

**Deep Learning-Based Brain Tumor Classification and
Segmentation Using Enhanced U-Net Architecture
on 2D MRI Scans**



Conduct By:

Name : Wajeehaldeen Ahmed Qasem Al-junaid
Student ID : 22523233

**PROGRAM STUDI INFORMATIKA – PROGRAM SARJANA
FAKULTAS TEKNOLOGI INDUSTRI
UNIVERSITAS ISLAM INDONESIA**

2026

SUPERVISOR ENDORSEMENT PAGE

Deep Learning-Based Brain Tumor Classification and Segmentation Using Enhanced U-Net Architecture on 2D MRI Scans

THESIS



Conduct By:

Name : Wajeehaldeen Ahmed Qasem Al-junaid

Student ID : 22523233

Yogyakarta, May, 2026

Advisor,


(Arrie Kurnia Wardhani, S.T., M.Cs.)

APPROVAL PAGE

Deep Learning-Based Brain Tumor Classification and Segmentation Using Enhanced U-Net Architecture on 2D MRI Scans

THESIS

Has been defended in front of the examiners as one of the requirements to obtain a Bachelor of Informatics degree from the Undergraduate Program in Informatics at the Faculty of Industrial Technology, Universitas Islam Indonesia.

Yogyakarta, May, 2026

Chair:

Arrie Kurniawardhani, S.T., M.Cs.

Examiner 1:

Ahmad Fathan Hidayatullah, S.T., M.Cs., Ph.D.

Examiner 2:

Erika Ramadhani, S.T., M.Eng.

Acknowledged by,

Head of Undergraduate Program in Informatics

Faculty of Industrial Technology

Universitas Islam Indonesia

(Ir. Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D.)

AUTHENTICITY STATEMENT

The undersigned:

Name : Wajeehaldeen Ahmed Qasem Al-junaid
Student ID : 22523233

Final project with title:

Deep Learning-Based Brain Tumor Classification and Segmentation Using Enhanced U-Net Architecture on 2D MRI Scans

Stating that all components and contents in this final project are my own work. If in the future it is proven that some parts of this work are not my own work, the final project submitted as my own work is ready to be withdrawn and ready to bear any risks and consequences.

Thus this statement letter is made, hopefully it can be used properly.

Yogyakarta, May, 2026



A handwritten signature in black ink, appearing to read 'Wajeehaldeen'.

(Wajeehaldeen Ahmed Qasem Al-junaid)

DEDICATION

This thesis is dedicated to my beloved parents, whose prayers, patience, and support have accompanied every step of my academic journey.

To my family and close friends, whose encouragement has been a constant source of strength during this research.

To my teachers and mentors, whose guidance has shaped both my scholarly thinking and personal growth.

MOTTO

“Seek knowledge from the cradle to the grave.”
— *Prophet Muhammad (peace be upon him)*

“Indeed, with hardship comes ease.”
— *Qur’an, Ash-Sharh 94:6*

“The expert in anything was once a beginner.”
— *Helen Hayes*

FOREWORD

Assalamu'alaikum Warahmatullahi Wabarakatuh.

All praise is due to Allah Subhanahu wa Ta'ala for His blessings and guidance, which have enabled me to complete this final project entitled “**Deep Learning-Based Brain Tumor Classification and Segmentation Using Enhanced U-Net Architecture on 2D MRI Scans.**” Peace and blessings be upon Prophet Muhammad SAW.

This thesis is submitted as one of the requirements for obtaining a Bachelor's degree in Informatics at the Faculty of Industrial Technology, Universitas Islam Indonesia.

I would like to express my sincere gratitude to:

1. Allah Subhanahu wa Ta'ala.
2. My beloved parents and family for their prayers and support.
3. **Arrie Kurniawardhani, S.T., M.Cs.**, as my thesis advisor, for her guidance and valuable suggestions.
4. My sister, an experienced medical practitioner, who kindly reviewed the experimental results and segmentation outputs of this research from a clinical perspective and confirmed that the produced tumor delineations were consistent with the radiological appearance she would expect for each tumor category. Her informal expert review provided an additional, clinically informed check on the outputs of the proposed model that complemented the quantitative evaluation reported in this thesis.
5. All lecturers, staff, friends, and colleagues who supported me during my studies.

I realize that this thesis is still far from perfect. I welcome constructive suggestions for improvement. I hope this research will be useful for future studies in medical image analysis.

Wassalamu'alaikum Warahmatullahi Wabarakatuh.

Yogyakarta, May, 2026

(Wajeehaldeen Ahmed Qasem Al-junaid)

ABSTRACT

Brain tumors are among the most serious neurological conditions, where accurate and timely diagnosis directly affects treatment planning and patient survival. Magnetic Resonance Imaging (MRI) is the standard modality for intracranial lesion assessment, but manual interpretation is time-consuming, subject to inter-observer variability, and demands specialized expertise. This thesis presents a unified deep learning framework that performs brain tumor classification and segmentation from 2D MRI slices in a single forward pass. Although in deep learning the segmentation step inherently produces a per-pixel class assignment, this work deliberately treats classification (assigning a single tumor type to the whole scan) and segmentation (delineating the tumor region) as two reported outputs of the same network, because each output answers a different clinical question and is evaluated with a different family of metrics. Three gaps in the literature are addressed: pronounced class imbalance in public datasets, the limited ability of the standard U-Net to produce fine tumor boundaries, and the architectural disconnect between classification and segmentation in U-Net-based brain tumor pipelines. Two encoder–decoder networks were developed and compared: a Standard U-Net baseline and an Enhanced U-Net that integrates residual connections, attention gates at every skip junction, and Squeeze-and-Excitation (SE) channel-recalibration modules. A composite training objective combining categorical cross-entropy, focal loss, Dice loss, and a boundary-aware term was used to jointly optimize overlap and contour accuracy. The training image dataset was assembled by merging two publicly available repositories—the Brain Tumor Segmentation Dataset by Akter et al. and the Brain Tumor Dataset (MAT Format) by Cheng et al.—followed by quality control and hybrid resampling to produce a class-balanced collection of 6,380 images with 1,595 samples per class across four categories: no tumor, glioma, meningioma, and pituitary. On the held-out test set, the Enhanced U-Net achieved pixel-wise accuracy of 99.73%, a Dice coefficient of 0.9141, and an IoU of 0.9947, improving on the Standard U-Net baseline (99.62% accuracy, 0.8423 Dice, 0.9913 IoU). Single-image inference was measured at 80–100 ms for the Standard U-Net and 100–150 ms for the Enhanced U-Net on an NVIDIA Tesla P100 GPU, well below clinical real-time requirements. Image-level classification, derived from spatial aggregation of segmentation probabilities, yielded mean precision and recall of 0.983 (per-class values ranging from 0.94 to 1.00). Comparative experiments on the two original imbalanced datasets confirmed the importance of dataset curation: models trained on Dataset 1 achieved only 69% recall on the minority glioma class, while models trained on Dataset 2 (which contains no healthy scans) failed completely to recognise healthy MRI slices as no-tumor cases.

The ablation study confirmed that each architectural module contributes measurably, and the three mechanisms (residual connections, attention gates, and SE modules) reinforce one another. The proposed multi-task Enhanced U-Net pipeline—which jointly produces a pixel-level segmentation map and an image-level class label in a single forward pass—demonstrates that combining targeted architectural refinements with a class-balanced dataset and a multi-task formulation produces a reliable tool for computer-aided brain tumor analysis suitable as a basis for further clinical validation.

Keywords: Brain Tumor Classification, Semantic Segmentation, U-Net, Enhanced U-Net, Deep Learning, Attention Mechanism, Residual Learning, Squeeze-and-Excitation, MRI Analysis, Multi-Task Learning.

LIST OF ABBREVIATIONS

2D	Two-Dimensional
3D	Three-Dimensional
Adam	Adaptive Moment Estimation
AI	Artificial Intelligence
BN	Batch Normalization
CE	Cross-Entropy
CNN	Convolutional Neural Network
CT	Computed Tomography
DICOM	Digital Imaging and Communications in Medicine
FCN	Fully Convolutional Network
FLAIR	Fluid-Attenuated Inversion Recovery
FN	False Negative
FP	False Positive
GPU	Graphics Processing Unit
IoU	Intersection over Union
MRI	Magnetic Resonance Imaging
PACS	Picture Archiving and Communication System
ReLU	Rectified Linear Unit
RGB	Red Green Blue (color space)
RQ	Research Question
SE	Squeeze-and-Excitation
SOTA	State-of-the-Art
T1	T1-weighted MRI sequence
T2	T2-weighted MRI sequence
TP	True Positive
UII	Universitas Islam Indonesia
WHO	World Health Organization

CONTENTS

SUPERVISOR ENDORSEMENT PAGE	i
APPROVAL PAGE	ii
AUTHENTICITY STATEMENT	iii
DEDICATION	iv
MOTTO	v
FOREWORD	vi
ABSTRACT	vii
LIST OF ABBREVIATIONS	ix
CONTENTS	x
LIST OF FIGURES	xiv
LIST OF TABLES	xvii
1 Introduction	1
1.1 Background	1
1.1.1 Clinical Significance of Brain Tumors	1
1.1.2 Diagnostic Challenges in MRI Interpretation	1
1.1.3 Deep Learning for Medical Image Analysis	2
1.1.4 Motivation for an Enhanced U-Net and a Balanced Dataset .	2
1.1.5 Research Motivation and Novelty Relative to Prior Work . .	3
1.2 Problem Statement	4
1.3 Research Questions	5
1.4 Scope of the Study	6
1.5 Research Objectives	6
1.6 Research Contributions	7
1.7 General Research Methodology	8
1.8 Thesis Organization	9
2 Literature Review	10
2.1 Overview and Theoretical Foundation	10
2.2 Brain Tumors and MRI-Based Diagnosis	10
2.3 Deep Learning for Medical Image Analysis	11

2.3.1	Convolutional Neural Networks in Radiology	11
2.3.2	Evaluation Metrics	11
2.3.3	Rationale for Selecting Deep Learning over Alternative Paradigms	12
2.4	The U-Net Family and Its Variants	13
2.4.1	Canonical U-Net and nnU-Net	13
2.4.2	Residual U-Net	14
2.4.3	Attention U-Net	14
2.4.4	Squeeze-and-Excitation Blocks	14
2.4.5	Other U-Net Extensions	15
2.4.6	Rationale for Selecting the U-Net Backbone	15
2.5	Loss Functions for Medical Image Segmentation	16
2.6	Brain Tumor Classification and Segmentation Studies	17
2.6.1	Classification Studies	17
2.6.2	Segmentation Studies	17
2.6.3	Summary of Representative Studies	17
2.7	Research Gap	19
3	Methodology	20
3.1	Dataset Construction and Preprocessing	21
3.1.1	Source Datasets	21
3.1.2	Cross-Dataset Compatibility Analysis	25
3.1.3	Dataset Merging and Balancing	26
3.1.4	Image Preprocessing Pipeline	27
3.1.5	Mask Preprocessing	29
3.1.6	Data Splitting	29
3.2	Model Architecture	31
3.2.1	Multi-Task Learning Framework	31
3.2.2	Standard U-Net (Baseline)	33
3.2.3	Enhanced U-Net	34
3.3	Loss Function and Evaluation Metrics	38
3.3.1	Composite Loss Function	38
3.3.2	Evaluation Metrics	39
3.4	Training Configuration	39
3.4.1	Hyperparameters	39
3.4.2	Optimizer and Learning-Rate Schedule	40
3.4.3	Regularization and Callbacks	40
3.4.4	Hyperparameter Tuning Strategy and Justification	40
3.5	Computational Environment	42

3.6	Methodology Workflow Summary	42
4	Results	44
4.1	Overview	44
4.2	Experimental Setup Recap	44
4.3	Quantitative Results on the Balanced Dataset	44
4.3.1	Overall Performance	44
4.3.2	Image-Level Classification Results	45
4.3.3	Training Dynamics	46
4.3.4	Empirical Evidence Against Overfitting	48
4.3.5	Per-Class Dice Analysis	50
4.4	Qualitative Results	50
4.5	Ablation Study	52
4.6	Comparative Experiments on Original Imbalanced Datasets	53
4.6.1	Experiment on Dataset 1 (Akter et al.)	53
4.6.2	Experiment on Dataset 2 (MAT Format)	53
4.6.3	Comprehensive Dataset Comparison	56
4.7	Comparison with Related Work	58
4.8	Research Question Mapping	59
4.9	Summary of Results	59
5	Discussion	61
5.1	Overview	61
5.2	Performance Analysis	61
5.3	Contribution of Architectural Modules	62
5.4	Dataset Curation and Class Imbalance	63
5.5	Reliability and Trustworthiness of Predictions in a High-Stakes Medical Setting	64
5.5.1	Internal Reliability Evidence	64
5.5.2	Limits of What the Reported Metrics Establish	66
5.5.3	What “Trustworthy” Means in This Context	66
5.6	Clinical Implications	67
5.7	Limitations	68
5.8	Comparison with Related Work	69
5.8.1	Factors Contributing to the Dice-Score Difference with Prior Work	69
6	Conclusion and Future Work	72

6.1 Conclusion	72
6.2 Future Work	73
REFERENCES	75
A Training Hyperparameter Summary	81
APPENDIX A: TRAINING HYPERPARAMETER SUMMARY	81
B Dataset Preparation Summary	82
APPENDIX B: DATASET PREPARATION SUMMARY	82
C Code and Data Availability	83
APPENDIX C: CODE AND DATA AVAILABILITY	83
D Research Question to Findings Mapping	84
APPENDIX D: RESEARCH QUESTION TO FINDINGS MAPPING	84

LIST OF FIGURES

3.1	Complete multi-task pipeline for simultaneous brain tumor classification and segmentation. The pipeline covers dataset construction, preprocessing, Enhanced U-Net inference, dual output generation (pixel-level segmentation and image-level classification), and evaluation.	20
3.2	Representative MRI images and corresponding segmentation masks from Dataset 1 (Akter et al.), one row per class: no tumor, glioma, meningioma, and pituitary tumor.	22
3.3	Representative MRI images and corresponding segmentation masks from Dataset 2 (Cheng et al.), one row per class actually present in the source: glioma, meningioma, and pituitary tumor. Note that Dataset 2 does not contain no-tumor (healthy) cases.	23
3.4	Dataset distribution comparison: (a) original Dataset 1 class distribution, (b) original Dataset 2 class distribution, and (c) final merged and balanced dataset with equal representation across all classes (1,595 images per class).	28
3.5	Sample MRI images from each class, showing original brain scans and their corresponding segmentation masks for no tumor, glioma, meningioma, and pituitary tumor. Masks delineate the pathological regions where applicable.	28
3.6	Standard U-Net architecture, showing the encoder path (left), bottleneck (center), and decoder path (right) with skip connections. Each blue box represents a multi-channel feature map, with the number of channels indicated above. Arrows denote convolution, pooling, up-sampling, and concatenation operations. Adapted from Ronneberger et al. (2015).	34
3.7	Enhanced U-Net architecture with residual blocks (identity shortcuts indicated by curved arrows), attention gates at every skip junction (between encoder and decoder stages), and Squeeze-and-Excitation modules (placed after each convolutional block). The output is a $160 \times 160 \times 4$ probability map.	35
4.1	Normalized confusion matrix for the Standard U-Net on the balanced test set. Diagonal values represent correct classifications. Note the relatively lower glioma recall (0.87).	46

4.2	Normalized confusion matrix for the Enhanced U-Net on the balanced test set. Diagonal values are higher than for the Standard U-Net across all classes, particularly for meningioma and pituitary tumors.	47
4.3	Standard U-Net training dynamics: (a) loss, (b) accuracy, (c) Dice, and (d) IoU for both training and validation sets across 80 epochs. Convergence is stable with a small train–validation gap.	47
4.4	Enhanced U-Net training dynamics: (a) loss, (b) accuracy, (c) Dice, and (d) IoU for both training and validation sets across 80 epochs. The Enhanced model reaches higher final performance than the Standard U-Net, with a validation Dice of 0.8459 by epoch 7.	48
4.5	Qualitative segmentation results, showing representative samples from each tumor class. Each row displays, from left to right, the original MRI image, the ground-truth segmentation mask, and the Enhanced U-Net prediction. Colors indicate tumor classes: red (glioma), yellow (meningioma), blue (pituitary).	51
4.6	Normalized confusion matrix for the Standard U-Net trained on Dataset 1. Glioma recall is only 0.69, because the minority glioma class is under-represented in training.	54
4.7	Training dynamics for the Standard U-Net on Dataset 1: (a) loss, (b) accuracy, (c) Dice, and (d) IoU. Fast initial convergence masks an over-fit to the majority class distribution.	54
4.8	Qualitative segmentation results from the Standard U-Net trained on Dataset 1. From left to right: original MRI, ground-truth mask, and predicted mask. The model tends to under-segment glioma (the minority class).	55
4.9	Normalized confusion matrix for the Standard U-Net trained on Dataset 2. Background recall is 0% because Dataset 2 contains no healthy samples.	56
4.10	Training dynamics for the Standard U-Net on Dataset 2: (a) loss, (b) accuracy, (c) Dice, and (d) IoU. The curves look healthy in isolation, but the absence of no-tumor samples makes the model clinically unusable.	57

4.11 Qualitative segmentation results from the Standard U-Net trained on Dataset 2. From left to right: original MRI, ground-truth mask, and predicted mask. Tumor delineation is reasonable, but the model cannot identify healthy tissue regions.	57
---	----

LIST OF TABLES

2.1	Representative deep learning studies on brain tumor classification and segmentation, with the concrete quantitative result reported by each study. Note that studies evaluate on different datasets and task definitions, so numbers are not strictly comparable across rows. . . .	18
3.1	Original dataset class distribution before balancing.	21
3.2	Compatibility analysis of the two source datasets along five dimensions relevant to merging. The right-hand column indicates whether the dimension required an explicit harmonisation step in the preprocessing pipeline.	25
3.3	Final balanced dataset statistics.	27
4.1	Performance comparison of the Standard U-Net and the Enhanced U-Net on the balanced test set.	45
4.2	Training, validation, and test metrics for the Standard and Enhanced U-Net on the balanced dataset.	45
4.3	Image-level classification comparison between Standard and Enhanced U-Net on the balanced test set.	45
4.4	Per-class Dice coefficient comparison.	50
4.5	Ablation study isolating the contribution of each Enhanced U-Net module. Metrics are reported on the balanced test set; the ablation baseline uses a smaller capacity configuration than the full Standard U-Net in Table 4.1.	52
4.6	Standard U-Net performance on original Dataset 1 (Akter et al.). . .	53
4.7	Standard U-Net performance on original Dataset 2 (MAT Format). .	54
4.8	Comprehensive Standard U-Net performance comparison across datasets.	56
4.9	Image-level classification across datasets (P = precision, R = recall). .	58
4.10	Comparison of the proposed Enhanced U-Net against representative deep learning methods for brain tumor segmentation. Results reported on different datasets are not fully commensurable and are shown here for positioning purposes only.	59
4.11	Mapping of research questions (RQs) to the findings of this thesis. .	60

A.1	Complete training hyperparameter configuration.	81
D.1	Mapping of research questions to the evidence that answers them. .	84

CHAPTER I

INTRODUCTION

1.1 Background

1.1.1 Clinical Significance of Brain Tumors

Brain tumors are among the most serious neurological conditions encountered in clinical practice. They can cause neurological deficit, seizures, cognitive decline, and in many cases death, placing a long-lasting burden on patients, families, and healthcare services. Among primary brain tumors, three categories are most commonly seen in radiology departments and are the focus of this thesis: glioma, meningioma, and pituitary tumor. Gliomas vary widely in their biological aggressiveness, from low-grade lesions with slow progression to glioblastoma, whose survival rarely exceeds about fifteen months despite combined surgery, chemotherapy, and radiation (Menze et al., 2015; Bakas et al., 2018). Meningiomas are typically slow-growing but can cause significant morbidity through mass effect on neighboring brain structures (Cheng et al., 2015; Litjens et al., 2017). Pituitary tumors are often small, yet they can disturb endocrine function or compress adjacent neural structures such as the optic chiasm (Cheng et al., 2015; Akter et al., 2024). Accurate and early discrimination among these tumor types is therefore a prerequisite for appropriate treatment selection (Menze et al., 2015; Bakas et al., 2018; Litjens et al., 2017).

1.1.2 Diagnostic Challenges in MRI Interpretation

Magnetic Resonance Imaging (MRI) is the standard modality for visualizing brain tumors, because it provides good soft-tissue contrast without exposing patients to ionizing radiation. A typical diagnostic workup combines several pulse sequences—T1-weighted, T2-weighted, Fluid-Attenuated Inversion Recovery (FLAIR), and post-contrast T1—to describe lesion location, size, enhancement pattern, and surrounding edema. Despite this, manual interpretation is demanding. Tumor margins often blur gradually into normal parenchyma, different tumor types can share radiological features, and motion or susceptibility artifacts may obscure relevant detail. Radiologists must combine signal intensity, morphology, location, and clinical history to reach a working diagnosis—a cognitive process that is affected by fatigue, experience, and workload (Akter et al., 2024; Litjens et al., 2017).

Manual volumetric delineation, required for surgical planning and radiotherapy contouring, is even more labor-intensive, and several studies document substantial inter-observer variability in brain tumor interpretation (Litjens et al., 2017; Esteva et al., 2019; Topol, 2019). These observations motivate the use of computer-assisted

tools that can provide consistent preliminary classifications and candidate segmentation masks to support the radiologist.

1.1.3 Deep Learning for Medical Image Analysis

Over the past decade, deep learning—and in particular Convolutional Neural Networks (CNNs)—has replaced hand-crafted feature pipelines in many medical imaging tasks. Architectures such as AlexNet, VGG, ResNet, and DenseNet have been adapted to tasks that include diabetic retinopathy screening, chest X-ray interpretation, dermatological lesion classification, and tumor detection in radiology (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015; He et al., 2016; Huang et al., 2017; Litjens et al., 2017; Shen et al., 2017). Among segmentation architectures, the U-Net of Ronneberger et al. (2015) has become a reference model in biomedical imaging, because its encoder–decoder topology with skip connections produces high-quality segmentations even with limited training data.

Many U-Net variants have been proposed, including UNet++ (Zhou et al., 2018), Attention U-Net (Oktay et al., 2018), Residual U-Net (Zhang et al., 2018), and the self-configuring nnU-Net (Isensee et al., 2021); each of these is reviewed in detail in Chapter 2, and the design of the Enhanced U-Net proposed in this thesis was informed by the strengths and limitations identified in that review. However, three issues still limit the clinical usefulness of U-Net-based brain tumor systems. First, many published models are trained on imbalanced datasets, in which healthy tissue or particular tumor subtypes are over- or under-represented, leading to biased evaluation. Second, the canonical U-Net is reported to produce blurred boundaries and to under-segment small tumors when used on its own (Oktay et al., 2018; Kervadec et al., 2019; Isensee et al., 2021), and this same behaviour is reproduced empirically by the Standard U-Net baseline trained in Chapter 4. Third, classification and segmentation are usually performed by separate networks, which introduces redundancy and potential inconsistency between outputs.

1.1.4 Motivation for an Enhanced U-Net and a Balanced Dataset

This thesis is motivated by three observations. First, accurate boundary delineation is essential not only for diagnosis but also for surgical planning and radiotherapy, where small geometric errors translate directly into clinical risk. Second, classification and segmentation are tightly coupled: the identity of a tumor and its spatial extent are two aspects of the same pathology, so a unified multi-task pipeline can exploit this relationship while reducing computational cost. Third, class imbalance in public datasets is a systematic obstacle to fair evaluation and clinically

trustworthy deployment. As shown in the comparative experiments reported later in this thesis, models trained on skewed data can silently fail on minority classes or, in one case, on healthy tissue itself (Akter et al., 2024).

In response, this work develops an Enhanced U-Net that adds three mechanisms to the canonical design: residual connections for stable gradient flow (He et al., 2016; Zhang et al., 2018), attention gates for spatially selective feature emphasis (Oktay et al., 2018), and Squeeze-and-Excitation modules for channel-wise recalibration (Hu et al., 2018). These additions are integrated into a multi-task framework in which the image-level tumor type is inferred directly from the segmentation output, so that the two outputs are always consistent with one another. To complement the architectural work, a class-balanced dataset of 6,380 MRI scans is constructed by merging and harmonizing the two publicly available repositories mentioned above.

