2026-08

Comparative analysis of deep learning encoders for binary glioblastoma segmentation on post-treatment MRI

Sarker, Rudrodeb

Independent University, Bangladesh (IUB)

Comparative Analysis of Deep Learning Encoders for Binary Glioblastoma Segmentation on Post-Treatment MRI

August 2026

Prepared by:

Rudrodeb Sarker (ID: 2220985)

Moaz (ID: 2220368)

Department of Computer Science and Engineering

Independent University, Bangladesh

Supervised by:

Dr. Saadia Binte Alam

Associate Professor

Department of Computer Science and Engineering

Independent University, Bangladesh

A Final Year Design Project (FYDP) Report submitted for the Degree of Bachelor's of Computer Science & Engineering

Attestation

We are aware of the fact that plagiarism is strictly prohibited and it conflicts with our university's rules and regulations. We guarantee the authenticity of our work. We have used some copyrighted material and models of others in our project, which has been properly cited following international standards and proper guidelines.

Author Name:

Rudrodeb Sarker

. . . . . .

Signature:

. . . . . .

Author Name:

Moaz

. . . . . .

Signature:

. . . . . .

Evaluation Committee

Supervisor

Name: _____ Signature: _____

Internal Examiner 1

Name: _____ Signature: _____

Internal Examiner 2

Name: _____ Signature: _____

External Examiner

Name: _____ Signature: _____

Acknowledgement

We would like to express our sincere gratitude to all those who have contributed to the successful completion of this thesis.

First and foremost, we are deeply grateful to our thesis supervisor, Dr. Saadia Binte Alam, Associate Professor, Department of Computer Science and Engineering, Independent University, Bangladesh, for the invaluable guidance, continuous support, and constructive feedback throughout this research. We also thank our co-supervisor, Dr. Md Rashedur Rahman, Assistant Professor, Department of Computer Science and Engineering, Independent University, Bangladesh, for his thoughtful feedback and guidance. Their expertise in deep learning and medical image analysis has been instrumental in shaping the direction of this work.

We would also like to acknowledge the technical support provided by Mr. Fatin Israq, Mr. Siam Tahsin Bhuiyan, and Ms. Halima Khatun, whose contributions helped carry this project forward.

We would like to thank our families and friends for their unwavering support, patience, and encouragement throughout our academic journey. Their belief in our abilities has been a constant source of motivation.

Finally, we used Large Language Models (LLM), specifically Claude Opus 4.6 from Anthropic, and Gemini 3 Pro from Google, to improve the clarity and language of our writing. These tools were used only for editorial purposes, and we confirm that the method, hypothesis, data analysis, result generation, and figures were produced entirely by the authors.

List of Publications

1. **M. Moaz**, **R. Sarker**, R. Rahman, F. Israq, S. T. Bhuiyan, R. Islam, A. Islam, and S. B. Alam, "Systematic evaluation of encoder backbones and preprocessing strategies for post-treatment glioblastoma segmentation on T1-weighted MRI," in *Proc. 1st Int. Conf. Next-Generation Electrical & Electronics, Computer Systems, and Technologies (iCONEECT)*, Chittagong, Bangladesh, 2026. (Under Review)

Abstract

Glioblastoma (GBM), classified as World Health Organization (WHO) Grade IV, is the most common and aggressive primary malignant brain tumor, accounting for approximately 50.1% of all malignant central nervous system tumors. Despite multimodal treatment combining surgical resection, radiation therapy, and temozolomide chemotherapy, the median overall survival remains a dismal 12–15 months. Accurate tumor delineation in post-treatment magnetic resonance imaging (MRI) is critical for monitoring treatment response, distinguishing true progression from pseudoprogression or radiation necrosis, and guiding clinical decisions. However, manual delineation by neuroradiologists is time-consuming, subjective, and susceptible to significant inter- and intra-observer variability, underscoring the urgent need for reliable automated tools.

Our study presents a systematic evaluation of six encoder backbones within the U-Net architecture for binary whole-tumor segmentation on T1-weighted pre-contrast MRI from the University of California San Diego Post-Treatment Glioblastoma (UCSD-PTGBM) dataset, publicly available through The Cancer Imaging Archive (TCIA). The encoders are ResNet18, ResNet50, DenseNet121, MobileNetV2, Mix Transformer B2 (MiT-B2), and EfficientNet-B3, which span diverse architectural paradigms including residual learning, dense connectivity, depthwise separable convolutions, compound scaling, and vision transformers. Unlike prior studies that optimize for accuracy alone, this work jointly assesses three dimensions: segmentation performance (Dice coefficient and Intersection over Union), model complexity (parameter count), and computational efficiency (inference time), providing practical guidance for deployment in resource-constrained environments.

Beyond the encoder comparison, this study investigates the impact of three image preprocessing strategies: Contrast-Limited Adaptive Histogram Equalization (CLAHE), standard contrast enhancement (CE), and a novel RGB channel-stacking approach that combines the original, CLAHE-enhanced, and contrast-enhanced versions of a single MRI slice into three color channels. All models are trained using 5-fold patient-wise cross-validation with Dice loss and the Adam optimizer, and preprocessing-enhanced models are tested on unprocessed images to assess cross-domain robustness. Experimental results demonstrate that EfficientNet-B3 achieves the best overall performance with a Dice coefficient of 0.763, IoU of 0.618, a moderate parameter count of 13.16 million, and a clinically acceptable inference time of 119.60 milliseconds—outperforming both larger models (ResNet50 with 32.52M parameters, MiT-B2 with 27.48M parameters) and lightweight alternatives (MobileNetV2 with 6.63M parameters), thereby validating compound scaling for medical image analysis. The preprocessing analysis reveals that training on raw images yields superior results compared to all enhanced strategies, with up to 14.5% IoU degradation when preprocessing applied during training is absent at inference. These findings establish that EfficientNet-B3 with raw image training provides the most efficient and accurate pipeline for post-treatment glioblastoma delineation on T1-weighted MRI, eliminating preprocessing overhead while maintaining robust performance.

Contents

List of Figures

List of Tables

Chapter-1: Introduction

1.1 Background

Brain tumors represent one of the most devastating categories of neoplasms affecting the central nervous system (CNS), with profound implications for patient morbidity and mortality. Among this diverse spectrum, glioblastoma (GBM), classified as World Health Organization (WHO) Grade IV, stands as the most common and aggressive primary malignant brain tumor [3]. According to the Central Brain Tumor Registry of the United States (CBTRUS), GBM accounts for approximately 50.1% of all malignant CNS tumors, with a global incidence rate of 3–5 cases per 100,000 individuals [3]. The disease predominantly affects adults, with a median age at diagnosis of approximately 64 years, and exhibits a slight male predominance [4].

The standard of care for GBM follows the Stupp protocol, based on the results of a pivotal study conducted in 2005 [4]. The treatment program involves multimodality with maximum resection surgery, followed by radiation therapy and adjuvant temozolomide chemotherapy. Nevertheless, despite this aggressive strategy, the prognosis for the disease is extremely unfavorable; the median overall survival (OS) is 12-15 months, while 5-year survival does not exceed 5% [4]. Such a universal fatal prognosis is due to the invasive characteristics of GBM, as well as its molecular heterogeneity and development of resistance to therapy.

Magnetic resonance imaging (MRI) is essential for diagnosing GBM and subsequent management of the disease, including surgery planning, treatment evaluation, and response assessment [5]. Multiple MRI sequences are routinely acquired, including T1-weighted pre-contrast (T1pre), T1-weighted post-gadolinium contrast-enhanced (T1 CE), T2-weighted, and T2-weighted Fluid-Attenuated Inversion Recovery (FLAIR). All of these sequences provide information about different compartments of the tumor and the surrounding tissue environment. Among these, T1 pre-contrast images are important for obtaining anatomical information, which makes them available at all institutions and a standard method for brain tumors.

Accurate tumor segmentation, which refers to the demarcation of the boundaries of the tumors and the separation of neoplastic tissues from normal brain tissues, is a crucial component in the process. It is very important for measuring the tumor volume, assessing the progression of the disease, evaluating the response to treatment using the RANO criteria, and for surgical and radiotherapy planning. [5]. In the post-treatment setting, however, the task becomes substantially more challenging. Complex imaging appearances

arising from surgery, radiation, and chemotherapy—including pseudoprogression, radiation necrosis, surgical cavities, blood products, and treatment-related edema—can closely mimic true tumor progression on conventional MRI [6, 7], making reliable differentiation critical for appropriate decision-making.

Manual tumor delineation by experienced neuroradiologists is widely regarded as the gold standard for identifying tumor regions. However, the process is labor-intensive and can take approximately 30–60 minutes for a single scan [7]. The resulting annotations may also vary substantially across radiologists and, in some cases, between repeated assessments by the same radiologist. This issue is particularly evident when tumor boundaries are unclear or when post-treatment changes make it difficult to distinguish residual tumor from surrounding tissue. Such inconsistencies can affect the reliability of tumor volume measurements and may have implications for treatment planning and clinical decision-making. With the growing number of patients requiring imaging and the increasing need for repeated assessments over the course of treatment, the limitations of manual delineation have become more apparent. These challenges emphasize the need for automated methods that can provide accurate, consistent, and efficient tumor delineation.

1.2 Problem Statement

Despite substantial progress in deep learning-based tumor segmentation over the past decade, important challenges still hinder the adoption of these methods in clinical practice, particularly for post-treatment glioblastoma (GBM).

First, the majority of existing studies focus on pre-treatment or de novo tumors, where the tumor margins are generally easier to identify on contrast-enhanced MRI [5, 8]. In contrast, segmentation after treatment is considerably more challenging because therapy-related changes can resemble residual or recurrent tumor, making the two difficult to distinguish [6]. One major example is pseudoprogression, which develops in approximately 20–30% of patients during the first few months following chemoradiation and may appear as new or increased contrast enhancement that closely resembles true tumor progression [7]. Additional factors, including radiation necrosis, postoperative cavities, and imaging artifacts, further increase the ambiguity of post-treatment MRI. Consequently, segmentation models designed for this setting must be robust enough to account for these complex and often overlapping imaging characteristics.

Second, although numerous encoder backbone architectures have been proposed for natural image tasks, their comparative assessment when it comes to medical image segmentation, especially in the post-treatment brain tumor context, remains limited. Current literature mainly evaluates encoders based solely on their accuracy measures without any consideration for other constraints like size and speed of inference of the model [9]. Considering the constraints of the environment when it comes to resource-constrained settings, including developing nations, the efficiency of the model is just as important as its accuracy.

Third, the influence of image preprocessing strategies on segmentation quality, particularly when combined with transfer learning from ImageNet pre-trained encoders, has not been rigorously studied for post-treatment GBM. While techniques such as CLAHE and

contrast enhancement have shown benefits in other medical imaging applications [10], their effectiveness on post-treatment MRI data remains unexplored. Moreover, the approach of combining multiple preprocessing variants into RGB channels to enrich the feature space from a single modality has not been compared against standard single-channel training.

Finally, the risk of domain mismatch—where preprocessing applied during training is absent during inference—has not been examined in this context. Understanding how models trained with specific preprocessing strategies behave when tested on unprocessed images is crucial for practical deployment, where maintaining complex preprocessing pipelines may not always be feasible.

1.3 Motivation

This research is driven by the convergence of pressing needs in both neuro-oncology and medical image analysis.

From a clinical standpoint, the rising incidence of GBM, combined with growing reliance on serial MRI monitoring for response assessment, has created an unsustainable demand for manual tumor delineation by neuroradiologists. Automated tools that provide accurate, reproducible, and efficient boundary extraction have the potential to reduce radiologist workload, minimize inter-observer variability, and enable more consistent and timely follow-up evaluations.

From a technical standpoint, the rapid proliferation of deep learning architectures has created a paradox of choice for practitioners. With dozens of encoder backbones available—each embodying different design philosophies, parameter counts, and computational demands—selecting the most suitable architecture for a given application is a non-trivial decision. Larger and more complex models may yield marginal accuracy gains yet prove impractical in environments with limited computational resources or real-time processing requirements. A comparative analysis that provides clear guidance on the accuracy–efficiency trade-off is therefore of considerable practical value.

Moreover, post-treatment GBM segmentation has received markedly less attention than its pre-treatment counterpart, despite being equally—if not more—important in practice. Because monitoring continues throughout a patient's disease course, requiring repeated delineation at every follow-up imaging timepoint, efficient automated tools validated specifically for post-treatment data constitute a critical unmet need.

Finally, the availability of the UCSD-PTGBM dataset [11, 12], a high-quality, publicly available collection of post-treatment GBM scans with expert annotations, presents a unique opportunity to conduct rigorous and reproducible research on this problem. Leveraging this resource to establish performance benchmarks and identify optimal architectures can accelerate both future investigation and translation into practice.

1.4 Objectives

This study pursues four interrelated research aims:

1. **Multi-Dimensional Encoder Benchmarking:** To compare six encoder backbone architectures—ResNet18, ResNet50, DenseNet121, MobileNetV2, MiT-B2, and EfficientNet-B3—within the U-Net framework for binary whole-tumor segmentation on post-treatment GBM T1-weighted pre-contrast MRI, evaluating each across three dimensions: segmentation accuracy (Dice coefficient and IoU), model complexity (parameter count), and computational cost (inference time).
2. **Preprocessing Impact Investigation:** To examine how three image preprocessing strategies—CLAHE, contrast enhancement (CE), and RGB channel stacking—affect segmentation quality relative to training on raw images, and to determine whether multi-representation inputs offer an advantage over single-channel training.
3. **Domain Robustness Evaluation:** To measure how preprocessing-trained models degrade when inference is performed on unprocessed images, thereby quantifying the domain mismatch risk inherent in preprocessing-dependent pipelines.
4. **Deployment Suitability Analysis:** To synthesize findings across accuracy, efficiency, model size, and robustness to identify the most suitable encoder–training configuration for integration into post-treatment monitoring workflows.

1.5 Key Contributions

This thesis delivers the following novel contributions to the field:

- **Comprehensive and efficiency-oriented encoder evaluation for post-treatment GBM:** A benchmarking study on the UCSD-PTGBM dataset comparing six architectures—spanning residual, dense, lightweight, compound-scaled, and transformer-based paradigms—across accuracy, model size, and inference time. Unlike prior work that reports accuracy alone, this study jointly profiles all three dimensions to guide architecture selection under real-world constraints.
- **Investigate that multi-representation RGB stacking does not outperform single-channel training:** An empirical finding showing that combining original, CLAHE, and contrast-enhanced images into an RGB input fails to improve Dice or IoU over raw image training on post-treatment MRI, while isolated contrast enhancement yields the best cross-domain transfer among the preprocessing strategies tested.
- **Quantification of domain mismatch degradation:** A robustness analysis demonstrating that preprocessing shifts the learned feature distribution, producing up to 14.5% IoU degradation when the preprocessing used at training time is omitted during inference—a finding with direct implications for pipeline reliability in practice.

- **Identification of the optimal encoder–training configuration:** EfficientNet-B3 trained on raw images achieves the highest Dice coefficient (0.763) and IoU (0.618) with a moderate footprint (13.16M parameters) and clinically acceptable inference latency (119.60 ms), outperforming both larger and smaller alternatives and eliminating preprocessing dependency.

1.6 Thesis Organization

The remainder of this thesis is organized as follows:

Chapter 2 presents a comprehensive review of the related literature, covering brain tumor segmentation methods from traditional approaches to modern deep learning architectures, the evolution of encoder backbone designs, transfer learning in medical image analysis, preprocessing strategies for MRI, the BraTS challenge benchmarks, and the specific challenges of post-treatment glioblastoma segmentation.

Chapter 3 describes the UCSD-PTGBM dataset used in this study, including its demographics, imaging protocol, annotation standards, and the data selection and partitioning strategies employed.

Chapter 4 details the proposed methodology, encompassing the overall pipeline, preprocessing strategies, model architectures, training protocol, and evaluation framework.

Chapter 5 presents the experimental results and analysis, covering the encoder backbone comparison, preprocessing strategy evaluation, segmentation visualization, and comparison with existing literature.

Chapter 6 provides an in-depth discussion of the findings, interpreting the results in the context of architectural design principles, clinical implications, and comparisons with related work.

Chapter 7 concludes the thesis with a summary of findings, acknowledgement of limitations, directions for future research, analysis of the social Effects, environmental impact analysis and sustainability and ethical issues.

Chapter-2: Literature Review

This chapter reviews the existing body of research on brain tumor segmentation in MRI, tracing the evolution from traditional image processing techniques through classical machine learning methods to modern deep learning architectures. The review is organized to first survey prior studies across three methodological paradigms, then identify the key challenges that persist in this domain, and finally articulate the specific research gaps that this thesis aims to address.

2.1 Prior Studies

Research in brain tumor segmentation has progressed through three distinct methodological eras. Early approaches relied on manual delineation and hand-crafted image processing rules, followed by classical machine learning classifiers that introduced data-driven decision boundaries, and culminating in the deep learning revolution that enabled end-to-end learned representations from raw image data. The following subsections survey representative and influential works from each era, tracing the evolution in accuracy, automation, and clinical applicability.

2.1.1 Brain Tumor Segmentation

Accurate segmentation of brain tumors from MRI is a fundamental requirement for diagnosis, surgical planning, treatment monitoring, and response assessment. The complexity of this task arises from the inherent heterogeneity of tumor appearance, the variability across patients and imaging protocols, and the ambiguous boundaries between neoplastic and healthy tissue—challenges that are further compounded in the post-treatment setting. This subsection surveys the evolution of brain tumor segmentation methods across three major paradigms.

2.1.1.1 Traditional Methods

Before the advent of machine learning, tumor boundary extraction relied on manual delineation by expert neuroradiologists and classical image processing techniques. Manual segmentation remains the clinical gold standard, yet the process is extremely time-

consuming, typically requiring 30–60 minutes per scan, and is subject to considerable inter-observer and intra-observer variability [5, 7]. Even experienced radiologists can produce substantially different delineations of the same tumor, particularly in regions with ambiguous boundaries or complex post-treatment appearances, introducing uncertainty into volumetric measurements and potentially affecting treatment decisions. Studies have shown that inter-rater agreement for glioma delineation can vary by as much as 20–30% in terms of the Dice coefficient depending on the tumor sub-region being annotated [5], underscoring that the variability is not simply random noise but reflects genuine diagnostic uncertainty at diffuse tumor margins. These inherent limitations of manual approaches motivated the development of semi-automatic and fully automatic methods.

A comprehensive survey of MRI brain tumor segmentation methods was conducted by Gordillo et al. [13], who systematically catalogued the classical techniques available at the time, categorizing them into region-based, boundary-based, and classification-based paradigms, and identifying the key challenges including intensity heterogeneity, poorly defined boundaries, and the wide variability in tumor appearance across patients. The survey established that no single classical method could reliably handle all tumor types and imaging conditions, motivating continued development of more robust approaches. Published in *Magnetic Resonance Imaging* in 2013, this work provided an important baseline assessment of the state of the art prior to the deep learning revolution. **Thresholding-based methods** represent the simplest class of automatic segmentation techniques, identifying pixels whose intensity values fall within a predefined range. Global thresholding applies a single intensity cutoff to the entire image, while adaptive variants such as Otsu's method [14] compute optimal thresholds by maximizing inter-class variance. While computationally simple, thresholding is highly sensitive to intensity variations across different MRI scanners and acquisition protocols, and struggles with the heterogeneous intensity distributions characteristic of glioblastoma [15]. Brain tumors frequently exhibit overlapping intensity ranges with healthy gray and white matter on T1-weighted sequences, meaning that no single threshold can reliably separate tumor from non-tumor tissue. Its reliance on fixed intensity ranges limits adaptability to the variable tissue appearances encountered in post-treatment imaging.

Region growing algorithms start from a user-defined seed point and iteratively include neighboring pixels that satisfy certain homogeneity criteria, such as intensity similarity or gradient magnitude constraints. Although more adaptive than thresholding, region growing requires manual seed selection and is prone to leaking into adjacent tissues with similar intensity values—a limitation particularly problematic in the post-treatment setting, where tissue boundaries are often ambiguous [5]. The dependence on user initialization also hinders reproducibility and throughput. Variants employing multiple seed points or incorporating confidence-connected criteria improved robustness but could not fully overcome the fundamental sensitivity to initialization and tissue homogeneity assumptions. **Watershed segmentation** treats the gradient magnitude image as a topographic surface and identifies catchment basins separated by watershed lines that correspond to object boundaries. While effective at detecting local intensity boundaries, watershed segmentation is notoriously prone to over-segmentation, producing far more regions than actual anatomical structures [13]. Marker-controlled variants mitigate this issue by constraining the segmentation to user-defined or automatically detected markers, but the quality of results remains critically dependent on marker placement and gradient quality. In brain tumor applications, the low contrast between tumor tissue and surrounding edema on T1-weighted pre-contrast images frequently produces weak gradients that lead to inadequate

boundary detection.

Morphological operations—including erosion, dilation, opening, and closing—were often employed as pre- or post-processing steps to refine initial segmentations produced by other methods. Morphological reconstruction and hit-or-miss transforms were used to remove small spurious regions, fill holes within tumor masks, and smooth irregular boundaries. While effective at reducing noise-induced false positives, morphological operations cannot recover information that was lost during the initial segmentation step, and their structuring element size requires careful tuning for different tumor sizes and shapes.

Atlas-based methods use a reference brain atlas to identify deviations from normal anatomy, assuming that tumor regions will deviate significantly from the expected tissue distribution. However, these approaches require accurate registration between the patient image and the atlas, which becomes challenging in the presence of large tumors that distort normal anatomy or surgical cavities that alter the expected brain structure [15]. Post-treatment changes further compromise registration accuracy, limiting clinical utility in follow-up imaging. The computational cost of deformable registration is also substantial, particularly for volumetric data, making atlas-based methods slow compared to simpler alternatives.

Active contour (snake) models and **level set methods** define an evolving curve or surface driven toward tumor boundaries by image-based energy functions. While capable of producing smooth and topologically consistent delineations, these methods are sensitive to initialization, can become trapped in local minima, and often require manual parameter tuning for different patients and tumor types. Their reliance on gradient-based edge detection makes them particularly vulnerable to the diffuse and irregular boundaries characteristic of infiltrative gliomas. Geodesic active contour models incorporated edge-based stopping functions that halted curve evolution at strong intensity gradients, but the absence of strong gradients at infiltrative tumor boundaries frequently caused the contour to pass through the true boundary and converge on nearby anatomical structures instead. **Markov Random Fields (MRFs)** provided a probabilistic framework for incorporating spatial context into pixel-wise classification by modeling the joint probability of neighboring pixel labels. MRF-based methods encode the prior assumption that neighboring pixels are likely to share the same label, discouraging isolated misclassifications and producing spatially coherent segmentations. The energy minimization was typically performed using iterative algorithms such as iterated conditional modes (ICM) or graph cuts. While MRFs improved spatial consistency compared to purely pixel-wise methods, the optimization was computationally expensive, and the pairwise potentials often over-smoothed fine-grained boundary details that are critical for accurate tumor delineation [13].

