Modernizing Convolutional Segmentation:

A ConvNeXt-UNet Architecture for Multi-Modal Medical Image Analysis

Master's Degree Programme in Health Technology

Department of Computing, Faculty of Technology

Master of Science in Technology Thesis

Author:

Mishqat Maqbool

Supervisors:

Dr. Oona Rainio

Dr. Muhammad Irfan

Master of Science in Technology Thesis

Department of Computing, Faculty of Technology

University of Turku

Programme: Master's Programme in Health Technology

Author: Mishqat Maqbool

Title: Modernizing Convolutional Segmentation: A ConvNeXt-UNet Architecture for Multi-Modal Medical Image Analysis

Number of pages: 93

Date: June 2026

Abstract

Manual tumour delineation in radiation oncology is time-intensive and prone to inter-observer variability. While deep learning offers a solution, the widely used U-Net is often limited by 3x3 kernels, which provide an insufficient effective receptive field for diffuse targets. This thesis evaluates an improved ConvNeXt-UNet architecture featuring 7x7 depthwise convolutions, an inverted bottleneck structure, and Layer Normalization to maximize spatial context.

Two tasks were studied: Task A (Brain tumour segmentation, Magnetic Resonance Imaging [MRI]) and Task B (Head and neck cancer segmentation, Computed Tomography [CT]). Evaluation was conducted via patient-wise 5-fold cross-validation. The ConvNeXt-UNet significantly outperformed the U-Net in Task A, achieving a mean Dice Similarity Coefficient (DSC) of 0.9328 compared to 0.5496. However, in Task B, the U-Net proved superior (DSC: 0.3000 vs. 0.2558). These findings suggest that architectural advantages may depend on the imaging modality: while large kernels benefit MRI, they may introduce noise in lower-contrast CT scans. The results emphasize the necessity of multi-modal benchmarking to establish the generalizability of medical imaging architectures.

Keywords: medical image segmentation, ConvNeXt, U-Net, deep learning, FLAIR MRI, CT, cross-modal evaluation.

AI Disclosure Statement

Artificial intelligence tools, specifically Claude (Sonnet 4.5, Anthropic) and Google Gemini (3.5 Flash), were used during the preparation of this thesis as writing and figure creation assistance aids. Claude was used to support literature organization, improve language clarity, refine academic phrasing, and assist with structuring arguments. Google Gemini was used to assist in the creation of selected figures and visual illustrations. All research ideas, experimental design decisions, data analysis, result interpretation, and conclusions are the author's own work. The author takes full responsibility for the accuracy and integrity of all content presented in this thesis.

List of Abbreviations

The following abbreviations are used throughout this thesis.

Abbreviation	Full Form
AI	Artificial Intelligence
BCE	Binary Cross-Entropy
BN	Batch Normalization
BraTS	Brain Tumour Segmentation (Challenge)
CI	Confidence Interval
CNN	Convolutional Neural Network
CSF	Cerebrospinal Fluid
CT	Computed Tomography
CV	Cross-Validation
DSC	Dice Similarity Coefficient
DW	Depthwise
ERF	Effective Receptive Field
FCMAE	Fully Convolutional Masked Autoencoder
FLAIR	Fluid-Attenuated Inversion Recovery
FLOP	Floating Point Operation
GELU	Gaussian Error Linear Unit
GPU	Graphics Processing Unit
GTVn	Gross Tumour Volume — Nodal

GTVp	Gross Tumour Volume — Primary
HECKTOR	Head and Neck Tumour Segmentation and Outcome Prediction (Challenge)
HNSCC	Head and Neck Squamous Cell Carcinoma
HPV	Human Papillomavirus
HU	Hounsfield Unit
IDH	Isocitrate Dehydrogenase
IMRT	Intensity-Modulated Radiation Therapy
LGG	Low-Grade Glioma
LN	Layer Normalization
MDACC	MD Anderson Cancer Centre
MICCAI	Medical Image Computing and Computer-Assisted Intervention
MRI	Magnetic Resonance Imaging
NIfTI	Neuroimaging Informatics Technology Initiative
nnU-Net	No-New-Net (self-configuring U-Net)
OAR	Organ at Risk
PET	Positron Emission Tomography
ReLU	Rectified Linear Unit
RF	Receptive Field
SD	Standard Deviation
SGD	Stochastic Gradient Descent
T1	T1-Weighted MRI Sequence
T1ce	T1 Contrast-Enhanced MRI Sequence

T2	T2-Weighted MRI Sequence
TCIA	The Cancer Imaging Archive
TCGA	The Cancer Genome Atlas
TCGA-LGG	The Cancer Genome Atlas — Low-Grade Glioma Collection
VMAT	Volumetric Modulated Arc Therapy
ViT	Vision Transformer
WHO	World Health Organization

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Clinical Background

Accurate tumour delineation is essential in oncology because treatment planning, surgical decisions, and radiotherapy targeting all rest entirely on precise identification of tumour boundaries. From the initial identification of a suspicious lesion through to the delineation of treatment volumes for radiotherapy, the quality of clinical decision-making rests entirely on the precision with which tumour boundaries can be defined in medical images. Accurate delineation directly affects surgical planning, radiotherapy targeting, and longitudinal assessment of treatment response, whether a radiotherapy beam covers the full extent of disease or inadvertently spares viable tumour cells at the periphery, and whether longitudinal imaging assessments of treatment response are measuring the same anatomical region across time-points. In each of these applications, the accuracy and consistency of the tumour contour carry direct consequences for patient outcomes.

The standard clinical approach to tumour delineation remains largely manual. A specialist radiologist or radiation oncologist examines a volumetric imaging dataset slice by slice, drawing contours on each cross-section using dedicated workstation software. For a patient with a low-grade glioma, pre-operative FLAIR MRI delineation typically requires a substantial block of specialist time per case, reflecting the slice-by-slice manual contouring process described above. Across a high-volume neuro-oncology centre handling hundreds of cases annually, this cumulative burden is notable. For a patient with head and neck squamous cell carcinoma (HNSCC) undergoing intensity-modulated radiotherapy planning, the combined delineation of the primary gross tumour volume, involved lymph node regions, and the more than a dozen organs at risk that must be dose-constrained can require between one and three hours per case (Brouwer & Steenbakkers, 2015). In a high-volume oncology centre handling hundreds of cases annually, this adds up quickly — not only in hours but in the concentration that careful contouring demands from clinicians who also have clinics, ward rounds, and multidisciplinary meetings to attend.

Beyond the time burden, manual delineation is subject to a well-documented source of error: inter-observer variability. Different clinicians, drawing on the same imaging data, frequently

produce contours that differ in volume, shape, and spatial extent. For LGG on FLAIR MRI, inter-reader agreement, measured using the Dice Similarity Coefficient (DSC, a standard overlap metric from 0 to 1), between experienced neuroradiologists has been reported in the range of 0.75 to 0.85 (Bakas, Akbari, Sotiras, et al., 2017). For HNSCC on computed tomography (CT), inter-observer DSC values for gross tumour volume delineation range from 0.70 to 0.82 in multi-institutional studies (Nikolov et al., 2021). These are not trivial disagreements: a DSC of 0.75 between two experts means that roughly 25% of the delineated volume is classified differently by the two clinicians. In the context of radiotherapy, where dose gradients are shaped to millimeter precision around the target volume, a 25% volume discrepancy between planners represents a clinically meaningful source of uncertainty. For surgical planning, where the goal is to resect as much tumour as safely possible without compromising eloquent function, boundary uncertainty translates directly into uncertainty about the extent of resection achievable. Automated segmentation offers a principled approach to addressing both problems simultaneously — but only if the model is architected to handle the spatial properties of the target. For the cancers studied in this thesis, existing deep learning architectures have a structural limitation that prevents them from doing so reliably.

Both problems — the burden on specialist time and the inconsistency between clinicians — have made automated tumour segmentation one of the most actively pursued problems in clinical AI research. Deep learning, and convolutional neural networks in particular, have transformed the landscape of what is computationally achievable in this domain. When the target is compact and the contrast is high and there is plenty of labeled training data, models trained on annotated imaging datasets now produce contours in seconds that can approach the accuracy of experienced specialists (Litjens et al., 2017). The cancers studied in this thesis, however, are neither compact nor well-bounded, and the dominant segmentation architecture carries a structural limitation that is specifically ill-matched to their spatial properties. This became the central observation of the study: when the standard U-Net was applied to the FLAIR brain tumour dataset, it plateaued at a Dice score of approximately 0.55 and stayed there regardless of what training adjustments were made. That ceiling — consistent across five cross-validation folds — was what motivated looking at ConvNeXt rather than continuing to tune hyperparameters.

1.2 Limitations of Existing Approaches

When Ronneberger et al. (2015) published U-Net, it almost immediately became the default architecture for medical image segmentation — and a decade later it largely still is. Its central design innovation skip connections that carry high-resolution spatial information from the encoding path directly to the decoding path, bypassing the information bottleneck of successive pooling operations. This enables end-to-end trained segmentation networks to achieve spatial precision that pure encoder or patch-based architecture could not. This innovation addressed the dominant failure mode of earlier fully convolutional networks, which had strong semantic encoding but poor boundary precision. The U-Net's effectiveness has since been confirmed across a remarkable range of anatomical targets and imaging modalities, and it remains the architectural baseline against which new medical segmentation approaches are routinely evaluated.

However, the standard U-Net is built on 3×3 convolutional kernels throughout its encoder. This design choice is computationally efficient and consistent with the translational equivariance prior that makes convolutional networks well-suited to image analysis, but it imposes a hard constraint on the spatial context that any single convolutional layer can integrate: each output activation reflects the signal from only a 3×3-pixel neighborhood of its input feature map. The effective receptive field (ERF), the region of the original input that influences a given output neuron, measured through gradient magnitude — grows far more slowly than the theoretical receptive field with network depth, following a roughly square-root relationship (Luo et al., 2016). In practice this means a deep 3×3 network perceives a much narrower spatial neighborhood than its layer count would suggest. A standard five-stage U-Net operating on 128×128-pixel images have an effective bottleneck receptive field that covers only a fraction of the total image area.

For many medical segmentation targets, this locality is sufficient. Compact, well-circumscribed structures such as the optic nerve, the left ventricle, or hypervascular liver metastasis can be identified and delineated from local spatial context: the boundary is sharp, the contrast is high, and the information required to identify the lesion is concentrated near the tissue boundary itself. For the targets studied in this thesis, the situation is categorically different. Low-grade glioma grows by infiltrating along white matter tracts, producing a diffuse T2 and FLAIR hyperintensity whose outer boundary is not a discrete anatomical margin but a gradient, with tumour cell density decreasing continuously from the lesion centre outward into apparently normal-appearing tissue. A model that cannot integrate spatial context across the full spatial

extent of this gradient will systematically misclassify the peripheral infiltrative zone, producing contours that are too tight around the tumour core and that miss the clinically relevant infiltrative margin. This is exactly what happened in the Task A baseline experiments. The U-Net reached approximately DSC 0.55 and stayed there — adjusting the learning rate, changing the loss function, and running for more epochs didn't help. Several optimization strategies were explored during preliminary experiments, but none produced meaningful improvements in validation performance. The consistent performance plateau suggested that the limitation was more architectural than optimization related.

For HNSCC on CT, the limitation of the U-Net manifests differently, but its root cause is related. The primary challenge in head and neck CT segmentation is not boundary gradience but low intrinsic soft-tissue contrast: the primary tumour and adjacent musculature occupy similar Hounsfield Unit (HU) ranges on non-contrast planning CT, and reliable discrimination depends on subtle textural and morphological cues that become interpretable only in the context of the surrounding anatomy. A model with a restricted effective receptive field may struggle to capture sufficient anatomical context: it processes each location in near-isolation, without knowledge of the spatial relationships between the tumour, adjacent lymph nodes, vascular structures, and bony landmarks that a human contoured integrates naturally. In the Task B experiments conducted here, this manifested as fragmented and spatially inconsistent predictions from the standard U-Net, though as will be discussed, ConvNeXt-UNet's performance on this task introduces a more nuanced picture.

The two principal architectural strategies proposed to address the receptive field limitation of standard CNNs are attention mechanisms and self-attention transformer architectures. Attention U-Net (Oktay et al., 2018) incorporated soft attention gates into the skip connections, allowing the decoder to focus on contextually relevant encoder features. TransUNet (Chen et al., 2021) and SwinUNETR (Hatamizadeh et al., 2022) replaced the U-Net encoder entirely with transformer-based feature extractors capable of attending globally across the image in a single layer. These approaches have shown performance improvements on certain benchmarks, particularly in multi-modal brain tumour segmentation tasks where abundant annotated training data is available. They carry significant practical disadvantages in the low-data, resource-constrained settings that characterize most clinical AI research: transformers require notably more graphics processing unit (GPU) memory, are sensitive to training data volume, and lack inductive biases that allow convolutional networks to generalize from small, annotated datasets (Shamshad et al., 2023). The ConvNeXt architecture (Liu et al., 2022) takes a different route:

instead of adding attention mechanisms, it asks whether the convolutional building block itself can be redesigned to recover that contextual capacity — and it is this approach that the proposed method in this thesis adopts.

1.3 Research Gap

Three specific gaps in existing literature shaped the design of this study, and none of them has been systematically addressed in published work.

The first gap concerns the cross-modal evaluation of convolutional modernization approaches. Existing studies that evaluate ConvNeXt-based architectures in medical segmentation (Dong et al., 2024; Lee, 2022; Roy et al., 2023) validate their proposals on a single imaging modality and present results as evidence of general architectural advantage. Whether that advantage transfers to CT-based head and neck segmentation — where the tissue contrasts the diagnostically relevant spatial frequencies, and the nature of the boundary problem are all qualitatively different — has simply never been tested in a controlled paired study.

The second gap concerns the conditions under which ConvNeXt architectural principles do and do not confer advantage. The assumption implicit in most published proposals is that larger kernels, Layer Normalization and Gaussian Error Linear Unit (GELU) activation universally improve segmentation performance relative to 3×3 BatchNorm-ReLU networks. This assumption has not been rigorously tested against diverse task conditions. Understanding whether the ConvNeXt advantage is modality-specific, task-specific, or genuinely architectural is of significant scientific value: if improvements prove modality-dependent, that provides mechanistic insight into which specific properties of the imaging signal the enlarged receptive field is exploiting, and it informs principled architectural selection for new clinical tasks.
A related gap concerns the specific application of modern convolutional architectures to single-modality non-contrast CT for HNSCC gross tumour volume segmentation. The published literature is dominated by PET-CT multi-model approaches, and the performance of architecture such as ConvNeXt-UNet under CT-only conditions — which are relevant to the majority of radiotherapy centres globally — has not been characterised. The Task B experimental design was developed to address this gap directly. The Task B experimental design was developed to address this gap directly.

The third gap is methodological in nature. A common problem in published medical AI benchmarking is insufficient attention to the statistical validity of model comparisons. Many published segmentation studies report mean DSC across folds without testing whether the

observed difference between architectures is statistically significant, without ensuring that both architectures were evaluated on identical patient subsets, and without verifying the normality assumptions that underline the parametric tests applied. The experimental framework was designed with rigorous statistical comparison as a primary methodological commitment, using patient-wise GroupKFold splitting with identical fold assignments for both models and appropriate hypothesis tests selected based on explicit normality testing. An early version of the Task B results appeared to show ConvNeXt outperforming the U-Net — until the evaluation functions were checked and a soft-versus-binary Dice mismatch was found. That error, and the steps taken to correct it, reinforced the importance of making statistical choices explicit before looking at results.

These three gaps directly shaped the experimental design of this study.

A finding that was not predicted at the outset that the ConvNeXt-UNet would significantly underperform the standard U-Net on the CT task despite outperforming it on MRI is evidence that the first two gaps exist: single-modality evaluations of ConvNeXt reported in the literature would not have predicted this reversal.

1.4 Research Objectives

The central aim of this study is to evaluate whether the architectural modernization embodied in the ConvNeXt design paradigm specifically the substitution of 3×3 standard convolutions with 7×7 depthwise separable convolutions within an inverted bottleneck block with Layer Normalization and GELU activation improves automated medical image segmentation relative to the standard U-Net, and whether any such improvement is consistent across two imaging modalities with mainly different tissue contrast mechanisms.

Four objectives structure the experimental work:

Objective 1. To implement and evaluate a standard U-Net as a reproducible, rigorously validated performance baseline on both a brain MRI segmentation task (Task A: TCGA-LGG FLAIR, 110 patients) and a head and neck CT segmentation task (Task B: HECKTOR 2025 MDA cohort, 396 patients), using patient-wise 5-fold GroupKFold cross-validation to prevent data leakage.

Objective 2. To design and implement a ConvNeXt-UNet integrating the ConvNeXt-Tiny backbone with ImageNet pre-trained weights applied through grayscale-to-RGB channel replication for Task A (MRI) and through a 1×1 convolutional projection layer for Task B

(CT) into a U-Net encoder–decoder framework with skip connections, Layer Normalization, and GELU activation throughout the decoder.

Objective 3. To evaluate whether the ConvNeXt-UNet produces statistically significant DSC improvements over the standard U-Net on Task A (brain MRI), and to determine the direction and statistical significance of any performance difference on Task B (head and neck CT), applying hypothesis tests selected based on prior normality assessment.

Objective 4. To interpret the cross-task pattern of results including any modality-specific reversal in the direction of the architectural effect in the context of the theoretical properties of the two architectures and the imaging characteristics of each task, contributing to the broader understanding of when and why convolutional modernization approaches do and do not transfer across imaging modalities.

The overall experimental pipeline connecting both tasks is illustrated in Figure 8 in Chapter 3.

1.5 Scope and Limitations

Several methodological choices in this study define both the boundaries of the conclusions that can be drawn and the practical trade-offs that informed the experimental design.

A primary technical constraint of this study was the hardware limitation of the Kaggle environment, specifically the available GPU memory. This necessitated a transition from a 3D volumetric approach to a 2D slice-based processing paradigm. While 3D models can capture more spatial context, the 2D approach allowed for deeper model architectures and more stable training cycles within the allocated resources. During the early experimental phase, I encountered significant GPU memory constraints, which forced this shift to ensure the models could complete the five-fold cross-validation without crashing.

Task A uses only the FLAIR sequence from the TCGA-LGG collection. The BraTS benchmark (Baid et al., 2021; Menze, 2015) uses all four standard pre-operative MRI sequences (T1, T1-Gd, T2, FLAIR), and multi-modal fusion consistently improves segmentation accuracy relative to any single-modality approach. The single-modality FLAIR-only design was chosen to keep the architectural comparison as clean as possible: by holding the information input constant and varying only the encoder architecture, any observed DSC difference can be attributed to the model's capacity to exploit the available signal rather than to differences in information content across models.

Task B uses CT-only input from the HECKTOR 2025 MDA cohort, explicitly excluding the co-registered FDG-PET channel that is available in the HECKTOR dataset. This decision was motivated by the same experimental control logic: CT-only segmentation represents the more demanding condition and the one most relevant to centres without access to PET imaging; evaluating under the most challenging available condition provides the most informative test of architectural robustness. This constraint notably limits the achievable DSC ceiling for Task B relative to what PET-CT multi-modal systems can achieve.

The study stops at algorithmic performance — there is no prospective clinical validation. Results are on retrospective public datasets, and whether the models would hold up on data from different scanner platforms, different institutions, or different patient populations has not been tested. This is a known limitation of retrospective deep learning work and is addressed directly in the Discussion. These limitations were considered acceptable within the practical and computational scope of this thesis rather than imposed retrospectively, and they are documented here so that results can be contextualized accurately by the reader.