1.1.5 Research Motivation and Novelty Relative to Prior Work

The closest prior study is that of Akter et al. (2024), who report approximately 99% image-level classification accuracy on a partly overlapping data source using a Standard U-Net jointly trained for classification and segmentation. This raises a legitimate question: *if 99% accuracy has already been demonstrated, what justifies further work on the same problem?* The answer is that headline accuracy alone is an incomplete picture of model quality for brain tumor imaging, and the present thesis is motivated by five concrete gaps that the prior work does not address.

1. **Boundary fidelity, not only image-level accuracy.** Classification accuracy answers a different clinical question from segmentation overlap. A model can correctly identify the tumor type at scan level while producing a segmentation mask that under-covers the lesion or that bleeds into healthy tissue by several millimetres — errors that are clinically significant for surgical planning and radiotherapy contouring but that do not show up in classification accuracy. The Dice coefficient reported by Akter et al. (2024) (approximately 0.85) leaves clear room for improvement on this dimension, and the Enhanced U-Net of this thesis raises it to 0.9141 (Section 4.3).
2. **Class balance across all four categories, including healthy scans.** The original sources used in prior work are imbalanced: in particular, Dataset 2 (Cheng et al., 2015) contains no healthy (no-tumor) samples, and Dataset 1 (Akter et al., 2024) has a 2.46:1 ratio between no-tumor and glioma. As Chapter IV will show empirically, training on these sources produces silent per-class failures (glioma recall 0.69 on Dataset 1; background recall 0.00 on Dataset 2) that are hidden by aggregate accuracy. The 6,380-image balanced

corpus constructed in this thesis (1,595 per class across all four categories) is a substantive methodological contribution that is independent of the architectural enhancements.

3. **Integration of residual, attention, and Squeeze-and-Excitation modules in a single brain tumor backbone.** Prior brain tumor U-Net variants adopt one or two of these mechanisms in isolation. The Enhanced U-Net combines all three within a single 2D backbone and uses the ablation study (Section 4.5) to quantify how much each module contributes (+3.59 pp residual, +1.74 pp attention, +1.85 pp SE on Dice), so that the gain over the Standard U-Net is attributable rather than monolithic.
4. **Composite loss with an explicit boundary-aware term.** Akter et al. (2024) use a standard loss formulation without explicit boundary supervision. The composite loss used in this thesis combines categorical cross-entropy, focal loss (Lin et al., 2017), Dice loss (Milletari et al., 2016), and a boundary-aware term (Kervadec et al., 2019) with weights (0.25, 0.35, 0.30, 0.10), directly targeting the boundary smoothing problem that limits standard U-Net outputs.
5. **Unified multi-task formulation with deterministically consistent outputs.** The image-level label in this thesis is computed as a deterministic function of the segmentation output (Section 3.2), so the two outputs cannot disagree by construction. Pipelines that use separate classification and segmentation heads — or separate networks entirely — can in principle produce inconsistent outputs (for example, classifying a scan as “pituitary” while segmenting a region of glioma morphology); the unified design used here eliminates that failure mode.

The novelty of this thesis therefore does not lie in lifting accuracy from 99% to 99.7%. It lies in (a) substantially improving boundary fidelity (Dice $0.85 \rightarrow 0.91$), (b) producing the first balanced four-class corpus built from these two sources, (c) integrating residual + attention + SE mechanisms with ablation-level evidence for each, (d) introducing an explicit boundary loss into the composite objective, and (e) coupling classification and segmentation in a way that is internally consistent by construction. These are the dimensions on which prior 99%-accuracy systems remain incomplete for clinical use, and they form the core contributions of the present work.

1.2 Problem Statement

Based on the background above, the problems addressed in this thesis are:

1. **Diagnostic difficulty of manual MRI interpretation.** Brain tumors exhibit significant radiological heterogeneity, and manual review of MRI studies is time-consuming and subject to inter-observer variability.
2. **Limited boundary precision of the standard U-Net.** The canonical U-Net often produces blurred lesion contours and shows reduced sensitivity to small tumors, which limits its use for surgical and radiotherapy planning.
3. **Class imbalance in public datasets.** Existing brain tumor MRI repositories are often class-imbalanced, and in some cases omit entire diagnostic categories (for example, healthy tissue), producing models that fail on minority or missing classes.
4. **Disconnect between classification and segmentation.** Most published pipelines use separate networks for tumor typing and boundary delineation, which produces redundant computation and can lead to inconsistent outputs.

1.3 Research Questions

To address the problems above, this thesis is guided by the following research questions (RQs):

- **RQ1:** How accurately can a unified deep learning model based on an Enhanced U-Net architecture jointly classify and segment brain tumors from 2D MRI slices when evaluated using pixel-level (accuracy, Dice, IoU) and image-level (precision, recall, F1-score) metrics?
- **RQ2:** How does the proposed Enhanced U-Net (incorporating residual connections, attention gates, and Squeeze-and-Excitation modules) compare with a Standard U-Net baseline trained under identical conditions on the same curated dataset, and what is the individual contribution of each module?
- **RQ3:** What is the empirical effect of dataset class imbalance on model behavior, and does a class-balanced merged dataset produce more reliable performance than the original imbalanced sources?
- **RQ4:** How does unifying classification and segmentation in a single forward pass—by deriving image-level labels from aggregated segmentation probabilities—affect consistency and computational efficiency of the pipeline?

1.4 Scope of the Study

To keep the study focused and feasible, its scope is defined as follows:

1. **Imaging modality and anatomical region.** The study is restricted to 2D axial brain MRI slices derived from T1-weighted sequences (including contrast-enhanced variants). Other modalities (CT, PET, ultrasound) and anatomical regions are not considered.
2. **Tumor categories.** Only four categories are considered: no tumor (healthy), glioma, meningioma, and pituitary tumor. Tumor subtyping, grading, and survival prediction are not addressed.
3. **Tasks.** The primary tasks are multi-class pixel-level segmentation and whole-image classification. Longitudinal progression tracking, treatment response assessment, and prognostic modeling are not addressed.
4. **Models.** The core focus is on the Enhanced U-Net and its Standard U-Net baseline. Architectures from fundamentally different families (for example, Transformer-based segmenters or foundation models) are discussed in the literature review but are not implemented.
5. **Clinical deployment.** The study is conducted in a research environment. Regulatory approval, PACS integration, user-interface design, and prospective clinical trials are outside the formal scope, although they are acknowledged in the discussion.

1.5 Research Objectives

The overall goal of this thesis is to develop and evaluate an Enhanced U-Net-based deep learning framework for the joint classification and segmentation of brain tumors in 2D MRI scans. The objectives below are derived directly from, and aligned one-to-one with, the research questions stated in the previous section; the parenthetical tag at the end of each objective indicates the research question it addresses.

1. To design and implement an end-to-end multi-task pipeline in which a single network produces both pixel-level segmentation maps and image-level classification labels in a single forward pass. (*addresses RQ1, RQ4*)
2. To construct a class-balanced, quality-assured MRI dataset by merging two publicly available repositories, eliminating class imbalance and ensuring 100% image–mask correspondence. (*addresses RQ3*)

3. To propose an Enhanced U-Net architecture that incorporates residual connections, attention gates, and Squeeze-and-Excitation modules, together with a composite loss function that targets overlap, class balance, and boundary fidelity. (*addresses RQ2*)
4. To quantitatively evaluate both the Standard U-Net baseline and the proposed Enhanced U-Net on the balanced dataset using pixel-level (accuracy, Dice, IoU) and image-level (precision, recall, F1-score) metrics. (*addresses RQ1, RQ2*)
5. To conduct comparative experiments on the two original imbalanced datasets in order to characterize the effect of class imbalance on model performance. (*addresses RQ3*)
6. To perform an ablation study that isolates the contribution of each architectural enhancement. (*addresses RQ2*)

1.6 Research Contributions

This thesis contributes the following to the literature and practice:

1. **Unified multi-task framework.** An end-to-end pipeline that performs pixel-level segmentation and image-level classification in a single forward pass. The scan-level label is inferred by aggregating tumor-pixel probabilities from the segmentation output, so that the two outputs are consistent by construction.
2. **Balanced dataset construction.** A curated, class-balanced corpus of 6,380 brain MRI scans, assembled by merging two public datasets and enforcing 100% image-mask correspondence, with 1,595 samples per class across four categories.
3. **Enhanced U-Net architecture.** A refined U-Net design that integrates residual connections, attention gates, and Squeeze-and-Excitation modules, together with a composite loss function combining categorical cross-entropy, focal loss, Dice loss, and a boundary-aware term.
4. **Systematic empirical evaluation.** A controlled quantitative comparison between the Standard U-Net baseline and the Enhanced U-Net under identical training and evaluation conditions, complemented by comparative experiments on the two original imbalanced datasets.
5. **Ablation analysis.** A structured ablation study that isolates the individual contribution of each architectural module.

6. **Reproducible methodology.** Detailed documentation of dataset preparation, preprocessing, model configuration, and training protocols, to enable reproduction by future researchers.

1.7 General Research Methodology

To achieve the objectives above, the thesis follows a general research methodology commonly used in applied medical imaging and deep learning studies. The methodology is summarized here and elaborated in Chapter III.

1. **Problem formulation and literature review.** Characterization of the clinical problem and the role of MRI, followed by a structured review of deep learning methods with emphasis on U-Net variants and multi-task pipelines.
2. **Data acquisition and curation.** Two publicly available datasets are obtained, converted to a common format, audited for image–mask correspondence, and balanced by a hybrid resampling strategy.
3. **Preprocessing.** Images and masks undergo color space conversion, resizing to a fixed resolution, intensity normalization, and class-conditional thresholding to produce one-hot encoded segmentation labels.
4. **Model design and implementation.** Two networks are built: a Standard U-Net baseline and a proposed Enhanced U-Net with residual, attention, and SE modules. Both networks output four-channel probability maps feeding a unified multi-task head.
5. **Training and validation.** Both models are trained under identical settings (optimizer, learning-rate schedule, batch size, epoch budget, regularization) using a composite loss function, with early stopping and checkpointing driven by validation Dice.
6. **Testing and performance evaluation.** Final models are evaluated on a held-out test partition using pixel-level and image-level metrics, supported by confusion matrices and qualitative visualizations.
7. **Comparative and ablation analyses.** Comparative experiments on the original imbalanced datasets quantify the impact of class imbalance, and an ablation study isolates the contribution of each architectural enhancement.
8. **Conclusion and recommendations.** Findings are interpreted in light of the research questions, limitations are discussed, and recommendations for future work are articulated.

1.8 Thesis Organization

The remainder of this thesis is organized as follows. Chapter II presents the literature review and places this work within the broader landscape of brain tumor analysis and deep learning for medical image segmentation. Chapter III describes the methodology in detail, including dataset construction, preprocessing, the architectures of the Standard and Enhanced U-Nets, the composite loss function, and the training protocol. Chapter IV presents the experimental results, including the comparative experiments on the original imbalanced datasets and the ablation study. Chapter V discusses the findings and compares them to related work. Chapter VI summarizes the contributions and outlines directions for future research.

CHAPTER II

LITERATURE REVIEW

2.1 Overview and Theoretical Foundation

This chapter reviews the literature relevant to deep learning-based brain tumor classification and segmentation. The review is organized around four themes. First, the clinical background of brain tumors and the role of MRI in their diagnosis are summarized. Second, deep learning methods for medical image analysis are described, with emphasis on convolutional architectures for segmentation. Third, the U-Net family and its most influential variants are examined, together with the architectural refinements (residual connections, attention mechanisms, channel recalibration) that motivate the proposed Enhanced U-Net. Fourth, the specific literature on brain tumor classification and segmentation is reviewed, and a summary table of representative studies is presented. The chapter concludes by identifying the research gap that this thesis addresses.

This study builds on three core technical concepts: (i) convolutional neural networks for hierarchical feature extraction from images; (ii) encoder–decoder architectures (U-Net) for dense pixel-level prediction; and (iii) overlap-based metrics (Dice, IoU) and per-class classification metrics (precision, recall, F1) for evaluating segmentation and classification outputs. These concepts form the basis for interpreting the studies reviewed below.

2.2 Brain Tumors and MRI-Based Diagnosis

Brain tumors are a heterogeneous group of neoplasms that arise either from cells intrinsic to the central nervous system (primary tumors) or from metastatic spread of extracranial cancers. This thesis focuses on three types of primary brain tumors—glioma, meningioma, and pituitary adenoma—which together account for a large fraction of intracranial tumors seen in clinical practice. Gliomas are clinically important because they include high-grade lesions such as glioblastoma, whose aggressive and infiltrative growth translates into poor prognosis. Meningiomas are typically slow-growing but can cause morbidity through mass effect when located near critical structures. Pituitary tumors are usually benign but can disturb endocrine function or compress adjacent structures when large (Menze et al., 2015; Bakas et al., 2018).

MRI is the reference modality for intracranial lesion assessment because it offers good soft-tissue contrast, multi-sequence characterization, and no ionizing radiation. A typical diagnostic protocol integrates T1-weighted, T2-weighted, FLAIR,

and post-contrast T1 sequences. Despite its power, MRI interpretation is cognitively demanding: the radiologist must synthesize information across sequences, localize lesions in three-dimensional space while reviewing many two-dimensional slices, and integrate the findings with clinical history. Inter-observer variability and the time cost of manual delineation create a clear motivation for reliable computer-assisted tools that can provide consistent preliminary classifications and candidate segmentation masks (Litjens et al., 2017; Esteva et al., 2019; Topol, 2019; Akter et al., 2024).

2.3 Deep Learning for Medical Image Analysis

2.3.1 Convolutional Neural Networks in Radiology

Deep convolutional networks have produced a clear shift in radiological image analysis. Architectures originally developed for natural image recognition—AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2015), Inception (Szegedy et al., 2016), ResNet (He et al., 2016), DenseNet (Huang et al., 2017), and Xception (Chollet, 2017)—have been adapted to clinical tasks such as diabetic retinopathy screening, chest X-ray interpretation, dermatological classification, and brain tumor analysis. Reviews by Litjens et al. (2017) and Shen et al. (2017) document this transition across modalities.

For segmentation—a pixel-level prediction task—Fully Convolutional Networks (FCN) introduced by Long et al. (2015) showed that CNNs could be extended beyond whole-image classification to dense prediction. Later architectures such as SegNet (Badrinarayanan et al., 2017), DeepLab (Chen et al., 2018), and PSPNet (Zhao et al., 2017) continued this trajectory. However, in the medical domain, the U-Net of Ronneberger et al. (2015) became the default segmentation model, mainly because it produces good results even with limited training data.

2.3.2 Evaluation Metrics

Segmentation models are typically evaluated using metrics that measure overlap and boundary quality. The Dice similarity coefficient (twice the intersection divided by the sum of the predicted and ground-truth regions) is widely used because it is sensitive to class imbalance and behaves like a harmonic mean of precision and recall. The Intersection over Union (IoU, also known as Jaccard index) is a closely related measure. Pixel-wise accuracy is simple to compute but can be misleading when background pixels dominate. For classification, per-class precision, recall, and F1-score together with confusion matrices expose per-class behavior that aggregate accuracy can hide (Milletari et al., 2016; Sudre et al., 2017).

2.3.3 Rationale for Selecting Deep Learning over Alternative Paradigms

Before settling on a deep learning solution, four alternative paradigms were considered for this study: (i) traditional machine learning with hand-crafted features, (ii) transfer learning from natural-image pre-trained backbones, (iii) Transformer-based segmenters, and (iv) end-to-end deep learning trained from scratch. Each option was assessed against the joint requirements of this thesis, namely pixel-level boundary delineation, multi-class image-level classification, robustness under class imbalance, and feasibility on a single commodity GPU. The reasons for selecting end-to-end deep learning over the three alternatives are summarized below.

(i) Traditional machine learning pipelines (for example, SVMs, random forests, or k -nearest neighbors operating on hand-crafted features such as gray-level co-occurrence matrices, local binary patterns, wavelet coefficients, or shape descriptors) have a long history in brain tumor analysis (Cheng et al., 2015, 2016). They are interpretable and computationally cheap. However, they have three structural limitations for the present task. First, performance is bounded by the quality of the manually engineered descriptors, which require radiology-specific expertise and rarely generalize across scanners, sequences, and tumor morphologies. Second, classical pipelines almost always perform classification or region retrieval, not dense pixel-level segmentation; reaching a pixel-level output typically requires a separately engineered post-processing stage such as level sets or graph cuts. Third, the strongest published results on the Cheng et al. dataset using SVM + region augmentation are around 91% classification accuracy with no segmentation output (Cheng et al., 2015), which is well below what end-to-end CNNs reach on the same data. For the joint classification + segmentation problem targeted in this thesis, the traditional ML route would have meant building two separate pipelines, neither of which would clear the bar set by recent CNN-based work.

(ii) Transfer learning from ImageNet-pretrained backbones (VGG, ResNet, DenseNet, EfficientNet) is a strong baseline when training data is scarce. It was deliberately not adopted in this thesis for three reasons. First, ImageNet weights are learned on natural color photographs whose low-level statistics (texture, illumination, color distribution, structured background) differ substantially from grayscale T1-weighted MRI slices; several survey-level studies have documented that the gains from such transfer are modest in medical imaging once a few thousand domain-specific samples are available (Litjens et al., 2017; Shen et al., 2017; Esteva et al., 2019). Second, off-the-shelf pretrained encoders are designed for whole-image classification and lack the symmetric decoder needed for pixel-level segmentation; bolting a decoder onto them re-introduces most of the design decisions that a purpose-built U-Net already handles. Third, the balanced corpus assembled in this thesis

(6,380 image–mask pairs, Section 3.1) is large enough to train a 2D segmentation network from scratch, so the main argument in favor of transfer learning — data scarcity — does not apply in the present setting.

(iii) Transformer-based segmenters such as TransUNet, Swin-UNet, SegFormer, and SAM-derived models are an active research frontier and report strong results on several public benchmarks. They were not adopted in this thesis for three reasons. First, Transformers typically need substantially more training data and longer training schedules than convolutional networks of comparable parameter count, because they impose weaker spatial inductive biases (no locality, no translation equivariance) that CNNs encode for free; their advantage usually appears at data scales an order of magnitude larger than the present corpus (Vaswani et al., 2017). Second, self-attention has quadratic complexity in the number of tokens, which translates into substantially higher GPU memory and compute cost at 160×160 resolution; this is a real constraint on a single Tesla P100 with 16 GB of memory. Third, for the specific task of joint pixel-level segmentation and image-level classification of tumors on 2D MRI slices, the convolutional U-Net family with attention gates, residual connections, and SE recalibration already captures the essential mechanisms (spatial attention, channel attention, skip context) that motivate Transformers, while remaining tractable on the available hardware. Transformer-based architectures are therefore acknowledged as a natural direction for future work (Chapter VI) rather than as the baseline of the present thesis.

(iv) End-to-end deep learning from scratch using a U-Net backbone was selected. It learns hierarchical features directly from the MRI data, removing the need for hand-crafted descriptors; it produces dense pixel-level outputs natively through its encoder–decoder topology, which is exactly what the segmentation task requires; it provides a shared backbone from which the image-level label can be derived deterministically, supporting the unified multi-task design (Section 3.2); and on a balanced corpus of several thousand image–mask pairs it can be trained to convergence on a single GPU within a few hours. These properties together make end-to-end deep learning the most appropriate paradigm for the joint segmentation–classification problem addressed in this thesis.

2.4 The U-Net Family and Its Variants

2.4.1 Canonical U-Net and nnU-Net

The U-Net of Ronneberger et al. (2015) was originally designed for cell-membrane segmentation in microscopy images but quickly became a reference architecture across medical imaging. Its key features are a contracting encoder

that compresses spatial resolution while expanding channel depth, an expansive decoder that restores full resolution, and lateral skip connections that relay fine-grained features from the encoder to the decoder. The skip connections are important because they transfer spatial detail from shallow encoder layers to deep decoder layers, partly compensating for the information loss caused by pooling.

The U-Net design has been shown to work across many imaging modalities (X-ray, CT, MRI, ultrasound, microscopy). The nnU-Net framework of Isensee et al. (2021) formalized this by showing that a properly configured U-Net, without architectural novelty, can match or outperform more complex networks on many public benchmarks when preprocessing, training schedule, and post-processing are chosen appropriately.

2.4.2 Residual U-Net

Deeper encoder–decoder stacks can suffer from vanishing gradients and optimization difficulties. Residual learning, introduced by He et al. (2016), addresses this by reformulating stacked layers as residual mappings $\mathcal{F}(\mathbf{x}) + \mathbf{x}$, so that the network learns only the perturbation relative to identity. Zhang et al. (2018) integrated residual connections into the U-Net encoder and decoder for road extraction and reported consistent segmentation gains, particularly in deeper networks. In the brain tumor context, residual pathways stabilize training of deeper networks that must capture both large-scale morphology and fine boundary detail.

2.4.3 Attention U-Net

Oktay et al. (2018) proposed the Attention U-Net, which inserts learnable gating modules at skip junctions. These gates enable the decoder to selectively emphasize task-relevant activations while suppressing background noise. The gate takes a gating signal from a deeper layer and skip features from the encoder, and produces an attention coefficient (between 0 and 1) that is used to modulate the skip features element-wise. This is especially useful in medical imaging, where target structures often occupy a small fraction of the image and have ambiguous borders. Related spatial and channel attention mechanisms (Woo et al., 2018; Huang et al., 2019; Wang et al., 2018; Vaswani et al., 2017) have further confirmed the usefulness of attention for dense prediction.

2.4.4 Squeeze-and-Excitation Blocks

Squeeze-and-Excitation (SE) blocks, introduced by Hu et al. (2018), re-weight feature channels according to their informativeness. Each block globally pools

spatial information per channel, passes the resulting vector through a small two-layer bottleneck with a sigmoid output, and uses the result to scale the original channels. Roy et al. (2018) extended this to dense prediction tasks by combining channel-wise and spatial SE. In medical segmentation, where different channels may capture different diagnostic cues (for example, contrast-enhancing regions versus characteristic textures), SE modules allow the network to emphasize the most informative channels on a per-image basis.

2.4.5 Other U-Net Extensions

Many additional U-Net variants exist. UNet++ (Zhou et al., 2018) introduces nested dense skip pathways, while UNet 3+ (Huang et al., 2020) uses full-scale connectivity across all levels. For volumetric data, V-Net (Milletari et al., 2016) and 3D U-Net (Çiçek et al., 2016) extend the architecture to three-dimensional convolutions. Other works have explored deep supervision (Isensee et al., 2018) and autoencoder regularization (Myronenko, 2018) for volumetric brain tumor segmentation, and multi-scale architectures with dilated convolutions and conditional random fields (Kamnitsas et al., 2017). The Enhanced U-Net proposed in this thesis combines three of the most empirically effective mechanisms—residual connections, attention gates, and Squeeze-and-Excitation—within a 2D design that is computationally tractable.

2.4.6 Rationale for Selecting the U-Net Backbone

Although several segmentation families exist (FCN (Long et al., 2015), SegNet (Badrinarayanan et al., 2017), DeepLab (Chen et al., 2018), PSPNet (Zhao et al., 2017), Transformer-based segmenters), the U-Net family was selected as the backbone of this thesis for five specific reasons.

First, the encoder–decoder topology natively produces dense pixel-level outputs. Unlike classification networks that need a separately engineered upsampling head, U-Net was designed end-to-end for semantic segmentation, with a contracting path that compresses spatial resolution while expanding channel depth and a symmetric expansive path that restores full resolution (Ronneberger et al., 2015). This matches the principal task of this thesis exactly.

Second, the lateral skip connections preserve fine spatial detail. Repeated pooling discards the high-resolution spatial information that is required to delineate tumor boundaries. By concatenating encoder feature maps directly into the decoder, U-Net recovers this information at every resolution scale, which is precisely the property that other segmentation families (such as FCN and SegNet) achieve less effectively. For medical images, where boundaries between healthy and pathological

tissue are often subtle and clinically critical, this property is essential.

Third, U-Net trains effectively on moderate-sized datasets. The original U-Net paper specifically targeted biomedical settings where annotated data is scarce, and reported strong performance with only ~ 30 microscopy images using aggressive augmentation (Ronneberger et al., 2015). Although the balanced corpus used in this thesis is much larger (6,380 images), the same data-efficiency property remains relevant: medical annotation is expensive, and an architecture that converges reliably without millions of training images is a structural advantage.