Overall, traditional methods provided the initial tools for tumor boundary extraction, but their reliance on hand-crafted features, sensitivity to parameter settings, requirement for manual initialization, and inability to capture complex spatial and textural patterns limited both their accuracy and clinical applicability [5, 15, 13]. Gordillo et al. [13] reported that traditional methods typically achieved Dice coefficients in the range of 0.60–0.75 for whole tumor segmentation on small single-institution datasets, with sensitivity values between 0.65 and 0.80 and specificity above 0.90 owing to the dominance of background voxels. However, these results were obtained under controlled conditions with carefully selected cases, and performance degraded substantially on more heterogeneous datasets with diverse tumor types and imaging protocols. The survey concluded that none of

these methods individually achieved sufficient accuracy for clinical deployment, and that combining multiple techniques in hybrid pipelines offered only incremental improvements. These limitations motivated the shift toward data-driven approaches capable of learning discriminative representations directly from annotated training data.

2.1.1.2 Machine Learning-Based Methods

The introduction of machine learning marked a significant advancement in brain tumor segmentation, replacing fixed rules with models that learn decision boundaries from annotated training data. These methods offered improved accuracy over traditional approaches by leveraging statistical learning from examples rather than relying on manually specified parameters. The typical pipeline involved two distinct stages: feature extraction, where handcrafted descriptors were computed from the MRI volumes, followed by classification, where a supervised learning algorithm used these features to assign labels to individual voxels or image patches.

The **feature engineering stage** was central to the success of machine learning approaches and typically involved computing a rich set of descriptors at each voxel. Common features included first-order intensity statistics (mean, variance, skewness, kurtosis) computed within local neighborhoods, texture descriptors derived from Gabor filter banks and gray-level co-occurrence matrices (GLCM), gradient magnitude and orientation histograms, spatial location coordinates relative to brain landmarks, and multi-scale features extracted at different Gaussian pyramid levels. Some approaches also incorporated symmetry-based features that compared corresponding locations in the left and right brain hemispheres, exploiting the observation that tumors typically produce asymmetric intensity patterns. The quality and diversity of these handcrafted features directly determined the ceiling of achievable segmentation performance, creating a labor-intensive bottleneck that required significant domain expertise to design effectively.

Random forests became one of the most successful pre-deep-learning approaches for brain tumor segmentation. These ensemble methods combine multiple decision trees, each trained on random subsets of features and data, to produce robust voxel-wise predictions. Within the context of the BraTS Challenge, random forest classifiers trained on multi-modal MRI features formed the backbone of several top-performing submissions in the early years of the competition [5]. Notably, Tustison et al. [16] won the BraTS 2013 challenge using a random forest framework with optimally-tuned features extracted from multi-modal MRI, achieving Dice scores of 0.87 for complete tumor, 0.78 for tumor core, and 0.74 for enhancing tumor, along with sensitivity of 0.86 and positive predictive value (precision) of 0.88 for whole tumor on the challenge test set. The strength of random forests lay in their ability to handle high-dimensional feature spaces without overfitting, their inherent feature importance ranking that provided interpretability, and their relatively fast training and inference compared to other ensemble methods. However, performance remained fundamentally constrained by the expressiveness of manually engineered features, which could not capture the full complexity of tumor appearance variations.

Support Vector Machines (SVMs) were also applied to brain tumor segmentation, using kernel functions to project feature vectors into high-dimensional spaces where linear separation between tumor and non-tumor classes becomes possible. The radial

basis function (RBF) kernel was most commonly used, as it could model non-linear decision boundaries between tumor and non-tumor tissue classes. In the BraTS 2013 challenge, SVM-based submissions achieved Dice scores in the range of 0.83–0.85 for whole tumor, with sensitivity of 0.80–0.88 and specificity above 0.95 [5]. SVMs demonstrated competitive accuracy when combined with carefully engineered feature sets, particularly for binary voxel-wise classification tasks. However, their computational cost scales poorly with dataset size—the training complexity of standard SVMs is $O(n^2)$ to $O(n^3)$ in the number of training samples—limiting applicability to large-scale clinical datasets and high-resolution volumetric MRI data where millions of voxels require classification.

k-Nearest Neighbors (k-NN) classifiers assigned each voxel the majority class label among its k nearest neighbors in the feature space. While conceptually simple and requiring no explicit training phase, k-NN's computational cost during inference scales linearly with the size of the training set, making it impractical for large volumetric datasets. Furthermore, k-NN is highly sensitive to the choice of distance metric and the curse of dimensionality, where performance degrades as the number of features increases and the feature space becomes sparse. **Conditional Random Fields (CRFs)** were employed as post-processing modules or combined with other classifiers to enforce spatial consistency in the segmentation output. By modeling spatial dependencies between neighboring voxels through pairwise potentials that penalize label disagreements between adjacent voxels, CRFs help reduce isolated false positive predictions and produce more anatomically plausible results. Fully connected CRFs extended this concept by considering pairwise relationships between all voxels in the image, not just immediate neighbors, enabling long-range label consistency. CRFs would later prove valuable as post-processing modules for deep learning methods as well, as demonstrated by Kamnitsas et al. [17], where the integration of a fully connected CRF improved Dice scores by 0.3–0.5 percentage points on the BraTS 2015 test set. **Ensemble and cascaded approaches** combined multiple classifiers or sequential processing stages to improve segmentation robustness. Cascaded pipelines typically used a coarse initial segmentation to identify candidate tumor regions, followed by a refined second-stage classifier operating only within the candidate regions. This two-stage strategy reduced class imbalance by excluding the majority of background voxels before the final classification, and improved boundary precision by allowing the second stage to focus computational resources on ambiguous regions. Bagging and boosting strategies were also applied to combine predictions from multiple classifiers trained on different feature subsets or data partitions.

While machine learning methods offered improved accuracy over traditional approaches, they remained limited by the quality and completeness of handcrafted features. The manual feature engineering process was labor-intensive, domain-specific, and inherently incomplete—unable to represent the full complexity of tumor appearance across different grades, locations, treatment histories, and imaging protocols. Features that were discriminative for one tumor type or MRI protocol often failed to generalize to others, and there was no systematic way to determine the optimal feature set for a given task without extensive experimentation. A comprehensive survey by Litjens et al. [15], published in *Medical Image Analysis* in 2017, documented this transition and identified the shift to deep learning as the most significant methodological change in medical image analysis, as deep networks automatically discover hierarchical feature representations directly from raw image data, eliminating the bottleneck of manual feature design.

2.1.1.3 Deep Learning-Based Methods

The application of deep learning to medical image analysis, beginning around 2013–2014, represented a paradigm shift in brain tumor segmentation. These methods learn feature hierarchies directly from raw data, eliminating manual feature engineering and enabling the capture of complex spatial patterns at multiple scales. This subsection reviews the key deep learning contributions that have shaped the field, organized from foundational architectures through recent innovations.

Fully Convolutional Networks (FCNs), proposed by Long et al. [18], established the broader foundation for dense prediction tasks. By replacing the fully connected layers of classification networks with convolutional layers, FCNs enabled end-to-end dense prediction, processing the entire image in a single forward pass to produce a segmentation map of the same spatial dimensions as the input. The introduction of skip connections—combining coarse semantic information from deeper layers with fine-grained spatial detail from shallower layers—was a key innovation. Multi-scale feature fusion variants (FCN-32s, FCN-16s, FCN-8s) demonstrated progressively improved boundary precision. Published at IEEE CVPR in 2015, FCN's conceptual contributions established the encoder-decoder paradigm that remains foundational to all subsequent segmentation architectures. Building directly on this foundation, Ronneberger et al. [1] proposed **U-Net** at MICCAI 2015, introducing a symmetric encoder-decoder architecture with concatenation-based skip connections specifically designed for biomedical image segmentation. The architecture consists of an encoder (contracting path) that progressively reduces spatial resolution while increasing feature channels, a decoder (expanding path) that progressively recovers spatial resolution, and skip connections that concatenate feature maps from each encoder stage directly to the corresponding decoder stage. A critical contribution was demonstrating that data augmentation combined with skip connections could yield excellent performance even with very small annotated datasets—a property of particular importance in medical imaging where expert annotations are scarce and expensive. U-Net won the ISBI 2015 cell tracking challenge and rapidly became the de facto standard for biomedical image segmentation. Its influence on brain tumor segmentation has been profound, with subsequent research focusing largely on improving U-Net through better encoders, attention mechanisms, and architectural modifications.

The extension of U-Net to volumetric data was addressed by Çiçek et al. [19], who proposed **3D U-Net** at MICCAI 2016. By replacing all 2D operations with their 3D counterparts, the architecture could process volumetric MRI data directly, capturing inter-slice spatial context that is lost when processing individual 2D slices independently. Importantly, 3D U-Net was designed to learn from sparsely annotated volumes, where only a subset of slices carry ground truth labels, making it practical for settings where complete volumetric annotation is prohibitively expensive. Concurrently, Milletari et al. [20] proposed **V-Net** at IEEE 3DV 2016, another 3D fully convolutional architecture that introduced the **Dice loss function**—a training objective derived directly from the Dice Similarity Coefficient. By optimizing the Dice coefficient during training rather than cross-entropy, V-Net addressed the severe class imbalance inherent in medical image segmentation, where the foreground (tumor) region typically occupies a small fraction of the image. Dice loss has since become the standard training objective for brain tumor segmentation, including in this thesis. Among the first to specifically design deep architectures for brain tumor segmentation, Havaei et al. [21] proposed a **two-pathway convolutional neural network** that processed input patches through parallel local and

global pathways, capturing both fine-grained boundary details and broader contextual information simultaneously, published in *Medical Image Analysis* in 2017. The local pathway used small receptive fields to capture detailed local patterns, while the global pathway employed larger receptive fields to incorporate surrounding anatomical context. A two-phase cascaded training strategy was introduced where the output of a first-pass network was used to provide additional input features to a second-pass network, enabling iterative refinement of predictions. Evaluated on the BraTS 2013 challenge dataset, the method achieved Dice scores of 0.88 for complete tumor, 0.79 for tumor core, and 0.73 for enhancing tumor, with corresponding sensitivity (recall) of 0.89, 0.79, and 0.73, and positive predictive value (precision) of 0.87, 0.79, and 0.73 on the test set, establishing a strong benchmark for deep learning-based brain tumor segmentation.

Pereira et al. [22] investigated the use of **small** 3×3 **convolutional kernels** throughout the network architecture for brain tumor segmentation, published in *IEEE Transactions on Medical Imaging* in 2016. Inspired by VGGNet's design philosophy, they demonstrated that stacking multiple small-kernel layers achieves the same effective receptive field as larger kernels while providing deeper feature hierarchies and requiring fewer parameters. Intensity normalization was applied to each MRI modality independently, and data augmentation through rotation was used to address the limited training set size. Evaluated on the BraTS 2013 challenge data, the method achieved Dice scores of 0.88 for complete tumor, 0.83 for core tumor, and 0.77 for enhancing tumor, with positive predictive values (precision) of 0.88, 0.87, and 0.85, and sensitivity (recall) of 0.89, 0.83, and 0.72 respectively, demonstrating that careful architectural design with small kernels could match or exceed the performance of more complex multi-pathway approaches. Kamnitsas et al. [17] proposed **DeepMedic**, an efficient multi-scale 3D CNN architecture for brain lesion segmentation, published in *Medical Image Analysis* in 2017. The architecture employed parallel processing pathways operating at different input resolutions—a high-resolution pathway for local detail and a low-resolution pathway for broader context—enabling efficient multi-scale feature extraction without the computational burden of processing large 3D receptive fields at full resolution. A key contribution was the integration of a fully connected CRF as a post-processing step to enforce spatial consistency and remove anatomically implausible false positive predictions. Evaluated on the BraTS 2015 challenge data, DeepMedic with CRF post-processing achieved Dice scores of 0.901 for whole tumor, 0.755 for tumor core, and 0.728 for enhancing tumor, with sensitivity of 0.892, 0.755, and 0.718 and precision of 0.912, 0.783, and 0.773 on the test set. The architecture demonstrated strong performance with moderate computational requirements and was subsequently adopted as a competitive baseline in numerous brain tumor segmentation studies.

Recognizing that standard skip connections in U-Net transmit all encoder features indiscriminately, including potentially irrelevant or noisy activations, Oktay et al. [23] proposed **Attention U-Net** at MIDL 2018. This architecture introduced attention gate modules at each skip connection that learn to selectively weight encoder feature maps based on their relevance to the current decoding stage. The attention gates produce spatial attention coefficients that suppress feature responses in irrelevant background regions while highlighting salient features in the regions of interest, enabling the decoder to focus computational resources on the most informative features. This mechanism is particularly beneficial for segmenting structures with variable sizes and shapes, where the relevant spatial locations change significantly across different scales. Attention U-Net demonstrated improved segmentation accuracy on abdominal organ segmentation tasks and has since been widely adopted in brain tumor segmentation pipelines.

Zhou et al. [24] proposed **UNet++**, which redesigned the skip connection topology through nested, dense skip pathways, published in *IEEE Transactions on Medical Imaging* in 2020. Rather than direct connections between encoder and decoder at corresponding levels, UNet++ introduces a series of nested dense convolutional blocks between them, progressively bridging the semantic gap between encoder and decoder feature maps. The dense skip pathways aggregate features at multiple semantic levels before passing them to the decoder, reducing the learning difficulty at each decoder stage. Deep supervision was applied to intermediate outputs at different semantic levels, enabling the model to be pruned at inference time for improved efficiency. UNet++ demonstrated consistent improvements over standard U-Net across multiple medical image segmentation benchmarks, confirming that bridging the semantic gap in skip connections is an effective strategy for improving segmentation precision.

The **Brain Tumor Segmentation (BraTS) Challenge**, organized annually since 2012, has been instrumental in driving algorithmic innovation and establishing standardized evaluation benchmarks for the field [5]. The challenge provides large-scale, multi-institutional MRI datasets with expert annotations and evaluates algorithms using consistent metrics including Dice coefficient, Hausdorff distance, sensitivity, and specificity across three tumor sub-regions: enhancing tumor (ET), tumor core (TC), and whole tumor (WT). The **BraTS 2021** benchmark [8] represented a significant scale-up, utilizing 2,000 pre-operative multi-parametric MRI scans (T1, T1-CE, T2, FLAIR) from multiple institutions. Winning solutions typically achieved Dice scores above 0.90 for whole tumor, with performance varying considerably across sub-regions. Most recently, the **BraTS 2024 Post-Treatment Challenge** [25] extended the scope to post-treatment GBM, introducing sub-region annotations including resection cavities and reflecting the growing recognition of the importance of follow-up assessment. Notably, both BraTS 2021 and BraTS 2024 relied on multi-parametric input requiring all four MRI modalities, and top-performing methods universally depended on this multi-modal availability—a constraint that limits applicability when only a subset of sequences is available in practice.

Myronenko [26] won the **BraTS 2018 Challenge** with a 3D encoder-decoder architecture incorporating variational autoencoder (VAE) regularization, published in the BrainLes Workshop at MICCAI 2019. The approach used a large asymmetric encoder-decoder with a shared encoder for both the segmentation and reconstruction tasks. The VAE branch reconstructed the input image from the bottleneck representation, serving as a regularization mechanism that forced the encoder to learn more generalizable features by preventing overfitting to the segmentation-specific training signal. The model was trained with a combination of Dice loss for the segmentation branch and L_2 reconstruction loss plus KL divergence for the VAE branch. On the BraTS 2018 validation dataset, the method achieved Dice scores of 0.9084 for whole tumor, 0.8597 for tumor core, and 0.8154 for enhancing tumor, with Hausdorff95 distances of 4.48 mm, 8.28 mm, and 3.81 mm respectively, and sensitivity of 0.9206, 0.8749, and 0.8487 across the three sub-regions. These results secured first place in the competition and established a strong benchmark for subsequent work. Isensee et al. [27] introduced **nnU-Net**, a self-configuring framework for deep learning-based biomedical image segmentation, published in *Nature Methods* in 2021. Rather than proposing a novel architecture, nnU-Net systematically automates the key design decisions in the segmentation pipeline—including preprocessing, network architecture (2D, 3D, or cascaded), training configuration, and post-processing—based on dataset fingerprinting and a set of heuristic rules derived from extensive empirical analysis. When applied to brain tumor segmentation on BraTS data, nnU-Net achieved

performance competitive with or exceeding specialized architectures, despite using no task-specific architectural innovations. Across 23 diverse biomedical segmentation datasets, nnU-Net demonstrated consistent top-tier performance, establishing that systematic pipeline configuration is at least as important as architectural innovation. However, nnU-Net's self-configuring pipeline does not allow systematic comparison of individual encoder architectures, as the architecture is determined automatically—a limitation that motivates the controlled encoder comparison conducted in this thesis.

The introduction of transformer architectures to computer vision opened new possibilities for capturing long-range spatial dependencies that purely convolutional architectures may miss due to their limited receptive fields. Wang et al. [28] proposed **TransBTS** at MICCAI 2021, which embedded a transformer module within a 3D CNN encoder-decoder framework for multimodal brain tumor segmentation. The 3D CNN encoder first extracted local volumetric features, which were then processed by transformer layers to capture global context through self-attention, before being decoded back to full resolution. By combining the local feature extraction strengths of CNNs with the global context modeling of transformers, TransBTS achieved competitive performance on BraTS benchmarks, demonstrating the complementary benefits of these two paradigms. Building on the success of Swin Transformer in natural image tasks, Hatamizadeh et al. [29] proposed **Swin UNETR** in the BrainLes Workshop at MICCAI 2022. This architecture employed a Swin Transformer encoder with shifted window self-attention to efficiently model both local and global spatial relationships, connected to a CNN-based decoder through skip connections. The hierarchical shifted window approach reduced the quadratic complexity of standard self-attention to linear complexity while maintaining the ability to capture cross-window interactions. On the BraTS 2021 validation dataset, Swin UNETR achieved Dice scores of 0.928 for whole tumor, 0.890 for tumor core, and 0.823 for enhancing tumor, with Hausdorff95 distances of 5.37 mm, 7.14 mm, and 12.92 mm respectively, establishing transformer-based architectures as competitive alternatives to purely convolutional approaches for brain tumor segmentation.

Zeineldin et al. [30] introduced **TransXAI**, an explainable AI architecture integrating CNNs and Vision Transformers for multimodal glioma segmentation, published in *Scientific Reports* in 2024. Their approach combined convolutional feature extraction with transformer-based global context modeling while providing interpretable visualizations of the model's decision process through attention maps and gradient-based explanations. The explainability component addressed a growing concern in clinical AI deployment—the need for clinicians to understand and trust model predictions. However, TransXAI relied on multi-parametric MRI input (all four BraTS modalities) and did not evaluate computational efficiency metrics such as model size or inference time, leaving open the question of practical deployability in resource-constrained settings.

Zhang et al. [31] proposed **DeepGlioSeg**, a framework combining CNN and Transformer branches with region-based weighting for advanced glioma MRI data, published in *Frontiers in Oncology* in 2025. The architecture employed parallel CNN and Transformer pathways that extracted complementary local and global features, which were then fused using a region-based weighting strategy that assigned different importance to different spatial regions based on their proximity to tumor boundaries. This strategy improved sub-region delineation accuracy, particularly for small and irregular enhancing tumor regions. However, their evaluation was limited to pre-treatment cases with multi-modal input, leaving the performance on the more challenging post-treatment scans unexplored.

The effectiveness of pre-trained encoder backbones for brain tumor segmentation was demonstrated by Naser and Deen [32], who employed a **U-Net with VGG16 encoder** for simultaneous glioma segmentation and grading, published in *Computers in Biology and Medicine* in 2020. Using T1 and FLAIR-weighted MRI as input, their model achieved a mean Dice Similarity Coefficient (DSC) of 0.84 for tumor segmentation, with reported sensitivity of 0.82 and specificity of 0.96. Their work validated the effectiveness of ImageNet pre-trained encoders for medical imaging tasks, demonstrating that features learned from natural images transfer meaningfully to the medical domain. However, the study used only a single encoder architecture (VGG16) without comparing alternatives, and did not assess inference speed or model compactness—limiting the practical guidance available for architecture selection. Wu et al. [33] proposed a dual-branch decoder architecture for simultaneous brain tumor segmentation and grading, published in *Scientific Reports* in 2025. The shared encoder extracted features used by two specialized decoder branches, one for pixel-wise segmentation and one for grade classification, enabling multi-task learning that leveraged the synergy between segmentation and grading objectives. Sun et al. [34] employed an ensemble of three different 3D CNN architectures with majority voting for brain tumor segmentation and survival prediction using multimodal MRI scans, published in *Frontiers in Neuroscience* in 2019. Their ensemble strategy combined complementary architectural strengths to improve robustness, and the survival prediction component demonstrated the clinical utility of segmentation-derived volumetric features. Both studies focused exclusively on pre-treatment data with multi-modal input, and neither examined how their architectures perform on the more challenging post-treatment scans. Kumari et al. [10] proposed an enhanced glioma tumor detection and segmentation framework combining modified deep learning with edge fusion and frequency features, published in *Scientific Reports* in 2025. Their approach integrated edge-enhanced representations and frequency domain features extracted through wavelet transforms into the segmentation pipeline, enriching the feature space beyond what standard spatial convolutions alone could capture. The edge fusion component specifically targeted the improvement of boundary delineation, which is a persistent challenge in glioma segmentation due to the infiltrative nature of these tumors.