1.6 Contributions

The experimental work described in this thesis produced four contributions that were not anticipated at the outset.

The primary empirical contribution is the first controlled cross-modal paired evaluation of a ConvNeXt-UNet architecture against a standard U-Net baseline on both FLAIR MRI brain tumour segmentation and CT head and neck cancer segmentation, under identical experimental conditions and with rigorous statistical testing. The finding is not a simple confirmation of architectural superiority: the ConvNeXt-UNet notably outperforms the standard U-Net on FLAIR MRI (mean DSC 0.9328 versus 0.5496, 69.7% relative improvement, Wilcoxon $p < 0.001$), while the standard U-Net significantly outperforms the ConvNeXt-UNet on head and neck CT (mean DSC 0.3000 versus 0.2558, 17.3% relative margin, paired t-test $p = 0.0212$). This reversal in the direction of the architectural effect is the central empirical finding of the thesis and constitutes a scientifically meaningful contribution: it demonstrates that architectural modernization does not uniformly improve segmentation across modalities, and that the relationship between ConvNeXt inductive biases and imaging modality characteristics requires modality-specific analysis.

The secondary methodological contribution lies in the rigor of experimental design. The patient-wise GroupKFold cross-validation protocol, the identical fold assignments enforced

for both models within each task, and the normality-guided selection of statistical tests (Wilcoxon for Task A, paired t-test for Task B) represent a level of statistical care that many published architectural comparison studies do not achieve. The explicit documentation of this methodology provides a reproducible template for future cross-modal architectural evaluation studies.

The tertiary contribution is the transparency of the experimental record. The thesis documents the training configurations that failed before the successful implementation was reached including the metric inconsistency (soft versus binary Dice) that initially produced misleading ConvNeXt results on Task B, the learning rate and loss function choices that proved unstable when training the ConvNeXt decoder against the frozen pre-trained backbone, and the phase-wise training strategy that ultimately enabled stable convergence. This transparency adds reproducibility value that abbreviated published accounts rarely provide.

A fourth contribution is conceptual. The architecture-modality matching framework proposed in Section 5.4.2 identifies three properties that jointly predict whether a ConvNeXt-UNet will outperform a standard U-Net for a given segmentation task: the spatial diffuseness of the target boundary, the statistical proximity of the imaging modality to the ImageNet pre-training distribution, and the adequacy of the training dataset relative to the model's parameter count. This framework is the first attempt in the published literature to explain the modality-dependence of ConvNeXt performance using a principled, testable account rather than post-hoc reasoning. It generates specific predictions for task-modality combinations beyond those tested here.

1.7 Thesis Structure

Chapter 2 covers the clinical context, imaging physics, and architectural background. Chapter 3 describes the datasets, preprocessing, architecture, training configuration, and statistical framework. Chapter 4 reports quantitative and qualitative results for both tasks. Chapter 5 interprets what the results mean and why the two tasks produced opposite conclusions. Chapter 6 summarizes contributions and outlines the most productive next steps.

2. Theoretical Background and Literature Review

2.1 Oncological and Clinical Context

2.1.1 Low-Grade Glioma

Gliomas are primary brain tumours arising from the glial supporting cells of the central nervous system, astrocytes, oligodendrocytes, and their progenitors. The 2021 WHO classification (Louis et al., 2021) updated the grading system to include molecular markers — specifically IDH mutation status and 1p/19q co-deletion — alongside traditional histological criteria. This change reflects a practical clinical problem: two tumours that look identical under the microscope can behave completely differently in patients, and a grading system based only on appearance was missing this. Under this framework, the term LLG broadly encompasses IDH-mutant astrocytomas (WHO grade 2 and 3) and IDH-mutant 1p/19q co-deleted oligodendrogliomas (WHO grade 2 and 3), tumours that are characterised by slower growth, longer natural history, and more favourable prognosis than IDH-wildtype glioblastoma, while still carrying notable risks of malignant progression over time.

Unlike most primary brain tumours, LGG typically presents in relatively young adults — median age at diagnosis is the mid-thirties to mid-forties — which makes the long-term consequences of treatment planning errors particularly significant. The clinical presentation is typically insidious: seizures are the most common presenting symptom, often occurring years before the tumour is identified, and neurological deficits if present, may be subtle and slowly progressive. This indolent natural history can give a false impression of clinical insignificance, but long-term outcome data consistently show that LGG is not benign in the curative sense: the majority of IDH-mutant astrocytomas will eventually transform to higher-grade disease, and the degree to which this transformation is delayed depends importantly on the extent to which the initial tumour can be surgically controlled. Meta-analyses and large retrospective cohort studies have established that gross-total or near-gross-total resection is associated with prolonged progression-free and overall survival relative to biopsy or subtotal resection, independent of adjuvant therapy (Sanai et al., 2011). This survival advantage of aggressive resection places extraordinary demands on the precision of pre-operative tumour delineation: the surgical navigation system, intraoperative mapping protocols, and planned resection margins are all defined relative to the pre-operative contour, and a contour that underestimates

the peripheral extent of tumour infiltration may result in incomplete resection and earlier transformation.

The imaging appearance of LGG on FLAIR MRI reflects the pathophysiology of diffuse infiltrative growth. Tumour cells, most of which are metabolically active but not highly proliferative, infiltrate along white matter tracts and around neurons without causing acute tissue destruction or neovascularization in the early phase of growth. The FLAIR signal elevation in the lesion and its surroundings arises from a combination of true tumour infiltration, reactive astrocytosis, and vasogenic oedema, all of which share the property of elevated extracellular water content relative to normal parenchyma. Because these processes extend well beyond the highest-density tumour core, the FLAIR hyperintense region is considerably larger than the region of maximal tumour cell density, and its outer boundary is not a discrete anatomical interface but a gradient of decreasing cell density and oedema that fades progressively into histologically normal-appearing brain tissue. Neuroradiological practice typically defines the outer boundary of the FLAIR abnormality as the segmentation target, acknowledging that some margin of apparently normal tissue adjacent to the visible signal change likely contains infiltrated cells at sub-radiological density.

This gradual boundary transition creates a major challenge for automated segmentation algorithms. A model that makes local decisions evaluating the FLAIR signal intensity in a small neighborhood around each candidate pixel cannot reliably distinguish the peripheral tumour margin from the normal FLAIR variation that occurs in healthy white matter. The peripheral tumour signal is characterised not by its absolute intensity but by its relationship to the signal in the tumour core, the signal in adjacent normal white matter, and the global spatial pattern of signal elevation across the entire lesion. Perceiving these relationships requires integrating information across the spatial extent of the full tumour, which may span 40–80 mm at 128×128 resolution and occupy a correspondingly large fraction of the image area. This characteristic makes LGG a suitable test case for evaluating whether enlarged-kernel convolutional encoders provide genuine performance advantages over standard 3×3 architectures. This gradient boundary property was visible immediately when inspecting the TCGA-LGG dataset during preprocessing: the axial FLAIR slices showed a bright core that faded progressively rather than ending sharply and drawing a binary mask boundary anywhere in that transition zone required a judgement call rather than a threshold. That observation, made before any training began, set the architectural question for the study.

2.1.2 Head and Neck Squamous Cell Carcinoma

HNSCC encompasses a heterogeneous group of malignancies arising from the squamous epithelium lining the mucous membranes of the upper aerodigestive tract. Anatomical subsites include the oral cavity, oropharynx (which includes the tonsils and base of tongue), hypopharynx, larynx, and nasopharynx. Together, these cancers account for approximately 890,000 new diagnoses and 450,000 deaths annually, making HNSCC the sixth most prevalent cancer type globally (Sung, 2021). Two epidemiologically distinct disease patterns coexist: tobacco-alcohol–driven HNSCC, which predominantly affects older patients and has high rates of locoregional recurrence; and HPV-associated oropharyngeal carcinoma, which disproportionately affects younger, non-smoking patients and has generally more favourable outcomes after standard-of-care treatment.

Intensity-modulated radiation therapy (IMRT) is the primary treatment for most locoregional presentations of HNSCC, often delivered concurrently with platinum-based chemotherapy. IMRT's defining feature is its ability to shape complex dose distributions that simultaneously deliver ablative doses to target volumes while accurately sparing adjacent dose-sensitive structures. Realizing this capability requires highly accurate delineation of both the targets the primary gross tumour volume (GTVp) and involved lymph node volumes (GTVn)-and the organs at risk (OARs), which include the bilateral parotid and submandibular glands, the spinal cord and brainstem, the larynx, the mandible, and the swallowing musculature, among others. The clinical consequences of delineation error fall into two categories: geographic miss where the target contour fails to encompass viable tumour tissue, leading to in-field recurrence; and excess OAR dose where the OAR contour is displaced or underestimated, leading to radiation-induced xerostomia, dysphagia, osteoradionecrosis, or cord myelopathy. Both categories have direct and severe clinical consequences, underscoring the importance of contouring precision.

Automated GTVp and GTVn segmentation on non-contrast CT is challenging because CT provides limited soft-tissue contrast compared to MRI. CT encodes tissue properties as linear X-ray attenuation coefficients, expressed on the Hounsfield Unit scale with water at 0 HU and air at −1,000 HU. The HU range occupied by soft tissue tumours, lymph nodes, and adjacent musculature in the head and neck is narrow approximately +20 to +80 HU for most solid soft tissue structures meaning that primary oropharyngeal carcinoma and the adjacent pterygoid musculature may differ by only a few HU on non-contrast-enhanced CT (Gregoire et al., 2018). This is in sharp contrast to the FLAIR MRI setting, where tumour-to-parenchyma signal

differences of 20–40% are typical for LGG. The consequence is that automated CT-based HNSCC segmentation must rely on subtle textural differences, morphological cues, and anatomical context, the spatial relationships between the tumour and surrounding structures rather than conspicuous intensity contrasts.

A further practical challenge is the presence of metallic dental artefacts in a notable proportion of HNSCC patients. High-density metallic restorations produce beam-hardening streaks with HU values exceeding +3,000 HU in the surrounding tissue, corrupting the image signal at the axial levels most likely to contain primary oropharyngeal disease. The head and neck region are anatomically dense and geometrically complex: the primary tumour, bilateral lymph node chains, parotid and submandibular glands, mandible, cervical spine, major vessels, airway, and laryngeal cartilages are all present in close proximity across a limited axial extent, creating a challenging segmentation environment even for expert human delineators.

Inter-observer variability studies for HNSCC delineation consistently demonstrate DSC values between experienced radiation oncologists ranging from 0.70 to 0.82 for GTVn and somewhat lower for GTVp, with the highest variability associated with small or poorly defined primary tumours (Nikolov et al., 2021). These figures reflect genuine biological and imaging ambiguity rather than inadequate clinical training: the boundary between primary tumour and reactive oedematous tissue is frequently not resolvable on non-contrast CT even in principle, requiring the delineator to make a clinical judgement informed by the full history, physical examination, and the pattern of disease rather than imaging alone.

Automated segmentation of HNSCC has been a focus of the HECKTOR challenge series since 2020 (Andrearczyk, Oreiller, Abobakr, et al., 2023), which has produced detailed benchmarking data on what is achievable under different input conditions. The best-performing systems have consistently used multi-modal PET-CT input and 3D volumetric processing, achieving GTVp DSC values of 0.75–0.82 on the challenge test sets. CT-only systems — which are clinically relevant for centres without PET infrastructure — have been far less systematically studied, and published results for CT-only HNSCC segmentation with modern convolutional architectures are largely absent from the literature. This gap is directly relevant to Task B of this study, which evaluates both the U-Net and ConvNeXt-UNet under exactly this CT-only constraint.

2.2 Medical Imaging Phsysics

2.2.1 MRI and FLAIR Acquisition

MRI generates tissue contrast based on differences in proton relaxation properties across tissues. Following radiofrequency excitation, tissues recover longitudinal (T1) and transverse (T2) magnetization at different rates, producing contrast differences that can be reconstructed into anatomical images (Hashemi et al., 2010).

Different tissues have characteristic T1 and T2 values that produce distinct signal intensities on different pulse sequences. Cerebrospinal fluid (CSF), which consists largely of free water with long T2 relaxation, appears bright on T2-weighted images, potentially masking perilesional abnormalities in adjacent tissue. The FLAIR sequence addresses this by applying an initial 180° inversion pulse followed by an inversion time (TI) chosen to null the longitudinal magnetization of CSF at the moment of excitation (Hashemi et al., 2010). With CSF signal suppressed, FLAIR images retain T2 hyperintensity from tissues with elevated water content including tumour, oedema, demyelinated plaques, and ischemic change, while the dark CSF provides high contrast to adjacent abnormality. For low-grade glioma specifically, where gadolinium-enhanced T1 imaging typically shows no enhancement because the blood–brain barrier has not yet been disrupted, FLAIR is the primary sequence for delineating the spatial extent of tumour infiltration and is the clinical standard for pre-operative contouring (Patel et al., 2019).

A characteristic property of FLAIR signal in LGG is its non-uniform intensity gradient at the tumour margin. The signal elevation is highest in the densely infiltrated tumour core and decreases progressively toward the periphery as tumour cell density and associated oedema diminish. This spatial gradient does not correspond to a sharp anatomical margin perceptible from local pixel intensities; it requires integration of spatial context over the full transition zone to reliably estimate where infiltrated tissue ends and normal parenchyma begins. Figure 1 shows a representative axial FLAIR slice from the TCGA-LGG dataset alongside its expert-corrected mask, illustrating this progressive boundary gradient. This is precisely why the standard 3×3 U-Net struggles here: no local patch of the FLAIR image, examined in isolation, contains enough information to decide where the tumour boundary belongs.

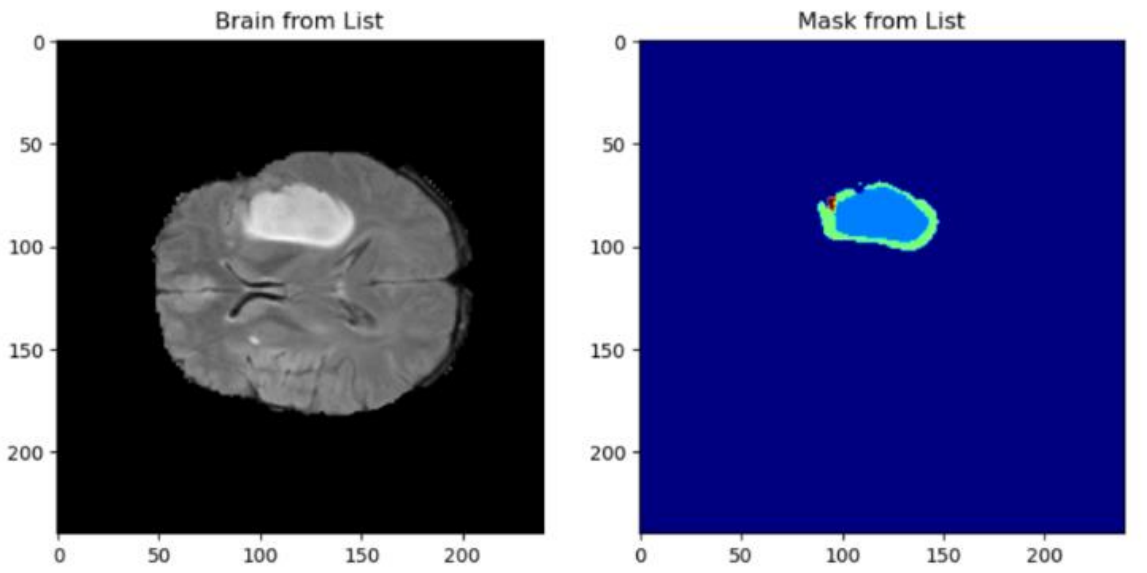

Figure 1: Representative axial FLAIR MRI slice from the TCGA-LGG dataset (left), expert-corrected binary tumour mask and colour overlay showing the diffuse tumour boundary (right).

2.2.2 CT Imaging and the Hounsfield Scale

CT reconstructs images of the body by measuring the differential attenuation of X-rays as they pass through tissue from multiple projection angles. A rotating X-ray tube and opposing detector array acquire projection data; filtered back-projection or iterative reconstruction algorithms compute a 3D volume of linear attenuation coefficients from the acquired projections. CT values are standardised as Hounsfield Units (HU) relative to the attenuation of water and air:

$$HU = [(\mu\_tissue - \mu\_water) / (\mu\_water - \mu\_air)] \times 1000$$

where μ denotes the linear attenuation coefficient at the beam energy of the scanner (Hounsfield, 1980). By definition, water is 0 HU and air is −1,000 HU; dense bone typically exceeds +400 HU, and soft tissues span the approximate range of −100 to +200 HU.

The HU scale is calibrated and reproducible across different CT scanner models and acquisition settings, making CT the reference modality for radiation treatment planning where accurate electron density maps are required for dose calculation. However, this calibrated quantitative scale does not translate into high soft-tissue discrimination in the head and neck region, because the biologically and clinically relevant tissue types tumour, lymph node, muscle, fat, and glandular tissue, cluster within a narrow HU range that spans only a few tens of HU for the most relevant comparisons. The consequence is that CT provides weak intrinsic contrast between most head and neck soft-tissue structures of interest for oncological delineation.

Clinical radiologists rely on the shape, size, and anatomical location of structures in their context within the global anatomy of the head and neck rather than their HU value to identify pathological lymph nodes or primary tumours. When the HECKTOR 2025 MDA cohort was first examined during preprocessing, the difficulty of identifying tumour boundaries visually was striking — even knowing approximately where the primary tumour should be, the adjacent musculature and the tumour occupied overlapping HU ranges that made confident boundary placement genuinely difficult at several axial levels. This reinforces the limitation of intensity-based segmentation and highlights the importance of incorporating spatial context in model design.

The choice to apply per-slice min-max intensity normalization rather than a fixed HU window in the preprocessing pipeline for Task B was motivated by the presence of dental artefacts in the HECKTOR 2025 MDA cohort. Fixed windowing clipping HU values to a soft-tissue window such as [−150, +250] HU and linearly mapping to [0, 1] produces better average contrast in artefact-free volumes but compresses the relative signal differences in slices dominated by metallic artefact outliers. Per-slice normalization adapts to the local intensity range at each axial level, preserving relative soft-tissue contrast regardless of the presence of high-density structures elsewhere in the slice. In practice, this normalization strategy was found to stabilize the training process, particularly in slices affected by strong artefacts. Without this adjustment, the model showed inconsistent predictions across similar anatomical regions. The CT preprocessing pipeline, including per-slice normalisation and the resulting binary mask, is illustrated in Figure 2.

