing standards.

The collaboration with AI tools, particularly ChatGPT, was dialog-based and iterative, enabling me to improve the overall quality of my thesis while ensuring that the final work reflects my own critical thinking and research efforts.

Übersicht verwendeter Hilfsmittel

In dieser Diplomarbeit kamen generative KI-Tools als ergänzende Hilfsmittel im Schreib-, Überarbeitungs- und Strukturierungsprozess zum Einsatz. Der intellektuelle und kreative Beitrag bleibt dabei überwiegend mein eigener. Alle von der KI vorgeschlagenen Inhalte wurden kritisch geprüft und angepasst, um wissenschaftliche Genauigkeit, Kohärenz und die Übereinstimmung mit den Forschungszielen sicherzustellen.

Die KI-Unterstützung wurde gezielt in Bereichen wie Textoptimierung, der Verbesserung der Verständlichkeit, inhaltlicher Umstrukturierung und Verbesserung des wissenschaftlichen Schreibstils eingesetzt. Die verwendeten generativen KI-Tools sind wie folgt:

- **Tool:** ChatGPT (OpenAI)
  - **Version:** GPT-4 (Stand: September 2024)
  - **Verwendung:** ChatGPT diente als dialogbasierter Assistent, um verschiedene Abschnitte der Arbeit schrittweise zu überarbeiten und zu verfeinern. Im kontinuierlichen Austausch konnte ich Ideen klären, Inhalte neu strukturieren und eine konsistente Ausdrucksweise sicherstellen. Konkrete Beiträge beinhalten:
    * Umstrukturierung komplexer Abschnitte zur Verbesserung des logischen Flusses.
    * Vereinfachung komplizierter Sätze unter Beibehaltung wissenschaftlicher Präzision.
    * Vorschläge für prägnante und beschreibende Kapitelüberschriften.
    * Verfeinerung von Erklärungen zu Methoden, Ergebnissen und Visualisierungen.
    * Alternativformulierungen für redundante oder unklare Textpassagen.
  - **Kritische Prüfung:** Alle KI-generierten Vorschläge wurden sorgfältig überprüft und angepasst, um die wissenschaftliche Integrität und Originalität der Arbeit zu gewährleisten.
- **Tool:** Grammarly (Grammarly, Inc.)
  - **Verwendung:** Grammarly wurde für die Prüfung von Grammatik, Rechtschreibung und Stil eingesetzt. Es half dabei, kleinere Fehler zu identifizieren

und die Klarheit sowie Kohärenz des Textes zu verbessern, um den Anforderungen des akademischen Schreibens zu entsprechen.

Die Zusammenarbeit mit KI-Tools, insbesondere ChatGPT, erfolgte dialogbasiert und iterativ. Dieser Ansatz ermöglichte es mir, die Qualität der Diplomarbeit zu steigern, während die endgültige Arbeit meinen eigenen kritischen Denkprozess und Forschungsbeitrag widerspiegelt.

List of Figures

List of Tables

Acronyms

2-HG 2-Hydroxyglutarate. 14

3T 3 Tesla. 48

AM Additive manufacturing. 42

AMD Advanced Micro Devices, Inc. 88, 143

ANOVA Analysis of Variance. 98

ASNR American Society of Neuroradiology. 4, 45, 90

BADF Boosted Anisotropic Diffusion Filter. 146

BraTS Brain Tumor Segmentation. xi, xiii, 4, 25, 27, 40, 45–47, 50, 54, 57, 67, 69, 70, 90, 91, 125, 144, 149, 157

BraTumIA Brain Tumor Image Analysis. 27, 28

BS Block Size. 72

CBF Cerebral blood flow. 12

CBV Cerebral blood volume. 12

cCT Cranial Computed Tomography. 10, 155

Cho Cholin. 15, 155

CL Clip Limit. 72

CLAHE Contrast-Limited Adaptive Histogram Equalization. 6, 8, 57, 67, 68, 70–76, 145–147, 155–157

CNN Convolutional Neural Network. 2, 8, 17, 20, 25

CNS Central Nervous System. 45

CPU Central Processing Unit. 88, 143

Cr Creatin. 15, 155

CSV Comma Separated Values. 86

CT Computed Tomography. 10, 13, 17, 22, 72

CUDA Compute Unified Device Architecture. 87, 143

DCE Dynamic Contrast Enhanced. 12

DDR Double Data Rate. 88, 143

DICOM Digital Imaging and Communications in Medicine. 54, 69

DRL Deep reinforcement learning. 83

DSC Dynamic Susceptibility Contrast. 12

DSC Dice Similarity Coefficient. 29–32, 35–38, 50, 61

DTI Diffusion tensor imaging. 10, 11

EES Extravascular Extracellular Space. 13

FA Fractional Anisotropy. 11

FDG 18F-Fluorodeoxyglucose. 13

FET 18F-Fluoroethyltyrosine. 10, 13, 155

FFE Fast Field Echo. 48, 49, 52, 55, 80, 145, 156

FLAIR Fluid Attenuated Inversion Recovery. xi, xiii, 4–6, 10, 18, 27, 47–55, 59, 60, 67, 69, 74, 75, 77–80, 82, 85, 86, 122, 124, 145, 155–157, 160

FN False Negative. 31, 35

FOV Field of View. 48

FP False Positive. 31, 35

GAN Generative Adversarial Networks. 24–26, 146, 147

GB Gigabyte. 87, 88, 142, 143

GBM Glioblastoma multiforme. xi, xiii, 1, 2, 9, 10, 27, 45, 47, 50, 92, 148, 150

GDDR Graphics Double Data Rate. 88

GDS Generalized Dice Score. 35, 36

GPU Graphics Processing Unit. 5, 62, 69, 82, 86–88, 142, 143, 149

HBM2 High Bandwidth Memory Type 2. 87

HD-BET High-definition Brain Extraction Tool. 57, 70

HD95 95% Hausdorff Distance. 29, 33, 37

HGG High-Grade Glioma. 45

HSD Honestly Significant Difference. 98

IDH Isocitrate Dehydrogenase. 10, 14, 92, 155

IoU Intersection over Union. xi, xiii, 3–5, 8, 29–32, 37, 62–65, 69, 86, 97, 98, 100–104, 108–111, 113, 117–123, 125–131, 133–139, 158–161, 163

IQR Interquartile Range. 54, 64, 90, 92, 117–119, 159

K^{trans} volume transfer constant. 13

LGG Low-Grade Glioma. 45, 46

MAE Mean Absolute Error. 25

MGMT O(6)-methylguanine-DNA methyltransferase. 14

MICCAI Medical Image Computing and Computer-Assisted Interventions. 4, 45, 90

MRI Magnetic Resonance Imaging. xiii, 1, 2, 4–6, 10–14, 17, 18, 22, 26–28, 47, 48, 50, 54, 57, 60, 67–72, 74, 75, 77, 79, 80, 82, 85, 92, 94, 122, 124, 142, 145, 155–157, 160

MRS Magnetic Resonance Spectroscopy. 14

MTT Mean Transit Time. 12

NAA N-Acetyl Aspartate. 14, 15, 155

NaN Not a Number. 6, 68, 70, 71, 143, 145, 155

NAS Neural Architecture Search. 7

NIfTI Neuroimaging Informatics Technology Initiative. 47, 54, 69, 88

OOD Out-of-Distribution. 60

OS Overall survival. 27

PET Positron Emission Tomography. 13, 17

PET-CT Positron Emission Tomography-Computed Tomography. 10, 13, 155

ppm parts per million. 14, 15, 155

RAM Random Access Memory. 88, 143

rCBV Relative Cerebral Blood Volume. 12

ReLU Rectified Linear Unit. 18, 20, 81

ROI Region of Interest. 27

RSNA Radiological Society of North America. 4, 45, 90

SPSS Statistical Package for Social Sciences. 54

SRI Stanford Research Institute. 6, 57, 61, 67, 69, 70, 145, 155

SUV Standardized uptake value. 10, 155

T1ce T1 post-contrast. 6, 27, 51, 52, 55, 56, 59, 60, 78–80, 82, 85, 86, 122, 124, 145, 155–157, 160

TE Echo Time. 48

TI Inversion Time. 48

TN True Negative. 35

TNR True Negative Rate. 35

TP True Positive. 35

TPR True Positive Rate. 35

TR Repetition Time. 48

TSE Turbo Spin Echo. 48, 49, 80, 145

VRAM Video Random Access Memory. 142

WHO World Health Organization. 9, 10, 14, 45–47, 92

WSL2 Windows Subsystem for Linux 2. 143

Bibliography

[AAHK24] Mohd Ali, Mehboob Ali, Mubashir Hussain, and Deepika Koundal. Generative Adversarial Networks (GANs) for Medical Image Processing: Recent Advancements. *Archives of Computational Methods in Engineering*, 2024. doi: https://doi.org/10.1007/s11831-024-10174-8.

[AFdMSP18] Paulo Amorim, Thiago Franco de Moraes, Jorge Silva, and Helio Pedrini. *3D Adaptive Histogram Equalization Method for Medical Volumes*. SciTePress, Jan 2018. doi: https://doi.org/10.5220/0006615303630370.