Several studies have specifically addressed the challenges of **post-treatment brain tumor segmentation**. Gagnon et al. [6] demonstrated deep learning-based segmentation of infiltrative and enhancing cellular tumor at pre- and post-treatment multishell diffusion MRI of glioblastoma, published in *Radiology: Artificial Intelligence* in 2024. Their work was among the first to specifically evaluate deep learning segmentation performance on post-treatment GBM imaging, revealing the substantial performance degradation that occurs when models encounter treatment-induced imaging changes. Helland et al. [7] investigated the segmentation of glioblastomas in early post-operative multi-modal MRI with deep neural networks, published in *Scientific Reports* in 2023. Their study demonstrated the feasibility of automated segmentation in the immediate post-operative setting, though the complex imaging appearances arising from surgery, blood products, and early treatment effects posed significant challenges for accurate delineation. Hannisdal et al. [35] demonstrated the feasibility of deep learning-based delineation for radiotherapy planning in grade-4 glioma using the nnU-Net framework on multi-parametric MRI, published in *Neuro-Oncology Advances* in 2023. The study achieved median Dice scores of 0.81–0.82 with median Hausdorff95 distances of 10.1–12.7 mm, validating that automated delineation can reach clinically acceptable accuracy for treatment planning while also revealing the remaining boundary imprecision through the Hausdorff metric. However, the nnU-Net's self-configuring pipeline did not allow systematic comparison of individual

encoder architectures, and the evaluation relied on multi-parametric input. Gkournelos et al. [36] studied how BraTS-trained models generalize to clinical data with missing MRI sequences, using multi-class segmentation on multi-parametric input, published in *Scientific Reports* in 2023. They found considerable performance variation when individual modalities were absent, highlighting the importance of evaluating models under realistic deployment conditions where not all MRI sequences may be available.

Table 2.1 summarizes the key deep learning-based studies discussed in this review, presenting their methods, datasets, quantitative results, and publication venues.

Table 2.1: Summary of Key Deep Learning Studies in Brain Tumor Segmentation

Study	**Method**	**Dataset**	**Key Results (Dice)**
Long et al. [18]	FCN	PASCAL VOC	–
Ronneberger et al. [1]	U-Net	ISBI 2015	–
Havaei et al. [21]	Two-pathway CNN	BraTS 2013	WT: 0.88; TC: 0.79; ET: 0.73
Pereira et al. [22]	Small-kernel CNN	BraTS 2013	WT: 0.88; TC: 0.83; ET: 0.77
Kamnitsas et al. [17]	DeepMedic (3D CNN + CRF)	BraTS 2015	WT: 0.90; TC: 0.76; ET: 0.73
Myronenko [26]	3D U-Net + VAE	BraTS 2018	WT: 0.91; TC: 0.86; ET: 0.82
Isensee et al. [27]	nnU-Net	23 datasets	Top-tier across all
Hatamizadeh et al. [29]	Swin UNETR	BraTS 2021	WT: 0.93; TC: 0.89; ET: 0.82
Naser & Deen [32]	U-Net + VGG16	Custom LGG	Mean: 0.84
Zeineldin et al. [30]	TransXAI (CNN + ViT)	BraTS	–
Hannisdal et al. [35]	nnU-Net	Custom GBM	Median: 0.81–0.82
Helland et al. [7]	DNN	Post-op GBM	–

Several comprehensive reviews have synthesized the state of the field. Dorfner et al. [9] conducted a comprehensive review of deep learning for brain tumor analysis in MRI, published in *npj Precision Oncology* in 2025, surveying architectures, training strategies, and evaluation methodologies across dozens of studies. They identified generalizability and clinical integration as key barriers to widespread adoption, noting that most studies report accuracy metrics in isolation without considering computational efficiency or deployment feasibility. Buchlak et al. [37] reviewed multimodal deep learning-based prognostication in glioma patients, published in *Cancers* in 2023, demonstrating that combining clinical, genomic, and imaging data through multimodal approaches outperforms unimodal methods. Singh et al. [38] reviewed radiomics and radiogenomics approaches in gliomas, published in the *British Journal of Cancer* in 2021, identifying standardization as a key challenge where different feature extraction toolkits and preprocessing pipelines produce inconsistent results. Ren et al. [39] surveyed artificial intelligence-based MRI radiomics and radiogenomics in glioma, published in *Cancer Imaging* in 2024, noting that most AI pipelines assume clean, preprocessed input without examining the consequences of preprocessing inconsistency.

2.2 Challenges Faced

Brain tumor segmentation presents multiple well-documented technical obstacles that make automated delineation inherently difficult, as evidenced by the persistent performance limitations observed across the studies reviewed above. First is the problem of severe class imbalance, where the tumor region typically occupies less than 5–10% of the total brain volume in a given MRI scan while the vast majority of voxels represent healthy brain tissue or background. This imbalance creates a straightforward statistical problem for learning algorithms: a classifier that predicts every voxel as non-tumor achieves over 90% accuracy, yet provides zero clinical utility. The studies by Milletari et al. [20] and Myronenko [26] specifically addressed this challenge through Dice-based loss functions and auxiliary regularization branches respectively, yet small tumor sub-regions—particularly the enhancing tumor class—remain the most difficult to segment accurately across all reviewed methods, with Dice scores for enhancing tumor consistently 10–15 percentage points lower than whole tumor scores [21, 17, 29]. Beyond class imbalance, the visual characteristics of brain tumors on MRI introduce substantial segmentation complexity. On T1-weighted pre-contrast images, tumor tissue frequently appears iso-intense or only mildly hypo-intense relative to surrounding gray and white matter, making intensity-based discrimination unreliable when the tumor has similar brightness to healthy tissue. Boundary ambiguity arises because glioblastomas are infiltrative neoplasms that do not form discrete capsulated masses but instead invade along white matter tracts and perivascular spaces, producing diffuse and ill-defined boundaries that lack the sharp intensity gradients required by edge-based segmentation methods [13, 5]. This infiltrative growth pattern means that even expert neuroradiologists disagree on the precise extent of tumor involvement, as reflected in the considerable inter-observer variability documented by Menze et al. [5] and Helland et al. [7]. Standard architectures that rely on local intensity gradients or learned boundary features struggle in these regions, as Kamnitsas et al. [17] and Havaei et al. [21] both noted that the majority of segmentation errors occur at these ambiguous tumor-to-normal tissue transitions.

Tumor appearance itself varies substantially across patients, grades, and locations within the brain. High-grade gliomas such as glioblastoma exhibit necrotic cores, enhancing rims, and surrounding edema that each present distinct intensity characteristics across MRI sequences, while lower-grade gliomas may appear as homogeneous masses without enhancement. Intra-patient heterogeneity is also pronounced, with different regions of the same tumor exhibiting different imaging characteristics depending on local cellularity, vascularity, and necrosis [32]. This heterogeneity means that segmentation models must handle an extreme range of appearances within a single tumor, a challenge that Pereira et al. [22] addressed through data augmentation and Ronneberger et al. [1] mitigated through skip connections that preserve multi-scale spatial information. Additionally, tumor size variability is extreme, ranging from sub-centimeter residual nodules to large masses occupying entire lobes, requiring architectures to operate effectively across this scale range—a property explicitly targeted by multi-scale approaches such as DeepMedic [17] and attention-based methods [23].

Multi-modal dependency represents a critical practical challenge identified across the reviewed literature. The most successful segmentation methods on BraTS benchmarks [5, 8, 26, 29] universally require all four standard MRI sequences (T1, T1-CE, T2, FLAIR) as input, exploiting the complementary information that each sequence provides about

different tumor compartments. However, Gkournelos et al. [36] demonstrated that removing even a single modality causes considerable performance degradation, revealing that these models do not learn robust representations of tumor appearance but instead rely on the specific multi-modal intensity patterns present during training. In clinical practice, complete multi-modal acquisitions are not always available due to patient intolerance, time constraints, or equipment limitations, making multi-modal dependency a direct obstacle to real-world deployment. Post-treatment imaging complexity introduces a distinct class of segmentation challenges that fundamentally differ from those encountered with de novo tumors. After surgical resection, radiation therapy, and chemotherapy, the brain exhibits treatment-induced changes that can closely mimic tumor on MRI [6, 7]. Pseudoprogression occurs in approximately 20–30% of GBM patients within the first months after chemoradiation, producing new or enlarging contrast enhancement that represents inflammatory response rather than true tumor growth but is virtually indistinguishable from genuine progression on conventional sequences. Radiation necrosis presents as a delayed complication with similar imaging characteristics to recurrent tumor. Surgical cavities, blood products, and post-operative edema further alter the expected tissue appearance and create additional intensity heterogeneity that confounds automated algorithms. Models trained exclusively on pre-treatment data, where tumor boundaries tend to be more clearly defined, experience substantial performance degradation when applied to this post-treatment landscape [25], as the learned feature representations do not capture these treatment-specific imaging patterns.

Domain shift across institutions and acquisition protocols compounds these segmentation challenges. MRI intensity values are not standardized across different scanners, field strengths, and acquisition parameters, meaning that the same tissue type can produce substantially different signal intensities depending on the imaging hardware and protocol used [9, 38]. Models trained on data from one institution frequently fail to generalize to external data, as the intensity distributions, noise characteristics, and spatial resolution differ from what was seen during training. Ren et al. [39] further noted that inconsistencies in preprocessing pipelines between training and inference exacerbate this domain shift, as the learned feature distributions become misaligned with the input statistics encountered during deployment. This sensitivity to acquisition conditions means that a model demonstrating strong performance on a benchmark dataset may produce unreliable results when applied to data from a different clinical site without careful domain adaptation or retraining.

2.3 Research Gap and Scope of Study

Despite the substantial body of work reviewed above, several critical gaps remain that motivate this thesis:

1. **Lack of multi-dimensional encoder comparison for post-treatment GBM:** Prior studies have evaluated encoder architectures primarily on pre-treatment data using accuracy metrics alone [5, 8, 32], without jointly considering model size and inference time. No study has benchmarked a diverse set of encoders—spanning residual (ResNet18, ResNet50), dense (DenseNet121), lightweight (MobileNetV2), compound-scaled (EfficientNet-B3), and transformer-based (MiT-B2) paradigms—

specifically on post-treatment GBM data with efficiency profiling. While Isensee et al. [27] demonstrated the importance of systematic pipeline design, their nnU-Net framework does not permit controlled encoder-level comparison. This thesis fills this gap by evaluating six architectures across three dimensions: segmentation accuracy (Dice coefficient and IoU), model complexity (parameter count), and computational cost (inference time).

2. **Limited investigation of preprocessing impact with pre-trained encoders:** While CLAHE and contrast enhancement have shown benefits in other medical imaging contexts [10], their interaction with ImageNet pre-trained encoders on post-treatment MRI has not been studied. Furthermore, the RGB channel stacking approach—combining different preprocessing variants (original, CLAHE, and contrast-enhanced) as multi-channel input—has not been evaluated against standard single-channel training in this domain. Singh et al. [38] identified preprocessing standardization as a key challenge, and Ren et al. [39] noted that most AI pipelines assume clean input without examining preprocessing consequences. This thesis provides the first systematic comparison of preprocessing strategies for post-treatment brain tumor segmentation with pre-trained encoders.

3. **Absence of cross-domain robustness evaluation:** The practical scenario where preprocessing applied during training is omitted during inference (domain mismatch) has not been quantified for brain tumor delineation, despite being a realistic concern when complex preprocessing pipelines cannot be guaranteed in deployment. This thesis measures this degradation explicitly, providing evidence-based guidance on the robustness implications of preprocessing choices.

4. **Single-modality performance benchmarks:** Most state-of-the-art methods rely on multi-parametric MRI input [35, 30, 31, 29, 26]. However, as demonstrated by Gkournelos et al. [36], performance degrades substantially when modalities are missing, yet no study has established deliberate single-modality baselines for post-treatment GBM. T1-weighted pre-contrast images are the most universally available MRI sequence; establishing benchmarks using only this modality provides practical lower-bound estimates and deployment guidance for the most constrained clinical scenarios.

This thesis addresses these gaps through a holistic evaluation framework that jointly considers segmentation accuracy, model efficiency, preprocessing effects, and cross-domain robustness, providing evidence-based recommendations for deploying deep learning-based delineation in post-treatment GBM workflows.

Chapter-3: Dataset

This chapter provides a detailed description of the dataset used in this study, including its source, demographics, imaging protocol, annotation methodology, and the data preparation strategies employed for the experiments.

3.1 Dataset Description

This study utilizes the University of California San Diego Post-Treatment Glioblastoma Multimodal MRI (UCSD-PTGBM) dataset [11, 12], a comprehensive, publicly available medical imaging dataset hosted on The Cancer Imaging Archive (TCIA). The UCSD-PTGBM dataset was specifically designed to support research in post-treatment glioblastoma analysis, addressing the critical need for annotated post-treatment imaging data in the neuro-oncology research community.

The dataset comprises 243 imaging timepoints collected from 178 unique subjects with histopathologically confirmed high-grade gliomas, all classified as World Health Organization (WHO) Grade IV—the most aggressive and malignant grade of brain tumors. All subjects underwent post-treatment follow-up imaging, meaning the MRI scans were acquired after the patients had received their initial treatment regimen (typically maximal safe surgical resection followed by concurrent chemoradiation per the Stupp protocol). This post-treatment context is critical because the imaging characteristics of post-treatment glioblastoma are fundamentally different from de novo (untreated) tumors, exhibiting complex appearances due to treatment-induced changes such as surgical cavities, post-surgical enhancement, pseudoprogression, and radiation effects.

The imaging was performed at the University of California San Diego using 3T clinical MRI scanners with an advanced brain tumor imaging protocol, ensuring high-quality, clinically representative data. The use of 3T field strength provides improved signal-to-noise ratio and spatial resolution compared to lower-field scanners, which is important for accurate tumor delineation.

3.2 Demographics and Clinical Metadata

Table 3.1 summarizes the key demographic characteristics of the patient population included in the UCSD-PTGBM dataset.

Table 3.1: UCSD-PTGBM Dataset Demographics

Characteristic	**Value**
Total Subjects	178
Total Imaging Timepoints	243
Age (mean ± SD)	56 ± 13 years
Sex (Male / Female)	116 / 62
Tumor Grade	WHO Grade IV
Scanner Field Strength	3T

The patient population exhibits a mean age of 56 years with a standard deviation of 13 years, which is consistent with the known epidemiology of glioblastoma that predominantly affects adults in their fifth to seventh decades of life. The sex distribution shows a male predominance (116 males, 62 females; ratio of approximately 1.87:1), which is also consistent with the established epidemiological pattern of glioblastoma showing a slight male preponderance [3].

The dataset includes additional clinical metadata for a subset of patients, including IDH (Isocitrate Dehydrogenase) mutation status, MGMT (O6-Methylguanine-DNA Methyltransferase) promoter methylation status, overall survival (OS), and progression-free survival (PFS). These molecular biomarkers are of significant prognostic importance in glioblastoma management: IDH-wildtype tumors (the majority) have a substantially worse prognosis than IDH-mutant variants, while MGMT promoter methylation is associated with improved response to temozolomide chemotherapy. While these clinical metadata are not directly used in the segmentation experiments of this thesis, their availability in the dataset enables potential future extensions to prognostic prediction and correlation analysis.

The fact that the dataset includes multiple imaging timepoints for some patients (243 timepoints from 178 subjects, indicating that some patients contributed multiple follow-up scans) is important for two reasons. First, it provides a longitudinal perspective that reflects the real-world clinical workflow of serial tumor monitoring. Second, it necessitates careful data partitioning to ensure patient-level splits (see Section 3.6), preventing data leakage from the same patient appearing in both training and test sets.

3.3 Imaging Protocol and MRI Sequences

The UCSD-PTGBM dataset provides a comprehensive set of MRI sequences at each imaging timepoint, reflecting the advanced brain tumor imaging protocol used in clinical practice:

- **T1-weighted pre-gadolinium (T1pre):** Acquired before the administration of gadolinium contrast agent, T1pre images provide fundamental anatomical information and baseline tissue contrast. In T1-weighted imaging, cerebrospinal fluid (CSF) appears dark, white matter appears bright, and gray matter appears intermediate. Tumors typically appear as areas of altered signal intensity, though the contrast between tumor and surrounding tissue is generally less pronounced compared to post-contrast imaging.
- **T1-weighted post-gadolinium (T1-CE):** Acquired after intravenous administration of gadolinium-based contrast agent, T1-CE images highlight regions where the blood-brain barrier has been disrupted, which is a hallmark of high-grade gliomas. Enhancing tumor regions appear as bright areas on T1-CE images, providing important information about the active tumor component.
- **T2-weighted FLAIR:** Fluid-Attenuated Inversion Recovery (FLAIR) images suppress the signal from cerebrospinal fluid while maintaining T2-weighted tissue contrast. FLAIR images are particularly valuable for visualizing peritumoral edema and non-enhancing tumor components, which appear as hyperintense (bright) regions.
- **Restricted Spectrum Imaging (RSI):** An advanced diffusion-weighted imaging technique that provides information about tissue cellularity and microstructure, aiding in the differentiation of tumor from treatment effects.
- **Dynamic Susceptibility Contrast (DSC) perfusion:** Provides information about tumor vascularity and blood flow, which can help distinguish highly vascularized tumor regions from treatment-related changes.
- **Arterial Spin Labeling (ASL):** A non-contrast perfusion technique that measures cerebral blood flow, providing complementary perfusion information without the need for contrast injection.

For this study, only the T1-weighted pre-contrast (T1pre) sequence is used as input to the segmentation models. This deliberate choice serves multiple purposes: (1) T1pre is the most universally available MRI sequence across clinical settings, including resource-constrained environments where advanced sequences may not be available; (2) using a single modality establishes a lower-bound performance benchmark, providing context for evaluating the marginal benefit of additional modalities; and (3) it simplifies the preprocessing and model pipeline, which is advantageous for clinical deployment.

3.4 Annotation Standards

Each imaging timepoint in the UCSD-PTGBM dataset includes neuroradiologist-approved voxelwise tumor segmentations that follow the BraTS (Brain Tumor Segmentation) annotation standard [6, 11]. The BraTS annotation protocol defines four tumor compartments:

1. **Enhancing Tumor (ET):** Regions showing contrast enhancement on T1-CE images, typically representing the most active tumor component with disrupted blood-brain barrier.

2. **Non-Enhancing/Necrotic Tumor Core (NETC):** Regions within the tumor core that do not show contrast enhancement, including areas of necrosis (cell death) that are commonly found in the center of glioblastoma tumors.

3. **Surrounding FLAIR Abnormality / Non-Enhancing FLAIR Hyperintensity (SNFH):** Peritumoral regions showing abnormal hyperintensity on FLAIR images, which may represent a combination of vasogenic edema and infiltrating tumor cells.

4. **Resection Cavity (RC):** Fluid-filled spaces resulting from surgical removal of tumor tissue, which are persistent features in post-treatment imaging.

The availability of expert neuroradiologist-approved annotations ensures high-quality ground truth labels for training and evaluation. The use of the established BraTS annotation standard facilitates comparison with other studies and benchmarks in the brain tumor segmentation literature.

3.5 Data Selection and Binary Mask Generation

For the binary segmentation task in this thesis, the multi-compartment BraTS annotations are merged into a single binary mask by combining all four tumor sub-regions (ET, NETC, SNFH, and RC) into a unified foreground class. This binary formulation simplifies the segmentation task to a two-class problem: tumor (foreground) versus non-tumor (background). While this binary approach loses the sub-regional information that may carry different prognostic implications, it provides a clear and well-defined evaluation framework for comparing encoder backbones and preprocessing strategies, which is the primary focus of this thesis.

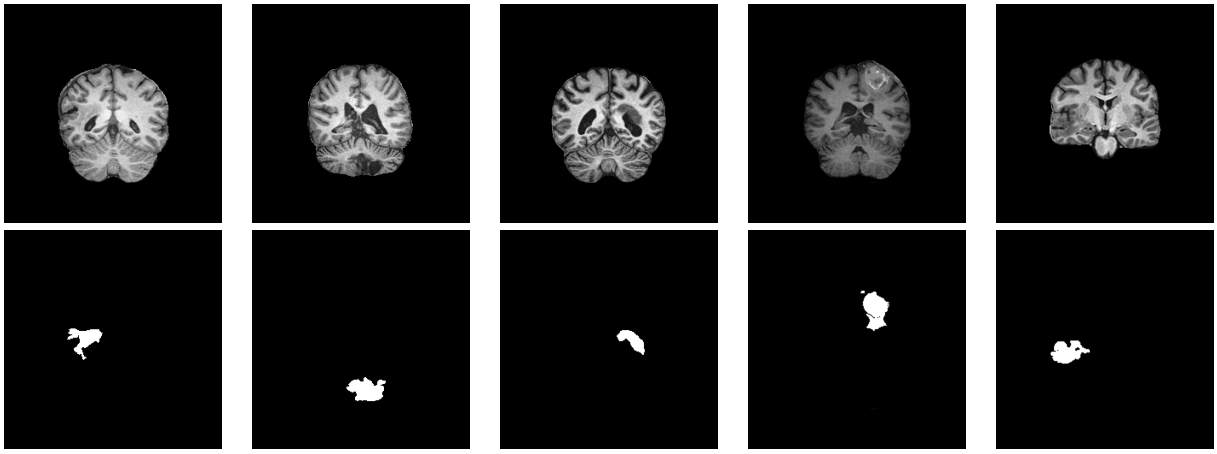


Figure 3.1: Sample T1pre MRI slices (top row) and corresponding binary tumor masks (bottom row) from the UCSD-PTGBM dataset, illustrating the variability in tumor size, shape, and location across different patients.

The 3D NIfTI volumes (the standard file format for neuroimaging data) are sliced along the axial plane to generate 2D slices, which are then converted to 8-bit normalized PNG images with intensities rescaled to the $[0, 255]$ range. The corresponding binary masks

are processed identically to maintain spatial alignment between MRI slices and their segmentation labels.

Figure 3.1 shows representative examples from the dataset, illustrating the T1pre MRI slices and their corresponding binary tumor masks. The examples demonstrate the variability in tumor size, shape, and location across different patients, as well as the relatively subtle contrast between tumor and surrounding tissue on T1pre images, which makes this segmentation task particularly challenging.

3.6 Data Partitioning Strategy

The dataset is partitioned using **5-fold cross-validation** with a strict **patient-wise split** to ensure rigorous and unbiased evaluation. In patient-wise splitting, all imaging slices from a given patient are assigned exclusively to either the training set or the validation set within each fold—no patient's data appears in both sets simultaneously. This is critical because different slices from the same patient share similar anatomical structures, tumor characteristics, and imaging conditions. If slices from the same patient were allowed to appear in both training and validation sets, the model could memorize patient-specific features rather than learning generalizable tumor segmentation patterns, leading to artificially inflated performance metrics (data leakage).

The 5-fold cross-validation procedure divides the patient pool into five approximately equal-sized, non-overlapping subsets. In each fold, four subsets are used for training and one subset is used for validation. The process is repeated five times, with each subset serving as the validation set exactly once. Final performance metrics are computed as the average across all five folds, providing a robust estimate of model performance that accounts for the variability inherent in any single train-test split.