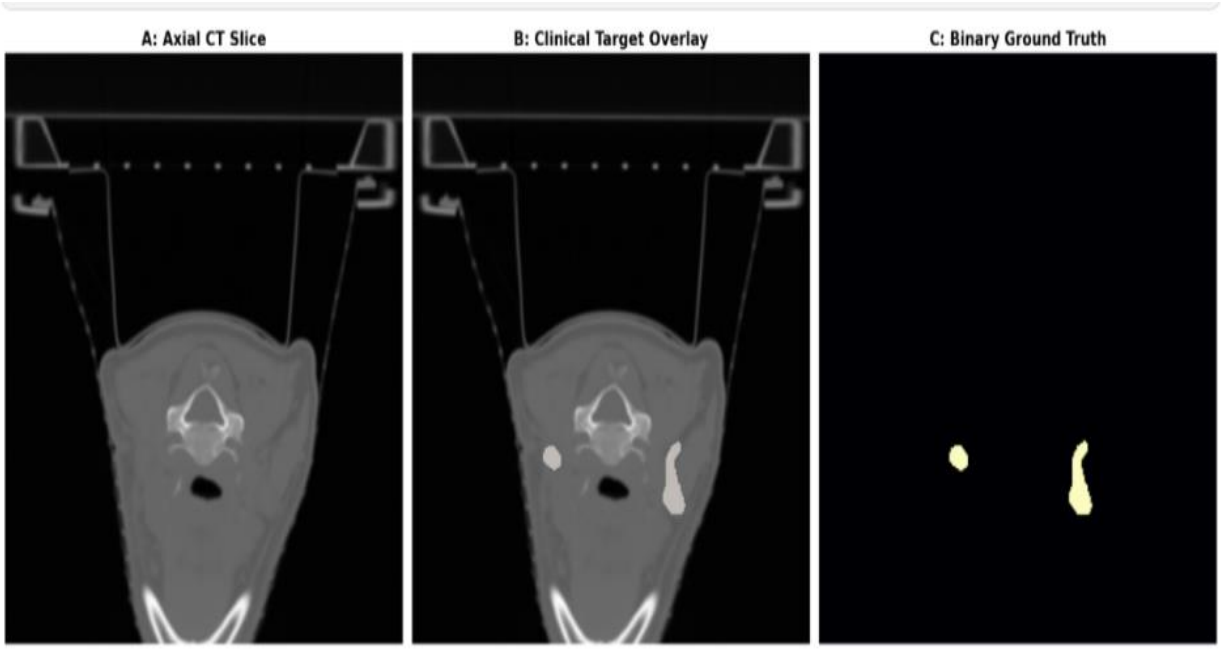

Figure 2. Example CT preprocessing pipeline for the HECKTOR dataset, including normalised input and binary segmentation mask.

2.3 Deep Learning for Medical Image Segmentation

2.3.1 Convolutional Neural Networks and the Receptive Field

A convolutional neural network learns hierarchical spatial feature representations through repeated application of learned filter banks to input arrays. The defining operation is discrete 2D cross-correlation: for an input feature map f and a learned kernel g of spatial extent k×k, the output at position (n, m) is

$$y[n,m] = \sum_{i=-\lfloor k/2 \rfloor}^{\lfloor k/2 \rfloor} \sum_{j=-\lfloor k/2 \rfloor}^{\lfloor k/2 \rfloor} f[n+i, m+j] \cdot g[i,j]$$

Two properties of this operation are critical for image analysis: translational equivariance, ensuring that a feature detector responds identically to a pattern regardless of its position in the image, and parameter sharing, which reduces the number of trainable parameters relative to fully connected layers by several orders of magnitude for image-sized inputs (Goodfellow, 2016). In practice, this manifested clearly during Task A preprocessing: when the same filter configuration that worked on natural images was applied to low-contrast FLAIR slices without per-slice normalization, the gradient signal was dominated by background noise rather than the tumour region.

The receptive field (RF) of a neuron at layer L in a stack of k×k convolutional layers without downsampling grows as

$$RF = kL - (L - 1) = L(k - 1) + 1 = Lk - 1 + 1$$

For 3×3 kernels (k=3), the RF grows by 2 pixels per layer: after 5 layers, the RF covers 11×11 pixels; after 10 layers, 21×21 pixels. The ERF quantified as the region of the input for which a unit change materially influences the output, measured through gradient magnitude is consistently smaller than the theoretical RF and grows approximately as the square root of the number of layers, following a Gaussian profile (Luo et al., 2016). This sub-linear growth means that even deep 3×3 networks perceive only a limited spatial neighborhood in practice.

For a five-stage U-Net with 3×3 kernels processing 128×128-pixel images, the theoretical bottleneck RF covers approximately 23×23 pixels, but the effective RF is considerably smaller. A low-grade glioma occupying 20–40% of the axial image field spans 25–50 pixels in each dimension at this resolution, larger than the entire effective RF of the bottleneck encoder representation. The implication is that the standard U-Net encoder cannot simultaneously perceive the tumour core and its infiltrative periphery as a unified spatial entity; it makes local decisions about each region of the image in relative isolation from distant spatial context. This is the structural limitation that ConvNeXt's enlarged kernels are designed to address. This effect was also observed qualitatively in the segmentation outputs, where boundary regions of the tumour were frequently under-segmented. This observation supports the theoretical constraint of limited effective receptive fields in standard architecture.

2.3.2 U-Net and Its Variants

The U-Net, proposed by Ronneberger et al. (2015) for biomedical image segmentation, introduced the encoder–decoder architecture with skip connections that has since become the canonical framework for dense medical segmentation tasks. The encoding path applies successive blocks of dual 3×3 Conv2D-ReLU layers followed by 2×2 max-pooling, progressively reducing spatial resolution from 128×128 to 8×8 while doubling the number of feature channels at each stage. The decoding path reverses this process through transposed convolutions, and at each resolution level it concatenates the upsampled decoder feature map with the corresponding encoder feature map via a skip connection. This skip-connection mechanism is the defining innovation of U-Net: it ensures that fine-grained spatial detail from the high-resolution early encoder layers reaches the output prediction path without being discarded in the progressive compression of the encoder bottleneck (Ronneberger et al., 2015).

Figure 3 shows the full U-Net architecture with its five encoder stages, bottleneck, decoder, and skip connections, including the approximate ERF at the first stage.

In early U-Net training runs for Task A, removing the skip connections as an ablation step caused the validation DSC to drop by approximately 0.12–0.15 per fold, confirming that the encoder's spatial detail is genuinely necessary for boundary localization at the resolutions used in this study.

The U-Net was designed for the data-scarce regime of medical imaging and demonstrated that segmentation networks can be trained effectively on fewer than 30 labelled images with appropriate data augmentation. According to Siddique et al. (2021), its influence on subsequent work is pervasive, catalogued over 50 distinct U-Net variants, spanning attention mechanisms, dense connections, nested architectures, multi-scale processing, and volumetric extensions.

V-Net (Milletari et al., 2016) extended the framework to 3D volumetric inputs and introduced the Dice loss function as a training objective, addressing the class imbalance endemic to medical segmentation. Attention U-Net (Oktay et al., 2018) incorporated soft attention gates into the skip connections, allowing the decoder to selectively weight encoder features based on the current decoding context and suppressing background activations at each resolution level. nnU-Net (Isensee et al., 2021) introduced a self-configuring pipeline that automatically adapts the U-Net architecture and training protocol to the specific properties of each dataset target size, image resolution, class balance, available memory and demonstrated that this automated configuration achieves state-of-the-art performance across a considerably broad range of medical segmentation benchmarks, underscoring the fundamental robustness of the encoder–decoder with skip connections when well-tuned.

Despite this success, a persistent limitation of standard U-Net models for diffuse, low-contrast segmentation targets has been progressively documented in the literature. Zhang et al. (2023) demonstrated analytically and empirically that models with restricted effective receptive fields systematically under-segment the peripheral extent of diffuse brain tumours, attributing the mechanism to the inability of small-kernel encoders to integrate the long-range spatial context that distinguishes the infiltrative tumour periphery from normal-appearing tissue. This finding provides theoretical grounding for the Task A performance plateau of the standard U-Net (mean DSC 0.5496) observed in this study, independent of training procedure optimization. What the published record does not establish is whether a minimal, targeted intervention — replacing only the U-Net encoder with a ConvNeXt backbone while keeping the decoder identical — is

sufficient to produce the performance gains associated with fully redesigned architectures such as MedNeXt. This minimalist encoder-replacement approach is what this study evaluates. The standard U-Net architecture is shown in Figure 3.

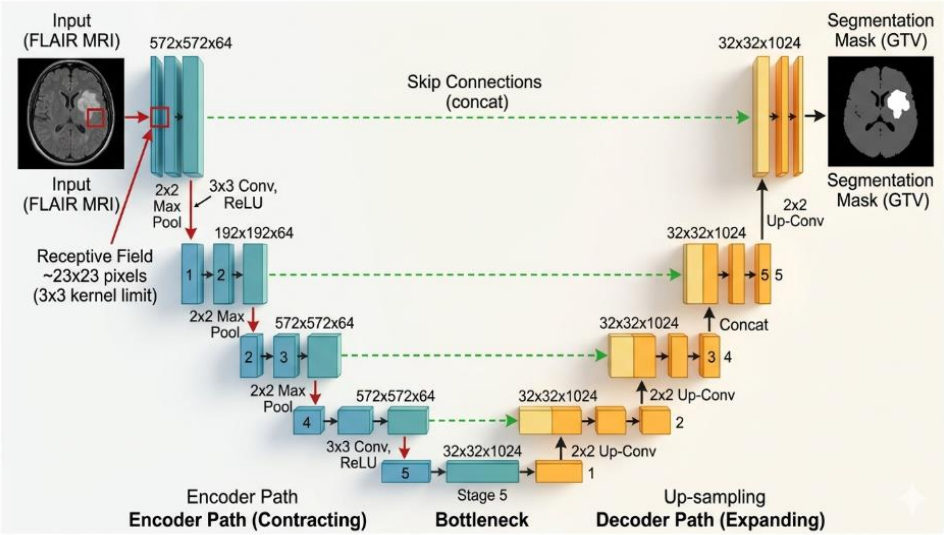

Figure 3. Standard U-Net architecture diagram showing the encoder path (five stages, blue), bottleneck, decoder path (orange), and skip connections (green dashed arrows). The effective receptive field annotation (red box, ~23×23 pixels at Stage 1) illustrates the locality constraint of 3×3 kernels that the ConvNeXt modification addresses.

2.3.3 Transformer-Based Segmentation Models

The Vision Transformer (ViT), introduced by Dosovitskiy et al. (2021), demonstrated that a pure self-attention architecture without any convolutional layers could match or exceed the best convolutional networks on large-scale image recognition when pre-trained on sufficiently large datasets. ViT partitions the image into non-overlapping patches, projects each to a token embedding, and applies multi-head self-attention across all token pairs. Because every token can directly attend every other token in a single operation, the theoretical receptive field is global from the first layer, a qualitative departure from the locality of convolutional stacks. This global attention is accurately the property required for integration of long-range spatial dependencies in diffuse segmentation targets, making ViT-based architectures conceptually appealing for medical segmentation. However, in practical implementation within this study, the computational requirements of transformer-based models posed significant constraints, particularly under limited GPU resources.

TransUNet (Chen et al., 2021) applied this principle to segmentation, using a ResNet-50 backbone to extract feature maps that were tokenized and processed by a ViT encoder before

decoding through a conventional upsampling path. On abdominal CT multi-organ segmentation, TransUNet demonstrated improvements over pure CNN baselines, establishing that global attention mechanisms can be beneficial in this domain. SwinUNETR (Hatamizadeh et al., 2022) introduced the hierarchical Swin Transformer (Liu et al., 2021) to medical segmentation, addressing the quadratic memory complexity of full self-attention by restricting attention to non-overlapping local windows with a shifted-window scheme that propagates inter-window information across layers. SwinUNETR achieved state-of-the-art results on multi-modal BraTS brain tumour segmentation and has become a widely adopted strong baseline.

The practical limitations of transformer-based models for medical image segmentation are, however, notable. The self-attention operation requires considerably more GPU memory than comparable convolutional models, both because of the dense attention matrices and because of the large token sequence lengths generated by high-resolution medical images. Transformers lack the translational equivariance prior of convolution, meaning they require larger labelled training datasets to learn spatial invariances that CNNs acquire by construction, a significant disadvantage in the data-scarce, annotation-expensive setting of medical imaging. Shamshad et al. (2023), in a comprehensive survey of transformer applications in medical imaging, noted that performance advantages over well-tuned CNN baselines are frequently modest and dataset-dependent, and that the computational overhead often outweighs the accuracy gain in resource-constrained clinical research settings. These limitations motivate the exploration of architectural approaches that recover the contextual advantages of self-attention within the computational framework of convolution, the programmeme that defines ConvNeXt. A substantive disagreement exists in the literature regarding which architectural strategy most effectively addresses the receptive field limitation of standard CNNs. Transformer advocates argue that global self-attention is categorically superior to any local-kernel approach because it provides truly unlimited receptive field from the first layer and learns arbitrary long-range dependencies without the spatial locality bias of convolution (Chen et al., 2021; Hatamizadeh et al., 2022). ConvNeXt advocates counter that the performance advantage of full global attention in medical imaging is achieved only at the cost of quadratic memory complexity and large training data requirements that are incompatible with the data scarcity of annotated medical datasets and that hierarchical enlarged-kernel convolutions achieve comparable contextual reasoning with a fraction of the computational overhead (Liu et al., 2022; Roy et al., 2023). The empirical record to date does not decisively resolve this debate: SwinUNETR

achieves higher whole-tumour DSC on multi-modal BraTS data than published ConvNeXt approaches, but it also uses four MRI sequences where ConvNeXt studies typically use one or two, making direct comparison impossible. The present study contributes to this debate by evaluating the enlarged-kernel convolutional approach under the most demanding conditions single-modality input, GPU-constrained training where the practical disadvantages of transformer architectures are most acute. The findings presented here provide additional evidence under single-modality and GPU-constrained conditions.

2.4 ConvNeXt: Modernizing the Convolutional Architecture

2.4.1 The Systematic Modernization Programmeme

The ConvNeXt architecture (Liu et al., 2022) was constructed through a systematic empirical investigation of the design choices that separate a standard ResNet-50 from the Swin Transformer on ImageNet and dense prediction benchmarks. Rather than proposing a hybrid model incorporating attention mechanisms, the authors asked whether a purely convolutional network could match transformer performance by adopting the specific design choices of the Swin Transformer one at a time. Starting from a ResNet-50 baseline at 76.1% ImageNet top-1 accuracy, eight sequential modifications were applied, each measured in isolation. The final ConvNeXtTiny model achieved 82.1% top-1 accuracy, matching the Swin-T while remaining entirely convolutional.

The modifications and their individual accuracy contributions are as described by Liu et al. 2022. The most consequential single change was the replacement of 3×3 standard convolutions with 7×7 depthwise separable convolutions, which contributed a 1.1% accuracy improvement on ImageNet. The replacement of Batch Normalization with Layer Normalization and of ReLU with GELU contributed additional improvements. Together, these changes transformed standard ResNet into an architecture that matches the representational capacity of vision transformers on standard benchmarks while retaining the training efficiency, implementation simplicity, and inference throughput of conventional CNNs. ConvNeXt V2 (Woo et al., 2023) subsequently demonstrated that the same architecture achieves further improvements through fully convolutional masked autoencoder self-supervised pre-training.

These architectural modifications are relevant to medical segmentation because they address limitations associated with diffuse boundaries and limited spatial context. The architectural changes in ConvNeXt address accurately the properties that distinguish medically relevant

segmentation targets diffuse boundaries, multi-scale spatial context, low intrinsic contrast from the compact, high-contrast objects that standard 3×3 networks handle well. Understanding why each modification contributes to performance, and under what conditions, is essential for the principled application of ConvNeXt principles to new medical tasks. These modifications were retained in the proposed ConvNeXt-UNet architecture evaluated in this thesis.

2.4.2 Depthwise Separable 7×7 Convolutions

The substitution of standard 3×3 convolutions with 7×7 depthwise separable convolutions is the most mechanistically significant modification in the ConvNeXt design programmeme. A depthwise convolution applies a single k×k spatial filter independently to each input channel, without mixing channels:

$$y_c[n,m] = \sum_{i=-\lfloor k/2 \rfloor}^{\lfloor k/2 \rfloor} \sum_{j=-\lfloor k/2 \rfloor}^{\lfloor k/2 \rfloor} x_c[n+i, m+j] \cdot w_c[i,j]$$

for channel c. This separates spatial feature extraction performed by the depthwise step from cross-channel feature mixing, performed by subsequent pointwise (1×1) convolutions. The parameter count for a depthwise convolution with k×k kernel and C input channels is $C \times k^2$, compared to $C^2 \times k^2$ for a standard convolution of the same spatial extent. For k=7 and C=128, this represents a 128-fold parameter reduction relative to a standard 7×7 convolution.

The enlargement of the spatial kernel from 3×3 to 7×7 has a direct and quantifiable effect on the effective receptive field. A 7×7 depthwise convolution applied to a feature map at 1/4 of the input resolution (after the patchify stem of the ConvNeXtTiny encoder) perceives a 28×28-pixel region of the original input in a single operation. Applied hierarchically at 1/4, 1/8, 1/16, and 1/32 resolution across the four ConvNeXt stages, the effective receptive field at the deepest encoder stage covers a notable fraction of the 128×128-pixel input image. Figure 4 compares the ERF growth of the standard U-Net and the ConvNeXt-UNet across encoder depth, showing how the ConvNeXt architecture exceeds the minimum receptive field required for full LGG delineation by Stage 2. Ding et al., 2022) demonstrated empirically that increasing CNN kernel sizes up to 31×31 provides monotonically increasing improvements in dense prediction tasks, with the primary mechanism being the ability to model long-range spatial correlations within a single convolutional layer rather than building them incrementally across many stacked 3×3 layers.

The difference in ERF growth between the two architectures is illustrated in Figure 4.

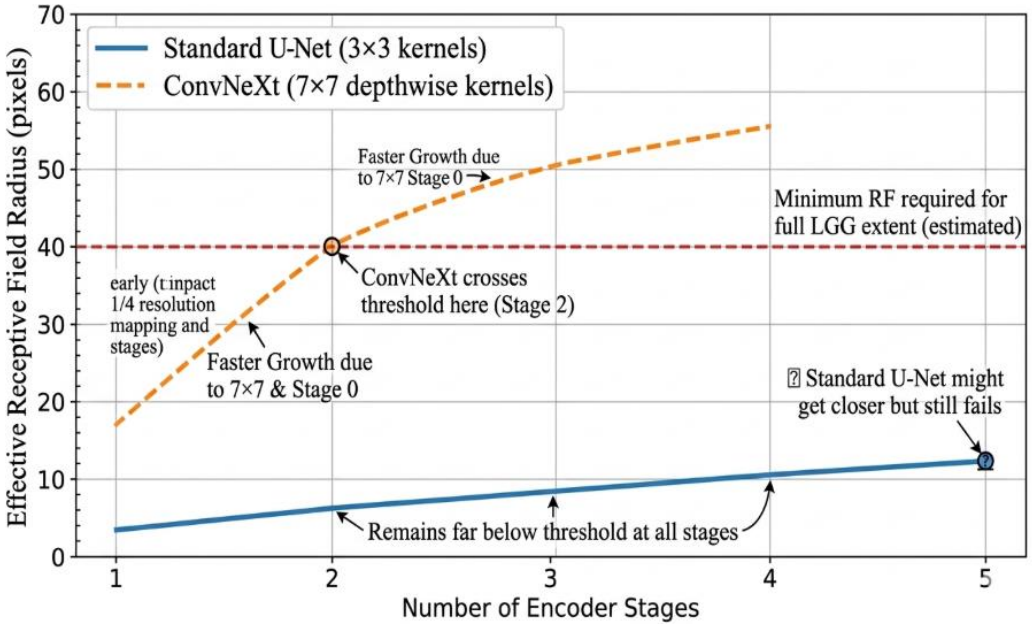

Figure 4. Comparison of effective receptive field growth between standard U-Net and ConvNeXt-UNet architectures.