[AKBD22] Ayman Al-Kababji, Faycal Bensaali, and Sarada Prasad Dakua. Scheduling Techniques for Liver Segmentation: ReduceLROnPlateau vs OneCycleLR. In Akram Bennour, Tolga Ensari, Yousri Kessentini, and Sean Eom, editors, *Intelligent Systems and Pattern Recognition*, pages 204–212. Springer International Publishing, Jun 2022. doi: https://doi.org/10.1007/978-3-031-08277-1_17.

[AMS23] Peshraw Abdalla, Bashdar Mohammed, and Ari Saeed. The Impact of Image Augmentation Techniques of MRI Patients in Deep Transfer Learning Networks for Brain Tumor Detection. *Journal of Electrical Systems and Information Technology*, 10, Nov 2023. doi: https://doi.org/10.1186/s43067-023-00119-9.

[ASM+21] Laith Alzubaidi, J. Santamaria, Mohamed Manoufali, Beadaa J. Mohammed, Mohammed Abdulraheem Fadhel, Jinglan Zhang, Ali H. Altimemy, Omran Al-Shamma, and Ye Duan. MedNet: Pre-trained Convolutional Neural Network Model for the Medical Imaging Tasks. *ArXiv*, abs/2110.06512, 2021. doi: http://doi.org/10.48550/arXiv.2110.06512.

[ATH+21] Orhun Utku Aydin, Abdel Aziz Taha, Adam Hilbert, Ahmed A. Khalil, Ivana Galinovic, Jochen B. Fiebach, Dietmar Frey, and Vince Istvan Madai. An Evaluation of Performance Measures for Arterial Brain Vessel Segmentation. *BMC Medical Imaging*, 21(1):113, Jul 2021. doi: https://doi.org/10.1186/s12880-021-00644-x.

[AWS+16] Nathalie L. Albert, Michael Weller, Bogdana Suchorska, Norbert Galldiks, Riccardo Soffietti, Michelle M. Kim, Christian la Fougere, Whitney Pope, Ian Law, Javier Arbizu, Marc C. Chamberlain, Michael Vogelbaum, Ben M. Ellingson, and Joerg C. Tonn. Response Assessment in Neuro-Oncology working group and European Association for Neuro-Oncology recommendations for the clinical use of PET imaging in gliomas. *Neuro Oncol*, 18(9):1199–1208, 2016. doi: https://doi.org/10.1093/neuonc/now058.

[BBF+23] Spyridon Bakas, Ujjwal Baid, Keyvan Farahani, Jake Albrecht, James Eddy, Timothy Bergquist, Thomas Yu, Verena Chung, Russell (Taki) Shinohara, Michel Bilello, Suyash Mohan, Satyam Ghodasara, Ahmed W. Moawad, Jeffrey Rudie, Luiz Otavio Coelho, Elka Miller, Fanny E. Morón, Mark C. Oswood, Robert Y. Shih, . . . , and Zeke Meier. The International Brain Tumor Segmentation (BraTS) Cluster of Challenges. International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2023 (MICCAI 2023). 2023. doi: https://doi.org/10.5281/zenodo.7837974.

[BDEY01] Jonathan H. Burdette, David D. Durden, Allen D. Elster, and Yi-Fen Yen. High B-Value Diffusion-Weighted MRI of Normal Brain. *J Comput Assist Tomogr*, 25(4):515–9, Jul-Aug 2001. doi: https://doi.org/10.1097/00004728-200107000-00002.

[BEB+19] Jeroen Bertels, Tom Eelbode, Maxim Berman, Dirk Vandermeulen, Frederik Maes, Raf Bisschops, and Matthew B. Blaschko. Optimizing the Dice Score and Jaccard Index for Medical Image Segmentation: Theory and Practice. 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 92–100. Springer International Publishing, 2019. doi: https://doi.org/10.1007/978-3-030-32245-8_11.

[BGM+23] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe Campos Kitamura, Sarthak Pati, Luciano Prevedello, Jeffrey Rudie, Chiharu Sako, Russell Shinohara, Timothy Bergquist, Rong Chai, James Eddy, Julia Elliott, Walter Reade, Thomas Schaffter, Thomas Yu, Jiaxin Zheng, 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-André Weber, Abhishek Mahajan, Bjoern Menze, Adam E. Flanders, and

Spyridon Bakas. RSNA-ASNR-MICCAI-BRATS-2021, 2023. doi: https://doi.org/10.7937/JC8X-9874.

[BGPB14] Harun Badakhshi, Reinhold Graf, Vikas Prasad, and Volker Budach. The impact of 18 F-FET PET-CT on target definition in image-guided stereotactic radiotherapy in patients with skull base lesions. *Cancer Imaging*, 14(1):25, 2014. doi: https://doi.org/10.1186/1470-7330-14-25.

[BML94] Peter J. Basser, James Mattiello, and Denis J. LeBihan. Estimation of the Effective Self-Diffusion Tensor from the NMR Spin Echo. *J Magn Reson B*, 103(3):247–54, 1994. doi: https://doi.org/10.1006/jmrb.1994.1037.

[BP96] Peter J. Basser and Carlo M. Pierpaoli. Microstructural and Physiological Features of Tissues Elucidated by Quantitative-Diffusion-Tensor MRI. *J Magn Reson B*, 111(3):209–19, Jun 1996. doi: https://doi.org/10.1006/jmrb.1996.0086.

[BPE+23] Josef A. Buchner, Jan C. Peeken, Lucas Etzel, Ivan Ezhov, Michael Mayinger, Sebastian M. Christ, Thomas B. Brunner, Andrea Wittig, Björn Menze, Claus Zimmer, Bernhard Meyer, Matthias Guckenberger, Nicolaus Andratschke, Rami A El Shafie, Jürgen Debus, Susanne Rogers, Oliver Riesterer, Katrin Schulze, Horst J. Feldmann, Oliver Blanck, Constantinos Zamboglou, Konstantinos Ferentinos, Angelika Bilger, Anca L. Grosu, Robert Wolff, Jan S. Kirschke, Kerstin A. Eitz, Stephanie E. Combs, Denise Bernhardt, Daniel Rückert, Marie Piraud, Benedikt Wiestler, and Florian Kofler. Identifying core MRI sequences for reliable automatic brain metastasis segmentation. *Radiotherapy and Oncology*, 188:109901, 2023. doi: https://doi.org/10.1016/j.radonc.2023.109901.

[BRJ+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 Zhaolin Chen. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. *ArXiv*, 2019. doi: https://doi.org/10.48550/arXiv.1811.02629.

[BRK+90] John W. Belliveau, Bruce R. Rosen, Howard L. Kantor, Richard R. Rzedzian, David N. Kennedy, Robert C. McKinstry, James M. Vevea, Mark S. Cohen, Ian L. Pykett, and Thomas J. Brady. Functional Cerebral Imaging by Susceptibility-Contrast NMR. *Magn Reson Med*, 14(3):538–46, 1990. doi: https://doi.org/10.1002/mrm.1910140311.

[BRT+20] Ujjwal Baid, Swapnil U. Rane, Sanjay Talbar, Sudeep Gupta, Meenakshi H. Thakur, Aliasgar Moiyadi, and Abhishek Mahajan. Overall survival prediction in glioblastoma with radiomic features using machine learning. *Frontiers in Computational Neuroscience*, 14, 2020. doi: https://doi.org/10.3389/fncom.2020.00061.

[BTR+20] Ujjwal Baid, Sanjay Talbar, Swapnil Rane, Sudeep Gupta, Meenakshi H. Thakur, Aliasgar Moiyadi, Nilesh Sable, Mayuresh Akolkar, and Abhishek Mahajan. A Novel Approach for Fully Automatic Intra-Tumor Segmentation With 3D U-Net Architecture for Gliomas. *Frontiers in Computational Neuroscience*, 14, 2020. doi: https://doi.org/10.3389/fncom.2020.00010.

[BV09] Dani S. Bidros and Michael A. Vogelbaum. Novel drug delivery strategies in neuro-oncology. *Neurotherapeutics*, 6(3):539–46, Jul 2009. doi: https://doi.org/10.1016/j.nurt.2009.04.004.

[CAB+04] Robert W. Cox, John Ashburner, Hester Breman, Kate Fissell, Christian Haselgrove, Colin J. Holmes, Jack L. Lancaster, David E. Rex, Stephen M. Smith, and Jeffrey B. Woodward. A (Sort of) New Image Data Format Standard: NiFTI-1. In *10th Annual Meeting of the Organization for Human Brain Mapping*, volume 22, page 01, 2004.

[CAL+16] Özgün Cicek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432. Springer International Publishing, 2016. doi: https://doi.org/10.1007/978-3-319-46723-8_49.

[CAMS+23] Pierre-Henri Conze, Gustavo Andrade-Miranda, Vivek K. Singh, Vincent Jaouen, and Dimitris Visvikis. Current and Emerging Trends in Medical Image Segmentation With Deep Learning. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 7(6):545–569, 2023. doi: https://doi.org/10.1109/TRPMS.2023.3265863.

[CB13] Charles A. Cuenod and Daniel Balvay. Perfusion and Vascular Permeability: Basic Concepts and Measurement in DCE-CT and DCE-MRI. *Diagn Interv Imaging*, 94(12):1187–204, 2013. doi: https://doi.org/10.1016/j.diii.2013.10.010.