Run \ Fold	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4
Training run 1	Training	Training	Training	Validation	Test
Training run 2	Training	Validation	Training	Test	Training
Training run 3	Validation	Training	Test	Training	Training
Training run 4	Training	Test	Training	Training	Validation
Training run 5	Test	Training	Validation	Training	Training

Figure 3.2: Illustration of the 5-fold patient-wise cross-validation.

This rigorous evaluation methodology ensures that the reported performance metrics reflect the model's ability to generalize to unseen patients, which is the clinically relevant measure of performance. All models and preprocessing strategies in this thesis are evaluated using the same 5-fold cross-validation splits to ensure fair and consistent comparison.

Chapter-4: Methodology

This chapter provides a detailed description of the complete methodology employed in this thesis, including the overall experimental pipeline, data preprocessing strategies, model architectures, training protocol, and evaluation framework.

4.1 Overall Pipeline Overview

The experimental methodology follows a systematic two-phase evaluation approach designed to provide comprehensive insights into both encoder backbone selection and preprocessing strategy effectiveness for post-treatment glioblastoma segmentation.

Phase 1: Encoder Backbone Comparison. In the first phase, all six encoder backbones (ResNet18, ResNet50, DenseNet121, MobileNetV2, MiT-B2, and EfficientNet-B3) are trained and evaluated on original (unprocessed) T1-weighted pre-contrast MRI images using 5-fold patient-wise cross-validation. Each encoder is assessed across three dimensions: segmentation performance (Dice coefficient and IoU), model complexity (parameter count), and computational efficiency (inference time). This phase identifies the best overall encoder backbone.

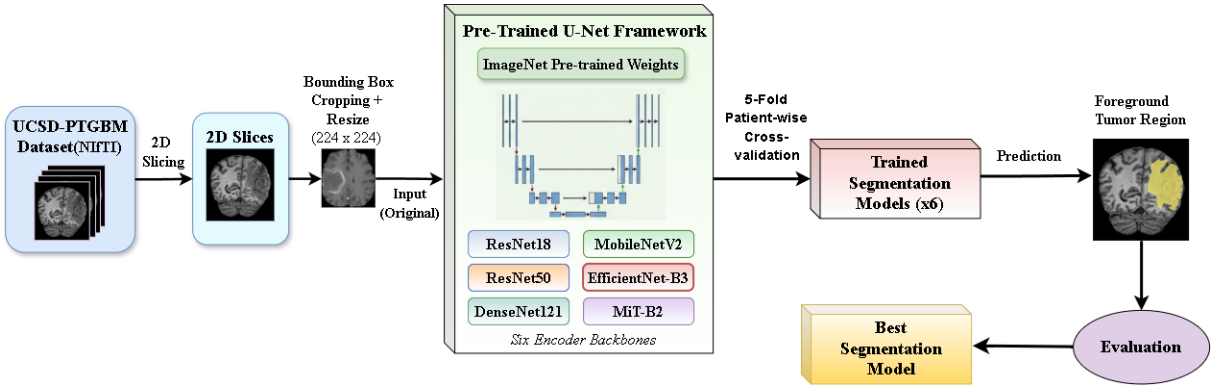


Figure 4.1: Encoder Backbone Comparison on original images.

Phase 2: Preprocessing Strategy Evaluation. In the second phase, the best-performing encoder identified in Phase 1 is used to evaluate three preprocessing strategies (CLAHE, contrast enhancement, and RGB channel stacking) against the baseline of raw image training. Models trained on preprocessed images are tested on original (unprocessed) images to assess cross-domain robustness, simulating the realistic clinical scenario where preprocessing may not be consistently maintained during deployment.

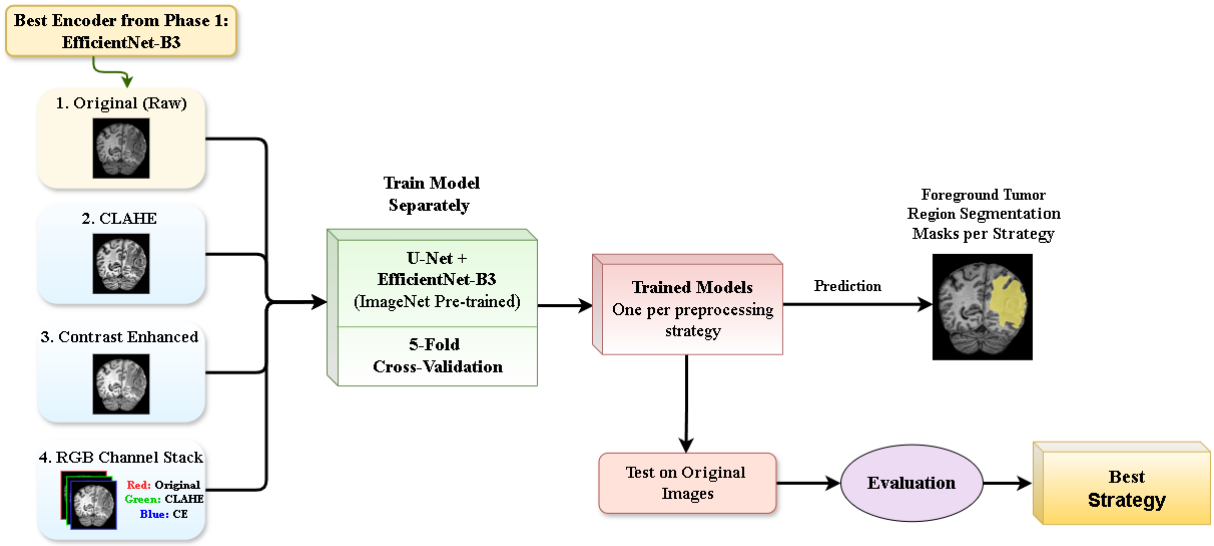


Figure 4.2: Encoder Backbone Comparison on original images.

Figure 4.10 illustrates the complete methodology pipeline showing the flow from data preprocessing through model training and evaluation.

4.2 Data Preprocessing

Data preprocessing transforms the raw 3D MRI volumes into 2D input images suitable for the segmentation models. The preprocessing pipeline consists of several sequential steps, with optional additional enhancement strategies applied after the base preprocessing.

4.2.1 Volume Slicing and Normalization

The original MRI data is stored in NIfTI format (.nii or .nii.gz), a standard neuroimaging file format that encodes 3D volumetric data along with spatial metadata (voxel dimensions, orientation, coordinate system). The preprocessing pipeline extracts 2D slices from the 3D volumes along the axial (transverse) plane, which is the most commonly used orientation for brain tumor assessment in clinical practice.

Each extracted axial slice undergoes intensity normalization to convert the raw voxel intensity values (which can vary widely in range and distribution across different patients and scanners) into standardized 8-bit images with intensities in the range $[0, 255]$. This normalization is performed using min-max scaling:

$$I_{\text{norm}}(x, y) = \frac{I(x, y) - I_{\min}}{I_{\max} - I_{\min}} \times 255 \tag{4.1}$$

where $I(x, y)$ is the original voxel intensity at position (x, y), $I_{\min}$ and $I_{\max}$ are the minimum and maximum intensity values of the slice, and $I_{\text{norm}}(x, y)$ is the normalized intensity value.

The resulting normalized images are saved as PNG format for efficient storage and loading during training.

The corresponding binary segmentation masks undergo the same slicing process to maintain spatial alignment. Mask values are binarized such that any voxel belonging to any of the four tumor sub-regions (ET, NETC, SNFH, RC) is assigned a value of 1 (foreground), while all other voxels are assigned a value of 0 (background).

4.2.2 Bounding Box Cropping

To reduce the computational burden and focus the model's attention on the relevant brain region, a bounding box cropping strategy is applied. For each patient, the bounding box enclosing the largest foreground (tumor) region across all axial slices is computed. This bounding box defines the spatial extent of the tumor region with sufficient margin to include relevant surrounding context.

The computed bounding box is then applied uniformly to all MRI slices and corresponding masks for that patient, ensuring spatial consistency across slices. This cropping step removes excessive background (non-brain regions, such as the surrounding air and skull) while preserving the tumor and its immediate anatomical context. The uniform application of the same bounding box across all slices of a patient prevents inconsistent cropping that could introduce artificial spatial shifts between adjacent slices.

After cropping, all images are resized to a fixed resolution of 224×224 pixels to match the input size expected by the ImageNet pre-trained encoder backbones. This resizing step uses bilinear interpolation for MRI images and nearest-neighbor interpolation for binary masks (to preserve the discrete label values).

4.2.3 CLAHE Preprocessing

Contrast-Limited Adaptive Histogram Equalization (CLAHE) [40] is applied as the first preprocessing strategy. CLAHE divides the image into small non-overlapping rectangular regions called tiles and performs histogram equalization independently on each tile. The key parameter is the **clip limit**, which restricts the maximum height of the histogram bins before equalization. Histogram values exceeding the clip limit are redistributed uniformly across all bins, preventing the over-amplification of contrast that can occur with standard adaptive histogram equalization.

The mathematical formulation of CLAHE can be described as follows. For each tile, the cumulative distribution function (CDF) of the local histogram is computed:

$$\mathrm{CDF}(i) = \sum_{j=0}^{i} p(j) \tag{4.2}$$

where $p(j)$ is the clipped and redistributed probability of intensity level j. The equalized

intensity value is then:

$$I_{\text{CLAHE}}(x, y) = \text{round}\left(\text{CDF}(I(x, y)) \times (L - 1)\right) \tag{4.3}$$

where L is the number of intensity levels (256 for 8-bit images). Bilinear interpolation is applied at tile boundaries to eliminate artificial block artifacts, producing a smooth, locally contrast-enhanced output image.

CLAHE is particularly effective for enhancing the visibility of subtle tissue boundaries and internal tumor heterogeneity in MRI images, where the contrast between different tissue types may be limited on T1-weighted pre-contrast sequences. Figure 4.3 shows the effect of CLAHE on a representative MRI slice.

4.2.4 Contrast Enhancement

Standard contrast enhancement (CE) is applied as the second preprocessing strategy. This technique stretches the intensity range of the image to utilize the full dynamic range, improving the visual distinction between different tissue types. The contrast stretching operation is defined as:

$$I_{\text{CE}}(x, y) = \frac{I(x, y) - I_{\text{low}}}{I_{\text{high}} - I_{\text{low}}} \times 255 \tag{4.4}$$

where I_{low} and I_{high} are the lower and upper intensity boundaries used for stretching (which may be set to specific percentile values to exclude outliers). Values below I_{low} are clipped to 0, and values above I_{high} are clipped to 255.

Unlike CLAHE, which operates locally on individual tiles, contrast enhancement is a global operation that applies the same transformation uniformly across the entire image. This preserves the relative intensity relationships between different brain regions while expanding the overall dynamic range, which may be beneficial for maintaining compatibility with ImageNet pre-trained encoders that have learned to interpret global intensity patterns from natural images.

4.2.5 RGB Channel Stacking

The RGB channel stacking strategy is the third and most novel preprocessing approach investigated in this thesis. Instead of training on a single grayscale image (replicated across three channels), this approach combines three different representations of the same MRI slice into the three color channels of a single composite image:

- **Red channel:** Original grayscale MRI image
- **Green channel:** CLAHE-enhanced version of the MRI image

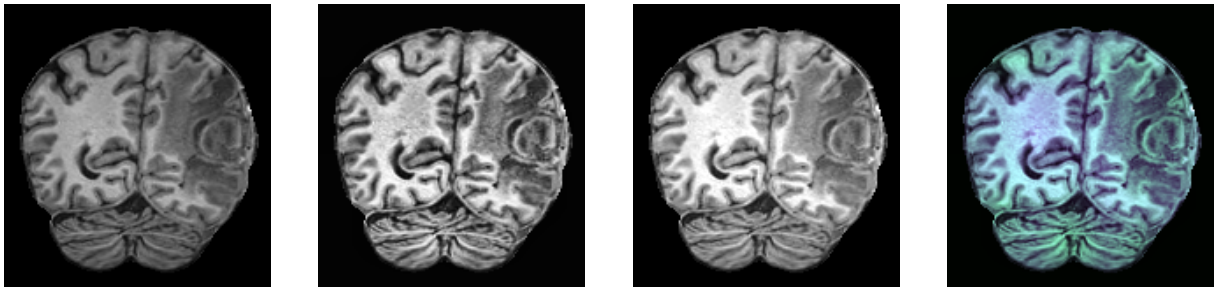
- **Blue channel:** Contrast-enhanced version of the MRI image

The rationale behind this approach is to enrich the feature space available to the encoder by providing complementary views of the same anatomical information. Each preprocessing variant emphasizes different aspects of the image—the original preserves natural intensity relationships, CLAHE enhances local contrast and texture details, and CE expands the global dynamic range—and combining them into RGB channels allows the model to leverage all three representations simultaneously.

This strategy is inspired by the multi-modal MRI approach [34], where different MRI sequences (T1, T2, FLAIR) are stacked as input channels. The key difference is that RGB stacking uses different preprocessed versions of a single modality rather than different physical modalities, making it applicable even when only one MRI sequence is available.

For the other three input strategies (original, CLAHE-only, CE-only), the single-channel grayscale images are replicated across all three channels to create a 3-channel input that matches the RGB format expected by ImageNet pre-trained encoders.

Figure 4.3 illustrates the visual appearance of all four input representations for a representative T1-weighted MRI slice.



(a) Original (b) CLAHE (c) CE (d) RGB Stack

Figure 4.3: Preprocessing variants applied to a representative T1-weighted MRI slice: (a) Original grayscale image, (b) CLAHE-enhanced image showing improved local contrast, (c) Contrast-enhanced image with expanded dynamic range, (d) RGB channel-stacked composite image combining all three representations.

4.3 Model Architecture

The segmentation framework in this thesis uses the U-Net architecture with interchangeable encoder backbones. This section describes the overall framework and the specific characteristics of each encoder backbone evaluated.

4.3.1 U-Net Framework

U-Net [1] serves as the base segmentation architecture for all experiments. The U-Net framework consists of an encoder (contracting path) that progressively extracts hierarchical features through downsampling operations, and a decoder (expanding path) that

progressively recovers spatial resolution through upsampling operations. Skip connections between corresponding encoder and decoder stages concatenate feature maps, enabling the decoder to access high-resolution spatial information for precise boundary delineation.

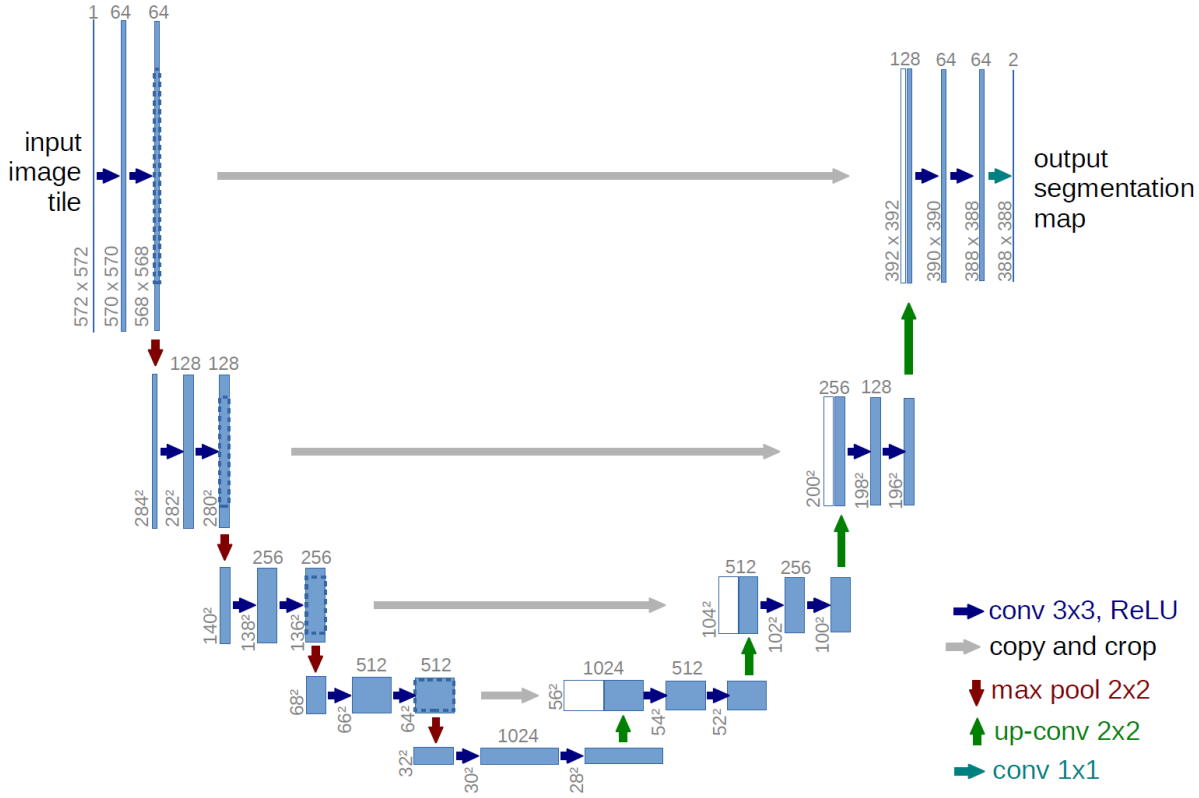


Figure 4.4: U-Net Architecture [1].

In the implementation used in this thesis, the encoder portion of U-Net is replaced with each of the six pre-trained backbone architectures, while the decoder structure remains consistent across all experiments. This modular design ensures that differences in segmentation performance can be attributed directly to the encoder backbone's feature extraction capabilities rather than variations in the decoder architecture.

All encoder backbones are initialized with weights pre-trained on the ImageNet dataset (ILSVRC-2012) [41], which contains approximately 1.2 million images across 1,000 object categories. The rationale for this transfer learning approach is that low-level features learned from natural images—such as edges, textures, and gradients—are sufficiently generic to be useful for medical image feature extraction, while higher-level features can be adapted to the medical domain through fine-tuning [15]. The pre-trained weights thus provide a strong initialization for the feature extraction layers, enabling effective training even with the relatively limited annotated medical imaging data available, accelerating convergence, and often leading to better final performance than training from scratch.

Table 4.1 provides a summary comparison of the six encoder architectures evaluated in this thesis.

Table 4.1: Summary of Encoder Backbone Architectures

Encoder	Paradigm	Key Design	Params (M)
ResNet18	Residual	Basic blocks, skip connections	14.33
ResNet50	Residual	Bottleneck blocks	32.52
DenseNet121	Dense	Dense connectivity	13.61
MobileNetV2	Lightweight	Depthwise separable conv.	6.63
EfficientNet-B3	Compound	Compound scaling, MBConv	13.16
MiT-B2	Transformer	Efficient self-attention	27.48

4.3.2 ResNet18 Encoder

ResNet18 [2] is employed as a lightweight residual encoder. The architecture consists of an initial layer of 7×7 convolutions with a stride of 2, followed by max pooling, and then four residual stages. Each stage contains two basic residual blocks, with each block consisting of two 3×3 convolutional layers with batch normalization and ReLU activation, along with an identity shortcut connection. The four stages produce feature maps at progressively reduced spatial resolutions (56×56, 28×28, 14×14, 7×7 for 224×224 input) with increasing channel depths (64, 128, 256, 512).

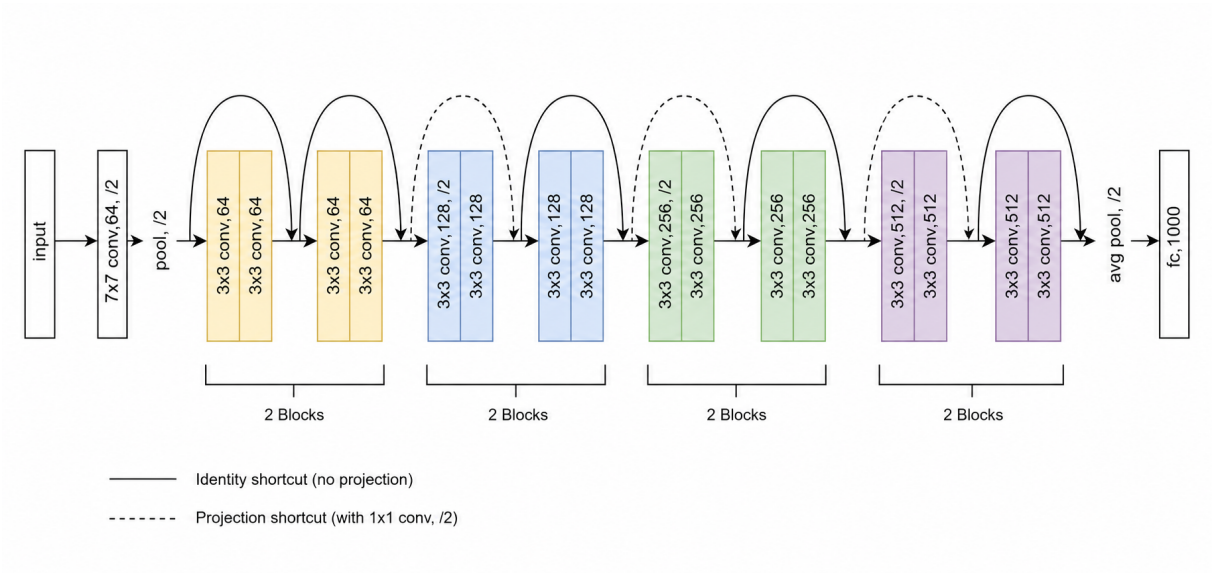


Figure 4.5: ResNet-18 Architecture.

As a U-Net encoder, ResNet18 contributes 14.33M parameters to the overall segmentation model. Its relatively shallow depth and simple block design make it computationally efficient, with the fastest inference time among the non-lightweight encoders, making it suitable for applications where processing speed is a priority.

4.3.3 ResNet50 Encoder

ResNet50 [2] represents a deeper and more capable variant of the residual network family. It shares the same four-stage structure as ResNet18 but replaces basic blocks with bottleneck blocks. Each bottleneck block consists of three layers: a 1×1 convolution that reduces the channel dimension (bottleneck), a 3×3 convolution that performs spatial feature extraction in the reduced dimension, and a 1×1 convolution that restores the original channel dimension. This bottleneck design reduces computational cost while enabling deeper architectures with richer feature representations.