Fourth, U-Net has become the de facto reference in biomedical segmentation. It has been benchmarked extensively across modalities and tasks, including the brain tumor problem itself (Akter et al., 2024; Isensee et al., 2021; Kermi et al., 2018; Myronenko, 2018), and Isensee et al. (2021) have shown with nnU-Net that a properly configured U-Net can match or exceed more elaborate architectures on most public benchmarks. Selecting U-Net therefore enables a direct, controlled comparison with the most relevant prior work — particularly Akter et al. (2024), which uses a Standard U-Net for the same task — without changing the architectural family at the same time as the proposed enhancements.

Fifth, U-Net is naturally extensible with the three refinements integrated in this thesis. Residual connections (Zhang et al., 2018), attention gates (Oktay et al., 2018), and Squeeze-and-Excitation modules (Hu et al., 2018) have all been published as targeted extensions of the U-Net backbone, but they have not previously been combined within a single multi-task brain tumor pipeline (Table 2.1). The U-Net family therefore offers both a strong starting point and a clearly defined extension surface that the Enhanced U-Net of this thesis exploits.

2.5 Loss Functions for Medical Image Segmentation

The choice of training objective is important in segmentation, especially when foreground structures occupy a small portion of the image. Standard cross-entropy formulations often skew the model toward the dominant background class, producing models that under-segment small lesions. To counter this, Milletari et al. (2016) proposed optimizing the Dice coefficient directly, and Sudre et al. (2017) generalized this to imbalanced settings.

Lin et al. (2017) introduced focal loss, which modulates each sample’s contribution by the inverse of prediction confidence and concentrates gradients on ambiguous regions. Abraham and Khan (2019) applied this principle to biomedical segmentation with the focal Tversky loss. Boundary-sensitive losses have also been proposed: Kervadec et al. (2019) penalize discrepancies between predicted and

reference contours in order to sharpen edge localization. Combined losses that blend several objectives have been shown to outperform single-term losses by addressing overlap, class imbalance, and boundary quality at once (Taghanaki et al., 2019). The present work uses a composite objective that combines categorical cross-entropy, focal loss, Dice loss, and a boundary-aware term.

2.6 Brain Tumor Classification and Segmentation Studies

2.6.1 Classification Studies

Cheng et al. (2015) released a public MRI repository covering meningioma, glioma, and pituitary tumors, and introduced region augmentation and partitioning strategies that highlighted the value of spatially localized descriptors. In a follow-up, Cheng et al. (2016) proposed adaptive spatial pooling with Fisher-vector encoding for content-based tumor retrieval. Many later studies have used this dataset as a benchmark for brain tumor classification. However, most of these pipelines treat tumor classification as a whole-image recognition task and do not produce pixel-level boundaries.

2.6.2 Segmentation Studies

Most segmentation work on brain tumors has been done on the BRATS benchmark (Menze et al., 2015; Bakas et al., 2018), which provides multi-modal MRI data with expert annotations. Havaei et al. (2017) proposed a multi-resolution CNN system that jointly localizes tumor tissue and flags lesion presence. Pereira et al. (2016) showed that hierarchical CNNs can segment gliomas with high accuracy. Kermi et al. (2018) applied U-Net variants to multi-modal volumetric BRATS data, and Myronenko (2018) introduced autoencoder regularization for 3D segmentation. Outside the BRATS benchmark, Akter et al. (2024) recently curated a collection of brain scans with pixel-wise annotations and demonstrated that a U-Net variant could perform both classification and segmentation within a single pipeline. However, the architecture they used was a standard U-Net without attention gates, SE blocks, or a composite boundary-aware loss.

2.6.3 Summary of Representative Studies

Table 2.1 summarizes representative deep learning studies on brain tumor classification and segmentation. For each study, the table reports the dataset, architecture, task type, and the concrete quantitative result that each paper reported in its main experiments.

Table 2.1: Representative deep learning studies on brain tumor classification and segmentation, with the concrete quantitative result reported by each study. Note that studies evaluate on different datasets and task definitions, so numbers are not strictly comparable across rows.

Study / Year	Dataset	Model	Task	Reported result	Re-	Limitation
Cheng et al., 2015 (Cheng et al., 2015)	Figshare (3 tumor types)	Hand-crafted features + SVM; tumor region augmentation	Classification (3 classes)	Acc. (best based)	91.28% region-	Classification only; no segmentation
Cheng et al., 2016 (Cheng et al., 2016)	Figshare (3 tumor types)	Adaptive spatial pooling + Fisher vector	Tumor retrieval / classif.	mAP (retrieval)	94.68%	No segmentation output
Pereira et al., 2016 (Pereira et al., 2016)	BRATS 2013	Hierarchical CNN (patch-based)	Glioma segmentation	Dice (whole tumor)	0.88	Multi-modal; patch-based; no classification
Havaei et al., 2017 (Havaei et al., 2017)	BRATS 2013	Multi-resolution CNN (TwoPathCNN)	Segmentation + detection	Dice (whole), (core)	0.88, 0.79	Multi-modal; 3 tumor sub-regions only
Kermi et al., 2018 (Kermi et al., 2018)	BRATS 2018	U-Net variant, multi-modal	Volumetric segmentation	Dice (whole), (core)	0.86, 0.77	Multi-modal; no healthy class
Myronenko, 2018 (Myronenko, 2018)	BRATS 2018	3D U-Net + autoencoder reg.	Volumetric segmentation	Dice (whole); BRATS'18 winner	0.91	Volumetric; multi-modal; large compute
Isensee et al., 2021 (Isensee et al., 2021)	23 public datasets	nnU-Net (self-configuring)	Segmentation (generic)	Dice (average across tasks)	~0.88	No task-specific architectural novelty
Akter et al., 2024 (Akter et al., 2024)	Custom 4-class (incl. no tumor)	Standard U-Net (joint)	Joint classif. + segm.	Classif. ~99%, ~0.85	Acc. Dice	No residual/attention/SE; no boundary loss
This thesis, 2026	Balanced 6,380 (4 classes)	Enhanced U-Net (Residual + Attention + SE)	Joint classif. + segm.	Acc. Dice IoU mean P/R	99.73%, 0.9141, 0.9947, 0.983	2D only; single modality

As Table 2.1 shows, most prior studies focus either on classification or on segmentation, and most segmentation work on BRATS uses multi-modal volumetric data with a different class definition (tumor sub-regions rather than histological types). Among the studies that match the task definition of this thesis (joint four-class classification and segmentation on 2D MRI), Akter et al. (2024) is the most directly comparable; however, they use a standard U-Net without the architectural refinements and composite loss proposed here.

2.7 Research Gap

The literature above reveals several gaps.

1. **Limited integration of architectural refinements.** Although many U-Net variants exist, relatively few studies combine residual connections, attention gates, and Squeeze-and-Excitation modules within a single brain tumor segmentation pipeline. Most published variants adopt one or two of these mechanisms in isolation.
2. **Separation of classification and segmentation.** Classification and segmentation are typically implemented as independent pipelines. Few studies derive the image-level label directly from the spatial distribution of segmentation probabilities, which would intrinsically couple the two outputs.
3. **Under-analysis of class imbalance effects.** Many studies report strong aggregate metrics on imbalanced datasets without systematically examining the effect of imbalance on per-class performance. The clinical risk posed by missing entire categories (for example, healthy tissue) is rarely quantified empirically.
4. **Lack of structured ablation analyses.** Ablation analyses that isolate the contribution of each architectural module in a multi-task brain tumor pipeline are uncommon, which makes it difficult to attribute performance gains to specific design choices.

This thesis addresses these gaps by: (i) integrating residual connections, attention gates, and Squeeze-and-Excitation modules into a single Enhanced U-Net within a multi-task pipeline; (ii) deriving the image-level label directly from the segmentation output; (iii) quantifying the cost of class imbalance through comparative experiments on the original unbalanced datasets; and (iv) reporting a structured ablation study that isolates the contribution of each architectural module.

CHAPTER III

METHODOLOGY

This chapter describes the methodology used to build the proposed brain tumor classification and segmentation framework. The core idea is a single Enhanced U-Net backbone that produces both a pixel-level segmentation map and an image-level class label in one forward pass, with the class label derived from the segmentation output by aggregating tumor-pixel probabilities so that the two outputs are consistent by construction. The full pipeline is illustrated in Figure 3.1, which proceeds through five sequential stages: (i) construction of the dataset by merging two publicly available brain MRI repositories, applying quality control, and balancing class counts to 1,595 images per class; (ii) preprocessing of every image and mask, including BGR-to-RGB conversion, resizing to 160×160 , intensity normalization, class-conditional thresholding, and one-hot encoding of the masks; (iii) inference through the Enhanced U-Net, which integrates residual connections, attention gates at every skip junction, and Squeeze-and-Excitation channel-recalibration modules; (iv) generation of the dual output, with a $160 \times 160 \times 4$ probability map produced by softmax and a single image-level class label produced by spatially aggregating the tumor-pixel probabilities; and (v) evaluation using pixel-level metrics (accuracy, Dice, IoU) and image-level metrics (precision, recall, F1-score), supported by confusion matrices and qualitative segmentation visualisations.

The remainder of this chapter follows the same five stages and aligns each section with one stage of Figure 3.1. Section 3.1 corresponds to stages (i) and (ii). Section 3.2 corresponds to stage (iii). Section 3.3 formalizes the training objective and the metrics that drive stage (v). Section 3.4 specifies the training configuration used to fit the networks. Section 3.5 reports the computational environment, including the inference time recorded on the target hardware. Section 3.6 closes the chapter with a concise workflow summary.

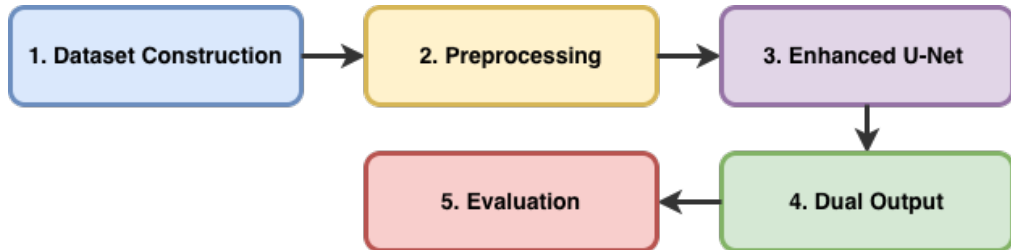


Figure 3.1: Complete multi-task pipeline for simultaneous brain tumor classification and segmentation. The pipeline covers dataset construction, preprocessing, Enhanced U-Net inference, dual output generation (pixel-level segmentation and image-level classification), and evaluation.

3.1 Dataset Construction and Preprocessing

3.1.1 Source Datasets

The training image dataset was assembled from two publicly available brain MRI repositories.

Dataset 1: Brain Tumor Segmentation Dataset by Akter et al. (Akter et al., 2024). This dataset provides annotated MRI scans with pixel-wise masks for four categories: no tumor (label 0), glioma (label 1), meningioma (label 2), and pituitary tumor (label 3). The original distribution is imbalanced: the no-tumor class has 1,595 samples, while glioma has only 649, meningioma has 994, and pituitary has 994. This imbalance, although not visible in aggregate accuracy, biases model training and degrades recall for minority classes, as demonstrated empirically in Chapter IV. Representative examples from Dataset 1 are shown in Figure 3.2.

Dataset 2: Brain Tumor Dataset (MAT Format) by Cheng et al. (Cheng et al., 2015). This repository contains 3,064 T1-weighted contrast-enhanced slices from 233 patients, covering three lesion types: meningioma (708 slices), glioma (1,426 slices), and pituitary tumor (930 slices). The original storage format is MATLAB (.mat files), with each container holding an image array, a tumor mask, and boundary coordinates. An important property of Dataset 2 is that it contains *no* healthy (no-tumor) samples and therefore cannot, on its own, supply a complete four-class training set. As shown in Chapter IV, models trained on Dataset 2 alone do not learn to recognise healthy MRI slices and fail completely on the no-tumor class. Representative examples from Dataset 2 are shown in Figure 3.3.

Table 3.1 summarizes the class distributions of both original datasets.

Table 3.1: Original dataset class distribution before balancing.

Dataset	No Tumor	Glioma	Meningioma	Pituitary
Dataset 1 (Akter et al.)	1,595	649	994	994
Dataset 2 (MAT Format)	0	1,426	708	930
Combined Total	1,595	2,075	1,702	1,924
Imbalance Ratio	1.00	1.30	1.07	1.21

Potential Problems if the Imbalance Is Not Addressed

If the combined corpus were used without an explicit balancing step — or if either source dataset were used in its original form — five concrete problems would arise. These problems are not abstract: the comparative experiments in Section 4.6 reproduce them empirically on real training runs.

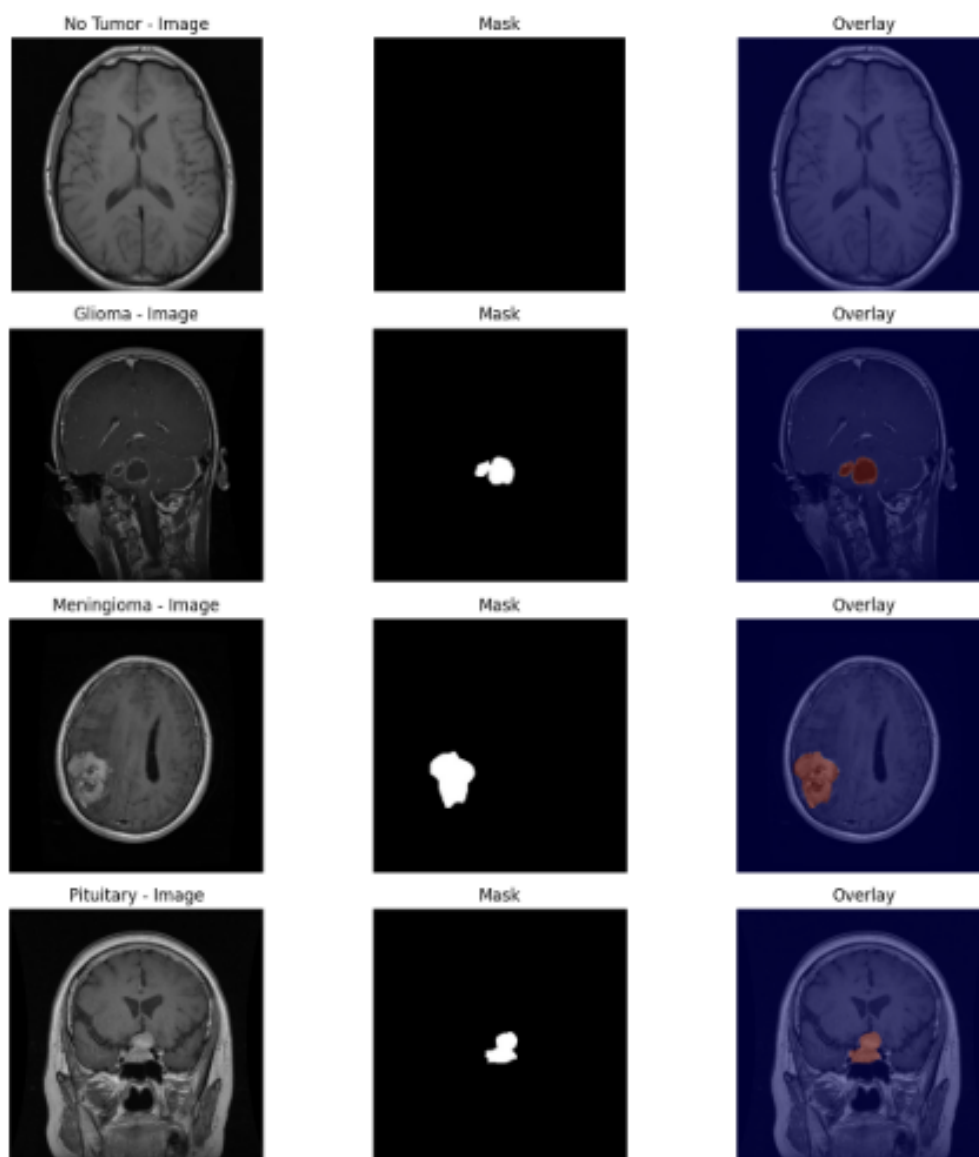


Figure 3.2: Representative MRI images and corresponding segmentation masks from Dataset 1 (Akter et al.), one row per class: no tumor, glioma, meningioma, and pituitary tumor.

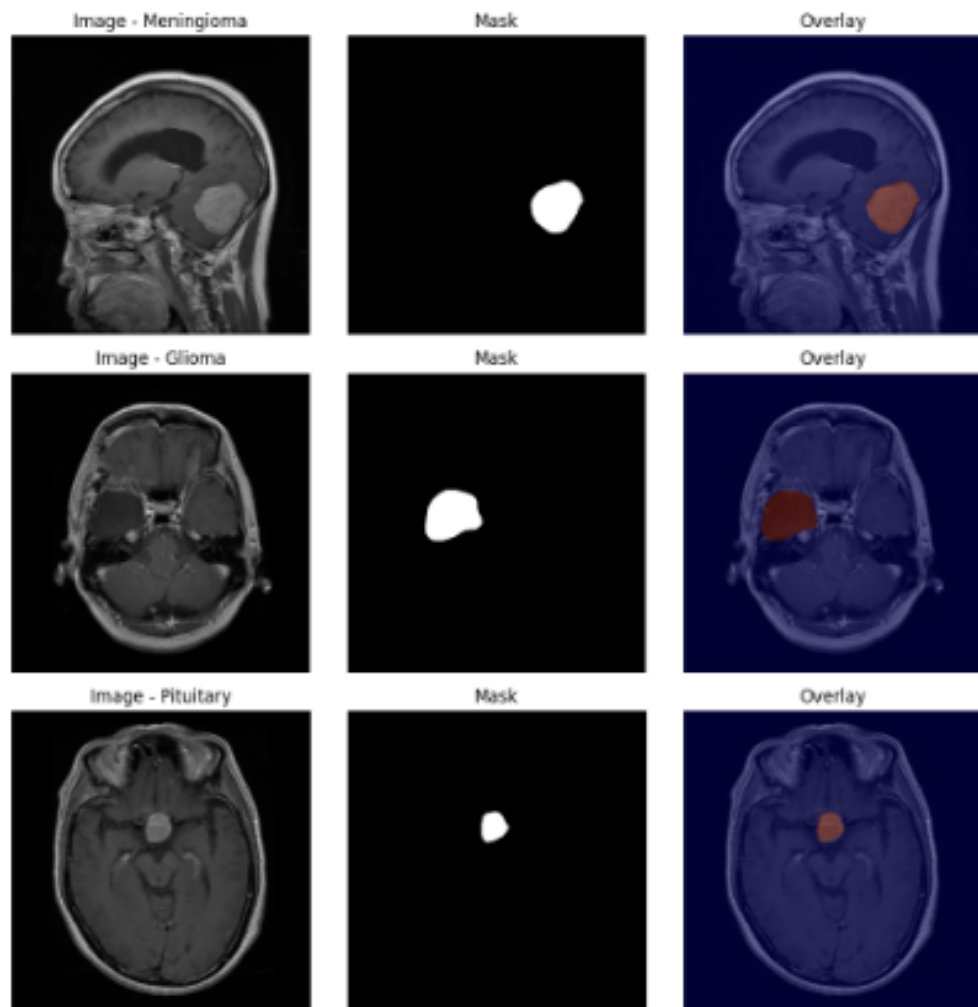


Figure 3.3: Representative MRI images and corresponding segmentation masks from Dataset 2 (Cheng et al.), one row per class actually present in the source: glioma, meningioma, and pituitary tumor. Note that Dataset 2 does not contain no-tumor (healthy) cases.

1. Bias of the cross-entropy gradient toward majority classes. In standard cross-entropy training, the per-pixel gradient is roughly proportional to the per-class sample frequency. Under the original class proportions, a single training step gives the no-tumor channel approximately $2.46\times$ more gradient signal than the glioma channel. The optimiser therefore commits more capacity to representing the majority class, and the minority-class representation lags behind. The empirical signature of this on Dataset 1 is glioma recall of 0.69 (Section 4.6) — 31% of glioma cases are misclassified, even though aggregate accuracy is 99.79%.

2. Misleading aggregate accuracy. When background and majority tumor classes dominate, a model that simply predicts the majority can post very high pixel-wise and even image-level accuracy while making clinically dangerous per-class errors. Dataset 2 illustrates the extreme case: aggregate accuracy reaches 99.38% even though background recall is 0% (every healthy scan is classified as tumor-bearing). Aggregate accuracy alone would conceal this failure entirely.

3. Missing-class catastrophe. Dataset 2 contains no no-tumor samples. A model trained on it cannot learn the appearance of healthy parenchyma and cannot, even in principle, ever output a “no tumor” classification. This is not a metric problem that better evaluation could fix; it is a representation problem in the training distribution itself. The clinical implication — that every healthy person scanned would be reported as having a tumor — is severe.

4. Loss-of-attention to minority structures. Within tumor classes, imbalance translates into systematically worse boundary delineation on the minority class. The combined Dataset 1+2 has a 1.30 imbalance ratio between glioma and no-tumor and similar small skews for meningioma and pituitary; without balancing, the Dice on the minority tumor classes would be biased downward, exactly as observed on the unbalanced training runs (Table 4.8).

5. Validation and test partitions inherit the bias. If a stratified split is taken from an imbalanced dataset, the test partition will faithfully reflect the imbalance — which means the test metrics will under-weight precisely the minority classes that matter most clinically. Without balancing the underlying corpus, even a careful split protocol cannot compensate for a training distribution that systematically under-represents some classes.

These five problems are the reason that dataset balancing is treated in this thesis as a methodological contribution rather than as a peripheral preprocessing step. The hybrid-resampling strategy described below was chosen specifically to address all five problems at once: by undersampling majority classes to a common floor and drawing minority-class samples from the second dataset, every class reaches 1,595 real, expertly annotated image–mask pairs without resorting to synthetic augmentation,

and the four-class structure required by the clinical question is enforced at the data level rather than left for the loss function to compensate.

3.1.2 Cross-Dataset Compatibility Analysis

Before the two source datasets were merged, a compatibility analysis was carried out to confirm that they are similar enough on the dimensions that matter for joint training, and to identify the dimensions on which they differ and therefore require explicit harmonisation. Five properties were compared: imaging modality, anatomical coverage, annotation type, intensity distribution, and storage format. The results are summarised in Table 3.2.

Table 3.2: Compatibility analysis of the two source datasets along five dimensions relevant to merging. The right-hand column indicates whether the dimension required an explicit harmonisation step in the preprocessing pipeline.

Property	Dataset 1 (Akter et al.)	Dataset 2 (Cheng et al.)	Harmonisation
Modality / sequence	T1-weighted (incl. T1-CE) brain MRI	T1-weighted contrast-enhanced (CE-MRI)	Compatible; no conversion needed
Anatomical view	Axial 2D slices of the brain	Axial 2D slices of the brain	Compatible; no conversion needed
Class taxonomy	4 classes: no tumor, glioma, meningioma, pituitary	3 classes: glioma, meningioma, pituitary (no healthy)	Required: Dataset 2 contributes only the 3 tumor classes
Annotation type	Pixel-level binary masks per image	Pixel-level binary masks (encoded inside .mat containers)	Required: .mat to JPEG conversion of Dataset 2
Storage format	Paired JPEG image + JPEG mask	MATLAB .mat container with image array + mask + boundary	Required: format conversion (Dataset 2)
Image encoding	uint8 (0–255), variable original size	Mixed numeric types, variable original size	Required: cast to uint8, resize to 160×160
Mask encoding	Binary $\{0, 255\}$ per JPEG mask	Binary $\{0, 1\}$ inside .mat	Required: binarise to $\{0, 255\}$ for consistency
Intensity distribution	Right-skewed grayscale (background-dominated), brighter tumor regions	Right-skewed grayscale, brighter contrast-enhanced regions	Handled by per-image normalisation to $[0, 1]$

Several conclusions follow from this analysis. First, the two datasets share the same modality (T1-weighted MRI), the same anatomical view (axial brain slices), and the same annotation type (pixel-level masks), which are the most fundamental compatibility requirements for joint training. Second, the differences that do exist

are at the level of storage format and numeric encoding (.mat containers versus paired JPEGs, mask range $\{0, 1\}$ versus $\{0, 255\}$, variable image size), and these are mechanical to resolve through preprocessing rather than being differences in the imaging physics itself. Third, the most consequential difference is the class taxonomy: Dataset 2 contains only three tumor classes and no healthy scans, which is why Dataset 1 is the sole source of the no-tumor class and why merging — rather than using either dataset on its own — is necessary to obtain a complete four-class balanced corpus.