[CDT16] Ahmad Chaddad, Christian Desrosiers, and Matthew Toews. Radiomic Analysis of Multi-Contrast Brain MRI for the Prediction of Survival in Patients with Glioblastoma Multiforme. *Annu*

Int Conf IEEE Eng Med Biol Soc, pages 4035–4038, 2016. doi: https://doi.org/10.1109/EMBC.2016.7591612.

[CdVSG+24] Maria Correia de Verdier, Rachit Saluja, Louis Gagnon, Dominic LaBella, Ujjwall Baid, Nourel Hoda Tahon, Martha Foltyn-Dumitru, Jikai Zhang, Maram Alafif, Saif Baig, Ken Chang, Gennaro D'Anna, Lisa Deptula, Diviya Gupta, Muhammad Ammar Haider, Ali Hussain, Michael Iv, Marinos Kontzialis, Paul Manning, and Jeffrey Rudie. *The 2024 Brain Tumor Segmentation (BraTS) challenge: glioma segmentation on post-treatment MRI.* May 2024. doi: http://dx.doi.org/10.48550/arXiv.2405.18368.

[CGD+12] Changho Choi, Sandeep K. Ganji, Ralph J. DeBerardinis, Kimmo J. Hatanpaa, Dinesh Rakheja, Zoltan Kovacs, Xiao-Li Yang, Tomoyuki Mashimo, Jack M. Raisanen, Isaac Marin-Valencia, Juan M. Pascual, Christopher J. Madden, Bruce E. Mickey, Craig R. Malloy, Robert M. Bachoo, and Elizabeth A. Maher. 2-Hydroxyglutarate Detection by Magnetic Resonance Spectroscopy in IDH-Mutated Patients with Gliomas. *Nat Med*, 18(4):624–9, 2012. doi: https://doi.org/10.1038/nm.2682.

[Cho15] Francois Chollet. Keras, 2015. Accessed from: https://keras.io. Last accessed 03/16/2024.

[CIF+23] Maurizio Cè, Giovanni Irmici, Chiara Foschini, Giulia Maria Danesini, Lydia Viviana Falsitta, Maria Lina Serio, Andrea Fontana, Carlo Martinenghi, Giancarlo Oliva, and Michaela Cellina. Artificial Intelligence in Brain Tumor Imaging: A Step toward Personalized Medicine. *Current Oncology*, 30(3):2673–2701, 2023. doi: https://doi.org/10.3390/curroncol30030203.

[CMB+22] Edward Chan, Philip Martin, Caterina Brighi, Sugendran Pillay, Lois Holloway, Peter Metcalfe, and Eng-Siew Koh. NIMG-75. Repeatability of Manual Segmentation of Glioblastoma on MRI - Quality Assurance for a Quantitative MRI Radiomics Repeatability Study. *Neuro-Oncology*, 24(Issue Supplement_7):p182, Nov 2022. doi: https://doi.org/10.1093/neuonc/noac209.693.

[Col23] Google Colab. Google Colaboratory, 2023. Accessed from: https://colab.research.google.com. Last accessed 03/16/2024.

[Con20] The MONAI Consortium. Project MONAI. Zenodo, 2020. Accessed from: https://doi.org/10.5281/zenodo.4323059. Last accessed 03/16/2024.

[Cor24] IBM Corp. IBM SPSS Statistics for Windows (Version 29.0) [Computer software]. NY: IBM Corp. 2024. Accessed from: https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-29. Last accessed 03/16/2024.

[Cos23] Manuel Cossio. *Augmenting Medical Imaging: A Comprehensive Catalogue of 65 Techniques for Enhanced Data Analysis*. Mar 2023. doi: https://doi.org/10.48550/arXiv.2303.01178.

[CRF23] Adrian Celaya, Béatrice M. Rivière, and David T. Fuentes. A Generalized Surface Loss for Reducing the Hausdorff Distance in Medical Imaging Segmentation. 2023. doi: https://doi.org/10.48550/arXiv.2302.03868.

[CWS+22] Boyuan Chen, Mingzhi Wen, Yong Shi, Dayi Lin, Gopi Krishnan Rajbahadur, and Zhen Ming Jiang. Towards Training Reproducible Deep Learning Models, May 2022. doi: https://doi.org/10.1145/3510003.3510163.

[Dev16] TensorFlow Developers. TensorFlow (v2.15.0). Zenodo. 2016. Accessed from: https://doi.org/10.5281/ZENODO.4724125. Last accessed 03/16/2024.

[Dhe14] Frederic Dhermain. Radiotherapy of high-grade gliomas: Current standards and new concepts, innovations in imaging and radiotherapy, and new therapeutic approaches. *Chin J Cancer*, 33(1):16–24, Jan 2014. doi: https://doi.org/10.5732/cjc.013.10217.

[DLCG16] Pedro D. Delgado-López and Eva M. Corrales-García. Survival in Glioblastoma: A Review on the Impact of Treatment Modalities. *Clinical and Translational Oncology*, 18(11):1062–1071, 2016. doi: https://doi.org/10.1007/s12094-016-1497-x.

[DMFAMJRV19] Coral Durand-Munoz, Eduardo Flores-Alvarez, Sergio Moreno-Jimenez, and Ernesto Roldan-Valadez. Pre-Operative Apparent Diffusion Coefficient Values and Tumor Region Volumes as Prognostic Biomarkers in Glioblastoma: Correlation and Progression-Free Survival Analyses. *Insights Imaging*, 10(1):36, Mar 2019. doi: https://doi.org/10.1186/s13244-019-0724-8.

[DPSS19] Alessio D'Alessio, Gabriella Proietti, Gigliola Sica, and Bianca Maria Scicchitano. Pathological and Molecular Features of Glioblastoma and Its Peritumoral Tissue. *Cancers*, 11(4):469, 2019. doi: https://doi.org/10.3390/cancers11040469.

[DV16] Vincent Dumoulin and Francesco Visin. A Guide to Convolution Arithmetic for Deep Learning. *ArXiv*, Mar 2016. doi: https://doi.org/10.48550/arXiv.1603.07285.

[DYL+17] Hao Dong, Guang Yang, Fangde Liu, Yuanhan Mo, and Yike Guo. Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks. May 2017. doi: https://doi.org/10.48550/arXiv.1705.03820.

[EOPB+22] Youssef El Ouadih, Bruno Pereira, Julian Biau, Beatrice Claise, Remi Chaix, Pierre Verrelle, Toufik Khalil, Xavier Durando, and Jean-Jacques Lemaire. DTI Abnormalities Related to Glioblastoma: A Prospective Comparative Study with Metastasis and Healthy Subjects. *Curr Oncol*, 29(4):2823–2834, Apr 2022. doi: https://doi.org/10.3390/curroncol29040230.

[FADK+18] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Marianne Amitai, Jacob Goldberger, and Hayit Greenspan. GAN-based Synthetic Medical Image Augmentation for increased CNN Performance in Liver Lesion Classification. *Neurocomputing*, 321:321–331, 2018. doi: https://doi.org/10.1016/j.neucom.2018.09.013.

[FBKC+12] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, John Buatti, Stephen Aylward, James V. Miller, Steve Pieper, and Ron Kikinis. 3D Slicer as an Image Computing Platform for the Quantitative Imaging Network. *Magn Reson Imaging*, 30(9):1323–41, 2012. doi: https://doi.org/10.1016/j.mri.2012.05.001.

[FCA+87] Franz Fazekas, John B. Chawluk, Abass Alavi, Howard I. Hurtig, and Robert A. Zimmerman. MR Signal Abnormalities at 1.5 T in Alzheimer's Dementia and Normal Aging. *AJR Am J Roentgenol*, 149(2):351–6, Aug 1987. doi: https://doi.org/10.2214/ajr.149.2.351.

[FHP+21] Austin-John Fordham, Caitlin-Craft Hacherl, Neal Patel, Keri Jones, Brandon Myers, Mickey Abraham, and Julian Gendreau. Differentiating Glioblastomas from Solitary Brain Metastases: An Update on the Current Literature of Advanced Imaging Modalities. *Cancers (Basel)*, 13(12):2960, Jun 2021. doi: https://doi.org/10.3390/cancers13122960.

[FLG+24] Haiqing Fan, Yilin Luo, Fang Gu, Bin Tian, Yongqin Xiong, Guipeng Wu, Xin Nie, Jing Yu, Juan Tong, and Xin Liao. Artificial Intelligence-Based MRI Radiomics and Radiogenomics in Glioma. *Cancer Imaging*, 24(1):36, 2024. doi: https://doi.org/10.1186/s40644-024-00682-y.

[FSL+24] 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*, volume abs/2402.17317. 2024. doi: https://doi.org/10.48550/arXiv.2402.17317.

[FT18] Farzan Farnia and David Tse. A Convex Duality Framework for GANs. *ArXiv*, abs/1810.11740, 2018. doi: https://doi.org/10.48550/arXiv.1810.11740.

[FTPM20] Xue Feng, Nicholas J. Tustison, Sohil H. Patel, and Craig H. Meyer. Brain Tumor Segmentation Using an Ensemble of 3D U-Nets and Overall Survival Prediction Using Radiomic Features. *Frontiers in Computational Neuroscience*, 14, 2020. doi: https://doi.org/10.3389/fncom.2020.00025.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. Adaptive Computation and Machine Learning series. MIT Press, London, England, 2016.