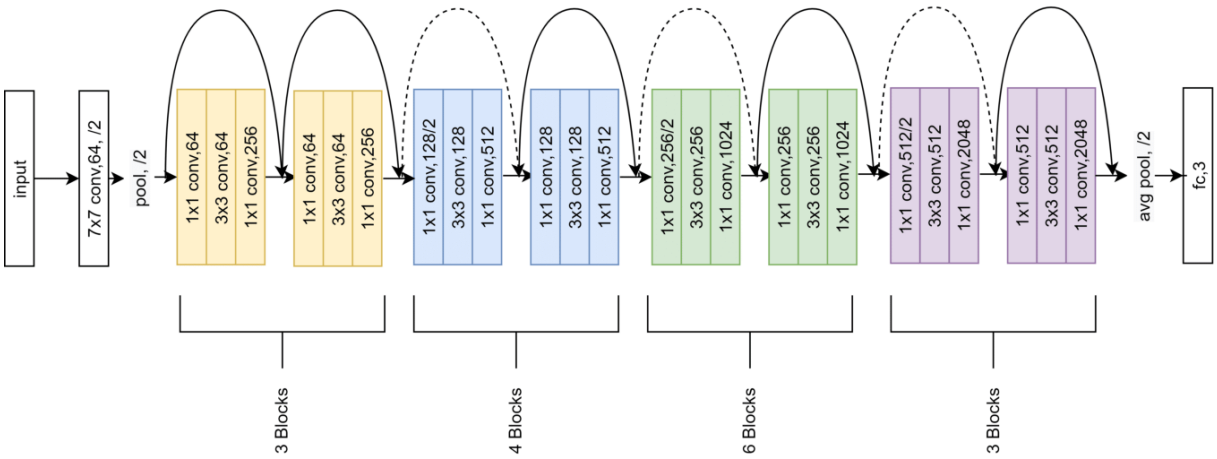


Figure 4.6: ResNet-50 Architecture [2].

The four stages of ResNet50 contain 3, 4, 6, and 3 bottleneck blocks respectively, producing feature maps with the same spatial resolutions as ResNet18 but with channel depths of 256, 512, 1024, and 2048. As a U-Net encoder, ResNet50 contributes 32.52M parameters, making it the second-largest model evaluated. The substantially higher parameter count and deeper architecture provide greater representational capacity but at the cost of increased computational requirements.

4.3.4 DenseNet121 Encoder

DenseNet121 [42] employs dense connectivity within each of its four dense blocks. In a dense block, every layer receives the concatenated feature maps of all preceding layers as input and passes its own feature maps to all subsequent layers. DenseNet121's four dense blocks contain 6, 12, 24, and 16 layers respectively (totaling 121 layers including transition layers and the initial convolution).

Each layer within a dense block consists of batch normalization, ReLU activation, and a 3×3 convolution that produces a fixed number of new feature channels (the growth rate, $k = 32$). Transition layers between dense blocks perform downsampling through batch

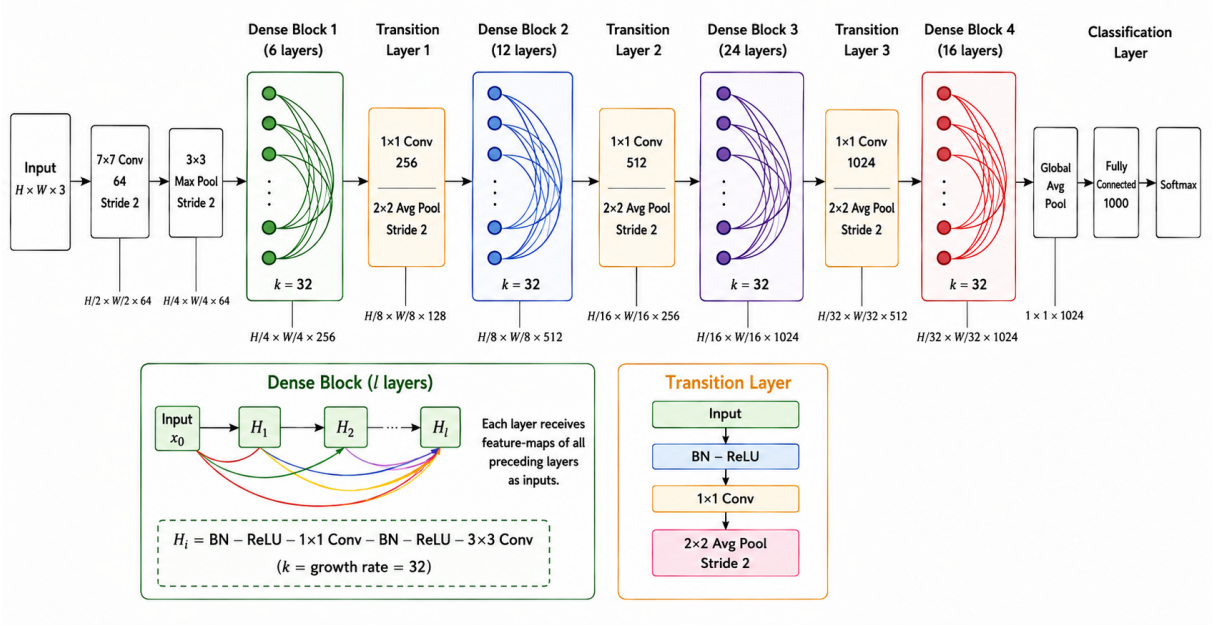


Figure 4.7: DenseNet-121 Architecture.

normalization, 1×1 convolution (which halves the number of channels for compression), and 2×2 average pooling.

As a U-Net encoder, DenseNet121 contributes 13.61M parameters. The dense connectivity pattern promotes feature reuse and gradient flow, achieving competitive accuracy with fewer parameters than comparably deep residual networks. However, the need to store and concatenate feature maps from all preceding layers increases memory consumption during training and contributes to longer inference times compared to architectures of similar parameter count.

4.3.5 MobileNetV2 Encoder

MobileNetV2 [43] is the most lightweight encoder evaluated, employing depthwise separable convolutions and inverted residual blocks with linear bottlenecks. The architecture consists of an initial standard convolution followed by a series of inverted residual blocks organized into seven stages with varying expansion ratios and output channel dimensions.

Each inverted residual block follows the sequence: (1) 1×1 pointwise convolution to expand the channel dimension by a specified expansion ratio (typically 6); (2) 3×3 depthwise convolution applied independently to each expanded channel; (3) 1×1 pointwise convolution to project back to a narrow bottleneck dimension; and (4) a residual shortcut connection between the narrow bottleneck representations (only when input and output dimensions match).

The use of linear (non-ReLU) activations in the final pointwise projection is a critical design choice that prevents information loss in the low-dimensional bottleneck representations. As a U-Net encoder, MobileNetV2 contributes only 6.63M parameters, making it the smallest model evaluated and the fastest in terms of inference time. However, the aggressive parameter reduction may limit its ability to capture the complex and subtle features

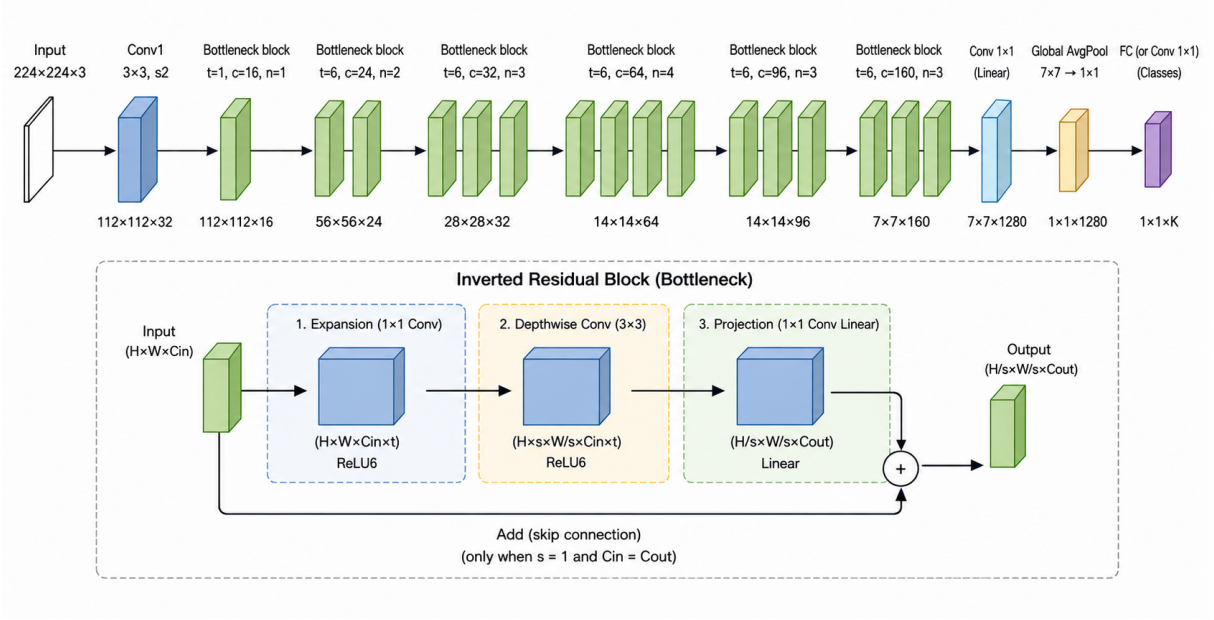


Figure 4.8: MobileNetV2 Architecture.

necessary for accurate brain tumor segmentation.

4.3.6 EfficientNet-B3 Encoder

EfficientNet-B3 [44] is derived from the EfficientNet family through compound scaling of the baseline EfficientNet-B0 architecture. The compound scaling method uniformly scales network depth ($d = 1.4$), width ($w = 1.2$), and input resolution ($r = 1.15$) relative to B0, using the compound coefficient $\phi = 3$.

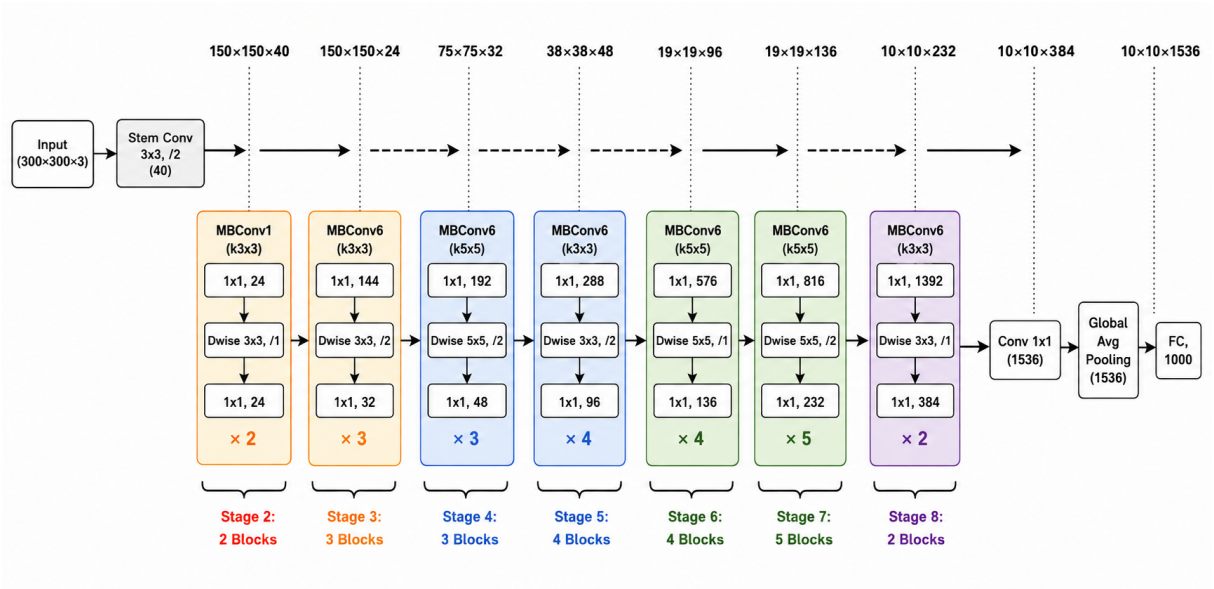


Figure 4.9: EfficientNet-B3 Architecture.

The building blocks of EfficientNet are Mobile Inverted Bottleneck Convolution (MBConv) blocks, which combine depthwise separable convolutions with squeeze-and-excitation (SE)

optimization [45]. The SE module adaptively recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels through a global average pooling, followed by two fully connected layers (reduction and excitation), and finally a sigmoid gating function that scales each channel's feature map.

As a U-Net encoder, EfficientNet-B3 contributes 13.16M parameters. The compound scaling approach ensures that depth, width, and resolution are balanced, avoiding the suboptimal configurations that result from scaling any single dimension in isolation. This balanced scaling philosophy enables EfficientNet-B3 to achieve accuracy competitive with much larger models while maintaining moderate computational requirements.

4.3.7 MiT-B2 (SegFormer) Encoder

MiT-B2 [46] (Mix Transformer B2) is the encoder component of SegFormer, designed specifically for efficient semantic segmentation. Unlike the convolutional encoders described above, MiT-B2 uses a hierarchical transformer architecture that processes the input image through four transformer stages at progressively reduced spatial resolutions.

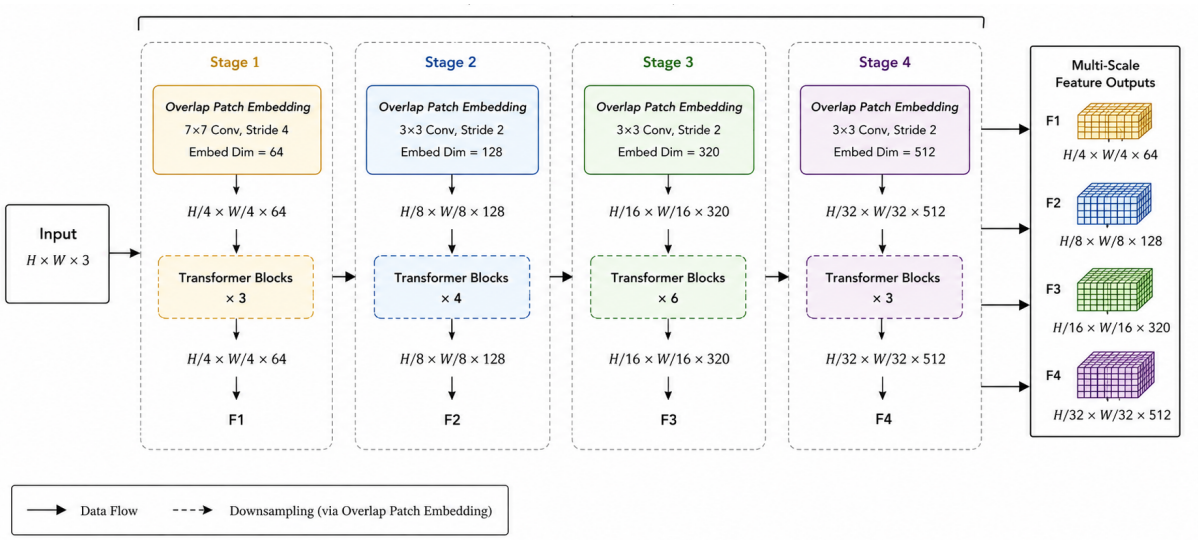


Figure 4.10: MiT-B2 Architecture.

Each transformer stage consists of three main components: (1) **overlapping patch embedding** that splits the input into overlapping patches and projects them into a sequence of embedding vectors, preserving local continuity across patch boundaries; (2) **efficient self-attention** that reduces the spatial dimensions of keys and values by a reduction ratio R, reducing the computational complexity from $O(N^2)$ to $O(N^2/R)$, where N is the sequence length; and (3) **Mix-FFN** (Feed-Forward Network) that incorporates a 3×3 depthwise convolution within the standard FFN to encode positional information without requiring explicit positional encoding.

As a U-Net encoder, MiT-B2 contributes 27.48M parameters. Its transformer-based design enables the capture of long-range spatial dependencies that may be missed by purely convolutional encoders, which have limited receptive fields determined by their kernel sizes and network depth. This capability is potentially valuable for brain tumor segmentation,

where tumor regions may have complex shapes and spatial relationships with surrounding anatomical structures.

4.4 Training Protocol

This section describes the training configuration and hyperparameters used for all experiments.

4.4.1 Loss Function: Dice Loss

All models are trained using **Dice Loss** in multiclass mode computed from logits. Dice Loss is derived from the Dice Similarity Coefficient and directly optimizes the overlap between the predicted segmentation and the ground truth. The Dice Loss for binary segmentation is defined as:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon} \tag{4.5}$$

where p_i is the predicted probability for pixel i, g_i is the ground truth label for pixel i, and ϵ is a small constant added for numerical stability to prevent division by zero. The summation is computed over all pixels in the image.

Dice Loss is preferred over standard cross-entropy loss for medical image segmentation because it directly optimizes the metric of interest (Dice coefficient) and is inherently robust to class imbalance. In brain tumor segmentation, the tumor region typically occupies a small fraction of the total image area, creating a severe class imbalance. Cross-entropy loss can be dominated by the large number of easily classified background pixels, while Dice Loss gives equal weight to the overlap between prediction and ground truth regardless of the relative sizes of foreground and background regions.

4.4.2 Optimizer and Learning Rate

The **Adam optimizer** [47] (Adaptive Moment Estimation) is used for all experiments with a learning rate of 2×10^{-4}. Adam combines the benefits of two popular optimization techniques: momentum (which accelerates convergence by accumulating exponentially decaying averages of past gradients) and RMSProp (which adapts the learning rate for each parameter based on the magnitude of recent gradients). The Adam update rule is:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t \tag{4.6}$$

where θ_t are the model parameters at step t, $\eta = 2 \times 10^{-4}$ is the learning rate, $\hat{m}_t$ is the

bias-corrected first moment estimate (mean of gradients), and $\hat{v}_t$ is the bias-corrected second moment estimate (mean of squared gradients).

The chosen learning rate of 2×10^{-4} represents a moderate value that balances convergence speed with training stability, suitable for fine-tuning pre-trained models where the initial weights are already close to a good solution and overly large learning rates could destabilize the pre-trained representations.

4.4.3 Early Stopping

Training runs for a maximum of **100 epochs** with **early stopping** based on validation loss. Early stopping monitors the validation loss at the end of each epoch and terminates training when the validation loss has not improved for a specified number of consecutive epochs (patience). The model weights from the epoch with the best validation loss are restored as the final model.

This mechanism serves two important purposes: (1) it prevents overfitting by stopping training before the model begins to memorize the training data at the expense of generalization performance; and (2) it reduces unnecessary computational expense by terminating training once further improvement is unlikely.

4.4.4 ImageNet Normalization

All input images are normalized using the **encoder-specific ImageNet mean and standard deviation** values. For each encoder, the pixel values in each channel are standardized as:

$$I^{(c)}_{\text{normalized}} = \frac{I^{(c)} - \mu^{(c)}}{\sigma^{(c)}} \tag{4.7}$$

where $I^{(c)}$ is the pixel value in channel c, $\mu^{(c)}$ is the ImageNet mean for channel c, and $\sigma^{(c)}$ is the ImageNet standard deviation for channel c. The standard ImageNet normalization values are $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$ for the RGB channels.

This normalization ensures that the input distribution matches the distribution expected by the pre-trained encoder weights, which is essential for effective transfer learning. Without this normalization, the pre-trained features would be misaligned with the input statistics, potentially degrading performance.

Table 4.2 summarizes all training hyperparameters used in the experiments.

Table 4.2: Training Hyperparameters

Hyperparameter	**Value**
Loss Function	Dice Loss (multiclass, from logits)
Optimizer	Adam
Learning Rate	2×10^{-4}
Maximum Epochs	100
Early Stopping	Yes (based on validation loss)
Input Resolution	224×224 pixels
Input Channels	3 (grayscale replicated or RGB stack)
Normalization	Encoder-specific ImageNet mean/std
Pre-trained Weights	ImageNet (ILSVRC-2012)
Cross-Validation	5-fold, patient-wise split
Data Augmentation	None

4.5 Evaluation Framework

A comprehensive evaluation framework is essential to rigorously assess and compare segmentation models, particularly in the medical imaging domain where performance implications extend directly to patient care. This section describes the four standard metrics used to evaluate segmentation performance: Intersection over Union (IoU), Dice Coefficient (F1 Score), Precision, and Recall. These four metrics were selected because they collectively characterize different aspects of segmentation quality—spatial overlap (IoU, Dice), false positive control (Precision), and false negative control (Recall)—and are the most widely adopted metrics in both the BraTS challenge benchmarks [5] and the broader medical image segmentation literature [15], enabling direct comparison with existing work.

All metrics are computed exclusively on the **foreground (tumor) class** to provide a clinically meaningful assessment that is not biased by the dominant background class. In brain tumor segmentation, the background class typically occupies more than 90% of the image, and including it in the metric computation would artificially inflate performance scores. For instance, a trivial model that predicts every pixel as background would achieve pixel-wise accuracy exceeding 90% on a typical brain MRI, yet provide no clinical utility whatsoever. By restricting evaluation to the foreground class, the metrics directly reflect the model's ability to detect and delineate the tumor region itself, which is the clinically actionable output.

4.5.1 Intersection over Union (IoU)

Intersection over Union (IoU), also known as the Jaccard Index, measures the overlap between the predicted segmentation and the ground truth as a ratio of their intersection to their union:

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} = \frac{TP}{TP + FP + FN} \tag{4.8}$$

where P is the set of predicted foreground pixels, G is the set of ground truth foreground pixels, TP is the number of true positives (correctly predicted foreground pixels), FP is the number of false positives (background pixels incorrectly predicted as foreground), and FN is the number of false negatives (foreground pixels incorrectly predicted as background).

IoU ranges from 0 (no overlap) to 1 (perfect overlap) and is a strict metric that heavily penalizes both over-segmentation (high FP) and under-segmentation (high FN). Compared to the Dice coefficient, IoU applies a harsher penalty for errors: for a given prediction, the IoU denominator (TP + FP + FN) counts error pixels only once, whereas the Dice denominator (2TP + FP + FN) includes true positives twice, effectively diluting the error contribution. As a result, IoU provides a more conservative and stringent assessment of overlap quality.

An important practical consideration is the handling of **edge cases**. When both the ground truth and the prediction contain no foreground pixels (i.e., $|P| = |G| = 0$), IoU is undefined (0/0). In this study, such cases are assigned an IoU of 1.0, because the model correctly predicts the absence of tumor. Conversely, when only one of P or G is empty, the IoU is 0, reflecting a complete miss or complete false alarm. These conventions ensure that the metric remains well-defined across the entire dataset, including slices that contain no visible tumor.