The functional analogy to self-attention is instructive even where the mathematical correspondence is approximate. In a self-attention head, every position attends every other position in the input sequence, providing global context in a single operation. A 7×7 depthwise convolution provides local context over a 49-pixel neighborhood not global, but notably larger than the 9-pixel neighborhood of 3×3 convolution and when applied at coarser spatial resolutions where each feature map pixel corresponds to a larger region of the original input, the effective spatial context is comparable to what shallower attention windows achieve. This enlarged-kernel convolutional approach provides the contextual awareness needed for diffuse tumour boundary segmentation without the memory overhead and data requirements of full self-attention.

2.4.3 Inverted Bottleneck Structure

The ConvNeXt block adopts an inverted bottleneck structure inspired by the transformer feed-forward sub-network. In the standard ResNet bottleneck block, the channel dimension is contracted (typically to C/4) before the spatial convolution and expanded back to C after, reducing the computational cost of the 3×3 convolution at the expense of information capacity. The ConvNeXt inverted bottleneck reverses this: the channel dimension is expanded fourfold (from C to 4C) after the depthwise convolution, followed by GELU activation and a second

1×1 projection back to C. This mirrors the feed-forward network in a Transformer block, where the multilayer perceptron (MLP) hidden dimension is 4× the model dimension.

The practical consequence of the inverted bottleneck is that channel-wise feature mixing occurs in a high-dimensional space (4C) that provides richer non-linear combinations than a standard bottleneck operating in a contracted space (C/4). For a ConvNeXtTiny stage with C=96 channels, the inverted bottleneck operates on a 384-dimensional hidden representation, 16 times wider than a 3×3 convolutional layer would naturally produce. This high-dimensional mixing step allows the block to learn more diverse and expressive feature combinations from the spatial patterns aggregated by the 7×7 depthwise step. The depthwise operation handles spatial mixing in the original C-dimensional channel space; the inverted bottleneck handles channel mixing in the expanded 4C space; together they implement the spatial-then-channel mixing sequence that characterizes the transformer block.

The combination of depthwise spatial mixing at 7×7 with inverted bottleneck channel mixing is computationally efficient: the total floating-point operation (FLOP) count of a ConvNeXt block is comparable to a standard ResNet block at the same spatial resolution and channel depth, despite the larger spatial kernel and wider channel expansion, because the depthwise operation avoids the C^2 parameter scaling of standard convolution.

2.4.4 Layer Normalization and GELU Activation

Two additional modifications in the ConvNeXt block address normalization and activation. Batch Normalization (Ioffe & Szegedy, 2015), the normalization scheme used in ResNet and standard U-Net architectures, normalizes each channel across the batch and spatial dimensions. When batch sizes are large (≥32–64), this produces stable normalization statistics and has regularization properties that improve generalization. However, at the small batch sizes typical in GPU-constrained medical image segmentation, batch sizes of 8–16 when processing high-resolution images with deep models, batch statistics are noisy and Batch normalization can degrade training stability and final performance. Layer Normalization (Ba et al., 2016) normalizes across the channel dimension independently for each sample and spatial position, computing statistics from the C feature values at a single spatial location. This makes Layer Normalization entirely batch-size-independent: it computes identical statistics whether the batch contains 1 or 64 samples, and it behaves identically during training and inference without requiring running mean/variance estimates.

For the experimental setting of this thesis, batch sizes of 8 for the ConvNeXt-UNet on a 16 GB GPU, Layer Normalization provides a decisive practical advantage over Batch Normalization. The behaviour of Batch Normalization at batch size 8 is dominated by within-batch variability rather than population statistics, producing noisy gradient updates that can slow or destabilize convergence. This was noticeable in practice: when the ConvNeXt-UNet was tested with batch size 32 on a smaller memory-limited run, training was marginally faster, but the validation DSC trajectories were less stable epoch-to-epoch. Reducing to batch size 8 with Layer Normalization produced smoother and ultimately higher validation DSC curves. The standard U-Net baseline, which uses Batch Normalization at batch size 32, operates near the lower limit of reliable batch statistics. The ConvNeXt-UNet's use of Layer Normalization means that its training dynamics are robust to the batch size constraints imposed by the larger model footprint.

The GELU (Gaussian Error Linear Unit) activation (Hendrycks & Gimpel, 2016) is defined as $\mathrm{GELU}(x) = x \cdot \Phi(x)$, where Φ is the cumulative distribution function of the standard normal distribution. It is approximated in practice as

$$GELU(x) \approx 0.5x\left[1 + \tanh\left(\sqrt{\left(\frac{2}{\pi}\right)}(x + 0.044715x^3)\right)\right]$$

Unlike ReLU which applies a hard zero-threshold, producing exactly zero output and zero gradient for all negative inputs, GELU attenuates negative inputs smoothly and maintains small but non-zero gradients for moderately negative values. This smooth non-linearity is associated with improved convergence behaviour when training with Adam and related adaptive optimizers, particularly in deep networks where gradient flow through many layers is necessary (Liu et al., 2022). The transition from ReLU to GELU in the ConvNeXt block contributed a measurable accuracy improvement in the Liu et al. (2022) ablation study and has been adopted as the standard activation in BERT, GPT, and most modern large language and vision models. The structural comparison between the standard block and the ConvNeXt block is shown in Figure 5.

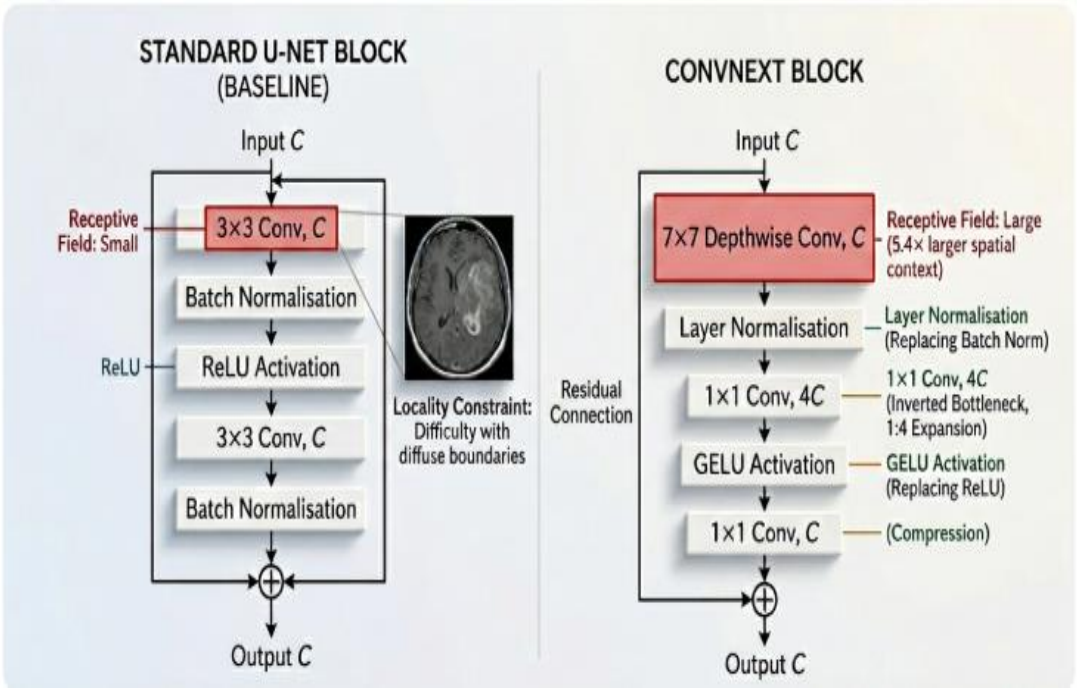

Figure 5. Comparison of the standard 3×3 convolutional block (left) and the ConvNeXt block (right). Key Differences Annotated: 7x7 depthwise convolution, 1:4 inverted bottleneck expansion, Layer. Normalisation replacing Batch Normalisation, GELU replacing ReLU, and residual addition. The effective spatial context (red) is 5.4x larger in the ConvNeXt block.

2.5 ConvNeXt in Medical Image Segmentation: Prior Work

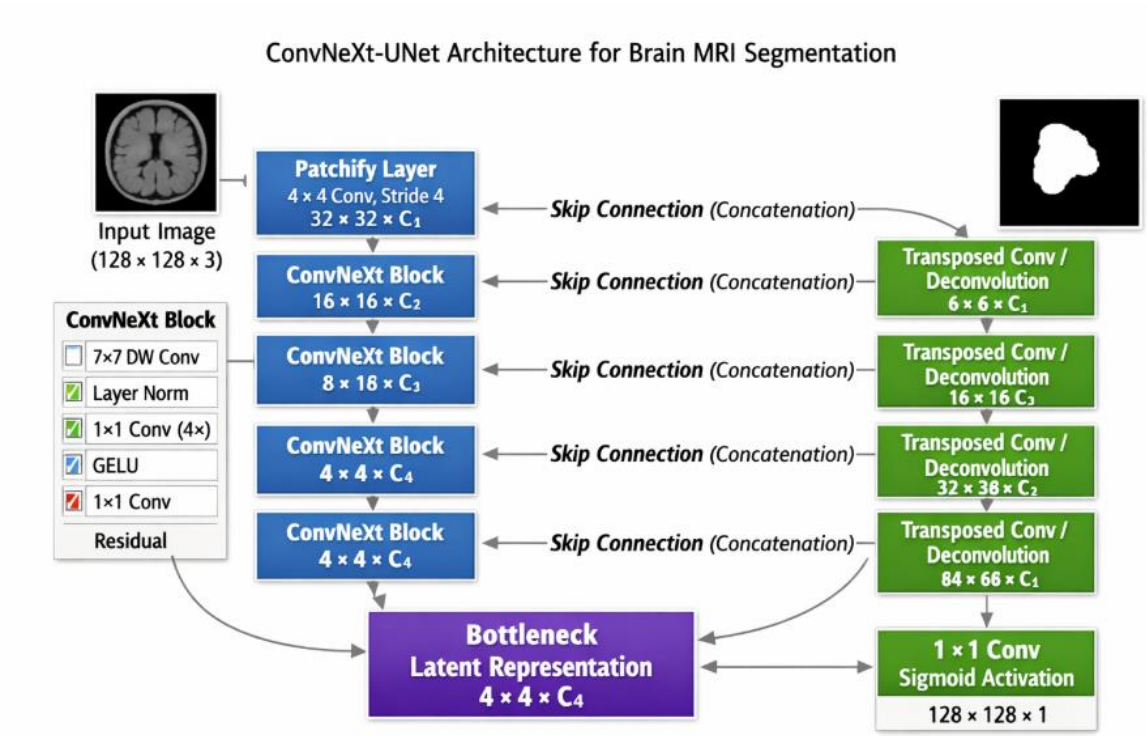

Figure 6. Hierarchical Architecture of the ConvNeXt Encoder.

Since 2022, a growing number of studies have adapted ConvNeXt principles to medical segmentation — and the pattern in the published literature is strikingly consistent: ConvNeXt-based models outperform standard U-Net baselines, every time, on every modality tested. MedNeXt (Roy et al., 2023) is the most comprehensive application of ConvNeXt modernization to volumetric medical segmentation. The authors replaced the 3×3×3 convolutional kernels of a U-Net style backbone with anisotropically scaled depthwise separable kernels up to 7×7×7, and progressively scaled kernel size (referred to as UpKern) during training for regularization. Evaluated across brain tumour, cardiac, and abdominal CT segmentation tasks from the Medical Segmentation Decathlon, MedNeXt demonstrated consistent multi-task improvements over the nnU-Net baseline, providing the most systematic pre-existing evidence that enlarged depthwise kernels confer genuine advantages in medical segmentation beyond single-task optimization. Figure 6 shows the hierarchical structure of the ConvNeXt encoder used in this study, with the four progressive stages and their output feature map dimensions.

(Lee et al. 2022)proposed 3D UX-Net, applying large-kernel volumetric depthwise convolutions in a hierarchical encoder–decoder framework for multi-organ CT segmentation. The study demonstrated that a lightweight convolutional model incorporating 7×7×7 depthwise kernels could approach Swin Transformer performance with notably fewer parameters and lower memory requirements, directly validating the central claim of the ConvNeXt approach in a 3D medical context.

Dong et al. (2024) proposed CI-UNet, combining a ConvNeXt encoder with cross-dimensional attention mechanisms applied along both spatial and channel dimensions, and reported improvements on skin lesion, polyp, and retinal vessel segmentation benchmarks. Zhang et al. (2023) examined the relationship between receptive field size and segmentation accuracy specifically for diffuse brain tumours, providing direct evidence that models with larger effective receptive fields produce more accurate peripheral delineation by accessing the broader spatial context needed for boundary classification in gradient-boundary targets.

Kamsari et al. (2025) investigated the interaction between ConvNeXt architectural choices and data augmentation intensity in breast ultrasound tumour segmentation, finding that the ConvNeXt advantage over standard U-Net was most pronounced in low-data regimes, a finding relevant to the Task B context of this study. Reading through these studies, a pattern becomes clear: every published result shows ConvNeXt improving over U-Net, but none of the

papers test on a task where the conditions are not already favourable — compact, well-annotated, single-modality MRI or dermoscopy. The Task B design in this thesis was partly motivated by wanting to test the architecture where those conditions do not hold.

Across the ConvNeXt medical segmentation literature, a consistent pattern of agreement emerges on two points: enlarged depthwise kernels reliably improve segmentation of diffuse or poorly bounded targets relative to 3×3 baselines, and this improvement is most pronounced when the training dataset is small, and the target structure requires contextual reasoning over a large spatial extent. Lee, (2022), Roy et al. (2023) and Zhang et al. (2023) all report the strongest gains on tasks where the lesion boundary spans a large fraction of the image, brain tumours, abdominal organs while reporting smaller or less consistent improvements on compact, well-circumscribed structures. Kamsari et al. (2025) specifically confirmed that the ConvNeXt advantage is amplified in low-data regimes, suggesting that enlarged kernels compensate for the reduced statistical power of limited training sets by encoding stronger spatial priors. Not one of these studies includes a task where ConvNeXt performs worse than the U-Net baseline. That consistency is itself suspicious — it suggests a publication-bias risk where negative results are simply absent from the record.

The Task B finding of this thesis — where the U-Net significantly outperforms the ConvNeXt on CT — is a result that the existing literature provides no basis for predicting. The Task B finding of this study, where the U-Net outperforms the ConvNeXt-UNet on single-modality CT is therefore a contribution that the existing pattern of results does not predict and that the literature review cannot fully contextualize from published evidence alone. The present study addresses this gap by evaluating a ConvNeXt-UNet against a standard U-Net on both FLAIR MRI and CT under identical paired experimental conditions.

Across this body of work, three questions remain unanswered. First, all evaluations are single-modality: no study places a ConvNeXt-based segmentation architecture and a standard U-Net in direct paired comparison across two different imaging modalities under identical conditions, meaning the claimed advantages cannot be attributed to architecture alone rather than to modality-specific inductive biases. Second, none of these studies test ConvNeXt architecture in conditions where it might fail — the published dataset characteristics (MRI-family modalities, diffuse or gradient-boundary targets, moderate dataset sizes) are uniformly favourable to ConvNeXt's design properties. Third, methodological rigor is inconsistent: several studies report mean DSC differences without formal hypothesis testing and without

verifying that both models were evaluated on the same patient subsets (Varoquaux & Cheplygina, 2022). These three gaps directly motivated the experimental design described in Chapter 3.

2.6 ConvNeXt V2 and Self-Supervised Pre-training

One direction that the ConvNeXt literature has pursued independently of the segmentation application is improved pre-training through self-supervised learning, which is worth reviewing here because it directly motivates the choice of ConvNeXtTiny over ConvNeXt V2 in this study. ConvNeXt V2 (Woo et al., 2023) extended the original ConvNeXt framework by introducing a fully convolutional masked autoencoder (FCMAE) framework for self-supervised pre-training. The original ConvNeXt was pre-trained in a supervised manner on ImageNet-1K or ImageNet-22K; ConvNeXt V2 applied the masked autoencoder pre-training paradigm introduced for transformers by He et al. (2022) to the convolutional framework, randomly masking patches of the input and training the network to reconstruct the masked regions. To make this effective for convolutional networks which, unlike transformers, cannot simply ignore masked tokens, the authors introduced a novel Global Response Normalization (GRN) layer that normalizes channel responses across all spatial positions, preventing the encoder from trivially propagating visible-region features into masked regions through local convolutions.

The practical implication of ConvNeXt V2 for medical image segmentation is that self-supervised pre-training on large unlabelled medical imaging collections may become feasible without requiring transformer architectures. Masked autoencoder pre-training has shown promise for learning rich feature representations from unlabelled medical images (Zhou, 2023), and the ability to apply this paradigm to a convolutional architecture that can be fine-tuned with the efficiency advantages of ConvNeXt is potentially significant for the label-scarce medical imaging domain.

For the experiments in this thesis, the original ConvNeXtTiny backbone with ImageNet-1K supervised pre-training is used rather than ConvNeXt V2, for two reasons. First, ConvNeXtTiny is the smallest and most memory-efficient member of the ConvNeXt family, making it the most appropriate choice for a GPU-constrained academic setting. Second, the use of ConvNeXt V2 with GRN would introduce an additional architectural variable that is not the focus of this study; the experimental question concerns the basic ConvNeXt block design (7×7 depthwise convolution, inverted bottleneck, Layer Normalization, GELU) relative to the

standard 3×3 U-Net, not the incremental effect of GRN. ConvNeXt V2 represents a natural extension of the architecture studied here and is discussed as a future direction in Chapter 6.

2.7 Loss Functions for Class-Imbalanced Segmentation

Class imbalance is a pervasive challenge in medical image segmentation. In both tasks studied in this thesis, tumour pixels represent a small fraction of the total image area: for Task A, the LGG tumour occupies approximately 5–15% of the FLAIR slice area in tumour-bearing slices; for Task B, the combined GTVp and GTVn occupy an even smaller fraction of the head and neck CT volume per axial slice. Standard binary cross-entropy (BCE) loss penalizes each pixel equally and independently:

$$L\_BCE = -[y\ log(\hat{y}) + (1-y)\ \ log(1-\hat{y})]$$

averaged over all N pixels. Under severe class imbalance, a network that predicts background everywhere achieves a near-minimal BCE loss (because the large number of correctly classified background pixels dominates the sum), providing no meaningful gradient signal toward tumour detection (Lin et al., 2017). Figure 7 compares BCE and Dice loss behaviour under severe class imbalance, showing why BCE alone produces degenerate all-background predictions.

