



POLITECNICO
MILANO 1863

**SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE**

EXECUTIVE SUMMARY OF THE THESIS

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

LAUREA MAGISTRALE IN HIGH PERFORMANCE COMPUTING ENGINEERING

Author: LUIS FELIPE EPIA REALPE

Advisor: PROF. VALENTINA CORINO

Co-advisor: MERI FERRETTI

Academic year: 2025–2026

1. Introduction

Atypical Lipomatous Tumor (ALT) is the most common soft tissue sarcoma of the extremities. Wide surgical excision with negative margins is the primary treatment, and accurate three-dimensional delineation of the tumour on MRI is a prerequisite for planning the resection. Manual segmentation by expert radiologists is time-consuming — up to four hours per large-volume case — and subject to inter-operator variability [1, 4]. No validated fully automatic 3D segmentation pipeline for ALT previously existed; published approaches either address broader lipomatous tumour categories or operate slice-by-slice in 2D, falling short of the volumetric setting required for pre-surgical planning [4].

ALT segmentation directly conditions therapy planning. The resection margin must encircle the entire tumour volume in three dimensions, and an incomplete or inaccurate delineation risks either positive margins — associated with local recurrence — or unnecessarily wide excision that sacrifices healthy tissue. A reliable automated draft would reduce radiologist burden from several hours to a quality-controlled correction task, allowing the same

specialist capacity to serve a larger patient population. It would also reduce inter-operator variability: studies report non-trivial disagreement between experienced annotators on large, irregularly shaped lesions of the type common in ALT [1].

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**, built on the nnU-Net v2 self-configuring framework [2] and adapted to the specific challenges of this task through three custom components.

2. Segmentation Challenges

Automatic 3D segmentation of ALT involves several compounding difficulties.

Small dataset. The annotated cohort comprises 46 patients — realistic for a rare tumour at a single institution, but imposing strict constraints on model complexity and evaluation reliability. With five-fold cross-validation, each validation fold contains approximately nine patients, producing high fold-to-fold variance in all reported metrics.

Tumour heterogeneity. ALT tumours span three orders of magnitude in volume (from a few

cm³ to several hundred cm³), grow around neurovascular bundles rather than forming compact masses, and vary internally between purely lipomatous and fibrous or myxoid regions. This variability makes both detection and precise boundary delineation challenging.

Gibbs ringing artifacts. k-space truncation at high-contrast fat-muscle interfaces produces oscillatory boundary artifacts whose severity varies across scanners and acquisition protocols [3]. A model that exploits these artifacts as localisation cues will fail to generalise across acquisition settings.

Multimodal complementarity. T1-weighted images provide the fat-specific hyperintense signal that distinguishes the tumour from surrounding muscle. T2-weighted images reveal internal heterogeneity and peritumoral oedema. T1 and T2 acquisitions in this dataset were not always acquired in compatible orientations, requiring case-by-case geometric alignment before a dual-channel pipeline could be applied. Combining both modalities during training — rather than aggregating independent predictions at inference — allows the model to learn cross-modal correlations inaccessible to independently optimised models.

3. Methods

3.1. Dataset and Evaluation Protocol

The dataset comprises 46 patients with both T1- and T2-weighted limb MRI volumes and expert annotations. Dataset construction required handling heterogeneous acquisition orientations: of the 12 patients with T1/T2 orientation mismatches, eight were fused by resampling the T2 onto the T1 grid, while three (IOG7, IOG28, IOG45) could not be fused because their acquisitions also differed in anatomical coverage, making the merged volume degenerate. Three further cases (IOG26, IOG29, IOG35) were excluded from the fused dataset due to low inter-rater annotation agreement, to avoid propagating uncertain labels into the dual-channel ground truth. These six cases are instead contributed as modality-specific derived cases to the single-channel datasets. The T1T2 dual model is therefore evaluated on the 40 geometrically aligned cases, while the baseline models use all available cases (T1: $n = 46$; T2:

$n = 45$), where n denotes the sample size.

All models are evaluated under five-fold patient-level cross-validation. The primary metric is the Dice Similarity Coefficient (DSC); the Intersection over Union (IoU) is reported as a secondary metric. Statistical comparisons use the Wilcoxon Signed-Rank test, paired on the case intersection of the two models being compared ($p < 0.05$: *, $p < 0.01$: **).