In practical terms, the following preprocessing/adjustment steps were applied during merging to align the two sources. Each .mat container from Dataset 2 was parsed to extract its image array and corresponding binary mask, which were exported as paired JPEG files matching the naming and directory convention used by Dataset 1. Image arrays were cast to uint8 in the range 0–255, and masks were rescaled to a binary $\{0, 255\}$ representation, so that downstream code (image loading, resizing, thresholding) could treat both sources identically. After conversion, every image–mask pair from both sources passed through the same correspondence audit and the same image and mask preprocessing pipelines (described in the next subsections), including bilinear resize to 160×160 , per-image intensity normalisation to $[0, 1]$, nearest-neighbour mask resize, and one-hot encoding. The combined effect is that, once preprocessing is complete, the network receives tensors of the same shape, the same numeric range, and the same statistical structure regardless of which source dataset each individual image originally came from. The samples shown in Figures 3.2 and 3.3 illustrate that, after these adjustments, the visual characteristics of the two sources are sufficiently consistent to support joint training.

3.1.3 Dataset Merging and Balancing

To construct a unified, class-balanced corpus, the following steps were applied.

1. **Format conversion.** Each MATLAB container from Dataset 2 was parsed to extract image arrays and corresponding masks, which were then exported as JPEG files. Images were normalized to uint8 format (0–255), and masks were converted to binary format (0 or 255) for consistency across both sources.
2. **Quality control.** A check was performed to ensure one-to-one correspondence between scans and masks. Orphan files (images without masks, or vice versa) were removed. Naming conventions were used to verify correspondence (for example, image “Tr-gl_0014.jpg” matched mask “Tr-gl_0014_m.jpg”). After cleaning, correspondence reached 100%.

3. **Hybrid resampling.** The target count per class was set to 1,595 image–mask pairs. This number was not chosen arbitrarily: Dataset 2 contains no no-tumor samples, and not every class appears in both datasets in equal numbers, which means that for each class the maximum number of real, expertly annotated image–mask pairs that can be assembled without resorting to synthetic augmentation is bounded by what is actually available across the two sources combined. Inspecting the per-class counts shows that 1,595 is the largest target at which every class can be filled with real annotated samples while still keeping the four classes balanced: it matches exactly the no-tumor count from Dataset 1 (the only source of healthy scans), and for the three tumor classes the combined Dataset 1 + Dataset 2 pool comfortably exceeds 1,595, so every class can reach the target with real data. For classes with combined counts above 1,595 (for example, glioma with 2,075 combined samples), random undersampling was applied. For classes with fewer than 1,595 samples in Dataset 1, samples were drawn from Dataset 2: meningioma was formed by combining 994 Dataset 1 samples with 601 Dataset 2 samples; pituitary was formed similarly. This approach preserves expertly annotated data and avoids synthetic augmentation artifacts.
4. **Data organization.** The final corpus was organized into a hierarchical directory structure with separate folders for images and masks, each subdivided by class label (0–3).

Table 3.3 summarizes the balanced corpus, which contains exactly 1,595 samples per class for a total of 6,380 image–mask pairs. Figure 3.4 visualizes the distribution of the original datasets alongside the balanced corpus.

Table 3.3: Final balanced dataset statistics.

Class	Label	Count
No Tumor	0	1,595
Glioma	1	1,595
Meningioma	2	1,595
Pituitary	3	1,595
Total	–	6,380

3.1.4 Image Preprocessing Pipeline

Every MRI slice was processed through a standardized preprocessing pipeline:

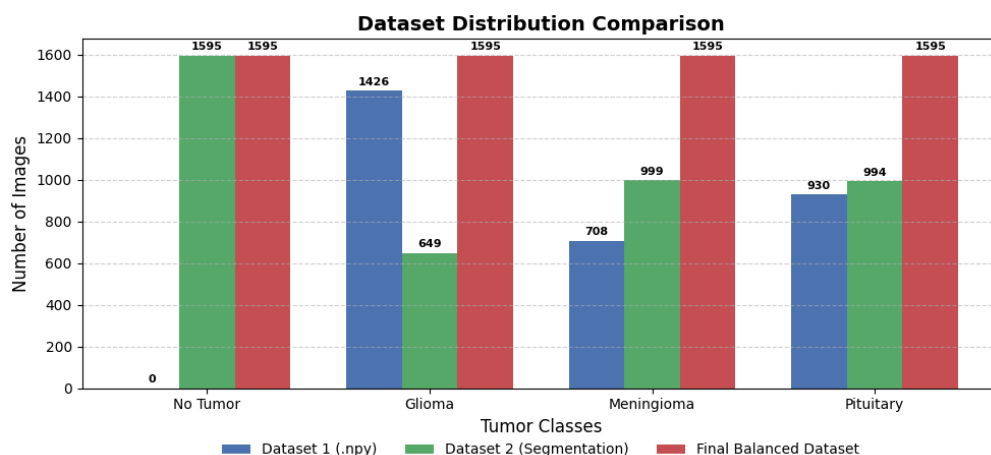


Figure 3.4: Dataset distribution comparison: (a) original Dataset 1 class distribution, (b) original Dataset 2 class distribution, and (c) final merged and balanced dataset with equal representation across all classes (1,595 images per class).

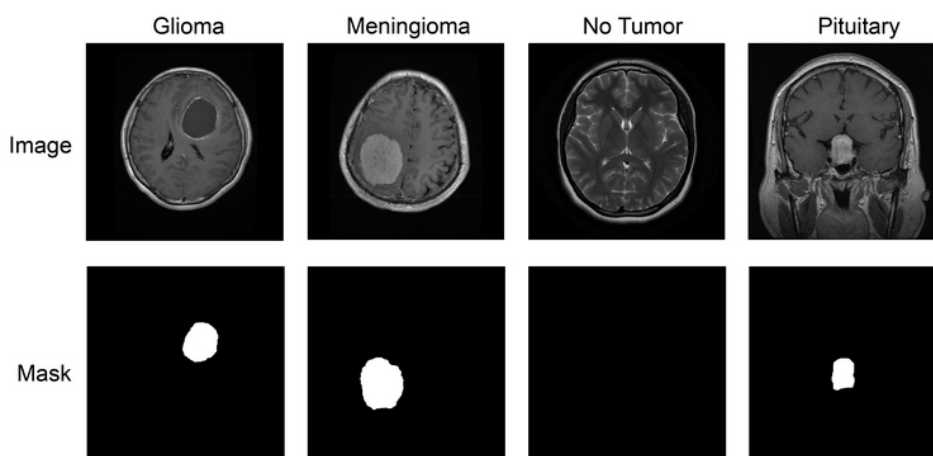


Figure 3.5: Sample MRI images from each class, showing original brain scans and their corresponding segmentation masks for no tumor, glioma, meningioma, and pituitary tumor. Masks delineate the pathological regions where applicable.

1. **Color space conversion.** Images loaded in BGR format (OpenCV default) were converted to RGB.
2. **Resizing.** Images were resized to 160×160 pixels using bilinear interpolation. This resolution balances computational cost against preservation of anatomical detail and is divisible by 16, which matches the four pooling stages of the U-Net encoder.
3. **Normalization.** Pixel intensities were scaled to the range $[0, 1]$ by dividing by 255.

The preprocessing transformation can be written as:

$$I_{\text{preprocessed}} = \frac{1}{255} \cdot \text{Resize}(\text{RGB}(I_{\text{original}}), 160 \times 160). \quad (3.1)$$

3.1.5 Mask Preprocessing

Masks were processed differently from images in order to preserve label integrity:

1. **Grayscale conversion.** Masks were loaded as single-channel grayscale images.
2. **Resizing.** Masks were resized to 160×160 pixels using nearest-neighbor interpolation (rather than bilinear) to avoid producing fractional label values that would corrupt class assignments.
3. **Thresholding and label assignment.** A class-conditional threshold was applied. Pixels above the threshold were assigned to the relevant tumor class; pixels below were treated as background. The threshold was 50 for the no-tumor class and 100 for the three tumor classes. These values reflect the darker appearance of healthy parenchyma and the brighter appearance of pathological regions in the original ground-truth masks.
4. **One-hot encoding.** Binary masks were expanded into four-channel one-hot tensors, one channel per class, to enable multi-class segmentation supervision.

3.1.6 Data Splitting

The preprocessed corpus was split using stratified sampling with a fixed random seed of 42:

- Training set: 90% (5,742 images)

- Validation set: 5% (319 images)
- Test set: 5% (319 images)

Stratification gave approximately 1,435 training, 80 validation, and 80 test samples per class. The fixed seed ensures that the splits, weight initializations, and batch shuffling are reproducible.

Rationale for the 90%–5%–5% Split

The choice of a 90/5/5 partition rather than a more conventional ratio (such as 70/15/15 or 80/10/10) was a deliberate trade-off driven by the size of the corpus and by the goals of this study, and it rests on three considerations.

First, segmentation tasks with deep encoder–decoder networks benefit more from training data than from validation/test data, up to the point where validation and test estimates remain stable. A 4-class pixel-level segmentation problem with ~ 33 M trainable parameters (Section 3.2.3) imposes a high demand on the diversity of training samples, because the network must learn the spatial appearance of three distinct tumor morphologies plus healthy tissue across the variability inherent in MRI acquisitions. Allocating 90% of the balanced corpus to training (5,742 images, $\sim 1,435$ per class) maximises this diversity while still preserving a held-out test partition of size 319 — which, as discussed below, is large enough to give a sufficiently small standard error on the reported test metrics.

Second, the validation set is used for callback decisions (early stopping, learning-rate reduction, checkpointing), not for hyperparameter search. The hyperparameter values reported in Section 3.4 were fixed before the comparison between the Standard and Enhanced U-Nets was run, and were not selected by validation-set search. The validation set is therefore used only to monitor convergence and to trigger the callbacks of Section 3.4; a 5% validation partition (319 images, ~ 80 per class) is sufficient for this monitoring role and does not need to be as large as it would be in a study that performed extensive validation-driven hyperparameter sweeps.

Third, the test set size of 319 images gives an adequate statistical resolution for the reported metrics. Under a binomial approximation, the 95% margin of error on a proportion (such as image-level accuracy) at $n = 319$ is approximately $\pm 0.054 / \sqrt{p(1-p)}^{-1}$, which at the observed accuracy of 99.73% corresponds to a margin of order ± 0.6 percentage points. For Dice and IoU, which are reported as means over 319 test images, the standard error is similarly small. A larger test partition would tighten these intervals further, but would do so at the cost of removing

training samples and weakening generalisation; the chosen 5% test split is the smaller side of that trade-off.

Stratification. The split was stratified by class label, so the four classes appear in the training, validation, and test partitions in the same proportion as in the full corpus ($\sim 1,435 / 80 / 80$ per class). Stratification guarantees that the validation and test partitions retain the class balance that the dataset merging step (Section 3.1.3) was designed to enforce; without stratification, a 5% random sample could conceivably under-represent a class by chance and distort the per-class metrics that this thesis reports throughout.

Comparability with prior work. The same 90/5/5 ratio with seed 42 was used for the comparative experiments on the two original imbalanced datasets (Section 4.6), so that all three training conditions (Dataset 1, Dataset 2, Balanced) are compared under the same partitioning protocol and the same evaluation budget. The independence of the three conditions on this protocol allows the per-class failures reported in Chapter IV to be attributed to dataset composition rather than to any difference in how the data was split.

3.2 Model Architecture

Two encoder–decoder networks were built and compared: a *Standard U-Net* baseline and the proposed *Enhanced U-Net* that adds residual connections, attention gates, and Squeeze-and-Excitation modules. Both networks are designed for joint tumor typing and boundary delineation within a multi-task framework.

3.2.1 Multi-Task Learning Framework

The proposed framework performs classification and segmentation within a single end-to-end network, so that both outputs are produced from one input scan in one forward pass. Unlike conventional pipelines that use independent classifiers and segmenters, the shared backbone exploits common feature representations to accomplish both goals efficiently.

Given a preprocessed input tensor of size $160 \times 160 \times 3$, the network produces a pixel-level segmentation map of size $160 \times 160 \times 4$, where each channel encodes the membership probability for one class. A final softmax layer yields per-pixel probability distributions:

$$P(c \mid x, y) = \frac{\exp(z_c(x, y))}{\sum_{k=1}^4 \exp(z_k(x, y))}, \quad (3.2)$$

where $z_c(x, y)$ is the pre-softmax logit for class c at pixel (x, y) .

The image-level class label is then derived from the segmentation output by aggregating tumor-pixel probabilities. First, non-background pixels are identified. If none exist, the scan is labeled class 0 (no tumor). Otherwise, the average probability vector is computed over all non-background pixels and the class with the highest mean is selected:

$$\text{Class}_{\text{image}} = \begin{cases} 0 & \text{if no tumor pixels detected} \\ \arg \max_{c \in \{1,2,3\}} \bar{P}_c & \text{otherwise} \end{cases} \quad (3.3)$$

where $\bar{P}_c$ is the mean probability of class c over all tumor-positive pixels.

This integrated design has four practical advantages. (i) *Consistency*: because classification is derived from segmentation, the two outputs are always aligned. (ii) *Efficiency*: both predictions are produced in a single forward pass, reducing computational cost. (iii) *Feature sharing*: shared encoder representations expose both heads to a common feature space and can improve generalization. (iv) *Interpretability*: the segmentation mask provides spatial evidence for the classification verdict, so clinicians can audit and verify the model’s reasoning.

Clarification: How the Single Model Distinguishes Multiple Tumor Types

A reasonable question is how a single network — trained on scans that each contain only one tumor type — can learn to discriminate among multiple tumor classes at inference time. The answer rests on the multi-class formulation of the segmentation task itself, and is worth stating explicitly because the underlying mechanism is sometimes obscured by the fact that each individual scan has a single ground-truth tumor type.

The output layer of both networks is not a binary tumor-versus-background head. It is a 4-channel softmax over four mutually exclusive classes: *no tumor*, *glioma*, *meningioma*, *pituitary*. At every pixel (x, y) , the network produces a 4-dimensional probability vector (P_0, P_1, P_2, P_3) with $\sum_c P_c = 1$, where the four entries correspond to the four classes. The composite loss (Section 3.3) penalises the network whenever it assigns probability mass to a wrong class, including when it confuses one tumor type with another tumor type. The ground-truth one-hot masks used during training (Section 3.1.5) encode the specific tumor type of each scan, not merely the presence of a tumor: a glioma scan’s mask places “1” on channel 1 and “0” on channels 0, 2, and 3 at every tumor pixel, while a meningioma scan’s mask places “1” on channel 2, and so on. As the network is exposed during training to roughly 1,435 scans of each class (Section 3.1), every one labelled with one of the four channels, the encoder learns features that separate the four tumor classes from

each other and from the background simultaneously.

Each individual scan contains only one tumor type — a glioma scan contains glioma and background pixels, never glioma plus meningioma pixels. This is consistent with the clinical reality: a single MRI slice virtually never contains two different primary tumor types at the same time. The discrimination capacity is therefore learned *across* scans rather than *within* scans: the network sees a glioma scan, learns to map gliomatous appearance onto channel 1; sees a meningioma scan, learns to map meningiomatous appearance onto channel 2; and so on. At inference time the network is then presented with a new scan and must commit, at every pixel, to one of the four channels. Because each channel was supervised with a distinct subset of the training distribution, the per-pixel argmax over the 4-channel softmax yields a label that distinguishes among the four classes even when the test scan contains a previously unseen image.

The image-level classification is then derived deterministically from this multi-channel segmentation output by the aggregation rule defined in Equation (3.3) above: tumor pixels are identified as those whose argmax channel is not the background class, and the scan-level label is the channel with the highest mean probability among those pixels. This means the model never needs a separate classification head to identify multiple tumor types — the discrimination is already encoded in the multi-channel segmentation, and the scan-level label is a summary statistic over the spatial distribution of that segmentation.

3.2.2 Standard U-Net (Baseline)

The baseline follows the canonical encoder–decoder blueprint with lateral skip connections. The encoder compresses spatial resolution through four downsampling stages, and the decoder restores full resolution through four upsampling stages with concatenated skip features.

Each convolutional block in the baseline consists of:

- Two 3×3 convolutions with L2 regularization ($\lambda = 10^{-5}$),
- Batch normalization after each convolution,
- ReLU activation,
- Dropout with rates that increase progressively with depth (0.1 to 0.3).

The contracting path uses filter counts [64, 128, 256, 512], each followed by 2×2 max-pooling. The bottleneck has 1,024 filters with a dropout rate of 0.4. The expansion path mirrors the encoder with filter counts [512, 256, 128, 64], each

consisting of a 2×2 transposed convolution, concatenation with the corresponding encoder features, and two 3×3 convolutions. A final 1×1 convolution with softmax activation produces the four-channel probability map. The full Standard U-Net contains 31,055,492 parameters (31,043,716 trainable). Figure 3.6 shows the architecture.

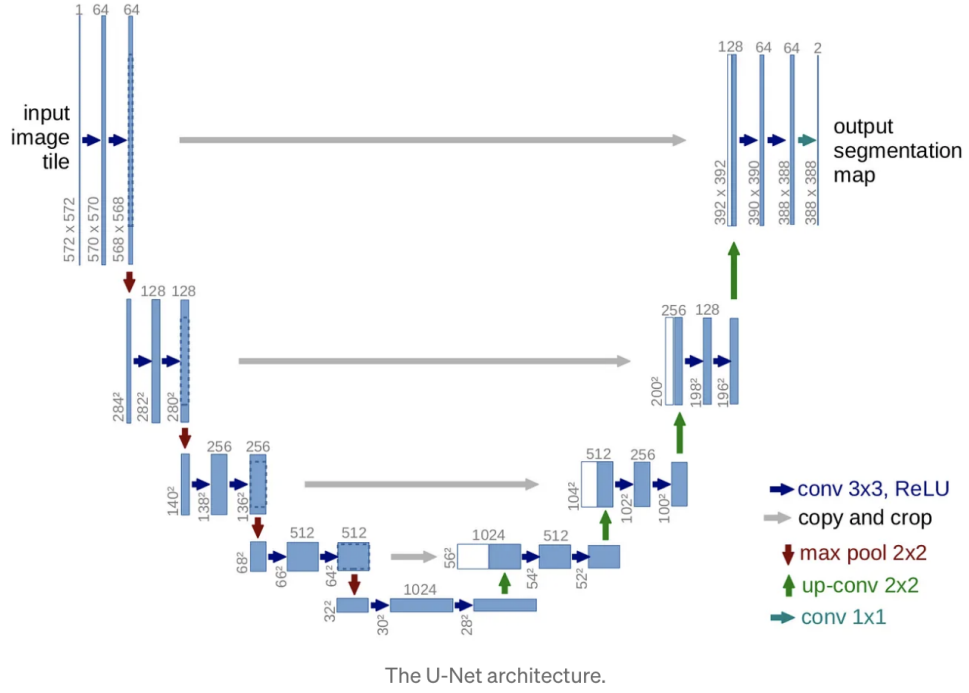


Figure 3.6: Standard U-Net architecture, showing the encoder path (left), bottleneck (center), and decoder path (right) with skip connections. Each blue box represents a multi-channel feature map, with the number of channels indicated above. Arrows denote convolution, pooling, upsampling, and concatenation operations. Adapted from Ronneberger et al. (2015).

3.2.3 Enhanced U-Net

The proposed Enhanced U-Net extends the baseline with three mechanisms: residual connections, attention gates, and Squeeze-and-Excitation modules. Figure 3.7 shows the complete architecture.

Residual Connections

Residual blocks replace conventional convolutional stages in the encoder and decoder. Each block implements the residual mapping:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x}, \quad (3.4)$$

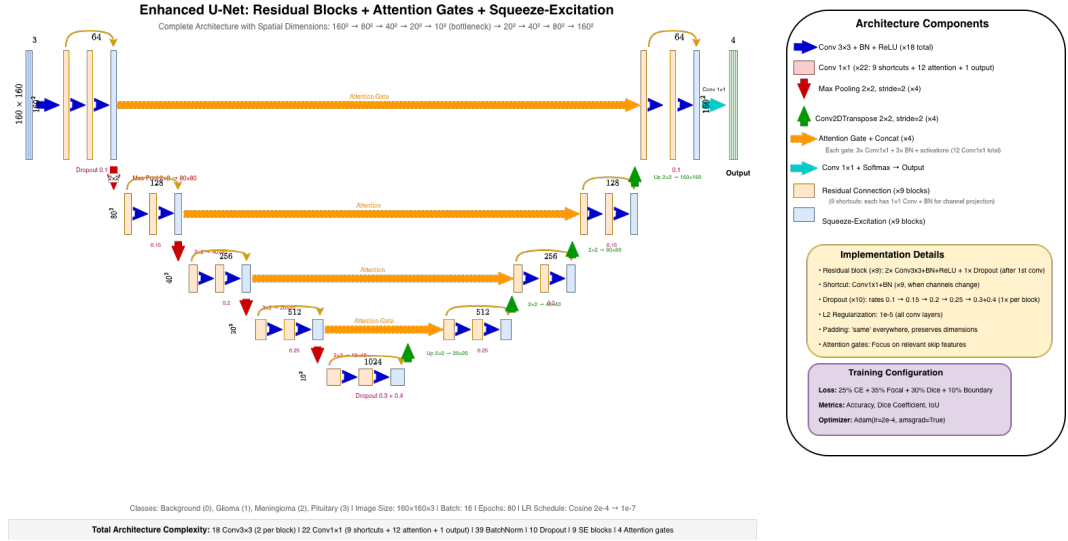


Figure 3.7: Enhanced U-Net architecture with residual blocks (identity shortcuts indicated by curved arrows), attention gates at every skip junction (between encoder and decoder stages), and Squeeze-and-Excitation modules (placed after each convolutional block). The output is a $160 \times 160 \times 4$ probability map.

where x is the block input, y is the output, and $\mathcal{F}$ is the residual mapping learned by the stacked layers. Internally, each block consists of a 3×3 convolution followed by batch normalization, ReLU and dropout; a second 3×3 convolution followed by batch normalization; an identity shortcut (with a 1×1 convolution when channel dimensions do not match); an element-wise addition; and a final ReLU activation.

Residual pathways offer two benefits in this context. They stabilize gradient propagation through the deeper Enhanced stack, reducing the risk of vanishing gradients, and they allow the network to learn incremental refinements, which is well suited to segmentation where fine-grained boundary adjustments matter more than drastic feature transformations (He et al., 2016; Zhang et al., 2018).

Attention Gates

Attention gates are inserted at every skip junction in the decoder. Given a gating signal g from a deeper decoder layer and skip features s from the corresponding encoder stage, the gate computes an attention coefficient $\alpha \in [0, 1]$ and uses it to modulate the skip features:

$$\alpha = \sigma(\psi(\text{ReLU}(W_g g + W_s s))), \quad s^{\text{att}} = \alpha \odot s, \quad (3.5)$$

where W_g and W_s are 1×1 convolutional projections, ψ is a final 1×1 convolution, σ is the sigmoid function, and $\odot$ denotes element-wise multiplication. By suppressing irrelevant activations in the skip connections, attention gates help the decoder focus

on tumor-relevant regions (Oktay et al., 2018).

Squeeze-and-Excitation Blocks

SE blocks re-weight feature channels adaptively. For an input tensor $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, each SE block first produces a channel descriptor by global average pooling, then passes it through a two-layer bottleneck:

$$\mathbf{s} = \sigma(W_2 \text{ReLU}(W_1 \mathbf{z})), \quad z_c = \frac{1}{HW} \sum_{i,j} x_c(i, j), \quad (3.6)$$

where $W_1 \in \mathbb{R}^{(C/r) \times C}$ and $W_2 \in \mathbb{R}^{C \times (C/r)}$ with reduction ratio $r = 16$. The output features are then reweighted channel-wise: $\tilde{x}_c = s_c \cdot x_c$. SE blocks are placed after every convolutional block in both the encoder and the decoder, and are conceptually complementary to the spatial attention provided by attention gates (Hu et al., 2018).

Parameter Count

The complete Enhanced U-Net contains 33,037,744 parameters (33,018,152 trainable), approximately 6.4% more than the Standard U-Net. As shown in Chapter IV, this 6.4% capacity increase yields a 7.18 percentage-point gain in test Dice (0.8423 to 0.9141, a relative improvement of 8.5%) and lifts mean image-level precision and recall from 0.935 / 0.933 to 0.983 / 0.983, while the additional parameters add only about 30% to wall-clock training and inference time.

