Project Design & Analysis Report

University/Department	California State University, Dominguez Hills — Computer Science Department
Course/Thesis/Project Title	CSC 590/595 — Master’s Project/Thesis
Semester/Year	Fall 2025
Project Title	Brain Tumor Segmentation with U■Net & Fine■Tuned Foundation Models (Me
Student Name	[Your Name]
Student ID	[Your Student ID]
Instructor / Committee Chair	Dr. Sahar Hooshmand
Date	October 12, 2025

Note: This Report 0 (Design & Analysis) aligns with the syllabus checkpoints and Cedars■Sinai brief.

Abstract

Accurate brain tumor segmentation from multi■modal MRI is a critical step in neuro oncology, informing diagnosis, surgical planning, and therapy response assessment. This project investigates two complementary approaches: (1) a convolutional U■Net baseline trained end to end and (2) a fine■tuned foundation model (MedSAM) adapted to the same data. Using the Medical Segmentation Decathlon Task01_BrainTumour dataset (FLAIR, T1, T1■contrast■enhanced, T2), we target the standard subregions—whole tumor (WT), tumor core (TC), and enhancing tumor (ET)—and evaluate performance with the Dice similarity coefficient and 95th percentile Hausdorff distance (HD95), alongside runtime and GPU memory footprint.

Our objectives are to (i) establish a reproducible U■Net baseline, (ii) fine tune MedSAM for volumetric tumor delineation, and (iii) analyze accuracy efficiency trade offs and robustness under common clinical style perturbations (e.g., intensity shift, slice thickness variability). Expected contributions include a transparent pipeline (preprocessing, training, inference, post■processing), an apples to apples comparison between a specialized CNN and a promptable foundation model, and practical guidance on when each approach is preferable under constrained compute. This work directly follows the Cedars Sinai brief emphasizing a U■Net baseline and a fine■tuned foundation model on Task01_BrainTumour and is structured to scale into the 40–50 page final report required by the course.

Keywords: Brain MRI, Tumor Segmentation, U■Net, MedSAM, Dice, HD95, Foundation Models

Introduction

1.1 Background & Motivation

Primary and secondary brain tumors exhibit substantial heterogeneity in shape, location, and appearance across MRI modalities. Manual delineation of tumor subregions—edema/infiltration, necrotic/non enhancing core, and enhancing tumor—remains time consuming and operator dependent, impacting reproducibility and clinical throughput. Deep learning has delivered strong results in medical image segmentation through encoder–decoder architectures such as **U■Net**, while foundation models (e.g., SAM variants and **MedSAM**) promise rapid adaptation and promptable segmentation with fewer labeled samples. For a graduate capstone setting, contrasting a purpose built CNN with a fine■tuned foundation model offers both pedagogical value and clinically relevant insights. This project follows the Cedars Sinai topic specification to build a **U■Net** baseline and then fine tune a pre trained model (**MedSAM**) on the **Medical Segmentation Decathlon Task01_BrainTumour** dataset.

1.2 Problem Statement

Given multi■modal brain MRI volumes (FLAIR, T1, T1■contrast■enhanced, T2), automatically segment three clinically meaningful subregions: **Whole Tumor (WT)**, **Tumor Core (TC)**, and **Enhancing Tumor (ET)**. The model must generalize across patients and scans and output voxel wise labels suitable for volumetric analysis. We will quantify accuracy using **Dice** and **HD95** and report computational efficiency (inference time per volume, peak VRAM), enabling a rigorous accuracy vs. efficiency comparison between (i) a supervised **U■Net** and (ii) a fine■tuned **MedSAM**.

1.3 Objectives

1. Establish a **U■Net** baseline. Implement a strong, reproducible baseline with standardized preprocessing (spacing/intensity normalization), patch based training (2D/2.5D or memory aware 3D), and post■processing.

2. Fine tune **MedSAM**. Adapt **MedSAM** to brain MRI segmentation via prompt design and lightweight parameter■efficient fine tuning (e.g., adapters/**LoRA**), ensuring input/output compatibility with the baseline pipeline.

3. Evaluate rigorously. Use fixed data splits and report **Dice** and **HD95** per subregion, plus runtime and VRAM. Conduct paired statistics over cases to test significance of performance differences.

4. Analyze robustness. Stress test both models to modest domain shifts (e.g., intensity scaling, Gaussian noise) to examine generalization.