3.2. nnU-Net Framework

nnU-Net v2 [2] is a self-configuring segmentation framework that automatically determines pre-processing, patch size, batch size, network architecture, and training schedule from dataset fingerprint analysis. All models in this thesis use the 3d full-resolution configuration with `nnUNetPlans` auto-generated architecture and the DA5 augmentation pipeline. The network is a 3D encoder-decoder with skip connections and residual blocks, trained with a combination of Dice and cross-entropy loss and deep supervision on intermediate decoder outputs. Three custom components extend the default configuration.

3.3. Gibbs Ringing Augmentation

A custom trainer inserts a `GibbsRingingTransform` into the DA5 pipeline. For each input channel, with probability $p = 0.15$, the k-space representation is truncated by retaining only a random central fraction $\alpha \sim \mathcal{U}(0.5, 0.9)$:

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

where M_α zeros all k-space coefficients outside the central α -fraction. Lower α produces stronger ringing. By exposing the model to a range of synthetic artifact intensities at training time, the augmentation prevents reliance on boundary ringing as a localisation cue and improves robustness to scanner-specific acquisition variability. Ablation experiments confirmed that the Gibbs-augmented trainer consistently outperformed equivalent configurations trained without it.

3.4. Small-Dataset Training Adaptations

Two modifications adapt nnU-Net to the small cohort. First, the Dice loss is computed *per sample* rather than *per batch* (`batch_dice=False`),

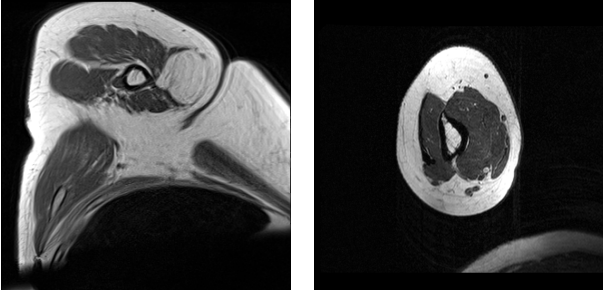


Figure 1: Representative T1-weighted limb MRI slices showing Gibbs ringing. Oscillating intensity bands (visible at fat–muscle boundaries) vary in severity across acquisitions and motivate the Gibbs augmentation strategy.

ensuring that the $660\times$ tumour-volume range does not concentrate gradients on the largest lesions. Second, the training schedule is fixed at 350 epochs — validated empirically via convergence monitoring — replacing the 1000-epoch default calibrated for much larger datasets.

3.5. Model Configurations

Four configurations are evaluated:

T1-only baseline model and **T2-only baseline model** are trained independently on a single modality. These establish the single-modality performance floor.

T1T2 ensemble model applies a voxel-level union of the T1-only and T2-only predictions at inference time, without any joint training.

T1T2 dual model feeds T1 and T2 as two input channels simultaneously. nnU-Net adapts the first convolutional layer to the two-channel input automatically. The model is trained end-to-end on both modalities, learning cross-modal correlations rather than combining independent predictions.

3.6. 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; at 350 epochs per fold this yields approximately 8.1 hours per fold and 40 hours for the complete five-fold run per dataset. Training used PyTorch via nnU-Net v2 with CUDA acceleration;

4. Results

4.1. Single-Modality Baselines

The T1-only baseline model 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 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 baselines is not statistically significant (Wilcoxon, $p = 0.346$, $n = 45$), indicating broadly comparable single-modality performance. Both distributions exhibit a wide IQR, driven by a small number of zero-Dice failures and anatomically challenging cases. Despite similar aggregate performance, the two baselines fail on largely non-overlapping subsets of the cohort.

4.2. T1T2 Dual Model

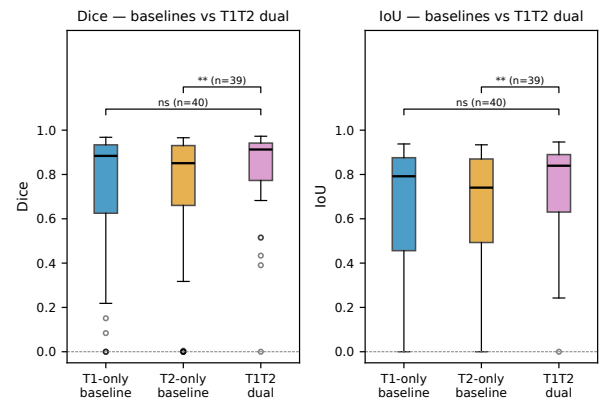


Figure 2: Dice (left) and IoU (right) for the T1-only baseline ($n = 46$), T2-only baseline ($n = 45$), and T1T2 dual model ($n = 40$). Brackets compare each baseline against the dual model; two-sided Wilcoxon test (ns: not significant, **: $p < 0.01$).

