



POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Automatic 3D Segmentation of Atypical Lipomatous Tumor in Limb MRI using nnU-Net

TESI DI LAUREA MAGISTRALE IN
HIGH PERFORMANCE COMPUTING

Luis Felipe Epia Realpe, 11008027

Advisor:
Prof. Valentina Corino

Co-advisor:
Meri Ferretti

Academic year:
2025-2026

Abstract: Atypical Lipomatous Tumor (ALT) is the most common soft tissue sarcoma of the extremities. Accurate three-dimensional delineation of the tumour on MRI is essential for surgical treatment, yet manual segmentation requires up to four hours per case and is subject to inter-operator variability. No validated fully automatic 3D segmentation pipeline for ALT has previously been reported.

This thesis presents a segmentation framework using nnU-Net v2 and evaluated through a five-fold cross-validation. We used a private dataset composed by 46 patients with T1- and T2-weighted MRI volumes. Three custom components adapt the framework to this task: a Gibbs ringing augmentation trainer that simulates k-space truncation artefacts during training to improve robustness to scanner variability; a custom nnU-Net trainer family with per-sample Dice loss and a training schedule calibrated to the small cohort. In addition to models employing T1-only and T2-only images, a dual-channel T1+T2 model that enables joint learning of cross-modal features was developed as well. Thus Four configurations have been evaluated: T1-only (mean Dice 0.731), T2-only (0.711), post-hoc union ensemble (0.768), and the dual-channel model, which achieves a mean Dice of 0.823 and a median of 0.913 on the patients whose T1 and T2 volumes share a compatible acquisition orientation ($n = 40$), significantly outperforming all alternatives ($p = 0.011$, Wilcoxon Signed-Rank test). Ablation experiments confirmed that Gibbs augmentation consistently improves over equivalent configurations without it. The results demonstrate that joint dual-channel training is the key driver of performance and that 3D nnU-Net with Gibbs augmentation produces clinically useful ALT segmentations for pre-surgical planning.

Key-words: medical image segmentation, nnU-Net, atypical lipomatous tumor, MRI, data augmentation, Gibbs ringing

1. Introduction

1.1. Clinical Context

Atypical Lipomatous Tumor (ALT) is the most common soft tissue sarcoma of the extremities. Wide surgical excision with negative margins is the primary treatment objective, and accurate three-dimensional delineation

of the tumour is a prerequisite for planning the resection. This delineation is currently performed manually by expert radiologists on MRI, a process that is time-consuming — up to four hours per large-volume case — and subject to inter-operator variability [5, 8]. No validated fully automatic 3D segmentation pipeline for ALT exists; the few published approaches address broader lipomatous tumour categories or operate in 2D [8], falling short of the full-resolution volumetric setting required for pre-surgical planning. A detailed account of ALT pathology, MRI characteristics, and related segmentation literature is given in Section 2.

1.2. The Segmentation Challenge

Automatic 3D segmentation of ALT in limb MRI presents a set of compounding difficulties that distinguish it from better-studied segmentation tasks.

Dataset scarcity. Collecting and annotating a large cohort of ALT cases requires specialist expertise and significant radiologist time. The dataset available for this thesis comprises 46 patients, a size that is realistic for a rare tumour type at a single institution but imposes strong constraints on the complexity of the model that can be reliably trained and evaluated.

Tumour heterogeneity. ALT tumours vary enormously in size — from a few cubic centimetres to several hundred cubic centimetres — and in shape, often growing around and between muscles and neurovascular bundles rather than forming compact, well-defined masses. Their MRI appearance varies with internal composition: purely lipomatous regions differ markedly from areas of fibrous or myxoid stroma [2, 5].

MRI acquisition variability. Limb MRI is acquired in a variety of orientations and with heterogeneous acquisition parameters across patients. Gibbs ringing, an artifact arising from k-space truncation at high-contrast fat-muscle interfaces, is particularly common in limb imaging and can alter the apparent sharpness of tumour boundaries in a scanner-dependent manner [7]. Models trained without exposure to this artifact risk exploiting boundary ringing as a localisation cue, which fails to generalise across acquisition settings.

Multimodal input. Both T1 and T2 sequences carry complementary information about ALT: T1 provides the fat-specific signal that distinguishes the tumour from muscle, while T2 reveals internal heterogeneity and peritumoral change. Exploiting this complementarity requires a principled strategy for combining the two modalities during training rather than at inference time.

1.3. Contributions

This thesis presents what is, to the best of our knowledge, the first fully automatic 3D segmentation pipeline specifically developed and evaluated for ALT in limb MRI. The pipeline is built on nnU-Net v2 [6], a self-configuring framework that automates preprocessing, architecture selection, and training configuration, and is adapted to the specific challenges of this task through three custom components.

1. **Gibbs ringing augmentation.** A custom trainer introduces random k-space truncation during data augmentation. By exposing the model to a range of synthetic Gibbs ringing intensities at training time, the augmentation prevents the model from relying on acquisition-specific boundary artifacts and improves robustness to scanner variability. Ablation experiments confirm that the Gibbs-augmented trainer outperforms equivalent configurations trained without this augmentation.
2. **Small-dataset training adaptation.** Per-sample Dice loss and an empirically calibrated training schedule adapt nnU-Net to the specific constraints of our dataset. The per-sample loss prevents the $660\times$ tumour-volume range from concentrating gradients on the largest lesions; the training schedule is set from convergence monitoring rather than the default budget calibrated for much larger datasets.
3. **Dual-channel T1+T2 pipeline.** T1 and T2 volumes acquisitions are presented to the model as two input channels, enabling joint end-to-end training on both modalities simultaneously. This allows the model to learn cross-modal features that are inaccessible to models trained on individual modalities and subsequently combined.

The best configuration — the T1T2 dual model with Gibbs augmentation trained for 350 epochs — achieves a median Dice of 0.913 (IQR 0.169) on geometrically aligned cases ($n = 40$), outperforming both the baseline models and the T1T2 ensemble model; the advantage over the T1T2 ensemble model is statistically significant ($p = 0.011$, Wilcoxon Signed-Rank test).

1.4. Thesis Structure

The remainder of this thesis is organised as follows.

Section 2 provides background on ALT pathology and MRI characteristics, the nnU-Net framework, MRI acquisition artifacts, and related work in soft tissue tumour segmentation.

Section 3 describes the patient cohort, the dataset construction procedure for each modality configuration, and the cross-validation protocol.

Section 4 details the custom components developed on top of nnU-Net: the Gibbs ringing augmentation trainer, the custom trainer family, and the dual-channel pipeline construction.

Section 5 reports quantitative results for all evaluated configurations, including per-case analysis and failure mode characterisation.

Section 6 interprets the results, discusses the absence of a held-out test set, and identifies the key limitations of the study.

Section 7 summarises the contributions and outlines directions for future work.

2. Background and Related Work

2.1. Atypical Lipomatous Tumor

Atypical Lipomatous Tumor (ALT) belongs to the family of lipomatous mesenchymal tumours and represents the most common soft tissue sarcoma of the extremities. Pathologically, it corresponds to a well-differentiated liposarcoma (WDLPS) grade 1, though the term ALT is preferred clinically when the tumour arises in surgically resectable locations such as the limbs, since the prognosis differs from deep-seated WDLPS. ALT is characterised by mature adipocytic cells with at least focal nuclear atypia and MDM2 gene amplification. Tumours frequently exceed 10 cm in their longest dimension at the time of diagnosis and can insinuate between tissue planes, making delineation challenging [4, 5].

ALT does not metastasise in its pure form, but carries a significant risk of local recurrence following incomplete resection; dedifferentiation to a higher-grade sarcoma is possible in a minority of cases and dramatically worsens prognosis. Wide negative surgical margins are therefore the primary treatment objective, requiring precise preoperative knowledge of the tumour’s three-dimensional extent and its relationship to adjacent neurovascular structures. Currently, expert radiologists perform manual delineation of ALT on MRI, a process that can take up to four hours per case and is subject to inter-operator variability [8].

MRI is the imaging modality of choice for evaluating ALT owing to its superior soft-tissue contrast [5]. On T1-weighted sequences, ALT appears as a predominantly bright mass consistent with mature fat, often traversed by fibrous septa of lower signal intensity. On T2-weighted sequences, the signal is heterogeneous: regions of myxoid change or fibrous stroma appear bright relative to the background fat, while areas of dedifferentiation may exhibit more pronounced contrast enhancement. These appearances broadly overlap with those of benign lipomas, and the distinction ultimately rests on pathological confirmation of MDM2 amplification [2, 4]. The heterogeneous internal composition and large, irregular shape make ALT a challenging target for automatic segmentation.

No standardised automated segmentation pipeline exists specifically for ALT. The combination of large and irregular tumour shape, variable internal signal composition, and the absence of fixed anatomical boundaries distinguishes ALT segmentation from organ segmentation tasks where priors on shape and location are readily available. This thesis addresses that gap by developing and evaluating a fully automatic 3D segmentation pipeline based on nnU-Net, described in the following sections.

2.2. Medical Image Segmentation

Before the advent of deep learning, medical image segmentation relied primarily on atlas-based registration and deformable model approaches. Atlas-based methods work by deforming a manually labelled template image to match a new unseen volume, transferring the template labels through the estimated deformation field. Deformable models — including level sets and active contours — represent boundaries as evolving curves or surfaces driven by image gradients and shape priors. While these approaches were grounded in well-understood mathematics, they required careful hand-crafted features tailored to the specific anatomy of interest and were sensitive to initialisation, image noise, and the considerable shape variability found across patients. Generalising such pipelines to new scanners, field strengths, or acquisition protocols typically demanded substantial re-engineering effort, limiting their practical scalability in clinical settings.

The introduction of deep convolutional neural networks (CNNs) transformed the field by replacing hand-crafted features with representations learned directly from data. Fully convolutional networks (FCNs) extended the classification CNN paradigm to dense prediction: by replacing the fully connected layers at the network head with convolutional ones, the network produces a spatial output map whose spatial dimensions match those of the input, allowing per-voxel class assignment in a single forward pass. This architectural change removed the need for explicit feature engineering and enabled end-to-end training from labelled image-mask pairs. Because

the convolutional layers share weights across all spatial positions, FCNs scale gracefully to images of varying size and are highly parameter-efficient relative to the dense predictions they produce.

U-Net [10] refined and extended the FCN concept by introducing a symmetric encoder-decoder architecture with skip connections, designed explicitly for biomedical image segmentation where labelled data is scarce. The encoder arm progressively contracts the spatial resolution through successive pooling operations, building a hierarchy of feature representations that capture contextual information at increasing scales. The decoder arm mirrors this hierarchy, recovering spatial resolution through transposed convolutions and thereby producing full-resolution output predictions. At each resolution level, skip connections transfer the corresponding encoder feature maps directly to the decoder, reintroducing fine-grained spatial detail that is inevitably compressed during the encoding pathway. This combination — global semantic context from the bottleneck representation and local structural precision from the skip-connected encoder features — gave U-Net a decisive advantage over earlier FCN variants on biomedical benchmarks and established it as the dominant architecture for medical image segmentation.

Extensions to volumetric data followed shortly after the original publication. The 3D U-Net [3] generalises all convolutional, pooling, and transposed-convolution operations from two dimensions to three, enabling direct processing of volumetric medical images such as MRI and CT scans without collapsing the depth dimension into a channel or processing axial slices independently. By treating the full volume as a three-dimensional tensor, the network can exploit inter-slice context and anisotropic structures that are invisible to 2D slice-wise approaches, which is particularly important for tumour segmentation tasks where lesion boundaries are defined in three dimensions.

Despite its success, U-Net requires task-specific tuning of architectural hyperparameters, augmentation strategies, and training schedules — choices that significantly affect performance and are typically made by domain experts. nnU-Net, described in the following section, automates this process.

2.3. nnU-Net v2

nnU-Net [6] is a self-configuring deep learning framework for biomedical image segmentation. Unlike conventional pipelines that require extensive manual hyperparameter tuning, nnU-Net reads the properties of a given dataset and automatically derives a complete, optimised training pipeline — encompassing preprocessing, network architecture, augmentation strategy, and postprocessing. The framework was developed at the German Cancer Research Center (DKFZ) and, since its introduction, has established state-of-the-art performance across a wide range of segmentation benchmarks, including the Medical Segmentation Decathlon. The second major version, nnU-Net v2, refactors the original codebase for modularity and extends support to additional modalities and multi-GPU training, while preserving the core self-configuration philosophy.

The key insight behind nnU-Net is that most design choices in medical image segmentation — patch size, network depth, feature map counts, batch size, resampling resolution, and normalisation scheme — are not universal constants but follow directly from the dataset itself. By formalising this dependency, nnU-Net replaces ad-hoc tuning with a deterministic rule set applied to a compact statistical summary of the training data, referred to as the *dataset fingerprint*.

Dataset Fingerprinting

Before any model is trained, nnU-Net analyses the raw training set to construct a dataset fingerprint. This fingerprint is a structured collection of per-dataset statistics that drives every downstream configuration decision. Concretely, the extracted quantities include: the median voxel spacing across all training volumes, per-channel intensity statistics (mean, standard deviation, and selected percentiles), the median image shape in voxels, and the total number of training cases.

The voxel spacing statistics are particularly important for volumetric medical images, where the physical resolution is encoded in the image header and varies substantially across scanners and acquisition protocols. For a dataset of N training images, each with voxel spacing $\mathbf{s}_i = (s_i^x, s_i^y, s_i^z)$, the fingerprint records the per-axis median:

$$\bar{s}^d = \text{median}_{i=1}^N s_i^d, \quad d \in \{x, y, z\}. \quad (1)$$

This median spacing $\bar{\mathbf{s}}$ later defines the target resolution for resampling, ensuring that all volumes are brought to a common physical scale before training.

Intensity statistics are collected after resampling and are stratified by channel. For MRI data, percentile-based clipping values (typically the 0.5th and 99.5th percentiles) are stored alongside the global mean and standard deviation of foreground voxels, providing the parameters required for normalisation. The median image shape after resampling informs patch size selection, since nnU-Net targets patches that are as large as possible while respecting GPU memory constraints.

Preprocessing

Given the dataset fingerprint, nnU-Net applies a deterministic preprocessing pipeline to every training and test image before any network interaction. Two decisions are central to this stage.

Resampling. All volumes are resampled to the median voxel spacing $\bar{s}$ extracted during fingerprinting. Resampling uses third-order spline interpolation for image intensities and nearest-neighbour interpolation for segmentation masks, preserving label boundaries. This step ensures that the network always operates on data with a consistent physical resolution, decoupling performance from the heterogeneity of acquisition protocols in the training set.

Intensity normalisation. The normalisation scheme is modality-dependent and is selected automatically from the fingerprint. For CT images, intensities are clipped to a fixed Hounsfield Unit window and then linearly rescaled. For MRI — which is the modality used in this thesis — nnU-Net applies per-channel z-score normalisation computed exclusively over foreground voxels:

$$\hat{x}_v = \frac{x_v - \mu_{fg}}{\sigma_{fg}}, \quad (2)$$

where x_v is the raw intensity of voxel v , and μ_{fg} and σ_{fg} are the mean and standard deviation of all foreground voxels in that channel. Using only foreground voxels excludes the large background regions common in MRI acquisitions, which would otherwise skew the statistics and reduce contrast in the region of interest.