5. Report and release. Produce a clear write up and a minimal, reproducible codebase that can scale into the course’s 40–50 page final report with code appendix and GitHub link, as required.

1.4 Scope & Limitations

Data scope. We focus on **MSD Task01_BrainTumour**; all experiments use only these labeled volumes to avoid data leakage.

Compute constraints. Training will respect a single GPU budget typical for graduate projects; we will prefer 2D or 2.5D training and memory aware 3D patches where feasible.

Model scope. We compare one representative CNN (**U■Net**) with one representative foundation model (**MedSAM**). Broader architecture sweeps (e.g., Swin **U■Net**, **TransUNet**) are noted for future work if time permits.

Clinical claims. The work is research oriented; no diagnostic claims are made. External clinical deployment, regulatory validation, and prospective testing are out of scope.

Timeline alignment. This Design & Analysis submission corresponds to the Oct 12, 2025 milestone; subsequent progress reports and the in person final presentation will build on these foundations.

1.5 Contributions (This Stage)

• Problem framing & requirements. A precise segmentation objective with metrics (**Dice**, **HD95**) and compute constraints defined up front.

• Design choices. Justified selection of **U■Net** baseline and **MedSAM** fine tuning consistent with the Cedars Sinai brief on **Task01_BrainTumour**.

• Planned pipeline. End to end plan for preprocessing, augmentation, training, inference, post■processing, and evaluation—structured to scale into the 40–50 page final report with a code appendix and GitHub link per course policy.

Literature Review

2.1 Public Datasets & Pre processing Norms

Datasets. We will use the **Medical Segmentation Decathlon** (**MSD**) **Task01_BrainTumour**, which reuses the **BraTS** multi site glioma MRI collection and labels (multi■modal T1, contrast enhanced T1 (T1ce/T1Gd), T2, and FLAIR). **MSD** standardizes task definitions and evaluation across organs and includes a permissive license for research use. Medical Decathlon

The **BraTS** challenges distribute MRI volumes pre■processed by the organizers: co■registered to a common anatomical atlas, resampled to 1 mm^3 isotropic resolution, and skull stripped (typical volume shape ≈240×240×155). Labels target three sub regions which yield the common aggregates **Whole Tumor (WT)**, **Tumor Core (TC)**, and **Enhancing Tumor (ET)**. Perelman School of MedicinePerelman School of

Pre processing norms in the literature. Common steps (when not already provided) include N4 bias field correction to address intensity inhomogeneity, brain extraction (e.g., FSL BET or learned SynthStrip), and rigid/affine registration + resampling to harmonize spacing and orientation (**SimpleITK**/**ITK** pipelines). Intensities are then normalized per volume (e.g., z score inside brain mask). **SimpleITK**+4IACL+4Massachusetts Institute of

All files use NIfTI; Python I/O through **NiBabel** is standard. nipy.org

Implication for this project. Because **MSD**/**BraTS** already supply co registration, isotropic spacing, and skull stripping, our pipeline will mainly apply orientation harmonization and intensity normalization, with optional N4 for ablations. Perelman School of Medicine

2.2 CNN Architectures for Medical Segmentation

U■Net (2D). The seminal **U■Net** couples a contracting path (context) with an expanding path (localization) via skip connections—effective with limited annotations by heavy augmentation. Its locality and efficient inference made it the de facto baseline across modalities.

3D **U■Net**. Extends **U■Net** to volumetric convolutions for better through plane context, trading higher GPU memory and longer training for improved cross slice consistency.

U■Net++ / **Attention U■Net**. **U■Net**++ narrows the semantic gap between encoder/decoder via nested dense skip pathways and deep supervision; **Attention U■Net** inserts attention gates to highlight salient structures, often improving **sensitivity** for small lesions. arXiv+2SpringerLink+2

Strengths/weaknesses summary.

Strengths: strong localization, data efficiency, simplicity, broad tooling support.

Weaknesses: limited global context modeling (pure CNN locality), potential fragmentation in 2D (no z context), large memory for 3D on full volumes.