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 and +6.2 points over the T2-only baseline. The narrower IQR (0.169 versus 0.309 and 0.270) confirms more consistent predictions across the patient cohort. Figure 2 compares the dual model against both baselines.

4.3. T1T2 Ensemble Model

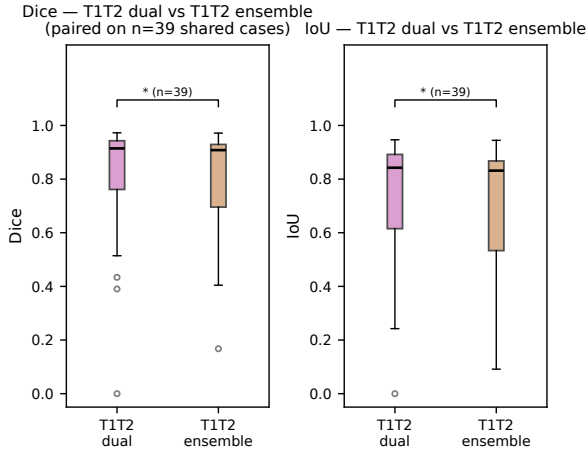


Figure 3: Dice (left) and IoU (right) distributions for the T1T2 dual model ($n = 40$) and T1T2 ensemble model ($n = 45$), paired on their $n = 39$ shared cases. Significance bracket: Wilcoxon Signed-Rank test ($p = 0.011$).

The T1T2 ensemble model (voxel-level union of the T1-only and T2-only predictions) achieves a median Dice of 0.883 (IQR 0.241) and median IoU of 0.790 over 45 cases, improving on both baseline models but falling below the T1T2 dual model. The performance gap between the ensemble and the dual model is statistically significant ($p = 0.011$, $n = 39$; Figure 3).

4.4. Ablation Summary

Table 1 collects aggregate performance across all configurations. The T1T2 dual model achieves the best median Dice and narrowest IQR across all metrics. Per-case Dice scores for the T1T2 dual model (Figure 5) show a bimodal distribution, with 28 of 40 cases exceeding Dice = 0.80. Tumour size and Dice score are not significantly correlated (Pearson $r = 0.059$, $p = 0.719$ on log-transformed volumes), indicating that the model does not systematically fail on small or large tumours.

Figure 4 shows patient IOG6, illustrating how joint dual-channel training outperforms both single-modality baselines and the output-level ensemble.

5. Discussion

Performance interpretation and failure modes. The two single-modality baselines per-

Table 1: Five-fold cross-validation results. T1T2 dual model on $n = 40$ aligned cases; baselines and ensemble on all available cases. Bold: best 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 ensemble	0.883	0.241	0.790	0.342
T1T2 dual	0.913	0.169	0.839	0.259

form comparably ($p = 0.346$), yet the T1T2 dual model significantly outperforms the T2-only baseline ($p = 0.008$) but not the T1-only baseline ($p = 0.572$) — reflecting complementary failure sets: cases where T1 fails (e.g. IOG28, zero Dice due to orientation mismatch) are precisely those where T2 succeeds, and vice versa. The per-case distribution is bimodal: a cluster of high performers (≥ 0.80) and a low-Dice tail driven by acquisition mismatch (IOG28), extreme small volume (IOG45, 2.3 cm^3), modality interference (IOG1), and ambiguous boundaries (IOG31). Segmentation difficulty is not correlated with tumour size ($r = 0.059$, $p = 0.719$), confirming that failure cases span the full volume range. The closest comparable published result is the Super Learner ensemble of Liu et al. [4] (mean DSC 0.803 ± 0.184 , 185 patients, multi-centre cohort).

Why joint dual-channel training outperforms fusion. The +3.0 percentage-point gap in median Dice between the T1T2 dual model and the T1T2 ensemble cannot be attributed to additional data access: both approaches draw on the same T1 and T2 signal. The T1-only model already handles the majority of aligned cases well, whereas the T2-only model has a heavier low-Dice tail that joint training corrects. Cross-modal gradients propagated jointly through both input streams allow the encoder to learn these correlations; output-level union cannot.