Architecture

The network architecture employed by nnU-Net is the **PlainConvUNet** [3, 10], an encoder-decoder convolutional network in which skip connections transfer encoder feature maps to the corresponding decoder stage by concatenation. The encoder progressively reduces spatial resolution through a sequence of convolutional stages; downsampling is implemented by a strided convolution at the entry of each stage, which simultaneously halves spatial dimensions and increases the number of feature maps. The decoder mirrors this structure using transposed convolutions to recover spatial resolution, and concatenates the corresponding encoder feature maps at each level via the skip connections. These connections reintroduce the fine-grained spatial detail that is suppressed during encoding, supporting precise boundary localisation in the decoded output.

The network depth and feature-map schedule are not fixed but are automatically derived from the dataset fingerprint. nnU-Net selects the maximum depth at which the spatial dimensions of the deepest feature maps remain above a minimum threshold (typically 4^3 voxels). The complete specification — kernel sizes, strides, feature-map counts, and normalisation parameters — is stored in `nnUNetPlans.json`, making every experiment fully reproducible.

Each convolutional stage in both encoder and decoder consists of two sequential units, each comprising a convolution, instance normalisation, and a leaky ReLU activation (negative slope 0.01). Instance normalisation is preferred over batch normalisation because the small batch sizes imposed by high-resolution 3D patches render batch statistics unreliable. Deep supervision is applied during training via auxiliary segmentation heads attached to every decoder stage; only the final full-resolution head is used at inference.

For the ALT dataset, the fingerprint-driven configuration yields a 6-stage encoder and 5-stage decoder with feature-map schedule $32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 320 \rightarrow 320$ (encoder) and $320 \rightarrow 256 \rightarrow 128 \rightarrow 64 \rightarrow 32$ (decoder), totalling approximately 30.7 million trainable parameters. Because the dataset has pronounced voxel anisotropy (median through-plane spacing 5.0 mm versus in-plane 0.78 mm, a ratio of $\approx 6.4\times$), the self-configuration pipeline adapts both kernel shapes and downsampling strides. The two shallowest encoder stages use $1 \times 3 \times 3$ kernels that apply no through-plane receptive-field extension, and the first two downsampling transitions use in-plane-only strides ($1 \times 2 \times 2$); the three deeper transitions switch to isotropic $3 \times 3 \times 3$ kernels and stride $2 \times 2 \times 2$. The decoder inverts this pattern with matching transposed-convolution strides. This anisotropy-aware design preserves through-plane context in the early feature maps and avoids collapsing the sparse depth dimension before meaningful 3D representations have been extracted. Figure 1 illustrates the resulting network topology.

Training Configuration

Once the architecture is fixed, nnU-Net determines the training patch size and batch size. The patch size is chosen to be as large as possible within the available GPU memory, subject to a minimum batch size of 2. Concretely, nnU-Net starts from the median image shape in the fingerprint and progressively reduces the patch dimensions until the required memory fits within the target budget. For the 3d full-resolution configuration used in this thesis, patches are drawn from full-resolution resampled volumes, which means patch sizes can reach several hundred voxels per side for datasets with coarse native spacing.

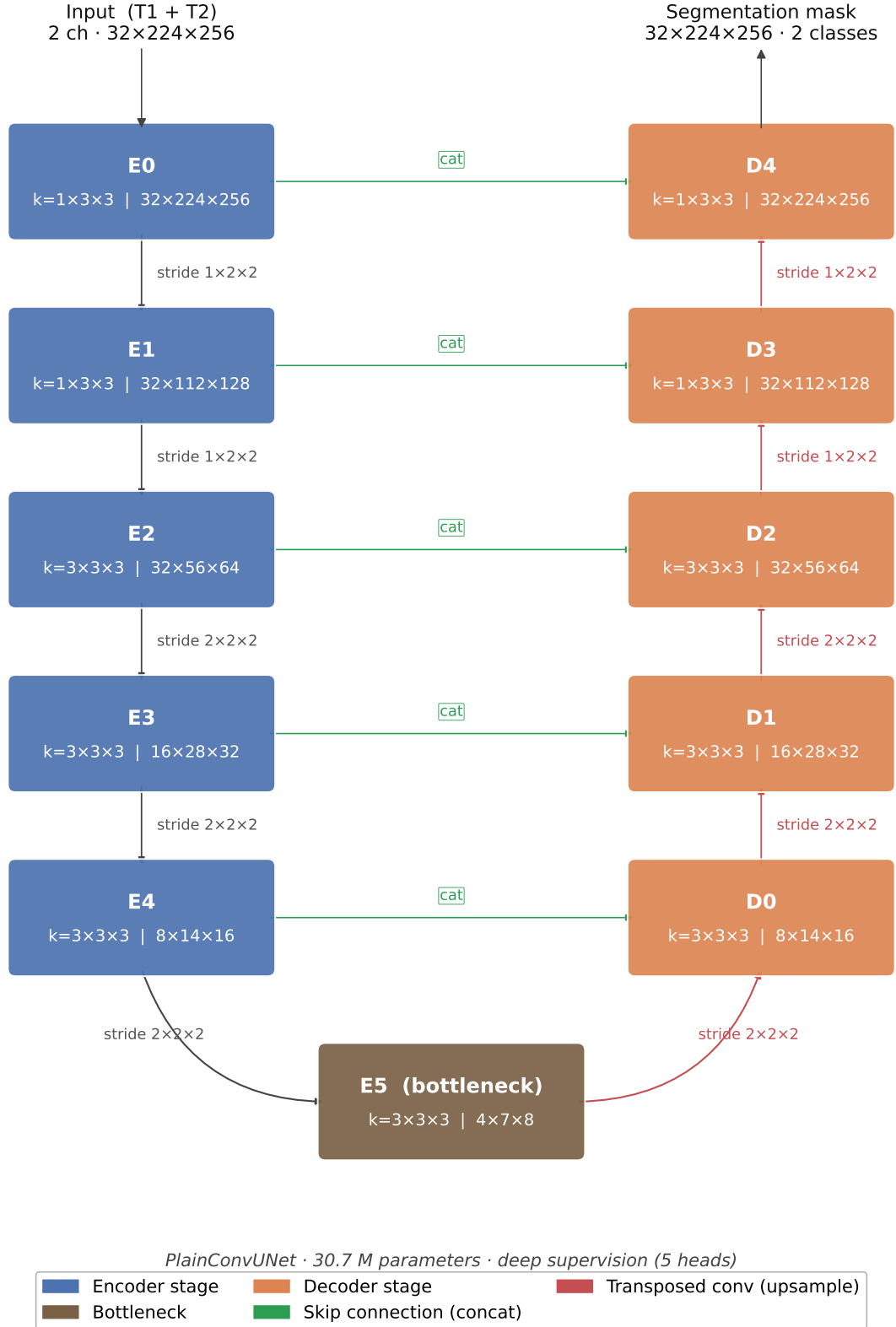


Figure 1: PlainConvUNet auto-configured by nnU-Net for the dual-channel T1+T2 dataset (3d full-resolution). Each box shows the spatial dimensions ($D \times H \times W$) at that resolution level. The first two downsampling transitions use anisotropic $1 \times 2 \times 2$ strides (in-plane only); the three deeper transitions are isotropic $2 \times 2 \times 2$.

Training proceeds for 1 000 epochs by default, where each epoch consists of 250 mini-batch gradient steps. The optimiser is stochastic gradient descent with Nesterov momentum ($\mu = 0.99$), and the learning rate follows a polynomial decay schedule:

$$\eta_k = \eta_0 \left(1 - \frac{k}{K}\right)^{0.9}, \quad (3)$$

where η_0 is the initial learning rate, k is the current epoch, and K is the total number of epochs.

Data augmentation is applied on-the-fly during training using the DA5 pipeline, which combines heavy spatial transforms with intensity perturbations. Spatial augmentations include random rotations in all three axes, random scaling, and elastic deformations with randomised grid coarseness and magnitude. Intensity augmentations include additive Gaussian noise, multiplicative brightness and contrast shifts, gamma correction, and simulation of low-resolution artefacts through downsampling and re-upsampling within the pipeline. This aggressive augmentation strategy is a primary reason for nnU-Net’s generalisation across diverse datasets, as it substantially increases the effective training set size and discourages the network from memorising acquisition-specific intensity patterns.

The default training objective is a composite loss combining the soft Dice loss and cross-entropy loss, formally defined in Section 2.6. This combination provides gradient signal from both overlap-based and pixel-wise perspectives, which is particularly beneficial for highly imbalanced segmentation tasks where the foreground structure occupies only a small fraction of the image volume.

Five-fold cross-validation is run by default, partitioning the training set into five equally sized folds and training five independent models. The ensemble of these five models is used as the final predictor, which typically improves generalisation compared to any single-fold model.

Inference

At test time, nnU-Net processes each volume through a sliding-window strategy because the full volume is generally too large to fit in GPU memory as a single patch. The window size matches the training patch size and is advanced with a 50% overlap in each spatial dimension. Predictions from overlapping windows are aggregated by weighted averaging, where voxels near the patch centre receive higher weight than those near the boundary — a Gaussian weighting scheme that reduces boundary artefacts common in hard-tiled prediction.

Test-time augmentation (TTA) via mirroring is applied by default. For each test volume, predictions are generated for the original orientation and for each of the possible mirror reflections along the spatial axes. The resulting predictions are averaged, which consistently improves segmentation quality at the cost of increased inference time.

Following prediction, nnU-Net evaluates whether a postprocessing step — retaining only the largest connected foreground component — improves cross-validation performance. If it does, it is applied automatically; otherwise, the raw model output is used. This data-driven decision ensures that postprocessing is only introduced when it genuinely helps, rather than being applied unconditionally.

Preliminary experiments comparing the 2d and 3d full-resolution configurations showed that 3d version consistently outperformed the 2D baseline; all results reported in this thesis therefore use the 3d full-resolution configuration exclusively.

2.4. MRI Artefacts: Gibbs Ringing

Gibbs ringing, also called the truncation artefact, manifests in MRI images as a series of oscillating intensity bands that appear near sharp transitions between tissues of very different signal intensities [7]. The bands are oriented parallel to the boundary, decrease in amplitude with increasing distance from the interface, and can obscure fine anatomical detail in the immediate neighbourhood of the transition.

The artefact arises from the finite sampling of k-space — the Fourier domain representation of the MRI signal — where truncating the high-frequency components before reconstruction introduces the Gibbs phenomenon in the spatial domain.

Gibbs ringing is particularly pronounced in limb MRI because of the large high-contrast interface between subcutaneous fat and muscle. On T1-weighted sequences, fat returns a very bright signal while muscle returns a substantially darker one, creating a steep intensity step at the fat-muscle boundary. The resulting oscillations extend into both the surrounding fat and the adjacent muscular tissue, and can produce intensity patterns that locally resemble those caused by pathological changes, complicating both visual assessment and automated analysis.

This characteristic of limb MRI motivated the Gibbs ringing augmentation strategy described in Section 4.2, which deliberately introduces synthetic ringing during training to improve the model’s robustness to this domain-specific artefact. Figure 2 shows a representative example from the dataset.



Figure 2: Gibbs ringing artefacts in three T1-weighted limb MRI slices from the dataset. Oscillating intensity bands appear along high-contrast fat-muscle boundaries, with varying severity across acquisitions.

2.5. Related Work

Only one published study has addressed automatic segmentation of lipomatous tumours in MRI to date. Liu et al. [8] proposed an ensemble framework combining 2D U-Net and attention U-Net models, each trained independently on either T1-weighted or proton-density MR images with multiple normalisation strategies, across 185 patients from four institutions. The models were combined into a Super Learner ensemble with optimised weights, achieving a mean Dice Similarity Coefficient (DSC) of 0.803 ± 0.184 on 20 held-out test cases. The study demonstrated that data heterogeneity — arising from multicenter acquisition with different scanners and protocols — is the primary challenge for segmentation of lipomatous tumours, and that ensemble learning over diverse normalisation strategies helps mitigate it. The authors noted as a limitation that their approach operates on 2D axial slices and does not model inter-slice spatial dependencies, leaving volumetric context unexploited. Beyond lipomatous tumours, automated segmentation of soft-tissue sarcomas in MRI remains an open problem. The field lacks the large, standardised benchmarks available for organ segmentation, and most published works deal with small, single-centre cohorts. nnU-Net has demonstrated strong performance on rare-tumour segmentation tasks precisely because its self-configuring pipeline avoids the need for large-scale hyperparameter searches [6], making it well-suited to the data-scarce regime characteristic of ALT studies.

The present work differs from Liu et al. [8] in three key respects: (i) it operates entirely in three dimensions, capturing inter-slice context through the 3D U-Net backbone of nnU-Net; (ii) it incorporates T2-weighted imaging as a second input channel alongside T1, providing complementary tissue characterisation; and (iii) it introduces a Gibbs ringing augmentation strategy tailored to the characteristic artefacts of limb MRI. Despite operating on a substantially smaller cohort (46 vs. 185 patients), the proposed pipeline achieves competitive segmentation performance.

2.6. Loss Functions

nnU-Net trains with a composite loss that combines a region-overlap term and a per-voxel classification term.

Dice Loss

The Dice loss maximises the overlap between predicted and reference segmentation masks; it naturally handles class imbalance by normalising by the total predicted and reference volume rather than counting absolute voxels.

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_i p_i g_i}{\sum_i p_i^2 + \sum_i g_i^2} \quad (4)$$

Here $p_i \in [0, 1]$ is the predicted probability and $g_i \in \{0, 1\}$ is the ground-truth label for voxel i [9].

Cross-Entropy Loss (CE)

The cross-entropy loss penalises per-voxel confidence errors and provides numerically stable gradient signal throughout training, including in early epochs when the Dice loss gradient can vanish if predictions are near-

uniform.

$$\mathcal{L}_{\text{CE}} = - \sum_i \left[g_i \log p_i + (1 - g_i) \log(1 - p_i) \right] \quad (5)$$

Composite Dice+CE Loss

The two losses have complementary failure modes — Dice optimises recall-precision balance globally but stalls when predictions are uniform; CE provides local gradient signal from the first iteration.

$$\mathcal{L} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}} \quad (6)$$

This composite loss is the default in nnU-Net and is used as the training objective in all experiments in this thesis. Evaluation, however, is performed using the Dice Similarity Coefficient (DSC) computed on hard predictions after thresholding — not the soft training loss.

3. Data and Problem Setting

3.1. Patient Cohort

The dataset comprises 46 patients with histologically confirmed Atypical Lipomatous Tumour of the extremities, identified by sequential identifiers IOG1 through IOG56 (non-consecutive). For each patient, a T1-weighted and a T2-weighted MRI volume were acquired in the same imaging session, providing inherently co-registered multi-modal data without the need for inter-session alignment.

Annotation Protocol

Tumour boundaries were delineated manually by a physician on a slice-by-slice basis, independently for each modality: each patient has a separate T1 annotation and a separate T2 annotation. The two annotations can differ in extent, as T1 and T2 sequences highlight different tissue characteristics of the tumour. The resulting masks and corresponding image volumes were stored as NeuroImaging Informatics Technology Initiative files (.NIfTI).

3.2. Dataset Construction

Three nnU-Net datasets were derived from the 46-patient cohort, each targeting a different input modality configuration. nnU-Net requires datasets to follow the naming convention `DatasetXXX_Name`; the identifiers 501–503 were assigned to these three configurations.

