VT1: Automated Segmentation of Parahippocampal Gyrus Subregions Using Deep Learning for Alzheimer's Disease Research

Martin Oswald

Eigenständigkeitserklärung

I, Martin Oswald, declare that this thesis titled, *VT1: Automated Segmentation of Parahippocampal Gyrus Subregions Using Deep Learning for Alzheimer's Disease Research* and the work presented in it is my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

Abstract

This thesis presents an automated approach for segmenting parahippocampal gyrus (PHG) subregions in MRI scans, focusing on early Alzheimer's disease biomarkers. Using 111 MRI scans from the AMBI dataset and 42 from BAMBI, we developed and evaluated multiple deep learning approaches for segmenting four critical regions: the entorhinal cortex (ERC), medial perirhinal cortex (mPRC), lateral perirhinal cortex (lPRC), and parahippocampal cortex (PHC).

Starting with a baseline 3D U-Net architecture, we investigated various loss function configurations and introduced a novel SegReg architecture combining segmentation with regression-based region localization. Our most successful approach, combining Dice and cross-entropy loss functions, achieved mean Dice Similarity Coefficient scores of 0.754 on the training set and 0.654 on the validation set while effectively addressing the challenge of extreme class imbalance.

The study demonstrates the feasibility of automated PHG subregion segmentation while highlighting the complexities of handling anatomically intricate structures. These findings provide valuable insights for developing clinical tools supporting early Alzheimer's disease detection and research.

Acknowledgements

I want to thank Dr. Manuel Dömer and Dr. Ahmed Abdulkadir for their valuable guidance and support throughout the project. Their insightful feedback and technical expertise were invaluable in shaping this research's direction and quality.

Table of contents

1. Project Charta

The Project Charta outlines the foundation and objectives of this study, starting with the problem definition, situation assessment, and project goals. It provides a structured overview to contextualize the challenges this research addresses, the methodologies employed, and the anticipated outcomes.

1.1. Problem Definition

Dementia is a general term covering several diseases impacting cognitive abilities like memory and thinking. Alzheimer's disease (AD) is the most common form of dementia, causing 60-70% of dementia cases [1].

People often use AD and dementia interchangeably, but the terms refer to different aspects of the same condition. The Alzheimer's Association Workgroup defines AD as a long-term brain disorder characterized by changes in brain biology and structure, such as the presence of Amyloid plaques (extracellular deposits of amyloid- (A) peptides) and Neurofibrillary tangles (intracellular aggregates of tau protein). On the other hand, dementia refers to the stage where cognitive decline becomes evident. This distinction is important because individuals can have AD without displaying symptoms of dementia [2]. There is a wide range of biomarkers to detect AD, including blood tests, plasma tests, cerebrospinal fluid tests, and imaging. The literature review describes dementia symptoms and biomarkers.

Krumm et al. investigate AD progression by measuring the thickness of different brain regions. The early stages of AD involve shrinking of the parahippocampal gyrus (PHG), which includes the entorhinal cortex (ERC), medial perirhinal cortex (mPRC), lateral perirhinal cortex (lPRC), and parahippocampal cortex (PHC), as illustrated in Figure 2.1. The researchers used Magnetic resonance imaging (MRI) to scan the participant's brain and segment the different regions of interest (ROIs) in the PHG. The segmentation was performed by an expert blind to the participant's diagnosis, ensuring an unbiased measurement [3].

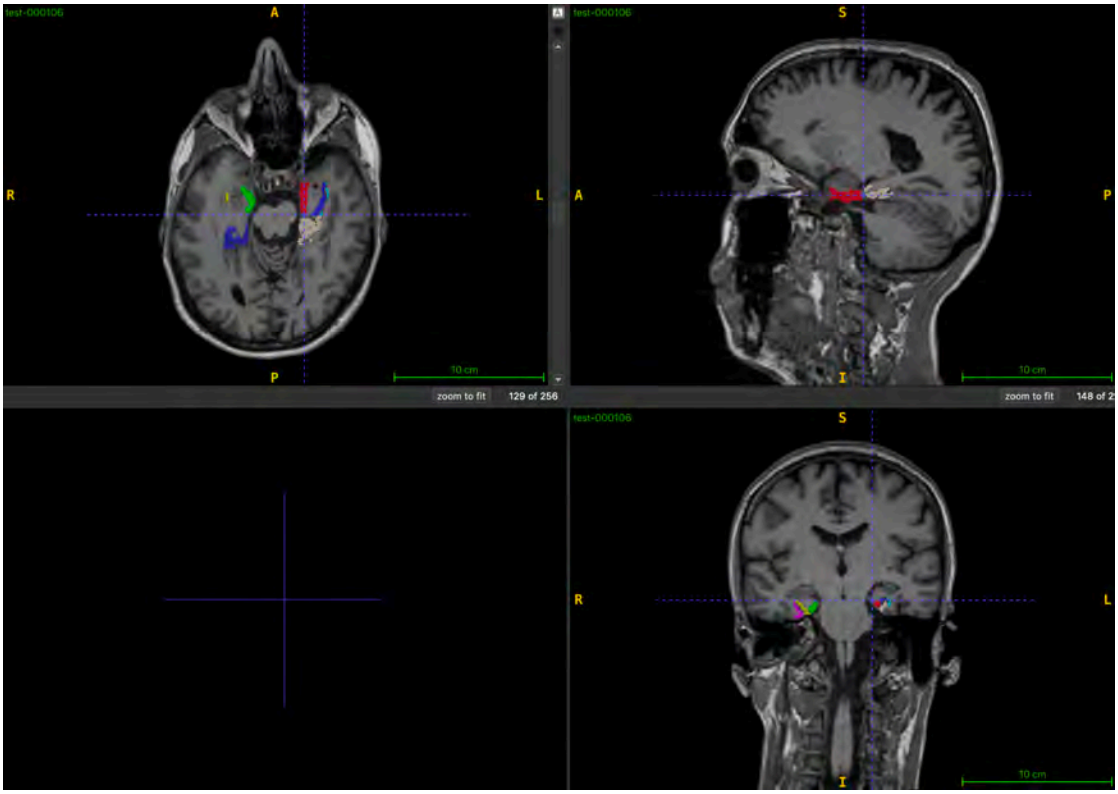


Figure 1.1.: Schematic illustration of PHG regions in an MRI sample. The regions of interest include the entorhinal cortex (red, green), medial perirhinal cortex (turquoise, pink), lateral perirhinal cortex (blue, yellow), and parahippocampal cortex (white, violet).

The study involved 121 participants, comprised of 64 healthy control participants and 57 individuals diagnosed with AD. In the AD group, 34 patients were diagnosed with dementia, and 23 had amnestic Mild Cognitive Impairment (aMCI). aMCI is a mild cognitive dysfunction in which individuals develop dementia, remain stable, or even return to normal [3][4].

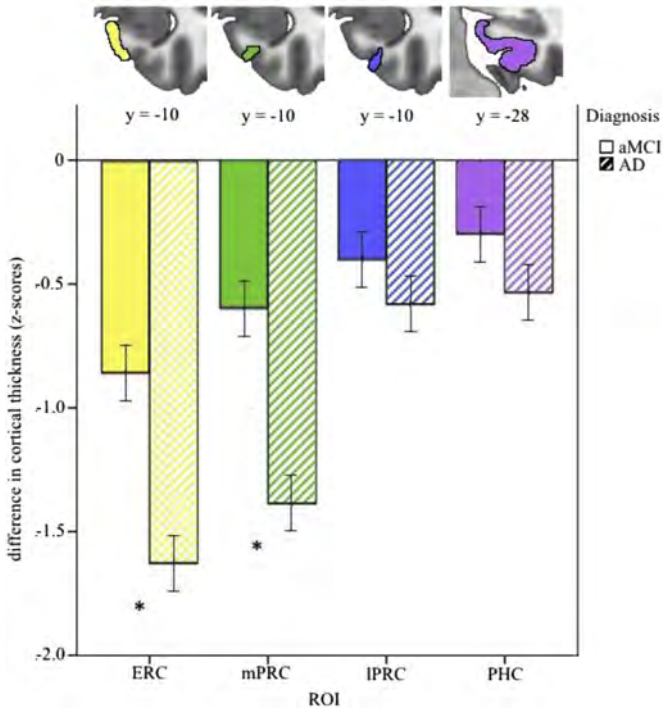


Figure 1.2.: mean thickness values for aMCI and AD participants for each ROI from [3].

Figure 2.2 shows a notable difference in mean thickness in AD and aMCI participants' ERC and mPRC ROIs. Krumm et al.'s research offers a novel view of the changes associated with AD progression. Understanding the biological changes in the brain is crucial for diagnosing and managing neurodegenerative diseases and developing treatments to slow or halt disease progression.

Manual segmentation of the PHG subregions is a time-consuming process that can only be performed by an expert and can introduce variability. A reliable automated segmentation method is essential to support research in early detection and intervention of AD. This VT aims to create an AI model capable of accurately segmenting the PHG regions in scans, thereby aiding the researchers in accelerating their analysis.

1.2. Situation Assessment

This semester-long project, conducted as part of the Master of Science in Engineering (MSE) program at the Zurich University of Applied Sciences (ZHAW), is conducted by Martin Oswald, under the supervision of Dr. Manuel Dömer and Dr. Ahmed Abdulkadir.

The research employs open-source Python libraries for data exploration and model development. Data is securely stored in the Center for Artificial Intelligence's (CAI) Ceph

cluster, while computationally intensive tasks, including model training and inference, leverage the CAI's GPU cluster to enable scalable processing.

The project focuses on developing a workflow for volumetric analysis to assess cortical thickness indirectly in the subregions of the PHG. While direct measurement of cortical thickness often requires complex surface-based processing that is beyond the project's scope, volumetric measures derived from segmentation masks provide a computationally efficient and meaningful alternative for extracting anatomical insights.

To ensure a structured and reproducible approach, the project follows the data science process outlined by Dömer et al. [5], including the stages of **project understanding**, **data acquisition and exploration**, **modeling**, **evaluation**, **implementation**, and **delivery**. However, deployment and monitoring phases are explicitly excluded to align with the semester timeline.

1.3. Project Goals and Success Criteria

The objective of this research is to develop an AI model capable of accurately segmenting the PHG subregions from MRI data. Building on the work by Krumm et al., who introduced an ensemble-based automated segmentation method where each model is specialized in a single region, this project aims to address key limitations of ensemble approaches. Specifically, the goals are to:

- **Develop a Leaner Model:** Integrate functionality into a single, compact AI model for segmentation, reducing complexity and resource overhead compared to ensembles of specialized models.
- **Enable Deployment for Research Use:** Providing a deployable solution that researchers can directly use in clinical or experimental workflows.

1.4. Use of Generative AI

By the supervisors' guidance, Large Language Models were utilized during the writing process to enhance academic tone, improve clarity, and correct grammatical errors. The factual content and intellectual contribution remain entirely the author's work.

2. Project Charta

2.1. Problem Definition

Dementia is a general term covering several diseases impacting cognitive abilities like memory and thinking. Alzheimer's disease (AD) is the most common form of dementia, causing 60-70% of dementia cases [1].

People often use AD and dementia interchangeably, but the terms refer to different aspects of the same condition. The Alzheimer's Association Workgroup defines AD as a long-term brain disorder characterized by changes in brain biology and structure, such as the presence of Amyloid plaques (extracellular deposits of amyloid- (A) peptides) and Neurofibrillary tangles (intracellular aggregates of tau protein). On the other hand, dementia refers to the stage where cognitive decline becomes evident. This distinction is important because individuals can have AD without displaying symptoms of dementia [2]. There is a wide range of biomarkers to detect AD, including blood tests, plasma tests, cerebrospinal fluid tests, and imaging. The literature review describes dementia symptoms and biomarkers.

Krumm et al. investigate AD progression by measuring the thickness of different brain regions. The early stages of AD involve shrinking of the parahippocampal gyrus (PHG), which includes the entorhinal cortex (ERC), perirhinal cortex (mPRC), lateral perirhinal cortex (IPRC), and parahippocampal cortex (PHC), as illustrated in Figure 2.1. The researchers used Magnetic resonance imaging (MRI) to scan the participant's brain and segment the different regions of interest (ROIs) in the PHG. The segmentation was performed by an expert blind to the participant's diagnosis, ensuring an unbiased measurement [3].

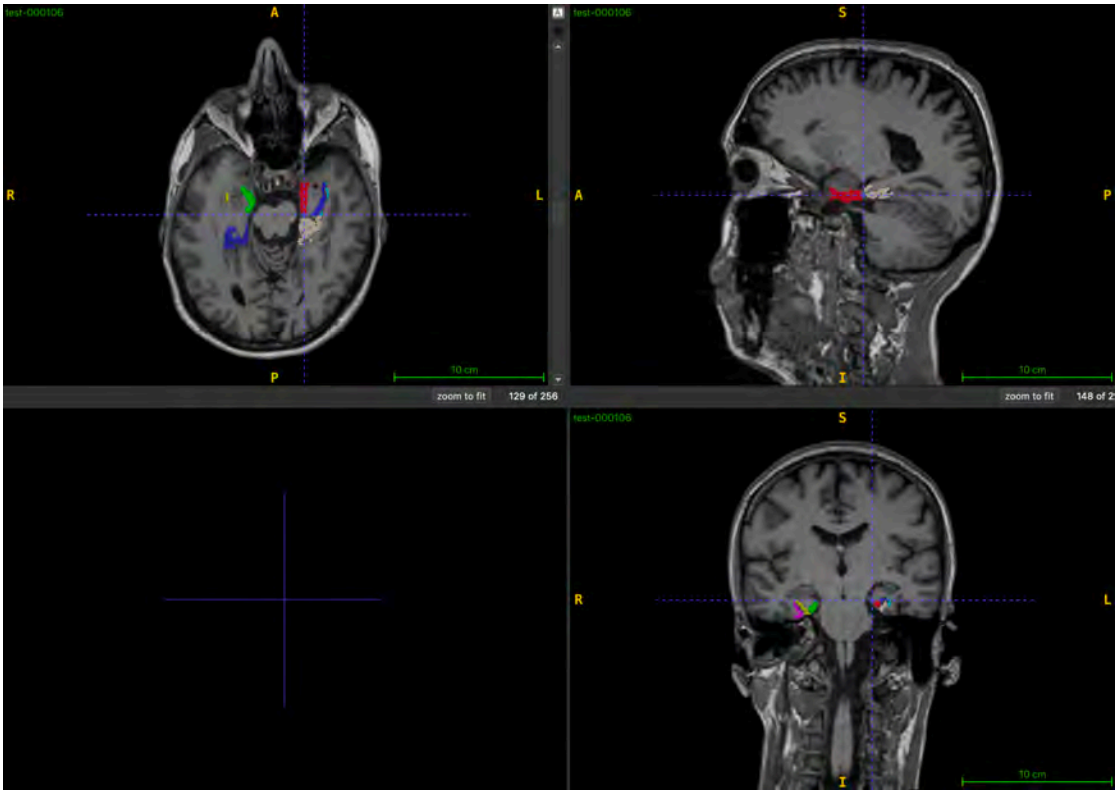


Figure 2.1.: Schematic illustration of the parahippocampal gyrus (PHG) regions in an MRI sample. The regions of interest include the entorhinal cortex (red, green), medial perirhinal cortex (turquoise, pink), lateral perirhinal cortex (blue, yellow), and parahippocampal cortex (white, violet).