IoU is widely used in the computer vision and medical imaging communities as a primary segmentation evaluation metric. In the specific context of brain tumor segmentation, IoU provides a complementary view to the Dice coefficient: while both metrics quantify overlap, IoU's stricter nature makes it more sensitive to moderate errors in boundary delineation, which is particularly relevant for assessing precise tumor margin detection in post-treatment MRI where boundaries are inherently ambiguous.

4.5.2 Dice Coefficient (F1 Score)

The Dice Coefficient, equivalent to the F1 Score, measures the harmonic mean of precision and recall, or equivalently, twice the ratio of the intersection to the total number of predicted and ground truth foreground pixels:

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|} = \frac{2TP}{2TP + FP + FN} \tag{4.9}$$

The Dice coefficient also ranges from 0 to 1, and is mathematically related to IoU: $\text{Dice} = 2 \times \text{IoU}/(1 + \text{IoU})$. Dice is always greater than or equal to IoU for the same prediction, and the two metrics are monotonically related. The Dice coefficient is the primary evaluation metric used in the BraTS Challenge [5] and is the most widely reported metric in the brain tumor segmentation literature, making it the preferred metric for comparison with existing studies.

An important distinction must be drawn between the Dice coefficient used as an **evaluation metric** (as described here) and the Dice loss used as a **training objective** (described in Section 4.4.1). As an evaluation metric, the Dice coefficient is computed on hard (binarized) predictions after thresholding the model's output probabilities at 0.5. As a training loss, the soft Dice loss operates on continuous probability values, enabling gradient-based optimization. While both are derived from the same mathematical formulation, their behavior differs: the soft Dice loss is differentiable and suitable for backpropagation, whereas the hard Dice evaluation metric provides a threshold-dependent assessment of the final segmentation quality.

The Dice coefficient is particularly sensitive to errors on **small objects**. When the ground truth tumor region is small (few pixels), even a modest number of false positive or false negative pixels can cause a disproportionately large drop in Dice. This property is relevant for post-treatment glioblastoma segmentation, where residual tumor volumes can be small following surgical resection. Models that achieve high mean Dice scores must therefore perform well not only on large, easily detectable tumors but also on small residual regions, making the metric an effective discriminator of clinically meaningful performance differences.

4.5.3 Precision

Precision measures the proportion of predicted foreground pixels that are actually foreground in the ground truth:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.10}$$

High precision indicates that the model has few false positive predictions, meaning that when the model predicts a pixel as tumor, it is likely to be correct. In clinical terms, high precision corresponds to few regions being incorrectly identified as tumor, minimizing the risk of unnecessary alarm or overtreatment. However, high precision alone is insufficient because a model that predicts very few pixels as tumor can achieve high precision while missing large portions of the actual tumor (low recall).

The interplay between precision and the decision threshold is worth noting: increasing the probability threshold above the default 0.5 would cause the model to predict tumor only for high-confidence pixels, thereby increasing precision at the expense of recall. Conversely, lowering the threshold would increase recall but reduce precision. In this study, a fixed threshold of 0.5 is used consistently across all models to ensure a fair comparison. Precision therefore reflects each encoder's inherent tendency toward conservative or aggressive prediction behavior at a standardized operating point.

In the clinical context of post-treatment glioblastoma, false positive predictions (low precision) may lead to overestimation of tumor burden, potentially causing unnecessary concern, additional imaging, or premature changes to the treatment plan. While over-detection is generally considered less harmful than under-detection (as it can be verified by radiologists during clinical review), consistently high false positive rates would undermine clinician confidence in the automated tool and increase the review burden.

4.5.4 Recall

Recall (also known as sensitivity or true positive rate) measures the proportion of ground truth foreground pixels that are correctly identified by the model:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.11}$$

High recall indicates that the model successfully identifies most of the actual tumor pixels, with few missed regions. In the clinical context of tumor segmentation, recall is particularly important because **under-detection of tumor (false negatives) carries greater clinical risk than over-detection (false positives)**. Missed tumor regions could lead to underestimation of tumor burden, potentially affecting treatment decisions and response assessment. Over-detection, while undesirable, can be more easily identified and corrected through clinical review.

The clinical asymmetry between false negatives and false positives is especially pronounced in the post-treatment monitoring setting, where the primary objective is to detect disease progression or recurrence early enough to modify therapy. A model with low recall risks missing early signs of tumor regrowth—small enhancing regions or subtle changes in tumor morphology—that may represent the earliest opportunity for therapeutic intervention. In contrast, false positive predictions, while increasing the review burden for neuroradiologists, are typically identified and dismissed during the standard clinical review process without adverse consequences to the patient.

Recall is also closely related to the concept of **negative predictive value (NPV)** in clinical decision-making: a model with high recall provides strong assurance that a negative prediction (no tumor in a region) is truly negative, which is essential for confident clinical decision-making in longitudinal monitoring.

The balance between precision and recall is captured by the Dice coefficient (F1 Score), which is the harmonic mean of the two. A model with high Dice achieves a good balance, while extreme values of either precision or recall at the expense of the other indicate suboptimal segmentation behavior. The precision-recall trade-off is a fundamental consideration in this study: encoders that achieve high Dice scores do so by maintaining both high precision and high recall simultaneously, rather than excelling in one at the expense of the other.

4.5.5 Metric Aggregation Strategy

The strategy for aggregating per-image metrics into summary statistics can significantly influence the interpretation of results. Two common approaches exist: **micro-averaging**, which pools all true positives, false positives, and false negatives across the entire dataset before computing the metric; and **macro-averaging**, which computes the metric independently for each image and then averages the per-image scores.

In this study, all metrics are computed per image and then **macro-averaged across**

the 5-fold cross-validation, providing a robust estimate of the model's generalization performance on unseen patients. Macro-averaging was chosen because it gives equal weight to every image regardless of the tumor size it contains. Under micro-averaging, images with large tumors (many foreground pixels) would dominate the aggregate metric, potentially masking poor performance on images with small tumors. Since small residual tumors are clinically significant in the post-treatment setting—often representing the critical boundary between stable disease and early progression—macro-averaging ensures that model performance on these challenging cases is not diluted by performance on easier, large-tumor images.

The 5-fold patient-wise cross-validation ensures that all slices from a given patient appear in the same fold, preventing data leakage across training and validation sets. This is critical because multiple 2D slices from the same patient share correlated anatomical features, and splitting them across folds would lead to overly optimistic performance estimates that do not reflect the model's ability to generalize to truly unseen patients.

4.5.6 Statistical Considerations for Imbalanced Data

Brain tumor segmentation is an inherently imbalanced classification problem at the pixel level: tumor pixels constitute a small minority of the total image pixels. This class imbalance has important implications for metric interpretation and comparison.

Standard pixel-wise accuracy is an inappropriate metric for this task because a trivial classifier that labels every pixel as background would achieve accuracy exceeding 90%, yet provide zero clinical utility. The four selected metrics—IoU, Dice, Precision, and Recall—are all designed to evaluate performance on the minority (tumor) class specifically, making them robust to the class imbalance problem.

However, even among these metrics, some are more informative than others in different contexts. IoU and Dice provide a comprehensive summary of overlap quality, while Precision and Recall offer diagnostic insight into the nature of errors. A model with high Dice but disproportionately low recall may be acceptable for screening applications but problematic for treatment planning, where complete tumor delineation is essential. Conversely, a model with high recall but low precision may be suitable for clinical workflows that include radiologist review, where over-detection can be easily corrected.

Reporting all four metrics together—as done in this study—provides a complete characterization of each model's segmentation behavior and supports informed decision-making about the most appropriate model for a given clinical application.

Chapter-5: Results and Analysis

This chapter presents the experimental results obtained from both phases of the evaluation framework. All metrics are reported on the foreground (tumor) class only and are macro-averaged across 5-fold patient-wise cross-validation. The results are organized into four sections: encoder backbone comparison (Phase 1), preprocessing strategy evaluation (Phase 2), segmentation visualization, and comparison with existing literature.

5.1 Phase 1: Encoder Backbone Comparison on Raw Images

In the first phase of experiments, all six encoder backbones are trained and evaluated on original (unprocessed) T1-weighted pre-contrast MRI images. This phase aims to identify the best overall encoder by jointly considering segmentation accuracy, model complexity, and inference speed.

5.1.1 Segmentation Performance Analysis

Table 5.1 presents the comprehensive performance comparison of all six encoder backbones trained and tested on original T1pre MRI images.

Table 5.1: Performance Comparison of U-Net Encoder Backbones on Raw T1pre MRI (5-Fold Cross-Validation, Foreground Class Only)

Encoder	IoU	Dice/F1	Precision	Recall	Parameters (M)	Inference Time (ms)
ResNet18	0.59642	0.74690	0.81074	0.69262	14.33	86.99
MiT-B2	0.61452	0.76066	0.81744	0.71288	27.48	146.27
DenseNet121	0.58352	0.73626	0.78900	0.70964	13.61	147.56
MobileNetV2	0.56884	0.72438	0.82564	0.64930	6.63	71.94
ResNet50	0.57328	0.72780	0.82236	0.65920	32.52	150.33
EfficientNet-B3	**0.61796**	**0.76348**	**0.81920**	**0.71560**	**13.16**	**119.60**

IoU and Dice Performance. The results reveal a clear performance hierarchy among the six encoders. EfficientNet-B3 achieved the highest IoU of 0.61796 and Dice coefficient of 0.76348, demonstrating the best overall segmentation accuracy. MiT-B2 followed closely

with an IoU of 0.61452 and Dice of 0.76066, trailing EfficientNet-B3 by only 0.34 percentage points in IoU and 0.28 percentage points in Dice. ResNet18 ranked third with an IoU of 0.59642 and Dice of 0.74690, representing a more substantial gap of 2.15 percentage points in IoU below EfficientNet-B3.

The lower tier of performance was occupied by DenseNet121 (IoU: 0.58352, Dice: 0.73626), ResNet50 (IoU: 0.57328, Dice: 0.72780), and MobileNetV2 (IoU: 0.56884, Dice: 0.72438). The spread between the best-performing encoder (EfficientNet-B3) and the worst-performing encoder (MobileNetV2) was 4.91 percentage points in IoU and 3.91 percentage points in Dice, indicating meaningful performance differences across the architectural paradigms.

Precision Analysis. Interestingly, the precision rankings did not align with the IoU/Dice rankings. MobileNetV2 achieved the highest precision (0.82564), followed by ResNet50 (0.82236), EfficientNet-B3 (0.81920), MiT-B2 (0.81744), ResNet18 (0.81074), and DenseNet-121 (0.78900). The high precision of MobileNetV2 and ResNet50 indicates that when these models predict a pixel as tumor, they are likely to be correct. However, this high precision comes at the cost of low recall, suggesting that these models adopt conservative prediction strategies that miss substantial portions of the actual tumor.

Recall Analysis. The recall rankings revealed a substantially different picture and are arguably more clinically important. EfficientNet-B3 led with a recall of 0.71560, followed closely by MiT-B2 (0.71288), DenseNet121 (0.70964), ResNet18 (0.69262), ResNet50 (0.65920), and MobileNetV2 (0.64930). The gap between the highest recall (EfficientNet-B3: 0.71560) and the lowest recall (MobileNetV2: 0.64930) was 6.63 percentage points—a substantial difference with direct clinical implications.

MobileNetV2's combination of the highest precision (0.82564) with the lowest recall (0.64930) reveals a highly conservative prediction pattern where the model predicts tumor only when it is highly confident, resulting in clean but incomplete segmentations. In clinical brain tumor segmentation, this conservative behavior represents a significant drawback: **under-detection of tumor tissue (false negatives) carries greater clinical risk than over-detection (false positives)**, because missed tumor regions could lead to underestimation of tumor burden and potentially inappropriate treatment decisions. Over-detection, while undesirable, can be more easily identified and corrected through clinical review by neuroradiologists.

Similarly, ResNet50—despite being the second-largest model (32.52M parameters)–achieved only the fifth-highest IoU and the second-lowest recall, indicating that its additional depth and capacity did not translate into better segmentation performance. This finding challenges the common assumption that larger models necessarily perform better and highlights the importance of architectural design principles beyond raw model capacity.

5.1.2 Model Complexity and Inference Time Analysis

Table 5.2 summarizes the model complexity (parameter count) and inference time for all six encoder backbones, providing a joint view of the efficiency characteristics that are critical for clinical deployment decisions.

Table 5.2: Model Complexity and Inference Time Comparison of Encoder Backbones

Encoder	**Parameters (M)**	**Inference Time (ms)**	**Dice Coefficient**
MobileNetV2	6.63	71.94	0.724
EfficientNet-B3	13.16	119.60	0.763
DenseNet121	13.61	147.56	0.736
ResNet18	14.33	86.99	0.747
MiT-B2	27.48	146.27	0.761
ResNet50	32.52	150.33	0.728

The parameter counts span a nearly 5-fold range from MobileNetV2 (6.63M) to ResNet50 (32.52M). Notably, the three models in the 13–14M parameter range (EfficientNet-B3, DenseNet121, ResNet18) exhibit significantly different performance levels, with EfficientNet-B3 achieving the highest Dice (0.763) and DenseNet121 achieving a considerably lower Dice (0.736). This demonstrates that parameter count alone is not a reliable predictor of segmentation performance; the architectural design principles (compound scaling for EfficientNet-B3 vs. dense connectivity for DenseNet121 vs. residual connections for ResNet18) play a more significant role.

ResNet50, despite having 2.5 times the parameters of EfficientNet-B3 (32.52M vs. 13.16M), achieved substantially lower Dice (0.728 vs. 0.763) and IoU (0.573 vs. 0.618). This result underscores that simply increasing model depth and capacity is not sufficient for improving segmentation performance on this task, and may even be counterproductive due to overfitting on the relatively limited training data.

Regarding inference time, MobileNetV2 was the fastest encoder at 71.94 ms, consistent with its design goal of mobile and edge deployment efficiency. However, this speed advantage is offset by its lowest segmentation accuracy, making it unsuitable as a standalone clinical tool for this task. EfficientNet-B3 achieved an inference time of 119.60 ms, which is approximately 66% faster than ResNet50 (150.33 ms) while simultaneously achieving substantially higher accuracy. For clinical deployment, an inference time of approximately 120 ms per slice is well within acceptable limits, as a typical brain MRI volume contains approximately 150–200 axial slices, resulting in a total processing time of approximately 18–24 seconds per volume—fast enough for real-time or near-real-time clinical workflow integration.

DenseNet121 and MiT-B2, despite having substantially fewer parameters than ResNet50, exhibited similar or slower inference times (147.56 ms and 146.27 ms vs. 150.33 ms). For DenseNet121, the dense concatenation operations contribute to memory-intensive feature map processing that slows inference. For MiT-B2, the self-attention computation, while theoretically efficient with the sequence reduction strategy, still involves substantial matrix operations that impact processing speed.

5.1.3 Joint Efficiency-Accuracy Trade-off

Considering all three evaluation dimensions jointly—segmentation accuracy, model complexity, and inference speed—EfficientNet-B3 emerges as the clear best overall encoder for post-treatment glioblastoma segmentation on T1pre MRI.

EfficientNet-B3 achieves the **highest IoU (0.618)** and **highest Dice (0.763)** with the **best recall (0.716)**, while maintaining a **moderate parameter count (13.16M)**—less than half of ResNet50 (32.52M) and MiT-B2 (27.48M)—and a **clinically acceptable inference speed (119.60 ms)** that is faster than three of the five competing encoders.

The compound scaling strategy of EfficientNet-B3 effectively balances depth, width, and resolution, achieving a configuration where:

- The network is deep enough to capture complex hierarchical features relevant to tumor segmentation, but not so deep that optimization becomes difficult or overfitting occurs.
- The network is wide enough to maintain sufficient representational capacity in each layer, but not so wide that parameters are wasted on redundant features.
- The input resolution is sufficient to preserve spatial details important for boundary delineation, while being computationally manageable.

This balanced configuration outperforms both larger models (ResNet50, MiT-B2) that sacrifice efficiency without proportional accuracy gains, and lightweight models (MobileNetV2) that sacrifice accuracy for speed beyond what is clinically necessary.

5.2 Phase 2: Preprocessing Strategy Evaluation

In the second phase, EfficientNet-B3 (identified as the best encoder in Phase 1) is used to evaluate three preprocessing strategies against the baseline of raw image training. Models are trained on preprocessed images and tested on original (unprocessed) images to assess cross-domain robustness.

5.2.1 Cross-Domain Testing Results

Table 5.3 presents the cross-domain results where models trained on preprocessed images are tested on original (unprocessed) images.

IoU and Dice Results. Among the three preprocessing strategies, contrast enhancement (CE) achieved the highest cross-domain performance with an IoU of 0.568 and Dice of 0.723. CLAHE followed with an IoU of 0.535 and Dice of 0.695, while RGB channel stacking performed worst with an IoU of 0.528 and Dice of 0.691.

Table 5.3: Impact of Preprocessing on Segmentation Performance (EfficientNet-B3, Cross-Domain Testing: Trained on Preprocessed, Tested on Original)

Training Input	**IoU**	**Dice**	**Precision**	**Recall**
Original (Baseline)	0.618	0.763	0.819	0.716
CLAHE	0.535	0.695	0.806	0.621
CE	**0.568**	**0.723**	**0.742**	**0.713**
RGB Stack	0.528	0.691	0.785	0.625

Recall Analysis. The recall results provide particularly interesting insights. CE achieved a recall of 0.713, which is remarkably close to the baseline recall of 0.716 (difference of only 0.003). This indicates that the CE-trained model maintains nearly the same sensitivity to tumor regions even when tested on original images, suggesting that contrast enhancement preserves the essential features that the model learns to associate with tumor tissue. In contrast, CLAHE (recall: 0.621) and RGB stacking (recall: 0.625) showed substantially reduced recall, indicating that the features learned from these preprocessing strategies are less transferable to original images.

Precision Analysis. For precision, the ranking reversed relative to the recall and Dice rankings. CLAHE achieved the highest precision (0.806), followed by RGB stacking (0.785) and CE (0.742). The baseline original training achieved 0.819 precision. The higher precision but lower recall of CLAHE and RGB-trained models indicates conservative predictions—these models predict fewer pixels as tumor overall, and while the pixels they do predict are mostly correct, they miss substantial tumor regions. This pattern is clinically undesirable for the same reasons discussed in the precision-recall analysis of Phase 1.

5.2.2 Comparison: Preprocessed vs. Raw Image Training

Comparing the best preprocessing result (CE: IoU 0.568, Dice 0.723) against the baseline (original training: IoU 0.618, Dice 0.763), raw image training outperforms the best preprocessed strategy by:

- **+0.050 IoU** (0.618 vs. 0.568, an improvement of 8.8%)
- **+0.040 Dice** (0.763 vs. 0.723, an improvement of 5.5%)
- **+0.077 Precision** (0.819 vs. 0.742, an improvement of 10.4%)
- **+0.003 Recall** (0.716 vs. 0.713, a marginal improvement of 0.4%)

These results clearly demonstrate that **training directly on original images with EfficientNet-B3 is the superior approach**, offering higher accuracy with the important practical advantages of zero preprocessing overhead, no additional computational cost for preprocessing during deployment, and no risk of domain mismatch degradation.

5.2.3 Domain Mismatch Analysis

The cross-domain degradation—defined as the performance loss when models trained on preprocessed images are tested on original images—provides critical insights for clinical deployment:

- **CE:** IoU dropped by 8.2% (0.618 → 0.568), Dice dropped by 5.3%
- **CLAHE:** IoU dropped by 13.3% (0.618 → 0.535), Dice dropped by 8.9%
- **RGB Stack:** IoU dropped by 14.5% (0.618 → 0.528), Dice dropped by 9.4%

The severity of this degradation varies substantially across preprocessing strategies. CE produces the least degradation, consistent with its more global and less disruptive transformation of the image statistics. CLAHE, which performs local contrast adjustment and can alter the relative intensity relationships between tissues, produces moderate degradation. The RGB stacking approach, which fundamentally changes the input representation from a replicated grayscale to a composite multi-representation image, produces the most severe degradation.

These findings carry a strong practical message: **if preprocessing is applied during training, the same preprocessing must be maintained during deployment to avoid significant performance degradation**. Any inconsistency between training-time and inference-time preprocessing pipelines will introduce a domain shift that degrades segmentation accuracy, potentially by more than 14% in IoU. For clinical deployment, this requirement adds complexity and potential points of failure to the processing pipeline, further supporting the recommendation to train directly on original images.

5.3 Segmentation Visualization

Figure 5.1 presents qualitative segmentation results comparing EfficientNet-B3 trained on original images (top row) and contrast-enhanced images (bottom row). The visualization shows the input MRI slices, ground truth masks, predicted segmentation masks, and overlay comparisons for representative test cases.

The visual comparison illustrates several key observations:

1. **Overall segmentation quality:** The original-trained model produces segmentations that closely match the ground truth in most cases, with good delineation of tumor boundaries and accurate identification of the tumor extent.
2. **Boundary precision:** The original-trained model tends to produce smoother and more coherent tumor boundaries compared to the CE-trained model, which occasionally shows more irregular boundary predictions.

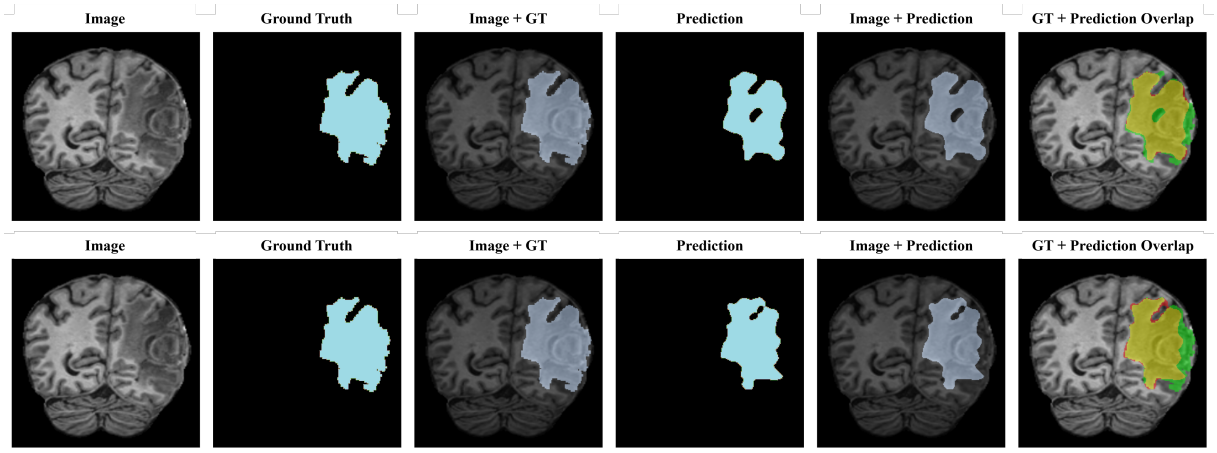


Figure 5.1: Segmentation output visualization comparing EfficientNet-B3 trained on original images (top row) and contrast-enhanced images (bottom row). Each column shows a different test case with the input MRI slice, ground truth mask, predicted mask, and overlay comparison.