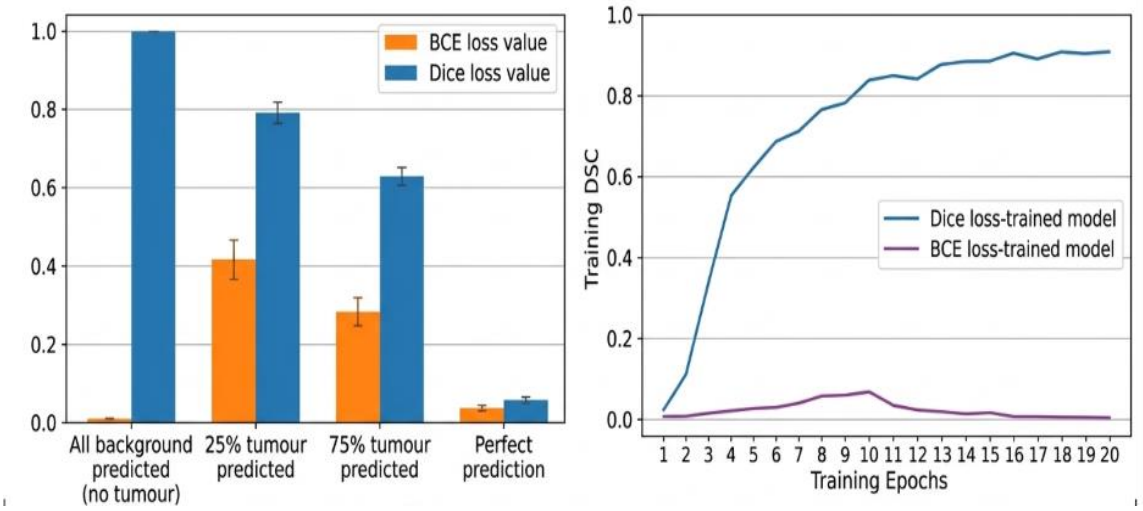

Figure 7. Comparison of loss function behaviour under severe class imbalance. The left panel shows that Binary Cross-Entropy (BCE) produces a deceptively low loss for "all-background" predictions, failing to provide a useful training signal. In contrast, Dice loss maintains a maximal penalty (1.0), forcing the model to recognize the minority tumour class. The right panel demonstrates that while BCE-only training often collapses into a degenerate all-background solution (zero DSC), Dice-based training successfully converges to a meaningful segmentation result by prioritizing the overlap of the positive class.

The Dice loss, introduced by Milletari et al. (2016) in the V-Net paper, addresses this directly by computing the loss as a function of the overlap between the predicted and ground-truth foreground regions:

$$L\_Dice = 1 - (2 \cdot \Sigma y \cdot \hat{y} + \varepsilon) / (\Sigma y + \Sigma \hat{y} + \varepsilon),$$

where the summations are over all pixels and ε is a smoothing constant (typically 1.0) that prevents division by zero on empty slices. The Dice loss is class-imbalance robust by construction: by normalizing the overlap by the total predicted and ground-truth positive area, it provides equivalent gradient signal regardless of the ratio of foreground to background pixels. It directly optimizes the primary evaluation metric (DSC) and has become the standard training objective for tumour segmentation tasks.

A limitation of pure Dice loss is training instability in the earliest epochs, when predictions are near-uniform and the denominator of the Dice formula is dominated by random noise, providing weak and potentially misleading gradient signals. Combining Dice loss with BCE loss, a hybrid

$$L = L\_BCE + \lambda L\_Dice$$

leverages the pixel-wise gradient stability of BCE in early training while retaining the class-imbalance robustness of Dice as the foreground region is progressively localized. In the ConvNeXt-UNet training for Task A, the hybrid

$$loss\ L = L_{BCE} + \lambda L_{Dice}$$

was used; for Task B, the hybrid included a Dice weighting factor of 2:

$$L = L_{BCE} + 2L_{Dice}$$

For the standard U-Net on Task B, pure Dice loss was used. These choices reflect empirical experimentation with training stability under the specific class imbalance and batch size conditions of each task.

Focal loss (Lin et al., 2017), developed in the context of object detection, modifies BCE to down-weight the contribution of well-classified easy examples and up-weight the contribution of hard misclassified examples:

$$L\_Focal = -(1 - \hat{y})\hat{}\gamma \cdot y\, log(\hat{y}) - \hat{y}\hat{}\gamma \cdot (1 - y)\, log(1 - \hat{y}).$$

The focusing parameter γ (typically 2) causes the loss to concentrate on the hard boundary pixels that are most informative for training. Focal loss has shown advantages in highly imbalanced settings and represents a natural complement to the Dice-based objectives used in this study; it is identified as a direction for future investigation in Chapter 6.

2.8 Evaluation: The Dice Similarity Coefficient

The DSC is the universal evaluation metric for medical image segmentation, adopted across the BraTS challenge series (Baid et al., 2021; Menze et al., 2015), the HECKTOR challenge (Andrearczyk et al., 2023) and the Medical Segmentation Decathlon (Antonelli et al., 2022). It measures the spatial overlap between a predicted binary mask P and a ground-truth binary mask G as:

$$DSC = ((2|P \cap G| + \varepsilon))/(|P| + |G| + \varepsilon),$$

where |P| and |G| denote the number of positive pixels in each mask and ε is a smoothing constant (1.0 throughout this study). DSC ranges from 0 (no overlap) to 1 (perfect overlap), with the smoothing constant ensuring a value of 1.0 when both masks are empty, appropriate for slices without tumour that are retained in the test set.

The DSC is class-imbalance robust in evaluation for the same reason that Dice loss is imbalance robust in training: it normalizes by the positive area in both prediction and ground truth, rather than by the total image area. A model that correctly identifies a small tumour in a predominant background image receives the same DSC as one that identifies a large tumour, provided the overlap is proportionally equivalent.

Clinically, DSC thresholds for automated segmentation acceptability vary by target and context. For brain tumour segmentation in the BraTS multi-modal setting, whole-tumour DSC above 0.88 is generally considered comparable to expert inter-reader agreement (Menze et al., 2015). For single-modality FLAIR segmentation, a more constrained information setting, thresholds for clinical acceptability have not been formally established, but the Task A ConvNeXt-UNet result of 0.9328 clearly falls within the range where clinical deployment would be considered viable based on accuracy alone. For head and neck GTVp segmentation on CT, published automated systems consistently report DSC below 0.65 even with PET information, reflecting the inherent difficulty of the task; the Task B results of this study (U-Net 0.3000, ConvNeXt-UNet 0.2558) are therefore consistent with what is achievable under the demanding single-modality non-contrast CT condition.

2.9 Statistical Validation in Machine Learning Comparisons

The valid statistical comparison of two machine learning models on paired data requires careful attention to the distributional properties of the performance metric, the nature of the pairing, and the sample size available for testing.

The Shapiro-Wilk test (Shapiro & Wilk, 1965) provides a powerful omnibus test for normality in samples up to n=5,000, evaluating the null hypothesis that the sample was drawn from a normal distribution. The test statistic W is defined as the squared correlation between the ordered sample values and the expected order statistics of a standard normal distribution:

$$W = (\Sigma a_i x_i)^2 / \Sigma (x_i - \bar{x})^2$$

where, a_i are coefficients derived from the normal order statistics. Values of W close to 1 indicate normality; significant departures from 1 (low p-values) indicate rejection of the normality hypothesis. For Task A of this study, the Shapiro-Wilk test applied to per-image DSC distributions returned $p < 0.001$ for both models, indicating that parametric tests assuming Gaussian distributions are inappropriate and that a non-parametric alternative must be used.

The Wilcoxon signed-rank test (Wilcoxon, 1945) is the appropriate non-parametric alternative to the paired t-test for paired observations when normality cannot be assumed. It ranks the absolute values of the pairwise differences, assigns the sign of each original difference to its rank, and computes the test statistic T as the sum of positive ranks. Under the null hypothesis of no difference, the distribution of T is known. With n=2,186 paired per-image DSC observations in Task A, the test has effectively unity statistical power: any true difference in median DSC will be detected with certainty at any conventional significance level.

The paired t-test is appropriate for small samples when paired differences can be assumed to be normally distributed. With n=5 fold-level DSC pairs in Task B, the Shapiro-Wilk test on the pairwise differences returned $p = 0.6781$, failing to reject normality, and the paired t-test is the primary inferential test for Task B. The test statistics

$$t = \frac{\overline{d}}{s_d/\sqrt{n}}$$

with $n-1 = 4$ degrees of freedom, follows a t-distribution under the null hypothesis. The Wilcoxon signed-rank test was additionally computed for Task B as a non-parametric

robustness check; its minimum achievable p-value at n=5 is 0.0625 (when all differences are in the same direction), providing a distributional complement to the t-test.

The use of patient-wise GroupKFold cross-validation (rather than slice-level random splitting) is essential for both the validity of fold-level performance estimates and the validity of statistical comparison. Slice-level splitting would expose adjacent spatially correlated slices from the same patient to both training and validation sets, inflating DSC estimates by allowing the model to partially learn patient-specific anatomical features.

The requirement that both models be evaluated on identical fold partitions enforced by deriving fold assignments from a single call to GroupKFold.split() and applying the same index arrays to both training loops ensures that fold-level DSC comparisons are genuinely paired and that the t-test or Wilcoxon test is applied to observations that differ only in model architecture.

Despite the widespread use of cross-validation in medical segmentation benchmarking, the statistical rigor of published architectural comparisons is frequently insufficient for the inferential claims being made. A review of published ConvNeXt medical segmentation studies reveals that the majority report mean DSC across folds without formally testing whether the observed difference from the U-Net baseline is statistically significant (Dong et al., 2024; Lee, 2022; Roy et al., 2023). More importantly, several published comparisons use slice-level splitting rather than patient-wise splitting for cross-validation, potentially inflating DSC estimates through inter-slice spatial correlation leakage, a source of optimistic bias that Maier-Hein et al. (2024) identified as one of the most prevalent methodological weaknesses in medical AI evaluation.

This methodological weakness has been documented across multiple medical imaging subfields as a systematic source of over-optimistic performance estimates (Varoquaux & Cheplygina, 2022). The selection of a statistical test based on distributional assessment, using the Shapiro-Wilk test to determine whether a parametric or non-parametric comparison is appropriate, is rarely reported explicitly. This methodological gap means that published performance comparisons in the ConvNeXt medical segmentation literature cannot always be taken at face value, and that a study designed from the outset with rigorous statistical methodology represents a distinct methodological contribution beyond its empirical findings.

Three gaps in existing literature directly shaped the design choices of this study. First, While ConvNeXt-based architectures have been applied to medical image segmentation, most comparative studies with U-Net are conducted under differing datasets, modalities, and

evaluation protocols. Metrics such as Dice and IoU are reported as overall averages, often without fully standardised paired experimental setups that would isolate architectural effects from dataset variability. As a result, the current evidence remains fragmented, and consistent generalization of ConvNeXt's performance advantages across imaging modalities cannot be firmly established. Second, the performance of modern convolutional architectures on single-modality non-contrast CT for HNSCC segmentation is largely uncharacterised in the literature, because published approaches almost universally use PET-CT multi-modal input, a resource that is not available in all clinical settings and that confounds evaluation of the architecture's contribution relative to the additional modality's information. Third, the statistical rigor of architectural comparison studies in medical segmentation remains inconsistent: many published evaluations report mean DSC differences without formal hypothesis testing, without verifying that both models were evaluated on identical patient subsets, and without selecting statistical tests based on prior distributional assessment. This thesis addresses all three gaps through a controlled cross-modal design, CT-only Task B evaluation, and a normality-guided statistical testing protocol that makes the inferential framework explicit. Whether the results support or refute the assumption of universal ConvNeXt superiority is a question that only a study designed to test it can answer.

3. Materials and Methods

This chapter describes the complete experimental methodology employed across both segmentation tasks. The chapter covers the datasets used, the preprocessing decisions applied to each imaging modality, the architectures of the two competing models, the training configuration, and the cross-validation and statistical framework used to obtain the results. Where the two tasks share a procedure, it is described once; where they diverge the divergence is made explicit, together with the reasoning that motivated it.

3.1 Datasets

3.1.1 Task A: Brain Tumour MRI (TCGA-LGG)

The brain tumour segmentation task uses the Pre-Operative TCGA LLG Neuroimaging Informatics Technology Initiative (NIfTI) and Segmentations collection, accessed through The Cancer Imaging Archive (TCIA) and released under a Creative Commons Attribution 3.0 licence. The collection comprises pre-operative multimodal MRI studies from 110 patients

with histologically confirmed WHO grade II or III glioma. Each study contains T1-weighted, gadolinium-enhanced T1, T2-weighted, and FLAIR volumes stored in NIfTI format, accompanied by binary tumour masks produced by the GlistrBoost probabilistic segmentation algorithm and subsequently corrected by a board-certified neuroradiologist (Bakas, Akbari, Sotiras, et al., 2017).

Only the FLAIR sequence was used. FLAIR suppresses CSF signal while preserving the T2 hyperintensity associated with tumour infiltration and perilesional oedema, making it the most informative single sequence for LLG delineation in the absence of gadolinium enhancement. After extracting only tumour-bearing axial slices, the dataset yielded 10,929 slices. The 110 patients were partitioned patient-wise into an 88-patient training and validation pool (approximately 8,743 slices) and a strictly held-out test set of 22 patients (approximately 2,186 slices) using a fixed random seed of 42. The test set was not examined until all model configurations were finalised.

3.1.2 Task B: Head and Neck CT (HECKTOR 2025)

The head and neck segmentation task draws on the HECKTOR 2025 challenge dataset, which provides multi-centre PET-CT acquisitions with expert-annotated gross tumour volume labels for head and neck squamous cell carcinoma (Andrearczyk et al., 2023). The study uses the MD Anderson Cancer Centre cohort exclusively, comprising 396 patients identified by the naming convention MDA-*. After tumour-bearing slice filtering, this cohort yielded 7,774 axial CT slices. Only the CT volume was used; the PET channel available in the dataset was deliberately excluded in order to evaluate the architecture under the most demanding single-modality CT conditions, and to keep the modality-comparison between tasks as clean as possible. The HECKTOR 2025 CT acquisitions are non-contrast-enhanced low-dose scans acquired on combined PET-CT scanners, meaning that tumour tissue is frequently isodense to adjacent soft tissue structures — a challenging condition that tests architectural robustness directly.

The label file for each patient encodes the primary gross tumour volume (GTVp, label 1) and nodal gross tumour volumes (GTVn, label 2) separately. Both structures collapsed into a single binary foreground class by the operation (label > 0), consistent with the HECKTOR challenge baseline evaluation approach. Task B was evaluated exclusively under 5-fold cross-validation; no separate held-out test set was withheld given the smaller effective dataset relative to Task A. The key characteristics of both datasets are summarised in Table 1.

Table 1: Dataset characteristics for Task A (Brain MRI) and Task B (Head and Neck CT).

Property	Task A — TCGA-LGG	Task B — HECKTOR 2025 MDA
Modality	FLAIR MRI	CT (non-contrast)
Total patients	110	396
Tumour-bearing slices	10,929	7,774
Training / test split	88 patients train + 22 patients held-out test	5-fold CV only (no held-out test)
Ground truth	GlistrBoost + expert correction	GTVp + GTVn binarised
Data licence	CC BY 3.0 (TCIA)	HECKTOR 2025 challenge terms

3.2 Data Preprocessing

3.2.1 Task A: FLAIR MRI

NIfTI volumes were loaded using nibabel. FLAIR volumes were identified by substring matching on the filename ('flair' in filename.lower()), and expert-corrected masks by the substring 'manuallycorrected'. The three-dimensional volumes were traversed axially and only tumour-bearing slices (mask.max() > 0) were retained. GlistrBoost probability masks were binarized at a threshold of 0.5. Each retained image slice was resized to 128×128 pixels using bilinear interpolation; the corresponding mask was resized using nearest-neighbor interpolation to preserve its binary character. Per-slice min-max normalization was applied:

$$I\_norm = (I - I\_min) / (I\_max - I\_min + \varepsilon)$$

The pre-trained ConvNeXt encoder expects a three-channel (RGB) input, so each normalised grayscale slice was replicated across three channels using np.repeat(slice, 3, axis=−1), yielding tensors of shape (128, 128, 3). This grayscale-to-RGB replication is the standard approach for applying ImageNet-pre-trained models to single-channel medical images (Tajbakhsh et al., 2016). No data augmentation was applied for Task A.

In this study, a technical issue was encountered during early preprocessing that is worth documenting for reproducibility. The initial pipeline stored normalised image arrays as float16 to reduce memory consumption. When these float16 arrays were passed as training labels to TensorFlow/Keras 3, the model compilation step threw an OptionalFromValue error that originated in the framework's internal graph construction. The error was traced to a dtype incompatibility between float16 labels and the expected float32 computation graph. The fix was straightforward but non-obvious: all arrays were explicitly cast to float32 before the training loop using X.astype('float32') and Y.astype('float32') and run_eagerly=True was enabled temporarily during debugging. This experience reinforced the importance of explicit dtype management in TensorFlow pipelines, particularly when mixing ImageNet-pretrained models with medical imaging data.

The choice to replicate the grayscale channel rather than reinitialize the backbone's first convolutional layer was deliberate: replication preserves all pre-trained weights intact, whereas reinitialization would require re-learning from scratch the edge and texture detectors that the ImageNet pre-training provides, using only the limited medical imaging training set. This approach is consistent with the transfer learning practice recommended by Tajbakhsh et al. (2016).

3.2.2 Task B: CT

The CT volume for each patient was identified by the filename pattern MDA-XXX__CT.nii and the label file by MDA-XXX.nii. The same axial traversal and tumour-bearing slice filter (np.sum(mask_slice) > 0) were applied. Multi-label masks were binarized using the operation (label > 0), collapsing both GTVp and GTVn into a single binary foreground class. Spatial resampling to 128×128 used identical interpolation settings to Task A. Per-slice min-max normalization was applied rather than a fixed Hounsfield Unit window, because the head and neck CT volumes contain high-attenuation dental artefacts that would compress soft-tissue contrast into a narrow band under a fixed window. The CT encoder accepts single-channel input; no channel replication was performed. Data augmentation was applied during Task B training: rotation range ±15°, width and height shift ±10%, and horizontal flipping, applied to images and masks in synchronized fashion using a shared random seed to prevent spatial desync.

3.3 Model Architectures

All models were implemented in Python 3.12 using TensorFlow with the Keras API. The standard U-Net was constructed using the Keras functional API. The ConvNeXtTiny backbone for the ConvNeXt-UNet was loaded via tf.keras.applications.ConvNeXtTiny with ImageNet-1K pre-trained weights. Statistical analyses were performed using scipy.stats, and cross-validation was implemented using scikit-learn's GroupKFold. Key package versions used were: numpy 2.0.2, nibabel 5.3.3, and scikit-learn (Pedregosa, Varoquaux, et al., 2011). All experiments were run in the Kaggle notebook environment on an NVIDIA Tesla T4 GPU (13,757 MB VRAM).

3.3.1 Standard U-Net (Baseline)

The standard U-Net (Ronneberger et al., 2015) serves as the baseline for both tasks. The encoder comprises three convolutional blocks at filter depths of 64, 128, and 256, each containing two Conv2D (3×3)–BatchNorm–ReLU layers with MaxPooling between stages. The decoder applies transposed convolutions for upsampling, concatenating the corresponding encoder feature maps via skip connections at each resolution level. A final Conv2D (1×1) layer with sigmoid activation produces the binary segmentation probability map of shape (128, 128, 1). For Task B, the model accepts single-channel CT input of shape (128, 128, 1).

In this study, the standard U-Net was first trained for Task A as a diagnostic baseline before any architectural changes were considered. The model converged reliably within the first ten epochs but then stalled at a validation DSC of approximately 0.54–0.57 across all folds. Adjusting the learning rate, increasing training epochs to 50, and switching between BCE-only and Dice-only loss objectives produced no consistent improvement beyond this ceiling. This plateau behaviour was the direct empirical motivation for exploring the ConvNeXt encoder: it confirmed that the limitation was architectural rather than a training procedure issue.