The study involved 121 participants, comprised of 64 healthy control participants and 57 individuals diagnosed with AD. In the AD group, 34 patients were diagnosed with dementia, and 23 had amnestic Mild Cognitive Impairment (aMCI). aMCI is a mild cognitive dysfunction in which individuals develop dementia, remain stable, or even return to normal [3][4].

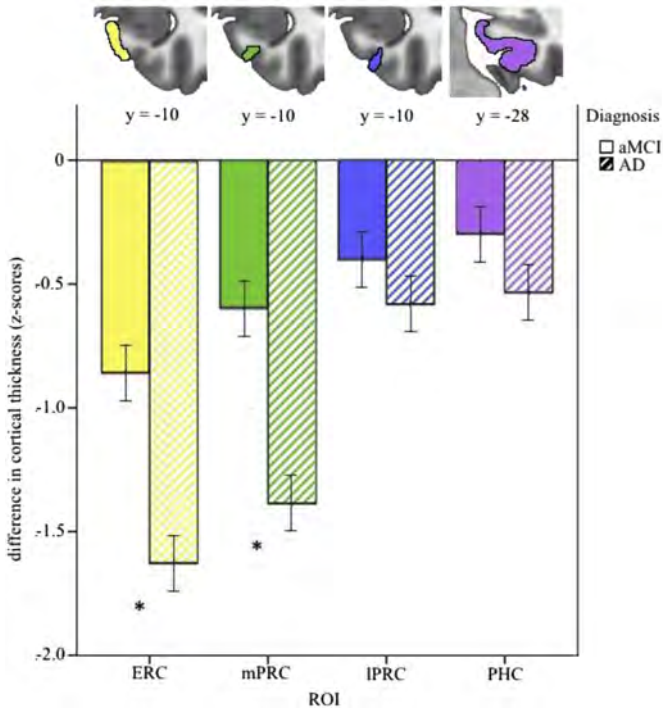


Figure 2.2.: mean thickness values for aMCI and AD participants for each ROI from [3].

Figure 2.2 shows a notable difference in mean thickness in AD and aMCI participants' ERC and mPRC ROIs. Krumm et al.'s research offers a novel view of the changes associated with AD progression. Understanding the biological changes in the brain is crucial for diagnosing and managing neurodegenerative diseases and developing treatments to slow or halt disease progression.

Manual segmentation of the PHG subregions is a time-consuming process that can only be performed by an expert and can introduce variability. A reliable automated segmentation method is essential to support research in early detection and intervention of AD. This VT aims to create an AI model capable of accurately segmenting the PHG regions in scans, thereby aiding the researchers in accelerating their analysis.

2.2. Situation Assessment

This VT is a semester-long project within the Master of Science in Engineering (MSE) program at the Zurich University of Applied Sciences (ZHAW). The project is being performed by Martin Oswald and supervised by Dr. Manuel Dömer and Dr. Ahmed Abdulkadir.

The project utilizes open-source Python libraries for data exploration and model development, the CAI's Ceph cluster for data storage, and the CAI's GPU cluster for model training and inference.

2.3. Project Goals and Success Criteria

This research aims to develop an AI model for precise segmentation of parahippocampal gyrus (PHG) subregions from MRI data, focusing on the ERC, mPRC, IPRC, and PHC regions. The project scope includes model development and data analysis but excludes deployment and monitoring. Unlike Krumm et al.'s cortical thickness approach, this study uses volumetric analysis as a proxy measure due to project constraints.

2.4. Data Mining Goals

The dataset provided by Krumm et al. includes MRI scans, expert-annotated ground truth segmentation masks, and additional participant metadata. The primary objective of this study is to address a complex segmentation task, where the model must precisely delineate small anatomical regions of interest (ROIs) while disregarding the predominant background. The extreme class imbalance compounds this challenge, as the background class is vastly overrepresented compared to the targeted regions, which comprise a minimal fraction of the image volume.

Additionally, while Krumm et al. utilized cortical thickness measurements in their analysis, this study focuses instead on deriving regional volumes as proxy measures. The conversion from segmentation masks to cortical thickness, while clinically significant, is beyond this project's scope. The Data Report elaborates on details regarding the dataset properties.

2.5. Use of Generative AI

By the supervisors' guidance, Large Language Models were utilized during the writing process to enhance academic tone, improve clarity, and correct grammatical errors. The factual content and intellectual contribution remain entirely the author's work.

3. Data Report

This data report compiles all information related to the datasets used in this project. By documenting the characteristics and preprocessing details of the data, we ensure traceability and reproducibility and provide a foundation for systematic expansion in future studies. The datasets analyzed in this work focus on 3D brain MRI scans and associated metadata for Alzheimer's disease-related segmentation tasks.

3.1. Raw Data

3.1.1. Dataset Overview

Table 3.1 presents the source datasets and their storage specifications.

Table 3.1.: Primary datasets utilized in this study.

Name	Source	Storage Location	Storage Path
AMBI	Provided by Krumm et al.	CAI cluster	/cluster/projects/movt1/ambi/
BAMBI	Provided by Krumm et al.	CAI cluster	/cluster/projects/movt1/ambi/

3.1.2. AMBI Dataset Specifications

The AMBI dataset consists of 111 unique MRI acquisitions from participants stratified across three diagnostic categories: Alzheimer's Dementia (AD; n=33), Normal Controls (NC; n=49), and amnestic Mild Cognitive Impairment (aMCI; n=21). The dataset encompasses T1-weighted three-dimensional brain MRI scans preprocessed via FreeSurfer for standardized head orientation alignment, accompanied by sparse segmentation masks delineating specific neuroanatomical regions.

The dataset originates from Krumm et al.'s investigation of cortical atrophy patterns in PHG regions, explicitly examining the ERC, mPRC, lPRC, and PHC. Image acquisition was performed on a 3T Siemens scanner with T1-weighted imaging parameters at isotropic 1 mm^3 voxel resolution [3]. A blinded rater manually annotated segmentation masks for ROIs to ensure unbiased assessment [3].

3.1.3. BAMBI Dataset Specifications

The BAMBI dataset comprises 42 MRI scans with corresponding sparse segmentation masks targeting ROIs identical to those of the AMBI dataset. While maintaining consistent imaging protocols and anatomical region specifications with AMBI, this dataset lacks participant metadata.

3.2. Processed Data

3.2.1. Data Processing Pipeline

The AMBI dataset was randomly split into training (n=83) and validation (n=20) subsets following standard machine learning practices. The BAMBI dataset serves as an independent test set. Raw data was transformed into the Hugging Face datasets format to optimize computational workflows and ensure reproducibility. This transformation facilitates standardized preprocessing pipelines, data transformations, and format normalization while maintaining the original feature structure.

Data governance and security protocols are strictly maintained, with both datasets securely stored on ZHAW's CAI Ceph cluster and access restricted to authorized research team members.

3.2.2. Feature Specifications

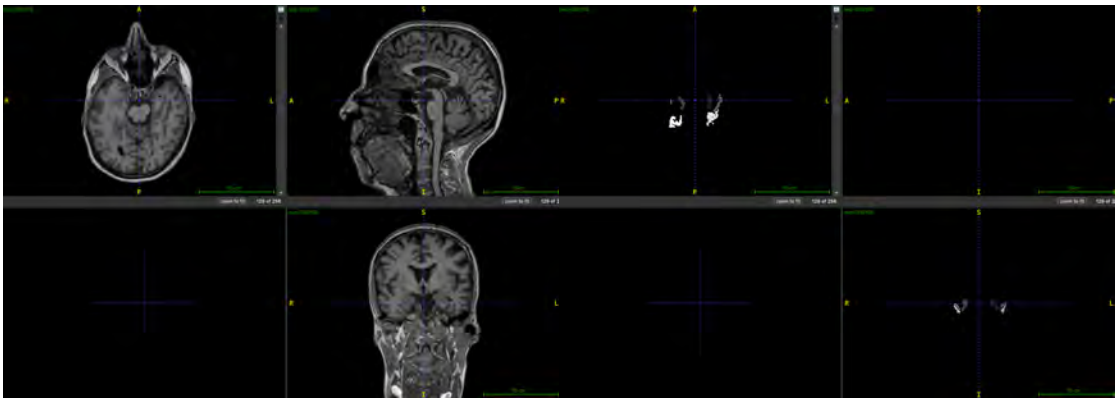
Table 3.2 details the complete feature set and corresponding specifications of the processed datasets.

Table 3.2.: Feature specifications of the processed datasets.

Feature Name	Data Type	Description
`mcid`	String	Unique subject identifier
`image`	3D Array (256x256x256, uint8)	3D brain MRI scan for each subject
`sparse`	3D Array (256x256x256, uint8)	Sparse segmentation mask; integers represent segmented regions (background = 0).
`affine`	2D Array (4x4, float64)	Affine transformation matrix for spatial alignment
`age`	Float32	Participant's age

Feature Name	Data Type	Description
`sex`	Categorical (0=male, 1=female)	Participant's biological sex
`diagnosis`	Categorical (0=AD, 1=NC, 2=aMCI)	Diagnosis (Alzheimer's Disease, Normal Control, or amnestic Mild Cognitive Impairment)
`mmse`	Int16	Mini-Mental State Examination score
`education`	Int16	Years of education

3.2.3. Visualization and Segmentation Labels



(a) Sample image (b) Sample sparse

Figure 3.1.: Dataset sample containing Figure 3.1a T1-weighted MRI scan and Figure 3.1b its corresponding sparse segmentation mask

Table 3.3.: Anatomical region encoding scheme for segmentation masks.

Label Value	Anatomical Region
0	Background
1	Left Entorhinal Cortex (ERC)
2	Right Entorhinal Cortex (ERC)
3	Left Lateral Perirhinal Cortex (lPRC)
4	Right Lateral Perirhinal Cortex (lPRC)
5	Left Medial Perirhinal Cortex (mPRC)
6	Right Medial Perirhinal Cortex (mPRC)
7	Left Parahippocampal Cortex (PHC)
8	Right Parahippocampal Cortex (PHC)

Figure 3.1 illustrates representative examples of the dataset components, showing both the raw T1-weighted MRI scan and its corresponding sparse segmentation mask. Table 3.3 provides the intensity value encoding scheme for anatomical regions in the segmentation masks.

3.3. Exploratory Data Analysis

This section presents key statistical analyses and distributions of the dataset features to provide insights into data characteristics and quality. Statistical distributions of all features are available in the supporting material.

3.3.1. Demographic Distribution

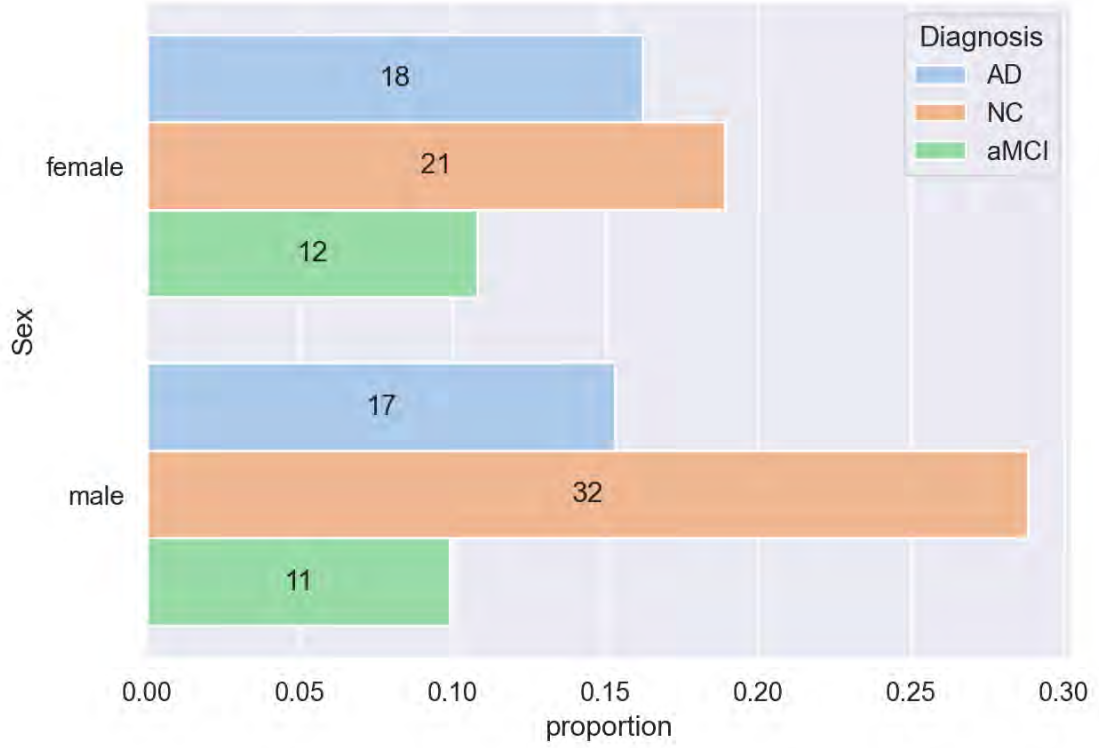


Figure 3.2.: Proportions and counts of diagnosis between male/female participants.

Figure 3.2 illustrates the sex distribution across diagnostic groups. Among AD patients, the distribution was relatively balanced, with 18 female and 17 male participants. The NC group showed more male participants (32 male vs. 21 female), while the aMCI group was approximately balanced (12 female vs. 11 male).