3. **Under-segmentation patterns:** Both models show some degree of under-segmentation in regions where the tumor boundary is ambiguous on T1pre images, which is expected given the inherently low contrast between tumor and surrounding tissue on non-contrast T1-weighted sequences.

4. **False positive regions:** The CE-trained model occasionally produces small isolated false positive regions that are not present in the original-trained model's predictions, suggesting that the contrast-enhanced training introduces some sensitivity to non-tumor contrast variations.

5.4 Comparison with Existing Literature

Table 5.4 contextualizes the results of this thesis against related work in brain tumor segmentation, particularly studies focusing on post-treatment glioblastoma.

Table 5.4: Comparison with Existing Post-Treatment GBM Segmentation Methods

Study	**Method**	**MRI Input**	**Dice**	**Dataset**
Hannisdal et al. [35]	nnU-Net	Multi-parametric	0.81–0.82	Institutional
Gagnon et al. [6]	Deep Learning	Multi-shell DSI	–	UCSD-PTGBM
BraTS 2024 [25]	Various	Multi-parametric	Varies	BraTS 2024
Naser & Deen [32]	U-Net + VGG16	T1 + FLAIR	0.84	Pre-treatment
Our Study	**U-Net + EfficientNet-B3**	**T1pre only**	**0.763**	**UCSD-PTGBM**

Our best result (Dice 0.763) is consistent with the performance range observed in the

post-treatment literature. Hannisdal et al. [35] achieved Dice scores of 0.81–0.82 using nnU-Net with multi-parametric MRI (T1, T1-CE, T2, FLAIR), which is a substantially richer input compared to our single T1pre sequence. The 0.05–0.06 Dice gap between our results and Hannisdal et al.'s results is attributable to the fundamental information advantage of multi-parametric MRI, where contrast-enhancing tumor boundaries are explicitly visible on T1-CE and edema is clearly delineated on FLAIR. Additionally, nnU-Net [27] employs automatic self-configuration of preprocessing, architecture topology, and training hyperparameters, which may further optimize performance beyond our fixed experimental setup. Importantly, despite these advantages, Hannisdal et al. did not report model complexity or inference time, leaving the practical deployability of their approach uncharacterized—a gap that our multi-dimensional evaluation explicitly addresses.

Helland et al. [7] investigated deep neural network-based segmentation of glioblastoma in early post-operative multi-modal MRI, reporting that post-treatment segmentation accuracy varies considerably depending on the tumor sub-region and post-surgical timepoint. Their work underscored the difficulty of delineating residual tumor from surgical artifacts and treatment-related changes—challenges that are equally present in the UCSD-PTGBM dataset used in our study. The consistency of these challenges across independent datasets reinforces the validity of our findings and highlights that the performance we achieve with a single non-contrast modality is meaningful given the inherent difficulty of the task.

Gagnon et al. [6] explored deep learning segmentation using multi-shell diffusion spectrum imaging (DSI) on the same UCSD-PTGBM patient cohort. While their study focused on infiltrative and enhancing cellular tumor delineation using advanced diffusion MRI sequences (which capture microstructural tissue properties beyond conventional contrast), it did not report standard Dice scores for direct comparison with conventional segmentation approaches. Nevertheless, their work demonstrates that the UCSD-PTGBM cohort presents a challenging segmentation target regardless of the imaging modality used, further contextualizing our results.

The BraTS 2024 challenge [25] extended its scope to include post-treatment glioma segmentation for the first time, reflecting the growing recognition that follow-up imaging presents distinct challenges from pre-treatment assessment. The challenge utilized multi-parametric MRI (T1, T1-CE, T2, FLAIR) with sub-region annotations including resection cavities, making it a more complex multi-class problem compared to our binary whole-tumor segmentation. While direct numerical comparison with BraTS 2024 is not possible due to differences in task definition, dataset composition, and evaluation protocol, the inclusion of post-treatment cases in the BraTS benchmark validates the clinical relevance and timeliness of our research focus.

Naser and Deen [32] achieved a higher Dice of 0.84, but their study used a combination of T1 and FLAIR sequences on pre-treatment (de novo) tumors, which present clearer tumor boundaries compared to the complex post-treatment imaging landscape of our study. The approximately 0.08 Dice gap is consistent with the general finding in the literature that post-treatment segmentation is inherently more challenging than pre-treatment segmentation: treatment-induced changes such as surgical cavities, radiation necrosis, pseudoprogression, and blood products create ambiguous tissue boundaries that are absent in de novo tumors. Moreover, the addition of FLAIR provides explicit sensitivity to peri-tumoral edema that is not available in our T1pre-only pipeline.

Zeineldin et al. [30] proposed TransXAI, a hybrid CNN-Transformer architecture achieving competitive accuracy on multi-modal glioma segmentation with explainability features. Similarly, Zhang et al. [31] introduced DeepGlioSeg with region-based weighting for advanced glioma MRI segmentation. Both studies relied on multi-parametric MRI input and focused on pre-treatment cases, and neither reported model size or inference time. These architectural innovations may offer accuracy benefits, but without efficiency profiling, their suitability for resource-constrained clinical deployment remains uncharacterized—a limitation our study specifically addresses through joint accuracy-efficiency evaluation.

Beyond accuracy alone, a critical differentiator of this study is the explicit characterization of **model complexity and inference efficiency**. While all comparison studies in Table 5.4 report segmentation accuracy, none provide a joint analysis of parameter count, inference latency, and segmentation quality. Our finding that EfficientNet-B3 achieves the best accuracy with only 13.16M parameters and 119.60 ms inference time—outperforming the 2.5× larger ResNet50—provides actionable guidance for deployment that is absent from the existing literature. This multi-dimensional perspective aligns with the call by Dorfner et al. [9] for more holistic evaluation frameworks in medical image segmentation research.

It is also worth noting that our study operates under a significantly more constrained data regime than the studies listed in Table 5.4. The UCSD-PTGBM dataset contains 77 patients, from which 2D axial slices are extracted—a substantially smaller data pool compared to the BraTS benchmarks (2,000+ cases in BraTS 2021) or proprietary institutional datasets used by Hannisdal et al. Despite this data limitation, EfficientNet-B3 achieves competitive Dice scores, suggesting that the combination of ImageNet pre-training and compound-scaled architecture is particularly effective in low-data medical imaging regimes. Furthermore, the reliance on a single non-contrast MRI sequence makes this pipeline applicable in clinical scenarios where contrast agent administration is contraindicated—such as patients with renal insufficiency, contrast allergy, or during pregnancy—broadening the potential clinical utility beyond what multi-modal approaches can offer.

These comparisons collectively demonstrate that our EfficientNet-B3 result represents a strong and competitive baseline for single-modality post-treatment GBM segmentation. The achieved Dice of 0.763 on T1pre-only input occupies a reasonable position relative to studies using richer multi-modal inputs, and provides a lower-bound estimate of achievable performance that can guide the development of more advanced approaches incorporating additional modalities, architectural innovations, and domain-specific optimization strategies.

Chapter-6: Discussion

This chapter provides an in-depth discussion of the experimental findings, interpreting the results in the context of architectural design principles, the nature of post-treatment glioblastoma imaging, and the practical implications for clinical deployment.

6.1 Why EfficientNet-B3 Outperforms Other Encoders

The superior performance of EfficientNet-B3 across all primary segmentation metrics (IoU, Dice, and Recall) can be attributed to several interconnected factors related to its architectural design and the specific characteristics of the post-treatment glioblastoma segmentation task.

Compound scaling advantage. The fundamental insight of EfficientNet is that network depth, width, and resolution are interdependent dimensions that should be scaled together rather than independently. When ResNet50 scales only depth (from 18 to 50 layers) while keeping width and resolution fixed, the additional layers provide diminishing returns because the increased depth is not matched by corresponding increases in width and resolution needed to fully exploit the additional feature extraction capacity. EfficientNet-B3's compound scaling ensures that all three dimensions grow proportionally, creating a balanced architecture where each layer has sufficient width to capture diverse features, the depth is adequate for hierarchical feature learning, and the resolution preserves spatial details critical for precise boundary delineation.

Squeeze-and-excitation attention. The SE modules in EfficientNet-B3's MBConv blocks provide an implicit attention mechanism that adaptively emphasizes the most informative feature channels while suppressing less relevant ones. For brain tumor segmentation, where the discriminative features may be distributed across a subset of channels depending on the local tissue context, this channel-wise attention can help the model focus on the most relevant representations at each spatial location.

Optimal accuracy-efficiency balance. EfficientNet-B3 occupies a sweet spot in the accuracy-efficiency space for this particular task. The post-treatment glioblastoma segmentation problem, while challenging, is a binary classification task at each pixel (tumor vs. background) that may not require the extreme depth and capacity of ResNet50 (32.52M parameters) or the global self-attention of MiT-B2 (27.48M parameters). EfficientNet-B3's 13.16M parameters provide sufficient capacity to capture the relevant features without the overfitting risk or optimization difficulty associated with larger models trained on

relatively limited medical imaging data.

6.2 The Role of Compound Scaling in Medical Image Segmentation

The success of EfficientNet-B3 validates the compound scaling principle for medical image segmentation and has broader implications for architecture selection in this domain.

Traditional approaches to improving model performance in medical imaging have often focused on increasing model depth (e.g., moving from ResNet18 to ResNet50 to ResNet101) under the assumption that deeper models can capture more complex features. Our results challenge this assumption: ResNet50 (32.52M parameters, 50 layers) performed substantially worse than both EfficientNet-B3 (13.16M parameters) and even the much shallower ResNet18 (14.33M parameters, 18 layers) in terms of IoU and Dice.

This finding suggests that for medical image segmentation tasks with limited training data, the optimal model capacity exists at a lower point than what is typically assumed from natural image tasks. The ImageNet dataset contains approximately 1.2 million training images across 1,000 categories, providing ample data to train very deep networks. In contrast, medical imaging datasets are typically orders of magnitude smaller, and the increased capacity of deeper models may lead to overfitting rather than improved generalization.

The compound scaling approach offers a principled methodology for finding this optimal capacity: by scaling depth, width, and resolution together, it avoids the suboptimal configurations that arise from scaling any single dimension to excess. This finding has practical implications for future medical image segmentation studies, suggesting that compound-scaled architectures should be considered as strong baselines before exploring more complex or larger alternatives.

6.3 Why Preprocessing-Enhanced Training Underperforms

The finding that all three preprocessing strategies degraded performance when models were tested on original images—with degradation ranging from 8.2% (CE) to 14.5% (RGB stacking) in IoU—requires careful interpretation in the context of both the preprocessing methods and the nature of ImageNet pre-training.

Feature distribution shift. Each preprocessing strategy modifies the statistical properties of the input images. CLAHE alters local intensity distributions and can change the relative contrast between tissues in a spatially non-uniform manner. Contrast enhancement linearly stretches the global intensity range. RGB stacking creates a fundamentally different input representation where each channel contains a different view of the same anatomical information. When models trained on these modified distributions are tested

on original images, the input statistics do not match what the model has learned to process, leading to activation patterns that deviate from the optimal operating point and result in degraded predictions.

ImageNet pre-training redundancy. A key factor may be that ImageNet pre-training already provides sufficient low-level feature adaptation for the task. The early layers of ImageNet pre-trained encoders have learned general-purpose feature detectors (edges, textures, gradients, color patterns) that are applicable to a wide range of visual tasks [15]. These pre-learned features may already capture the visual information that preprocessing strategies aim to enhance, making the additional preprocessing redundant or even counterproductive when it shifts the input away from the natural image statistics that the encoder was originally trained on.

Post-treatment imaging heterogeneity. The limited benefit of preprocessing in the post-treatment setting contrasts with pre-treatment studies where CLAHE has shown improvements [10]. This discrepancy may reflect the unique imaging characteristics of post-treatment glioblastoma. Pre-treatment tumors typically present with relatively uniform contrast patterns that can be systematically enhanced by preprocessing. In contrast, post-treatment imaging exhibits highly heterogeneous intensity distributions due to surgical cavities, blood products, radiation effects, and pseudoprogression, which create a complex mosaic of intensity patterns that is not uniformly improved by any single preprocessing strategy. The heterogeneity may even be exacerbated by preprocessing, as CLAHE and CE can amplify treatment-related artifacts alongside tumor features.

RGB stacking limitations. The particularly poor performance of RGB stacking (worst among all strategies) challenges the intuition that providing more information through additional channels should improve performance. Unlike true multi-modal MRI stacking, where each modality provides physically distinct information (T1 contrast, T2 contrast, FLAIR sensitivity), the three channels of the RGB stack contain highly correlated information—they are all derived from the same underlying MRI slice through different intensity transformations. The model may struggle to learn useful channel-specific features from these correlated representations, and the composite RGB input may confuse the early convolutional filters that were pre-trained to expect the color distributions of natural RGB images.

6.4 Clinical Implications

The findings of this thesis have several direct implications for the clinical deployment of deep learning-based segmentation tools for post-treatment glioblastoma:

Pipeline simplicity. The recommendation to train on original images with EfficientNet-B3 translates to the simplest possible deployment pipeline: raw MRI images can be directly fed into the model without any intermediate preprocessing steps. This simplicity reduces the number of potential failure points in the clinical workflow, eliminates the computational overhead of preprocessing, and avoids the need to maintain and validate preprocessing software alongside the segmentation model.

Inference speed feasibility. EfficientNet-B3's inference time of 119.60 ms per slice enables processing of a complete brain MRI volume (approximately 150–200 axial slices) in approximately 18–24 seconds. This processing speed is compatible with clinical workflows where real-time feedback during image review is valuable but not strictly required. For batch processing scenarios (e.g., processing multiple patients' scans overnight), the throughput would allow processing of approximately 150–200 patients per hour, which is sufficient for most clinical centers.

Resource-constrained deployment. EfficientNet-B3's moderate parameter count (13.16M) makes it deployable on standard clinical computing hardware without requiring specialized GPU servers. This is particularly important for hospitals and clinics in low- and middle-income countries, where access to high-end computing infrastructure may be limited but the clinical need for automated tumor monitoring tools is equally pressing.

Single-modality robustness. The demonstration that meaningful segmentation accuracy (Dice 0.763) can be achieved using only T1-weighted pre-contrast images is clinically significant because T1pre is the most universally available MRI sequence across clinical settings. In situations where the full multi-parametric MRI protocol is not available (due to scanner limitations, time constraints, or patient-related factors such as contrast agent contraindications), the T1pre-only model can still provide useful tumor segmentation.

Preprocessing consistency requirement. The severe cross-domain degradation observed when preprocessing is inconsistent between training and deployment is an important practical warning. Clinical deployment of any preprocessing-dependent model requires rigorous validation that the preprocessing pipeline is identical to the one used during training, including all parameter settings. Any deviation—intentional or accidental—can result in clinically significant performance degradation.

6.5 Performance in Context of Post-Treatment Literature

Positioning our results within the broader post-treatment glioblastoma literature provides important context for interpreting the achieved performance level. Our best Dice coefficient of 0.763 using only T1pre input is consistent with the performance range observed across different studies and modalities.

Hannisdal et al. [35] achieved median Dice scores of 0.81–0.82 using nnU-Net with the full multi-parametric MRI protocol (T1, T1-CE, T2, FLAIR). The approximately 0.05–0.06 Dice gap between our results and theirs can be attributed primarily to the information advantage of multi-parametric MRI, particularly the T1-CE sequence where contrast-enhancing tumor boundaries are explicitly visible through gadolinium uptake. Additionally, nnU-Net incorporates automatic configuration of preprocessing, architecture, and training hyperparameters, which may provide further optimization beyond our fixed hyperparameter configuration.

The comparison with Naser and Deen [32], who achieved Dice of 0.84 using U-Net + VGG16 on T1 and FLAIR sequences for pre-treatment glioma, highlights the additional challenge

of post-treatment segmentation. Pre-treatment tumors present with clearer tissue contrast and more regular boundaries compared to post-treatment tumors that are surrounded by surgical changes, radiation effects, and treatment-related artifacts. The approximately 0.08 Dice gap between our post-treatment results and their pre-treatment results is consistent with the general finding in the literature that post-treatment segmentation is more challenging than pre-treatment segmentation.

The limited benefit of preprocessing in our study also contrasts with findings in other medical imaging domains. For instance, Kumari et al. [10] reported substantial improvements from CLAHE-based preprocessing for glioma detection (achieving 99.1% accuracy). However, their study focused on tumor detection (a classification task) rather than precise segmentation, and used pre-treatment data where tumor contrast patterns are more uniform and consistently enhanced by CLAHE. These contextual differences highlight the importance of evaluating preprocessing strategies specifically for the target task (post-treatment segmentation) rather than extrapolating results from related but different applications.

Overall, our results establish a strong and reproducible baseline for single-modality post-treatment glioblastoma segmentation, and the multi-dimensional evaluation framework (accuracy, efficiency, robustness) provides practical guidance that goes beyond what is typically available from studies that report only accuracy metrics.

6.6 Addressing the Research Questions

This section consolidates the experimental findings by explicitly addressing each of the four research objectives stated in Chapter 1, ensuring that every objective is answered with evidence from the experimental results.

Research Objective 1: Multi-Dimensional Encoder Benchmarking. The first objective was to compare six encoder backbone architectures—ResNet18, ResNet50, DenseNet121, MobileNetV2, MiT-B2, and EfficientNet-B3—within the U-Net framework across three dimensions: segmentation accuracy, model complexity, and computational cost. This objective was comprehensively addressed in Phase 1 of the experiments (Table 5.1). The evaluation revealed that EfficientNet-B3 achieves the highest Dice coefficient (0.763) and IoU (0.618) while maintaining a moderate parameter count (13.16M) and clinically acceptable inference time (119.60 ms). Critically, the multi-dimensional profiling uncovered insights that would not have emerged from accuracy-only evaluation: ResNet50, the largest model (32.52M parameters), performed worse than the 2.5× smaller EfficientNet-B3, demonstrating that model size is not a reliable proxy for segmentation quality and validating the importance of jointly assessing all three dimensions. The success of compound scaling (EfficientNet-B3) over depth-only scaling (ResNet50), dense connectivity (DenseNet121), aggressive compression (MobileNetV2), and self-attention (MiT-B2) provides actionable guidance for architecture selection in medical image segmentation under real-world constraints.

Research Objective 2: Preprocessing Impact Investigation. The second objective was to examine how three preprocessing strategies—CLAHE, contrast enhancement (CE), and RGB channel stacking—affect segmentation quality relative to raw image training, and

whether multi-representation inputs offer an advantage over single-channel training. Phase 2 experiments (Table 5.3) provided a definitive answer: training directly on original images with EfficientNet-B3 outperformed all three preprocessing strategies across all primary metrics, with advantages of +0.050 IoU and +0.040 Dice over the best preprocessing strategy (CE). The RGB channel stacking approach, despite combining three complementary representations of the same MRI slice, performed worst among the strategies (IoU: 0.528, Dice: 0.691), demonstrating that multi-representation inputs derived from preprocessing variants of a single modality do not provide the complementary information benefit observed in true multi-modal MRI stacking. The finding that ImageNet pre-trained encoders already capture sufficient low-level features for the task, rendering additional preprocessing redundant or counterproductive, is a novel contribution with implications beyond this specific application.

Research Objective 3: Domain Robustness Evaluation. The third objective was to measure how preprocessing-trained models degrade when inference is performed on unprocessed images, quantifying the domain mismatch risk inherent in preprocessing-dependent pipelines. The cross-domain testing protocol—where models trained on preprocessed images were evaluated on original images—directly addressed this objective and yielded one of the most practically important findings of this thesis. Domain mismatch degradation ranged from 8.2% IoU loss for contrast enhancement, to 13.3% for CLAHE, and up to 14.5% for RGB stacking. This gradient of degradation correlates with the degree to which each preprocessing strategy alters the input distribution: CE applies a relatively mild global intensity stretching that partially preserves natural image statistics, while CLAHE applies locally non-uniform transformations, and RGB stacking fundamentally restructures the input representation. These results establish that preprocessing choices create persistent feature distribution dependencies that cannot be ignored during deployment, and they carry a clear practical message: if preprocessing is used during training, it must be strictly replicated during inference, or performance degradation of clinical significance will result.

Research Objective 4: Deployment Suitability Analysis. The fourth objective was to synthesize findings across accuracy, efficiency, model size, and robustness to identify the most suitable encoder–training configuration for clinical deployment. Combining the evidence from all three preceding objectives yields a clear and actionable recommendation: EfficientNet-B3 trained on raw (unprocessed) T1-weighted pre-contrast images is the optimal configuration for post-treatment glioblastoma segmentation. This recommendation rests on four converging lines of evidence: (1) it achieves the highest segmentation accuracy among all tested configurations (Dice: 0.763, IoU: 0.618); (2) its moderate parameter count (13.16M) enables deployment on standard clinical computing hardware without specialized GPU infrastructure; (3) its inference speed (119.60 ms per slice, approximately 18–24 seconds per volume) is compatible with clinical workflows; and (4) by eliminating preprocessing, it avoids the domain mismatch risks (up to 14.5% IoU degradation) that would compromise reliability in real-world deployment. This synthesis provides the holistic, evidence-based deployment guidance that Dorfner et al. [9] identified as lacking in the current literature.

Chapter-7: Conclusion

This chapter summarizes the key findings of this thesis, acknowledges the limitations of the study, and outlines directions for future research.

7.1 Summary of Findings

This thesis presented a systematic and comprehensive evaluation of six encoder backbones within the U-Net architecture for binary whole-tumor segmentation of post-treatment glioblastoma on T1-weighted pre-contrast MRI from the UCSD-PTGBM dataset. The study was conducted in two phases, yielding several important findings:

Finding 1: EfficientNet-B3 is the optimal encoder for post-treatment GBM segmentation. Among the six encoders evaluated—ResNet18, ResNet50, DenseNet121, MobileNetV2, MiT-B2, and EfficientNet-B3—EfficientNet-B3 achieved the best overall performance across all primary segmentation metrics, with the highest Dice coefficient of 0.76348 and IoU of 0.61796. Critically, EfficientNet-B3 accomplished this with a moderate parameter count of 13.16 million (less than half of ResNet50's 32.52M and MiT-B2's 27.48M) and a clinically acceptable inference time of 119.60 ms per slice. The compound scaling strategy of EfficientNet-B3 effectively balances depth, width, and resolution, validating this design principle for medical image segmentation applications.