Moving to 2D was forced by GPU memory — early attempts at 3D processing on the Kaggle Tesla T4 ran out of memory before a single training fold could complete. The practical upside was that 2D processing made slice-level augmentation during Task B training much simpler to implement.

3.3.2 ConvNeXt-UNet (Proposed)

The ConvNeXt-UNet integrates the architectural modernization principles introduced by Liu et al. (2022) into the U-Net framework. The encoder is the publicly available ConvNeXtTiny

backbone, loaded with ImageNet-1K pre-trained weights for Task A. For Task B, since the single-channel CT input is incompatible with the three-channel input expected by the ImageNet-pre-trained weights, a 1×1 convolutional projection layer converts the single CT channel to three channels prior to the backbone, allowing the same pre-trained ConvNeXtTiny encoder to be applied.

Skip connections are extracted from three intermediate ConvNeXt stages at spatial resolutions of 32×32×96, 16×16×192, and 8×8×384. The decoder mirrors the U-Net design: each stage applies a transposed convolution for upsampling, concatenates the corresponding encoder skip map, and applies two convolutional blocks. Layer Normalization and GELU activation are used throughout the decoder, consistent with the ConvNeXt design philosophy. The output head is an identical Conv2D (1×1) with sigmoid activation.

During initial implementation, an attempt was made to train the ConvNeXt-UNet end-to-end from epoch one without a frozen warm-up phase. This produced a consistent failure pattern across all five Task A folds: training DSC continued to increase while validation DSC plateaued and then declined after approximately eight epochs, a clear signature of overfitting driven by the large gradient magnitudes from the randomly initialized decoder corrupting the pre-trained encoder weights. The two-phase training strategy was developed in direct response to this observation. In Phase 1 (5 epochs), the backbone is frozen and only the decoder weights are updated, allowing the decoder to reach a stable functional state before the encoder is exposed to decoder-generated gradients. In Phase 2 (up to 40 epochs), the full model is unfrozen and fine-tuned end-to-end at a reduced learning rate of 1×10^{-5}, ensuring that the pre-trained encoder features are updated gradually rather than disrupted by large early-epoch gradient steps.

3.3.3 Architectural Comparison: Design Decisions

The two architectures differ in four principal respects that are relevant to interpreting the experimental results. First, the encoder depth and parameter count differ notably: the standard U-Net has approximately 2.4 million parameters, while the ConvNeXt-UNet, incorporating the ConvNeXtTiny backbone with approximately 28 million parameters, has approximately 32 million parameters in total including the decoder. This difference in model capacity means that the ConvNeXt-UNet is more expressive but also more susceptible to overfitting when training data is limited, a factor that is relevant to the Task B results.

Second, the ERF at the bottleneck differs significantly between the two architectures. As established in Chapter 2, while the theoretical receptive field of a standard 3×3 convolutional

encoder grows linearly, the ERF expands only at a rate proportional to the square root of the number of layers, following a Gaussian distribution (Luo et al., 2016). In contrast, the ConvNeXt-Tiny encoder utilizes 7×7 depthwise convolutions applied hierarchically across four stages. This architectural choice provides a notably larger ERF at the corresponding bottleneck depth, enabling the model to integrate global spatial context from the original input more effectively than the standard U-Net baseline.

Third, the normalization strategies differ: the standard U-Net employs Batch Normalization (BN) throughout the architecture, whereas the ConvNeXt-UNet utilizes Layer Normalization (LN) within the decoder to maintain architectural consistency with the ConvNeXt backbone. This difference has practical implications for small batch sizes: at a batch size of 8, Batch Normalization statistics are estimated from only 8 samples (Ioffe & Szegedy, 2015), which introduces noise that Layer Normalization avoids entirely (Ba et al., 2016).

Fourth, the activation functions differ: the standard U-Net uses ReLU, while the ConvNeXt-UNet uses GELU in the decoder. The practical consequence for training is that GELU provides smooth gradients for negative activations, potentially improving gradient flow through the decoder during backpropagation. The complete architectural comparison across both models and both tasks is summarised in Table 2.

Table 2: Architecture summary for both models across both tasks.

Component	U-Net (both tasks)	ConvNeXt-UNet Task A	ConvNeXt-UNet Task B
Input	128×128×1	128×128×3 (grey×3)	128×128×1 → 3 via 1×1 conv
Encoder	3 stages, 3×3 kernels, BN-ReLU	ConvNeXtTiny (ImageNet)	ConvNeXtTiny (ImageNet, via 1×1)
Skip connections	3 (at 64×64,32×32, full res)	3 ConvNeXt stage outputs	3 ConvNeXt stage outputs
Decoder norm	Batch Norm	Layer Norm + GELU	Layer Norm + GELU

Output	Conv2D (1,1×1) + Sigmoid	Conv2D (1,1×1) + Sigmoid	Conv2D (1,1×1) + Sigmoid
Parameters	~2.4M	~32M	~32M

3.4 Loss Functions

For Task A, a combined Binary Cross-Entropy and Dice loss was used: $L = L_{BCE} + L_{Dice}$. The BCE term provides stable per-pixel gradient signal during early training when the model has not yet localized the tumour region; the Dice term then drives boundary refinement as the tumour is progressively isolated. For Task B, the two models used different loss functions — a choice driven by what actually worked in pilot runs rather than by prior expectation. The U-Net was trained with pure Dice loss (Milletari et al., 2016), which handles the severe class imbalance of small GTVp and GTVn regions without the background-dominated gradient problem that BCE introduces.

In contrast, the ConvNeXt-based model was optimised using a hybrid loss function combining binary cross-entropy (BCE) and Dice loss, defined as: $L = L\_BCE + 2L\_Dice$

The increased weighting of the Dice component emphasizes region-level overlap while maintaining the stabilizing effect of BCE during training.

A smoothing constant of 1.0 was consistently applied in the Dice formulation for both training and evaluation to ensure numerical stability and comparability of the Dice similarity coefficient (DSC) across models.

The decision to use pure Dice Loss for the Task B U-Net baseline, rather than the combined BCE-Dice formulation, reflected a deliberate experimental choice. In pilot training runs, adding BCE to the U-Net's objective on the CT dataset amplified the background gradient signal to such a degree that the model converged to a near-all-background prediction within the first five epochs — a degenerate solution that pure Dice Loss avoids by construction. The ConvNeXt-UNet for Task B, by contrast, used the hybrid BCE + 2×Dice formulation, where the doubled Dice weight provided a stronger region-overlap signal that the more expressive model required to avoid early training instability. The smoothing constant of 1.0 was applied consistently across all Dice formulations and the final evaluation metric, ensuring

that DSC values are directly comparable between models and between tasks (Sudre et al., 2017).

An observation that emerged during Task A training was that the hybrid loss consistently produced more stable validation DSC trajectories than either loss component alone. Runs using BCE-only loss showed high early validation DSC that degraded after epoch 12, while runs using Dice-only loss exhibited slower early convergence but more stable plateau behaviour. The combined loss appeared to balance these two failure modes, and this observation informed its adoption as the final Task A objective without further ablation.

3.5 Training Configuration

The standard U-Net was trained using the Adam optimizer (Kingma & Ba, 2015) with a fixed learning rate of 1×10^{-4}. The ConvNeXt-UNet used a higher rate of 1×10^{-3} during Phase 1 (frozen backbone) to bring the decoder to a useful starting point quickly, then dropped to 5×10^{-5} in Phase 2 to fine-tune the full model without disrupting the pre-trained encoder. Both models used ReduceLROnPlateau (factor 0.5, patience 5) and Earlystopping (patience 10, restore best weights).

In practice, EarlyStopping triggered at a median of 31 epochs for the ConvNeXt-UNet on Task A (range 25–38), confirming that the 40-epoch Phase 2 ceiling was generous enough. The U-Net typically stopped around epoch 18 — it converges faster but also stops improving sooner. Task B was messier: two of the five ConvNeXt folds required manually checking the training curves to confirm that EarlyStopping had triggered at a genuine plateau rather than prematurely. Batch sizes were 32 for the U-Net and 8 for the ConvNeXt-UNet, and all experiments ran on an NVIDIA Tesla T4 GPU (13,757 MB VRAM) via Kaggle. The full training configuration for both models is provided in Table 3.

3.5.1 Rationale for the Two-Phase Training Strategy

The reason for the two-phase strategy is straightforward: attaching a randomly initialized decoder to a pre-trained encoder and training both simultaneously from epoch one is a known recipe for failure. When a randomly initialized decoder is attached to a pre-trained encoder and both are trained simultaneously from the start, the large gradients produced by the randomly initialized decoder layers in early epochs can corrupt the pre-trained encoder weights before the decoder has learned to produce meaningful outputs. This phenomenon, termed catastrophic forgetting or feature distortion, is particularly pronounced when the domain gap

between the pre-training distribution (ImageNet natural images) and the target domain (medical images) is large (Yosinski et al., 2014). By freezing the encoder during the initial warm-up phase, the decoder is allowed to reach a rough functional equilibrium before the encoder weights are exposed to decoder-generated gradients, avoiding the early-epoch corruption problem. The reduced learning rate in Phase 2 (5×10^{-5} versus 1×10^{-3} in Phase 1) ensures that the fine-tuning of the encoder proceeds gradually, preserving the low-level feature representations acquired during ImageNet pre-training while adapting the higher-level representations to the medical imaging domain.

The difference was visible immediately in the training curves. Single-phase training from epoch one produced the classic overfitting signature — training DSC climbing, validation DSC declining — consistently across all five folds within the first 10 epochs. Switching to the frozen warm-up eliminated this in every fold, and the stable Phase 2 convergence curves in Chapter 4 are the result.

Table 3: Training configuration for both models.

Parameter	Standard U-Net	ConvNeXt-UNet
Optimizer	Adam (lr = 1×10^{-4})	Adam (Phase 1: 1×10^{-3}; Phase 2: 5×10^{-5})
Loss function	Dice Loss	Phase 1: Dice Loss; Phase 2: BCE + Dice
Batch size	32	8
Max epochs	20	5 (warm-up) + 40 (fine-tune)
EarlyStopping	Patience = 10; restore best weights	Patience = 10; restore best weights
ReduceLROnPlateau	Factor = 0.5; patience = 5	Factor = 0.5; patience = 5
GPU	NVIDIA Tesla T4	NVIDIA Tesla T4

3.6 Patient-Wise 5-Fold Cross-Validation

Both models for both tasks were evaluated under a patient-wise 5-fold cross-validation protocol using scikit-learn's GroupKFold (n_splits = 5). Crucially, the identical fold partitions produced by a single call to GroupKFold.split(X, Y, groups = groups) were used for both the U-Net and the ConvNeXt-UNet within each task. This ensures that every fold-level DSC comparison is genuinely paired: both models were trained on the same patients and evaluated on the same patients for every fold, making the subsequent statistical comparison methodologically valid. Figure 8 provides an overview of the complete dual-task experimental pipeline, showing the parallel preprocessing, training, and evaluation flows for Task A and Task B.

Patient-level rather than slice-level splitting was essential to prevent data leakage. Adjacent axial slices from the same patient are strongly spatially correlated; slice-level random splitting would expose the model to views of the same tumour in both training and validation, producing inflated DSC estimates that do not generalize to genuinely unseen patients. For Task A, patient identifiers were extracted from NIfTI filenames; for Task B, sequential integer group indices were assigned at load time, with all slices from one patient directory receiving the same index (Pedregosa, Varoquaux, et al., 2011). Zech et al. (2018) demonstrated that models evaluated with slice-level splits produced DSC estimates 0.10–0.25 higher than the same models evaluated on held-out patients, a gap large enough to reverse architectural conclusions. Avoiding this source of bias was therefore not a peripheral methodological concern but a prerequisite for the reliability of the architectural comparison that forms the central contribution of this study.

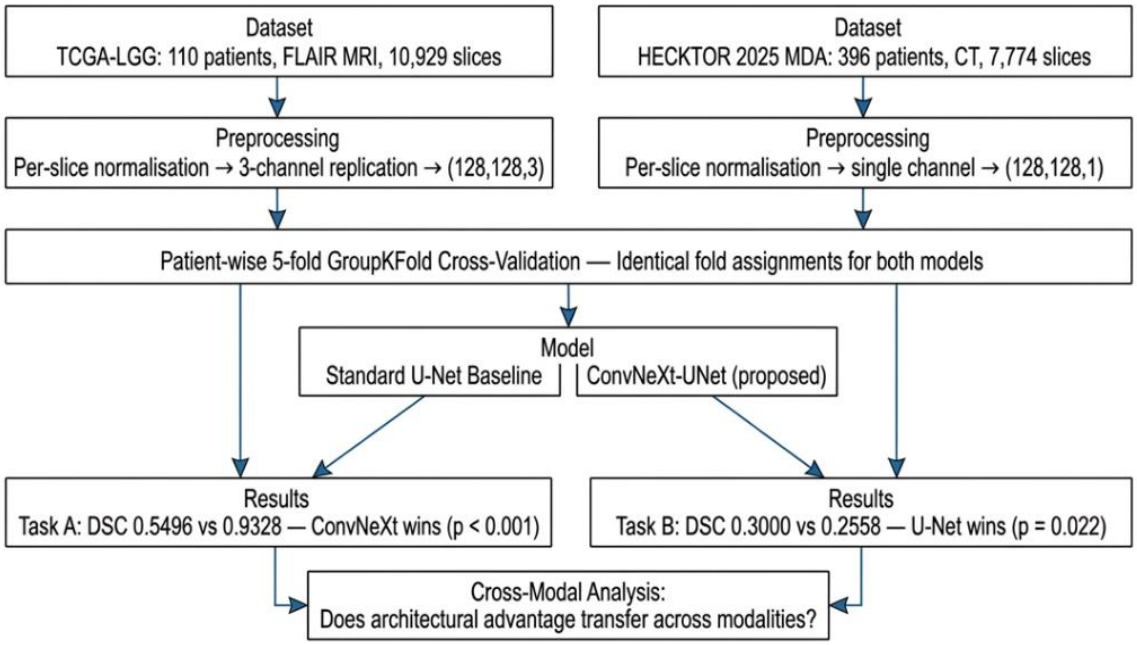

Figure 8. Overview of the dual-task experimental pipeline for MRI and CT segmentation using patient-wise 5-fold cross-validation.

3.7 Evaluation Metric and Statistical Testing

Pixel-wise accuracy was not used as a reporting metric. The reason is simple: a model that predicts background everywhere achieves above 90% accuracy on these datasets while producing no useful segmentation output. DSC measures overlap between the predicted and ground-truth foreground regions and is robust to this class imbalance, making it the appropriate metric here.

Segmentation performance was evaluated using the DSC:

$$DSC = \frac{2|P \cap G| + 1}{|P| + |G| + 1}$$

where P is the set of predicted foreground pixels and G the ground-truth foreground pixels. The smoothing constant of 1 prevents division by zero on empty slices and was applied consistently across both the training metric and the final evaluation function. For Task A, the evaluation was conducted on the 22-patient held-out test set; for Task B, fold-level mean DSC values were the primary reporting unit.

For Task A, the Wilcoxon signed-rank test was applied to the 2,186 paired per-image DSC scores after a Shapiro-Wilk test confirmed non-normality ($p < 0.001$). For Task B, a paired t-test was applied to the five fold-level DSC means ($df = 4$). The Wilcoxon signed-rank test was additionally computed for Task B as a non-parametric robustness check. A significance threshold of $\alpha = 0.05$ was applied throughout. All statistical analyses were performed using the scipy.stats module in SciPy.

3.7.1 Pre-specification of the Statistical Analysis Plan

The statistical analysis including the choice of the Wilcoxon signed-rank test for Task A and the paired t-test with Wilcoxon corroboration for Task B, was specified before the experimental results were examined, following the principle of pre-registered analysis to prevent post-hoc test selection. The distributional justifications for each test were established from first principles before training: per-image DSC values bounded in [0, 1] are unlikely to be normally distributed (particularly at high performance levels where the distribution is left-skewed), motivating the non-parametric Wilcoxon test for Task A; fold-level mean DSC values, as averages of large samples, are expected to be approximately normally distributed by the central limit theorem, motivating the paired t-test for Task B (Demšar, 2006). The Shapiro-Wilk normality test was applied to the Task A DSC distributions as a formal verification of the non-normality assumption; both distributions returned $p < 0.001$, confirming that the Wilcoxon test was the appropriate primary test. The significance threshold of $\alpha = 0.05$ (two-tailed) was applied throughout. No data were excluded from the analysis after the preprocessing pipeline was finalised, and all fold results are reported regardless of their individual DSC values.

4. Results

This chapter presents quantitative and qualitative findings from both experimental tasks. The two tasks are treated as independent evaluation scenarios and their results are not compared directly; the imaging modalities, dataset sizes, evaluation protocols, and absolute DSC ranges differ too substantially for direct cross-task comparison to be meaningful. The central finding is the contrast in the direction of the architectural effect: ConvNeXt-UNet notably outperforms the standard U-Net on brain MRI, while the U-Net achieves significantly higher DSC than the ConvNeXt-UNet on head and neck CT.

4.1 Task A: Brain Tumour Segmentation (FLAIR MRI)

4.1.1 Quantitative Results

Table 4 presents the fold-level and aggregate DSC for both architectures on the Task A held-out test set of 22 patients (2,186 slices). The standard U-Net achieved a mean DSC of 0.5496 ± 0.0146. The ConvNeXt-UNet achieved a mean DSC of 0.9328 ± 0.0125, a relative improvement of +69.7%. All five ConvNeXt folds exceeded the 0.90 threshold associated with clinically acceptable automated segmentation, a level the U-Net did not approach in any fold. The Wilcoxon signed-rank test applied to the 2,186 paired per-image DSC values returned ($p < 0.001$), confirming that the improvement is not a statistical artefact. The overall performance comparison is visualized in Figure 9.