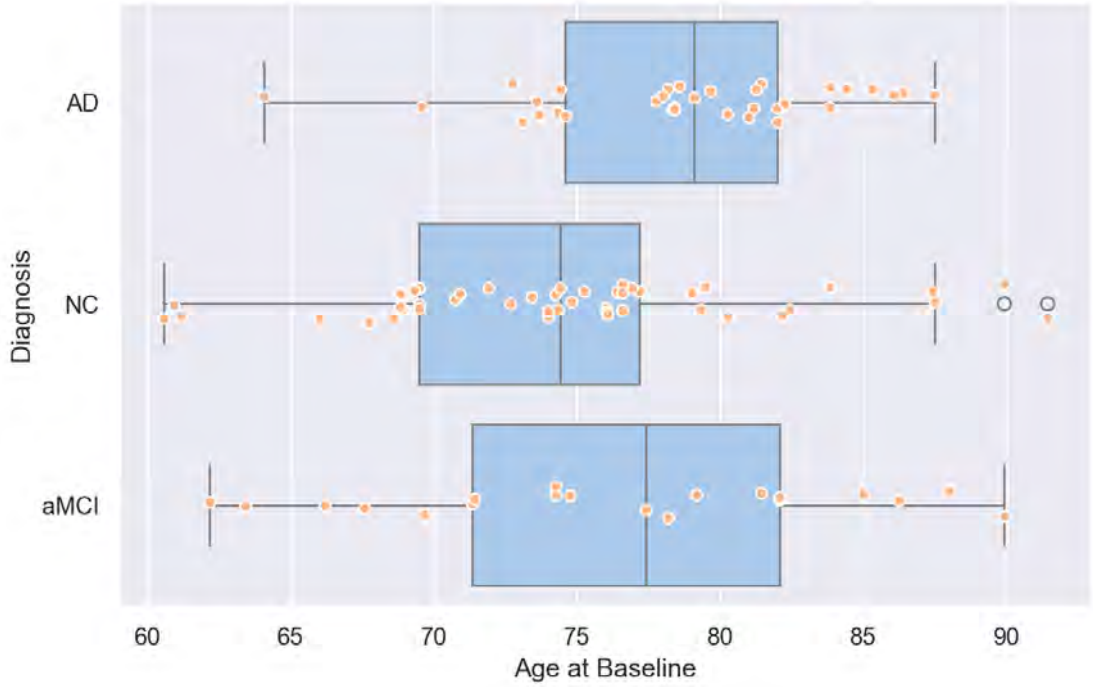
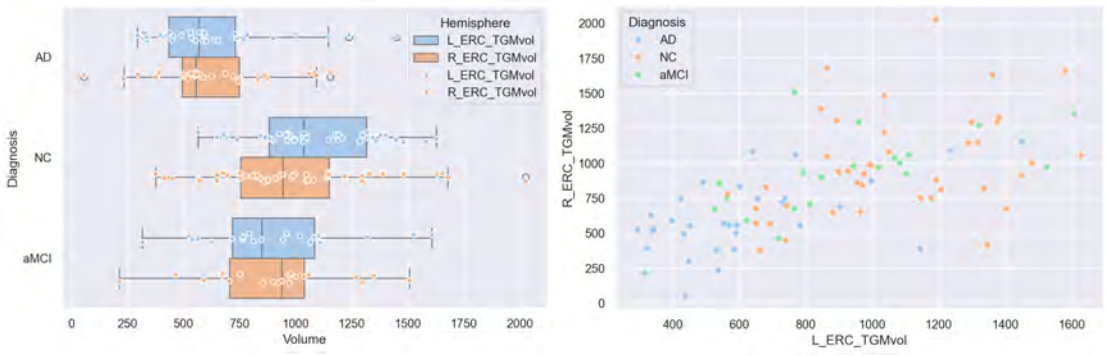


Figure 3.3.: Boxplot of age distribution across diagnostic groups.

Figure 3.3 demonstrates the age distribution across diagnostic groups. The AD group showed the highest mean age (78.95 ± 5.17 years), followed by the aMCI group (76.78 ± 8.31 years), while the NC group had the lowest mean age (74.67 ± 6.93 years). The overall age range across all participants was 60.58 to 91.42 years, with a mean of 76.47 ± 6.93 years. The age distributions show considerable overlap between groups, though AD patients tend to be slightly older.

3.3.2. Anatomical Measurements

Entorhinal Cortex

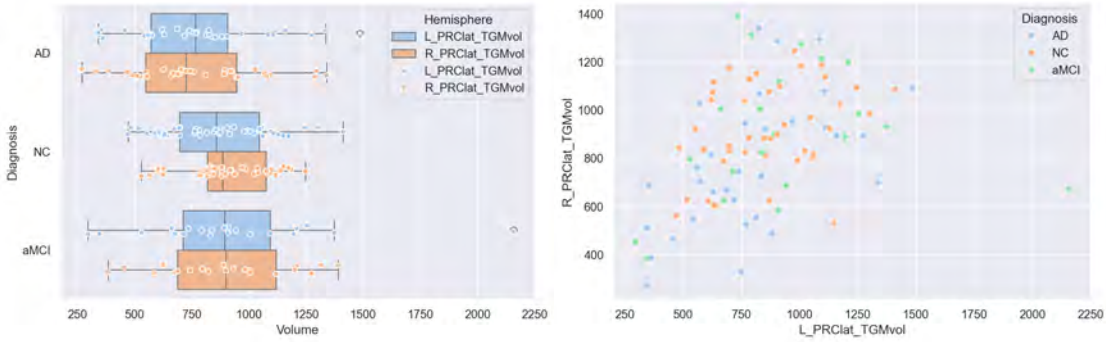


(a) boxplot (b) scatterplot

Figure 3.4.: Entorhinal cortex volumes across diagnostic groups shown as Figure 3.4a box and Figure 3.4b scatter plots

As shown in Figure 3.4, ERC volumes were markedly reduced in AD patients compared to healthy controls, with mean volumes of 613.3 mm³ (±275.1) and 613.4 mm³ (±254.8) for left and right hemispheres respectively, versus 1074.9 mm³ (±269.1) and 994.4 mm³ (±346.3) in controls. The aMCI group showed intermediate values, suggesting a progressive pattern of atrophy.

Perirhinal Cortex lateral



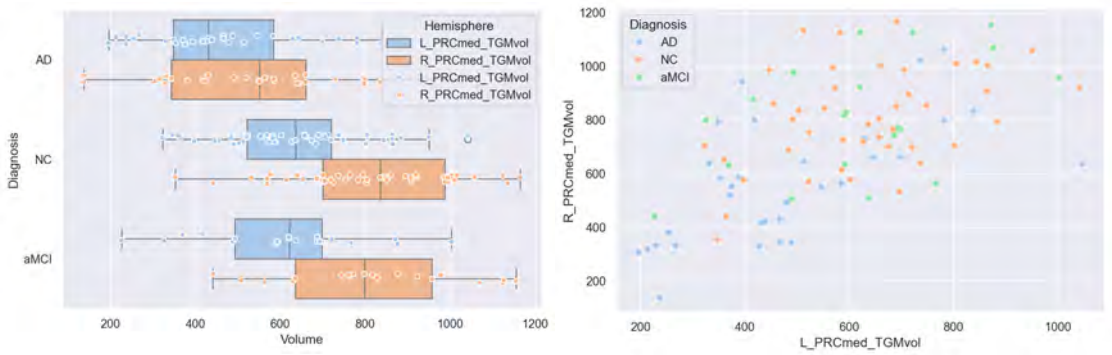
(a) boxplot (b) scatter

Figure 3.5.: Lateral perirhinal cortex volumes across diagnostic groups shown as Figure 3.5a box and Figure 3.5b scatter plots.

Figure 3.5 demonstrates that PRClat volumes were less severely affected than the ERC. AD patients showed mean volumes of 775.5 mm³ (±299.5) and 762.3 mm³ (±288.4) for left and right hemispheres, compared to control values of 862.5 mm³ (±222.1) and 911.6 mm³ (±181.7). Notably, aMCI patients maintained relatively preserved volumes. The scatter plot in Figure 3.5b reveals that most data points lie above the 45-degree line

(y=x), indicating systematically larger right hemisphere volumes than left hemisphere volumes across all diagnostic groups.

Perirhinal Cortex medial

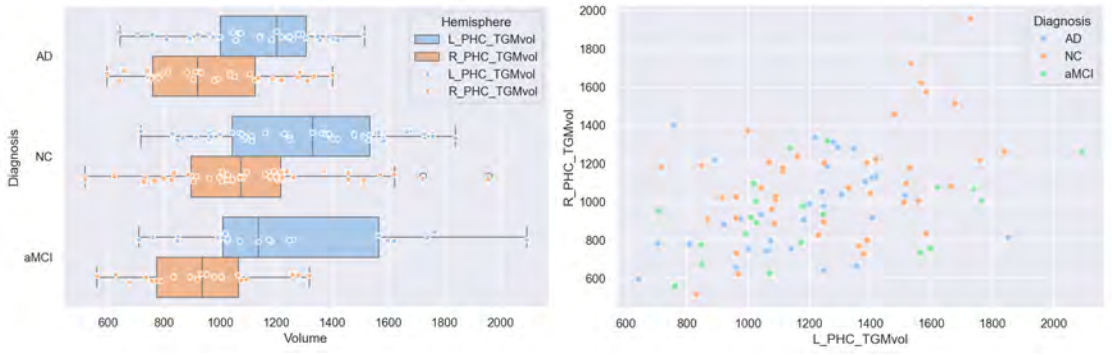


(a) boxplot (b) scatter

Figure 3.6.: Medial perirhinal cortex volumes across diagnostic groups shown as Figure 3.6a box and Figure 3.6b scatter plots.

The PRCmed measurements, illustrated in Figure 3.6, revealed a significant volume reduction in AD patients (left: 473.0 mm^3 $\pm$204.8; right: 554.4 mm^3 $\pm$233.7) compared to controls (left: 636.1 mm^3 $\pm$162.8; right: 822.1 mm^3 $\pm$189.8). The aMCI group showed intermediate atrophy, particularly in the right hemisphere. The scatter plot in Figure 3.6b reveals a consistent hemispheric asymmetry, with right PRCmed volumes systematically larger than left PRCmed volumes across all diagnostic groups, a pattern also observed in the lateral perirhinal cortex measurements.

Parahippocampal Gyrus



(a) boxplot (b) scatter

Figure 3.7.: Parahippocampal cortex volumes across diagnostic groups shown as Figure 3.7a box and Figure 3.7b scatter plots.

Figure 3.7 shows PHC volumes were relatively preserved compared to other regions, with less pronounced differences between groups. AD patients showed mean volumes of 1161.7 mm^3 ($\pm$252.4) and 954.0 mm^3 ($\pm$226.7) for left and right hemispheres, versus 1288.1 mm^3 ($\pm$303.4) and 1103.5 mm^3 ($\pm$324.5) in controls. The aMCI group demonstrated similar patterns to controls, suggesting this region may be affected later in disease progression.

These volumetric analyses support the pattern of differential regional vulnerability in early AD, with the ERC and PRCmed showing earlier and more pronounced atrophy compared to the PRClat and PHC, consistent with the findings of Krumm et al.

4. Modelling Report

This report documents the development and evaluation of deep learning models for automated segmentation of parahippocampal gyrus (PHG) regions from MRI scans. It documents a baseline and two improvement approaches that build upon each other.

4.1. Evaluation Metrics

We employ two complementary metrics to assess segmentation performance: the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU). Both metrics range from 0 to 1, where 1 indicates perfect segmentation overlap, and 0 indicates no overlap between predicted and ground-truth segmentations.

The Dice Similarity Coefficient, also known as the F1-score for binary classification, measures spatial overlap between predicted and ground truth segments [6]. It is calculated as:

- $DSC = 2|X\ Y| \ / \ (|X| + |Y|)$

Intersection over Union (IoU), also known as the Jaccard index, calculates the ratio of overlap to the total region encompassed by both segmentations [6]:

- $IoU = |X\ Y| \ / \ |X\ Y|$

Both metrics are particularly suitable for evaluating medical image segmentation tasks, as they are sensitive to over- and under-segmentation while robust to class imbalance [6] [7].

4.2. Establishing a Baseline

Given the data mining goals outlined in the project charta, which focus on accurate segmentation of PHG regions from MRI scans, establishing a baseline model is crucial for subsequent improvements and comparisons.

4.2.1. model Architecture

The baseline implementation utilizes a 3D U-Net architecture, as described in [8], implemented through the MONAI framework [9]. The model architecture consists of a single input channel and nine output channels - eight for the anatomical ROIs detailed in the data report and one for the background class. The network features three encoding/decoding levels, maintaining similarity with the original U-Net design, with two residual units per level. To mitigate overfitting, dropout regularization is applied with a probability of 0.3.

4.2.2. Data Preprocessing and Training Protocol

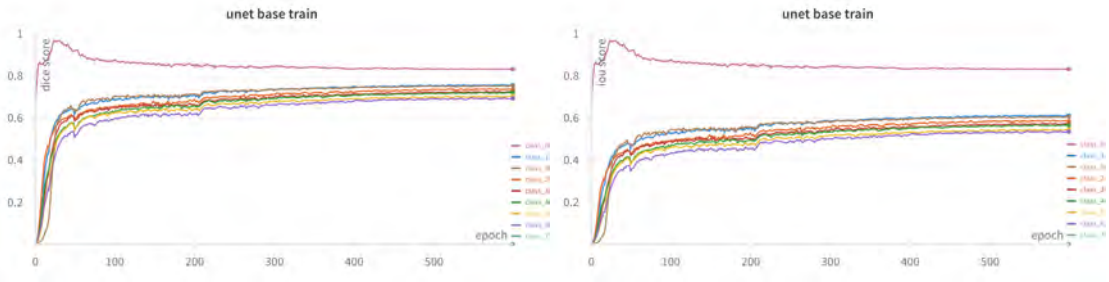
The training pipeline incorporates the following preprocessing steps:

1. Spatial cropping to 128x128x128 from the image center, ensuring ROI inclusion
2. Data type conversion from `uint8` to `float32`
3. One-hot encoding of sparse masks into nine classes
4. Input intensity normalization through linear scaling from [0, 255] to [-1, 1]

The training protocol employs a dice loss function from [10], notably excluding the background class (class 0) from loss calculation to address the severe class imbalance. The training was conducted with batch size 16 and utilized a cosine learning rate scheduler, starting at 0.1 and decreasing to 1e-7 over 600 epochs. The model was trained on an NVIDIA V100 GPU with 32GB VRAM, requiring approximately 48 hours for completion. The pytorch-lightning [11] library is used to run and seed the experiments with seed 0, ensuring that all experiments are reproducible.

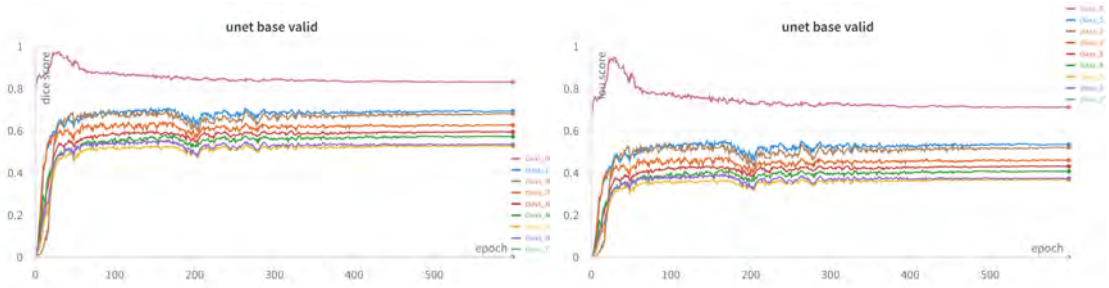
4.2.3. results

Figures Figure 4.1 and Figure 4.2 visualize the training and validation metrics during model training. These figures present DSC and IoU scores across training epochs.



(a) dice score (b) iou score

Figure 4.1.: Training performance metrics over 600 epochs showing the progression of Figure 4.1a Dice Similarity Coefficient and Figure 4.1b Intersection over Union scores for each anatomical region.



(a) dice score (b) iou score

Figure 4.2.: Validation performance metrics over 600 epochs showing the progression of Figure 4.2a Dice Similarity Coefficient and Figure 4.2b Intersection over Union scores for each anatomical region.

A notable observation from these progression curves is the paired learning pattern for bilateral anatomical regions. This pattern is particularly evident in Figure 4.2, where metric trajectories for corresponding left and right structures closely mirror each other throughout the training process. This behavior suggests the model successfully captures the anatomical symmetry inherent in brain structures.

Table 4.1.: Baseline U-Net performance metrics (DSC and IoU) for each anatomical region in both training and validation sets.