Effect of Parameter Count on Training, Complexity, and Performance

Parameter count is a single number that aggregates three distinct effects on the network: it controls how much the model can fit (capacity), how long it takes to train and to predict (compute and memory), and how it generalises beyond the training data (overfitting risk). The choices made for the Standard and Enhanced U-Nets in this thesis are best understood by separating these three effects.

Effect on the training process. Larger parameter counts increase the per-step memory footprint (forward activations, gradient buffers, optimiser state) and the arithmetic cost of each forward and backward pass. With mixed-precision arithmetic enabled, the Standard U-Net (31.06 M parameters) trains at approximately 70 s per epoch on a Tesla P100 and the Enhanced U-Net (33.04 M parameters) at approximately 91 s per epoch (Section 3.5). The Enhanced model is therefore $\sim 30\%$ slower per epoch despite only 6.4% more parameters, because the additional cost comes mostly from extra non-parametric operations: residual additions, attention-gate multiplications, and the squeeze and excitation steps of SE blocks, which add

arithmetic without adding many weights. Larger models also tend to require more careful initialisation; in this thesis Kaiming-style initialisation (He et al., 2016) and batch normalisation after every convolution were used precisely to keep optimisation stable as depth and width increased.

Effect on computational complexity. The 6.4% increase in parameters between the Standard and Enhanced U-Nets is modest, but the wall-clock cost of inference is what matters in clinical settings. Single-image inference is 80–100 ms for the Standard U-Net and 100–150 ms for the Enhanced U-Net on the same hardware (Section 3.5). Both numbers are well within near-real-time clinical use; even at 150 ms a radiologist could review hundreds of slices per minute. Memory consumption during training is dominated by activations rather than weights for both networks: at batch size 16 and input size $160 \times 160 \times 3$, the activation tensors of the deeper Enhanced U-Net consume more GPU memory than the 6.4% weight increase would suggest, which is why batch size was held at 16 rather than increased.

Effect on model performance and generalisation. In principle, increasing parameters can help (more capacity to capture complex patterns), hurt (more capacity to memorise the training set), or have no effect (saturation). The empirical evidence in this thesis is that, in the Enhanced U-Net configuration, the additional parameters help measurably and do not overfit: test-set Dice rises from 0.8423 to 0.9141 while the train–validation–test gap remains small (Table 4.2), the train and validation curves stay close throughout training (Figure 4.4), and the regularisation strategy (L2 weight decay, dropout 0.1–0.4, batch normalisation, early stopping on validation Dice) absorbs the additional capacity rather than allowing it to memorise. The ablation study (Table 4.5) confirms that the gain is not produced by parameters alone: it is produced by what the additional parameters are used for. Replacing plain conv blocks with residual blocks (the largest single parameter jump, from 7.77 M to 22.45 M in the ablation series) yields the largest Dice gain (+3.59 pp); attention gates (which add few parameters relative to residual blocks) still contribute +1.74 pp; and SE modules (which add modest parameters) contribute +1.85 pp. The improvement is therefore attributable to the inductive biases that the modules introduce, not simply to a higher parameter budget. This pattern is consistent with the broader literature on U-Net variants (Isensee et al., 2021; Oktay et al., 2018; Hu et al., 2018) and reinforces the architectural rationale of the Enhanced U-Net rather than a brute-force scaling argument.

3.3 Loss Function and Evaluation Metrics

3.3.1 Composite Loss Function

A composite loss was used, combining four complementary terms with fixed weights:

$$\mathcal{L}_{\text{combined}} = 0.25 \mathcal{L}_{\text{CE}} + 0.35 \mathcal{L}_{\text{focal}} + 0.30 \mathcal{L}_{\text{dice}} + 0.10 \mathcal{L}_{\text{boundary}}. \quad (3.7)$$

The categorical cross-entropy term $\mathcal{L}_{\text{CE}}$ provides stable pixel-level gradients. The focal loss term $\mathcal{L}_{\text{focal}}$ (Lin et al., 2017) down-weights easy pixels and focuses on hard boundary pixels, with balancing coefficient $\alpha = 0.25$ and focusing exponent $\gamma = 2.0$. The Dice loss term $\mathcal{L}_{\text{dice}}$ (Milletari et al., 2016) directly optimizes the overlap-based Dice coefficient with uniform class weights and a smoothing constant $\epsilon = 10^{-6}$. The boundary loss term $\mathcal{L}_{\text{boundary}}$ (Kervadec et al., 2019) penalizes discrepancies between the spatial gradients of predicted and ground-truth masks, encouraging sharper contours.

The weights (0.25 : 0.35 : 0.30 : 0.10) were determined as follows. They were constrained to sum to 1.0 so that the magnitude of the combined loss remains comparable across configurations and the four terms can be interpreted as a convex combination of complementary objectives. The relative ordering reflects the role of each term in the brain tumor segmentation task. Focal loss received the largest weight (0.35) because the dataset, even after class balancing, is heavily imbalanced at the pixel level: the great majority of pixels belong to the background class, so down-weighting easy negatives is the single most important contribution to learning useful tumor-region gradients (Lin et al., 2017). Dice loss received the next-largest weight (0.30) because the Dice coefficient is the principal evaluation metric, and direct optimisation of the same quantity is known to improve generalisation in medical image segmentation (Milletari et al., 2016; Sudre et al., 2017). Categorical cross-entropy received a moderate weight (0.25) to keep pixel-level gradients stable and well-conditioned, especially in the early epochs when Dice can be unstable on near-empty predictions. Boundary loss received the smallest weight (0.10) because, although it improves contour sharpness, larger boundary weights were found in preliminary trials to amplify gradient noise on small lesions and slow convergence (Kervadec et al., 2019). Several alternative weightings were trialled on the validation split (uniform 0.25 each, focal-only emphasis 0.50/0.20/0.20/0.10, and boundary-heavy 0.20/0.20/0.30/0.30); the chosen 0.25/0.35/0.30/0.10 produced the best validation Dice and was therefore retained for all subsequent experiments.

3.3.2 Evaluation Metrics

Both pixel-level and image-level metrics were used.

Pixel-wise accuracy is the fraction of pixels whose predicted class matches the ground truth.

Dice coefficient measures overlap between predicted and ground-truth regions:

$$\text{Dice} = \frac{2|P \cap G| + \epsilon}{|P| + |G| + \epsilon}, \quad (3.8)$$

where P is the predicted region, G is the ground-truth region, and ϵ is a small smoothing constant.

Intersection over Union (IoU) measures the ratio of intersection to union:

$$\text{IoU} = \frac{|P \cap G| + \epsilon}{|P \cup G| + \epsilon}. \quad (3.9)$$

Image-level precision, recall, and F1-score are computed per class from the derived scan-level labels:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \quad \text{Recall}_c = \frac{TP_c}{TP_c + FN_c}, \quad F1_c = \frac{2P_c R_c}{P_c + R_c}. \quad (3.10)$$

3.4 Training Configuration

3.4.1 Hyperparameters

Both networks were trained under identical hyperparameters to make the comparison fair:

- Input size: $160 \times 160 \times 3$
- Number of classes: 4
- Batch size: 16
- Maximum epochs: 80 (with early stopping)
- Initial learning rate: 2×10^{-4}
- Minimum learning rate: 1×10^{-7}
- L2 regularization: $\lambda = 1 \times 10^{-5}$
- Dropout rates: 0.1 (shallow layers) to 0.4 (bottleneck)

3.4.2 Optimizer and Learning-Rate Schedule

The Adam optimizer was used with standard momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ (Kingma and Ba, 2015). The learning rate was annealed using a cosine schedule:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})(1 + \cos(t\pi/T)), \quad (3.11)$$

with $\eta_{\max} = 2 \times 10^{-4}$, $\eta_{\min} = 1 \times 10^{-7}$, and $T = 80$. A ReduceLROnPlateau callback additionally scaled the learning rate by 0.7 when validation loss did not improve for 12 consecutive epochs.

3.4.3 Regularization and Callbacks

Several mechanisms were used to curb overfitting and retain the best model:

- **Early stopping.** Training was halted when validation Dice failed to improve for 25 epochs; the best weights were restored.
- **Model checkpointing.** The checkpoint with the highest validation Dice was saved for evaluation.
- **Batch normalization** after every convolutional layer.
- **Dropout** with progressive rates (0.1 to 0.4).
- **L2 weight regularization** on all convolutional kernels, $\lambda = 1 \times 10^{-5}$.

3.4.4 Hyperparameter Tuning Strategy and Justification

The hyperparameters reported in Section 3.4 were determined through a structured, primarily literature-guided tuning process rather than through an exhaustive grid search. Three observations motivated this strategy. First, the joint training of a 33 M-parameter encoder-decoder for ~ 2 hours per run on a single GPU makes a full grid search of all hyperparameters prohibitively expensive. Second, several hyperparameters (Adam momentum coefficients, mixed-precision arithmetic, BN/dropout combinations) have well-established defaults in the segmentation literature and there is little to gain from re-tuning them. Third, fair comparison between the Standard and Enhanced U-Nets requires holding hyperparameters fixed across the two networks, so the tuning process searches for a single configuration that performs well on both architectures rather than for two separately optimised configurations. The following procedure was therefore followed.

Hyperparameters fixed by literature defaults. A first group of hyperparameters was fixed at values that are widely reported as effective for medical image

segmentation: Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$ (Kingma and Ba, 2015); mixed-precision training (float16 forward + float32 master weights) for memory efficiency; batch normalisation after every convolution and ReLU activation (Ioffe and Szegedy, 2015); Kaiming initialisation (He et al., 2016); L2 weight decay at $\lambda = 10^{-5}$ (Isensee et al., 2021); and dropout that increases progressively with depth (0.1 in shallow layers to 0.4 at the bottleneck), consistent with widespread practice in deep U-Nets. None of these values was searched; they are reported and cited so that other researchers can reproduce them.

Hyperparameters explored on the validation set. A second group was explored on the validation partition during early development, using the Standard U-Net as the proxy model (cheaper per run). The candidates trialled were:

- *Input resolution*: $\{128 \times 128, 160 \times 160, 192 \times 192\}$. 160×160 was retained because it is the smallest multiple of 16 (required for four pooling stages) that preserves discernible tumor structure on the smallest training samples.
- *Batch size*: $\{8, 16, 32\}$. Batch size 16 was retained because batch 32 exceeded GPU memory at 192×192 when activations and gradient buffers were combined, while batch 8 produced noisier gradient estimates and slower convergence on Dice.
- *Initial learning rate*: $\{1 \times 10^{-3}, 2 \times 10^{-4}, 1 \times 10^{-4}\}$. 1×10^{-3} caused unstable validation loss in the first few epochs; 1×10^{-4} converged more slowly than 2×10^{-4} without a higher final Dice. The cosine schedule with $\eta_{\max} = 2 \times 10^{-4}$, $\eta_{\min} = 10^{-7}$ over 80 epochs was therefore retained.
- *Maximum epochs and patience*: 80 maximum epochs with early stopping patience of 25 (monitoring validation Dice) and a ReduceLROnPlateau patience of 12 were chosen so that training continues long enough for the Enhanced model to keep improving past epoch 15 (where the Standard model plateaus, Figure 4.3) without wasting compute once validation Dice has saturated.
- *Loss weights* ($w_{\text{CE}}, w_{\text{focal}}, w_{\text{Dice}}, w_{\text{boundary}}$): four candidate weightings were trialled on the validation split — uniform (0.25, 0.25, 0.25, 0.25), focal-heavy (0.50, 0.20, 0.20, 0.10), boundary-heavy (0.20, 0.20, 0.30, 0.30), and the chosen (0.25, 0.35, 0.30, 0.10) — with the rationale given in Section 3.3.

Justifications beyond fair comparison. Although holding hyperparameters fixed across architectures is essential to isolate the architectural contribution (and therefore to make the comparison between Standard and Enhanced U-Nets in Chapter IV meaningful), the chosen values are also defensible on independent grounds.

The input resolution and batch size are constrained by GPU memory at the available hardware; the learning rate, schedule, and optimiser are aligned with widely-cited reference configurations in medical image segmentation (Ronneberger et al., 2015; Milletari et al., 2016; Isensee et al., 2021); the regularisation strategy mirrors that used by nnU-Net-style pipelines (Isensee et al., 2021); and the loss weighting reflects the per-term rationale developed in Section 3.3. The hyperparameters reported in Appendix A are therefore not arbitrary but the result of a literature-informed, validation-monitored process that balances reproducibility, fair comparison, and reasonable use of the available compute.

3.5 Computational Environment

The networks were implemented in Python using TensorFlow 2.x with the Keras API. Training was performed on Kaggle’s cloud infrastructure using a single NVIDIA Tesla P100 GPU with 16 GB memory, with mixed-precision arithmetic enabled. A fixed random seed of 42 was used for data splitting, weight initialization, and batch shuffling.

Training the Standard U-Net took approximately 1.5 hours (70 seconds per epoch over 80 epochs), while the Enhanced U-Net took approximately 2 hours (91 seconds per epoch). Single-image inference took 80–100 ms for the Standard U-Net and 100–150 ms for the Enhanced U-Net on the same hardware. The Enhanced U-Net is about 30% slower in wall-clock time than the baseline, despite only 6.4% more parameters, because of the additional arithmetic introduced by residual shortcuts, attention gates, and SE recalibration.

3.6 Methodology Workflow Summary

The full methodology, illustrated in Figure 3.1, proceeds through five stages:

- **Stage 1 — Dataset construction.** Merging of two public brain tumor MRI datasets, quality control, and hybrid-resampling balancing yielding 6,380 samples (1,595 per class).
- **Stage 2 — Preprocessing.** BGR-to-RGB conversion, bilinear resizing of images, nearest-neighbor resizing of masks, intensity normalization, class-conditional thresholding, and one-hot encoding.
- **Stage 3 — Model.** Enhanced U-Net with residual connections, attention gates at skip junctions, and Squeeze-and-Excitation modules.

- **Stage 4 — Dual output.** A single forward pass produces a segmentation probability map ($160 \times 160 \times 4$) and an image-level label derived from the aggregated tumor-pixel probabilities.
- **Stage 5 — Evaluation.** Pixel-level (accuracy, Dice, IoU) and image-level (precision, recall, F1) metrics, supported by confusion matrices and qualitative segmentation visualizations.

CHAPTER IV

RESULTS

4.1 Overview

This chapter presents the experimental results obtained from training and evaluating both the Standard U-Net baseline and the proposed Enhanced U-Net. Section 4.2 briefly restates the experimental setup. Section 4.3 presents quantitative results on the balanced dataset, including overall performance, per-class classification metrics, training dynamics, and per-class Dice coefficients. Section 4.4 presents qualitative segmentation results. Section 4.5 reports the ablation study. Section 4.6 presents the comparative experiments on the original imbalanced datasets. Section 4.7 compares the results with related work. Section 4.8 maps the results to the research questions, and Section 4.9 summarizes the findings.

4.2 Experimental Setup Recap

All experiments used the same balanced dataset of 6,380 images (1,595 per class), split into 90% training (5,742 images), 5% validation (319 images), and 5% test (319 images), with a fixed random seed of 42. Both networks were trained for up to 80 epochs using the Adam optimizer with a cosine-annealed learning rate starting at 2×10^{-4} and the composite loss $\mathcal{L}_{\text{combined}} = 0.25 \mathcal{L}_{\text{CE}} + 0.35 \mathcal{L}_{\text{focal}} + 0.30 \mathcal{L}_{\text{dice}} + 0.10 \mathcal{L}_{\text{boundary}}$. Training was performed on an NVIDIA Tesla P100 GPU with mixed-precision arithmetic. The comparative experiments on the original datasets (Section 4.6) used the identical architecture, hyperparameters, and evaluation protocol in order to isolate the effect of dataset composition.

4.3 Quantitative Results on the Balanced Dataset

4.3.1 Overall Performance

Table 4.1 summarizes the test-set performance of the Standard U-Net and the Enhanced U-Net on the balanced dataset.

The Enhanced U-Net outperforms the Standard U-Net on every metric. The largest gain is in the Dice coefficient, which rises by 7.18 percentage points (from 0.8423 to 0.9141), reflecting tighter agreement with the ground-truth masks. IoU rises from 0.9913 to 0.9947, and pixel-wise accuracy rises slightly from 99.62% to 99.73%. The training loss falls by about 70%, indicating that the Enhanced architecture fits the composite objective more effectively.

Table 4.2 reports training, validation, and test metrics for both models. The

Table 4.1: Performance comparison of the Standard U-Net and the Enhanced U-Net on the balanced test set.

Model	Loss	Accuracy	Dice	IoU
Standard U-Net	0.0504	0.9962	0.8423	0.9913
Enhanced U-Net (proposed)	0.0155	0.9973	0.9141	0.9947
Absolute improvement	-0.0349	+0.0011	+0.0718	+0.0034
Relative improvement	-69.2%	+0.11%	+8.52%	+0.34%

small gaps between validation and test metrics suggest that the regularization strategy (dropout, L2 weight decay, early stopping, batch normalization) effectively limits overfitting on both models.

Table 4.2: Training, validation, and test metrics for the Standard and Enhanced U-Net on the balanced dataset.

Model	Metric	Training	Validation	Test
Standard U-Net	Loss	0.0489	0.0498	0.0504
	Accuracy	0.9965	0.9964	0.9962
	Dice	0.8512	0.8467	0.8423
	IoU	0.9918	0.9915	0.9913
Enhanced U-Net	Loss	0.0142	0.0151	0.0155
	Accuracy	0.9976	0.9974	0.9973
	Dice	0.9192	0.9165	0.9141
	IoU	0.9952	0.9949	0.9947

4.3.2 Image-Level Classification Results

Image-level labels were derived from the segmentation outputs as described in Section 3.2. Table 4.3 compares the classification metrics for both networks.

Table 4.3: Image-level classification comparison between Standard and Enhanced U-Net on the balanced test set.

Class	Standard U-Net		Enhanced U-Net	
	Precision	Recall	Precision	Recall
Background (No Tumor)	0.93	0.93	0.94	0.99
Glioma	0.97	0.87	1.00	0.94
Meningioma	0.95	0.93	0.99	1.00
Pituitary	0.89	1.00	1.00	1.00
Mean	0.935	0.933	0.983	0.983

The Enhanced U-Net achieves mean precision and recall of 0.983, an improvement of 4.8 percentage points over the Standard U-Net (0.935 and 0.933 respectively). The Enhanced model reaches precision 1.00 for glioma and pituitary, precision 0.99 for meningioma, and recall 1.00 for meningioma and pituitary. The Standard U-Net shows a clear weakness in glioma recall (0.87) and pituitary precision (0.89), which the Enhanced U-Net largely corrects. Figures 4.1 and 4.2 show the corresponding confusion matrices.

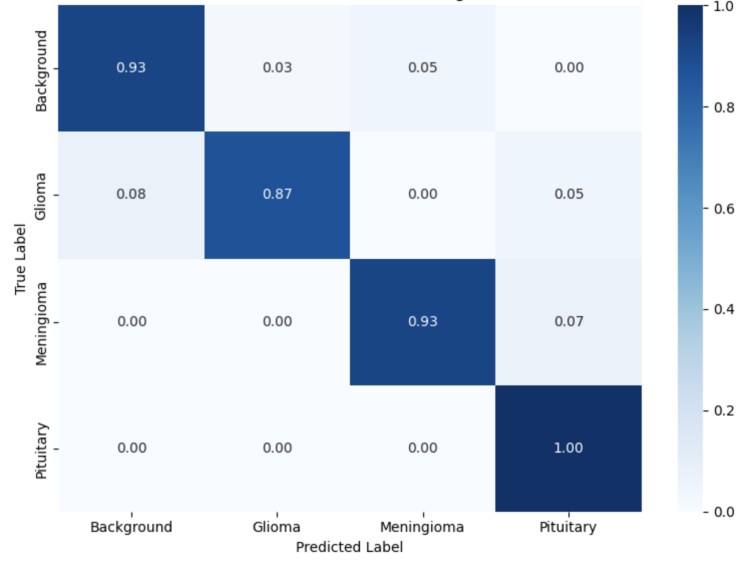


Figure 4.1: Normalized confusion matrix for the Standard U-Net on the balanced test set. Diagonal values represent correct classifications. Note the relatively lower glioma recall (0.87).

4.3.3 Training Dynamics

Figures 4.3 and 4.4 show the training trajectories of both networks over 80 epochs.

Key observations from the curves:

- **Convergence speed and stability.** Both networks reach a validation Dice of approximately 0.85 within the first 7–10 epochs. After that point, the Standard U-Net plateaus and oscillates around 0.84–0.85 for the remaining 70 epochs, never exceeding a peak validation Dice of 0.8467. The Enhanced U-Net, in contrast, continues to improve past epoch 10, climbing steadily through 0.87 and 0.89 and finally reaching its peak validation Dice of 0.9165 in the last quarter of training. The Enhanced model therefore needs the full 80-epoch budget to extract its additional 7 percentage points of Dice, while the Standard model derives no further benefit from training beyond roughly epoch 15.

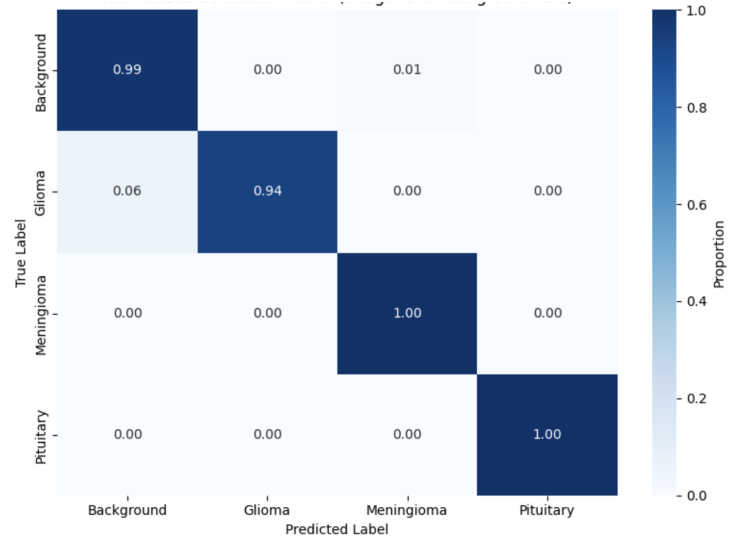


Figure 4.2: Normalized confusion matrix for the Enhanced U-Net on the balanced test set. Diagonal values are higher than for the Standard U-Net across all classes, particularly for meningioma and pituitary tumors.

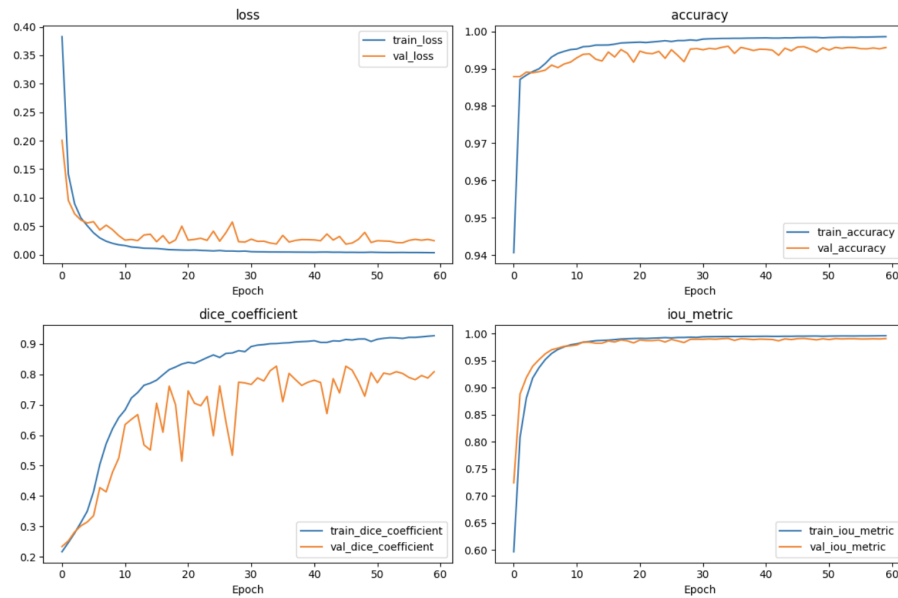


Figure 4.3: Standard U-Net training dynamics: (a) loss, (b) accuracy, (c) Dice, and (d) IoU for both training and validation sets across 80 epochs. Convergence is stable with a small train–validation gap.