[GBGC18] Megha Goyal, Bharat Bhushan, Shailender Gupta, and Rashmi Chawla. Contrast Enhancement Technique Based on Lifting Wavelet Transform. *3D Research*, 9(4):50, 2018. doi: https://doi.org/10.1007/s13319-018-0201-z.

[GKD+11] Norbert Galldiks, Lutz W. Kracht, Veronika Dunkl, Roland T. Ullrich, Stefan Vollmar, Andreas H. Jacobs, Gereon R. Fink, and Michael Schroeter. Imaging of Non- or Very Subtle Contrast-Enhancing Malignant Gliomas with [(11)C]-Methionine Positron Emission Tomography. *Mol Imaging*, 10(6):453–9, Dec 2011. PMID: 22201536.

[Goc23] Evgin Goceri. Medical Image Data Augmentation: Techniques, Comparisons and Interpretations. *Artificial Intelligence Review*, 56(11):12561–12605, 2023. doi: https://doi.org/10.1007/s10462-023-10453-z.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Y. Bengio. Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, 3, 2014. doi: https://doi.org/10.1145/3422622.

[GSD24] Eyad Gad, Seif Soliman, and M. Saeed Darweesh. Advancing Brain Tumor Segmentation via Attention-Based 3D U-Net Architecture and Digital Image Processing. In Mohamed Mosbah, Tahar Kechadi, Ladjel Bellatreche, and Faiez Gargouri, editors, *Model and Data Engineering*, pages 245–258. Springer Nature Switzerland, 2024. doi: https://doi.org/10.1007/978-3-031-49333-1_18.

[HHAH+23] Ghasem Hajianfar, Atlas Haddadi Avval, Seyyed Ali Hosseini, Mostafa Nazari, Mehrdad Oveisi, Isaac Shiri, and Habib Zaidi. Time-to-Event Overall Survival Prediction in Glioblastoma Multiforme Patients Using Magnetic Resonance Imaging Radiomics. *La radiologia medica*, 128(12):1521–1534, 2023. doi: https://doi.org/10.1007/s11547-023-01725-3.

[HHY+19] Chao Huang, Hu Han, Qingsong Yao, Shankuan Zhu, and S. Kevin Zhou. 3D U^2-Net: A 3D Universal U-Net for Multi-Domain Medical Image 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 291–299. Springer International Publishing, 2019. doi: https://doi.org/10.1007/978-3-030-32245-8_33.

[HJHK19] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *Journal of Digital Imaging*, 32(4):582–596, 2019. doi: https://doi.org/10.1007/s10278-019-00227-x.

[HKR93] Daniel P. Huttenlocher, Gregory A. Klanderman, and William J. Rucklidge. Comparing Images Using the Hausdorff Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993. doi: https://doi.org/10.1109/34.232073.

[HMAZ23] Aminou Halidou, Youssoufa Mohamadou, Ado Adamou Abba Ari, and Edinio Jocelyn Gbadoubissa Zacko. Review of Wavelet Denoising Algorithms. *Multimedia Tools and Applications*, 82(27):41539–41569, 2023. doi: https://doi.org/10.1007/s11042-023-15127-0.

[HMvdW+20] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández Del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. doi: https://doi.org/10.1038/s41586-020-2649-2.

[HNP+13] Markus Hutterer, Martha Nowosielski, Daniel Putzer, Nathalie L. Jansen, Marcel Seiz, Michael Schocke, Mark McCoy, Georg Gobel, Christian la Fougere, Irene J. Virgolini, Eugen Trinka, Andreas H. Jacobs, and Gunther Stockhammer. [18F]-Fluoro-Ethyl-L-Tyrosine PET: A Valuable Diagnostic Tool in Neuro-Oncology, but Not All

That Glitters Is Glioma. *Neuro Oncol*, 15(3):341–51, Mar 2013. doi: https://doi.org/10.1093/neuonc/nos300.

[HNT+21] 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*, pages 272–284. Springer-Verlag, Berlin, Heidelberg, 2021. doi: https://doi.org/10.1007/978-3-031-08999-2_22.

[HSF24] Wolfgang Karl Härdle, Léopold Simar, and Matthias Fengler. *Applied Multivariate Statistical Analysis*. Springer, 2024. doi: https://doi.org/10.1007/978-3-031-63833-6.

[HZRS16] 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. doi: https://doi.org/10.1109/CVPR.2016.90.

[IJK+21] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. *Nature Methods*, 18(2):203–211, 2021. doi: https://doi.org/10.1038/s41592-020-01008-z.

[IOK+16] Rika Inano, Naoya Oishi, Takeharu Kunieda, Yoshiki Arakawa, Takayuki Kikuchi, Hidenao Fukuyama, and Susumu Miyamoto. Visualization of Heterogeneity and Regional Grading of Gliomas by Multiple Features Using Magnetic Resonance-Based Clustered Images. *Sci Rep*, 6:30344, Jul 2016. doi: https://doi.org/10.1038/srep30344.

[ISP+19] 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, 2019. doi: https://doi.org/10.1002/hbm.24750.

[Jac01] Paul Jaccard. Etude de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37:547–579, 1901. doi: https://doi.org/10.5169/seals-266450.

[Jad20] Shruti Jadon. A Survey of Loss Functions for Semantic Segmentation. pages 1–7, Oct 2020. doi: https://doi.org/10.1109/CIBCB48159.2020.9277638.

[Jan15] Mariëlle Jansen. *Evaluation of intensity normalization methods for MR images*. Thesis, University Medical Center Utrecht, Jul 2015. doi: https://doi.org/10.13140/RG.2.2.19469.69606.

[JHG+19] Amod Jog, Andrew Hoopes, Douglas N. Greve, Koen Van Leemput, and Bruce Fischl. PSACNN: Pulse Sequence Adaptive Fast Whole Brain Segmentation. *Neuroimage*, 199:553–569, Oct 2019. doi: https://doi.org/10.1016/j.neuroimage.2019.05.033.

[JLOC14] Geon-Ho Jahng, Ka-Loh Li, Leif Ostergaard, and Fernando Calamante. Perfusion Magnetic Resonance Imaging: A Comprehensive Update on Principles and Techniques. *Korean J Radiol*, 15(5):554–77, Sep-Oct 2014. doi: https://doi.org/10.3348/kjr.2014.15.5.554.

[KBMB24] Teerath Kumar, Rob Brennan, Alessandra Mileo, and Malika Bendechache. Image Data Augmentation Approaches: A Comprehensive Survey and Future Directions. *IEEE Access*, 12:187536–187571, 2024. doi: https://doi.org/10.1109/ACCESS.2024.3470122.

[KBR20] Martin Kolarik, Radim Burget, and Kamil Riha. *Comparing Normalization Methods for Limited Batch Size Segmentation Neural Networks*. IEEE, Jul 2020. doi: https://doi.org/10.48550/arXiv.2011.11559.

[KBW+20] Florian Kofler, Christoph Berger, Diana Waldmannstetter, Jana Lipkova, Ivan Ezhov, Giles Tetteh, Jan Kirschke, Claus Zimmer, Benedikt Wiestler, and Bjoern H. Menze. BraTS Toolkit: Translating BraTS Brain Tumor Segmentation Algorithms Into Clinical and Scientific Practice. *Front Neurosci*, 14:125, Apr 2020. doi: https://doi.org/10.3389/fnins.2020.00125.

[KEB+21] Alexander Ke, William Ellsworth, Oishi Banerjee, A. Ng, and Pranav Rajpurkar. CheXtransfer: Performance and Parameter Efficiency of ImageNet Models for Chest X-Ray Interpretation. *Proceedings of the Conference on Health, Inference, and Learning*, Apr 2021. doi: https://doi.org/10.1145/3450439.3451867.

[KKR+18] Doo-Sik Kong, Junhyung Kim, Gyuha Ryu, Hye-Jin You, Joon Kyung Sung, Yong Hee Han, Hye-Mi Shin, In-Hee. Lee, Sung-Tae Kim, Chul-Kee Park, Seung Hong Choi, Jeong Won Choi, Ho Jun Seol, Jung-Il Lee, and Do-Hyun Nam. Quantitative Radiomic Profiling of Glioblastoma Represents Transcriptomic Expression. *Oncotarget*, 9(5):6336–6345, Jan 2018. doi: https://doi.org/10.18632/oncotarget.23975.

[KM04] Peter B. Kingsley and W. Gordon Monahan. Selection of the Optimum B Factor for Diffusion-Weighted Magnetic Resonance Imaging Assessment of Ischemic Stroke. *Magn Reson Med*, 51(5):996–1001, May 2004. doi: https://doi.org/10.1002/mrm.20059.

[KPG+24] Mert Karabacak, Shiv Patil, Zachary Charles Gersey, Ricardo Jorge Komotar, and Konstantinos Margetis. Radiomics-Based Machine

Learning with Natural Gradient Boosting for Continuous Survival Prediction in Glioblastoma. *Cancers*, 16(21):3614, 2024. doi: https://doi.org/10.3390/cancers16213614.

[KRA23] Gurinderjeet Kaur, Prashant Singh Rana, and Vinay Arora. Extracting Radiomic Features from Pre-Operative and Segmented MRI Scans Improved Survival Prognosis of Glioblastoma Multiforme Patients Through Machine Learning: A Retrospective Study. *Multimedia Tools and Applications*, 82(19):30003–30038, 2023. doi: https://doi.org/10.1007/s11042-022-14223-x.