Class	Region	train DSC	train IoU	valid DSC	valid IoU
0	Background	0.83085	0.7108	0.83156	0.71178
1	L-ERC	0.75618	0.61189	0.69266	0.53492
2	R-ERC	0.73326	0.5829	0.62648	0.46028
3	L-lPRC	0.72097	0.56864	0.59456	0.43116

Class	Region	train DSC	train IoU	valid DSC	valid IoU
4	R-lPRC	0.71879	0.56454	0.57239	0.40704
5	L-mPRC	0.70505	0.54763	0.52741	0.36722
6	R-mPRC	0.69292	0.53478	0.53464	0.37419
7	L-PHC	0.00216	0.00108	0.00234	0.00117
8	R-PHC	0.75274	0.60600	0.68090	0.52049
	Mean	0.65699	0.52536	0.56255	0.42314

The detailed performance metrics for each ROI presented in Table 4.1 reveal varying degrees of success across different regions. The model achieved strongest performance in segmenting the left entorhinal cortex (L-ERC) on the training set (DSC = 0.756) and maintained reasonable performance on the validation set (DSC = 0.693).

The right parahippocampal cortex (R-PHC) also showed robust performance (training DSC = 0.753, validation DSC = 0.681). However, the most concerning finding is the model's complete failure to segment the left parahippocampal cortex (L-PHC), achieving near-zero DSC scores (training = 0.002, validation = 0.002). This is particularly problematic given that the PHC represents one of the larger ROIs in the dataset, with volumes typically ranging from 1161.7 mm^3 ($\pm$252.4) to 1288.1 mm^3 ($\pm$303.4) across different diagnostic groups, as reported in the data report.

The remaining regions show moderate performance, with DSC values ranging between 0.69 and 0.72 on the training set and 0.52 and 0.59 on the validation set. While these results are promising for a baseline model, they indicate substantial room for improvement, particularly in addressing the L-PHC segmentation failure and enhancing the overall consistency of region detection across hemispheres.

4.3. Manipulating Losses

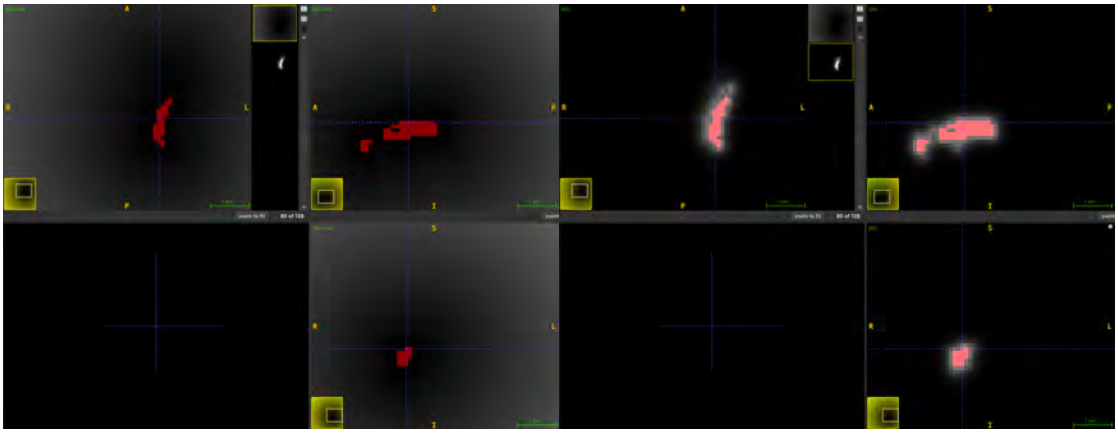
Following the baseline results, which showed promising performance but failed completely to segment the L-PHC, our first iteration focused on loss function modifications. This approach was mainly motivated by the extreme class imbalance inherent in our segmentation task, where the ROIs occupy only a tiny fraction of the total brain volume.

4.3.1. Approach

While maintaining the same model architecture, preprocessing pipeline, and training protocol as established in the baseline, we explored different loss function combinations to address the class imbalance and improve segmentation performance:

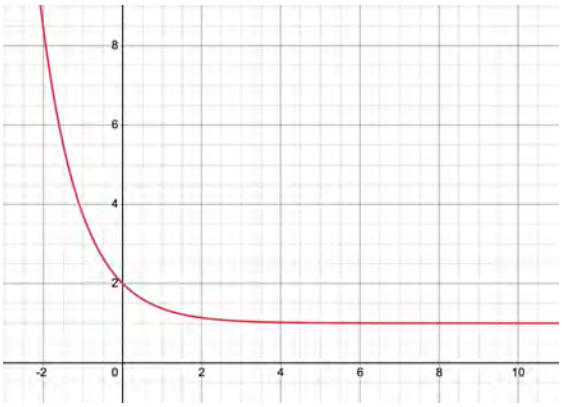
- **L1 Dice and Cross-Entropy Loss Combination:** Integration of the standard dice loss as described in [10] with weighted cross-entropy loss to leverage both regional overlap metrics and pixel-wise classification information. The cross-entropy loss was weighted to address class imbalance, applying a weight of 1.0 to all ROI classes and 0.001 to the background class.
- **L2 Focal Loss:** Implementation of the focal loss function to address the class imbalance by dynamically adjusting the contribution of hard-to-classify examples while down-weighting easy examples [12].
- **L3 Distance-Transform Enhanced Focal Loss:** Extension of the focal loss from [9] through the integration of distance transform information, applied as a custom post-criterion transform to the non-reduced focal loss output. Note that this enhancement could only be applied to focal loss due to shape compatibility requirements, as the dice loss output structure differs from the required input format for distance transformation.
- **L4 Dice and Focal Loss Combination:** Fusion of dice loss with focal loss to combine the benefits of regional overlap metrics with focused classification capabilities.

4.3.2. Training Protocol



(a) distance transformed sample

(b) distance transformed and decayed sample



(c) decay function

Figure 4.3.: Distance transform visualization showing: raw distance transform applied to a sparse sample Figure 4.3a, processed output with higher intensities indicating more substantial weighting near ROI boundaries Figure 4.3b, and the parametric decay function e^(-x) + used for transform modulation Figure 4.3c.

The distance transform enhancement leverages a multi-step approach to emphasize ROI boundaries during training. Initially, we compute the Euclidean distance transform using Scipy's [13] transformation method on one-hot encoded sparse matrices. This generates a tensor containing minimal distances from each voxel to its respective ROI boundary, as shown in Figure 4.3a.

To modulate the boundary emphasis, we apply a parametric decay function e^(-x) + , visualized in Figure 4.3c. We set =10, =0.8, and =0.01 for our experiments, creating a moderate exponential decay from the ROI boundaries. This parameter configuration

results in a maximum weight of 10 at the ROI boundaries, gradually decreasing to the minimum weight of 0.01 as distance increases, with =0.8 ensuring this decay occurs more gradually, allowing for a broader zone of influence around the ROI boundaries. The processed output, demonstrated in Figure 4.3b, creates focused emphasis zones around ROI borders, enabling the model to learn more precise segmentation boundaries while maintaining minimal influence from distant voxels.

For each loss function configuration (L1-L4), multiple training runs were performed with varying initial learning rates to determine optimal hyperparameters. Through this systematic exploration, we established optimal initial learning rates of 0.005 for configurations L1, L2, and L3, while L4 required a lower initial learning rate of 0.0001. All other training parameters remained consistent with the baseline configuration, enabling direct performance comparisons.

4.3.3. Results

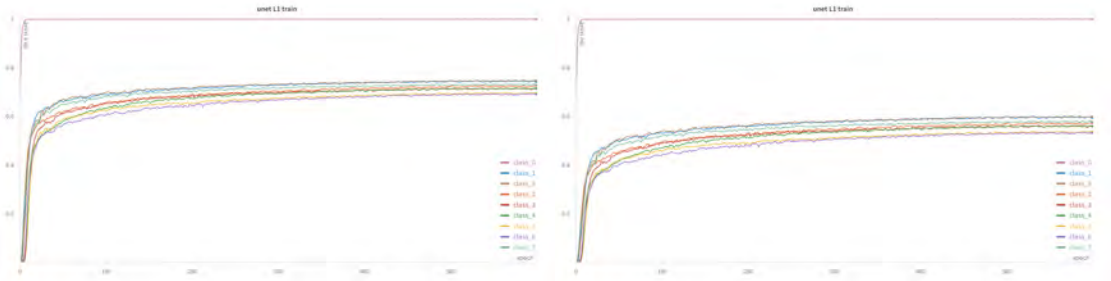
Table 4.2.: Performance metrics (DSC and IoU) across different loss functions (L1-L4) for each anatomical region in the training set.

Class	Region	L1/DSC	L1/IoU	L2/DSC	L2/IoU	L3/DSC	L3/IoU	L4/DSC	L4/IoU
0	Background	0.9992	0.9984	0.9985	0.9971	0.9991	0.9983	0.3577	0.2178
1	L-ERC	0.7402	0.5915	0.2718	0.1578	0.1974	0.1101	0.0007	0.0003
2	R-ERC	0.7249	0.5717	0.1685	0.0927	0.1744	0.0961	0.0009	0.0004
3	L-lPRC	0.7128	0.5584	0.0025	0.0012	0.2730	0.1636	0.0001	0.0
4	R-lPRC	0.7134	0.5585	0.0013	0.0006	0.1921	0.1098	0.0005	0.0002
5	L-mPRC	0.6959	0.5368	0.0147	0.0074	0.2889	0.1734	0.0007	0.0003
6	R-mPRC	0.6907	0.5317	0.0186	0.0094	0.2315	0.1351	0.0007	0.0003
7	L-PHC	0.7286	0.5760	0.3754	0.2329	0.2589	0.1499	0.0011	0.0005
8	R-PHC	0.7470	0.5987	0.0133	0.0067	0.2454	0.1406	0.0010	0.0005
	Mean	0.7503	0.6135	0.2072	0.1673	0.3179	0.2308	0.0404	0.0245

Table 4.3.: Performance metrics (DSC and IoU) across different loss functions (L1-L4) for each anatomical region in the validation set.

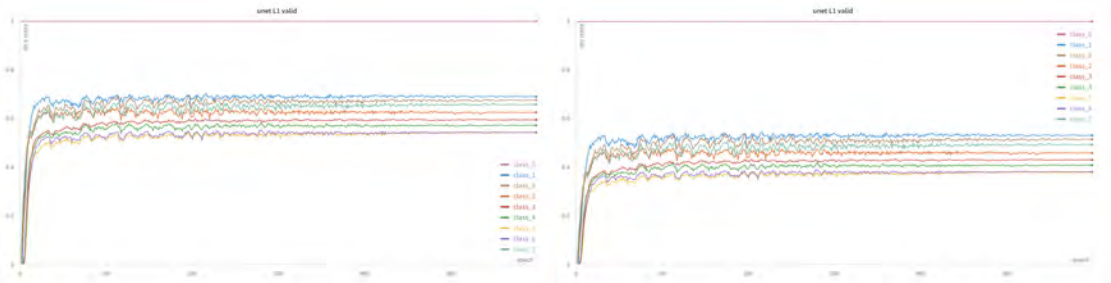
Class	Region	L1/DSC	L1/IoU	L2/DSC	L2/IoU	L3/DSC	L3/IoU	L4/DSC	L4/IoU
0	Background	0.9987	0.9975	0.9986	0.9972	0.9989	0.9979	0.3907	0.2427
1	L-ERC	0.6890	0.5298	0.2041	0.1141	0.1950	0.1087	0.0008	0.0003
2	R-ERC	0.6240	0.4576	0.1555	0.0847	0.1696	0.0936	0.0011	0.0005
3	L-lPRC	0.5940	0.4294	0.0	0.0	0.3046	0.1855	0.0	0.0
4	R-lPRC	0.5714	0.4078	0.0	0.0	0.1580	0.0894	0.0004	0.0002
5	L-mPRC	0.5408	0.3783	0.0004	0.0002	0.2403	0.1429	0.0007	0.0003

Class	Region	L1/DSC	L1/IoU	L2/DSC	L2/IoU	L3/DSC	L3/IoU	L4/DSC	L4/IoU
6	R-mPRC	0.5425	0.3807	0.0125	0.0063	0.1798	0.1027	0.0007	0.0003
7	L-PHC	0.6567	0.4922	0.4867	0.3251	0.2538	0.1460	0.0011	0.0005
8	R-PHC	0.6747	0.5141	0.0018	0.0009	0.2601	0.1498	0.0003	0.0001
	Mean	0.6547	0.5097	0.2066	0.1698	0.3067	0.2241	0.0439	0.0272



(a) dice score (b) iou score

Figure 4.4.: Training performance metrics for L1 model over 600 epochs showing the progression of Figure 4.4a Dice Similarity Coefficient and Figure 4.4b Intersection over Union scores for each anatomical region.



(a) dice score (b) iou score

Figure 4.5.: Validation performance metrics for L1 model over 600 epochs showing the progression of Figure 4.5a Dice Similarity Coefficient and Figure 4.5b Intersection over Union scores for each anatomical region.

The experimental results comparing different loss function configurations reveal significant performance variations across the tested approaches. Tables Table 4.2 and Table 4.3 present the comprehensive metrics for all configurations, with L1 (combined dice and cross-entropy loss) emerging as the most effective approach.

L1 achieved superior performance with mean DSC scores of 0.7503 and 0.6547 on training and validation sets, respectively. Notably, this configuration successfully addressed the L-PHC segmentation failure observed in the baseline model.

Figures Figure 4.4 and Figure 4.5 illustrate the learning progression over 600 epochs for the L1 configuration. The metrics exhibit bilateral symmetry in learning patterns, with paired performance trajectories for anatomical regions. While the training metrics converge to a narrower range, the broader dispersion in validation metrics suggests the model is learning robustly but also indicates potential sensitivity to variability in unseen data, warranting further evaluation for overfitting.

Alternative loss configurations showed less promising results. The focal loss approaches (L2 and L3) demonstrated mixed performance, with L3's distance transform enhancement showing some improvement over the standard focal loss (L2), though both underperformed compared to L1. The dice and focal loss combination (L4) showed particularly poor performance, with mean DSC scores below 0.05 across both datasets, suggesting challenges in loss component balancing.

The apendix contains detailed learning progression visualizations for configurations L2-L4 for comprehensive comparison.

4.4. Amplifying Segmentation with Regression

Building upon the insights from the loss function experiments, we developed a novel approach combining segmentation and regression with the goal of enhance ROI detection accuracy.