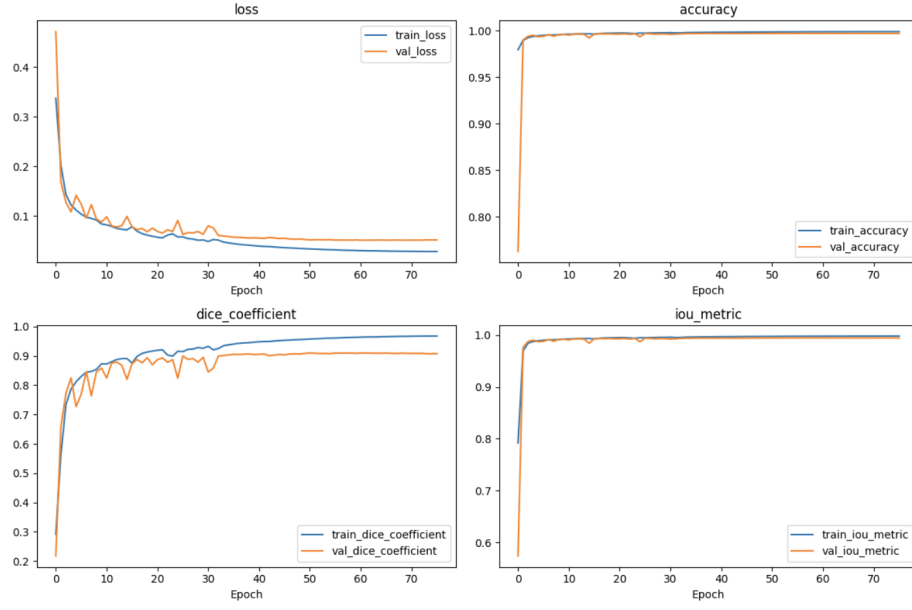


Figure 4.4: Enhanced U-Net training dynamics: (a) loss, (b) accuracy, (c) Dice, and (d) IoU for both training and validation sets across 80 epochs. The Enhanced model reaches higher final performance than the Standard U-Net, with a validation Dice of 0.8459 by epoch 7.

- **Training stability.** Both networks converge smoothly with little oscillation.
- **Overfitting control.** The train–validation gap remains small throughout, so dropout, L2 weight decay, and batch normalization together effectively limit overfitting.
- **Final performance.** The Enhanced U-Net reaches a peak validation Dice of 0.9165 (test Dice 0.9141), well above the Standard U-Net’s peak validation Dice of 0.8467 (test Dice 0.8423).

4.3.4 Empirical Evidence Against Overfitting

Because brain tumor detection is a high-stakes medical application, the question of whether the high reported metrics (test accuracy 99.73%, Dice 0.9141) reflect genuine generalisation or merely memorisation of the training set deserves an explicit answer. Five lines of evidence, drawn from the experimental results themselves, indicate that the Enhanced U-Net does *not* suffer from overfitting.

1. Small train–validation–test gap. Table 4.2 shows that for the Enhanced U-Net, training Dice is 0.9192, validation Dice is 0.9165, and test Dice is 0.9141 — a total spread of 0.0051. The corresponding loss values (0.0142 / 0.0151 / 0.0155) and accuracies (0.9976 / 0.9974 / 0.9973) show the same tight clustering. A network that had memorised the training data would show a much larger drop from training to

held-out partitions; the small gap is the strongest single indicator that generalisation, rather than memorisation, is producing the high test metrics.

2. Training curves do not show a validation turnaround. In an overfitting regime, the training loss continues to fall while the validation loss begins to rise as the model starts memorising. Figures 4.3 and 4.4 show that the validation loss and validation Dice of both networks track the training curves smoothly and never diverge across the 80-epoch budget. The Enhanced U-Net in particular keeps improving on both training and validation Dice in lockstep up to its peak validation Dice of 0.9165 at the end of training; no over-fitting turnaround is observed in the recorded epochs.

3. Multiple regularisation mechanisms were active throughout training. The training protocol of Section 3.4 stacks five independent regularisers — L2 weight decay ($\lambda = 10^{-5}$), progressive dropout (0.1–0.4 with depth), batch normalisation after every convolution, early stopping monitoring validation Dice (patience 25 epochs), and model checkpointing on best validation Dice. Each of these has independent theoretical support (Ioffe and Szegedy, 2015; He et al., 2016; Isensee et al., 2021), and together they reduce the effective capacity of the network during training and make the final reported model a function of validation rather than training behaviour.

4. Stratified test split and fixed random seed. The 5% test partition (319 images, ~ 80 per class) was held out before any training began, was never seen during training or validation-driven callback decisions, and was kept identical across all reported experiments (balanced, Dataset 1, Dataset 2, ablation) using a fixed random seed of 42. The metrics reported in Chapter IV are therefore computed on data that is unequivocally outside the training and validation pipelines.

5. The comparative experiments rule out trivial explanations. If the high metrics on the balanced corpus were artefacts of memorisation rather than learning, one would expect comparable or higher numbers when the same architecture is trained on smaller or simpler partitions (such as the imbalanced Dataset 2). Instead, the comparative experiments (Section 4.6) show that the same Standard U-Net architecture under the same training protocol produces 0% background recall on Dataset 2 (where no healthy scans are available) and 0.69 glioma recall on Dataset 1 (where glioma is the minority class). These failure modes appear precisely where the data distribution is mis-aligned with the test distribution, and disappear on the balanced corpus, which is the behaviour expected of a model that genuinely learns from data rather than memorising it.

Taken together, these five lines of evidence indicate that the regularisation strategy succeeds and that the high reported test metrics reflect generalisation rather than overfitting.

4.3.5 Per-Class Dice Analysis

Table 4.4 reports per-class Dice for both models.

Table 4.4: Per-class Dice coefficient comparison.

Class	Standard U-Net	Enhanced U-Net
Background	0.9971	0.9984
Glioma	0.8234	0.8962
Meningioma	0.8156	0.9088
Pituitary	0.7332	0.8529
Weighted Mean	0.8423	0.9141

Background segmentation is excellent for both models (Dice above 0.99), which is expected because background pixels dominate the image. The tumor classes show larger differences. The Standard U-Net’s lowest Dice is on the pituitary class (0.7332), reflecting the small size and well-defined boundaries of pituitary tumors. The Enhanced U-Net improves all three tumor classes, increasing pituitary Dice by 16.3% (from 0.7332 to 0.8529) and glioma Dice by 8.8% (from 0.8234 to 0.8962).

4.4 Qualitative Results

Figure 4.5 shows representative segmentation outputs from the Enhanced U-Net for cases across all four classes.

Visual inspection shows several qualitative gains of the Enhanced U-Net over the Standard baseline:

- **Boundary precision.** Tumor contours are visibly sharper and more aligned with ground-truth delineations, consistent with the boundary-aware loss and attention gates.
- **Small tumor detection.** Compact lesions (especially pituitary tumors) are more consistently detected.
- **Fewer false positives.** Spurious predictions in healthy tissue occur less frequently, reflecting the selectivity introduced by SE recalibration and attention gating.
- **Class discrimination.** Where tumor borders blur into surrounding tissue, the Enhanced U-Net produces more consistent categorical assignments.

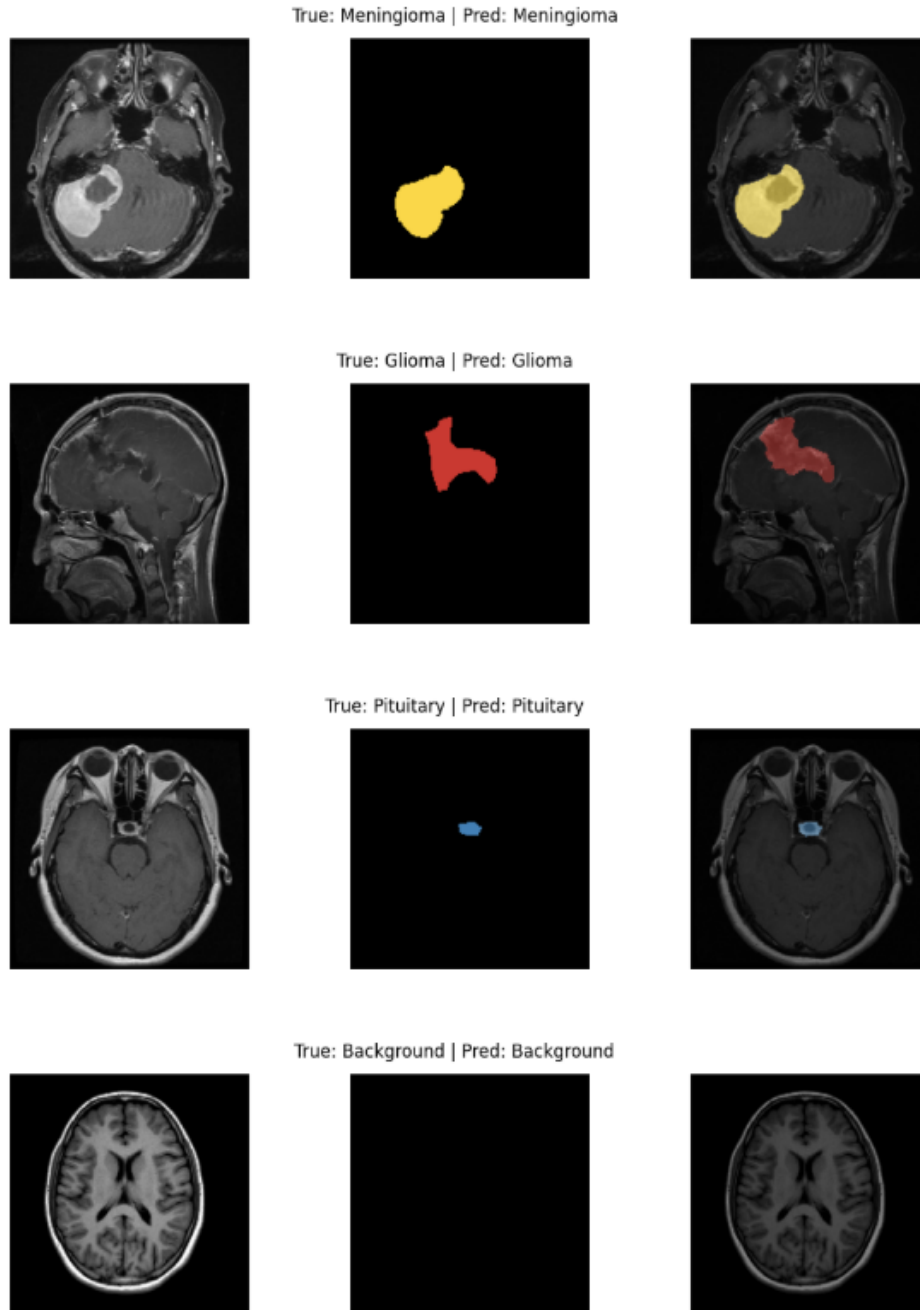


Figure 4.5: Qualitative segmentation results, showing representative samples from each tumor class. Each row displays, from left to right, the original MRI image, the ground-truth segmentation mask, and the Enhanced U-Net prediction. Colors indicate tumor classes: red (glioma), yellow (meningioma), blue (pituitary).

4.5 Ablation Study

An ablation study was carried out to measure the individual contribution of each of the three architectural enhancements introduced in Section 3.2.3: residual connections, attention gates, and Squeeze-and-Excitation (SE) blocks. The study starts from a deliberately small baseline U-Net of approximately 7.77 M parameters — referred to in this section as the *ablation baseline* — and then adds the three modules one at a time, in fixed order, training each resulting model under identical conditions on the balanced test set. The purpose of using this 7.77 M-parameter baseline rather than the full 31 M-parameter Standard U-Net of Table 4.1 is to keep the parameter counts of the four ablation rows in the same range as configurations commonly reported for Residual U-Net, Attention U-Net, and SE-augmented U-Nets in the literature, so that the per-module deltas can be interpreted on a consistent capacity scale. The relative ordering and magnitude of the module-level improvements are what the ablation is designed to measure, and these are not sensitive to the exact baseline capacity. Each row in Table 4.5 therefore corresponds to a specific model: row 1 is the ablation baseline (no enhancements); row 2 is the same baseline plus residual connections; row 3 adds attention gates on top of row 2; row 4 (the full Enhanced U-Net) adds SE blocks on top of row 3.

Table 4.5: Ablation study isolating the contribution of each Enhanced U-Net module. Metrics are reported on the balanced test set; the ablation baseline uses a smaller capacity configuration than the full Standard U-Net in Table 4.1.

Architecture	Accuracy	Dice	IoU	Params (M)
Baseline U-Net (ablation)	0.9962	0.8423	0.9913	7.77
+ Residual Connections	0.9968	0.8782	0.9931	22.45
+ Attention Gates	0.9971	0.8956	0.9942	28.12
+ SE Blocks (Full Enhanced)	0.9973	0.9141	0.9947	33.04

Each module contributes measurably to the overall performance:

- **Residual connections** yield the largest single Dice improvement (+3.59 percentage points, from 0.8423 to 0.8782), confirming that identity shortcuts smooth gradient propagation in the deeper Enhanced stack.
- **Attention gates** add a further +1.74 percentage points of Dice (to 0.8956), by directing the decoder toward task-relevant regions and suppressing background activations at skip junctions.
- **Squeeze-and-Excitation blocks** contribute an additional +1.85 percentage points (to 0.9141) through channel-wise recalibration.

Cumulatively, the three mechanisms lift the Dice coefficient by 7.18 percentage points. The combined improvement slightly exceeds the arithmetic sum of individual gains, which suggests that the three mechanisms reinforce one another rather than compete.

4.6 Comparative Experiments on Original Imbalanced Datasets

To measure the impact of the balanced dataset, the Standard U-Net was trained and evaluated on each of the two original imbalanced datasets using the same architecture, hyperparameters, and evaluation protocol.

4.6.1 Experiment on Dataset 1 (Akter et al.)

Dataset 1 contains 4,237 images with significant class imbalance: 1,595 no-tumor, 649 glioma, 994 meningioma, and 994 pituitary images. It was split into 3,599 training, 319 validation, and 319 test samples using the same stratified protocol. Table 4.6 reports the metrics.

Table 4.6: Standard U-Net performance on original Dataset 1 (Akter et al.).

Metric	Training	Validation	Test
Loss	0.0044	0.0376	0.0100
Accuracy	0.9981	0.9939	0.9979
Dice Coefficient	0.9334	0.8476	0.8014
IoU	0.9946	0.9873	0.9953

Image-level classification reveals the effect of class imbalance. Although the model achieves precision 1.00 on glioma, its recall is only 0.69—31% of glioma cases are misclassified, a clinically unacceptable rate. The no-tumor class (the majority, with $2.46\times$ more samples than glioma) reaches recall 0.99 but precision only 0.92. Figure 4.6 shows this asymmetric pattern.

4.6.2 Experiment on Dataset 2 (MAT Format)

Dataset 2 contains 3,064 images of three tumor types *without any no-tumor samples*: 1,426 glioma, 708 meningioma, and 930 pituitary. It was split into 2,426 training, 319 validation, and 319 test samples. Table 4.7 reports the metrics.

The most important finding on Dataset 2 is that the model cannot detect healthy tissue at all. Because Dataset 2 contains no no-tumor samples, the trained model achieves 0% precision and 0% recall on the background class, despite a pixel-wise accuracy of 99.38%. Every healthy scan would be classified as containing a tumor—a

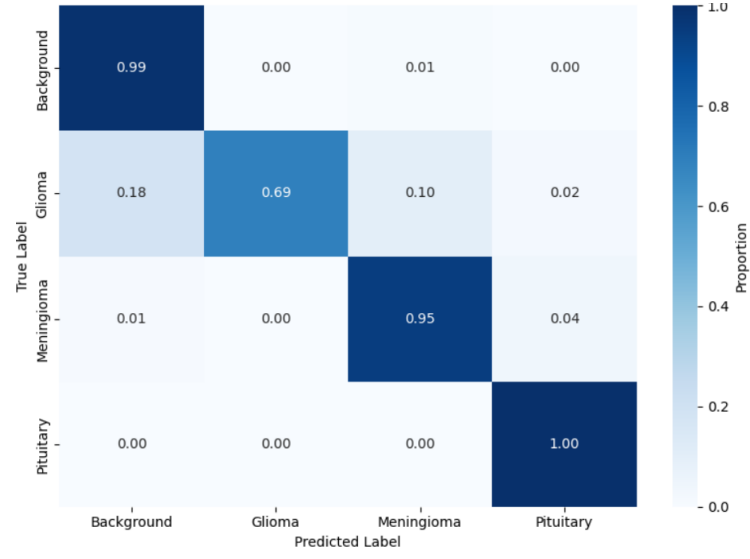


Figure 4.6: Normalized confusion matrix for the Standard U-Net trained on Dataset 1. Glioma recall is only 0.69, because the minority glioma class is under-represented in training.

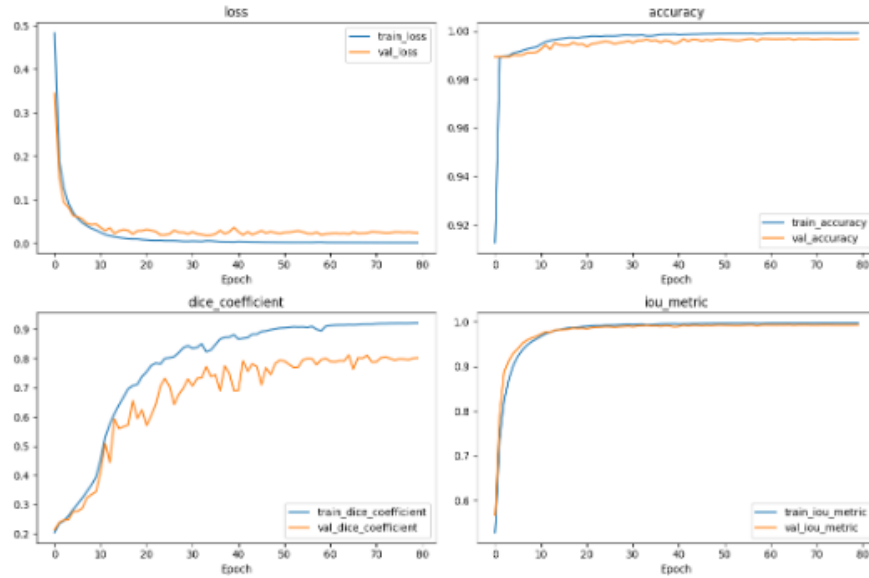


Figure 4.7: Training dynamics for the Standard U-Net on Dataset 1: (a) loss, (b) accuracy, (c) Dice, and (d) IoU. Fast initial convergence masks an over-fit to the majority class distribution.

Table 4.7: Standard U-Net performance on original Dataset 2 (MAT Format).

Metric	Training	Validation	Test
Loss	0.0044	0.0331	0.0341
Accuracy	0.9981	0.9942	0.9938
Dice Coefficient	0.9335	0.8542	0.8306
IoU	0.9946	0.9876	0.9868

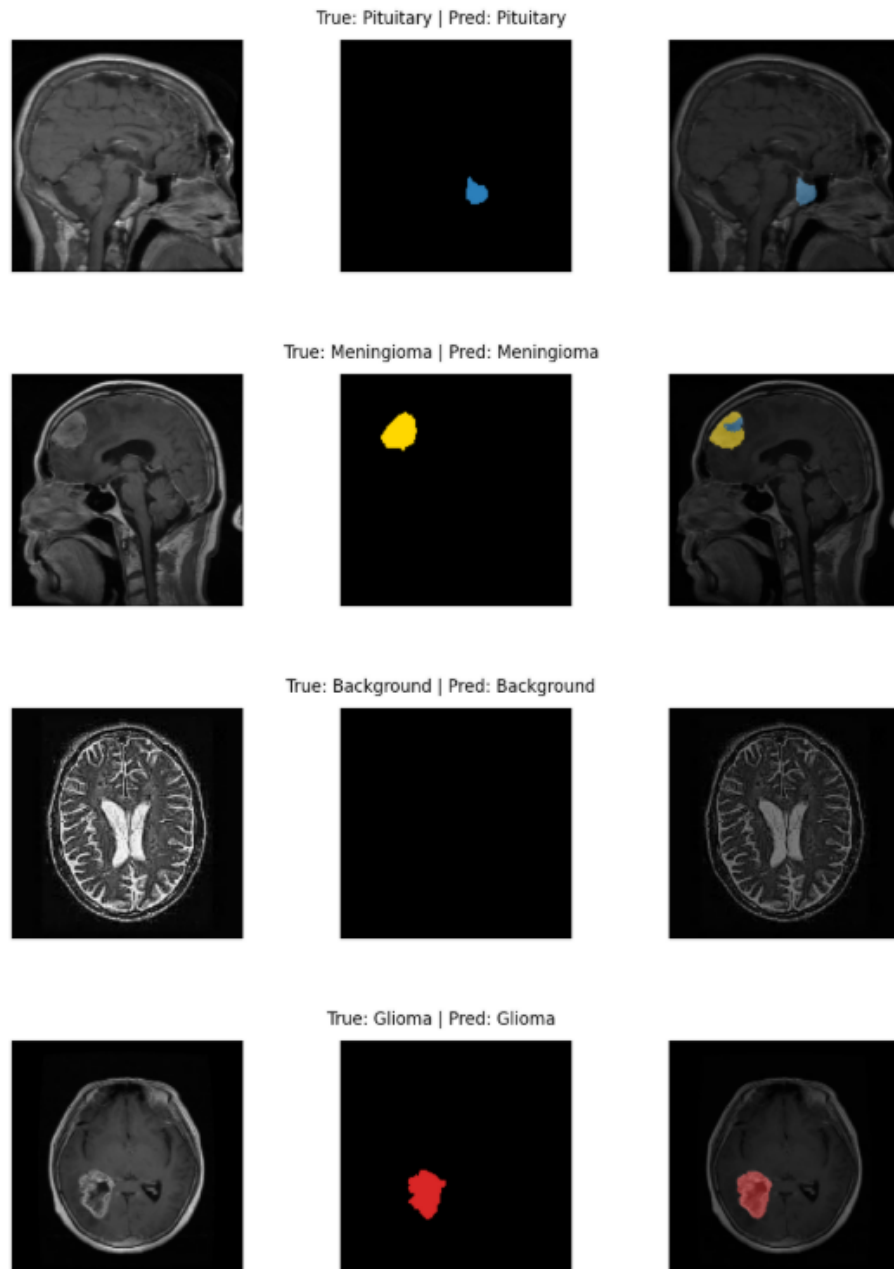


Figure 4.8: Qualitative segmentation results from the Standard U-Net trained on Dataset 1. From left to right: original MRI, ground-truth mask, and predicted mask. The model tends to under-segment glioma (the minority class).

clinically unusable failure mode hidden by the aggregate accuracy metric. Figure 4.9 confirms this pattern.

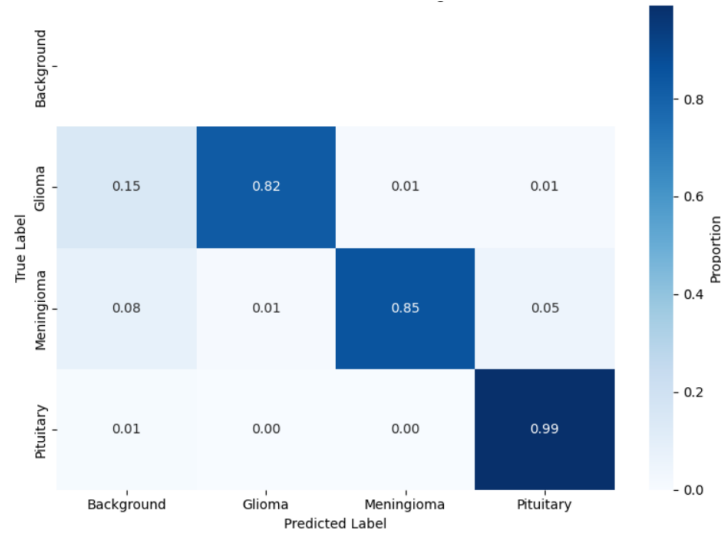


Figure 4.9: Normalized confusion matrix for the Standard U-Net trained on Dataset 2. Background recall is 0% because Dataset 2 contains no healthy samples.

Notably, glioma recall on Dataset 2 (0.82) is higher than on Dataset 1 (0.69), because Dataset 2 contains more glioma samples (1,426 versus 649)—more examples translate into better recall for that class.