2.3 Transformer & Hybrid Models

To capture long range dependencies, hybrids combine CNN decoders with Transformer encoders (e.g., **TransUNet**), or adopt pure Transformer U shapes (e.g., Swin **U■Net** using shifted window self attention). These improve context aggregation but can demand more data/compute and careful optimization on 3D medical volumes. SpringerLink+3Department of Computer Science+3arXiv+3

2.4 Foundation/Promptable Models

Segment Anything (SAM) (SAM) (SAM) introduced promptable segmentation (points/boxes/masks) trained on SA 1B (1B masks, 11M images) with strong zero■shot transfer on natural images. However, direct zero■shot SAM on medical images underperforms due to domain shift. **MedSAM** addresses this by (1) curating >1.5M medical image–mask pairs across 10 modalities, and (2) fine tuning SAM to the medical domain, showing broad gains across 2D tasks; recent extensions explore 3D/temporal cases (e.g., **MedSAM** 2/SAM 2). arXiv+4arXiv+4CVF Open

Fine tuning strategies. Full fine tuning is costly; parameter■efficient fine tuning (**PEFT**) such as **LoRA** inserts low rank adapters into attention/MLP layers, reducing trainable parameters while retaining performance—attractive for compute limited student projects and increasingly used with vision/medical foundation models.

2.5 Evaluation Standards

Primary metric: **Dice** Similarity Coefficient (overlap, 0–1). Secondary: **HD95** (95th percentile Hausdorff distance; boundary accuracy robust to outliers), plus **sensitivity**/**specificity** to detect over/under segmentation. **BraTS**/decathlon challenges canonicalized **Dice** + **HD95** as official metrics. Perelman School of

Caveat: **HD95** implementation details vary across toolkits; report voxel spacing and code to ensure comparability. papers.miccai.org

Validation protocols: 5 fold cross validation enhances reliability on small datasets; hold out tests are simpler but risk higher variance—**nnU■Net** popularized robust cross validated reporting with decathlon tasks. Nature

2.6 Gap Analysis

1. Label scarcity & cost → motivates transfer and **PEFT** (e.g., **LoRA**) for sample efficiency. OpenReview

2. Domain shift (scanner, protocol, site) continues to degrade generalization; active strands include domain adaptation/generalization and test time adaptation.

3. Inference cost for 3D/high capacity models → favors 2D/2.5D **U■Net** baselines and adapter based **MedSAM** fine tuning for practical deployment.

Why our plan fits the gap. A from scratch **U■Net** (data■efficient baseline) vs. **MedSAM**+**PEFT** (sample efficient transfer) lets us quantify accuracy–efficiency trade offs and generalization under controlled pre/post■processing on a standard dataset—matching the coach brief.

System Design

3.1 Overall Solution Overview

Pipeline: Data ingest → QC → Pre processing → Model training (baseline vs. **MedSAM** fine tune) → Validation/Test → Post processing → Reporting.

• Data ingest/QC: Verify modality presence/order (T1, T1ce, T2, FLAIR), header consistency, and spacing.

• Pre processing: Use **MONAI**/**ITK SimpleITK** for orientation (RAS/LPS), spacing confirmation, and z score normalization in brain mask; optional N4 ablation.

• Training: **U■Net** (2D/2.5D or patch based 3D) vs. **MedSAM** + **LoRA** adapters.

• **Evaluation**: **Dice** (per class + macro), **HD95**, **sensitivity**; runtime (ms/volume), VRAM footprint. Perelman School of Medicine

• Reporting: Tables/plots of metrics and qualitative overlays.

3.2 Architecture Diagrams (to include as figures)

• Fig. 1 — **U■Net** baseline. 4 level encoder–decoder with skip connections; 4 input channels (modalities); output logits for WT/TC/ET. arXiv

• Fig. 2 — **MedSAM** fine tuning. SAM image encoder + prompt encoder; **LoRA** adapters on attention blocks; point/box prompts formed from ground truth during training; inference with automatic box prompts from coarse proposals (optional). arXiv+2arXiv+2

3.3 Modules/Components

• Data module. NIfTI I/O (**NiBabel**), dataset registry, transforms (**MONAI** Spacingd, Orientationd, NormalizeIntensityd), class■balanced patch sampling for 3D.

• Model zoo. **U■Net** (2D/3D), **MedSAM** (checkpoint loader), **LoRA** injection.

• Trainer. Mixed precision, gradient clipping, early stopping, checkpointing; TensorBoard/W&B logging.

• Evaluator. Per case metrics (**Dice**/**HD95**); ensure consistent **HD95** implementation. papers.miccai.org

• Visualizer. Slice and 3D orthoview overlays; error maps (FP/FN).