4.4.1. Approach

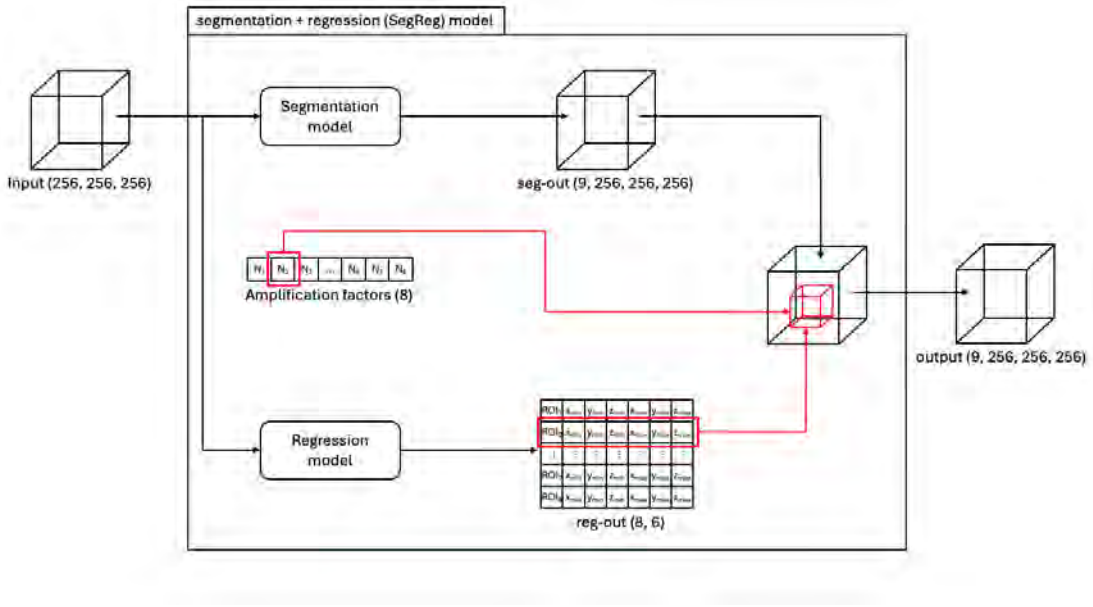


Figure 4.6.: SegReg model architecture.

We introduce SegReg, a novel architecture that processes input images in parallel through segmentation and regression pathways, as illustrated in Figure 4.6. The regression model shares the encoding path architecture of our previously used 3D U-Net but incorporates a dense layer at the lowest level to predict bounding boxes (x_min, y_min, z_min, x_max, y_max, z_max) for each ROI. The model's innovation lies in its amplification mechanism; the segmentation model's logits are selectively enhanced within predicted bounding box regions using trainable amplification parameters initialized at 1.0.

4.4.2. Regression Model Development

The regression component development involved evaluating two distinct approaches for ROI localization. The initial slice-based model predicted absolute slice positions (range: 0-255), while the alternative percentage-based model operated in normalized space (range: 0-1). Both models were trained with identical configurations: MSE loss, Adam optimizer with cosine learning rate decay from initial 0.0005 rate to 1e-7 over 600 epochs, and gradient clipping with a maximum norm of 0.5 to address stability issues.

The slice-based model achieved mean squared errors of 21.85588 (training) and 4.55073 (validation), while the percentage-based model demonstrated errors of 0.004412 (training) and 0.0671 (validation). The percentage-based approach produced more reliable

predictions through manual testing and validation, leading to its selection for integration into the final SegReg architecture.

4.4.3. Training Protocol

The training implementation maintained consistency with previous experiments while incorporating several key modifications. The preprocessing pipeline was enhanced to include ground truth conversion to bounding box tensors for regression training and integration of parallel processing paths for segmentation and regression.

Following the regression model training, two distinct configurations were evaluated:

- **M1:** Optimization of amplification parameters only (frozen regression and segmentation models)
- **M2:** Joint optimization of amplification parameters and segmentation model (frozen regression model)

M1 and M2 configurations underwent training for 100 epochs, with empirical testing determining an optimal learning rate of 0.0001. This structured approach allowed us to isolate the impact of the amplification mechanism while maintaining model stability.

4.4.4. Results

Table 4.4.: model 1 and model 2 on training set

Class	Region	mod-1/DSC	mod-1/IoU	mod-2/DSC	mod-2/IoU
0	Background	0.99987	0.99975	0.99952	0.99905
1	L-ERC	0.72989	0.57833	0.44440	0.28882
2	R-ERC	0.72231	0.56819	0.47745	0.31726
3	L-lPRC	0.74588	0.59890	0.42304	0.27233
4	R-lPRC	0.74456	0.59667	0.43974	0.28592
5	L-mPRC	0.70289	0.54445	0.35535	0.2181
6	R-mPRC	0.70360	0.54639	0.35810	0.22006
7	L-PHC	0.70675	0.54955	0.42636	0.27295
8	R-PHC	0.73029	0.57823	0.45223	0.29499
	Mean	0.75401	0.61783	0.48624	0.352164

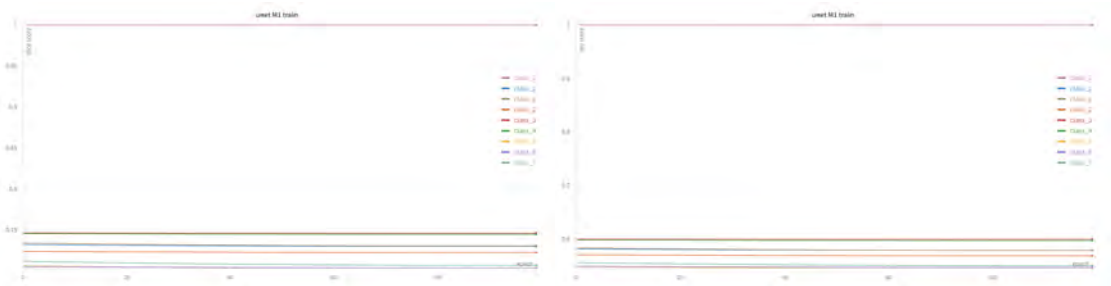
Table 4.5.: model 1 and model 2 on the validation set

Class	Region	mod-1/DSC	mod-1/IoU	mod-2/DSC	mod-2/IoU
0	Background	0.99985	0.99969	0.99933	0.99865

Class	Region	mod-1/DSC	mod-1/IoU	mod-2/DSC	mod-2/IoU
1	L-ERC	0.68782	0.52844	0.35051	0.21453
2	R-ERC	0.62395	0.45746	0.38803	0.24456
3	L-lPRC	0.59367	0.42904	0.34710	0.21315
4	R-lPRC	0.57127	0.40766	0.32157	0.19450
5	L-mPRC	0.54048	0.37791	0.24982	0.14478
6	R-mPRC	0.54178	0.38000	0.26319	0.15558
7	L-PHC	0.65547	0.49088	0.32794	0.19817
8	R-PHC	0.67451	0.51380	0.34869	0.21430
	Mean	0.65431	0.50943	0.39957	0.28647

Table 4.6.: amplification factors for each region in model 1 and model 2

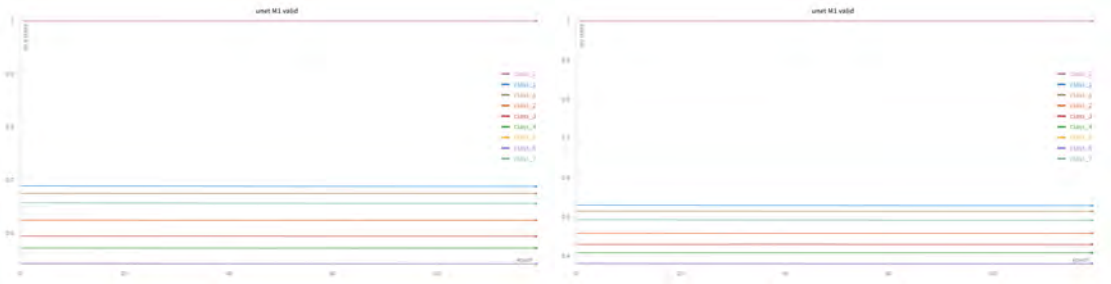
Class	Region	mod-1 amp	mod-2 amp
1	L-ERC	1.0385	1.0208
2	R-ERC	1.0414	1.0225
3	L-lPRC	1.0312	0.9853
4	R-lPRC	1.0453	1.0258
5	L-mPRC	1.0463	1.0151
6	R-mPRC	1.0366	1.0113
7	L-PHC	1.0482	1.0084
8	R-PHC	1.0484	1.0121



(a) dice score

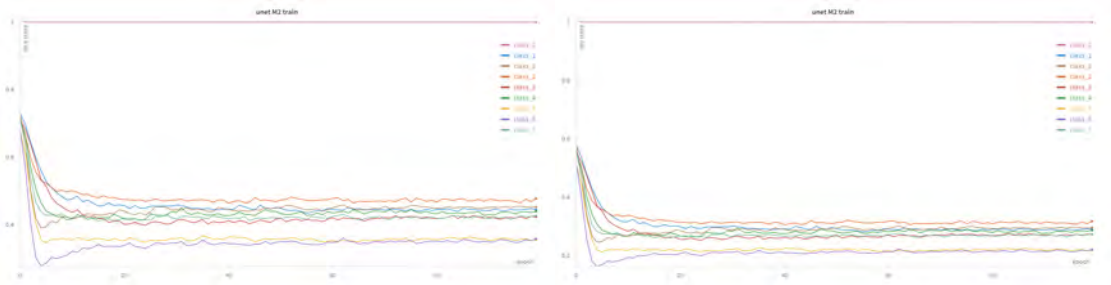
(b) iou score

Figure 4.7.: Training performance metrics for the M1 model over 100 epochs showing the progression of Figure 4.7a Dice Similarity Coefficient and Figure 4.7b Intersection over Union scores for each anatomical region.



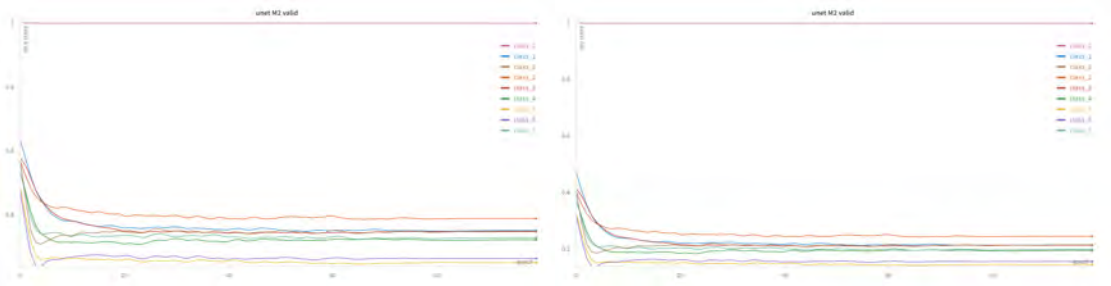
(a) dice score (b) iou score

Figure 4.8.: Validation performance metrics for the M1 model over 100 epochs showing the progression of Figure 4.8a Dice Similarity Coefficient and Figure 4.8b Intersection over Union scores for each anatomical region.



(a) dice score (b) iou score

Figure 4.9.: Training performance metrics for the M2 model over 100 epochs showing the progression of Figure 4.9a Dice Similarity Coefficient and Figure 4.9b Intersection over Union scores for each anatomical region.



(a) dice score (b) iou score

Figure 4.10.: Validation performance metrics for the M2 model over 100 epochs showing the progression of Figure 4.10a Dice Similarity Coefficient and Figure 4.10b Intersection over Union scores for each anatomical region.

The performance metrics for models M1 and M2 are presented in Tables Table 4.4 and Table 4.5. Model M1, focusing solely on amplification parameter optimization, achieved mean DSC scores of 0.754 (training) and 0.654 (validation), maintaining performance levels comparable to our best previous results.

The learned amplification parameters, detailed in Table 4.6, show minimal deviation from their initial values (range: 1.031-1.048 for M1), suggesting limited impact from pure amplification optimization. This observation is reflected in the training progression plots (Figure 4.7 and Figure 4.8), which show no metric improvements over the training period.

Model M2's attempt to jointly optimize amplification parameters and the segmentation model resulted in performance degradation, with mean DSC scores dropping to 0.486 (training) and 0.400 (validation). The training dynamics, visualized in Figure 4.9 and Figure 4.10, show an initial sharp decline in performance followed by stabilization at lower metric values.

These results suggest that while the SegReg architecture presents an innovative approach to combining segmentation and localization information, the current implementation's impact on segmentation performance is limited. Future work is needed to explore alternative integration strategies or more sophisticated amplification mechanisms.

4.5. Model Performance Comparison

The evaluation compares performance across different model configurations using the BAMBI test set. For each configuration, scatter plots are generated comparing predicted versus actual volumes for each anatomical region, with an x=y line indicating perfect prediction fit. The plots include key metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), R-squared (R^2), and count of zero predictions.

4.5.1. baseline:

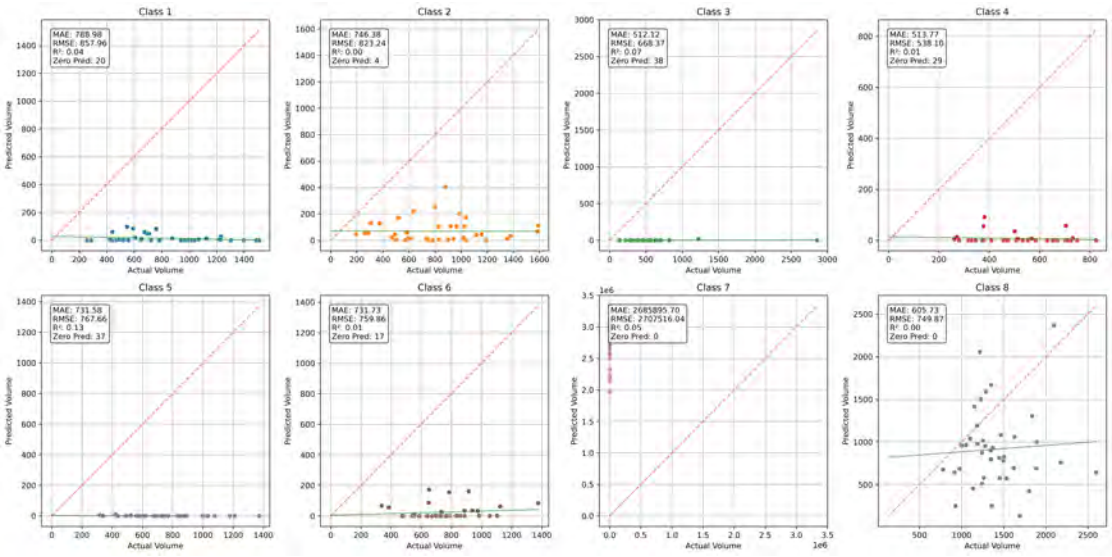


Figure 4.11.: Scatter plot comparing predicted vs. actual volumes for the baseline U-Net model.

The baseline U-Net, as shown in Figure 4.11, demonstrated significant shortcomings. Performance was poor across all regions, with high MAE values (512–788 mm^3) and minimal R^2 scores (<0.13). Zero predictions were notably frequent, particularly in the left lateral and medial perirhinal cortex (38 and 37 instances, respectively). While training performance was reasonable, the model generalized poorly to the test set. Severe undersegmentation affected most classes (1–6), often yielding zero predictions. Class 7 predictions were erratic and out of bounds, while class 8, although better, still exhibited undersegmentation with suboptimal metrics and low R^2 scores.