[KS20] Davood Karimi and Septimiu Salcudean. Reducing the Hausdorff Distance in Medical Image Segmentation With Convolutional Neural Networks. *IEEE Transactions on Medical Imaging*, 39(2):499–513, 2020. doi: https://doi.org/10.1109/TMI.2019.2930068.

[KSH17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM*, 60(6):84–90, 2017. doi: https://doi.org/10.1145/3065386.

[LGG+20] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. doi: https://doi.org/10.1109/TPAMI.2018.2858826.

[LJZZ23] Mengfang Li, Yuanyuan Jiang, Yanzhou Zhang, and Haisheng Zhu. Medical Image Analysis Using Deep Learning Algorithms. *Frontiers in Public Health*, 11, 2023. doi: https://doi.org/10.3389/fpubh.2023.1273253.

[LPO22] Anastasia-Maria Leventi-Peetz and Thomas Oestreich. *Deep Learning Reproducibility and Explainable AI (XAI).* 2022. doi: https://doi.org/10.48550/arXiv.2202.11452.

[LPW+21] David N. Louis, Arie Perry, Pieter Wesseling, Daniel J. Brat, Ian A. Cree, Dominique. Figarella-Branger, Cynthia Hawkins, H. K. Ng, Stefan M. Pfister, Guido Reifenberger, Riccardo Soffietti, Andreas von Deimling, and David W. Ellison. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol*, 23(8):1231–1251, Aug 2021. doi: https://doi.org/10.1093/neuonc/noab106.

[LQX+24] Xiaoyu Liu, Linhao Qu, Ziyue Xie, Jiayue Zhao, Yonghong Shi, and Zhijian Song. Towards More Precise Automatic Analysis: A Systematic Review of Deep Learning-Based Multi-Organ Segmentation. *BioMedical Engineering OnLine*, 23(1):52, Jun 2024. doi: https://doi.org/10.1186/s12938-024-01238-8.

[LWL+21] Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. *R-Drop: Regularized Dropout for Neural Networks*. 2021. doi: https://doi.org/10.48550/arXiv.2106.14448.

[LWS+84] Michael Laniado, Hans-Joachim Weinmann, Wolfgang Schorner, Rainer Felix, and Ulrich Speck. First Use of GdDTPA/Dimeglumine in Man. *Physiol Chem Phys Med NMR*, 16(2):157–65, 1984. PMID: 6505042.

[Mat24] Matplotlib. Visualization with Python., 2024. Accessed from: https://matplotlib.org. Last accessed 03/16/2024.

[MBC+08] Pratik Mukherjee, Jamie I. Berman, Sung W. Chung, Christopher P. Hess, and Roland G. Henry. Diffusion Tensor MR Imaging and Fiber Tractography: Theoretic Underpinnings. *AJNR Am J Neuroradiol*, 29(4):632–41, Apr 2008. doi: https://doi.org/10.3174/ajnr.A1051.

[MCV+97] Frederik Maes, Albert Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality Image Registration by Maximization of Mutual Information. volume 16, pages 187–198, 1997. doi: https://doi.org/10.1109/42.563664.

[MD22] Calum MacLellan and Feng Dong. *Hyper-Learning for Gradient-Based Batch Size Adaptation*. May 2022. doi: https://doi.org/10.48550/arXiv.2205.08231.

[MGM+24] Abiy Abinet Mamo, Bealu Girma Gebresilassie, Aniruddha Mukherjee, Vikas Hassija, and Vinay Chamola. Advancing Medical Imaging Through Generative Adversarial Networks: A Comprehensive Review and Future Prospects. *Cognitive Computation*, 16(5):2131–2153, 2024. doi: https://doi.org/10.1007/s12559-024-10291-3.

[MJB+15] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M.

Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. doi: https://doi.org/10.1109/TMI.2014.2377694.

[MKE+24] Sebastian Johannes Müller, Eya Khadhraoui, Marielle Ernst, Veit Rohde, Bawarjan Schatlo, and Vesna Malinova. Differentiation of Multiple Brain Metastases and Glioblastoma With Multiple Foci Using MRI Criteria. *BMC Med Imaging*, 24(1):3, Jan 2024. doi: https://doi.org/10.1186/s12880-023-01183-3.

[MKF+20] Razi Muzaffar, Elyse Koester, Sarah Frye, Saud Alenezi, Barbara B. Sterkel, and Medhat M. Osman. Development of Simple Methods to Reduce the Exposure of the Public to Radiation From Patients Who Have Undergone ^{18}F-FDG PET/CT. *Journal of Nuclear Medicine Technology*, 48(1):63–67, 2020. doi: https://doi.org/10.2967/jnmt.119.233296.

[MKL+16] Raphael Meier, Urspeter Knecht, Tina Loosli, Stefan Bauer, Johannes Slotboom, Roland Wiest, and Maurici Reyes. Clinical Evaluation of a Fully-automatic Segmentation Method for Longitudinal Brain Tumor Volumetry. *Sci Rep*, 6:23376, Mar 2016. doi: https://doi.org/10.1038/srep23376.

[MM22] Alhassan Mumuni and Fuseini Mumuni. Data Augmentation: A Comprehensive Survey of Modern Approaches. *Array*, 16:100258, 2022. doi: https://doi.org/10.1016/j.array.2022.100258.

[MMAC23] Ahmed Makhlouf, Marina Maayah, Nada Abughanam, and Cagatay Catal. The Use of Generative Adversarial Networks in Medical Image Augmentation. *Neural Computing and Applications*, 35(34):24055–24068, 2023. doi: https://doi.org/10.1007/s00521-023-09100-z.

[MOK+21] Kimberly D. Miller, Quinn T. Ostrom, Carol Kruchko, Nirav Patil, Tarik Tihan, Gino Cioffi, Hannah E. Fuchs, Kristin A. Waite, Ahmedin Jemal, Rebecca L. Siegel, and Jill S. Barnholtz-Sloan. Brain and Other Central Nervous System Tumor Statistics, 2021. *CA Cancer J Clin*, 71(5):381–406, Sep 2021. doi: https://doi.org/10.3322/caac.21693.

[MRV03] Miguel Murguía-Romero and Jose Villaseñor. Estimating the Effect of the Similarity Coefficient and the Cluster Algorithm on Biogeographic Classification. *Annales Botanici Fennici*, 40:415–421, Jan 2003.

[MSRK22] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. Towards a Guideline for Evaluation Metrics in Medical Image Segmentation. *BMC*

Research Notes, 15(1):210, 2022. doi: https://doi.org/10.1186/s13104-022-06096-y.

[MVL+14] Remco J. Molenaar, Dagmar Verbaan, Simona Lamba, C. Zanon, Judith W. Jeuken, Sandra H. Boots-Sprenger, Pieter Wesseling, Theo J. Hulsebos, Dirk Troost, Angela A. van Tilborg, Sieger Leenstra, W. Peter Vandertop, Alberto Bardelli, Cornelis J. van Noorden, and Fonnet E. Bleeker. The Combination of IDH1 Mutations and MGMT Methylation Status Predicts Survival in Glioblastoma Better Than Either IDH1 or MGMT Alone. *Neuro Oncol*, 16(9):1263–73, Sep 2014. doi: https://doi.org/10.1093/neuonc/nou005.

[MW47] Henry B. Mann and Donald R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, 11, 1947. doi: https://doi.org/10.1214/aoms/1177730491.

[NMS23] Venkata P. Nithya, Natarajan Mohanasundaram, and R. Santhosh. An Early Detection and Classification of Alzheimer's Disease Framework Based on ResNet-50. *Current Medical Imaging*, pages 1–20, Aug 2023. doi: https://doi.org/10.2174/1573405620666230825113344.

[NPBL22] Loris Nanni, Michelangelo Paci, Sheryl Brahnam, and Alessandra Lumini. Feature Transforms for Image Data Augmentation. *Neural Computing and Applications*, 34(24):22345–22356, 2022. doi: https://doi.org/10.1007/s00521-022-07645-z.

[NPN+21] Anh Nguyen, Khoa Pham, Dat Ngo, Thanh Ngo, and Lam Pham. An Analysis of State-of-the-art Activation Functions For Supervised Deep Neural Network. In *2021 International Conference on System Science and Engineering (ICSSE)*, pages 215–220, 2021. doi: https://doi.org/10.1109/ICSSE52999.2021.9538437.

[NPSA21] Nikhil Nasalwai, Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal. Addressing the Class Imbalance Problem in Medical Image Segmentation via Accelerated Tversky Loss Function. In Kamal Karlapalem, Hong Cheng, Naren Ramakrishnan, R. K. Agrawal, P. Krishna Reddy, Jaideep Srivastava, and Tanmoy Chakraborty, editors, *Advances in Knowledge Discovery and Data Mining*, pages 390–402. Springer International Publishing, 2021. doi: https://doi.org/10.1007/978-3-030-75768-7_31.

[NVI21] NVIDIA. NVIDIA A100 Tensor Core GPU, 2021. Accessed from: https://www.nvidia.com/en-us/data-center/a100/. Last accessed 03/16/2024.

[NWS14] Yi Niu, Xiaolin Wu, and Guangming Shi. Image Enhancement by Entropy Maximization and Quantization Resolution Upconversion. *IEEE Transactions on Image Processing*, pages 4047–4051, 2014. doi: https://doi.org/10.1109/ICIP.2014.7025822.