Finding 2: Larger models do not guarantee better segmentation. ResNet50, despite being the largest model evaluated (32.52M parameters), achieved only the fifth-highest IoU (0.573) and the second-lowest recall (0.659), performing substantially worse than EfficientNet-B3 and even the much shallower ResNet18 (14.33M parameters, IoU: 0.596). This finding challenges the assumption that increased model capacity leads to proportionally better performance and highlights the importance of architectural design principles over raw model size.

Finding 3: Lightweight models sacrifice critical recall. MobileNetV2, while being the smallest (6.63M parameters) and fastest (71.94 ms) model, achieved the lowest IoU (0.569) and, more importantly, the lowest recall (0.649). Its high precision (0.826) coupled with low recall indicates a conservative prediction strategy that misses substantial tumor regions—a clinically unacceptable behavior where under-detection carries greater risk than over-detection.

Finding 4: Raw image training outperforms all preprocessing strategies. Direct

training on original images with EfficientNet-B3 (Dice: 0.763, IoU: 0.618) outperformed the best preprocessing strategy, contrast enhancement, by 0.040 in Dice and 0.050 in IoU when tested on original images. RGB channel stacking, despite combining three representations, showed no meaningful advantage and performed worst among preprocessing strategies.

Finding 5: Preprocessing inconsistency causes severe performance degradation. Cross-domain testing revealed that preprocessing fundamentally shifts learned feature distributions, causing IoU degradation of 8.2% (CE), 13.3% (CLAHE), and 14.5% (RGB stacking) when preprocessing is applied during training but absent at inference. This finding has critical practical implications: if preprocessing is used during training, it must be strictly maintained during deployment.

Finding 6: Simple is best for clinical deployment. Combining Findings 1, 4, and 5, the most efficient and reliable clinical pipeline is EfficientNet-B3 trained directly on original T1pre images. This approach achieves the highest accuracy, requires no preprocessing overhead, eliminates domain mismatch risks, and operates within clinically acceptable inference time constraints.

7.2 Analysis of the Social Effects

The development and potential deployment of automated post-treatment glioblastoma segmentation tools carry significant social implications that extend beyond purely technical considerations.

Democratizing access to expert-level neuro-oncology care. Accurate tumor segmentation currently requires experienced neuroradiologists, a scarce resource concentrated in well-resourced academic medical centers. The finding that EfficientNet-B3 achieves clinically meaningful segmentation accuracy (Dice: 0.763) with a moderate computational footprint (13.16M parameters, 119.60 ms inference per slice) suggests that such tools could be deployed in community hospitals and clinics in low- and middle-income countries where specialist availability is limited. By providing automated tumor delineation as a decision-support tool, these models have the potential to reduce disparities in neuro-oncology care quality between well-resourced and under-resourced settings.

Reducing clinician burden and burnout. Manual tumor delineation requires 30–60 minutes per patient scan and must be repeated at each follow-up imaging timepoint (typically every 2–3 months) throughout a patient's disease course [7]. Automated segmentation that processes an entire brain volume in approximately 18–24 seconds can substantially reduce this workload, allowing neuroradiologists to allocate their time and expertise to higher-order clinical interpretation and decision-making rather than labor-intensive boundary tracing. This workload reduction is socially significant given the documented prevalence of burnout among radiologists and the growing demand for imaging services.

Enabling equitable single-modality care. The demonstration that meaningful segmentation can be achieved using only T1-weighted pre-contrast MRI—the most universally available and least expensive MRI sequence—has important equity implications. Patients who cannot receive gadolinium contrast agents (e.g., those with renal insufficiency, allergic

reactions, or pregnancy) or who are imaged at facilities lacking advanced MRI capabilities can still benefit from automated tumor monitoring. This single-modality capability removes a barrier that would otherwise exclude vulnerable patient populations from the benefits of AI-assisted care.

Trust and adoption challenges. Successful clinical deployment requires building trust among clinicians, patients, and healthcare administrators. The relatively moderate Dice coefficient of 0.763, while a strong baseline for single-modality post-treatment segmentation, may raise concerns about reliability in individual cases where segmentation errors could influence treatment decisions. Transparent communication about model capabilities and limitations, combined with human-in-the-loop validation workflows where clinicians review and can correct automated segmentations, will be essential for responsible adoption. Over-reliance on automated tools without adequate clinical oversight could lead to errors that erode both clinician competence and patient confidence in AI-assisted care.

7.3 Environmental Impact Analysis and Sustainability

The environmental footprint of deep learning research has become an increasingly important consideration in the AI community. This section analyzes the environmental implications of the research conducted in this thesis and the sustainability advantages of the recommended pipeline.

Computational cost of model training. Training six encoder backbones across 5-fold cross-validation, each for up to 100 epochs, represents a substantial computational investment. The parameter counts of the evaluated models span a nearly 5-fold range from MobileNetV2 (6.63M) to ResNet50 (32.52M), and training larger models generally requires more energy due to increased memory usage, longer per-epoch computation, and potentially more epochs to converge. Strubell et al. [48] demonstrated that training large deep learning models can produce carbon emissions comparable to the lifetime emissions of an automobile, highlighting the environmental cost of extensive model experimentation.

Efficiency-oriented architecture selection as Green AI. The core finding of this thesis—that EfficientNet-B3 achieves the best accuracy with a moderate parameter count (13.16M), outperforming both larger models (ResNet50 at 32.52M, MiT-B2 at 27.48M)—directly aligns with the "Green AI" paradigm advocated by Schwartz et al. [49]. Green AI emphasizes research that achieves strong results through computationally efficient methods rather than through brute-force scaling. By demonstrating that compound scaling achieves superior performance with fewer parameters and faster inference than simply increasing model depth or using attention-heavy architectures, this thesis provides evidence that efficiency-oriented design principles can benefit both accuracy and environmental sustainability.

Elimination of preprocessing overhead. The finding that raw image training outperforms all preprocessing strategies has a direct sustainability benefit: it eliminates the computational cost of applying CLAHE, contrast enhancement, or RGB channel stacking to every image during both training and inference. While individual preprocessing operations are computationally inexpensive, their cumulative cost across large-scale training

datasets, 5-fold cross-validation, and continuous clinical deployment is non-trivial. The recommended pipeline (EfficientNet-B3 on raw images) achieves the highest accuracy with zero preprocessing energy expenditure.

Transfer learning as a sustainable practice. The use of ImageNet pre-trained weights for all encoder backbones represents a sustainable practice that avoids the substantial computational cost of training from scratch. The pre-trained weights, computed once on the large-scale ImageNet dataset, are reused across countless downstream tasks, amortizing their computational cost. Furthermore, the use of the publicly available UCSD-PTGBM dataset [11, 12] from The Cancer Imaging Archive promotes reproducibility and eliminates the need for duplicate data collection efforts, reducing both the environmental and financial costs of medical imaging research.

7.4 Ethical Issues

The development and deployment of AI-based tumor segmentation tools raise several ethical considerations that must be addressed to ensure responsible and beneficial use in clinical practice.

Patient data privacy and consent. The UCSD-PTGBM dataset used in this research is fully de-identified and publicly available through The Cancer Imaging Archive (TCIA), in compliance with established data sharing standards for medical imaging research. However, deployment of segmentation models on clinical data requires strict adherence to patient privacy regulations, including the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in the European Union. Institutional ethics board approval must be obtained before any clinical deployment, and patients should be informed when AI tools are used in their care pathway [50].

Algorithmic bias and fairness. The UCSD-PTGBM dataset originates from a single institution (University of California San Diego) with a specific demographic profile: mean age of 56 $\pm$ 13 years and a male-to-female ratio of approximately 1.87:1 [11]. Models trained on this dataset may not generalize equitably across different ethnic, racial, age, or socioeconomic groups whose tumor imaging characteristics may differ due to biological variability or differences in imaging equipment and protocols [51]. The absence of multi-institutional validation in this thesis (acknowledged as a limitation) means that the extent of any demographic bias in the model's performance remains unknown. Before clinical deployment, comprehensive fairness audits across diverse patient populations are essential to identify and mitigate potential biases that could lead to disparate quality of care.

Clinical decision accountability. When AI-generated segmentation influences treatment decisions—such as determining tumor volume for response assessment, planning radiation therapy fields, or deciding whether imaging changes represent true progression versus pseudoprogression—the question of accountability for errors becomes critical. The World Health Organization has emphasized that AI tools in healthcare should function as decision-support systems that augment, rather than replace, clinical judgment [52]. The segmentation output should be presented to clinicians as a preliminary delineation

subject to expert review and modification, with the treating physician retaining ultimate responsibility for clinical decisions.

Risk of under-detection and clinical safety. The experimental results revealed that MobileNetV2, despite being the fastest and most lightweight model, achieved the lowest recall of 0.649, meaning that approximately 35% of actual tumor pixels were missed in the segmentation output. Deploying such a model without adequate safeguards could result in systematic under-estimation of tumor burden, potentially leading to under-treatment or premature discontinuation of therapy. This finding raises the ethical imperative to establish minimum performance thresholds—particularly for recall—before any segmentation model is approved for clinical use. The recommended EfficientNet-B3 configuration, with its highest recall of 0.716, represents a more acceptable trade-off but still requires careful clinical validation to ensure that the remaining 28.4% of missed tumor pixels do not systematically correspond to clinically significant regions.

Transparency and explainability. The "black box" nature of deep learning models can undermine clinical trust and hinder adoption [51]. Clinicians need to understand not only what the model predicts but also why specific regions are identified as tumor. Integrating explainability techniques such as Grad-CAM or attention visualization, as demonstrated by Zeineldin et al. [30] for brain tumor segmentation, would enhance transparency and enable clinicians to identify potential failure modes. Without such transparency, there is a risk that clinicians may either uncritically accept erroneous predictions or dismiss accurate predictions, both of which compromise patient care.

Equitable deployment considerations. While this thesis highlights the potential of lightweight, efficient models to democratize access to automated tumor monitoring, there is an ethical responsibility to ensure that deployment does not exacerbate existing healthcare inequalities. If AI tools are deployed only in well-resourced settings while under-resourced settings lack the infrastructure, training, or regulatory frameworks for adoption, the technology could widen rather than narrow the quality-of-care gap. Responsible deployment requires parallel investment in the computational infrastructure, clinical training, and regulatory capacity needed to support AI-assisted care in diverse healthcare settings [52, 50].

7.5 Limitations

This research faces several limitations that should be acknowledged and considered when interpreting the results:

Single MRI modality. Only T1-weighted pre-contrast (T1pre) MRI was used as input, whereas clinical practice typically employs multi-parametric MRI including T1, T1-CE, T2, and FLAIR sequences. Multi-modal approaches have been shown to achieve superior segmentation accuracy [37, 35] because different sequences provide complementary information about different tumor compartments. The results in this thesis represent a lower-bound performance achievable from a single non-contrast sequence, and the true clinical potential of these encoder architectures may be higher with multi-modal input.

Binary segmentation only. The segmentation task was limited to binary whole-tumor delineation (tumor vs. background), without sub-region differentiation. In clinical practice, distinguishing between enhancing tumor, non-enhancing/necrotic core, peritumoral edema, and resection cavity carries different prognostic and therapeutic implications. The binary formulation, while sufficient for comparing encoders and preprocessing strategies, does not provide the sub-regional information that may be needed for comprehensive clinical assessment.

Single-institution dataset. All experiments were conducted on the UCSD-PTGBM dataset from a single institution (University of California San Diego) using consistent 3T MRI scanners and acquisition protocols. The generalizability of the findings to data from different institutions, scanner manufacturers, field strengths (e.g., 1.5T), and acquisition protocols has not been evaluated. Multi-institutional studies have shown that domain shift across different centers can significantly affect model performance [36].

No data augmentation. No data augmentation techniques were applied during training. Data augmentation strategies such as random rotations, flipping, elastic deformations, intensity variations, and random cropping are widely used in medical image segmentation to improve model generalization, particularly when training data is limited. This may have limited the ability of the models to generalize across different tumor appearances, sizes, and orientations, and may partially explain the relatively moderate Dice scores compared to state-of-the-art approaches that employ extensive augmentation.

2D slice-based processing. The segmentation was performed on individual 2D axial slices rather than 3D volumetric processing. While 2D processing is computationally efficient and allows the use of ImageNet pre-trained encoders, it does not capture inter-slice spatial context that could improve segmentation consistency across the volume. Adjacent slices contain correlated anatomical information, and 3D approaches can leverage this information for more consistent and accurate volumetric segmentation.

7.6 Future Work

Based on the findings and limitations of this thesis, several promising directions for future research are identified:

Multi-modal input integration. Extending the framework to incorporate multi-parametric MRI input (T1-CE, T2, FLAIR) is the most immediate direction for improving segmentation accuracy. The encoder comparison framework established in this thesis can be applied to multi-modal inputs to determine whether the relative performance ranking of encoder backbones changes with richer input information. Multi-modal fusion strategies (early fusion, late fusion, attention-based fusion) can be systematically evaluated.

Multi-class sub-region segmentation. Extending from binary whole-tumor segmentation to multi-class BraTS sub-region segmentation (enhancing tumor, non-enhancing/necrotic core, surrounding FLAIR abnormality, resection cavity) would provide more clinically precise predictions. The sub-regional information is essential for comprehensive treatment response assessment and can be directly mapped to the RANO response criteria used in

clinical practice.

External multi-institutional validation. Conducting external validation on datasets from different institutions, scanner manufacturers, and acquisition protocols is essential for assessing the generalizability of the findings. Multi-center validation studies can identify potential domain shift issues and guide the development of domain adaptation strategies for robust cross-institutional deployment.

Advanced architectures. Evaluating state-of-the-art architectures such as nnU-Net (which automatically optimizes architecture and preprocessing for each task), Swin-UNet (which uses shifted window attention for efficient hierarchical feature extraction), and other recent innovations could provide further improvements. These architectures incorporate design principles that address some of the limitations identified in this study.

Systematic data augmentation. Implementing and evaluating systematic data augmentation strategies could improve model generalization and potentially boost performance for all encoder backbones. The interaction between data augmentation and encoder architecture could provide additional insights into the optimal training configuration for post-treatment brain tumor segmentation.

3D volumetric segmentation. Transitioning from 2D slice-based to 3D volumetric segmentation approaches (3D U-Net, V-Net, nnU-Net) would enable the model to leverage inter-slice spatial context for more consistent and accurate volumetric tumor segmentation. This extension would require different encoder architectures designed for 3D input and would not directly benefit from 2D ImageNet pre-training, representing a fundamentally different approach.

Clinical deployment and prospective validation. The ultimate goal is to deploy the segmentation model in a clinical environment and conduct prospective validation with neuroradiologists. Prospective studies would evaluate the model's performance in real-world clinical conditions, assess its impact on clinical workflow efficiency, and measure its agreement with expert segmentation in a controlled clinical setting. Such studies would bridge the gap between research performance and real-world clinical utility, which Dorfner et al. [9] identified as a key remaining challenge in the field.

Explainability and interpretability. Integrating explainable AI techniques (e.g., Grad-CAM, attention visualization) would provide insights into which image regions and features drive the model's predictions, enhancing clinical trust and facilitating the identification of potential failure modes. Zeineldin et al. [30] have demonstrated the value of such approaches for brain tumor segmentation.

Bibliography

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*. Springer, 2015, pp. 234–241.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.

[3] Q. T. Ostrom, M. Price, C. Neff, G. Cioffi, K. A. Waite, C. Kruchko, and J. S. Barnholtz-Sloan, "CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2015–2019," *Neuro-Oncology*, vol. 24, no. Suppl. 5, pp. v1–v95, 2022.

[4] R. Stupp, W. P. Mason, M. J. van den Bent *et al.*, "Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma," *New England Journal of Medicine*, vol. 352, no. 10, pp. 987–996, 2005.

[5] B. H. Menze, A. Jakab, S. Bauer *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[6] L. Gagnon, J. D. Rudie, T. M. Seibert *et al.*, "Deep learning segmentation of infiltrative and enhancing cellular tumor at pre- and posttreatment multishell diffusion MRI of glioblastoma," *Radiology: Artificial Intelligence*, vol. 6, no. 5, p. e230489, 2024.

[7] R. H. Helland, A. Ferles, A. Pedersen *et al.*, "Segmentation of glioblastomas in early post-operative multi-modal MRI with deep neural networks," *Scientific Reports*, vol. 13, p. 18897, 2023.

[8] U. Baid, S. Ghodasara, S. Mohan *et al.*, "The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv preprint arXiv:2107.02314*, 2021.

[9] F. J. Dorfner *et al.*, "A review of deep learning for brain tumor analysis in MRI," *npj Precision Oncology*, vol. 9, p. 2, 2025.

[10] A. Kumari *et al.*, "Enhanced glioma tumor detection and segmentation using modified deep learning with edge fusion and frequency features," *Scientific Reports*, vol. 15, p. 5281, 2025.

[11] L. Gagnon, D. Gupta, U. Nguyen *et al.*, "The University of California San Diego post-treatment glioblastoma (UCSD-PTGBM) annotated multimodal MRI dataset," *Scientific Data*, vol. 13, no. 1, 2026.

[12] L. Gagnon, D. Gupta, G. Mastorakos *et al.*, "The University of California San Diego annotated post-treatment high-grade glioma multimodal MRI dataset (UCSD-PTGBM)," Version 3 [dataset], The Cancer Imaging Archive, 2025.

[13] N. Gordillo, E. Montseny, and P. Sobrevilla, "State of the art survey on MRI brain tumor segmentation," *Magnetic Resonance Imaging*, vol. 31, no. 8, pp. 1426–1438, 2013.

[14] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[15] G. Litjens, T. Kooi, B. E. Bejnordi *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[16] N. J. Tustison, K. Shrinidhi, M. Wintermark, C. R. Durst, B. M. Kandel, J. C. Gee, M. C. Grossman, and B. B. Avants, "Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumor segmentation (simplified) with ANTsR," in *Neuroinformatics*, vol. 13, 2015, pp. 209–225.

[17] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61–78, 2017.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE CVPR*, 2015, pp. 3431–3440.

[19] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. MICCAI*. Springer, 2016, pp. 424–432.

[20] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. IEEE 3DV*, 2016, pp. 565–571.

[21] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical Image Analysis*, vol. 35, pp. 18–31, 2017.

[22] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[23] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," in *Proc. MIDL*, 2018, arXiv:1804.03999.

[24] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020.

[25] M. E. Shawky *et al.*, "The 2024 brain tumor segmentation (BraTS) challenge: Glioma segmentation on post-treatment MRI," *arXiv preprint arXiv:2405.18368*, 2024.

[26] A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in *Proc. BrainLes Workshop, MICCAI*. Springer, 2019, pp. 311–320.

[27] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[28] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "TransBTS: Multimodal brain tumor segmentation using transformer," in *Proc. MICCAI*. Springer, 2021, pp. 109–119.

[29] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Transformer-based semantic segmentation of brain tumors in MRI images," in *Proc. BrainLes Workshop, MICCAI*. Springer, 2022, pp. 272–284.

[30] R. A. Zeineldin, M. E. Karar, Z. Elshaer, J. Coburger, O. Burgert, and F. Mathis-Ullrich, "Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI," *Scientific Reports*, vol. 14, p. 2320, 2024.

[31] Y. Zhang *et al.*, "DeepGlioSeg: Advanced glioma MRI data segmentation with integrated local-global representation architecture," *Frontiers in Oncology*, vol. 15, p. 1449911, 2025.

[32] M. A. Naser and M. J. Deen, "Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images," *Computers in Biology and Medicine*, vol. 121, p. 103758, 2020.

[33] H. Wu *et al.*, "A deep ensemble learning framework for glioma segmentation and grading prediction," *Scientific Reports*, vol. 15, p. 4448, 2025.

[34] L. Sun, S. Zhang, H. Chen, and L. Luo, "Brain tumor segmentation and survival prediction using multimodal MRI scans with deep learning," *Frontiers in Neuroscience*, vol. 13, p. 810, 2019.

[35] M. H. Hannisdal, D. Goplen, S. Alam *et al.*, "Feasibility of deep learning-based tumor segmentation for target delineation and response assessment in grade-4 glioma using multi-parametric MRI," *Neuro-Oncology Advances*, vol. 5, no. 1, p. vdad037, 2023.

[36] K. Gkournelos *et al.*, "Multi-class glioma segmentation on real-world data with missing MRI sequences: Comparison of three deep learning algorithms," *Scientific Reports*, vol. 13, p. 18390, 2023.

[37] Q. D. Buchlak *et al.*, "Multimodal deep learning-based prognostication in glioma patients: A systematic review," *Cancers*, vol. 15, no. 2, p. 545, 2023.

[38] G. Singh, S. Manjila, N. Sakla *et al.*, "Radiomics and radiogenomics in gliomas: A contemporary update," *British Journal of Cancer*, vol. 125, no. 5, pp. 641–657, 2021.

[39] Y. Ren *et al.*, "Artificial intelligence-based MRI radiomics and radiogenomics in glioma," *Cancer Imaging*, vol. 24, no. 1, p. 36, 2024.

[40] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*. Academic Press, 1994, pp. 474–485.

[41] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, 2017, pp. 4700–4708.

[43] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE CVPR*, 2018, pp. 4510–4520.

[44] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019, pp. 6105–6114.

[45] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE CVPR*, 2018, pp. 7132–7141.

[46] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. NeurIPS*, 2021, pp. 12 077–12 090.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, arXiv:1412.6980.

[48] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proc. ACL*, 2019, pp. 3645–3650.

[49] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.

[50] E. Vayena, A. Blasimme, and I. G. Cohen, "Machine learning in medicine: Addressing ethical challenges," *PLoS Medicine*, vol. 15, no. 11, p. e1002689, 2018.

[51] D. S. Char, N. H. Shah, and D. Magnus, "Implementing machine learning in health care—addressing ethical challenges," *New England Journal of Medicine*, vol. 378, no. 11, pp. 981–983, 2018.

[52] World Health Organization, "Ethics and governance of artificial intelligence for health: WHO guidance," Geneva: World Health Organization, 2021. [Online]. Available: https://www.who.int/publications/i/item/9789240029200