4.5.2. Loss Manipulation Approaches

L1 Dice and Cross-Entropy Loss Combination

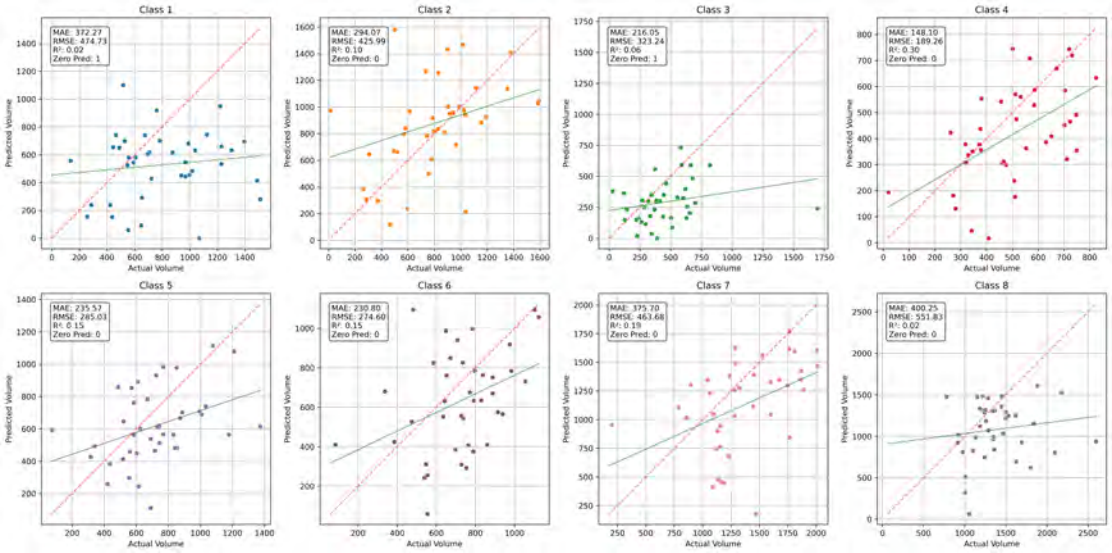


Figure 4.12.: Scatter plot of predicted vs. actual volumes for the L1 configuration.

The L1 configuration (Figure 4.12) demonstrated substantial improvements over the baseline. MAE values were significantly reduced (148–400 mm^3), and moderate improvement in R^2 scores was observed (up to 0.30 for R-lPRC). The most noteworthy enhancement was the near elimination of zero predictions, with only two remaining (one in L-ERC and one in L-lPRC). Furthermore, predictions for most classes aligned closely with the identity line, with class 4 achieving robust results. These findings indicate that the combined Dice and cross-entropy loss effectively addressed critical issues in the baseline model.

L2 Focal Loss

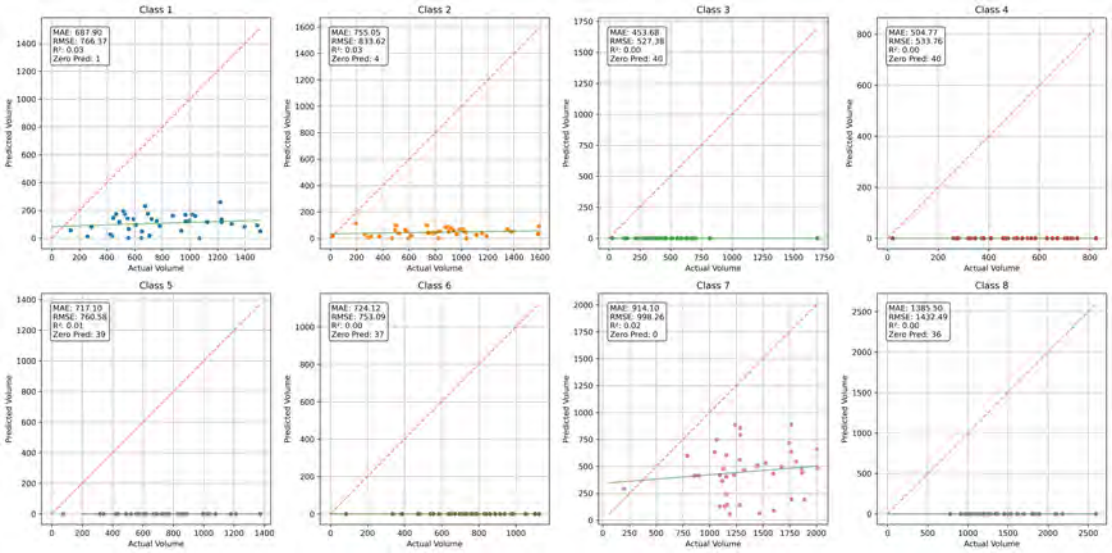


Figure 4.13.: Scatter plot of predicted vs. actual volumes for the L2 configuration.

The use of focal loss (Figure 4.13) resulted in significant performance degradation. The MAE values were substantially higher (453–1385 mm³), and R^2 scores were extremely low. Numerous zero predictions were reintroduced, particularly in medial regions (37–40 instances). Most classes (3, 4, 5, 6, and 8) suffered from severe undersegmentation, while class 7, though the best-performing, also exhibited underprediction.

L3 Distance-Transform Enhanced Focal Loss

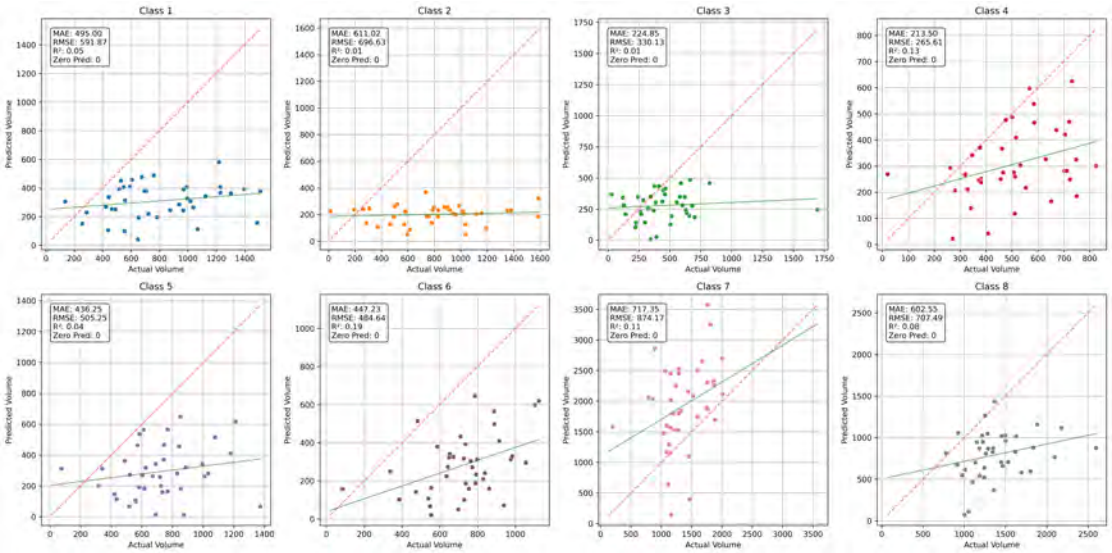


Figure 4.14.: Scatter plot of predicted vs. actual volumes for the L3 configuration.

The L3 configuration (Figure 4.14), which augmented focal loss with distance-transform weighting, showed modest improvements over L2. Zero predictions were entirely eliminated, and MAE values were reduced (213–717 mm^3). However, R^2 values remained weak (<0.19), reflecting limited predictive capability. While predictions were present for all classes, most except class 7 were still severely under segmented, as evidenced by the scarcity of data points above the identity line.

L4 Dice and Focal Loss Combination

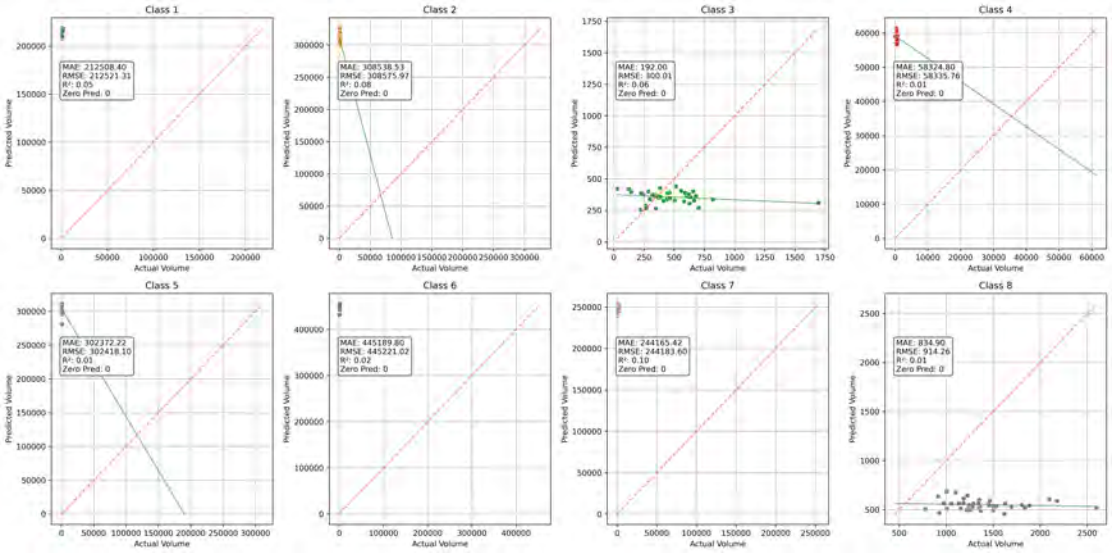


Figure 4.15.: Scatter plot of predicted vs. actual volumes for the L4 configuration.

The L4 configuration (Figure 4.15) performed poorly, producing highly unrealistic volume predictions. MAE values exceeded 200,000 mm^3 for some regions, reflecting fundamental issues with the combined Dice-Focal loss approach. Although zero predictions were avoided, most classes were wildly over segmented, undermining the model's practical utility.

4.5.3. Segmentation with Regression Approach

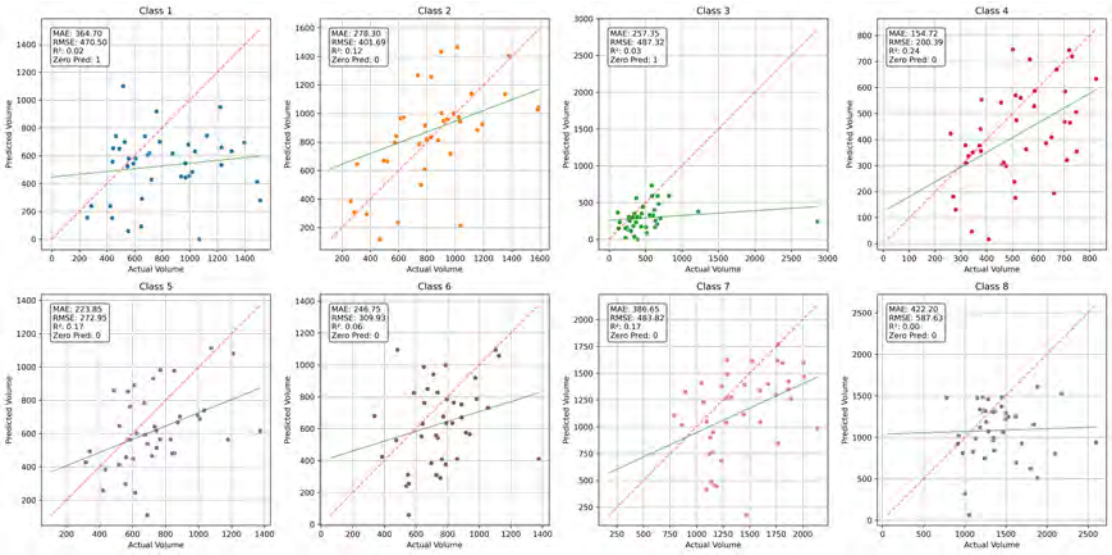


Figure 4.16.: Scatter plot of predicted vs. actual volumes for the M1 model.

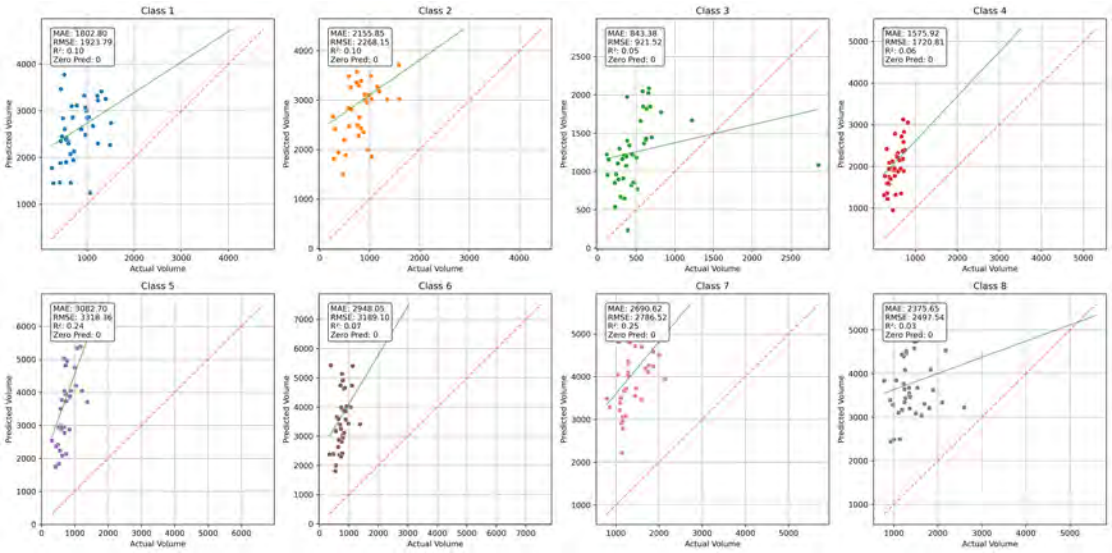


Figure 4.17.: Scatter plot of predicted vs. actual volumes for the M2 model.

The SegReg approach (Figure 4.16 Figure 4.17) yielded mixed results. The M1 configuration (amplification only) achieved performance comparable to L1, with similar MAE values and negligible zero predictions. Conversely, the M2 configuration (joint optimization) degraded performance significantly, yielding several orders of magnitude

higher MAE values. Notably, M2 exhibited pervasive over-segmentation, as nearly all predictions exceeded the identity line, highlighting instability in the joint learning mechanism.

4.6. Key Findings

The evaluation highlights critical insights into different model configurations' performance, limitations, and implications. This section synthesizes the findings to understand the factors influencing model effectiveness comprehensively.

- **Loss Function Impact:** The L1 configuration emerged as the most effective approach, successfully balancing prediction accuracy with model stability. Focal loss variants (L2, L3) showed limited success despite theoretical advantages in handling class imbalance. Combined loss approaches require careful calibration, as evidenced by L4's instability.
- **Architectural Innovations:** The SegReg approach demonstrated that complex architectural modifications might not necessarily improve performance compared to well-tuned loss functions. Amplification parameters showed minimal deviation from initial values (range: 1.031-1.048), suggesting limited impact. Joint optimization led to systematic over-segmentation, indicating potential instability in the combined learning process.
- **Clinical Implications:** Successful reduction of zero predictions in L1 and M1 configurations enhances clinical reliability. Improved R^2 scores, while still moderate, represent meaningful progress toward automated volume estimation. Consistent performance across bilateral structures suggests robust anatomical learning.
- **Technical Insights:** Simple loss function combinations outperformed complex architectural modifications. Distance transform enhancement showed promise in eliminating zero predictions. Model stability emerged as a critical factor in configuration selection.