Gibbs augmentation as a domain robustness mechanism. The Gibbs artifact is correlated with tumour boundary presence in the training data: ringing bands appear at fat-muscle interfaces that coincide with tumour margins. Without augmentation, the model can exploit boundary ringing as a localisation cue —

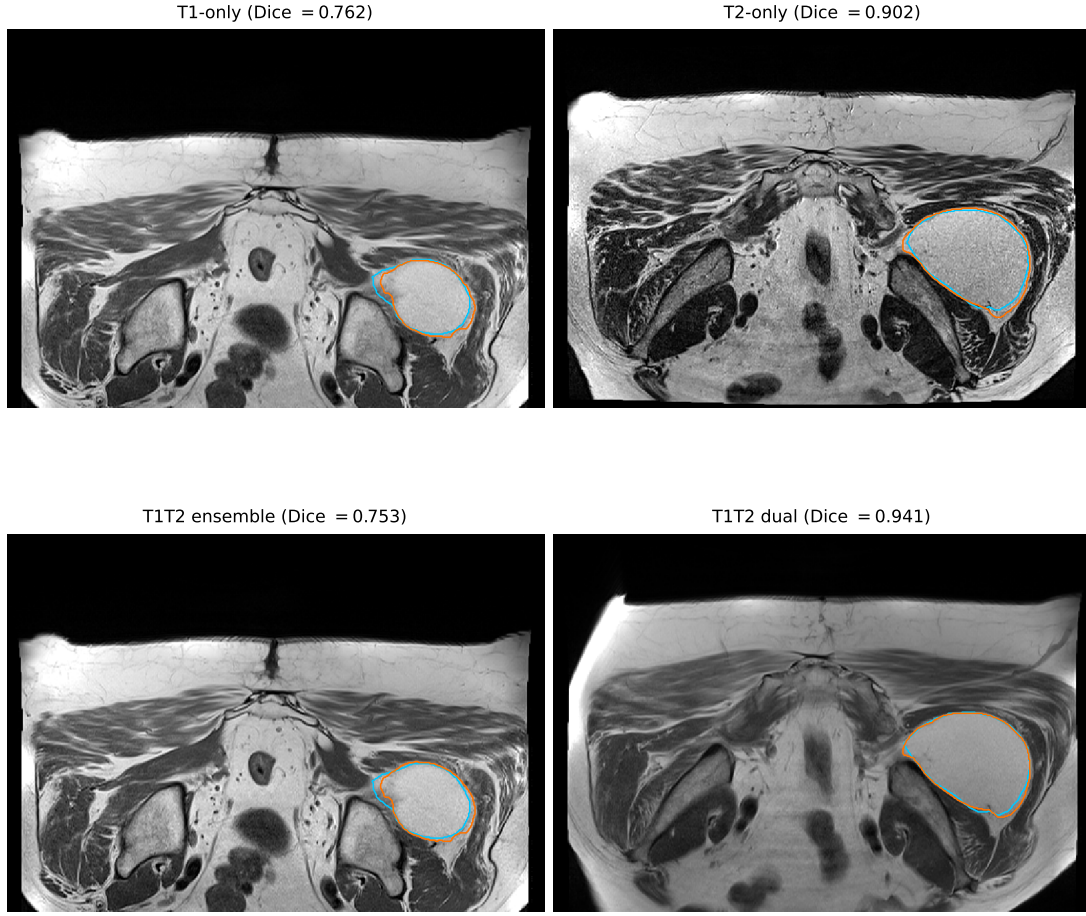


Figure 4: Qualitative ablation on patient IOG6 (axial slice 22). **Blue**: ground truth contour; **Orange**: model prediction contour. Dice scores: T1-only 0.762, T2-only 0.902, T1T2 ensemble 0.753, T1T2 dual 0.941. The ensemble underperforms T2-only because the union of both predictions expands false-positive regions; joint dual-channel training recovers the best boundary match.