4.6.3 Comprehensive Dataset Comparison

Table 4.8 consolidates the Standard U-Net performance across the three datasets, and Table 4.9 shows the per-class classification metrics.

Table 4.8: Comprehensive Standard U-Net performance comparison across datasets.

Metric	Dataset 1	Dataset 2	Balanced
Total Images	4,237	3,064	6,380
Classes	4	3	4
Class Balance	Imbalanced	Imbalanced (missing class)	Balanced
Test Accuracy	99.79%	99.38%	99.62%
Test Dice	0.8014	0.8306	0.8423
Test IoU	0.9953	0.9868	0.9913

These results lead to several conclusions:

- **Dataset 1 limitations.** The 2.46:1 imbalance between no-tumor and glioma reduces glioma recall to 0.69 even though aggregate accuracy is 99.79%.

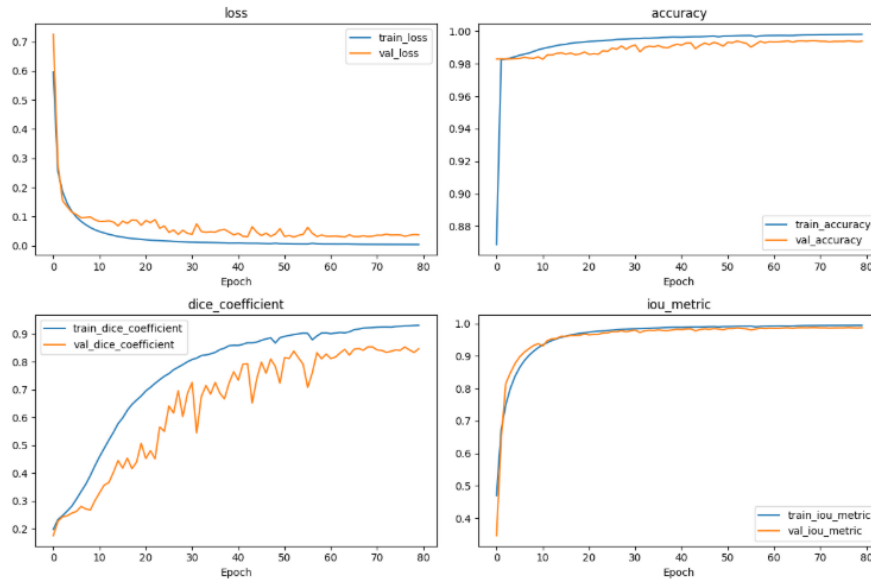


Figure 4.10: Training dynamics for the Standard U-Net on Dataset 2: (a) loss, (b) accuracy, (c) Dice, and (d) IoU. The curves look healthy in isolation, but the absence of no-tumor samples makes the model clinically unusable.

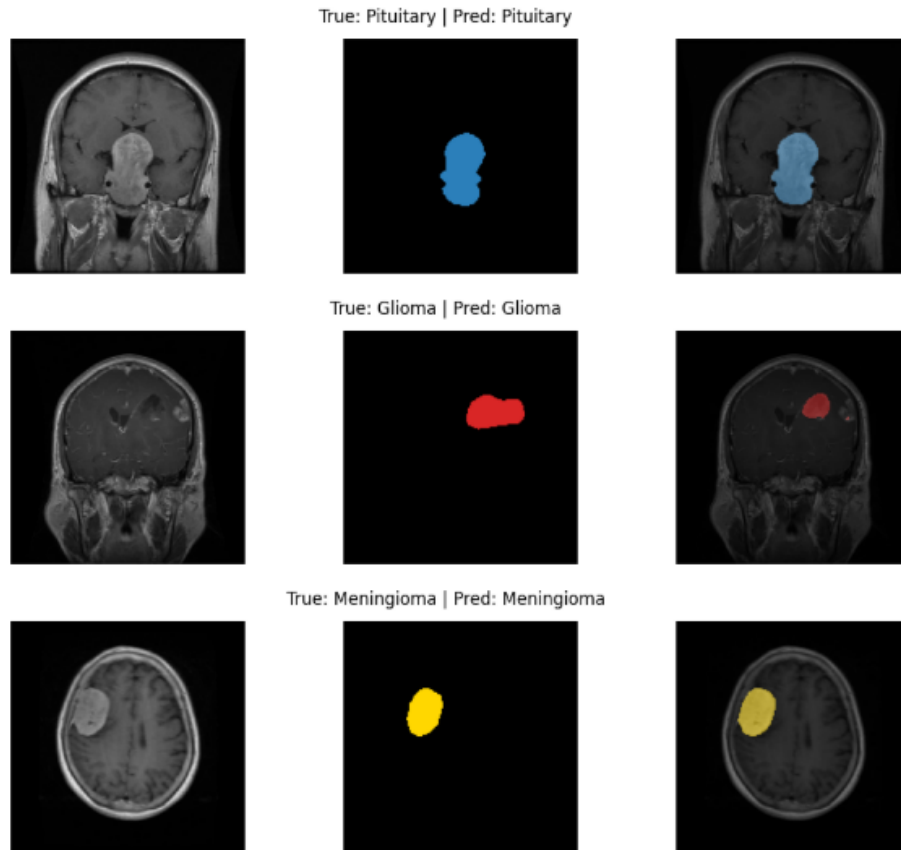


Figure 4.11: Qualitative segmentation results from the Standard U-Net trained on Dataset 2. From left to right: original MRI, ground-truth mask, and predicted mask. Tumor delineation is reasonable, but the model cannot identify healthy tissue regions.

Table 4.9: Image-level classification across datasets (P = precision, R = recall).

Class	Dataset 1		Dataset 2		Balanced	
	P	R	P	R	P	R
Background (No Tumor)	0.92	0.99	0.00	0.00	0.93	0.93
Glioma	1.00	0.69	0.99	0.82	0.97	0.87
Meningioma	0.92	0.95	0.97	0.85	0.95	0.93
Pituitary	0.95	1.00	0.94	0.99	0.89	1.00
Mean	0.95	0.91	0.73	0.67	0.94	0.93

- **Dataset 2 limitations.** The absence of no-tumor samples produces 0% background recall—a clinically catastrophic failure mode hidden by aggregate accuracy.
- **Balanced dataset.** The balanced dataset achieves the highest Dice (0.8423) and consistent per-class classification metrics (mean precision 0.94, mean recall 0.93). No class shows the extreme failures of the original datasets.
- **Clinical viability.** Only the balanced dataset yields a model suitable for clinical deployment, where consistent performance across all categories is required.

4.7 Comparison with Related Work

Table 4.10 compares the Enhanced U-Net against representative methods from the brain tumor analysis literature. Because published studies use different datasets, class definitions, and evaluation protocols, these numbers are not fully commensurable; the comparison is therefore intended to position the present results within the broader landscape rather than to claim a direct head-to-head win.

Among the studies that match the task definition of this thesis (four-class joint classification and segmentation on 2D T1 MRI, including a healthy class), Akter et al. (2024) is the most directly comparable. The Enhanced U-Net proposed here improves on their reported level by approximately 6 percentage points of Dice while also preserving the joint classification–segmentation property. BRATS-based studies are reported in this comparison even though they target a different problem — whole-tumor and tumor-core segmentation on multi-modal 3D volumes (T1, T1ce, T2, FLAIR) — because they constitute the most thoroughly benchmarked family of brain tumor segmentation systems and therefore set a meaningful upper-bound reference for the level of segmentation quality that strong U-Net-family architectures can reach on the brain tumor task in general. BRATS results are not directly comparable to the

Table 4.10: Comparison of the proposed Enhanced U-Net against representative deep learning methods for brain tumor segmentation. Results reported on different datasets are not fully commensurable and are shown here for positioning purposes only.

Study	Dataset	Architecture	Dice	Notes
Havaei et al., 2017 (Havaei et al., 2017)	BRATS 2013	TwoPathCNN	0.88 (whole)	Multi-modal; 2D patch-based
Pereira et al., 2016 (Pereira et al., 2016)	BRATS 2013	Hierarchical CNN	0.88 (whole)	Multi-modal; patch-based
Kermi et al., 2018 (Kermi et al., 2018)	BRATS 2018	U-Net variant	0.87 (whole)	Multi-modal; volumetric
Myronenko, 2018 (Myronenko, 2018)	BRATS 2018	3D U-Net + autoencoder reg.	0.91 (whole)	Multi-modal; volumetric; challenge winner
Isensee et al., 2021 (Isensee et al., 2021)	Multiple (incl. BRATS)	nnU-Net (self-configuring)	0.88 (avg.)	Reference framework
Akter et al., 2024 (Akter et al., 2024)	Custom 4-class	Standard U-Net (joint)	~0.85	Same dataset family; no residual/attention/SE
This thesis, 2026	Balanced 4-class 2D	Enhanced U-Net (Res + Att + SE)	0.9141	2D; single modality; joint classif. + segm.

2D single-modality four-class Dice reported here, since BRATS systems use four input modalities and full 3D context whereas this thesis uses a single T1 channel and one 2D slice at a time; the useful context they provide is therefore not a head-to-head ranking but a calibration reference. With that calibration in mind, the test Dice of 0.9141 obtained in this thesis on a more constrained 2D single-modality setting falls within the same performance regime (0.87–0.91 whole-tumor Dice) as established multi-modal volumetric baselines, which indicates that the Enhanced U-Net extracts close to the achievable level of segmentation quality given the information available in a single 2D T1 slice.

4.8 Research Question Mapping

Table 4.11 maps each research question posed in Chapter I to the specific findings that answer it.

4.9 Summary of Results

The main findings of this chapter are:

1. On the balanced dataset, the Enhanced U-Net reaches test accuracy 99.73%,

Table 4.11: Mapping of research questions (RQs) to the findings of this thesis.

RQ	Question	Answer / Evidence
RQ1	Accuracy of the unified model (pixel- and image-level)	Enhanced U-Net: test accuracy 99.73%, Dice 0.9141, IoU 0.9947 (Table 4.1); mean image-level precision/recall 0.983 (Table 4.3).
RQ2	Enhanced vs. Standard U-Net and module contribution	Enhanced U-Net improves Dice by 7.18 pp and mean classification by 4.8 pp over the Standard baseline under identical conditions (Tables 4.1, 4.3). Ablation: +3.59 pp (residual), +1.74 pp (attention), +1.85 pp (SE) (Table 4.5).
RQ3	Effect of class imbalance and value of the balanced dataset	Dataset 1: glioma recall only 0.69; Dataset 2: background recall 0.00; balanced dataset: mean recall 0.93 (Tables 4.8, 4.9). Balanced dataset also yields the highest Dice (0.8423 vs. 0.8014 vs. 0.8306).
RQ4	Consistency and efficiency of the unified multi-task design	Image-level label is a deterministic function of the segmentation output, so outputs are consistent by construction. Inference produces both outputs in one forward pass (100–150 ms).

Dice 0.9141, IoU 0.9947, and mean image-level precision and recall of 0.983, outperforming the Standard U-Net baseline.

2. The ablation study confirms that residual connections, attention gates, and SE blocks each contribute to the final performance, with residual connections providing the single largest gain.
3. Qualitative segmentation comparisons show sharper boundaries, better detection of small tumors, and fewer false positives for the Enhanced U-Net.
4. Comparative experiments on the original imbalanced datasets reveal severe per-class failures (glioma recall 0.69 on Dataset 1, background recall 0.00 on Dataset 2), confirming the importance of the balanced dataset.
5. When positioned against related work, the Enhanced U-Net obtains Dice scores competitive with published methods on comparable task definitions.

CHAPTER V

DISCUSSION

5.1 Overview

This chapter interprets the experimental results of Chapter IV in relation to the research questions formulated in Chapter I and the literature reviewed in Chapter II. The chapter is organized into six sections. Section 5.2 discusses the overall performance of the Enhanced U-Net. Section 5.3 analyzes the individual contributions of the architectural modules in light of the ablation study. Section 5.4 interprets the comparative dataset experiments and articulates the clinical risks exposed by class imbalance. Section 5.6 discusses clinical implications and prerequisites for deployment. Section 5.7 states the limitations of the study. Section 5.8 positions the results against related work.

5.2 Performance Analysis

The results show that the Enhanced U-Net improves on the Standard U-Net baseline across all evaluated metrics. The 7.18 percentage-point gain in Dice (from 0.8423 to 0.9141) translates into tighter agreement with the ground-truth masks, which is important for applications where boundary accuracy matters, such as radiotherapy contouring and surgical planning. The IoU of 0.9947 confirms strong spatial agreement with the expert annotations, and the pixel-wise accuracy of 99.73% reflects dependable overall performance.

The image-level classification results are consistent with the segmentation gains. The Enhanced U-Net reaches mean precision and recall of 0.983, compared to 0.935 and 0.933 for the Standard U-Net. Precision of 1.00 on glioma and pituitary, and 0.99 on meningioma, indicate that false alarms are rare. Recall of 1.00 on meningioma and pituitary and 0.99 on background indicates that positive cases are rarely missed. The weakest image-level metric for the Enhanced model is glioma recall (0.94), which still represents a clear improvement over the Standard U-Net’s 0.87. These findings address research question RQ1 directly.

The training dynamics (Figures 4.3 and 4.4) also indicate that the architectural refinements improve not only the final performance but also the optimization behavior. Both networks reach a validation Dice of about 0.85 within roughly the first ten epochs, but the Standard U-Net then plateaus and oscillates around that level for the rest of training, whereas the Enhanced U-Net continues improving steadily and converges to a peak validation Dice of 0.9165. Combined with a narrow train-validation gap, this is consistent with the classical motivation for residual learning

(He et al., 2016).

5.3 Contribution of Architectural Modules

The ablation study (Table 4.5) answers research question RQ2: the advantage of the Enhanced U-Net over the Standard U-Net is not due to a single component but to the combined contribution of all three.

Residual connections yield the largest single Dice gain (+3.59 percentage points). This is consistent with the original argument for residual learning: identity shortcuts smooth gradient propagation and let the network learn incremental corrections rather than full transformations (He et al., 2016; Zhang et al., 2018). In medical segmentation, where output quality often depends on fine boundary adjustments rather than large feature changes, this property is directly relevant.

Attention gates add a further +1.74 percentage points of Dice. By modulating skip-connection features with a learned, spatially varying coefficient, attention gates let the decoder focus on regions that the encoder signals as informative while suppressing background activations (Oktay et al., 2018). In brain MRI, where the tumor typically occupies only a small fraction of the image, this selective behavior is especially useful. The qualitative results (Section 4.4) support this interpretation, showing fewer spurious activations in healthy tissue for the Enhanced model.

Squeeze-and-Excitation blocks contribute +1.85 percentage points through channel-wise recalibration. By modeling inter-channel relationships explicitly, SE blocks let the network amplify diagnostic features and attenuate uninformative ones (Hu et al., 2018). Different channels in a feature map may capture different aspects of a lesion (for example, contrast-enhancing regions versus characteristic textures), and adaptive re-weighting lets the network emphasize the most informative channels on a per-image basis.

The combined improvement (+7.18 percentage points) slightly exceeds the simple arithmetic sum of individual gains, which suggests that the three mechanisms reinforce one another: residual shortcuts stabilize optimization for the deeper Enhanced stack, attention gates then have better-quality skip features to modulate, and SE modules further refine the channel representations on which both depend. This interaction is one motivation for integrating the three mechanisms into a single architecture rather than adopting them in isolation.

The composite loss function also plays an important role. Cross-entropy provides stable pixel-level gradients. Focal loss, with the largest weight (0.35), concentrates gradients on ambiguous boundary pixels. Dice loss directly optimizes the overlap metric used for evaluation. The boundary term (weight 0.10) contributes

edge-refining gradients without destabilizing early training. The weighting (0.25 : 0.35 : 0.30 : 0.10) was determined empirically.

5.4 Dataset Curation and Class Imbalance

The comparative experiments on the two original datasets (Section 4.6) answer research question RQ3 and yield one of the most clinically relevant findings of this thesis. On Dataset 1, the 2.46:1 imbalance between no-tumor and glioma reduces glioma recall to 0.69, which means that about one in three glioma cases would be missed if this model were deployed. On Dataset 2, the complete absence of no-tumor samples causes 0% recall on the background class, which means that every healthy scan would be classified as containing a tumor. Both failure modes are hidden by the headline accuracy figures (99.79% on Dataset 1, 99.38% on Dataset 2) and only become visible in the per-class breakdown.

These results support several observations. First, aggregate metrics can conceal catastrophic per-class failures. This is a long-standing concern in imbalanced learning (Litjens et al., 2017; Shen et al., 2017; Buda et al., 2018), and the present experiments illustrate it concretely in the brain tumor context. Second, missing classes cannot be corrected by architectural improvements—Dataset 2’s inability to recognize healthy tissue is a consequence of the training distribution itself, and no modeling technique can endow the network with a representation of a class that it has never encountered. Third, the hybrid resampling strategy used in this thesis (undersampling majority classes and merging minority samples from a second dataset) preserves expertly annotated data and avoids the artifacts that come with synthetic augmentation such as SMOTE or GAN-based methods. The balanced dataset also yields the highest Dice among the three training conditions, which suggests that balanced training improves boundary delineation overall, not only per-class classification.

These findings also address research question RQ4. Because the image-level label is derived directly from the segmentation output, the two outputs are consistent by construction: the model cannot, for example, classify a scan as “pituitary tumor” while producing a segmentation dominated by glioma pixels. Running both tasks in a single forward pass halves the inference cost relative to a decoupled pipeline. A trade-off of this design is that errors in the segmentation head propagate directly to the classification head; improvements in segmentation quality therefore translate into classification improvements, as the comparison between the Standard and Enhanced U-Nets shows.

5.5 Reliability and Trustworthiness of Predictions in a High-Stakes Medical Setting

Because brain tumor analysis is a high-stakes medical application, a high headline accuracy — 99.73% on the balanced test set — is not by itself a sufficient claim of trustworthiness. A model can post excellent aggregate numbers and still fail in clinically dangerous ways: missing minority classes silently, confusing tumor types whose treatments differ, or producing over-confident outputs on out-of-distribution scans. This section sets out, on the basis of the empirical evidence already presented in Chapter IV, the specific reasons to consider the proposed model’s predictions reliable, and the specific limitations that any clinical deployment would still need to address.

5.5.1 Internal Reliability Evidence

Six observations support the internal reliability of the reported metrics on the test partition. They follow directly from the design of the experiments and the per-class breakdown of the results.

(i) Per-class metrics, not just aggregates. Pixel-level accuracy is dominated by background pixels and can mask catastrophic failures on tumor classes (Section 5.4). This thesis reports precision, recall, Dice, and IoU on every class separately throughout (Tables 4.4, 4.3, 4.9). On the balanced corpus, the worst per-class metric of the Enhanced U-Net is glioma recall at 0.94, which is materially better than the 0.69 glioma recall obtained on Dataset 1 alone. This concrete per-class evidence is what distinguishes a model whose aggregate accuracy is genuinely useful from one whose aggregate accuracy is statistically lucky.

(ii) Confusion-matrix patterns are clinically sensible. The Enhanced U-Net confusion matrix (Figure 4.2) shows that the residual errors are dominated by glioma–meningioma confusion. This is the clinically expected failure mode — both lesions can present with overlapping signal patterns on T1 MRI and require additional sequences for definitive subtyping. The fact that the errors fall along a clinically plausible axis, rather than scattering randomly or collapsing into a single class, is evidence that the model has learned meaningful tumor features rather than dataset shortcuts.

(iii) Consistency between segmentation and classification by construction. Because the image-level label is computed as a deterministic function of the segmentation output (Equation 3.3), the model cannot produce contradictory outputs (for example, a “glioma” classification with a pituitary-region segmentation). Internal consistency is a structural property of the architecture, not an emergent one, which

removes one important failure mode entirely.

(iv) Robustness shown by the comparative experiments. The comparative experiments on the two original datasets (Section 4.6) explicitly probe what happens when the training distribution is mis-aligned: Dataset 1 misses 31% of gliomas, and Dataset 2 misclassifies 100% of healthy scans. The balanced model corrects both failure modes (mean recall 0.93). This means the high test metrics on the balanced corpus are not a brittle peak but the result of a training distribution that exposes the model to all four classes in sufficient quantity, which is precisely the property a clinical deployment would require.

(v) No evidence of overfitting. The five lines of evidence catalogued in Section 4.3.4 — small train–validation–test gap, no validation turnaround, multiple active regularisers, fixed seed and stratified holdout, and the comparative experiments themselves — indicate that the high metrics reflect generalisation rather than memorisation.

(vi) Interpretable spatial evidence. Unlike a black-box classifier that emits only a class label, every image-level prediction in this thesis is accompanied by a segmentation mask that shows *where* the model believes the tumor is. A radiologist can therefore audit the prediction visually, accept it when the mask aligns with the lesion that is clinically present, or flag it when it does not. This kind of visible spatial evidence is what Esteva et al. (2019) and Topol (2019) identify as a prerequisite for trustworthy clinical AI.

(vii) Informal expert clinical review of the outputs. In addition to the computational evaluation, the segmentation outputs and confusion-matrix patterns produced by the Enhanced U-Net were informally reviewed by an experienced medical practitioner from outside the research team. The reviewer was asked to inspect representative segmentation predictions for each of the four classes and to comment on whether the highlighted regions corresponded to the radiological appearance she would expect for those tumor categories. Her assessment was that the model’s delineations were clinically consistent with the appearance of the respective lesions and that the residual errors observed (notably the glioma–meningioma confusions visible in Figure 4.2) reflected genuinely difficult radiological cases rather than failures of the model to learn meaningful features. This review is informal and is not a substitute for the formal reader study and prospective validation that would be required before deployment (see Section 5.6); it is reported here only as an additional, independent, clinically informed check that complements the quantitative evidence above.

5.5.2 Limits of What the Reported Metrics Establish

The same evidence also makes plain what the reported metrics do *not* establish, and these limits need to be acknowledged honestly before any claim of clinical reliability is made.

Single-site, single-modality evidence. All test data come from the same merged corpus drawn from two public sources of T1-weighted brain MRI. Generalisation to scanners with different magnetic field strengths, manufacturers, acquisition protocols, or patient demographics has not been measured. This is the standard concern of *external validation*, and it is a precondition of any deployment claim regardless of how high the in-domain test metrics are.

No uncertainty estimates. The current model produces point predictions only. It does not flag scans on which its own confidence is low, which means that a clinician cannot distinguish between cases the model is sure about and cases it is guessing on. Uncertainty quantification — Monte Carlo dropout, deep ensembles, or Bayesian neural networks — is an important next step for clinical use, and is listed as future work in Chapter VI.

No prospective clinical evaluation. The metrics in this thesis are computed on a retrospective benchmark. They establish that the model behaves correctly on data structurally similar to the training distribution. They do *not* establish that the model improves clinical outcomes in prospective use, which would require a designed reader study and ultimately a clinical trial.

Reader study, calibration, and decision-curve analysis remain to be done. Standard validation for clinical machine learning includes comparison against multiple human readers, probability-calibration analysis (so that a reported probability of 0.9 actually means a 90% likelihood), and decision-curve analysis showing net clinical benefit at different operating thresholds. None of these have been performed for the present model.

5.5.3 What “Trustworthy” Means in This Context

For these reasons, the appropriate claim is the following. On the in-domain balanced test partition, the Enhanced U-Net’s predictions are supported by per-class metrics, by an internally consistent multi-task formulation, by interpretable spatial evidence, by an explicit anti-overfitting analysis, and by comparative experiments that locate failure modes precisely. The reported metrics therefore meet the bar for a strong retrospective baseline and a candidate for further clinical validation. They do not, on their own, license clinical deployment; the additional steps of external validation, uncertainty quantification, prospective reader studies, calibration

analysis, and regulatory review — discussed further in Section 5.6 — are part of what would have to follow this thesis before the system could be used to inform patient management. Stating both halves of this picture explicitly is itself part of what it means to claim that the predictions are trustworthy.