5. Evaluation

The evaluation phase assesses the developed AI model's performance against the success criteria outlined in the project charta. This chapter synthesizes findings related to model functionality, accuracy, and alignment with project goals. Key success factors include robust segmentation performance, handling class imbalance, and clinical applicability.

5.1. Success Criteria Overview

As defined in the project charta, the key success criteria are:

- **Accurate Segmentation**: The AI model should achieve reliable segmentation of all PHG subregions.
- **Clinical Usability**: Minimizing zero predictions and achieving consistent performance for bilateral anatomical regions.
- **Efficiency**: Delivering a lean, deployable solution for research use, avoiding excessive resource consumption.

5.2. Model Performance Evaluation

5.2.1. Quantitative Metrics

The **baseline model**, built on a 3D U-Net architecture, achieved moderate performance overall, with a mean validation DSC of 0.562. However, it failed to segment the Left Parahippocampal Cortex, indicating a critical shortcoming in handling certain regions. This highlighted the need for improved methods.

In efforts to improve segmentation, the **L1 Dice-Cross Entropy Loss Combination** proved the most effective. It raised the mean DSC to 0.654 on the validation set and successfully resolved the L-PHC segmentation failure, achieving a DSC of 0.656 for this challenging region.

Alternative loss functions were also explored. **Focal Loss (L2)** and **Distance-Transform Enhanced Focal Loss (L3)** showed mixed results. While some regions

recorded improved segmentation metrics, these approaches failed to generalize consistently across all regions. The **Dice-Focal Loss Combination (L4)** performed poorly, with substantial instability caused by an imbalance between the loss components.

Finally, the novel **Segmentation with Regression (SegReg)** architecture introduced two configurations. The **M1 (amplification-only)** configuration performed comparably to the L1 model, achieving a mean DSC of 0.654 on the validation set. Conversely, the **M2 (joint optimization)** configuration suffered severe performance degradation due to over-segmentation, revealing the instability of the joint-learning mechanism.

5.2.2. Anatomical Region Performance

Performance varied across anatomical regions, with significant improvements in key areas. Notably, the L1 configuration addressed the baseline model's failure in segmenting the Left Parahippocampal Cortex. This marked a substantial improvement, demonstrating the effectiveness of the L1 approach in handling challenging regions.

The model effectively captures anatomical symmetry, as metrics for bilateral structures consistently aligned. This reflects the model's ability to generalize across corresponding left and right regions, which is critical for robust segmentation.

Class imbalance issues posed challenges during baseline testing and were significantly mitigated using the L1 configuration. Zero predictions were notably reduced, and smaller regions benefited from more balanced segmentation accuracy across all classes, improving overall model reliability.

5.3. Alignment with Success Criteria

The project outcomes are evaluated against the success criteria in the project charta. The table below summarizes that the final models achieved notable success in meeting these criteria.

Success Criterion	Evaluation
Accurate Segmentation	Achieved with L1 and M1 configurations, which provided mean DSC scores above 0.65 and resolved segmentation failures.
Clinical Usability	L1 and M1 models demonstrated consistent performance and reduced failure cases, increasing clinical reliability.
Efficiency	The final models were computationally lightweight, balancing accuracy and resource efficiency.

This alignment demonstrates that the project successfully met its goals, addressed key challenges, and paved the way for further development and deployment.

5.4. Limitations

Despite notable achievements, the current implementation has limitations:

- Performance Gaps in Certain Regions: While L1 and M1 configurations outperformed the baseline, metrics for some regions (e.g., medial perirhinal cortex) remain suboptimal.
- Complexity of Joint Optimization: The SegReg M2 approach highlighted challenges in achieving stable joint optimization, leading to over-segmentation.

6. Deployment

The deployment phase of the project encompasses the design and implementation of a software framework to enable the practical application of the AI model. The deployment integrates a Python-based Command-Line Interface (CLI) and an Application Programming Interface (API) to ensure usability across a wide range of research environments. By supporting both local and distributed computing setups, the framework is designed to facilitate efficient, flexible, and reproducible workflows for end users. This chapter provides a detailed overview of the deployment architecture, implementation specifics, and current limitations.

6.1. Architecture

The deployment framework prioritizes accessibility and platform independence by packaging the CLI and API as cross-platform Python wheels. Each wheel can be installed on operating systems such as Windows, macOS, and Linux using standard Python package management tools. This approach ensures that researchers can integrate the software into their existing computational infrastructure without requiring extensive configuration. The wheels are available for download on GitHub.

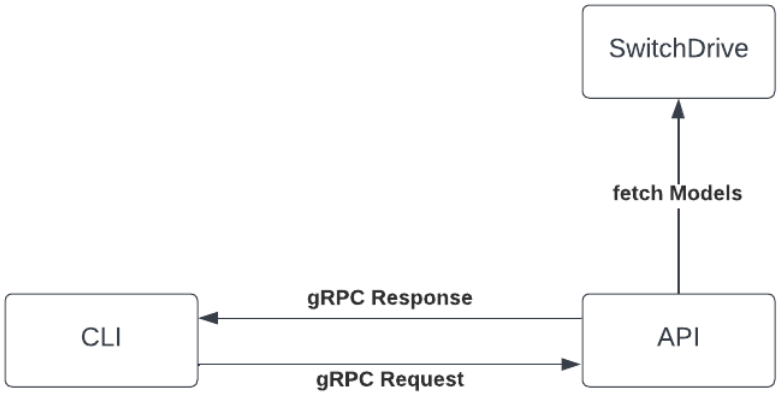


Figure 6.1.: Application Architecture

A high-level architectural diagram illustrating the interaction between the CLI, API, and external resources is provided in Figure 6.1. The CLI and API are designed to operate independently, with communication facilitated over the network using the gRPC protocol. Model files are stored on a public SwitchDrive folder, which keeps the size of the API's local footprint manageable while decoupling model updates from API versioning. Researchers can therefore update models without requiring reinstallation of the API.

To streamline deployment and enhance usability, Docker images have been created for both the CLI and API. These images, available on DockerHub, ensure execution in isolated and consistent environments, mitigating platform-specific issues. Additionally, a `docker-compose` configuration file is provided on GitHub to enable multi-container setups. This configuration facilitates horizontal scaling and improves fault tolerance, particularly in scenarios involving high resource demands or potential application crashes.

The API employs a First-Come, First-Serve (FCFS) worker queue, managed using gRPC's native threading capabilities. This mechanism ensures that while the server accepts all incoming requests, tasks exceeding the available queue capacity are deferred until workers become available. For resource efficiency, the API dynamically checks for the presence of CUDA-enabled GPUs. If a GPU is available, models are loaded onto the GPU to accelerate inference; otherwise, the API defaults to CPU execution.

6.2. API Implementation

The API forms the core of the deployment framework, enabling model inference through an efficient interface.

6.2.1. API Workflow

- **Model Management**: Upon initialization, the API fetches the pre-trained models from our SwitchDrive directory, which ensures that updates to models can be applied dynamically. The CLI also provides a command to trigger model updates during runtime, offering users control over model management without restarting the API.
- **Input Validation**: The API accepts 3D volumes with dimensions `256 x 256 x 256` as inputs, ensuring compatibility with the model's architecture. Custom validation checks are implemented to verify input input integrity and dimensionality.
- **Output Options**: By default, the API returns sparse segmentation masks, in which each voxel is assigned a discrete class label. Optionally, the API can generates probability masks, which provide the likelihood of each voxel belonging to a specific anatomical region.

6.2.2. API Configuration

The API configuration is managed through environment variables, offering users flexibility to adapt the deployment to their computational environment:

- **Port Configuration:** The API port can be customized to fit network requirements.
- **Worker Pool Size:** The number of concurrent workers may be adjusted based on hardware capabilities.

6.2.3. Limitations

The API currently lacks resource-awareness mechanisms to monitor or limit memory and computational usage. This limitation can cause crashes in environments where hardware resources are insufficient. In Dockerized environments, this issue is partially mitigated through the use of Docker Compose, which can automatically restart failed containers.

Security measures, such as API key authentication or SSL/TLS encryption, are currently not implemented. Consequently, the API is most suitable for deployment in trusted or restricted environments.

6.3. CLI Implementation

The CLI provides an interactive interface for researchers to manage their workflows and interact with the API. Built using Python's `typer` library, the CLI features a terminal-based user interface (TUI) with a focus on clarity and user-friendliness.

6.3.1. Commands and Options

The CLI supports the following commands:

- `models`:
  - `--list`: Lists the models available on the API server.
  - `--update`: Informs the API server to fetch the latest model versions from SwitchDrive.
- `segment`:
  - `--file` (`-f`, required): Specifies the path to the input NIfTI file for segmentation.
  - `--in-dir` (`-d`): Specifies a directory containing NIfTI files for batch processing.

– `--out-dir` (`-o`): Specifies the directory where output files will be saved.
– `--model` (`-m`, required): Specifies the model to be used for segmentation.
– `--probs` (`-p`): Toggles whether to save class probability masks or sparse segmentation masks.

6.3.2. User Interface Features

(a) CLI TUI (b) CLI Progress

Figure 6.2.: terminal UI screenshots showcasing the CLI interface with the `segment` command Figure 6.2a and segmentation progress tracking for global and per-file progress Figure 6.2b.

The CLI provides an interface for running segmentation tasks, as shown in Figure 6.2. While processing inputs, the CLI outputs real-time progress feedback. The global progress bar tracks the overall progress of the number of files, and individual progress bars display the status of each file being processed.

6.3.3. Limitations

The CLI currently lacks robust error handling for cases where the API is unavailable (e.g., network issues or server downtime). If the API server is unreachable, the CLI will time out without retry logic or fallback mechanisms.

7. Conclusion

This study aimed to develop a robust framework for automated segmentation of PHG subregions in MRI scans, addressing challenges like class imbalance, segmentation failures, and deployment usability. This chapter summarizes the key findings, highlights the project's contributions, and outlines directions for future research.

7.1. Summary of Findings

This project successfully developed and evaluated a deep learning-based framework for segmenting PHG subregions in MRI scans. By addressing the challenges posed by extreme class imbalance, anatomically intricate structures, and segmentation failures, the proposed solutions showed marked improvements in accuracy. The **L1 Dice-Cross Entropy Loss Combination** emerged as the most effective approach with a mean DSC of 0.654 on the validation set, resolving key limitations such as the failure to segment the L-PHC.

The evaluation demonstrated that the models performed consistently across bilateral anatomical regions, capturing symmetry and effectively reducing zero predictions. The lightweight and efficient deployment framework also ensured accessibility and potential for integration into neuroimaging workflows. These outcomes align with the success criteria, achieving accurate segmentation, clinical usability, and computational efficiency.

7.2. Key Contributions

1. **Improved Segmentation Performance**: The L1 configuration significantly enhanced segmentation accuracy, addressing critical failures in baseline models and mitigating the effects of class imbalance.
2. **Framework for Deployment**: A comprehensive deployment solution, including a Python-based API and CLI, was developed to support researchers in leveraging the model effectively in practical settings. Tools like Docker and gRPC were used to ensure platform independence and usability.
3. **Clinical Relevance**: By providing consistent segmentation results and eliminating zero predictions, the proposed solution sets the groundwork for potential clinical applications in Alzheimer's disease research and diagnosis.

7.3. Future Works

This section outlines directions for future research to address the limitations identified in this study and further enhance the proposed methods' performance and clinical applicability.

- **Enhancing Loss Function Design:** Future studies could explore the development of region-specific loss functions that account for heterogeneity across anatomical regions by assigning higher weights to more challenging regions. Additionally, adaptive multi-loss frameworks, such as dynamic weighting strategies or the integration of class-balanced Dice loss, may address the class imbalance and improve predictive performance [14], [15].
- **Refining Segmentation with Regression (SegReg):** The integration of regression as a complementary task to segmentation demonstrated potential but requires further refinement. Future research should investigate more advanced methods for amplifying segmentation logits within predicted bounding boxes, optimize joint-training strategies to enhance stability and explore lightweight architectures with attention mechanisms to balance efficiency and accuracy [16], [17].
- **Performing Conversion from Volume to Cortical Thickness:** To enhance the clinical utility of segmentation outputs, future efforts should develop a computational pipeline for translating voxel-wise segmentation volumes into cortical thickness measurements. These measurements would enable direct comparability with clinical biomarkers and studies, particularly in neurodegenerative disease research.
- **Optimizing Deployment:** Addressing resource management, security, and error handling challenges will be critical for achieving production-level robustness and broader applicability.

7.4. Final Remarks

This project demonstrates the feasibility of applying deep learning to automate PHG subregion segmentation while addressing key challenges in medical image segmentation. The results contribute valuable insights into improving segmentation accuracy and deploying clinically relevant tools. Although further development is necessary to refine these methods and expand their applicability, the outcomes of this study show promise for advancing AD research and diagnostics.

8. Bibliography

1. Greenblat C. Dementia [Internet]. 2023. Available from: https://www.who.int/news-room/fact-sheets/detail/dementia

2. Jack CR, Andrews JS, Beach TG, Buracchio T, Dunn B, Graf A, et al. Revised criteria for diagnosis and staging of alzheimer's disease: Alzheimer's association workgroup. Alzheimer's & Dementia. 2024 Aug;20:5143–69.

3. Krumm S, Kivisaari SL, Probst A, Monsch AU, Reinhardt J, Ulmer S, et al. Cortical thinning of parahippocampal subregions in very early alzheimer's disease. Neurobiology of Aging. 2016 Feb;38:188–96.

4. Kim YJ, Cho SK, Kim HJ, Lee JS, Lee J, Jang YK, et al. Data-driven prognostic features of cognitive trajectories in patients with amnestic mild cognitive impairments. Alzheimer's Research & Therapy. 2019 Dec;11:10.

5. Doemer M, Kempf D. Is it ops that make data science scientific? Archives of Data Science, Series A (Online First) [Internet]. 2022;8:12. Available from: https://publikationen.bibliothek.kit.edu/1000150238

6. Eelbode T, Bertels J, Berman M, Vandermeulen D, Maes F, Bisschops R, et al. Optimization for medical image segmentation: Theory and practice when evaluating with dice score or jaccard index. IEEE Transactions on Medical Imaging [Internet]. 2020 Nov;39(11):3679–90. Available from: http://dx.doi.org/10.1109/TMI.2020.3002417

7. Vlăsceanu GV, Tarbă N, Voncilă ML, Boiangiu CA. Selecting the right metric: A detailed study on image segmentation evaluation. BRAIN Broad Research in Artificial Intelligence and Neuroscience [Internet]. 2024; Available from: https://api.semanticscholar.org/CorpusID:274490895

8. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D u-net: Learning dense volumetric segmentation from sparse annotation. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, editors. Medical image computing and computer-assisted intervention – MICCAI 2016. Cham: Springer International Publishing; 2016. p. 424–32.