[NWS18] Prabhat Nagarajan, Garrett Warnell, and Peter Stone. Deterministic Implementations for Reproducibility in Deep Reinforcement Learning. *ArXiv*, abs/1809.05676, Sep 2018. doi: https://doi.org/10.48550/arXiv.1809.05676.

[OCW+21] Quinn T. Ostrom, Gino Cioffi, Kristin 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 2014-2018. *Neuro Oncol*, 23(12 Suppl 2):1–105, 2021. doi: https://doi.org/10.1093/neuonc/noab200.

[OdANC23] André Luiz C. Ottoni, Raphael M. de Amorim, Marcela S. Novo, and Dayana B. Costa. Tuning of Data Augmentation Hyperparameters in Deep Learning for Building Construction Image Classification With Small Datasets. *International Journal of Machine Learning and Cybernetics*, 14(1):171–186, 2023. doi: https://doi.org/10.1007/s13042-022-01555-1.

[OKA+19] Tomohiko Ozaki, Manabu Kinoshita, Hideyuki Arita, Naoki Kagawa, Yasunori Fujimoto, Yonehiro Kanemura, Mio Sakai, Yoshiyuki Watanabe, Katsuyuki Nakanishi, Eku Shimosegawa, Jun Hatazawa, and Haruhiko Kishima. Validation of Magnetic Resonance Imaging-Based Automatic High-Grade Glioma Segmentation Accuracy via (11)C-Methionine Positron Emission Tomography. *Oncol Lett*, 18(4):4074–4081, Oct 2019. doi: https://doi.org/10.3892/ol.2019.10734.

[Org24] Glioblastoma Research Organization. All about glioblastoma - Glioblastoma 101, 2024. Accessed from: https://www.gbmresearch.org/glioblastoma-101. Last accessed 08/10/2024.

[ORS+01] Sebastian Ourselin, Alexis Roche, Gérard Subsol, Xavier Pennec, and Nicholas Ayache. Reconstructing a 3D Structure from Serial Histological Sections. *Image and Vision Computing*, 19(1):25–31, 2001. doi: https://doi.org/10.1016/S0262-8856(00)00052-4.

[OSF+18] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, M. J. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention U-Net: Learning Where to

Look for the Pancreas. *ArXiv*, abs/1804.03999, Apr 2018. doi: http://doi.org/10.48550/arXiv.1804.03999.

[PAA+87] Stephen M. Pizer, E. Philip Amburn, John D. Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B. Zimmerman, and Karel Zuiderveld. Adaptive Histogram Equalization and Its Variations. *Computer Vision, Graphics, and Image Processing*, 39(3):355–368, 1987. doi: https://doi.org/10.1016/S0734-189X(87)80186-X.

[PBP+14] Nicole Porz, Stefan Bauer, Alessia Pica, Philippe Schucht, Jurgen Beck, Rajeev Kumar Verma, Johannes Slotboom, Mauricio Reyes, and Roland Wiest. Multi-Modal Glioblastoma Segmentation: Man Versus Machine. *PLoS One*, 9(5):e96873, May 2014. doi: https://doi.org/10.1371/journal.pone.0096873.

[PCCDM21] Davide Poggiali, Diego Cecchin, Cristina Campi, and Stefano De Marchi. Oversampling Errors in Multimodal Medical Imaging Are Due to the Gibbs Effect. *Mathematics*, 9:1348, 2021. doi: https://doi.org/10.3390/math9121348.

[PCP+22a] Alessia Pellerino, Mario Caccese, Marta Padovan, Giulia Cerretti, and Giuseppe Lombardi. Epidemiology, Risk Factors, and Prognostic Factors of Gliomas. *Clinical and Translational Imaging*, 10(5):467–475, 2022. doi: https://doi.org/10.1007/s40336-022-00489-6.

[PCP+22b] Sveinn Pálsson, Stefano Cerri, Hans S. Poulsen, Thomas Urup, Ian Law, and Koen Van Leemput. Predicting Survival of Glioblastoma From Automatic Whole-Brain and Tumor Segmentation of MR Images. *Sci Rep*, 12(1):19744, 2022. doi: https://doi.org/10.1038/s41598-022-19223-3.

[PDK+22] Tara Pattilachan, Ugur Demir, Elif Keles, Debesh Jha, Derk Klatte, Megan Engels, Sanne Hoogenboom, and Candice Bolan. *A Critical Appraisal of Data Augmentation Methods for Imaging-Based Medical Diagnosis Applications.* 2022. doi: https://doi.org/10.48550/arXiv.2301.02181.

[PGM+19] 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 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. 32:8024–8035, 2019. doi: https://doi.org/10.48550/arXiv.1912.01703.

[PHM+16] Nicole Porz, Simon Habegger, Raphael Meier, Rajeev Verma, Astrid Jilch, Jens Fichtner, Urspeter Knecht, Christian Radina, Philippe Schucht, Jurgen Beck, Andreas Raabe, Johannes Slotboom, Mauricio Reyes, and Roland Wiest. Fully Automated Enhanced Tumor Compartmentalization: Man vs. Machine Reloaded. *PLoS One*, 11(11):e0165302, Nov 2016. doi: https://doi.org/10.1371/journal.pone.0165302.

[PKF+11] Luciana Porto, Matthias Kieslich, Kea Franz, Thomas Lehrnbecher, Friedhelm Zanella, Ulrich Pilatus, and Elke Hattingen. MR Spectroscopy Differentiation Between High- and Low-Grade Astrocytomas: A Comparison Between Paediatric and Adult Tumours. *Eur J Paediatr Neurol*, 15(3):214–21, May 2011. doi: https://doi.org/10.1016/j.ejpn.2010.11.003.

[PLF19] Csaba Pinter, Andras Lasso, and Gabor Fichtinger. Polymorph Segmentation Representation for Medical Image Computing. *Computer Methods and Programs in Biomedicine*, 171:19–26, 2019. doi: https://doi.org/10.1016/j.cmpb.2019.02.011.

[PNA18] Sandur Poornachandra, C. Naveena, and Manjunath Aradhya. Intensity Normalization—A Critical Pre-processing Step for Efficient Brain Tumor Segmentation in MR Images. In Vikrant Bhateja, Bao Le Nguyen, Nhu Gia Nguyen, Suresh Chandra Satapathy, and Dac-Nhuong Le, editors, *Information Systems Design and Intelligent Applications*, pages 885–893. Springer Singapore, Jan 2018. doi: http://doi.org/10.1007/978-981-10-7512-4_87.

[POC14] Agus Pratondo, Sim Ong, and Chee-Kong Chui. *Region Growing for Medical Image Segmentation Using a Modified Multiple-seed Approach on a Multi-core CPU Computer*, volume 43, pages 112–115. Springer International Publishing, 2014. doi: https://doi.org/10.1007/978-3-319-02913-9_29.

[PS20] Vihari Piratla and Shiv Shankar. Untapped Potential of Data Augmentation: A Domain Generalization Viewpoint. *ArXiv*, 2020. doi: https://doi.org/10.48550/arXiv.2007.04662.

[PSR+19] Magdalini Paschali, Walter Simson, Abhijit Guha Roy, Muhammad Ferjad Naeem, Rüdiger Göbl, Christian Wachinger, and Nassir Navab. Data Augmentation with Manifold Exploring Geometric Transformations for Increased Performance and Robustness. *ArXiv*, abs/1901.04420, Jan 2019. doi: https://doi.org/10.48550/arXiv.1901.04420.

[Pyk17] Krystian Pyka. Wavelet-Based Local Contrast Enhancement for Satellite, Aerial and Close Range Images. *Remote Sensing*, 9:25, 2017. doi: https://doi.org/10.3390/rs9010025.

[QGT+19] Lianke Qin, Yifan Gong, Tianqi Tang, Yutian Wang, and Jiangming Jin. Training Deep Nets with Progressive Batch Normalization on Multi-GPUs. *International Journal of Parallel Programming*, 47(3):373–387, 2019. doi: https://doi.org/10.1007/s10766-018-0615-5.

[RB24] Novsheena Rasool and Javaid Iqbal Bhat. A Critical Review on Segmentation of Glioma Brain Tumor and Prediction of Overall Survival. *Archives of Computational Methods in Engineering*, pages 1–45, Oct 2024. doi: https://doi.org/10.1007/s11831-024-10188-2.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing, 2015. doi: https://doi.org/10.1007/978-3-319-24574-4_28.

[RIBM+23] Rehan Raza, Usama Ijaz Bajwa, Yasar Mehmood, Muhammad Waqas Anwar, and M. Hassan Jamal. dResU-Net: 3D Deep Residual U-Net Based Brain Tumor Segmentation From Multimodal MRI. *Biomedical Signal Processing and Control*, 79:103861, 2023. doi: https://doi.org/10.1016/j.bspc.2022.103861.

[RKF+22] Johannes Roth, Johannes Keller, Stefan Franke, Thomas Neumuth, and Daniel Schneider. Multi-plane UNet++ Ensemble for Glioblastoma Segmentation. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 285–294. Springer International Publishing, Jul 2022. doi: https://doi.org/10.1007/978-3-031-08999-2_23.