3.4 Algorithms/Techniques Considered

• Losses. **Dice**CE (**Dice** + Cross Entropy), Focal for class imbalance, Tversky for recall favoring, and Boundary loss to sharpen contours—evaluated per task. Proceedings of Machine Learning Research+3arXiv+3arXiv+3

• Augmentations. Spatial (flip/rotate/elastic), intensity (gamma, bias field), modality dropout; implemented via **MONAI** transforms. **MONAI**

• Post processing. Connected component filtering, small island removal, and optional DenseCRF refinement to reduce spurious edges. arXiv

Data Analysis & Requirements

4.1 Data Sources

Primary: **MSD Task01_BrainTumour**/**BraTS** derived volumes with expert labels for tumor sub regions; standardized NIfTI imaging and evaluation protocol (**Dice**/**HD95**). Medical

4.2 Collection/Cleaning & Pre processing

• Integrity checks: modality files per case, affine/spacing, non empty masks.

• Normalization: z score inside brain mask; keep modality channels aligned.

• Resampling/orientation: confirm isotropic 1 mm³ (**BraTS** pre■processed) and consistent orientation (RAS/LPS) for our pipeline. Perelman School of Medicine

• Optionals (for ablations): N4 bias correction; brain extraction if testing external data; registration with **SimpleITK** if needed. IACL+2Massachusetts Institute of

4.3 Tools/Libraries/Frameworks

PyTorch + **MONAI** for medical transforms/loaders; **NiBabel** for NIfTI; **SimpleITK/ITK** for registration/resampling; experiment tracking via TensorBoard/W&B.

4.4 Hardware/Software Requirements

• Compute: For 2D/2.5D **U■Net**, ≥12 GB VRAM suffices for batch sizes ≥8 (mixed precision). For 3D patch based (e.g., 128^3), ≥16–24 GB is preferable.

• Storage: ~10–20 GB for dataset + ~5–10 GB for checkpoints and logs (varies with runs).

• OS/Env: Linux, CUDA enabled GPU, Python 3.10+, **PyTorch**/**MONAI** latest stable.

4.5 Ethics, Privacy, Governance

The dataset is de identified and curated for research; decathlon tasks are distributed under a permissive CC BY SA license. No PHI is collected or processed. Medical Decathlon

Methodology

5.1 Workflow

1. Freeze splits (train/val/test) to ensure fair comparison.

2. Baseline first: train **U■Net** to convergence under the finalized pipeline.

3. Foundation model: fine tune **MedSAM** with **LoRA** adapters using identical pre/post■processing and splits; evaluate with/without prompts.

4. Hold out test + (optional) 5 fold CV to bound variance. Nature

5.2 Models Considered

• **U■Net**: start with 2D (2.5D) using all four modalities as channels (optionally stack k neighboring slices = 2.5D for context); then, if time permits, compare a 3D patch based variant for TC/ET improvements.

• **MedSAM**: initialize from public **MedSAM** weights; **PEFT** (**LoRA**) on attention blocks; prompts: box/points derived from ground truth during training, and at inference from a learned proposal (or a single center point) to study prompt **sensitivity**.

5.3 Training Setup

• Optimizer: AdamW; LR schedule: cosine decay with warmup; epochs: 200 (early stop on val **Dice**); batch size: tuned to VRAM; AMP mixed precision; augmentation: elastic, rotate, gamma, bias field. **MONAI**

• Regularization: weight decay; stochastic depth (if using Transformer blocks).

• Checkpointing: top k by mean **Dice** (WT/TC/ET).

5.4 Evaluation Criteria

• Primary: **Dice** per class and macro averaged.

• Secondary: **HD95**, **sensitivity**/**specificity**, inference latency per volume, VRAM peak, and #params/trainable params. Report confidence intervals via bootstrapping across cases; publish the metric implementation used (**HD95**). Perelman School of

5.5 Statistical Testing & Ablations

• Significance tests: paired Wilcoxon (non parametric) or paired t test on per case **Dice**.

• Ablations: (i) augmentation on/off; (ii) 2D vs. 2.5D vs. 3D; (iii) **MedSAM** prompt type/number; (iv) **LoRA** rank and which layers are adapted.

• Robustness: simulate domain shift via style/intensity transforms; test on held out sites if available. arXiv

Project Plan

6.1 Milestones & Timeline

Your course schedule specifies graded checkpoints and the Final Report (**40–50 pages**) + in person presentation. Map our milestones to the syllabus dates:

• By Oct 12: Report 0 – Design & Analysis (this document).