Dataset501 (T1-only). All 46 patients contribute one single-channel volume each. The T1-weighted image and the corresponding `_label.nii.gz` mask are converted to nnU-Net’s NIfTI format without modification. This dataset serves as the single-modality T1 baseline.

Dataset502 (T2-only). 45 patients contribute one single-channel T2 volume each.

Dataset503 (T1+T2 dual-channel). Each training sample is a two-channel volume in which channel 0 carries the T1 image and channel 1 carries the T2 image. The ground-truth label is the *union* of the T1 and T2 manual masks: a voxel is labelled positive if it is marked in either the T1 or the T2 annotation, capturing the full extent of the tumour as seen across both modalities.

3.3. Acquisition Heterogeneity

Voxel Spacing Variability

Table 1 summarises the voxel spacing statistics across all 46 patients for both modalities. In-plane resolution varies by a factor of $5.5\times$ for T1 and $5.0\times$ for T2, reflecting the heterogeneous acquisition protocols used across the cohort. Through-plane spacing is consistently coarser than in-plane spacing — the median through-plane value (5.0 mm) is approximately $6\times$ larger than the median in-plane value (0.78 mm), a characteristic anisotropy of clinical MRI acquisitions that prioritises anatomical coverage over depth resolution. This anisotropy is precisely what makes volumetric 3D processing valuable: a 3D model can exploit spatial context across consecutive thick slices that a 2D slice-wise approach would treat as independent. nnU-Net addresses the spacing heterogeneity by resampling all volumes to the dataset median spacing prior to training, ensuring the network always operates on data at a consistent physical resolution regardless of the original acquisition parameters.

Table 1: Voxel spacing statistics across all 46 patients (T1) and 45 patients (T2). *In-plane* spacing is the physical distance between adjacent pixels within a single image slice; *through-plane* spacing is the physical distance between consecutive slices. In-plane spacing is isotropic (identical in both height and width) for all acquisitions, so a single row covers both in-plane dimensions.

Modality	Direction	Min (mm)	Median (mm)	Max (mm)	Std (mm)
T1	In-plane	0.227	0.781	1.250	0.255
T1	Through-plane	3.300	5.000	7.200	1.033
T2	In-plane	0.286	0.755	1.438	0.273
T2	Through-plane	3.300	5.000	7.200	0.993

T1/T2 Orientation Mismatches

In 12 of the 46 patients (26%), the T1 and T2 volumes have different image orientations as encoded in their NIfTI affine matrices. The dominant T1 orientation is LAS¹ (41/46 cases); T2 volumes are distributed across three orientations: LAS (38), PSR (6), and LSP (2). The full list of mismatched cases is given in Table 2.

Table 2: Patients with T1/T2 orientation mismatches and their handling in Dataset503.

Patient	T1 orientation	T2 orientation	Dataset503 handling
IOG2	LSP	LAS	Fused (resampled)
IOG7	LSP	LAS	Split only
IOG10	LSP	LAS	Fused (resampled)
IOG27	LAS	PSR	Fused (resampled)
IOG28	LSP	PSR	Split only
IOG38	LSP	LAS	Fused (resampled)
IOG45	LAS	PSR	Split only
IOG46	LAS	PSR	Fused (resampled)
IOG47	LAS	LSP	Fused (resampled)
IOG51	LAS	PSR	Fused (resampled)
IOG55	LAS	PSR	Fused (resampled)

Other Edge Cases

IOG9 (degenerate T2). The T2 acquisition for IOG9 exhibits near-zero signal contrast across the tumour and surrounding tissue, rendering it uninformative. IOG9 is excluded from Dataset502 and contributes to Dataset503 with its T2 channel set to zero.

IOG26, IOG29, IOG35 (annotation uncertainty). These three cases have known inter-rater variability in their manual delineations. They are always emitted as split single-channel derived cases in Dataset503 to avoid propagating uncertain annotations into the fused dual-channel ground truth.

3.4. Cross-Validation Setup

All experiments use patient-level 5-fold cross-validation: all images belonging to a given patient appear in exactly one validation fold, preventing any data leakage between train and validation sets. The same fold assignments are applied identically to Dataset501, Dataset502, and Dataset503, ensuring that the T1, T2, and dual-channel models are evaluated on exactly the same patients in each fold.

Volume-Stratified Assignment

Fold assignment is stratified by tumour volume rather than random. Patients are sorted by their tumour voxel count and assigned to folds in round-robin order, so that each fold receives a representative sample of the full size distribution. This stratification is necessary because tumour volume spans an extreme range: the

¹Orientation codes follow the NIfTI convention: L/R (left/right), A/P (anterior/posterior), S/I (superior/inferior), indicating the positive direction of each image axis.

smallest tumour in the cohort contains approximately 405 voxels (corresponding to a sub-centimetre lesion) while the largest contains approximately 267,000 voxels — a factor of $660\times$. Under random assignment, a single fold receiving multiple small, low-contrast lesions (such as IOG1, IOG38, and IOG40) would collapse to near-zero Dice, artificially inflating variance across folds. Volume stratification ensures that each fold presents a comparable segmentation difficulty to the model.

3.5. Descriptive Statistics

Table 1 in Section 3.3 summarises voxel spacing. Beyond spacing, the raw image dimensions span a wide range: row counts from 256 to 704 pixels, column counts from 240 to 704 pixels, and slice counts from 12 to 69 — reflecting the variability in field-of-view and acquisition parameters across the cohort. nnU-Net’s preprocessing (resampling to the dataset median spacing, followed by patch-based training) abstracts away this variability, presenting the network with patches of consistent physical size regardless of the original image dimensions.

Tumour Volume Distribution

Table 3 summarises the tumour volume statistics for the cohort. The distribution is heavily right-skewed: the mean (179.5 cm^3) is more than twice the median (80.2 cm^3), driven by a small number of very large lesions that extend the upper tail. The $660\times$ voxel range motivates both the volume-stratified cross-validation described in Section 3.4 and the foreground oversampling strategy used during training (described in Section 4).

Table 3: Tumour volume summary statistics for the 46-patient cohort (T1 measurements).

n	Min	Q1	Median	Q3	Max	Mean $\pm$ Std
46	2.3	34.4	80.2	173.2	891.3	179.5 ± 233.1

All values in cm^3 .

4. Methods

4.1. Training Configuration

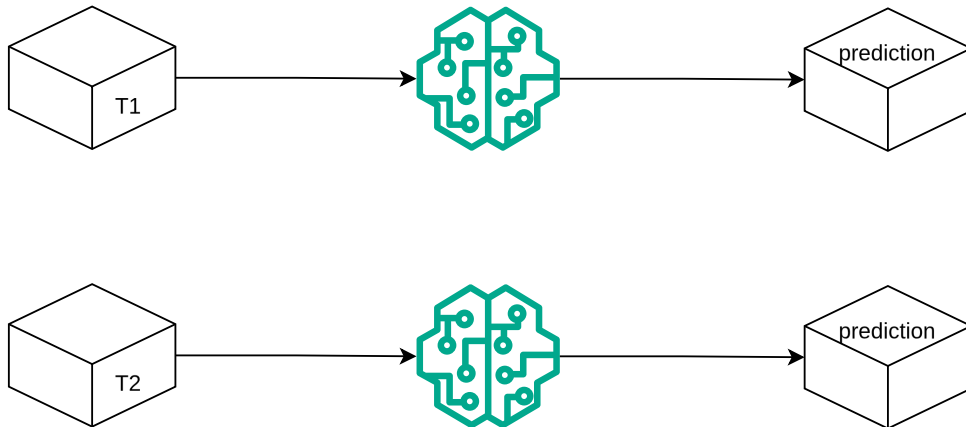


Figure 3: Input pipeline for the single-modality baseline models. Each modality is passed independently to its own nnU-Net instance.

All models in this thesis are trained using nnU-Net v2 with the 3d full-resolution configuration and `nnUNetPlans` auto-generated network plans, as described in Section 2.3. The DA5 augmentation pipeline and the Dice+CE composite loss (Section 2.6) are retained as defaults. The following modifications to the default training configuration were applied.

Training Duration

nnU-Net trains for 1 000 epochs by default. This work uses 350 epochs for all three datasets (T1-only baseline model, Dataset501; T2-only baseline model, Dataset502; and T1T2 dual model, Dataset503). The 46-patient

cohort is substantially smaller than the datasets for which the 1 000-epoch default was calibrated; preliminary runs showed that the pseudo-Dice EMA plateau is reached well before 1 000 epochs on this data. The 350-epoch budget was chosen to provide sufficient training for convergence while remaining computationally practical.

Per-Sample Dice Loss

nnU-Net v2 computes the Dice component of the composite loss at the batch level for the 3d full-resolution configuration. This is overridden to per-sample Dice (`batch_dice=False`), so that each case in a mini-batch receives an independent gradient signal. This is particularly important given the $660\times$ tumour volume range in this dataset: batch-level Dice can be dominated by the numerically largest tumours, suppressing the gradient contribution of small, difficult lesions.

Postprocessing

nnU-Net evaluates whether a largest-connected-component postprocessing step improves cross-validation performance, and applies it only if it does.

4.2. Gibbs Ringing Augmentation

Gibbs ringing is a structured, spatially coherent artefact characteristic of limb MRI (Section 2.4). The standard DA5 augmentation pipeline includes stochastic intensity transforms such as additive Gaussian noise and multiplicative brightness shifts, but none that replicate the oscillating band pattern produced by k-space truncation. A model trained without exposure to Gibbs ringing may conflate these intensity oscillations with genuine tissue boundaries or tumour heterogeneity, reducing its robustness to this systematic domain characteristic.

To address this, a custom augmentation transform — `GibbsRingingTransform` — is inserted after the `SpatialTransform` in the DA5 pipeline. The transform operates independently on each input channel with probability $p = 0.15$. For a selected channel with volume x , the augmentation proceeds as follows:

1. Compute the three-dimensional discrete Fourier transform to obtain the k-space representation $\mathcal{F}(x)$.
2. Apply a binary mask M_α that retains only the central fraction $\alpha \sim \mathcal{U}(0.5, 0.9)$ of k-space in each spatial dimension, discarding the outer high-frequency content.
3. Compute the inverse Fourier transform and take the real part:

$$\hat{x} = \text{Re}[\mathcal{F}^{-1}(\mathcal{F}(x) \cdot M_\alpha)] \quad (7)$$

Lower values of α produce stronger ringing. The range $[0.5, 0.9]$ was chosen to span mild to moderate ringing intensities representative of those observed in the dataset. At $\alpha = 0.9$, approximately 73% of k-space energy is retained and the resulting ringing is subtle; at $\alpha = 0.5$, only 12.5% is retained and the oscillations are pronounced.

4.3. T1T2 Dual Model

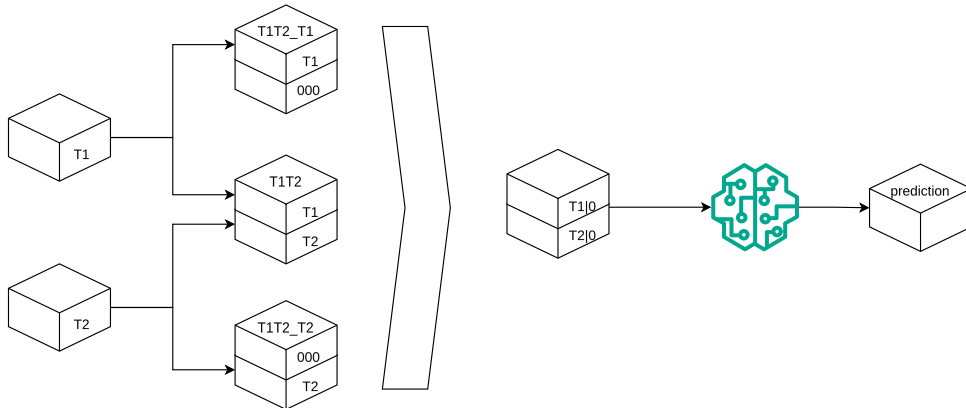


Figure 4: Input pipeline for the T1T2 dual model. T1 and T2 volumes are concatenated as two channels and passed jointly to a single nnU-Net. To produce the $T1T2_{T1}$ and $T1T2_{T2}$ degraded-input variants the missing channel is zeroed.

Dataset503 was constructed as described in Section 3.2: each sample is a two-channel NIfTI volume with T1 as channel 0 and T2 as channel 1, labelled by the union of the two modality-specific annotations. From a training perspective, nnU-Net treats this as a standard two-channel input without any architectural modification — the self-configuration pipeline reads the number of input channels from the dataset metadata and adjusts the first convolutional layer accordingly.

Per-channel intensity normalisation is applied independently: each channel is z-score normalised using the foreground voxel statistics of that channel, as is standard for MRI inputs in nnU-Net. This ensures that T1 and T2 intensity scales do not introduce modality-specific bias despite their different signal characteristics.

Handling of Split-Only Cases

As described in Section 3.3, 11 patients have T1 and T2 volumes in different image orientations (Table 2). Eight are handled by resampling the T2 onto the T1 grid. The remaining three (IOG7, IOG28, IOG45) are split into modality-specific derived cases (IOG7_T1, IOG7_T2, etc.) because their acquisitions also differ in anatomical coverage, making the fused volume degenerate.

IOG35 shares its orientation pair (LAS/LSP) with IOG47, which was fused successfully, so orientation is not the obstacle. It is contributed as split-only cases solely because its manual delineations have low inter-rater agreement: fusing an uncertain annotation into the dual-channel ground-truth union would propagate annotation noise into training.

All derived cases participate in training alongside the fused dual-channel cases without requiring any special handling by the trainer.

Dataset503 therefore produces three inference variants per patient. The *T1T2 dual* model receives both T1 and T2 inputs simultaneously and is evaluated on the 40 geometrically aligned cases. For patients where registration was not possible, the two modalities are passed as independent inputs: the *T1T2_{T1} dual* model supplies the T1 volume with the T2 channel set to zero, and the *T1T2_{T2} dual* model does the reverse. These degraded-input variants allow the model’s robustness to a missing modality to be assessed without retraining.

4.4. T1T2 Ensemble Model

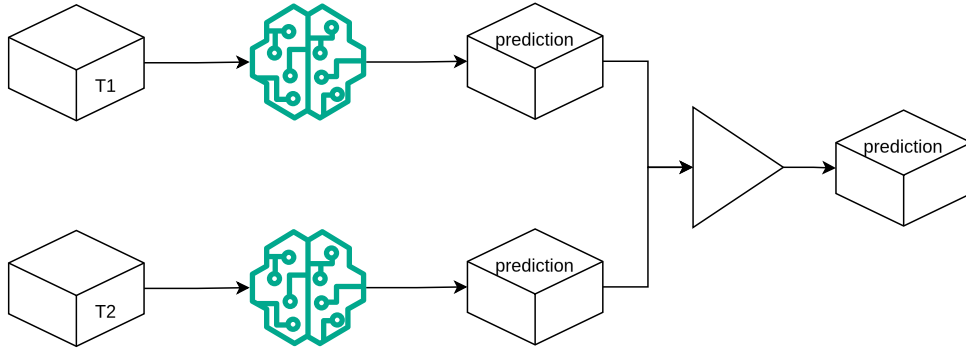


Figure 5: Input pipeline for the T1T2 ensemble model. The T1-only and T2-only baseline predictions are combined via voxel-level union at inference time; no joint training is performed.