[RKR+19] Michael Rebsamen, Urspeter Knecht, Mauricio Reyes, Roland Wiest, Raphael Meier, and Richard McKinley. Divide and Conquer: Stratifying Training Data by Tumor Grade Improves Deep Learning-Based Brain Tumor Segmentation. *Frontiers in Neuroscience*, 13, 2019. doi: https://doi.org/10.3389/fnins.2019.01182.

[RT12] Zoran Rumboldt and Majda Thurnher. *Glioblastoma Multiforme*, pages 317–318. Cambridge University Press, Cambridge, 2012. doi: https://doi.org/10.1017/CBO9781139030854.154.

[RZA22] Sivaramakrishnan Rajaraman, Ghada Zamzmi, and Sameer K. Antani. Novel Loss Functions for Ensemble-Based Medical Image Classification. *PLOS ONE*, 16(12):e0261307, 2022. doi: https://doi.org/10.1371/journal.pone.0261307.

[RZKB19] Maithra Raghu, Chiyuan Zhang, Jon M. Kleinberg, and Samy Bengio. Transfusion: Understanding Transfer Learning for Medical Imaging. In *Neural Information Processing Systems*, 2019. doi: https://doi.org/10.48550/arXiv.1902.07208.

[RZSP10] Torsten Rohlfing, Natalie M. Zahr, Edith V. Sullivan, and Adolf Pfefferbaum. The SRI24 Multichannel Atlas of Normal Adult Human Brain Structure. *Hum Brain Mapp*, 31(5):798–819, 2010. doi: https://doi.org/10.1002/hbm.20906.

[SDW+19] Hossein Shooli, Habibollah Dadgar, Yi-Xiang J. Wang, Manochehr Seyedi Vafaee, Saman Rassaei Kashuk, Reza Nemati, Esmail Jafari, Iraj Nabipour, Ali Gholamrezanezhad, Majid Assadi, and Mykol Larvie. An Update on PET-Based Molecular Imaging in Neuro-Oncology: Challenges and Implementation for a Precision Medicine Approach in Cancer Care. *Quant Imaging Med Surg*, 9(9):1597–1610, Sep 2019. doi: https://doi.org/10.21037/qims.2019.08.16.

[SHK+14] 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*, 15:1929–1958, Jan 2014.

[SK19] Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60, 2019. doi: https://doi.org/10.1186/s40537-019-0197-0.

[SKA+20] Yannick Suter, Urspeter Knecht, Mariana Alão, Waldo Valenzuela, Ekkehard Hewer, Philippe Schucht, Roland Wiest, and Mauricio Reyes. Radiomics for Glioblastoma Survival Analysis in Pre-Operative MRI: Exploring Feature Robustness, Class Boundaries, and Machine Learning Techniques. *Cancer Imaging*, 20(1):55, 2020. doi: https://doi.org/10.1186/s40644-020-00329-8.

[SKS+21] Simon Schmid, Johannes Krabusch, Thomas Schromm, Shi Jieqing, Stefan Ziegelmeier, Christian Ulrich Grosse, and Johannes Henrich Schleifenbaum. A New Approach for Automated Measuring of the Melt Pool Geometry in Laser-Powder Bed Fusion. *Progress in Additive Manufacturing*, 6(2):269–279, 2021. doi: https://doi.org/10.1007/s40964-021-00173-7.

[SLV+17] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and Manuel Jorge Cardoso. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. page 240–248, 2017. doi: https://doi.org/10.48550/arXiv.1707.03237.

[SMAS13] Jose Soares, Paulo Marques, Victor Alves, and Nuno Sousa. A Hitchhiker's Guide to Diffusion Tensor Imaging. *Frontiers in Neuroscience*, 7, 2013. doi: https://doi.org/10.3389/fnins.2013.00031.

[SMB10] Dominik Scherer, Andreas Müller, and Sven Behnke. *Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition*. Springer Berlin Heidelberg, 2010. doi: https://doi.org/10.1007/978-3-642-15825-4_10.

[SOM+20] Neetu Soni, Manish Ora, Namita Mohindra, Y. Menda, and Girish Bathla. Diagnostic Performance of PET and Perfusion-Weighted Imaging in Differentiating Tumor Recurrence or Progression from Radiation Necrosis in Posttreatment Gliomas: A Review of Literature. *AJNR Am J Neuroradiol*, 41(9):1550–1557, Sep 2020. doi: https://doi.org/10.3174/ajnr.A6685.

[SPPDSBO22] Fidan Seker-Polat, Nareg Pinarbasi Degirmenci, Ihsan Solaroglu, and Tugba Bagci-Onder. Tumor Cell Infiltration into the Brain in Glioblastoma: From Mechanisms to Clinical Perspectives. *Cancers*, 14(2):443, 2022. doi: https://doi.org/10.3390/cancers14020443.

[ST65] Edward O. Stejskal and John E. Tanner. Spin Diffusion Measurements: Spin Echoes in the Presence of a Time-Dependent Field Gradient. *The Journal of Chemical Physics*, 42(1):288–292, 1965. doi: https://doi.org/10.1063/1.1695690.

[Sta11] Richard P. Stanley. *Enumerative Combinatorics: Volume 1*. Cambridge University Press, USA, 2nd edition, 2011. doi: https://doi.org/10.1017/CBO9781139058520.

[STIM19] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How Does Batch Normalization Help Optimization? (No, It Is Not About Internal Covariate Shift). In *Neural Information Processing Systems*, Apr 2019. doi: https://doi.org/10.48550/arXiv.1805.11604.

[SZL+24] Yingxu Song, Yujia Zou, Yuan Li, Yueshun He, Weicheng Wu, Ruiqing Niu, and Shuai Xu. Enhancing Landslide Detection with SBConv-Optimized U-Net Architecture Based on Multisource Remote Sensing Data. *Land*, 13(6):835, Jun 2024. doi: http://doi.org/10.3390/land13060835.

[Sø48] Thorvald Julius Sørensen. *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*. I kommission hos E. Munksgaard, København, 1948.

[TB24] Eleonora Tiribilli and Leonardo Bocchi. Deep Learning-Based Workflow for Bone Segmentation and 3D Modeling in Cone-Beam CT Orthopedic Imaging. *Applied Sciences*, 14(17):7557, 2024. doi: https://doi.org/10.3390/app14177557.

[Ten24] TensorFlow. Tensorflow Mixed precision, 2024. Accessed from: https://www.tensorflow.org/guide/mixed_precision. Last accessed 03/16/2024.

[TH15] Abdel Aziz Taha and Allan Hanbury. Metrics for Evaluating 3D Medical Image Segmentation: Analysis, Selection, and Tool. *BMC Med Imaging*, 15:29, Aug 2015. doi: https://doi.org/10.1186/s12880-015-0068-x.

[TMC+11] Gerard Thompson, Samantha J. Mills, David J. Coope, James P. O'Connor, and Alan Jackson. Imaging Biomarkers of Angiogenesis and the Microvascular Environment in Cerebral Tumours. *Br J Radiol*, 84 Spec No 2(Spec Iss 2):S127–44, Dec 2011. doi: https://doi.org/10.1259/bjr/66316279.

[Tof97] Paul Tofts. Modeling Tracer Kinetics in Dynamic Gd-DTPA MR Imaging. *Journal of Magnetic Resonance Imaging*, 7:91–101, 1997. doi: https://doi.org/10.1002/jmri.1880070113.

[TRX23] Chao Tang, Rongcheng Ruan, and Zhaoying Xiong. Comparison Between [^{18}F]FET PET/MRI and [^{18}F]FET PET/CT in the Diagnosis of Glioma Recurrence: A Systematic Review and Meta-Analysis. 11(5):479–491, 2023. doi: https://doi.org/10.1007/s40336-023-00585-1.

[TWZ+22] Suqing Tian, Cuiying Wang, Ruiping Zhang, Zhuojie Dai, Lecheng Jia, Wei Zhang, Junjie Wang, and Yinglong Liu. Transfer Learning-Based Autosegmentation of Primary Tumor Volumes of Glioblastomas Using Preoperative MRI for Radiotherapy Treatment. *Frontiers in Oncology*, 12, 2022. doi: https://doi.org/10.3389/fonc.2022.856346.

[VARS24] Rao Vinay, Monika Agarwal, Geeta Rani, and Aparajita Sinha. Contrast Enhancement of Medical Images Using Otsu's Double Threshold. In Harish Sharma, Vivek Shrivastava, Ashish Kumar Tripathi, and Lipo Wang, editors, *Communication and Intelligent Systems*, pages 195–208. Springer Nature Singapore, May 2024. doi: https://doi.org/10.1007/978-981-97-2079-8_16.

[VCS24] Shradha Verma, Anuradha Chug, and Amit Prakash Singh. Revisiting Activation Functions: Empirical Evaluation for Image Understanding and Classification. *Multimedia Tools and Applications*, 83(6):18497–18536, 2024. doi: https://doi.org/10.1007/s11042-023-16159-2.

[VKG+22] Dominik Vilimek, Jan Kubicek, Milos Golian, Rene Jaros, Radana Kahankova, Pavla Hanzlikova, Daniel Barvik, Alice Krestanova, Marek Penhaker, Martin Cerny, Ondrej Prokop, and Marek Buzga. Comparative Analysis of Wavelet Transform Filtering Systems for Noise Reduction in Ultrasound Images. *PLOS ONE*, 17(7):e0270745, 2022. doi: https://doi.org/10.1371/journal.pone.0270745.