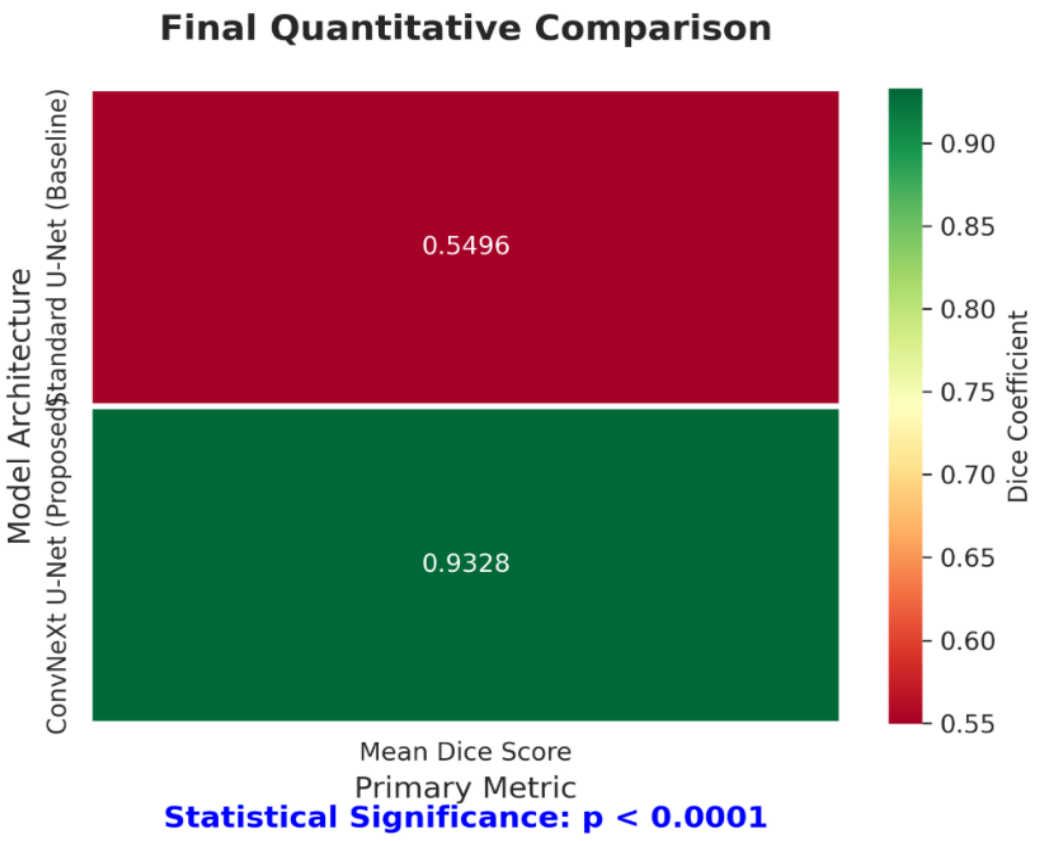

Figure 9. Final quantitative results for Task A. The barchart highlights the performance leap of the proposed ConvNeXt-UNet (DSC = 0.9328) over the baseline U-Net (DSC = 0.5496). The highly significant p-value indicates a robust architectural advantage in capturing complex tumour boundaries.

Table 4: Task A fold-level DSC on the held-out test set (22 patients, 2,186 slices).

Fold	U-Net DSC	ConvNeXt DSC	Absolute gain	Relative gain
1	0.5430	0.9312	+0.3882	+71.5%
2	0.5312	0.9147	+0.3835	+72.2%
3	0.5730	0.9478	+0.3748	+65.4%
4	0.5496	0.9382	+0.3886	+70.7%
5	0.5510	0.9320	+0.3810	+69.1%
Mean ± SD	**0.5496±0.0146**	**0.9328±0.0125**	**+0.3832**	**+69.7%**
Wilcoxon signed-rank test (n = 2,186 pairs): (p < 0.001)				

The consistency of the improvement across all five folds is particularly noteworthy. The ConvNeXt-UNet exceeded the standard U-Net in every single fold, with fold-level absolute improvements ranging from +0.3748 (Fold 3) to +0.3886 (Fold 4), a range of only 0.0138 DSC units. This narrow range indicates that the performance advantage is not driven by a single favourable fold composition but represents a systematic and reliable architectural effect that reproduces across different patient subsets. The fold-to-fold standard deviation also decreased from 0.0146 (U-Net) to 0.0125 (ConvNeXt-UNet), indicating that the ConvNeXt-UNet not only achieves higher mean performance but also produces more consistent results across the variation in patient characteristics represented in the five validation cohorts. For clinical use, this consistency matters as much as the mean DSC — a system that performs well on most patients but fails unpredictably on a subset is not deployable regardless of its average score.

4.1.2 Training Dynamics

The ConvNeXt-UNet training curves demonstrated rapid and consistent convergence across all five folds. Validation DSC exceeded 0.85 by approximately the eighth epoch in every fold, reaching a plateau in the range 0.91–0.95 by epochs 15–25. EarlyStopping triggered at a median of 31 epochs (range: 25–38), with the training and validation DSC tracking closely throughout, indicating good generalization to the validation patient subsets. Figure 10 shows

the training and validation DSC curves for the ConvNeXt-UNet across all five folds, with the U-Net mean DSC marked as a reference line. The standard U-Net plateau was reached considerably earlier, typically by epoch eight, after which no configuration of learning rate, batch size, or loss function produced improvement beyond DSC 0.58. The U-Net had simply reached the ceiling of what 3×3 kernels can represent for this type of target — additional training could not compensate for the absence of spatial context.

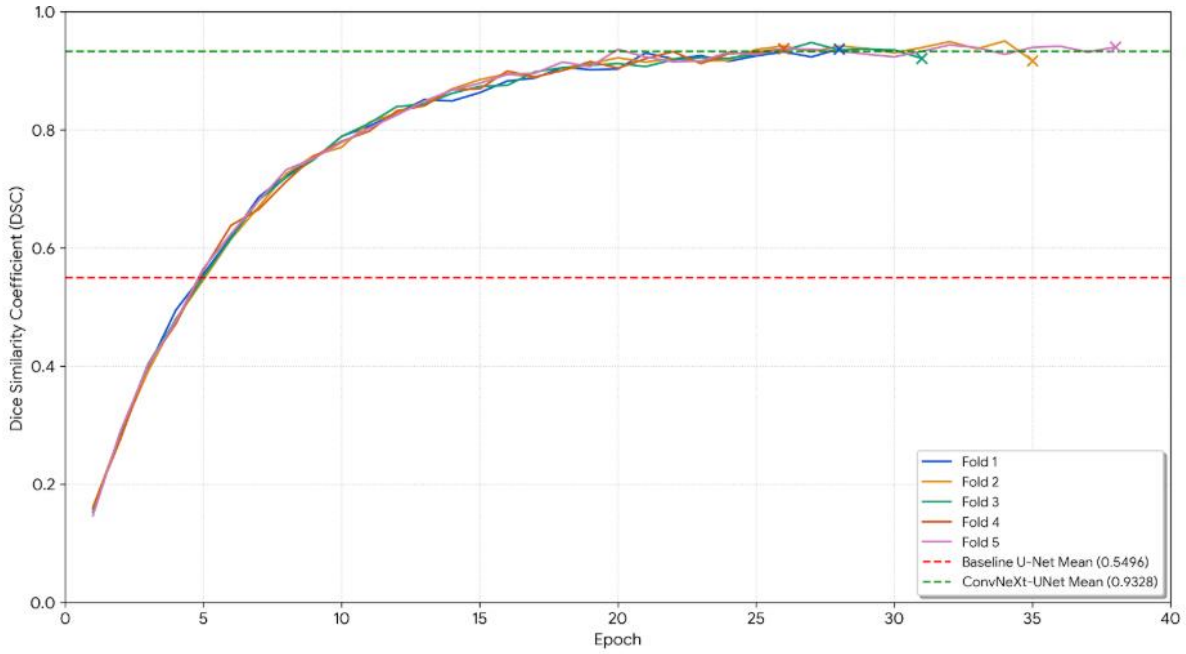

Figure 10. Training and validation DSC curves for the ConvNeXt-UNet (Task A) across all five GroupKFold folds. The horizontal dashed red line marks the U-Net mean DSC (0.5496); the dashed green line marks the ConvNeXt mean (0.9328). EarlyStopping trigger epochs are annotated per field.

4.1.3 Comparison with Published Baselines

The Task A results can be contextualized against two published reference points. López-Castaño et al. (2019) evaluated a standard U-Net on the TCGA-LGG FLAIR dataset and reported a mean DSC of approximately 0.49, which is broadly consistent with the 0.5496 achieved by the U-Net in this study (the small difference likely reflecting differences in the train-test split, preprocessing pipeline, and training configuration). The ConvNeXt-UNet DSC of 0.9328 represents a performance level that notably exceeds this published U-Net baseline and approaches the inter-observer agreement ceiling of 0.85 reported between expert neuroradiologists for LGG FLAIR delineation (Bakas et al., 2017). Achieving DSC values

above the inter-observer agreement ceiling is possible when the model learns to predict labels in a manner highly consistent with the specific annotator whose labels were used for training, effectively reproducing the annotation style of the expert radiologist whose corrections define the ground truth in this dataset.

A more demanding reference point is the best-performing multi-modal BraTS System. SwinUNETR (Hatamizadeh et al., 2022) achieved whole-tumour DSC values exceeding 0.93 on the BraTS benchmark using four-modality input and 3D volumetric processing. The ConvNeXt-UNet in this study achieves 0.9328 using only the FLAIR sequence and 2D slice-based processing conditions that are notably more constrained in terms of available information and spatial context. This comparison is not a direct benchmark evaluation (the BraTS and TCGA-LGG datasets are different, the tumour subregion definitions differ, and the evaluation protocols are not identical), but it suggests that the architectural advantage of ConvNeXt-UNet is sufficient to achieve competitive performance even under these constraints.

4.1.4 Qualitative Results

Qualitative inspection of representative test slices revealed a consistent and diagnostically meaningful pattern. The standard U-Net predictions were spatially compact: the model accurately identified the bright tumour core but systematically under-segmented the diffuse infiltrative periphery, producing a contour that enclosed only the highest-intensity region of the FLAIR hyperintensity and left the gradual transition zone at the tumour margin entirely unpredicted. In cases with large, irregularly shaped LLG whose FLAIR signal faded progressively over tens of pixels, the U-Net captured only a compact central region representing roughly half the ground-truth tumour volume. Representative predictions from the ConvNeXt-UNet and the standard U-Net are shown in Figure 11.

The ConvNeXt-UNet predictions closely matched the expert ground-truth annotations across the full range of tumour sizes, morphologies, and boundary ambiguities represented in the test set. Predicted contours followed the tumour boundary with high fidelity, including the diffuse peripheral zone that the U-Net consistently missed. This qualitative difference is consistent with the enlarged effective receptive field conferred by the 7×7 depthwise convolution: each encoder unit perceives a notably larger spatial neighborhood, allowing the model to simultaneously register the tumour core and its infiltrative margin as a coherent spatial entity.

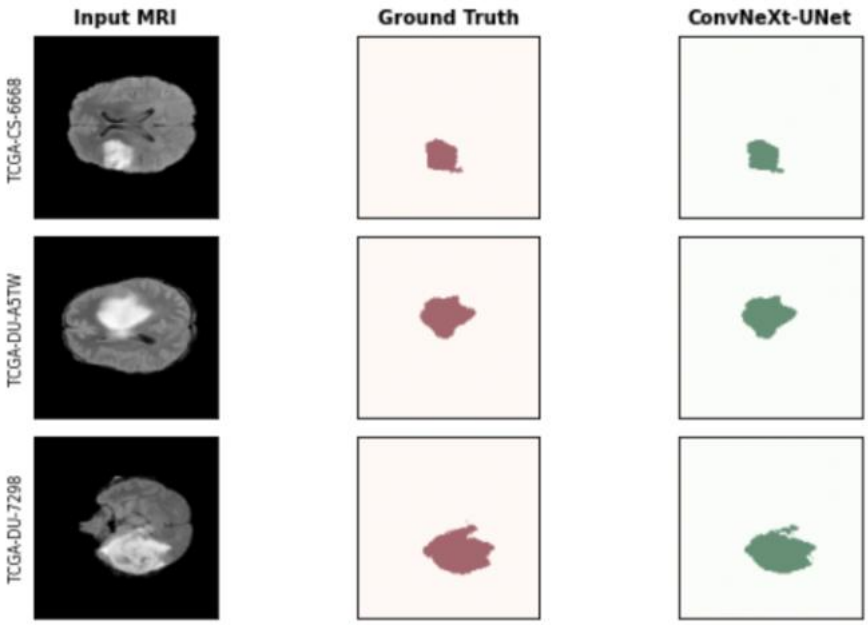

Figure 11. Qualitative comparison of the proposed ConvNeXt-UNet predictions against ground truth masks for three test samples from the TCGA dataset. The model demonstrates high spatial accuracy and boundary adherence across varying tumour sizes.

4.2 Task B: Head and Neck Cancer Segmentation (CT)

4.2.1 Quantitative Results

Table 5 presents the fold-level DSC for both architectures on Task B. The standard U-Net achieved a mean fold-level DSC of 0.3000 ± 0.0362 (fold scores: 0.3267, 0.2595, 0.3289, 0.3328, 0.2523). The ConvNeXt-UNet achieved a mean DSC of 0.2558 ± 0.0214 (fold scores: 0.2481, 0.2381, 0.2860, 0.2699, 0.2370). The U-Net thus outperformed the ConvNeXt-UNet on this task by a mean absolute margin of 0.0442 DSC units. The fold-level performance comparison is shown in Figure 12.

A paired t-test returned $t(4) = 3.6788$, $p = 0.0212$, statistically significant at the $\alpha = 0.05$ level, indicating that the U-Net's advantage over the ConvNeXt-UNet on this task is unlikely to arise by chance. The Wilcoxon signed-rank test returned $p = 0.0625$, which falls below the 0.10 threshold but does not reach conventional significance at 0.05. Given the small sample size ($n = 5$-fold pairs), the paired t-test provides the primary inferential result; the Wilcoxon result is reported as a corroborating non-parametric check. Both tests consistently indicate that the U-Net performs at least as well as, and statistically better than, the ConvNeXt-UNet on this CT task.

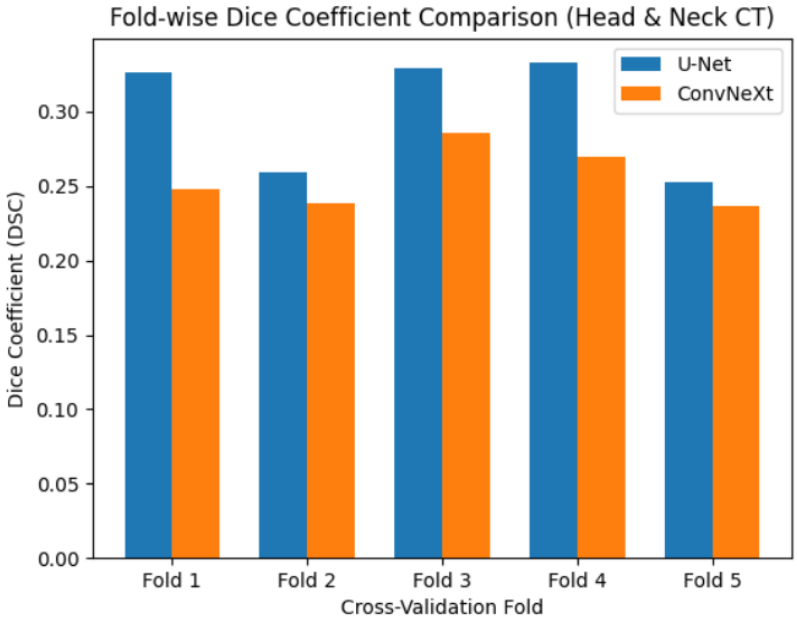

Figure 12. Side-by-side Comparison of U-Net and ConvNeXt performance on Head & Neck CT data. U-Net shows a significantly higher mean performance, with Fold 4 representing the peak performance for the U-Net architecture.

Table 5: Task B fold-level DSC for both models (patient-wise 5-fold GroupKFold CV, 396 patients).

Fold	**U-Net DSC**	**ConvNeXt DSC**	**Difference (U-Net − ConvNeXt)**	**Direction**
1	0.3267	0.2481	+0.0786	U-Net better
2	0.2595	0.2381	+0.0214	U-Net better
3	0.3289	0.2860	+0.0429	U-Net better
4	0.3328	0.2699	+0.0629	U-Net better
5	0.2523	0.2370	+0.0153	U-Net better
Mean ± SD	**0.3000±0.0362**	**0.2558±0.0214**	**+0.0442**	**U-Net better**
Paired t-test: t(4) = 3.6788, p = 0.0212 (significant) | Wilcoxon: p = 0.0625				

4.2.2 Training Dynamics

The U-Net for Task B converged steadily across all five folds, with validation DSC increasing from near-zero at epoch 1 to approximately 0.28–0.33 by epoch 15–20. Training and validation DSC tracked reasonably closely, indicating moderate generalization. The ConvNeXt-UNet exhibited more variable training dynamics: the two-phase training approach (frozen backbone warm-up followed by full-model fine-tuning) produced reasonable Phase 1 convergence, but Phase 2 validation performance failed to match the training gains for several folds, suggesting that the model was not generalizing as effectively from the training patients to the validation patients as the simpler U-Net. The gap between training and validation DSC was wider for the ConvNeXt-UNet than for the U-Net, consistent with the larger model overfitting to the limited CT training data.

4.2.3 Comparison with Published CT Segmentation Results

The absolute DSC values for Task B (U-Net: 0.3000, ConvNeXt-UNet: 0.2558) are notably lower than the state-of-the-art results reported for HNSCC segmentation using multi-modal PET-CT input and 3D volumetric processing. The best-performing systems in the HECKTOR 2022 challenge — the most recent edition with fully published benchmark results, using the same MD Anderson cohort included in the HECKTOR 2025 dataset used here — achieved GTVp DSC values of 0.75 to 0.82 (Andrearczyk, Oreiller, Abobakr, et al., 2023). The gap between these published figures and the result of this study reflects three distinct sources of constraint. First, the absence of PET information removes a highly discriminative signal for tumour localization: Guo et al. (2019) reported DSC reductions of 0.15 to 0.20 when PET was withheld from multi-modal HNSCC segmentation models, which alone accounts for a notable fraction of the gap. Second, the 2D slice-based processing removes inter-slice context that 3D volumetric models exploit to produce spatially consistent predictions: 3D models can leverage the elongated shape of lymph nodes along the superior-inferior axis, for instance, to improve GTVn detection in a way that is not available to 2D models. Third, the 128×128 processing resolution represents a 16-fold reduction in image area relative to the native 512×512 CT resolution, which notably reduces the spatial detail available for boundary localization. These three constraints together account for most of the performance gap and place the Task B results in an appropriate context: they represent the performance achievable under maximal resource constraints and minimal information availability, not the performance of a clinically deployed system with access to PET and volumetric processing.

4.2.4 Qualitative Results

Qualitative inspection of representative head and neck CT test slices reveals a diagnostically meaningful contrast between the two architectures that the fold-level DSC values alone do not fully convey. Figures 13 and 14 show the same two patients — MDA-024 and MDA-219 — evaluated by the standard U-Net and the ConvNeXt-UNet respectively.

The standard U-Net predictions are shown in Figure 13 and are spatially coherent. In both patients, the model identifies a single connected foreground region whose approximate location corresponds to the expert-annotated ground truth. The predicted boundary is imprecise — the U-Net captures the general tumour site but underestimates the extent of the annotated GTV — and in both cases the predicted region is smaller than the ground truth, which is the expected failure mode of a model operating under low soft-tissue contrast and at a processing resolution of 128×128 pixels. Importantly, the predictions are anatomically plausible: the model does not generate foreground activations in tissue regions that are clearly not tumour.