In addition to the T1T2 dual model, the T1T2 ensemble model combines predictions from the independently trained T1-only baseline model and T2-only baseline model by applying a union rule at the voxel level: a voxel is assigned to the foreground class if either model labels it as tumour. This strategy requires no joint training, introduces no additional parameters, and is motivated by the complementary contrast characteristics of T1 and T2 — a region missed by one model due to low contrast in that modality may exhibit higher contrast and better detectability in the other.

The T1T2 ensemble model is evaluated under the same 5-fold cross-validation framework employed for the individual models: for each fold, the corresponding fold models from the T1-only baseline and T2-only baseline are combined, ensuring that no test-set information is used in either component model.

4.5. Evaluation Protocol

All models are evaluated under the same five-fold patient-level cross-validation described in Section 3.4. The primary performance metric is the Dice Similarity Coefficient (DSC),

$$\text{DSC} = \frac{2|P \cap G|}{|P| + |G|}, \quad (8)$$

where P and G denote the predicted and ground-truth foreground voxel sets respectively. The Intersection over Union (IoU, Jaccard index) is reported as a secondary metric. Both the mean and the median Dice are reported throughout the results; a small number of cases produce zero-Dice predictions due to complete segmentation failures, which suppress the mean disproportionately, so the median better represents typical performance.

Statistical comparisons are performed using the Wilcoxon Signed-Rank test, paired on the intersection of cases shared by both models under comparison. All comparisons were pre-specified prior to analysis; no correction for multiple comparisons is applied. Significance thresholds: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***) ; otherwise not significant (ns).

Dataset503 produces three model variants per patient. The *T1T2 dual* model receives both T1 and T2 inputs simultaneously and is available only for the 40 geometrically aligned cases. For the remaining patients, the two modalities are supplied as independent inputs: the *T1T2_T1 dual* model uses the T1 volume with the T2 channel zeroed, and the *T1T2_T2 dual* model uses the T2 volume with the T1 channel zeroed. These degraded-input variants assess the T1T2 dual model’s robustness to missing modalities. Any comparison involving the T1T2 dual subset is restricted to its intersection with the other model’s case list; the resulting sample size (n) is stated explicitly in each caption.

5. Experiments and Results

5.1. Hardware and Training Details

All experiments were conducted on a single NVIDIA Quadro RTX 5000 GPU with 16 GB of VRAM. Per-epoch training time was approximately 83 seconds with the Gibbs augmentation trainer and 350 epochs, yielding approximately 8.1 hours per fold and 40 hours for the complete 5-fold run per dataset. Training used PyTorch via nnU-Net v2 with CUDA acceleration; `torch.compile` was enabled to reduce kernel launch overhead. Data augmentation was applied on-the-fly during training using background worker processes.

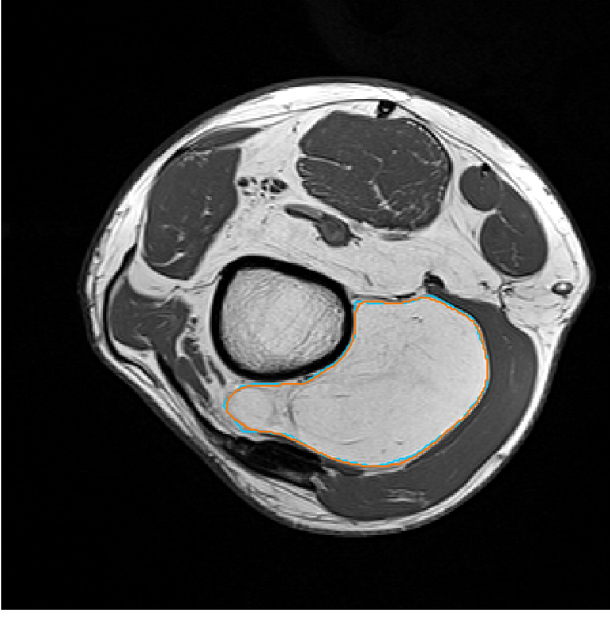
5.2. Single-Modality Baselines

Figure 6 shows a representative case (IOG56) for each single-modality baseline, illustrating the typical quality of segmentation on a well-represented patient.

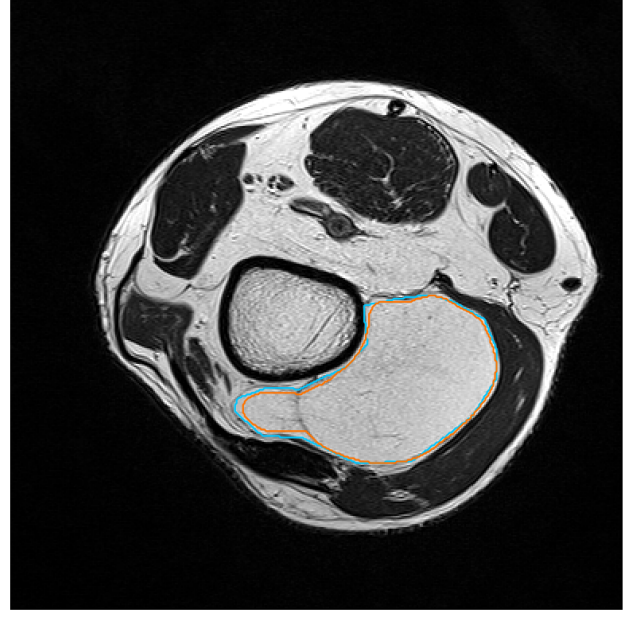
Table 4 summarises aggregate performance across all evaluated models. The T1-only baseline model (Dataset501) achieves a median Dice of 0.884 (IQR 0.309; Q1 = 0.625, Q3 = 0.934) and median IoU of 0.792 over 46 cases. The T2-only baseline model (Dataset502) achieves a median Dice of 0.851 (IQR 0.270; Q1 = 0.660, Q3 = 0.930) and median IoU of 0.741 over 45 cases. The difference between the two baseline models is not statistically significant (Wilcoxon, $p = 0.346$, $n = 45$), indicating broadly comparable discrimination ability.

Both single-modality models exhibit wide IQR, driven by a small number of zero-Dice failures (IOG28 on T1 due to acquisition orientation mismatch, IOG45 due to extremely small tumour volume) and anatomically challenging cases discussed in Section 6.4.

Figure 7 compares the Dice and IoU distributions of the T1-only baseline model and T2-only baseline model alongside the T1T2_T1 and T1T2_T2 dual models (discussed in Section 5.3). Figures 8 and 9 show per-case Dice scores for the T1-only baseline model and T2-only baseline model respectively, sorted in ascending order.



(a) T1-only baseline (Dice = 0.958).



(b) T2-only baseline (Dice = 0.954).

Figure 6: Representative segmentation on patient IOG56. **Blue:** ground-truth contour; **Orange:** model prediction contour. Both baseline models achieve near-perfect Dice on this case, illustrating the complementary appearance of T1 and T2 images and the high-quality predictions attainable on well-represented patients.

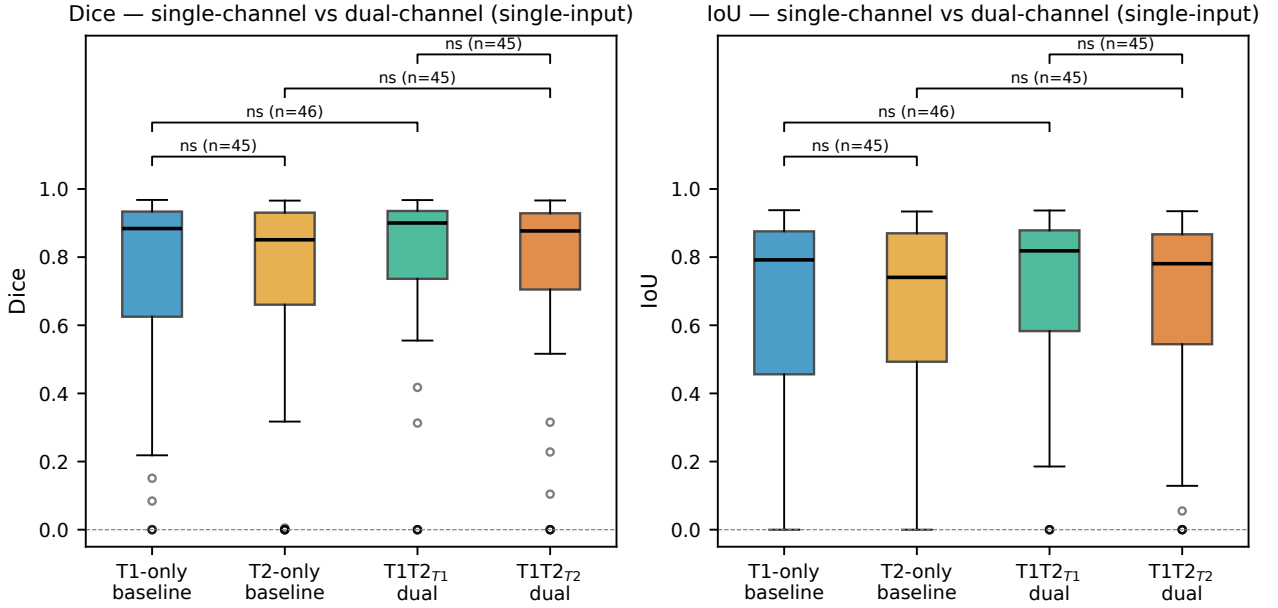


Figure 7: Dice (left) and IoU (right) distributions for the T1-only baseline model (DS501, $n = 46$), T2-only baseline model (DS502, $n = 45$), T1T2_{T1} dual model, and T1T2_{T2} dual model. Both metrics are bounded in $[0, 1]$. Significance brackets show Wilcoxon Signed-Rank test results (two-sided, paired on common cases; **ns**: not significant). Boxes show interquartile range; whiskers extend to $1.5 \times \text{IQR}$; outliers shown as dots.

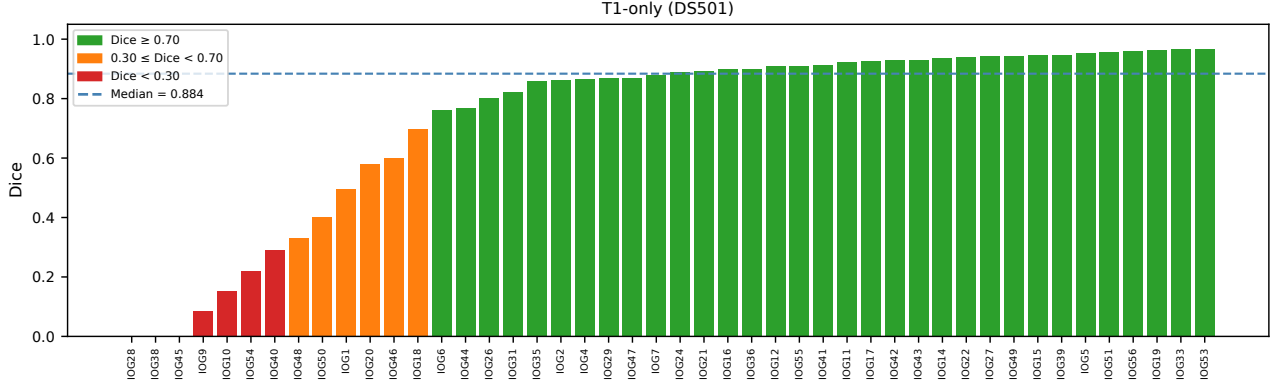


Figure 8: Per-case Dice scores for the T1-only baseline model (DS501, $n = 46$), sorted in ascending order. Colour indicates performance tier: green ≥ 0.70 , orange $0.30 \leq \text{Dice} < 0.70$, red < 0.30 . Dashed line marks the cross-validation median (0.884).

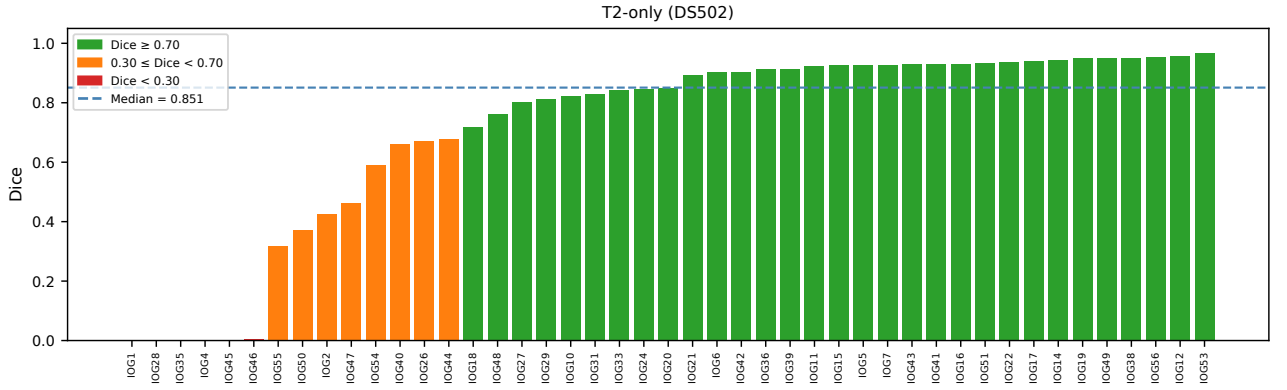


Figure 9: Per-case Dice scores for the T2-only baseline model (DS502, $n = 45$), sorted in ascending order. Colour coding as in Figure 8. Dashed line marks the cross-validation median (0.851).

Both distributions share the same bimodal shape: a large majority of cases cluster in the green zone (≥ 0.70), with a tail of difficult cases pulling the lower quartile down. The T1-only model has a slightly higher median (0.884 vs. 0.851) and a comparable green-zone count (33/46 vs. 31/45), yet its IQR is wider (0.309 vs. 0.270), driven by a heavier low-Dice tail.

5.3. T1T2 Dual Model

Figure 10 shows the three Dataset503 model variants on patient IOG33, illustrating the benefit of joint dual-channel training over single-channel degraded inputs.

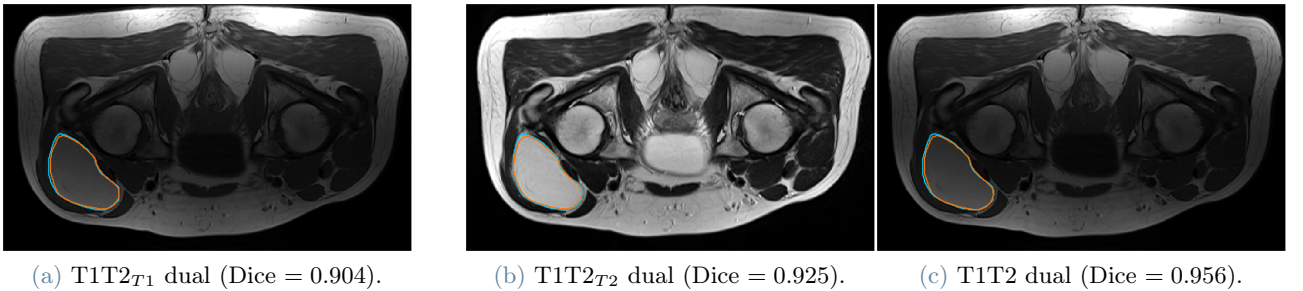


Figure 10: Dataset503 model variants on patient IOG33. **Blue**: ground-truth contour; **Orange**: model prediction contour. The T1T2 dual model (left) outperforms both degraded-input variants by +5.2 and +3.1 percentage points, demonstrating the gain from joint cross-modal training.