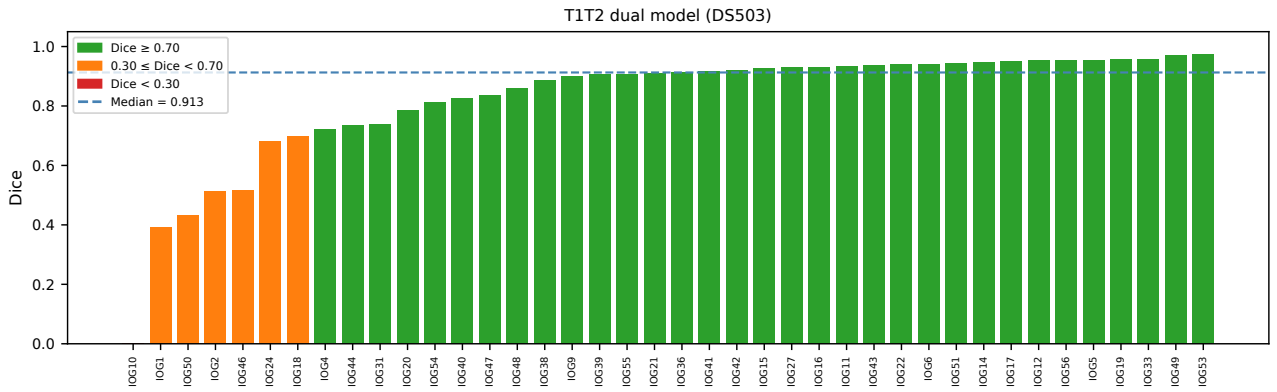


Figure 5: Per-case Dice scores for the T1T2 dual model ($n = 40$), sorted in ascending order. Colour indicates performance tier: green ≥ 0.70 , orange $0.30 \leq \text{Dice} < 0.70$, red < 0.30 . Dashed line marks the median (0.913).

a spurious correlation that collapses when a test scan is acquired with different k-space parameters or at a different field strength. By simulating a range of ringing intensities at training time (Eq. 1), the augmentation forces the model to rely on signal characteristics that are robust to truncation, primarily the T1 hyperintensity of adipose tissue relative to muscle.

Limitations. The cohort of 46 patients is small, resulting in high fold-to-fold variance and limiting statistical power. All results rely on five-fold cross-validation from a single institution. The dataset originates from a single centre with a consistent acquisition protocol; generalisation to different scanners, field strengths, or clinical workflows is unconfirmed. Three cases (IOG26, IOG29, IOG35) have low inter-rater agreement and are currently evaluated against a single annotator’s label, which underestimates model accuracy when the model detects real tissue absent from the ground truth.

6. Conclusions

The experiments demonstrate that all three custom components contribute measurably to the final result. Gibbs augmentation improves robustness across acquisition settings; per-sample Dice loss and the calibrated training budget ensure convergence on the small cohort; and joint dual-channel training captures cross-modal features that neither single-modality models nor post-hoc fusion can recover.

The T1T2 dual model achieves a median Dice of **0.913** (IQR 0.169) and median IoU of 0.839 on geometrically aligned cases ($n = 40$), significantly outperforming both the single-modality baselines and the post-hoc ensemble ($p = 0.011$). The narrower IQR relative to the baseline models (0.309 and 0.270) reflects more consistent predictions across the cohort, with 28 of 40 cases exceeding Dice = 0.80.

All reported figures are derived from five-fold patient-level cross-validation on a single-institution cohort. Generalisation to patients imaged at different institutions, on different scanners, or at different field strengths remains unconfirmed and constitutes the primary precondition for clinical translation.

Future work. Several directions would strengthen and extend these results. *Larger cohort:* expanding the population would reduce

fold-to-fold variance, allow a dedicated held-out test set, and enable clean ablation studies at matched epoch counts. *Multi-centre validation:* external evaluation on data from different scanners and field strengths is a necessary precondition for clinical deployment and would reveal whether Gibbs augmentation generalises as a domain robustness strategy. *Semi-supervised learning:* ALT annotation requires specialist expertise; methods such as mean-teacher training or consistency regularisation could leverage the large volume of unlabelled limb MRI in clinical archives to substantially expand the effective training set. *Consensus annotations:* obtaining a second annotation for the three low-agreement cases and computing STAPLE consensus labels would provide more reliable evaluation targets and clarify how much the reported Dice underestimates true accuracy.

References

- [1] Pankaj Gupta, Thomas A Potti, Scott D Wuertzer, Leon Lenchik, and David A Pachtolke. Spectrum of fat-containing soft-tissue masses at mr imaging. *Radiographics*, 36(3):753–766, 2016.
- [2] 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.
- [3] 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.
- [4] 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.