9. Cardoso MJ, Li W, Brown R, Ma N, Kerfoot E, Wang Y, et al. MONAI: An open-source framework for deep learning in healthcare [Internet]. 2022. Available from: https://arxiv.org/abs/2211.02701

10. Milletari F, Navab N, Ahmadi SA. V-net: Fully convolutional neural networks for volumetric medical image segmentation [Internet]. 2016. Available from: https://arxiv.org/abs/1606.04797

11. Falcon W, Borovec J, Wälchli A, Eggert N, Schock J, Jordan J, et al. PyTorchLightning/pytorch-lightning: 0.7.6 release [Internet]. Zenodo; 2020. Available from: https://doi.org/10.5281/zenodo.3828935

12. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection [Internet]. 2018. Available from: https://arxiv.org/abs/1708.02002

13. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods. 2020;17:261–72.

14. Chen Y, Yu L, Wang JY, Panjwani N, Obeid JP, Liu W, et al. Adaptive region-specific loss for improved medical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence [Internet]. 2023 Nov;45(11):13408–21. Available from: http://dx.doi.org/10.1109/TPAMI.2023.3289667

15. Golnari A, Diba M. Adaptive real-time multi-loss function optimization using dynamic memory fusion framework: A case study on breast cancer segmentation [Internet]. 2024. Available from: https://arxiv.org/abs/2410.19745

16. Wang X, Yu J, Zhang B, Huang X, Shen X, Xia M. LightAWNet: Lightweight adaptive weighting network based on dynamic convolutions for medical image segmentation. Journal of Applied Clinical Medical Physics [Internet]. 2024 Dec; Available from: http://dx.doi.org/10.1002/acm2.14584

17. Lu J, Chen J, Cai L, Jiang S, Zhang Y. H2ASeg: Hierarchical adaptive interaction and weighting network for tumor segmentation in PET/CT images [Internet]. 2024. Available from: https://arxiv.org/abs/2403.18339

18. Iaccarino L, Burnham SC, Dell'Agnello G, Dowsett SA, Epelbaum S. Diagnostic biomarkers of amyloid and tau pathology in alzheimer's disease: An overview of tests for clinical practice in the united states and europe. The Journal Of Prevention of Alzheimer's Disease. 2023;

19. Varesi A, Carrara A, Pires VG, Floris V, Pierella E, Savioli G, et al. Blood-based biomarkers for alzheimer's disease diagnosis and progression: An overview. Cells. 2022 Apr;11:1367.

20. Long JM, Coble DW, Xiong C, Schindler SE, Perrin RJ, Gordon BA, et al. Preclinical alzheimer's disease biomarkers accurately predict cognitive and neuropathological outcomes. Brain. 2022 Dec;145:4506–18.

21. Brickman AM, Manly JJ, Honig LS, Sanchez D, Reyes-Dumeyer D, Lantigua RA, et al. Plasma p-tau181, p-tau217, and other blood-based alzheimer's disease biomarkers in a multi-ethnic, community study. Alzheimer's & Dementia. 2021 Aug;17:1353–64.

22. Kim KY, Shin KY, Chang KA. GFAP as a potential biomarker for alzheimer's disease: A systematic review and meta-analysis. Cells. 2023 May;12:1309.

A. Appendix - Literature Review

A.1. Alzheimer's Disease and Dementia

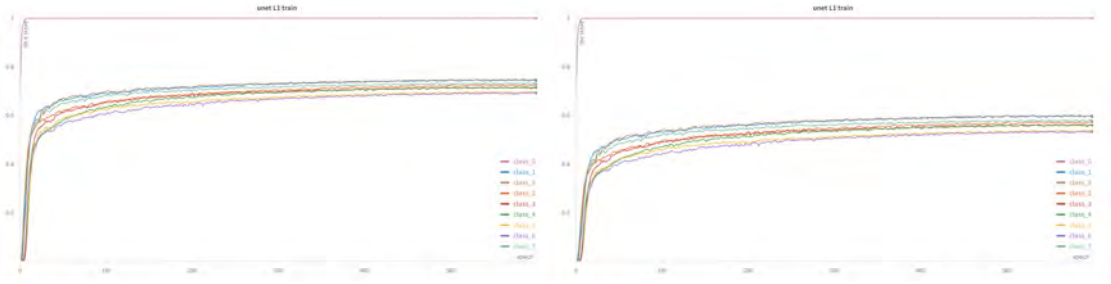
Dementia is a broad term covering various diseases impacting cognitive abilities, such as memory and thinking. According to the World Health Organization (WHO), symptoms of dementia include [1]:

- Forgetting recent events or information
- Losing or misplacing items
- Getting lost while walking or driving
- Experiencing confusion, even in familiar environments
- Losing track of time
- Having difficulty solving problems or making decisions
- Struggling to follow conversations or find the right words
- Facing challenges in performing familiar tasks
- Misjudging distances visually

While Amyloid plaques and Neurofibrillary tangles are hallmarks of AD [2], recent research has identified various biomarkers associated with the disease, including:

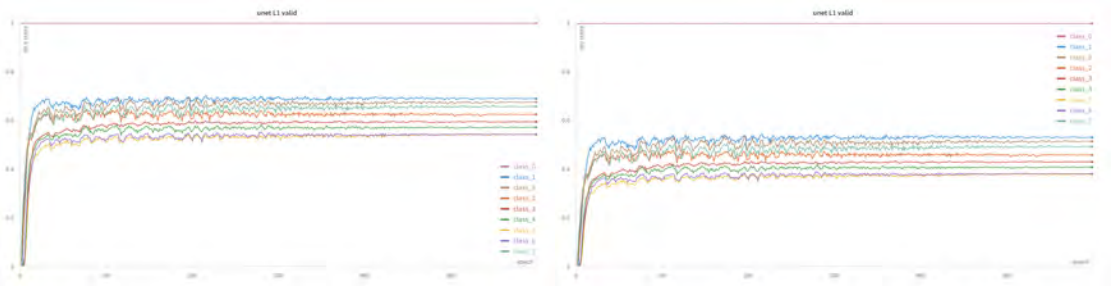
- **Amyloid- (A) peptides**: A 42 and A 40 in cerebrospinal fluid (CSF) and plasma are critical biomarkers for amyloid plaques [18][19][20].
- **Phosphorylated tau (p-tau) protein**: p-tau181 and p-tau217 in CSF and plasma are biomarkers for neurofibrillary tangles [21][18][19][20].
- **Total tau (t-tau) protein**: Elevated levels of t-tau in CSF are associated with AD [18][19].
- **Neurofilament light chain (NfL)**: NfL in CSF and blood is a marker of axonal damage [19].
- **Neurogranin**: Elevated levels of neurogranin in CSF are associated with synaptic dysfunction in AD [18]
- **YKL-40 (CHI3L1)**: YKL-40 is a marker of neuroinflammation in AD [18].
- **Glial fibrillary acidic protein (GFAP)**: GFAP in the blood is a potential biomarker for astrocytic activation in AD [22].
- **Lipid biomarkers**: Changes in lipid metabolism, such as sphingolipid and cholesterol metabolism alterations, are potential biomarkers for AD [19].

B. Appendix - Modelling Report



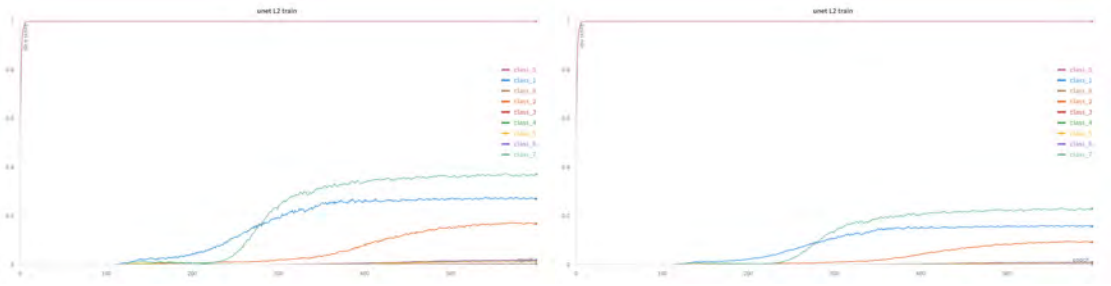
(a) dice score (b) iou score

Figure B.1.: Training performance metrics for L1 model over 600 epochs showing the progression of Figure B.1a Dice Similarity Coefficient and Figure B.1b Intersection over Union scores for each anatomical region.



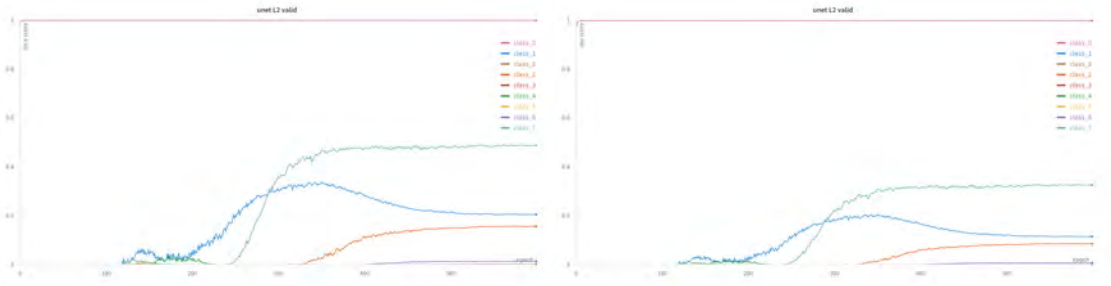
(a) dice score (b) iou score

Figure B.2.: Validation performance metrics for L1 model over 600 epochs showing the progression of Figure B.2a Dice Similarity Coefficient and Figure B.2b Intersection over Union scores for each anatomical region.



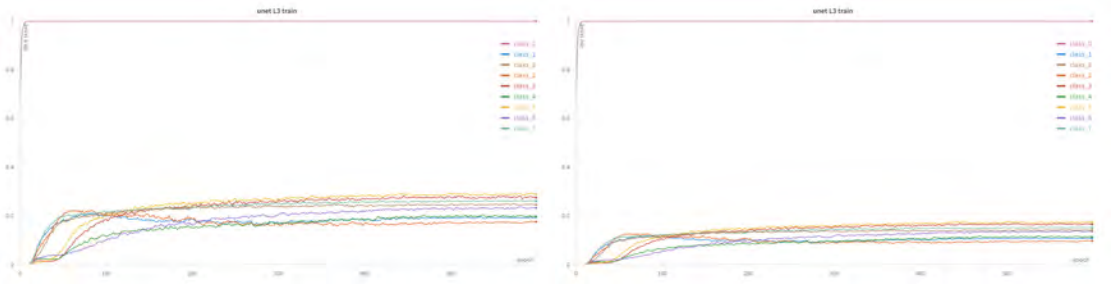
(a) dice score (b) iou score

Figure B.3.: Training performance metrics for L2 model over 600 epochs showing the progression of Figure B.3a Dice Similarity Coefficient and Figure B.3b Intersection over Union scores for each anatomical region.



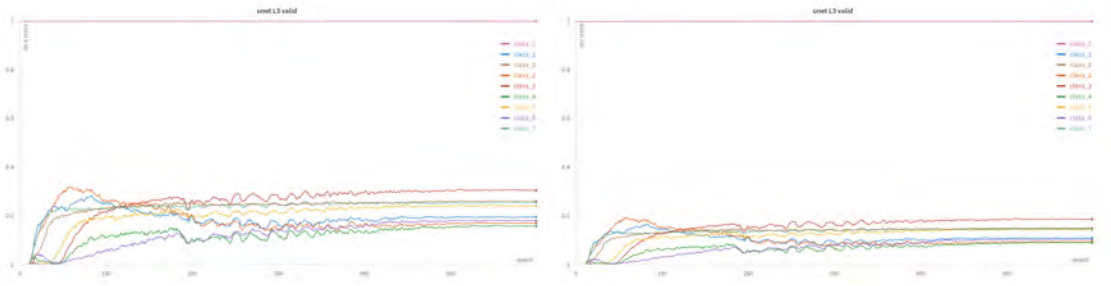
(a) dice score (b) iou score

Figure B.4.: Validation performance metrics for L2 model over 600 epochs showing the progression of Figure B.4a Dice Similarity Coefficient and Figure B.4b Intersection over Union scores for each anatomical region.



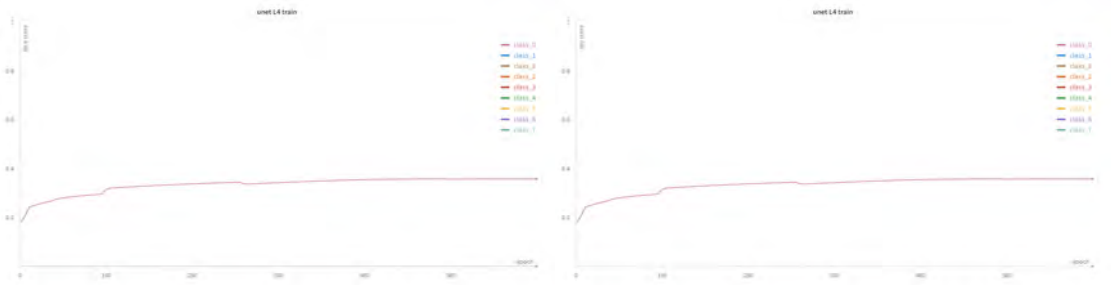
(a) dice score (b) iou score

Figure B.5.: Training performance metrics for L3 model over 600 epochs showing the progression of Figure B.5a Dice Similarity Coefficient and Figure B.5b Intersection over Union scores for each anatomical region.



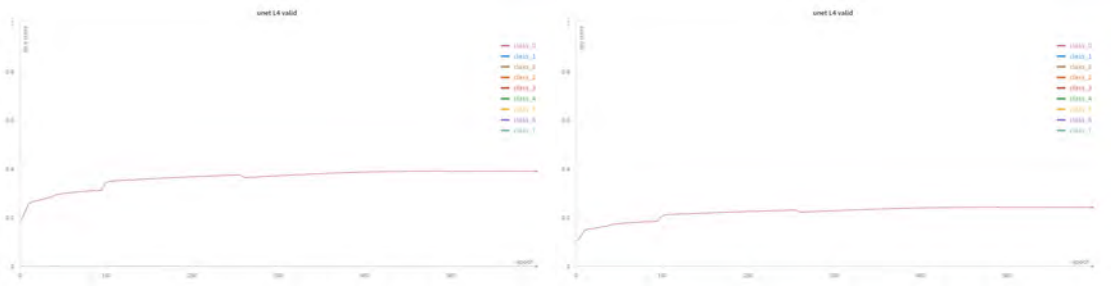
(a) dice score (b) iou score

Figure B.6.: Validation performance metrics for L3 model over 600 epochs showing the progression of Figure B.6a Dice Similarity Coefficient and Figure B.6b Intersection over Union scores for each anatomical region.



(a) dice score (b) iou score

Figure B.7.: Training performance metrics for L4 model over 600 epochs showing the progression of Figure B.7a Dice Similarity Coefficient and Figure B.7b Intersection over Union scores for each anatomical region.



(a) dice score (b) iou score

Figure B.8.: Validation performance metrics for L4 model over 600 epochs showing the progression of Figure B.8a Dice Similarity Coefficient and Figure B.8b Intersection over Union scores for each anatomical region.