The T1T2 dual model ($n = 40$) achieves a median Dice of 0.913 (IQR 0.169; $Q1 = 0.773$, $Q3 = 0.942$) and median IoU of 0.839, representing a gain of +2.9 percentage points in median Dice over the T1-only baseline model and +6.2 points over the T2-only baseline model. The narrower IQR (0.169 versus 0.309 and 0.270) confirms more consistent predictions across the patient cohort.

Benefit of dual-channel training. Figure 7 compares the T1T2_T1 and T1T2_T2 dual models against the dedicated baseline models. The T1T2_T1 dual model — operating on T1 input alone — does not significantly outperform the T1-only baseline model (Wilcoxon, $p = 0.583$, $n = 46$, not significant), nor does the T1T2_T2 dual model over the T2-only baseline model ($p = 0.169$, $n = 45$, ns). The two degraded-input variants themselves are also not significantly different ($p = 0.205$, $n = 45$, ns), indicating that neither modality dominates when the model is forced into single-channel mode.

Figure 11 compares the T1T2 dual model directly against both baselines. The T1T2 dual model significantly outperforms the T2-only baseline (Wilcoxon, $p = 0.008$, $n = 39$), while the improvement over the T1-only baseline does not reach significance ($p = 0.572$, $n = 40$). This asymmetry reflects the complementary failure sets identified in Section 5.2: the T1-only model already handles most of the aligned cases well, whereas the T2-only model has a heavier low-Dice tail that joint training corrects.

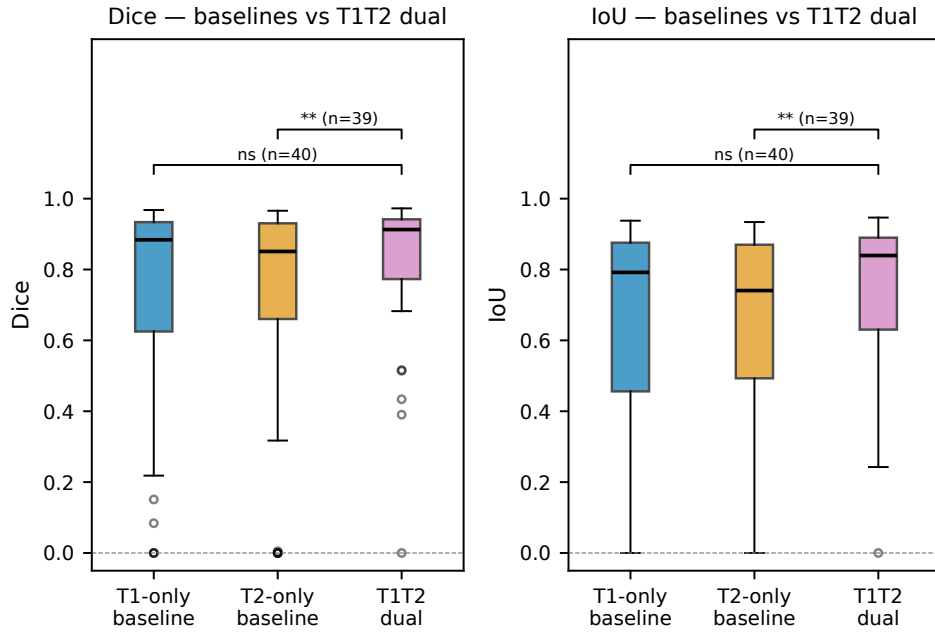


Figure 11: Dice (left) and IoU (right) distributions for the T1-only baseline ($n = 46$), T2-only baseline ($n = 45$), and T1T2 dual model ($n = 40$). Brackets compare each baseline against the T1T2 dual model (T1 vs dual: $n = 40$; T2 vs dual: $n = 39$). Two-sided Wilcoxon Signed-Rank test (ns: not significant, **: $p < 0.01$).

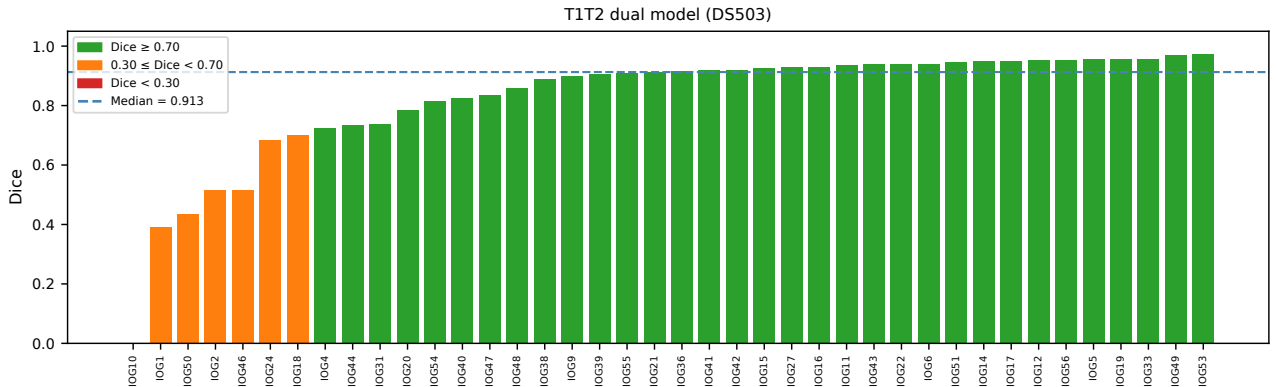


Figure 12: Per-case Dice scores for the T1T2 dual model (DS503, $n = 40$), sorted in ascending order. Colour coding as in Figure 8. Dashed line marks the cross-validation median (0.913).

5.4. T1T2 Ensemble Model

Figure 13 shows the constituent predictions and their union on patient IOG19, illustrating how the ensemble combines both modalities.

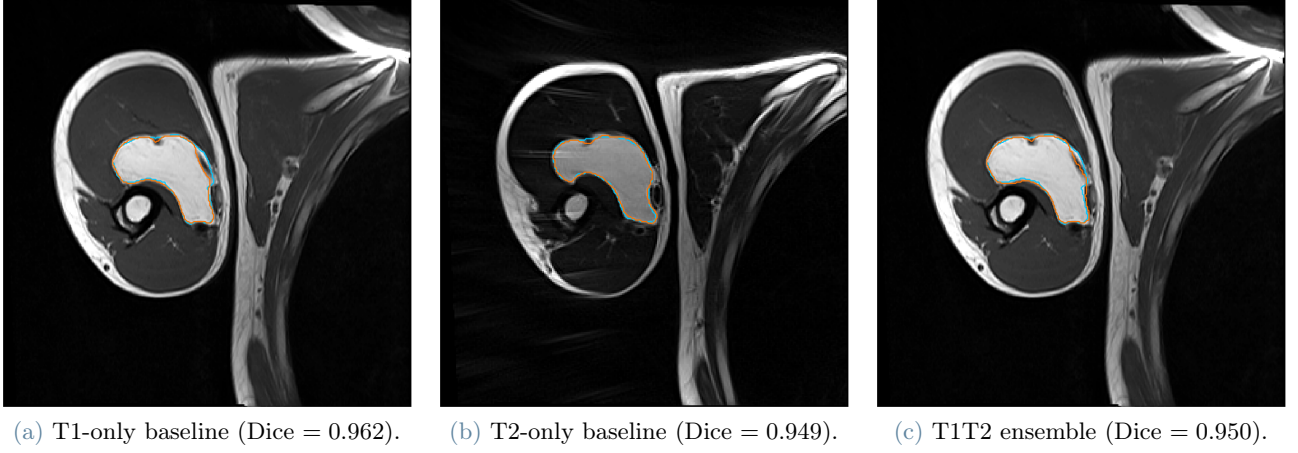


Figure 13: Ensemble model on patient IOG19. **Blue**: ground-truth contour; **Orange**: model prediction contour. The T1T2 ensemble (right) is the voxel-level union of the T1-only (left) and T2-only (centre) predictions.

The T1T2 ensemble model achieves a median Dice of 0.883 (IQR 0.241; Q1 = 0.686, Q3 = 0.927) and median IoU of 0.790 over 45 cases, outperforming both baseline models but falling below the T1T2 dual model.

Figure 15 compares the T1T2 ensemble model against the T1T2 dual model on the shared subset of cases ($n = 39$; Wilcoxon $p = 0.011$, *). The performance gap confirms that joint training on the concatenated T1+T2 input allows the model to learn cross-modal correlations that output-level combination cannot recover. The union strategy is also more prone to boundary false positives: when both baseline models predict overlapping but spatially offset regions, the union expands the predicted mask beyond either individual prediction. Figure 14 shows the per-case Dice breakdown for the T1T2 ensemble model, following the same layout as the baseline and T1T2 dual charts.

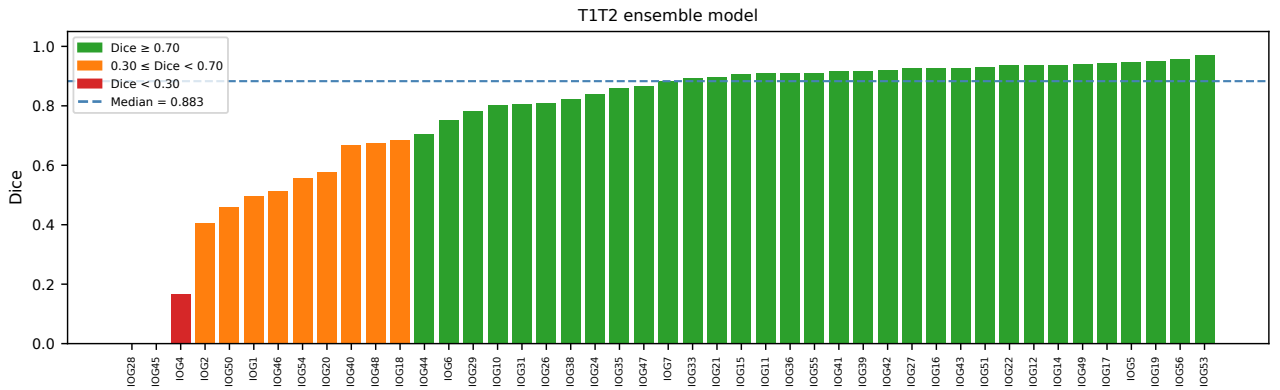


Figure 14: Per-case Dice scores for the T1T2 ensemble model ($n = 45$), sorted in ascending order. Median Dice = 0.883. Colour coding as in Figures 8–12: green ≥ 0.70 , orange 0.30–0.70, red < 0.30 .

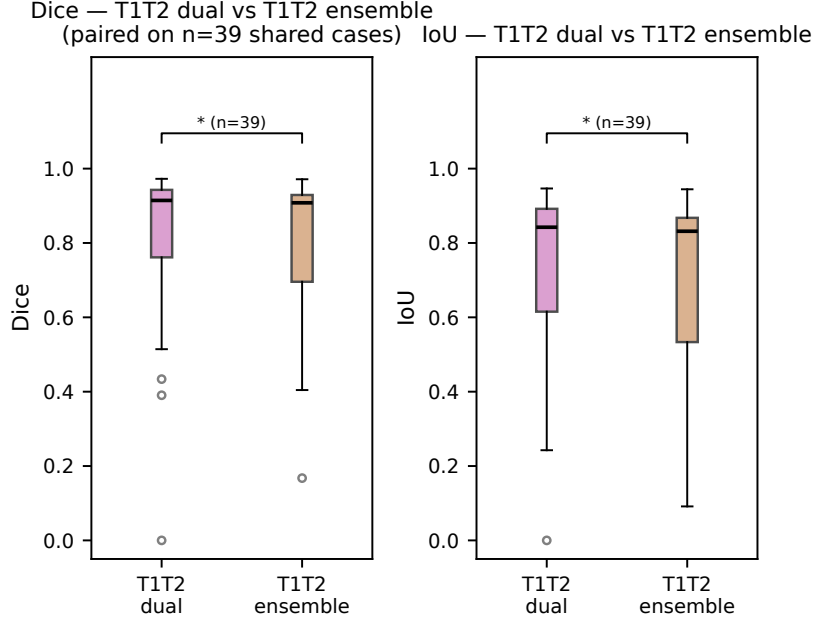


Figure 15: Dice (left) and IoU (right) for the T1T2 dual model ($n = 40$) and the T1T2 ensemble model ($n = 45$), paired on their shared case intersection ($n = 39$). Significance bracket shows two-sided Wilcoxon test result ($p = 0.011$).

5.5. Ablation Summary

Table 4 collects the aggregate performance of all evaluated configurations. The T1T2 dual model achieves the best median Dice and narrowest IQR across all metrics. The T1T2 ensemble model improves on the baseline models but does not match joint training. T1T2 dual model metrics are reported on the 40 geometrically aligned cases to ensure like-for-like comparability with the baseline models; the full 131-case median (including the T1T2_T1 and T1T2_T2 dual model variants) is 0.902.

Table 4: Five-fold cross-validation results for all model configurations. T1T2 dual model on $n = 40$ aligned cases; T1-only baseline $n = 46$; T2-only baseline $n = 45$; T1T2 ensemble $n = 45$. Bold indicates best value per column.

Model	Median Dice	IQR Dice	Median IoU	IQR IoU
T1-only baseline	0.884	0.309	0.792	0.419
T2-only baseline	0.851	0.270	0.741	0.377
T1T2 dual model	0.913	0.169	0.839	0.259
T1T2 ensemble model	0.883	0.241	0.790	0.342

Figure 16 shows a representative case (IOG40) with a clear step-wise progression from T1-only to the T1T2 dual model.