• By Oct 26: Progress Report 1 — **U■Net** baseline trained; initial **Dice**/**HD95** and sample overlays.

• By Nov 09: Progress Report 2 — **MedSAM**+**LoRA** implemented; early comparison to baseline.

• By Nov 23: Progress Report 3 — Full evaluation, ablations, robustness study; draft figures/tables.

• By Dec 07: Final Report & Presentation — complete write up (40–50 pp, 10 pt), code appendix + GitHub link.

Optional enrichment. The department notes an extra credit Cedars Sinai competition opportunity (keep focus on this project first).

6.2 Risk Management

• GPU memory/time limits: prefer 2.5D or 3D patches, mixed precision, gradient accumulation.

• Training instability (Transformer/**PEFT**): learning rate sweeps; scale **LoRA** rank; freeze more layers. arXiv

• Dataset quirks: enforce rigorous QC; lock splits; document metric code to avoid **HD95** inconsistencies. papers.miccai.org

Expected Outcomes

Technical. A reproducible pipeline with quantified performance deltas between a from scratch **U■Net** and **MedSAM**+**LoRA**, including accuracy (**Dice**/**HD95**), efficiency (latency/VRAM), and robustness under basic domain shifts.

Practical. Clear guidance on when a conventional **U■Net** suffices vs. when foundation model fine tuning pays off for brain MRI segmentation (compute/data budgets).

Challenges. Sensitivity to scanner differences, small enhancing lesions (ET), and prompt design for foundation models (quality/number of prompts). arXiv

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "**U■Net**: Convolutional Networks for Biomedical Image Segmentation," MICCAI, 2015. SpringerLink

[2] Ö. Çiçek et al., "3D **U■Net**: Learning Dense Volumetric Segmentation from Sparse Annotation," MICCAI, 2016. SpringerLink

[3] Z. Zhou et al., "**U■Net**++: A Nested **U■Net** Architecture for Medical Image Segmentation," MICCAI, 2018. SpringerLink

[4] O. Oktay et al., "**Attention U■Net**: Learning Where to Look for the Pancreas," arXiv:1804.03999, 2018. arXiv

[5] J. Chen et al., "**TransUNet**: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv:2102.04306, 2021. Department of Computer Science

[6] H. Cao et al., "Swin **U■Net**: Unet like Pure Transformer for Medical Image Segmentation," arXiv:2105.05537, 2021 / ECCV W 2022. GitHub

[7] A. Kirillov et al., "**Segment Anything (SAM)** (SAM)," ICCV 2023. CVF Open Access

[8] J. Ma et al., "**Segment Anything (SAM)** (SAM) in Medical Images (**MedSAM**)," arXiv:2304.12306, 2023. arXiv

[9] E. Tustison et al., "N4**ITK**: Improved N3 Bias Correction," ISBI, 2010. IACL

[10] S. M. Smith, "Fast Robust Automated Brain Extraction (BET)," Hum. Brain Mapp., 2002. Massachusetts Institute of Technology

[11] **SimpleITK** documentation: Resample/Registration.

[12] **NiBabel** documentation: Working with NIfTI images. nipy.org

[13] S. Bakas et al., "Advancing the TCGA Glioma MRI Collections with Expert Segmentation Labels," Scientific Data, 2017. Nature

[14] **BraTS** evaluation: **Dice** and **HD95**. Perelman School of Medicine

[15] F. Milletari et al., “V Net” (**Dice** Loss). arXiv:1606.04797, 2016. arXiv

[16] S. S. M. Salehi et al., “Tversky Loss,” MICCAI MLMI, 2017. rd.springer.com

[17] H. Kervadec et al., “Boundary Loss,” MIDL/MedIA, 2019–2020. Proceedings of Machine Learning Research

[18] P. Krähenbühl and V. Koltun, “Efficient Inference in Fully Connected CRFs,” NeurIPS, 2011/2012. NeurIPS Proceedings

[19] F. Isensee et al., “**nnU■Net**: A Self Configuring Method,” Nature Methods, 2021. Nature

[20] B. A. Reinke et al., “Towards a guideline for evaluation metrics in medical image segmentation,” BMC Res Notes, 2022; and HD dilemma (implementation variability). BioMed

Appendix- If any: In later stages you will have more to put here

Additional diagrams, preliminary results, supporting material.