[VMV+18] Martin Visser, Domenique Müller, Niels Verburg, Roelant Eijgelaar, Marnix Witte, Frederik Barkhof, Philip de Witt Hamer, and Jan de Munck. Inter-Observer Variation in Segmenting Glioma on MRI Before and After Resection. In Hannu Eskola, Outi Väisänen, Jari Viik, and Jari Hyttinen, editors, *EMBEC NBC 2017*, pages 161–164. Springer Singapore, 2018. doi: https://doi.org/10.1007/978-981-10-5122-7_41.

[VNL20] Minh H. Vu, Tufve Nyholm, and Tommy Löfstedt. *TuNet: End-to-End Hierarchical Brain Tumor Segmentation Using Cascaded Networks*, page 174–186. Springer International Publishing, 2020. doi: https://doi.org/10.1007/978-3-030-46640-4_17.

[VR09] Fred L. Van Rossum, Guido und Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[WC18] Pei Wang and Albert C. S. Chung. Focal dice loss and image dilation for brain tumor segmentation. In Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer Syeda-Mahmood, Anne Martel, Lena Maier-Hein, João Manuel R. S. Tavares, Andrew Bradley, João Paulo Papa, Vasileios Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayit Greenspan, and Anant Madabhushi, editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 119–127. Springer International Publishing, 2018. doi: https://doi.org/10.1007/978-3-030-00889-5_14.

[WHSX23] Fangyuan Wang, Ming Hao, Yuhai Shi, and Bo Xu. *ApproBiVT: Lead ASR Models to Generalize Better Using Approximated Bias-Variance Tradeoff Guided Early Stopping and Checkpoint Averaging*. Aug 2023. doi: https://doi.org/10.48550/arXiv.2308.02870.

[WKSM21] Brent D. Weinberg, Manohar Kuruva, Hyunsuk Shim, and Mark E. Mullins. Clinical Applications of Magnetic Resonance Spectroscopy in Brain Tumors: From Diagnosis to Treatment. *Radiol Clin North Am*, 59(3):349–362, May 2021. doi: https://doi.org/10.1016/j.rcl.2021.01.004.

[WKTLRR22] Michael Weller, Christiane B. Knobbe-Thomsen, Emilie Le Rhun, and Guido Reifenberger. Die WHO-Klassifikation der Tumoren des

zentralen Nervensystems 2021. *Der Onkologe*, 28(2):155–163, 2022. doi: https://doi.org/10.1007/s00761-021-01083-7.

[WLSC22] Yunze Wang, Qingyu Lin, Hongcheng Shi, and Dengfeng Cheng. Fluorine-18: Radiochemistry and Target-Specific PET Molecular Probes Design. *Frontiers in Chemistry*, 10:884517, Jun 2022. doi: https://doi.org/10.3389/fchem.2022.884517.

[WRM+14] Amanda J. Walker, Jake Ruzevick, Ashkan A. Malayeri, Daniele Rigamonti, Michael Lim, Kristin J. Redmond, and Lawrence Kleinberg. Postradiation imaging changes in the cns: how can we differentiate between treatment effect and disease progression? *Future Oncol*, 10(7):1277–97, 2014. doi: https://doi.org/10.2217/fon.13.271.

[WRP19] Martin Wistuba, Ambrish Rawat, and Tejaswini Pedapati. A Survey on Neural Architecture Search. *ArXiv*, abs/1905.01392, May 2019. doi: https://doi.org/10.48550/arXiv.1905.01392.

[WRP20] Yizhou Wan, Roushanak Rahmat, and Stephen J. Price. Deep Learning for Glioblastoma Segmentation Using Preoperative Magnetic Resonance Imaging Identifies Volumetric Features Associated with Survival. *Acta Neurochirurgica*, 162(12):3067–3080, 2020. doi: https://doi.org/10.1007/s00701-020-04483-7.

[WWZ20] Zhaobin Wang, E. Wang, and Ying Zhu. Image Segmentation Evaluation: A Survey of Methods. *Artif. Intell. Rev.*, 53(8):5637–5674, 2020. doi: https://doi.org/10.1007/s10462-020-09830-9.

[WZZ+13] Li Wan, Matthew Zeiler, Sixn Zhang, Yann Le Cun, and Rob Fergus. Regularization of Neural Networks Using DropConnect. *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 28(3):1058–1066, Jun 2013.

[WZZ+24] Hong Wei, Tianying Zheng, Xiaolan Zhang, Yuanan Wu, Yidi Chen, Chao Zheng, Difei Jiang, Botong Wu, Hua Guo, Hanyu Jiang, and Bin Song. MRI Radiomics Based on Deep Learning Automated Segmentation to Predict Early Recurrence of Hepatocellular Carcinoma. *Insights into Imaging*, 15(1):120, 2024. doi: https://doi.org/10.1186/s13244-024-01679-8.

[YBL+24] Wenjian Yao, Jiajun Bai, Wei Liao, Yuheng Chen, Mengjuan Liu, and Yao Xie. From CNN to Transformer: A Review of Medical Image Segmentation Models. *Journal of Imaging Informatics in Medicine*, 37(4):1529–1547, 2024. doi: https://doi.org/10.1007/s10278-024-00981-7.

[Yer47] Jacob Yerushalmy. Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-Ray Techniques. *Public Health Rep (1896)*, 62(40):1432–49, Oct 1947. PMID: 20340527.

[YMI+24] Yuki Yoshimi, Yuichi Mine, Shota Ito, Saori Takeda, Shota Okazaki, Takashi Nakamoto, Toshikazu Nagasaki, Naoya Kakimoto, Takeshi Murayama, and Kotaro Tanimoto. Image Preprocessing with Contrast-Limited Adaptive Histogram Equalization Improves the Segmentation Performance of Deep Learning for the Articular Disk of the Temporomandibular Joint on Magnetic Resonance Images. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, 138(1):128–141, 2024. doi: https://doi.org/10.1016/j.oooo.2023.01.016.

[YS20] Li Yang and Abdallah Shami. On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice. *Neurocomputing*, 415:295–316, 2020. doi: https://doi.org/10.1016/j.neucom.2020.07.061.

[YSSR22] Michael Yeung, Evis Sala, Carola-Bibiane Schönlieb, and Leonardo Rundo. Unified Focal Loss: Generalising Dice and Cross Entropy-Based Losses to Handle Class Imbalanced Medical Image Segmentation. *Computerized Medical Imaging and Graphics*, 95:102026, 2022. doi: https://doi.org/10.1016/j.compmedimag.2021.102026.

[YXZ+22] Suorong Yang, Wei-Ting Xiao, Mengcheng Zhang, Suhan Guo, Jian Zhao, and Shen Furao. Image Data Augmentation for Deep Learning: A Survey. *ArXiv*, abs/2204.08610, 2022. doi: https://doi.org/10.48550/arXiv.2204.08610.

[ZGJ23] Wenchao Zhang, Yu Guo, and Qiyu Jin. Radiomics and Its Feature Selection: A Review, Sep 2023. doi: https://doi.org/10.3390/sym15101834.

[ZKBMU22] 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. In Spyridon Bakas, Alessandro Crimi, Ujjwal Baid, Sylwia Malec, Monika Pytlarz, Bhakti Baheti, Maximilian Zenk, and Reuben Dorent, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 127–137. Springer Nature Switzerland, 2022. doi: https://doi.org/10.1007/978-3-031-33842-7_11.

[ZKMUB21] Ramy A. Zeineldin, Mohamed E. Karar, Franziska Mathis-Ullrich, and Oliver Burgert. Ensemble CNN Networks for GBM Tumors Segmentation Using Multi-parametric MRI. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and*

Traumatic Brain Injuries, pages 473–483. Springer International Publishing, 2021. doi: https://doi.org/10.1007/978-3-031-08999-2_41.

[ZLLW21] Yue Zhang, Shijie Liu, Chunlai Li, and Jianyu Wang. Rethinking the Dice Loss for Deep Learning Lesion Segmentation in Medical Images. *Journal of Shanghai Jiaotong University (Science)*, 26(1):93–102, 2021. doi: https://doi.org/10.1007/s12204-021-2264-x.

[ZQD+20] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, PP:1–34, 2020. doi: https://doi.org/10.1109/JPROC.2020.3004555.

[Zui94] Karel Zuiderveld. *Contrast Limited Adaptive Histogram Equalization. Graphics Gems IV.* Academic Press Professional, Inc., Aug 1994.

[ZVMB18] Chiyuan Zhang, Oriol Vinyals, Remi Munos, and Samy Bengio. A Study on Overfitting in Deep Reinforcement Learning. *ArXiv*, Apr 2018. doi: https://doi.org/10.48550/arXiv.1804.06893.

[ZYQ+23] Wang Zehao, Guo Yiwen, Li Qizhang, Yang Guanglei, and Zuo Wangmeng. DualAug: Exploiting Additional Heavy Augmentation with OOD Data Rejection. *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2310.08139.

[ZZG22] Ping Zheng, Xunfei Zhu, and Wenbo Guo. Brain Tumor Segmentation Based on an Improved U-Net. *BMC Medical Imaging*, 22(1):199, 2022. doi: https://doi.org/10.1186/s12880-022-00931-1.