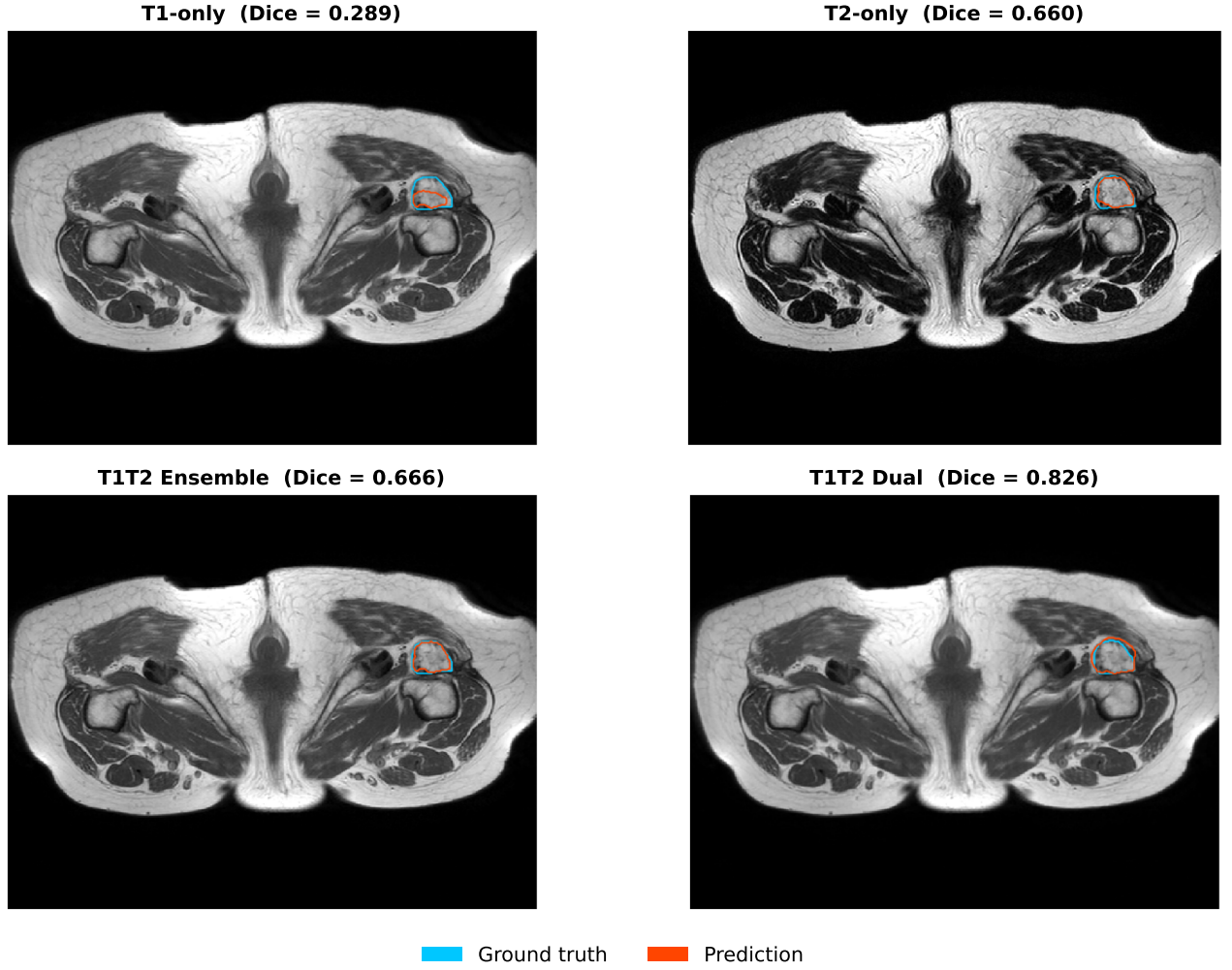


Figure 16: Qualitative ablation on patient IOG40, a pelvic ALT (axial slice 23/35). **Blue**: ground truth contour; **Orange**: model prediction contour. The T1-only model produces an under-segmented contour displaced from the ground-truth boundary (Dice = 0.289). The T2-only model recovers the lesion shape more accurately (Dice = 0.660), and the T1T2 ensemble achieves a marginal further gain by unioning both predictions (Dice = 0.666). The T1T2 dual model achieves the tightest boundary match (Dice = 0.826).

5.6. Tumor Volume and Segmentation Difficulty

Figure 17 shows the relationship between ground-truth tumour volume (cm^3) and Dice score for the T1T2 dual model ($n = 40$). Tumour volumes are computed from the nnU-Net preprocessed label masks as foreground voxel count multiplied by voxel volume derived from the image header.

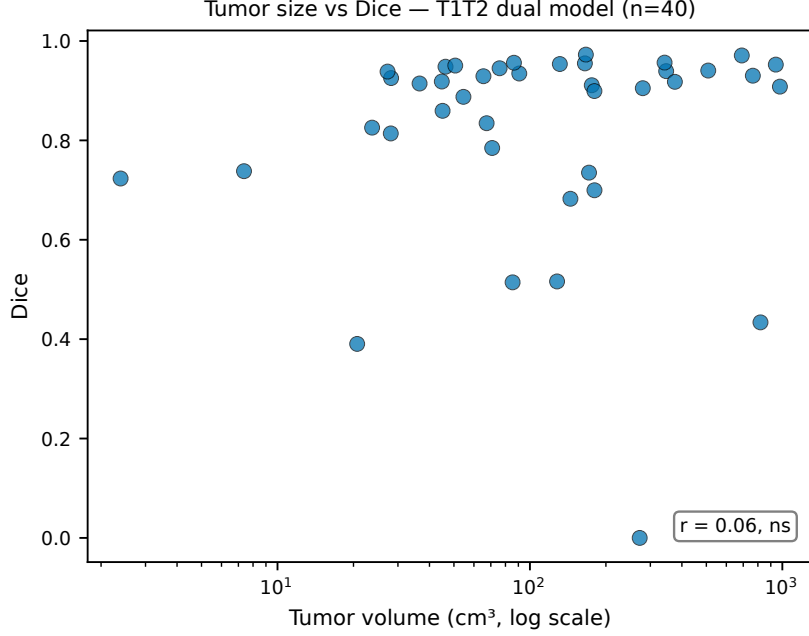


Figure 17: Ground-truth tumour volume (log scale) versus Dice score for the T1T2 dual model ($n = 40$). Pearson r computed on log-transformed volumes; the x-axis uses a logarithmic scale to accommodate the wide range of tumour sizes.

The Pearson correlation between log-tumour volume and Dice is $r = 0.059$ ($p = 0.719$, ns), indicating no significant relationship between tumour size and segmentation quality in this cohort. The failure cases discussed in Section 6.4 (IOG28, IOG45, and others) contribute low-Dice outliers across the full size range, confirming that poor performance is driven by acquisition artefacts and label quality issues rather than tumour size alone.

6. Discussion

6.1. Why Joint Dual-Channel Training Outperforms Output-Level Fusion

The +3.0 percentage-point gap in median Dice between the T1T2 dual model (0.913) and the T1T2 ensemble model (0.883) — a statistically significant difference ($p = 0.011$, Wilcoxon, $n = 39$) — cannot be attributed to access to additional data: both approaches draw on T1 and T2 signal from the same patients. The gap instead reflects a fundamental difference in how the two modalities are integrated during learning.

The T1T2 ensemble model combines T1-only baseline model and T2-only baseline model predictions at the output level, after each model has been independently optimised on a single input channel. By the time the union is computed, each constituent model has committed to a decision boundary in its own feature space; the union merely aggregates two independent binary maps. Disagreements between the two models are resolved by defaulting to positive, which benefits recall but increases false positives at the boundary — a structural limitation of any output-level fusion strategy.

Joint dual-channel training takes a categorically different approach. The T1T2 dual model receives T1 and T2 as two input channels simultaneously and is trained end-to-end to produce a single prediction. Gradients propagate jointly through both input streams, allowing the encoder to learn cross-modal correlations: which voxels are bright on T1 but suppressed on T2 (indicating lipomatous content), which regions exhibit T2 signal elevation without a T1 correlate (suggesting oedema or fibrous stroma), and how the spatial relationship between these patterns defines the tumour boundary. This cross-modal reasoning is invisible to independently trained models, yet it corresponds precisely to the multiparametric criteria used by human annotators when delineating ALT.

The asymmetry observed in Section 5.3 corroborates this interpretation. The T1T2 dual model significantly outperforms the T1T2_T2 dual model ($p = 0.003$, $n = 39$), but the improvement over the T1T2_T1 dual model does not reach significance ($p = 0.665$, $n = 40$). T2 signal is less discriminative in isolation — the T2 appearance of ALT is heterogeneous and overlaps with benign processes — but becomes more informative once T1 provides the fat-specific context. The T1T2 dual model has learned to exploit this synergy; the T1T2 ensemble model cannot.

This asymmetry is consistent with the MRI characteristics of ALT: the tumour contains adipose tissue that

appears hyperintense on T1-weighted images, providing strong and specific contrast against the surrounding muscle. T2 images offer complementary information about internal heterogeneity and oedema but carry less specific signal for the lipomatous component, making T1 the primary discriminative channel and T2 a contextual one.

6.2. Gibbs Ringing Augmentation as a Domain Robustness Mechanism

Gibbs ringing arises whenever an MRI acquisition uses a finite k-space sampling window [7]. High-contrast interfaces — such as the sharp fat-muscle boundary that defines the outer envelope of an ALT tumour — produce oscillatory overshoots and undershoots in the reconstructed image. In limb imaging, the axial field of view is narrow and k-space is commonly undersampled along the phase-encode direction, making Gibbs artifacts particularly prevalent at the tumour periphery. The severity of the artifact varies across scanners, field strengths, and acquisition protocols; the same patient imaged on different hardware can exhibit substantially different apparent boundary sharpness.

Standard augmentation pipelines, including nnU-Net’s DA5 baseline, apply spatial transforms, intensity jitter, and noise perturbations, but do not simulate k-space truncation effects. A model trained without Gibbs augmentation may inadvertently use boundary ringing patterns as a localisation cue: the artifact is correlated with tumour boundary presence in the training data, and the model can exploit this spurious correlation. However, the ringing pattern is unreliable across acquisition settings, causing unexpected failures when a test image is acquired with different k-space parameters that alter its character.

By inserting a random k-space truncation step — retaining a random central fraction of k-space in $[0.5, 0.9]$ with probability $p = 0.15$ per channel — the Gibbs augmentation trainer exposes the model to a wide range of synthetic ringing intensities during training. The model is thereby forced to rely on signal characteristics that are robust to truncation, primarily the T1 hyperintensity of adipose tissue, rather than the boundary artifact pattern itself. This regularisation is analogous to standard noise injection or dropout in that it prevents the network from exploiting fragile, acquisition-specific correlations.

The design of the augmentation is principled, and its empirical benefit is confirmed by ablation experiments comparing the Gibbs-augmented trainer against equivalent configurations without Gibbs augmentation. The Gibbs-augmented variant consistently achieved higher Dice across configurations and was therefore selected as the basis for all final model configurations reported in this thesis.

6.3. Evaluation Generalisability and the Absence of a Held-Out Test Set

All performance figures reported in this thesis are derived from five-fold patient-level cross-validation; no independent held-out test set was reserved. This design was unavoidable given the dataset size. With 46 patients, reserving even 10% as a held-out set would reduce the training pool to approximately 41 patients and the test set to 5 — too small to yield stable estimates of population-level performance. A randomly chosen five-patient holdout would vary by several Dice points across different random seeds, making it a less reliable performance proxy than five-fold CV, which averages over five complementary nine-patient validation folds that collectively cover the entire cohort.

Patient-level fold assignment guarantees that no patient contributes to both training and validation within a given fold, preventing direct data leakage. Five-fold cross-validation is therefore the standard evaluation strategy for small medical imaging datasets and provides a reasonable proxy for generalisation to new patients drawn from the same acquisition protocol [1].

6.4. Limitations

The per-case Dice distributions reveal a bimodal pattern common in small-dataset medical segmentation: a cluster of near-zero scores and a dense cluster above 0.80, with few intermediate results (0.30–0.70). Case difficulty is largely determined by anatomy-level factors — tumour size, acquisition alignment, boundary sharpness, and annotation quality — rather than a continuous difficulty gradient. The specific failure modes and one high-performance example are examined in the paragraphs below.

Notably, the failure sets of the T1-only and T2-only baseline models are largely non-overlapping. IOG28 and IOG45 are the only shared zero-Dice cases (orientation mismatch and near-degenerate tumour volume, respectively). Beyond these, T1-only fails substantially on cases such as IOG38 (Dice = 0.000) and IOG10 (0.151) that T2-only resolves at 0.952 and 0.823; conversely, T2-only produces zero-Dice predictions for IOG4, IOG35, and IOG1, all of which the T1-only model segments at 0.865, 0.860, and 0.495 respectively. This complementarity explains why both the ensemble and the dual-channel model outperform either single-modality baseline.

Validation-set overfitting. The absence of a held-out test set introduces a degree of optimistic bias in the reported figures. Because the final model configuration was selected by comparing cross-validation means across trainer variants and epoch counts, the reported performance includes bias proportional to the extent of that configuration search — sometimes referred to as *validation-set overfitting*. This effect is moderate here because the configuration space was intentionally constrained (two training durations, one augmentation variant per trainer family) and because cross-validation averages over multiple folds rather than a single validation split.

Generalisability to external data. The cohort is drawn from a single institution using a consistent acquisition protocol. Reported performance may not transfer directly to patients imaged on different hardware or with different field strengths. Prospective validation on an external multi-centre dataset is therefore a necessary precondition before clinical deployment.

Dataset size. The cohort of 46 patients is small by deep learning standards. While nnU-Net is specifically designed for small medical datasets and its self-configuring pipeline reduces overfitting risk through conservative patch sizes and calibrated augmentation, the wide IQR across models (T1-only: 0.309, T2-only: 0.270; even the best model reaches 0.169) reflects the disproportionate influence of individual hard cases on aggregate statistics. Adding more annotated cases could reduce this spread and improve confidence in the performance estimates.

Label quality and inter-rater variability. Tumour boundaries in ALT are inherently ambiguous: the infiltrative growth pattern and gradual signal transition between tumour and peritumoral fat create regions where experienced radiologists disagree on the correct delineation. Three patients (IOG26, IOG29, IOG35) were flagged as having low inter-rater agreement in boundary placement. These cases are included in both training and evaluation, and their annotations represent a single annotator’s judgement rather than a consensus label. Dice scores for these patients therefore measure agreement with one rater, not with an objective ground truth.

Annotation incompleteness. A related issue is annotation incompleteness. As shown in Figure 18, the model sometimes includes regions in its prediction that are visually consistent with ALT tissue but absent from the ground-truth label. In these cases, the reported Dice is a lower bound on true model accuracy: the metric penalises predictions that may in fact be correct. Patients IOG53, IOG56, and IOG33 achieve Dice above 0.95 on all model variants, corresponding to large, well-defined tumours with clear fat content and minimal imaging artefact. The prevalence of annotation incompleteness across the cohort is unknown, as it would require case-by-case radiological review of model predictions against the original imaging, which was outside the scope of this thesis.

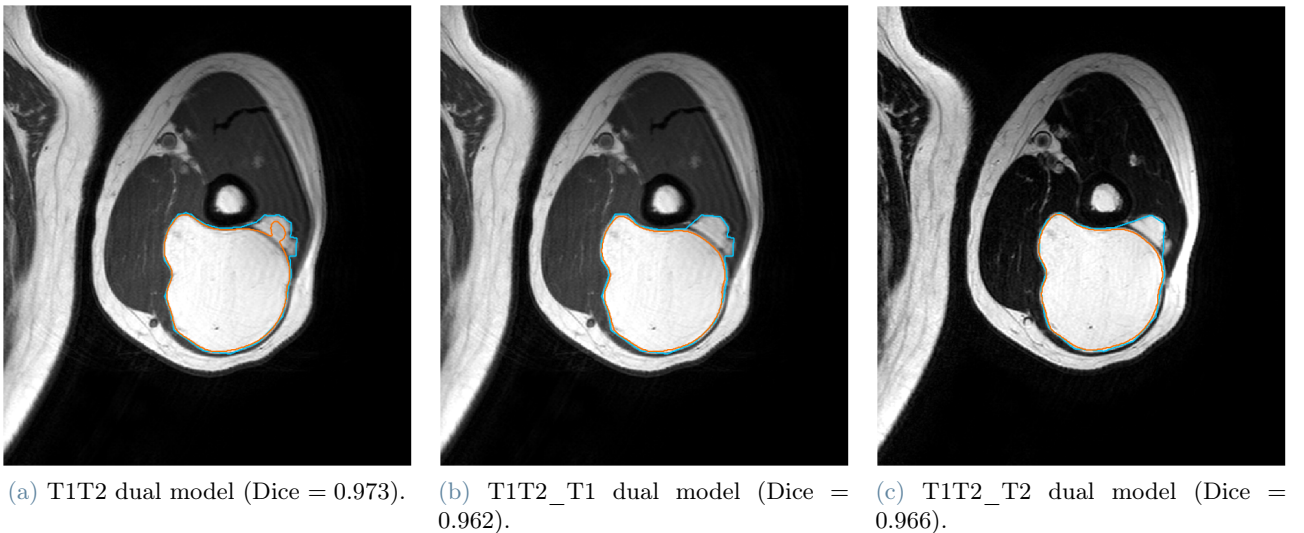


Figure 18: IOG53 high-performance example across all three T1T2 model variants. Blue contour: ground-truth label; red contour: model prediction. The tumour is large, hyperintense on T1 (adipose content), and sharply demarcated; the model segments it accurately regardless of which channel is used. A small prediction region absent from the ground-truth label likely represents an annotation omission, illustrating that Dice scores can underestimate true accuracy in cases of incomplete labelling.