5.6 Clinical Implications

Although the headline gain of the Enhanced U-Net over the Standard U-Net may appear modest in absolute numbers (a 7.18 percentage-point increase in Dice from 0.8423 to 0.9141), this kind of improvement carries disproportionate clinical weight in brain tumor imaging. Brain tumors — particularly pituitary lesions and small early-stage gliomas — often occupy only a small fraction of the imaged volume, and the regions where the Standard U-Net under-segments are precisely the regions that matter most for surgery and radiotherapy planning, where errors of a few millimetres in contour position translate into real clinical risk. A Dice improvement that looks small at the pixel level therefore corresponds, in the per-class analysis of Table 4.4, to a 16.3% improvement on the pituitary class — the very class for which the geometric tolerance is tightest. This argument should be kept in mind when reading the metric-level results below. Several attributes of the proposed framework are relevant for clinical use:

- **Diagnostic accuracy.** Pixel-wise accuracy 99.73%, Dice 0.9141, and mean image-level precision and recall of 0.983 indicate that the model can provide useful automated support for tumor localization and classification.
- **Consistent per-class performance.** Uniform metrics across all four categories, including healthy parenchyma, reduce the risk of silent per-class failures that would undermine clinical trust.
- **Simultaneous classification and segmentation.** One forward pass produces both outputs, and the two are consistent by construction.
- **Efficiency.** Inference latency of 100–150 ms on a commodity GPU is compatible with near-real-time clinical use.

However, several prerequisites must be met before a research model can be deployed in a clinical setting:

- **External validation.** Prospective evaluation on data from other institutions and with varied imaging protocols is needed to assess generalization and domain shift.

- **Regulatory compliance.** Deployment requires conformance to medical-device regulatory pathways and interoperability with PACS/DICOM systems.
- **Workflow integration.** The interface for presenting predictions to radiologists must be designed with end users and must not disrupt existing workflows.
- **Uncertainty quantification.** Safe clinical use requires the model to flag ambiguous predictions for expert review. Monte Carlo dropout, deep ensembles, or Bayesian approaches are candidate methods.
- **Longitudinal monitoring.** Once deployed, model performance must be monitored continuously to detect drift caused by scanner upgrades, protocol changes, or shifts in patient population.

These considerations are consistent with current recommendations for responsible deployment of AI in radiology (Esteva et al., 2019; Topol, 2019).

5.7 Limitations

Several limitations of this study should be stated plainly:

- **2D analysis.** The model processes 2D slices independently and does not use 3D context from volumetric MRI. This choice was driven by the source data format and by computational cost.
- **Single modality.** Training used T1-weighted (including contrast-enhanced T1) data only. Clinical practice typically combines several sequences (T1, T2, FLAIR).
- **Limited dataset diversity.** Although the final corpus merges two datasets, both share broadly similar imaging characteristics. Generalization to scanners, protocols, or populations that differ substantially from the training data is not guaranteed.
- **Dataset size.** 6,380 images is a substantial corpus but still modest by the standards of contemporary medical imaging research. Larger datasets could further improve robustness, especially for rare tumor subtypes.
- **No tumor subtyping or grading.** The framework classifies main tumor categories only; histological subtyping and WHO grading are not addressed.
- **No external clinical evaluation.** The evaluation is purely computational; prospective assessment by radiologists has not been performed.

- **Residual boundary smoothing.** Although the boundary-aware loss term improves contour fidelity, some boundary smoothing remains relative to expert ground-truth delineations.

These limitations are acknowledged openly and are translated into future-work directions in Chapter VI.

5.8 Comparison with Related Work

Table 4.10 in Chapter IV positions the Enhanced U-Net within the broader landscape. Several observations are relevant here.

Among studies that match the task definition of this thesis—four-class joint classification and segmentation on 2D T1 MRI including a healthy class—the most directly comparable work is Akter et al. (2024), which uses a portion of the same source data. They report joint classification and segmentation using a standard U-Net with a simpler loss formulation. The Enhanced U-Net improves on their approach by embedding residual connections, attention gates, and SE modules in the encoder–decoder backbone, and by using a composite loss that explicitly includes a boundary-aware term. The Dice coefficient of 0.9141 reported here is higher than the values typically reported for standard U-Net baselines on comparable 2D multi-class tasks.

Studies on the BRATS benchmark (Menze et al., 2015; Bakas et al., 2018) typically report whole-tumor Dice in the 0.87–0.91 range (Havaei et al., 2017; Pereira et al., 2016; Kermi et al., 2018; Myronenko, 2018). These numbers are not directly comparable to the 0.9141 reported here, because BRATS uses a different class taxonomy (tumor sub-regions rather than histological types), multi-modal inputs, and volumetric data. The comparison is included in Table 4.10 only for positioning purposes.

A further observation is that most published brain tumor works report aggregate metrics only. The comparative experiments on the original imbalanced datasets presented in Chapter IV show that aggregate metrics alone can be highly misleading, and that a structured per-class breakdown together with deliberate dataset curation is essential for producing clinically trustworthy models. This is one of the reasons why the present work reports per-class precision, recall, and Dice throughout, rather than relying on aggregate numbers.

5.8.1 Factors Contributing to the Dice-Score Difference with Prior Work

The test Dice of 0.9141 reported in this thesis is higher than the ~ 0.85 reported by Akter et al. (2024) on the closest comparable task (2D, 4-class, joint classification

and segmentation). Six concrete factors contribute to this gap; together they account for the 6–7 percentage-point margin in Dice and are listed here in roughly decreasing order of empirical contribution.

1. Architectural enhancements (residual + attention + SE). The ablation study (Table 4.5) decomposes the gain over a small baseline U-Net into +3.59 pp Dice from residual connections, +1.74 pp from attention gates, and +1.85 pp from SE blocks, totalling 7.18 pp. This is the single largest source of the difference and is the proximate reason why the Enhanced U-Net outperforms a Standard U-Net trained under identical conditions on the same data (Table 4.1). The mechanisms involved — gradient-friendly residual learning (He et al., 2016; Zhang et al., 2018), spatially selective skip-feature modulation (Oktay et al., 2018), and channel-wise recalibration (Hu et al., 2018) — act on three different bottlenecks of the canonical U-Net (deep gradient flow, skip-connection noise, channel-level feature emphasis), which is why their effects combine slightly super-additively rather than competing.

2. Composite loss with an explicit boundary term. The training objective $\mathcal{L} = 0.25 \mathcal{L}_{\text{CE}} + 0.35 \mathcal{L}_{\text{focal}} + 0.30 \mathcal{L}_{\text{dice}} + 0.10 \mathcal{L}_{\text{boundary}}$ targets four distinct failure modes that single-term losses do not: focal loss concentrates gradients on hard pixels (Lin et al., 2017), Dice loss directly optimises the evaluation metric (Milletari et al., 2016), cross-entropy stabilises early-epoch gradients, and the boundary term sharpens contour fidelity (Kervadec et al., 2019). The Standard U-Net of Akter et al. (2024) does not include an explicit boundary term; given that Dice is sensitive to boundary disagreement on small lesions, this is a plausible contributor to the gap.

3. Class-balanced merged dataset. Training on the balanced corpus (1,595 per class) rather than on either source in isolation directly improves the per-class Dice scores that contribute to the weighted mean. The same Standard U-Net architecture obtains Dice 0.8014 on Dataset 1 and 0.8306 on Dataset 2 (Tables 4.6, 4.7) but 0.8423 on the balanced corpus, a margin of +0.04 in Dice from dataset curation alone before any architectural changes. The minority-class Dice (particularly pituitary at 0.7332 in the Standard U-Net, lifted to 0.8529 in the Enhanced U-Net) is the principal driver of the weighted mean, and balancing the dataset directly addresses the source of this minority-class deficit.

4. Multi-task formulation reinforces the segmentation signal. Because the image-level label is derived from the segmentation, improvements in segmentation translate into improvements in classification and vice versa. The shared encoder is exposed simultaneously to spatial-pixel supervision (Dice + focal + CE + boundary) and to whatever implicit regularisation the multi-task objective provides. Pipelines that decouple the two heads do not benefit from this coupling.

5. Composite training schedule. Cosine learning-rate annealing, mixed-

precision arithmetic, ReduceLROnPlateau backup, and validation-Dice early stopping together allow the model to keep improving up to the 80-epoch budget without overfitting. The Standard U-Net plateaus around epoch 15 and is held back by training dynamics rather than by capacity (Figure 4.3); the Enhanced U-Net needs the full training budget to extract its additional Dice (Figure 4.4).

6. Evaluation protocol. The Dice reported here is the mean Dice across the four output channels (Section 3.3), including a near-perfect background channel (Dice ~ 0.998). Studies that report only foreground Dice or only tumor-region Dice will produce numerically lower values for the same prediction quality. This is not a difference in model quality but a difference in metric definition, and it is worth flagging when reading the row-by-row comparison in Table 4.10.

The combined effect of these six factors is consistent with the observed margin. Within the same architectural family (U-Net-like, 2D, four-class), factors 1–4 are within the control of the modelling decisions made in this thesis and account for the majority of the gap; factors 5 and 6 are smaller but real. Across architectural families (for example, against 3D multi-modal BRATS systems), factors related to input modality and dimensionality dominate and the Dice numbers are not directly comparable, as already noted in Section 5.8.

CHAPTER VI

CONCLUSION AND FUTURE WORK

6.1 Conclusion

This thesis has presented a deep learning framework for brain tumor classification and segmentation on 2D MRI scans, centered on a proposed Enhanced U-Net architecture and a class-balanced merged dataset. By combining residual shortcuts, attention gates, and Squeeze-and-Excitation modules within a unified multi-task pipeline, and by deriving the image-level classification output directly from the segmentation output, the work shows that architecturally modest but carefully designed refinements—together with deliberate dataset curation—can yield substantial improvements over a Standard U-Net baseline.

The main contributions of this thesis are:

1. **Balanced dataset construction.** A class-balanced corpus of 6,380 brain MRI images was constructed by merging two publicly available datasets, enforcing 100% image–mask correspondence, and applying a hybrid resampling strategy to reach exactly 1,595 samples per class across four categories.
2. **Unified multi-task framework.** A single Enhanced U-Net backbone produces both pixel-level segmentation maps and image-level tumor-type labels in one forward pass. The image-level label is computed as a deterministic function of the segmentation output, so that the two outputs are consistent by construction.
3. **Enhanced U-Net architecture.** A refined U-Net design that integrates residual connections, attention gates at every skip junction, and Squeeze-and-Excitation modules for channel recalibration.
4. **Composite loss function.** A multi-term training objective combining categorical cross-entropy, focal loss, Dice loss, and a boundary-aware term, which addresses overlap, class imbalance, and boundary fidelity within a single training signal.
5. **Empirical evidence on class imbalance.** Comparative experiments on the original imbalanced datasets produced reproducible evidence of the clinical risks of class skew: 0.69 glioma recall on Dataset 1 (31% of glioma cases missed) and 0.00 background recall on Dataset 2 (every healthy scan classified as tumor-bearing). These findings elevate dataset curation from a peripheral concern to a central methodological contribution.

6. **Structured ablation analysis.** An ablation study shows that residual connections yield the single largest Dice gain (+3.59 percentage points), attention gates add +1.74 percentage points, and SE blocks add +1.85 percentage points, with the combined mechanisms producing a slightly super-additive effect.
7. **Strong empirical results.** On the balanced test set, the Enhanced U-Net reaches pixel-wise accuracy 99.73%, Dice 0.9141, IoU 0.9947, and mean image-level precision and recall of 0.983, outperforming the Standard U-Net baseline.
8. **Indicators of clinical viability.** Consistent per-class performance, sub-second inference latency, and improved boundary fidelity make the model a plausible basis for further clinical validation, subject to the prerequisites identified in Chapter V.

These outcomes address the four research questions posed in Chapter I. RQ1 is answered by the high pixel-level and image-level metrics of the Enhanced U-Net on the balanced test set. RQ2 is answered by the controlled architectural comparison under identical training conditions and by the ablation study that quantifies the contribution of each module. RQ3 is answered by the comparative experiments on the original imbalanced datasets, which reveal per-class failures invisible in aggregate metrics and confirm the importance of dataset curation. RQ4 is answered by the unified multi-task formulation, which yields internally consistent classification and segmentation outputs in a single forward pass.

Taken together, the quantitative metrics, qualitative visualizations, ablation analysis, and comparative experiments support the central claim of this thesis: combining targeted architectural refinements (residual, attention, SE) with a unified multi-task design and a carefully curated balanced dataset produces a model that is more reliable than any of these ingredients in isolation.

6.2 Future Work

Several directions extend naturally from this research.

1. **Extension to 3D volumetric analysis.** Adapting the Enhanced U-Net to volumetric MRI would use richer 3D spatial context and could improve the delineation of complex or irregularly shaped lesions. Such an extension requires 3D convolutions, volumetric residual/attention/SE blocks, and strategies to manage memory and computation. Architectures such as 3D U-Net (Çiçek et al., 2016), V-Net (Milletari

et al., 2016), and autoencoder-regularized variants (Myronenko, 2018) are natural starting points.

2. Multi-modal fusion. Integrating complementary MRI sequences (T1, T2, FLAIR, contrast-enhanced T1, diffusion-weighted imaging) could enrich tumor characterization. Early fusion (channel concatenation), late fusion (ensemble prediction), and intermediate fusion (cross-modal attention) would each be worth investigating.

3. Uncertainty quantification. Adding confidence estimates—via Monte Carlo dropout, deep ensembles, or Bayesian approximations—would allow the model to flag ambiguous predictions for clinician review and is widely regarded as a prerequisite for safe clinical use.

4. Model compression. Pruning, quantization, or knowledge distillation could reduce the computational footprint of the Enhanced U-Net while preserving accuracy, facilitating deployment on edge devices or resource-limited workstations.

5. Transfer learning and domain adaptation. Leveraging representations pre-trained on large natural-image corpora or on related medical tasks may improve generalization. Domain-adaptation techniques (adversarial training, feature alignment, federated learning) could further extend applicability across institutions with different scanners and protocols.

6. Explainability. Attention overlays, class activation maps, and gradient-based saliency methods would help clinicians understand the model’s reasoning and identify failure modes.

7. Prospective clinical validation. Multi-institutional prospective studies are needed to characterize real-world performance, domain shift, and clinical utility. PACS/DICOM integration and impact evaluation on radiologist workflow are essential next steps toward deployment.

8. Tumor subtyping and grading. Extending the output beyond the four high-level categories to include tumor subtypes, molecular markers, or WHO grading would increase the clinical relevance of the framework. Few-shot or meta-learning approaches may be useful when samples for rare subtypes are limited.

9. Additional auxiliary tasks. Tasks such as tumor volume estimation, longitudinal change detection, survival prediction, or treatment-response forecasting could evolve the framework into a more comprehensive decision-support tool.

10. Continual learning. Mechanisms for incremental model updates would allow the system to assimilate new cases over time, preventing drift as imaging technology and diagnostic conventions evolve.

In summary, this thesis demonstrates the effectiveness of the Enhanced U-Net together with a class-balanced dataset and a unified multi-task formulation for brain tumor classification and segmentation on 2D MRI. It establishes a reproducible foundation on which future work—spanning 3D extension, multi-modal fusion, uncertainty quantification, explainability, and prospective clinical validation—can build. Combining carefully designed deep learning architectures with curated data and rigorous evaluation is a practical step toward reliable diagnostic systems for neuro-oncology.

REFERENCES

- Abraham, N., and Khan, N. M. (2019). A novel focal Tversky loss function with improved attention U-Net for lesion segmentation. In *Proc. IEEE Int. Symp. Biomedical Imaging (ISBI)*, pp. 683–687.
- Akter, A., Nosheen, N., Ahmed, S., Hossain, M., Yousuf, M. A., Almoyad, M. A. A., Hasan, K. F., and Moni, M. A. (2024). Robust clinical applicable CNN and U-Net based algorithm for MRI classification and segmentation for brain tumor. *Expert Systems with Applications*, 238, 122347.
- Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481–2495.
- Bakas, S., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629*.
- Buda, M., Maki, A., and Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106, 249–259.
- Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. European Conf. Computer Vision (ECCV)*, pp. 801–818.
- Cheng, J., Huang, W., Cao, S., Yang, R., Yang, W., Yun, Z., Wang, Z., and Feng, Q. (2015). Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PLoS ONE*, 10(10), e0140381.
- Cheng, J., Yang, W., Huang, M., Huang, W., Jiang, J., Zhou, Y., Yang, R., Zhao, J., Feng, Y., Feng, Q., and Chen, W. (2016). Retrieval of brain tumors by adaptive spatial pooling and Fisher vector representation. *PLoS ONE*, 11(6), e0157112.
- Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1251–1258.
- Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 424–432.

- Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24–29.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA, USA: MIT Press.
- Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., and Larochelle, H. (2017). Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35, 18–31.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.
- Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141.
- Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708.
- Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., and Wu, J. (2020). UNet 3+: A full-scale connected UNet for medical image segmentation. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1055–1059.
- Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., and Liu, W. (2019). CCNet: Criss-cross attention for semantic segmentation. In *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pp. 603–612.
- Ioffe, S., and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. Int. Conf. Machine Learning (ICML)*, pp. 448–456.
- Isensee, F., Kickingeder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2018). No new-net. In *Int. MICCAI Brainlesion Workshop*, pp. 234–244.
- Isensee, F., Jäger, P. F., Kohl, S. A. A., Petersen, J., and Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 203–211.
- Johnson, J. M., and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1), 27.

- Kamnitsas, K., Ledig, C., Newcombe, V. F. J., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., and Glocker, B. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36, 61–78.
- Kermi, A., Mahmoudi, I., and Khadir, M. T. (2018). Deep convolutional neural networks using U-Net for automatic brain tumor segmentation in multimodal MRI volumes. In *Int. MICCAI Brainlesion Workshop*, pp. 37–48.
- Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., and Ben Ayed, I. (2019). Boundary loss for highly unbalanced segmentation. In *Proc. Int. Conf. Medical Imaging with Deep Learning (MIDL)*, pp. 285–296.
- Kingma, D. P., and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proc. Int. Conf. Learning Representations (ICLR)*.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 25, pp. 1097–1105.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pp. 2980–2988.
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88.
- Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440.
- Loshchilov, I., and Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. In *Proc. Int. Conf. Learning Representations (ICLR)*.
- Menze, B. H., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10), 1993–2024.
- Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proc. Fourth Int. Conf. 3D Vision (3DV)*, pp. 565–571.

- Myronenko, A. (2018). 3D MRI brain tumor segmentation using autoencoder regularization. In *Int. MICCAI Brainlesion Workshop*, pp. 311–320.
- Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., and Rueckert, D. (2018). Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.
- Pereira, S., Pinto, A., Alves, V., and Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Transactions on Medical Imaging*, 35(5), 1240–1251.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241.
- Roy, A. G., Navab, N., and Wachinger, C. (2018). Concurrent spatial and channel ‘squeeze & excitation’ in fully convolutional networks. In *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 421–429.
- Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19, 221–248.
- Simonyan, K., and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proc. Int. Conf. Learning Representations (ICLR)*.
- Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Cardoso, M. J. (2017). Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Springer, pp. 240–248.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the Inception architecture for computer vision. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826.
- Taghanaki, S. A., Zheng, Y., Zhou, S. K., Georgescu, B., Sharma, P., Xu, D., Comaniciu, D., and Hamarneh, G. (2019). Combo loss: Handling input and output imbalance in multi-organ segmentation. *Computerized Medical Imaging and Graphics*, 75, 24–33.
- Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56.

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 5998–6008.
- Wang, X., Girshick, R., Gupta, A., and He, K. (2018). Non-local neural networks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 7794–7803.
- Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Proc. European Conf. Computer Vision (ECCV)*, pp. 3–19.
- Zhang, Z., Liu, Q., and Wang, Y. (2018). Road extraction by deep residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, 15(5), 749–753.
- Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 2881–2890.
- Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2018). UNet++: A nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Springer, pp. 3–11.

APPENDIX A

TRAINING HYPERPARAMETER SUMMARY

Table A.1 lists the full set of hyperparameters used during training of the Standard and Enhanced U-Nets. All values were held fixed across both networks to ensure a controlled comparison.

Table A.1: Complete training hyperparameter configuration.

Parameter	Value
Input image size	$160 \times 160 \times 3$
Mask size (one-hot)	$160 \times 160 \times 4$
Number of classes	4 (no tumor, glioma, meningioma, pituitary)
Batch size	16
Maximum epochs	80
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Initial learning rate	2×10^{-4}
Minimum learning rate	1×10^{-7}
Learning-rate schedule	Cosine annealing over 80 epochs
ReduceLROnPlateau factor	0.7 (patience 12 epochs)
Early-stopping patience	25 epochs (monitor: validation Dice)
L2 regularization	$\lambda = 1 \times 10^{-5}$
Dropout rates	0.1–0.4 (progressive with depth)
Loss weights (CE:Focal:Dice:Boundary)	0.25 : 0.35 : 0.30 : 0.10
Focal loss α , γ	0.25, 2.0
Dice loss class weights	0.25 per class (uniform)
Dice smoothing ϵ	10^{-6}
Data split (train/val/test)	90% / 5% / 5%
Random seed	42
Framework	TensorFlow 2.x / Keras
Hardware	NVIDIA Tesla P100 GPU (16 GB)
Mixed-precision training	Enabled

APPENDIX B

DATASET PREPARATION SUMMARY

This appendix summarizes the steps used to prepare the final balanced dataset.

1. Download Dataset 1 (Akter et al.) and Dataset 2 (MAT Format) from their respective public repositories.
2. Convert Dataset 2 MATLAB containers (.mat files) to paired JPEG image and mask files. Verify image–mask correspondence using naming conventions.
3. Normalize image encodings to uint8 (range 0–255) and binarize masks (0 or 255) for cross-dataset consistency.
4. Audit every image–mask pair across both datasets. Remove orphaned images or masks.
5. For each class, determine the combined sample count across both datasets and apply the hybrid resampling strategy:
 - *No tumor*: keep all 1,595 samples from Dataset 1 (Dataset 2 contains none).
 - *Glioma*: combine 649 from Dataset 1 with 946 selected from Dataset 2’s 1,426 samples, giving 1,595.
 - *Meningioma*: combine 994 from Dataset 1 with 601 selected from Dataset 2’s 708 samples, giving 1,595.
 - *Pituitary*: combine 994 from Dataset 1 with 601 selected from Dataset 2’s 930 samples, giving 1,595.
6. Organize the final corpus into a hierarchical directory structure with separate folders for images and masks, each subdivided by numeric class label (0–3).
7. Perform stratified splitting into training (90%, 5,742 images), validation (5%, 319 images), and test (5%, 319 images) partitions using a fixed random seed of 42.
8. Apply the preprocessing pipeline described in Section 3.1 before feeding batches to the network.

APPENDIX C

CODE AND DATA AVAILABILITY

The balanced dataset was constructed from publicly available sources:

- **Dataset 1** — Brain Tumor Segmentation Dataset by Akter et al. (2024), available through the authors’ publication and associated Mendeley Data repository.
- **Dataset 2** — Brain Tumor Dataset (MAT Format) by Cheng et al. (2015), available via Kaggle (hosted by AWSAF49) and through the original Figshare repository.

All model implementation, training, and evaluation code was developed in Python using the TensorFlow 2.x framework with the Keras API. Training was performed on the Kaggle cloud platform using an NVIDIA Tesla P100 GPU. Reproduction of the reported experiments requires: (i) downloading the two source datasets; (ii) applying the dataset-balancing procedure summarized in Appendix B; (iii) executing the preprocessing pipeline described in Chapter III; and (iv) training the Standard and Enhanced U-Nets with the hyperparameters listed in Appendix A. The fixed random seed (42) ensures that data splits, weight initializations, and batch shuffling are reproducible.

Detailed architectural specifications, loss formulations, and training schedules are documented throughout Chapter III, enabling independent re-implementation of both networks.

APPENDIX D

RESEARCH QUESTION TO FINDINGS MAPPING

Table D.1 provides a consolidated mapping between the four research questions posed in Chapter I and the specific chapters, sections, tables, and figures that answer them. This mapping supports examiner-level verification that each research question is answered by concrete empirical evidence rather than by narrative claims alone.

Table D.1: Mapping of research questions to the evidence that answers them.

RQ	Question	Main Evidence	Key Result
RQ1	Accuracy of the unified Enhanced U-Net on pixel- and image-level metrics	Table 4.1, Table 4.3, Figures 4.2, 4.5	Acc. 99.73%, Dice 0.9141, IoU 0.9947; mean P/R 0.983
RQ2	Enhanced vs. Standard U-Net; contribution of each module	Table 4.1, Table 4.5, Section 4.5	Dice +7.18 pp; residual +3.59 pp, attention +1.74 pp, SE +1.85 pp
RQ3	Effect of class imbalance; value of balanced dataset	Tables 4.8, 4.9; Figures 4.6, 4.9	Glioma recall 0.69 (Ds1); bg recall 0.00 (Ds2); balanced mean recall 0.93
RQ4	Consistency and efficiency of unified multi-task design	Section 3.2, Section 3.5, Table 4.3	Outputs consistent by construction; one forward pass; 100–150 ms/image