Inherent boundary ambiguity. Patient IOG44 illustrates a qualitatively different source of difficulty. This tumour arises within an anatomical region that is already predominantly fatty, so the tumour parenchyma produces a T1 signal nearly identical to the surrounding tissue. The boundary between tumour and background is imperceptible on inspection: there is no visible intensity step at the annotation edge, and the transition is continuous rather than abrupt. Even an expert annotator must rely on contextual and clinical knowledge rather than local image features to place the contour. Under these conditions, the Dice scores of 0.769 (T1-only baseline) and 0.735 (T1T2 dual) reflect genuine perceptual difficulty rather than model failure, and the reported numbers likely understate the true challenge. Figure 19 illustrates the near-absent boundary contrast.

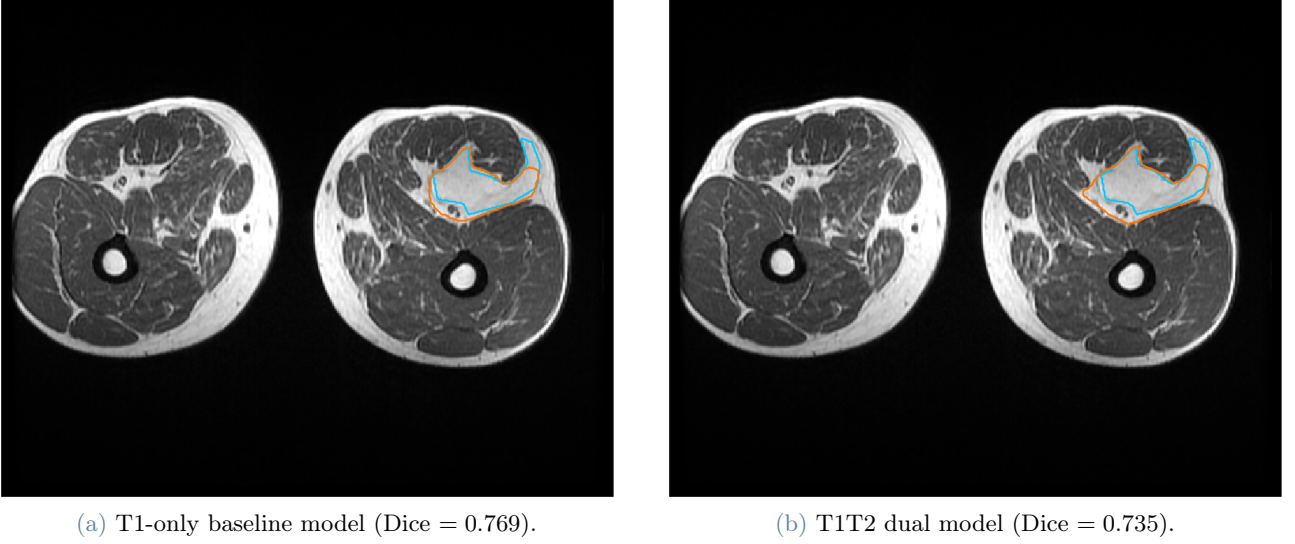


Figure 19: IOG44 boundary ambiguity. Blue contour: ground-truth label; red contour: model prediction. The tumour grows within an already-fatty anatomical region, producing near-zero T1 contrast between tumour and background. No clear intensity boundary is visible at the annotation edge; the model’s moderate Dice reflects genuine perceptual difficulty rather than a segmentation failure.

Acquisition alignment failures. The orientation-mismatched cases identified in Section 4.5 reduce the effective T1T2 dual model evaluation to 40 patients. Patient IOG28 illustrates the failure mode: its T1 volume was acquired in a coronal plane rather than the axial plane used for most cases. The T1 image carries no visible technical anomaly — signal and contrast look normal — but the view is entirely unlike what the model sees during training, yielding Dice = 0. The T2 acquisition is geometrically consistent with its label, however: the T1T2_T2 dual model achieves Dice = 0.795, confirming that the failure is confined to the T1 orientation mismatch. Figure 20 illustrates the contrast.

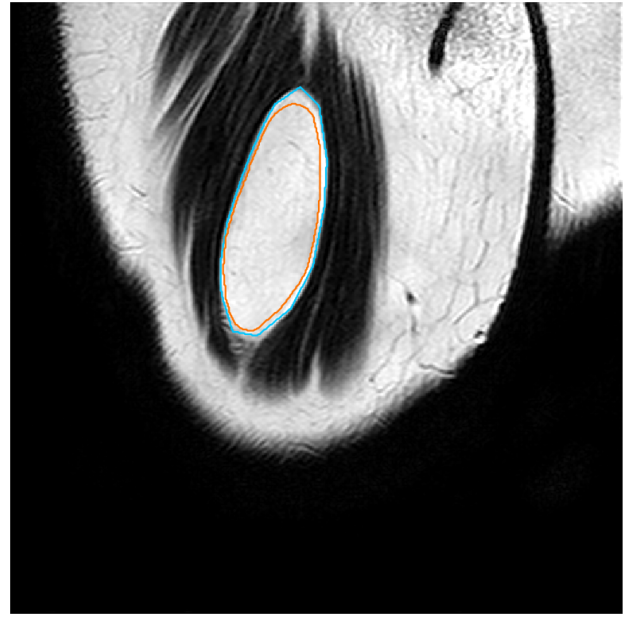
IOG28’s zero-Dice T1 failure suppresses the T1-only baseline mean; had alignment been corrected, the performance gap between baseline models and the T1T2 dual model might be more pronounced. For the eight mismatched cases that were fused by resampling, the approach works well for most (e.g. IOG27, IOG51, IOG55 achieve Dice ≥ 0.90), but IOG2 is an exception: the T1T2 dual model underperforms the T1-only baseline (Dice 0.514 vs 0.944), suggesting residual misalignment or field-of-view differences that resampling did not fully resolve. Alignment correction for the remaining mismatched cases is a tractable engineering improvement that could raise the reported mean and narrow the evaluation gap.

Small-object detection. The pipeline fails entirely on the smallest tumour in the cohort (IOG45, 2.3 cm³; cohort median 80.2 cm³). As Figure 21 shows, the ground-truth contour covers only a handful of voxels in the axial slice; at this scale, reliable segmentation requires dedicated small-object strategies — oversampling with smaller patch sizes or multi-scale detection heads — that standard nnU-Net does not provide. As the pipeline targets large-volume ALT for surgical planning, this limitation affects a narrow edge case; it should nonetheless be communicated to any prospective user.

Modality-specific generalisation failure. Patient IOG38 fails on the T1-only baseline model (Dice = 0) but achieves normal performance on the T2-only baseline and T1T2 dual models. The exact cause is not conclusively identified from visual inspection alone: the T1 image shows no obvious artefact but may present contrast or texture characteristics that fall outside the T1 training distribution. The T2 channel provides



(a) T1 slice — T1-only baseline model (Dice = 0).



(b) T2 slice — T1T2_T2 dual model (Dice = 0.795).

Figure 20: IOG28 orientation mismatch. Blue contour: ground-truth label; red contour: model prediction. The T1 acquisition is in an unusual coronal plane, causing complete failure on the T1-only baseline model. The T2 image is from a different acquisition plane but is geometrically consistent with its label; the T1T2_T2 dual model segments it successfully.

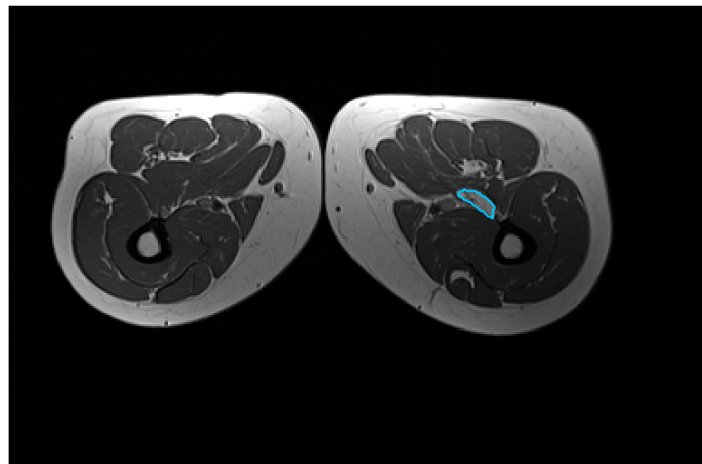
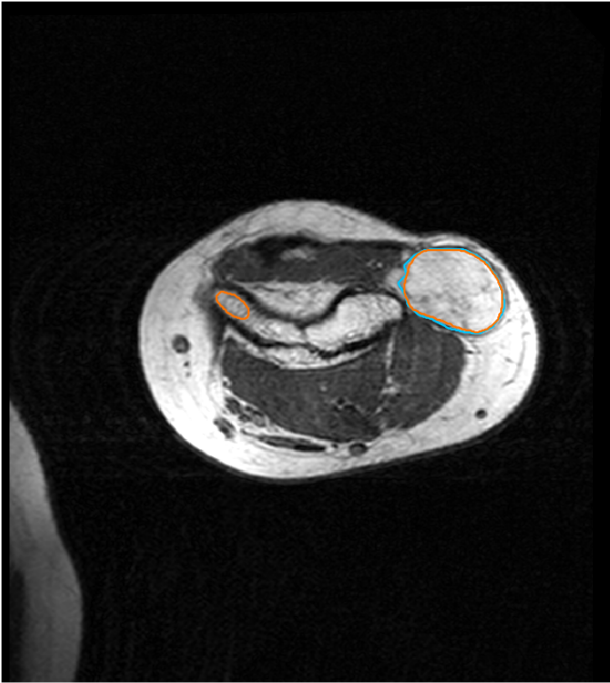


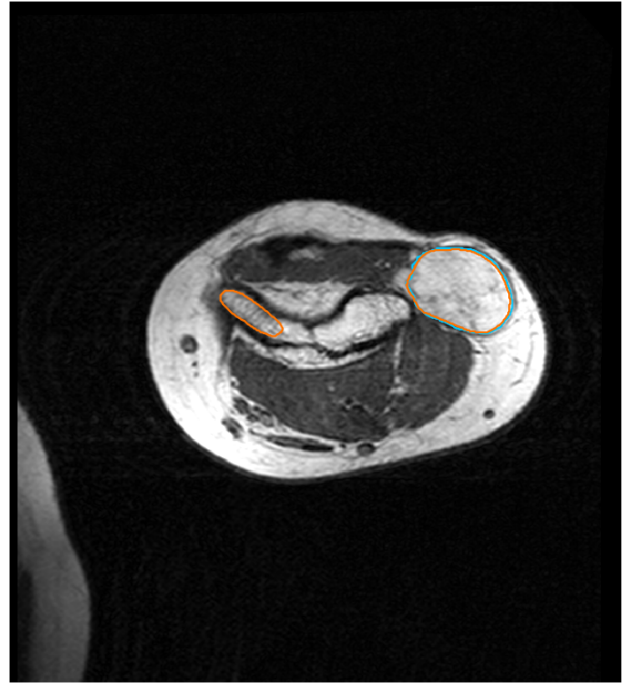
Figure 21: IOG45 failure case — T1 axial slice. Blue contour marks the ground-truth tumour (2.3 cm^3); no red prediction contour is visible because the model outputs an empty mask. The lesion occupies only a handful of voxels in the 3-D volume.

sufficient signal for the T1T2 dual model to succeed, and the T2-only baseline model also succeeds, ruling out anatomy or label quality as the primary cause. This case illustrates how dual-channel input can compensate for single-modality failures without any explicit missing-modality handling mechanism.

Anatomically implausible predictions and the case for post-processing. Patient IOG31 achieves good Dice scores on the baseline models (0.822 on the T1-only baseline, 0.829 on the T2-only baseline), yet visual inspection reveals a systematic error: the model includes a region of cortical bone within the predicted mask. Bone is unambiguously excluded from any soft tissue tumour by anatomy; no tumour tissue can occupy a cortical bone voxel. The inclusion nevertheless occurs because the model operates purely on local intensity patterns and has no mechanism to encode anatomical constraints. On T1-weighted images, cortical bone boundaries can produce intensity gradients that locally resemble tumour margins, misleading the model. This class of error is in principle straightforward to correct. A simple post-processing step that subtracts a bone mask — derived from the T1 image via thresholding or a lightweight bone segmentation model — from the predicted foreground would eliminate these implausible inclusions entirely, at negligible computational cost. The improvement in Dice would be modest on a case-by-case basis, since the bone region constitutes a small fraction of the total predicted volume, but the clinical benefit is disproportionate: a prediction that includes bone would require non-trivial manual correction before it could inform surgical planning. Figure 22 illustrates the bone inclusion on the T1T2 dual model prediction.



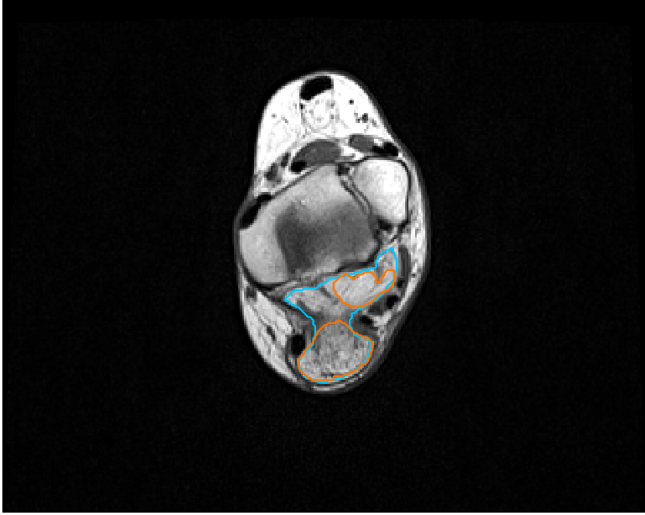
(a) T1-only baseline model (Dice = 0.822).



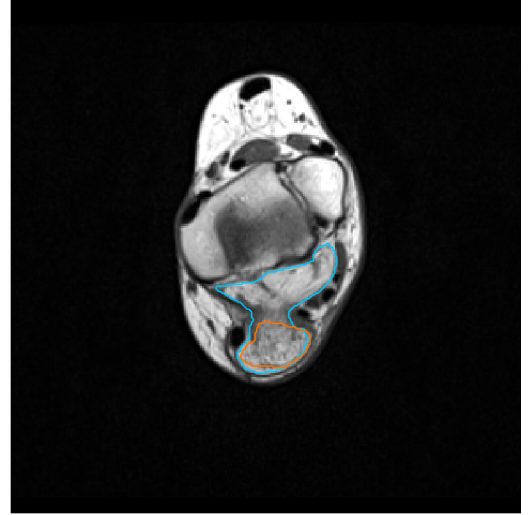
(b) T1T2 dual model (Dice = 0.738).

Figure 22: IOG31 anatomically implausible prediction. Blue contour: ground-truth label; red contour: model prediction. The predicted mask extends into a cortical bone region, which is anatomically impossible for soft tissue tumour. A bone-exclusion post-processing step would eliminate this class of error without requiring model retraining.

Weak T2 contrast and modality interference. Patient IOG1 illustrates a failure mode distinct from those above: the T2 image carries insufficient discriminative information for this case. The T2-only baseline model fails completely (Dice = 0). More critically, the T1T2 dual model performs *worse* than the T1-only baseline model (0.390 vs. 0.495), indicating that the weak T2 signal does not simply go unused — it actively interferes with the model’s decision. The T1-only baseline model, operating on a single channel, produces a partial segmentation; the T1T2 dual model, confronted with a T2 image that contradicts the T1-derived features, is misled into a less accurate prediction. Figure 23 illustrates the contrast between the two outputs. This case was not identified during dataset construction and therefore received no special handling, unlike IOG9 whose T2 channel was explicitly zeroed prior to training. It suggests that a quality-control step screening for



(a) T1-only baseline model (Dice = 0.495).



(b) T1T2 dual model (Dice = 0.390).

Figure 23: IOG1 modality interference. Blue contour: ground-truth label; red contour: model prediction. The T1-only baseline model produces a partial but recognisable segmentation. The T1T2 dual model, which also receives a T2 channel with near-absent contrast, produces a worse prediction, demonstrating that a low-quality input channel can actively degrade performance rather than being ignored.

low-T2-contrast cases — and zeroing the affected channel — should be applied as a standard preprocessing step before T1T2 dual model inference.

6.5. Clinical Implications

A median Dice of 0.913 (IQR 0.169) on the 40 aligned cases is clinically useful for pre-surgical planning. The primary role of automatic segmentation in this context is to delineate the tumour volume and its spatial relationship to adjacent neurovascular structures, informing the surgical approach and resection margin strategy. The T1T2 dual model’s predictions on well-represented cases closely match the expert annotation and would substantially reduce the radiologist’s manual delineation burden: from a full annotation generated from scratch — reported to take up to four hours per case for large-volume ALT [8] — to a quality-controlled correction of an automatically generated contour. Evidence across tumour segmentation workflows consistently shows that correcting an accurate automatic segmentation is faster than generating one from scratch, even when correction effort is non-trivial, freeing radiologist time for higher-complexity clinical decisions.

The model is not, however, a replacement for radiologist review. Three categories of case require human oversight before the output could inform any clinical action. First, cases where the T1 acquisition is geometrically atypical (as with IOG28) or where the T2 channel is near-zero (as with IOG9) fall outside the training distribution and may produce unreliable predictions. Second, very small tumours are not reliably detected. Third, in cases where annotation completeness is uncertain, the model may detect real ALT tissue not included in the reference label; a reviewing radiologist is needed to adjudicate whether model-positive, GT-negative regions are true tissue detections or false alarms.

A practical deployment would therefore require a radiologist-in-the-loop workflow in which the automatic segmentation is presented as a clearly labelled draft, with the radiologist confirming, correcting, or rejecting the contour before it influences any treatment decision. This workflow also provides a natural mechanism for accumulating additional high-quality annotations over time, directly addressing the dataset size limitation identified above.

6.6. Comparison with the Literature

Direct numerical comparison with published results in ALT segmentation is constrained by the scarcity of public benchmarks and the strong dependence of reported Dice on dataset-specific factors: tumour size distribution, annotation protocol, MRI scanner, and whether evaluation is performed on cross-validation folds or independent test sets.

The closest comparable study is Liu et al. [8], who described a Super Learner ensemble of 2D U-Nets for lipomatous tumour segmentation in T1-weighted or proton-density MRI. Evaluated on an independent test set of 20 patients drawn from a multi-centre cohort of 185, their ensemble achieved a mean DSC of 0.803 ± 0.184 across five tumour locations (limbs and trunk). Two differences limit direct numerical comparison. First, their pipeline operates slice-by-slice in 2D, foregoing inter-slice context; the authors identify 3D modelling as an explicit direction for future work. Second, their cohort covers the full spectrum of lipomatous tumours — benign lipomas through malignant liposarcomas — and includes trunk lesions, whereas this thesis addresses limb-confined ALT exclusively. Within these caveats, the T1T2 dual model achieves a median DSC of 0.913 and a mean of 0.823 — the latter comparable to, and slightly above, the Liu et al. ensemble mean — while operating in 3D and on T1+T2 dual-channel input rather than a single modality.

More broadly, reported Dice scores for soft tissue tumour segmentation in MRI span a wide range depending on task complexity. Organ segmentation tasks — where shape priors and anatomical location are stable — routinely achieve Dice above 0.90. Soft tissue tumours lack these priors: they appear at arbitrary locations within the limb, vary from a few cubic centimetres to several litres in volume, and their boundaries merge with surrounding structures of similar signal intensity. In this context, a median Dice of 0.913 on the 40 aligned cases compares favourably with results reported on comparably challenging tumour segmentation tasks, and particularly so given the small training cohort of 46 patients. The bimodal per-case distribution — most predictions either near-perfect or near-zero — is a well-documented characteristic of small-dataset medical segmentation [6] and is expected to resolve partially as the cohort grows, with the cluster of intermediate-difficulty cases becoming more populated and the model’s ability to handle outlier presentations improving.

7. Conclusions

7.1. Summary of Contributions

This thesis presented a fully automatic 3D segmentation pipeline for Atypical Lipomatous Tumor in limb MRI, addressing a task for which no validated automated solution previously existed. Manual delineation of ALT is time-consuming, subject to inter-operator variability, and a practical bottleneck in pre-surgical planning workflows. The pipeline developed here eliminates the cold-start burden, providing a complete draft segmentation that a radiologist can review and correct rather than generate from scratch.

The pipeline is built on nnU-Net v2 [6] and constitutes, to the best of our knowledge, the first fully automatic 3D segmentation system specifically developed and evaluated for ALT in limb MRI. Three custom components adapt the nnU-Net framework to the specific challenges of this task.

Gibbs ringing augmentation. The `nnUNetTrainerALT_os033_gibbs` trainer inserts a random k-space truncation step into the nnU-Net DA5 augmentation pipeline. During training, each input channel is independently corrupted by simulated Gibbs ringing — a k-space truncation to a random central fraction in $[0.5, 0.9]$ applied with probability $p = 0.15$ — preventing the model from exploiting the boundary ringing pattern as a localisation cue and improving robustness to scanner-specific acquisition variability. Ablation experiments confirmed that the Gibbs-augmented trainer outperformed equivalent configurations without this augmentation.

Custom trainer family. The `nnUNetTrainerALT` family introduces per-sample Dice loss (`batch_dice = False`), a fixed foreground oversampling rate, and an extended training schedule of 350 epochs. The 350-epoch schedule was determined empirically after observing that preliminary 250-epoch runs had not converged on this small dataset, and all final reported results use this extended schedule.

Dual-channel T1+T2 pipeline. T1 and T2 acquisitions are merged into a two-channel input, enabling the model to learn cross-modal correlations during training rather than combining independent predictions at inference time.

Performance. Systematic five-fold cross-validation over 46 patients established the following ranking. The T1-only baseline model achieves a median Dice of 0.884 (IQR 0.309) and the T2-only baseline model 0.851 (IQR 0.270); the difference is not statistically significant ($p = 0.346$). The T1T2 ensemble model improves on both, reaching a median of 0.883 (IQR 0.241), but is significantly outperformed ($p = 0.011$) by the T1T2 dual model trained jointly on T1 and T2 input. The T1T2 dual model on geometrically aligned cases ($n = 40$) achieves a median Dice of **0.913** (IQR 0.169), with a narrower IQR than the baseline models — indicating more consistent predictions across the patient cohort. The performance advantage of joint training over output-level fusion is attributed to the T1T2 dual model’s ability to learn cross-modal correlations during training, which are inaccessible to independently optimised models combined at inference time.

7.2. Future Work

Several directions would strengthen and extend the results presented in this thesis.

Larger annotated cohort. The most direct path to improved and more reliable performance is a larger dataset. The high fold-to-fold variance observed throughout this work is a consequence of 46 patients across five folds, where each validation set contains approximately nine patients. Expanding the cohort to 100–150 cases would substantially reduce variance, allow a dedicated held-out test set, and support more granular ablation studies — in particular, a clean comparison of Gibbs vs. non-Gibbs trainers at matched epoch counts on a fixed validation partition.

Held-out test set. All results in this thesis rely on cross-validation as a proxy for generalisation. A natural next step is to reserve a held-out set of patients at the point of dataset expansion, enabling an unbiased final evaluation of the selected pipeline and removing the residual optimistic bias introduced by configuration selection on the CV folds.

Multi-centre validation. The dataset is drawn from a single institution using a consistent acquisition protocol. External validation on data from different scanners and field strengths is a necessary precondition for clinical deployment. Multi-centre data would also reveal whether Gibbs augmentation generalises as a domain robustness strategy to acquisition variability beyond what is present in the current cohort.

Semi-supervised and annotation-efficient learning. ALT annotation is expensive and requires specialist expertise. Semi-supervised methods that leverage unlabelled MRI volumes — of which far more exist — could substantially expand the effective training set without a proportional increase in annotation cost. Approaches such as mean-teacher training or consistency regularisation are natural extensions of the nnU-Net backbone used here.

Annotation quality and consensus labels. The three patients with low inter-rater agreement (IOG26, IOG29, IOG35) are currently evaluated against a single annotator’s label. Obtaining a second annotation and computing a STAPLE or majority-vote consensus label for these cases would provide a more reliable evaluation target and a more accurate characterisation of the model’s agreement with human performance. More broadly, a systematic study of annotation incompleteness — the phenomenon observed for IOG53 where the model detects real tissue absent from the ground truth — would require radiological review of model predictions against the raw images and would clarify how much the reported Dice underestimates true accuracy.

References

- [1] Tyler J Bradshaw, Zachary Huemann, Junjie Hu, and Arman Rahmim. A guide to cross-validation for artificial intelligence in medical imaging. *Radiology: Artificial Intelligence*, 5(4):e220232, 2023.
- [2] M Brisson, T Kashima, D Delaney, R Tirabosco, A Clarke, S Cro, A M Flanagan, and P O’Donnell. MRI characteristics of lipoma and atypical lipomatous tumor/well-differentiated liposarcoma: Retrospective comparison with histology and MDM2 gene amplification. *Skeletal Radiology*, 42(5):635–647, 2013.
- [3] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 424–432. Springer, 2016.
- [4] Clyde M Gaskin and Clyde A Helms. Lipomas, lipoma variants, and well-differentiated liposarcomas (atypical lipomas): Results of mri evaluations of 126 consecutive fatty masses. *American Journal of Roentgenology*, 182(3):733–739, 2004.
- [5] Pankaj Gupta, Thomas A Potti, Scott D Wuertzer, Leon Lenchik, and David A Pacholke. Spectrum of fat-containing soft-tissue masses at mr imaging: the common, the uncommon, the characteristic, and the sometimes confusing. *Radiographics*, 36(3):753–766, 2016.
- [6] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

- [7] Elias Kellner, Bibek Dhital, Valerij G Kiselev, and Marco Reisert. Gibbs-ringing artifact removal based on local subvoxel-shifts. *Magnetic Resonance in Medicine*, 76(5):1574–1581, 2016.
- [8] Chih-Chieh Liu, Yasser G Abdelhafez, S Paran Yap, Francesco Acquafredda, Silvia Schirò, Andrew L Wong, Dani Sarohia, Cyrus Bateni, Morgan A Darrow, Michele Guindani, Sonia Lee, Michelle Zhang, Ahmed W Moawad, Quinn Kwan-Tai Ng, Layla Shere, Khaled M Elsayes, Roberto Maroldi, Thomas M Link, Lorenzo Nardo, and Jinyi Qi. AI-based automated lipomatous tumor segmentation in mr images: Ensemble solution to heterogeneous data. *Journal of Digital Imaging*, 36:1049–1059, 2023.
- [9] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.
- [10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

Abstract in lingua italiana

Il Tumore Lipomatoso Atipico (ALT, dall'inglese *Atypical Lipomatous Tumor*) è il sarcoma dei tessuti molli degli arti più comune. La delimitazione tridimensionale accurata del tumore su immagini MRI è fondamentale per la pianificazione chirurgica, tuttavia la segmentazione manuale richiede fino a quattro ore per caso ed è soggetta a variabilità inter-operatore. Non è stata precedentemente riportata alcuna pipeline completamente automatica validata per la segmentazione 3D dell'ALT.

Questa tesi presenta un framework di segmentazione basato su nnU-Net v2, valutato tramite validazione incrociata a cinque fold su un dataset privato di 46 pazienti con volumi MRI pesati in T1 e T2. Tre componenti personalizzati adattano il framework a questo compito: un trainer con augmentation per il ringing di Gibbs, che simula artefatti da troncamento dello spazio-k durante l'addestramento per migliorare la robustezza alla variabilità del sistema di acquisizione; una famiglia di trainer personalizzata con perdita Dice per campione e un programma di addestramento calibrato sulla coorte ridotta; e un modello a doppio canale T1+T2 che permette l'apprendimento congiunto delle correlazioni inter-modalità. Vengono confrontate quattro configurazioni: T1-only (Dice medio 0,731), T2-only (0,711), ensemble union post-hoc (0,768) e il modello a doppio canale, che ottiene un Dice medio di 0,823 e una mediana di 0,913 sui 40 pazienti i cui volumi T1 e T2 condividono un'orientazione di acquisizione compatibile, superando significativamente tutte le alternative ($p = 0,011$, test di Wilcoxon con ranghi con segno). Gli esperimenti di ablazione hanno confermato che l'augmentation di Gibbs migliora costantemente rispetto alle configurazioni equivalenti prive di essa. I risultati dimostrano che l'addestramento congiunto a doppio canale è il principale fattore di prestazione e che nnU-Net 3D con augmentation di Gibbs produce segmentazioni ALT clinicamente utili per la pianificazione pre-chirurgica.

Parole chiave: segmentazione di immagini mediche, nnU-Net, tumore lipomatoso atipico, MRI, data augmentation, ringing di Gibbs

Acknowledgements

I would like to express my sincere gratitude to Professor Corino and Professor Meri for their invaluable guidance throughout this thesis. The course in Applied AI for Bioinformatics that they teach was one of the most interesting and useful I encountered during my studies — Professor Corino's focus on theoretical foundations and Professor Meri's hands-on laboratory sessions gave me exactly the tools I needed to tackle this work. Their support made developing a thesis with them not only a natural choice, but a genuinely enjoyable experience, and I am proud of what we achieved together.

To my friends in Milan: thank you for celebrating every small triumph with me and for keeping a little piece of Colombia alive here. That warmth made all the difference.

To my brother, who always found a way to help even across very different fields, and to my sister, who has always been there to listen — thank you both.

I would also like to acknowledge Politecnico di Milano for providing an excellent academic environment and the resources that made this research possible.

Finally, I wish to thank the members of the thesis committee for their time and attention in evaluating this work.