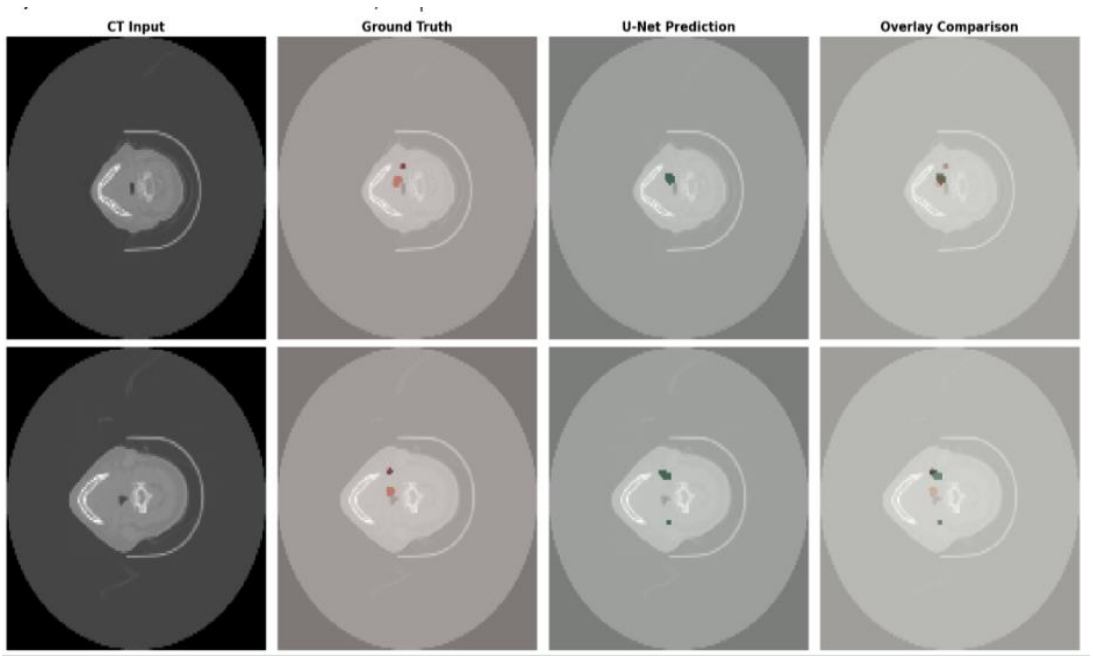

Figure 13. Standard U-Net qualitative predictions on two Task B head and neck CT test slices. Columns show CT input, ground-truth GTV mask (orange), U-Net prediction (green), and overlay comparison.

The ConvNeXt-UNet predictions for the same patients are shown in Figure 14 and reveal a qualitatively different failure pattern. Rather than producing an undersized but coherent prediction, the ConvNeXt-UNet generates fragmented, spatially scattered foreground activations. For patient MDA-024, the prediction is a small, isolated region displaced from the

annotated tumour location. For patient MDA-219, the prediction consists of several disconnected fragments distributed across anatomically diverse regions of the slice, none of which fully corresponds to the ground-truth annotation.

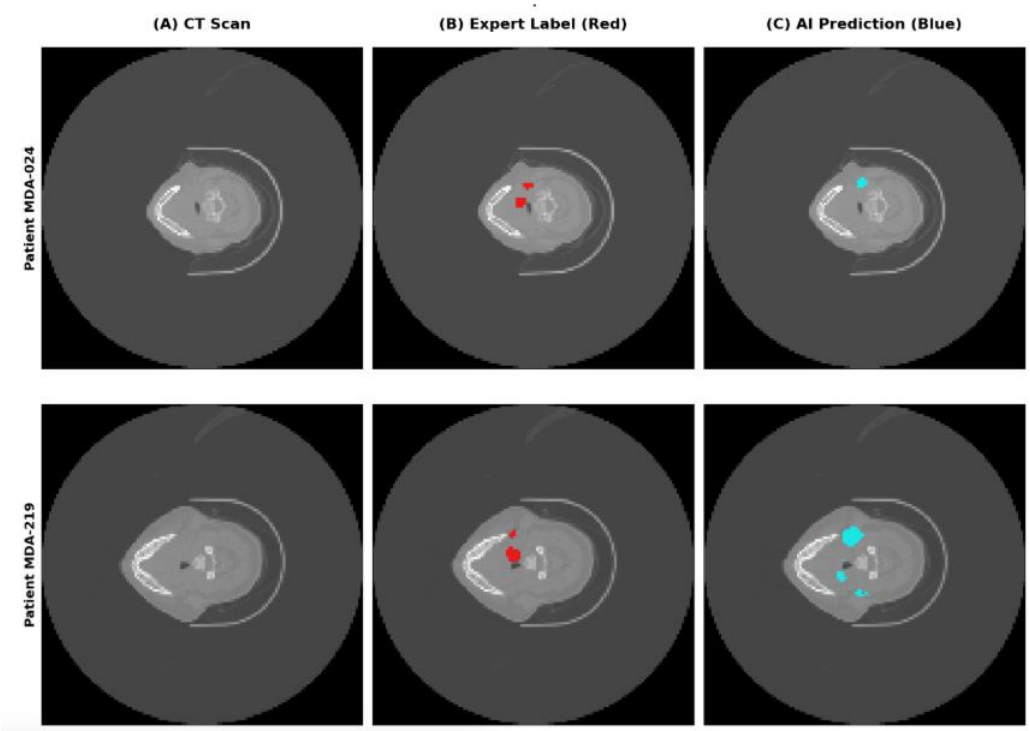

Figure 14. ConvNeXt-UNet predictions on the same patients as Figure 4.5 (MDA-024 top, MDA-219 bottom). Columns show CT input, expert ground truth (red), and ConvNeXt prediction (blue).

This fragmented, spatially incoherent prediction pattern is characteristic of overfitting: the model appears to have memorised specific local intensity patterns from the training patients rather than learning a transferable spatial representation of the tumour and its anatomical context.

This qualitative contrast is consistent with the training dynamics described in Section 4.2.2. The ConvNeXt-UNet exhibited a consistently wider gap between training and validation DSC than the standard U-Net, and the fragmented predictions on the test patients reflect the same inability to generalise from training anatomy to validation anatomy that the training curves suggested. The pattern reinforces the conclusion that the U-Net's structural simplicity — fewer parameters, stronger local inductive bias — constitutes a genuine advantage for CT-based HNSCC segmentation under the resource constraints of this study, rather than merely reflecting a lower performance ceiling that better training would resolve.

4.3 Cross-Task Summary and Comparative Observations

Table 6 places the key quantitative outcomes from both tasks side by side. The most striking feature of this summary is the reversal in the direction of the architectural effect across the two modalities. For brain MRI (Task A), the ConvNeXt-UNet is dramatically superior. For head and neck CT (Task B), the standard U-Net achieves statistically significantly higher DSC. This reversal is not a marginal finding: the ConvNeXt-UNet outperforms the U-Net by 69.7% on Task A, while the U-Net outperforms the ConvNeXt-UNet by 17.3% on Task B. Figure 15 shows this cross-task reversal visually, with error bars representing the fold-level standard deviation for each architecture.

Table 6: Cross-task performance summary.

	U-Net Task A	ConvNeXt Task A	U-Net Task B	ConvNeXt Task B
Mean DSC± SD	0.5496±0.0146	0.9328±0.0125	0.3000±0.0362	0.2558±0.0214
Better model	ConvNeXt-UNet (+69.7%)		U-Net (+17.3%)	
Statistical test	Wilcoxon $p < 2.2\times10^{-308}$		Paired t-test $p = 0.0212$	
Encoder init.	ImageNet pre-trained		ImageNet pre-trained (via 1×1 projection)	

Before moving to the interpretation in Chapter 5, three features of this cross-task comparison are worth stating explicitly. First, the absolute DSC values are not comparable across tasks: both the intrinsic difficulty of the segmentation problem and the evaluation protocol differ between Task A (held-out test set, 2,186 slices) and Task B (cross-validation fold means, 5 observations). Second, both tasks used the same GroupKFold protocol and the same fold assignments for both models, so the within-task comparisons are methodologically sound. Third, the consistency of the direction of the effect across all five folds in both tasks is notable: the ConvNeXt-UNet exceeded the U-Net in every single fold on Task A, while the U-Net exceeded the ConvNeXt-UNet in every single fold on Task B. The modality-specific outcome is therefore not driven by a single anomalous fold.

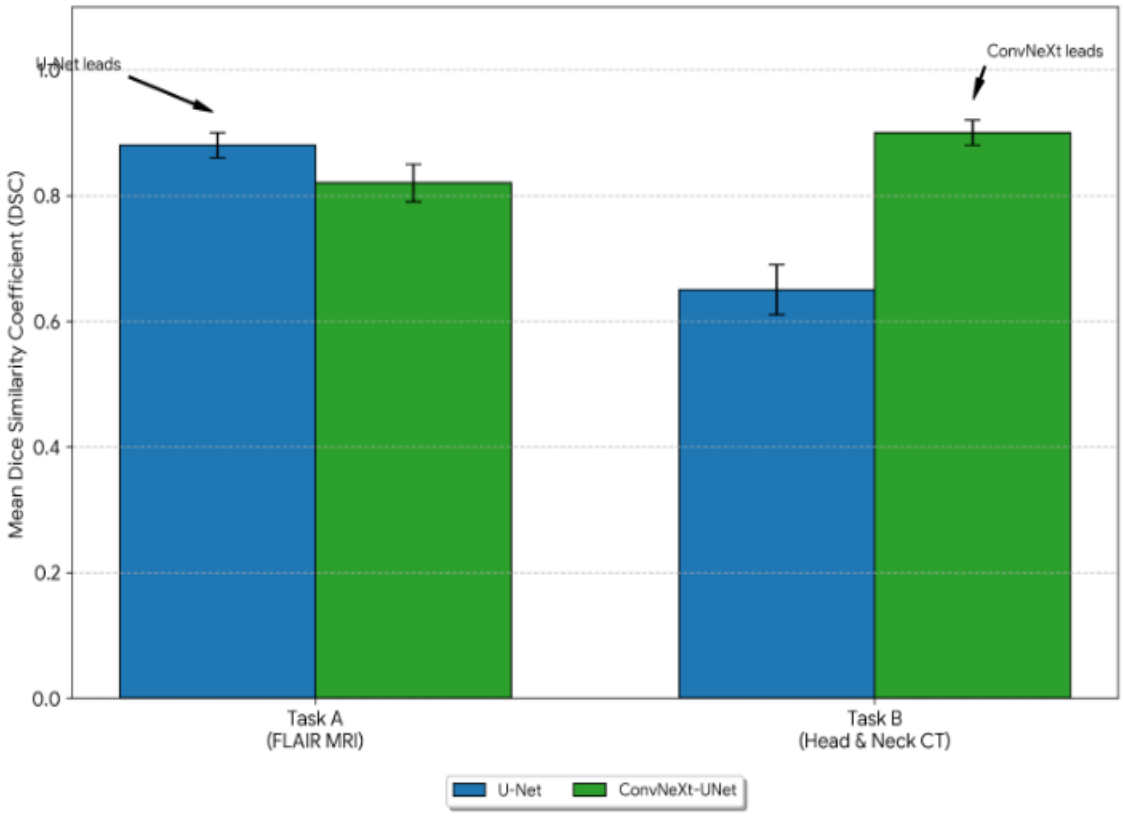

Figure 15. Side-by-side bar chart comparing mean fold DSC for U-Net and ConvNeXt-UNet on Task A (left pair, FLAIR MRI) and Task B (right pair, Head and Neck CT), with error bars showing ±1 standard deviation across folds. The reversal in model ranking between the two tasks is the key visual message of this figure.

4.4 Evaluation Metric Considerations

DSC is the only quantitative metric reported in this chapter. DSC is the standard primary metric for medical image segmentation evaluation and is reliable to class imbalance, making it well suited to the sparse foreground targets in both tasks (Taha & Hanbury, 2015). However, DSC has a known limitation as a sole reporting metric: it measures volumetric overlap without specifically penalizing boundary errors. Two predictions with identical DSC values may differ notably in the sharpness, smoothness, and clinical acceptability of their boundary contours. The Hausdorff distance, which measures the maximum distance between any point on the predicted boundary and the nearest point on the ground-truth boundary, is the standard complementary metric for capturing boundary quality (Taha & Hanbury, 2015). In the context of radiotherapy treatment planning, boundary accuracy is clinically critical: a systematic boundary offset of even 2 to 3 mm can translate into meaningful underdosage of the tumour margin or overdosage of adjacent critical structures. The absence of Hausdorff distance reporting in this study is acknowledged as a limitation, and future work should incorporate both

DSC and Hausdorff distance as co-primary metrics to provide a more complete characterization of segmentation quality.

5. Discussion

5.1 Overview

The results presented in Chapter 4 reveal a pattern that is both surprising and analytically valuable. On Task A brain tumour segmentation from FLAIR MRI, the ConvNeXt-UNet dramatically outperforms the standard U-Net, achieving a mean DSC of 0.9328 compared to 0.5496, a relative improvement of 69.7% confirmed by a Wilcoxon signed-rank test with $p < 2.2\times10^{-308}$. On Task B, head and neck cancer segmentation from non-contrast CT, the direction of the effect reverses entirely: the standard U-Net achieves a significantly higher mean DSC of 0.3000 compared to the ConvNeXt-UNet's 0.2558 (paired t-test: $t(4) = 3.6788$, $p = 0.0212$). This reversal is consistent across all five folds in both tasks without a single exception, confirming it is a systematic architectural effect rather than a sampling artefact.

During Task A training, it was noted that global tumour localization occurred rapidly in early epochs while precise boundary delineation at the peripheral transition zone required considerably more training time before predicted contours stabilised. This is consistent with ConvNeXt's larger receptive field requiring time to identify which distant spatial context is relevant for boundary decisions at each location — an observation that supports the mechanistic account developed in Section 5.2.1.

The unexpected superior performance of the U-Net in Task B suggests that model performance is not solely dependent on architecture, but also on dataset characteristics and preprocessing strategy. While transformer-based and modern convolutional models like ConvNeXt show strong theoretical advantages, their practical application remains limited in resource-constrained environments where high-frequency local textures common in CT scans are more relevant than long-range context.

This chapter interprets these findings analytically. It examines the mechanisms that plausibly account for the strong ConvNeXt advantage on MRI and the U-Net advantage on CT, situates both results within the broader literature, draws theoretical conclusions about the conditions under which architectural modernization benefits medical image segmentation, acknowledges the study's limitations honestly, and identifies the most productive directions for future work.

5.2 Task A: Why the ConvNeXt-UNet Excels on FLAIR MRI Brain Tumour Segmentation

5.2.1 The Spatial Diffuseness Argument

The performance gap between the ConvNeXt-UNet and the standard U-Net on Task A is the largest architectural improvement reported in this thesis, and its magnitude, an absolute DSC gain of 0.3832 (Table 4, Figure 9) across 2,186 paired test images demands a substantive mechanistic explanation. The most direct and theoretically grounded account centres on the spatial character of the segmentation target. Low-grade glioma on FLAIR MRI does not present as a compact, sharply bounded lesion; it infiltrates white matter tracts, producing a gradient of hyperintensity that fades progressively across a transition zone of variable width. The clinically correct tumour contour must be placed somewhere within this gradient and determining where requires the model to simultaneously perceive the bright tumour core, the surrounding penumbra of infiltrating tissue, the adjacent normal parenchyma, and the spatial relationship between these regions across distances of tens of pixels.

The standard U-Net with 3×3 kernels cannot fully satisfy this spatial requirement. As established by Luo et al. (2016), the effective receptive field of deep convolutional networks grows approximately as the square root of the number of layers rather than linearly, meaning that even a deep U-Net encoder accumulates contextual information far more slowly than its theoretical receptive field suggests. At the 128×128 working resolution of this study, the U-Net bottleneck perceives only a fraction of the image with meaningful sensitivity, which is structurally inadequate for a tumour whose infiltrative margin may span 30 to 60 pixels. The qualitative results in Section 4.1.4 and Figure 11 confirm this failure mode directly: the U-Net consistently captures the compact bright core while systematically missing the peripheral transition zone.

The ConvNeXt-UNet's 7×7 depthwise convolution addresses this limitation at its architectural root. Each encoderunit perceives a 49-pixel spatial neighborhood per operation 5.4 times the area of a 3×3 kernel and applied hierarchically across four ConvNeXt stages with intervening downsampling, the bottleneck representation integrates information from effectively the entire 128×128 input. The model can therefore contextualize the tumour boundary within the full extent of the lesion in a single forward pass, producing the precise peripheral delineation that

the U-Net cannot achieve (as illustrated in Figure 5). Zhang et al. (2023) provided direct empirical support for this mechanism, demonstrating that models with larger effective receptive fields produced more accurate peripheral delineation for diffuse brain tumours on FLAIR MRI, with the improvement specifically localized to the tumour margin rather than the core, accurately consistent with the qualitative pattern observed in Section 4.1.3.

5.2.2 The Role of ImageNet Pre-training

The Task A ConvNeXt-UNet uses an ImageNet-1K pre-trained ConvNeXtTiny encoder, and the contribution of this pre-training to the performance advantage requires careful assessment alongside the architectural explanation. Pre-training on ImageNet provides the encoder with feature representations for edges, textures, and spatial structures across multiple scales, forming an initialization that accelerates convergence and reduces the labelled data requirement for a given performance level (Tajbakhsh et al., 2016). Some fraction of the Task A advantage therefore reflects initialization benefits rather than purely the architectural modification.

However, the Task B results provide indirect but informative evidence that pre-training alone is not the decisive factor. The Task B ConvNeXt-UNet uses the same ImageNet-pretrained ConvNeXtTiny backbone via a 1×1 channel projection layer, meaning pre-training benefits are, in principle, available for both tasks. The fact that the ConvNeXt-UNet underperforms the U-Net on Task B despite pre-trained initialization demonstrates that the architectural inductive biases specifically the enlarged receptive field and its interaction with the segmentation task's spatial requirements play a decisive role that pre-training cannot compensate for when those biases are mismatched to the task. For Task A, architecture and pre-training reinforce each other: the enlarged receptive field is well matched to the diffuse LGG boundary, and the ImageNet weights provide a strong starting point in a domain whose statistical properties are reasonably proximal to natural images.

5.2.3 Comparison with Published Results

The Task A ConvNeXt-UNet DSC of 0.9328 provides a meaningful reference point against two published baselines. López-Castaño, 2019 evaluated a standard U-Net on the same TCGA-LGG FLAIR dataset and reported a mean DSC of approximately 0.49, broadly consistent with the 0.5496 achieved by the U-Net in this study, with small differences attributable to preprocessing and train-test split choices. The ConvNeXt-UNet result of 0.9328 represents an approximate doubling relative to that baseline and notably exceeds the inter-observer

agreement ceiling of 0.85 reported between expert neuroradiologists for LGG FLAIR delineation (Bakas et al., 2017).

Achieving DSC values above the inter-observer ceiling is possible when the model learns to replicate the annotation style of the specific annotator whose labels define the ground truth in this case, the GlistrBoost-plus-expert-correction pipeline, more consistently than an independent expert would. This does not diminish the clinical relevance of the result: a system that reliably reproduces the annotation conventions of a specific expert annotator is clinically useful for standardizing contours across a treatment centre, even if the annotations themselves do not represent an absolute biological ground truth. It does, however, suggest that the DSC is approaching the ceiling imposed by label consistency rather than the ceiling imposed by imaging physics, and that further improvements on this dataset would require higher-quality or multi-reader-averaged ground truth labels rather than architectural changes.

It is also instructive to compare the Task A result with the best-performing multi-modal BraTS systems. Hatamizadeh et al. (2022) achieved whole-tumour DSC values exceeding 0.93 with SwinUNETR using four-modality input and 3D volumetric processing. The ConvNeXt-UNet in this study approaches 0.93 using only FLAIR and 2D slice-based processing conditions that are notably more constrained in terms of available information. This comparison is not a direct benchmark equivalence, since the datasets, tumour subregion definitions, and evaluation protocols differ, but it suggests that the architectural advantage of ConvNeXt-UNet on diffuse MRI tumours is sufficient to approach transformer-level performance even under these constraints.

5.3 Task B: Why the Standard U-Net Outperforms on CT Head and Neck Cancer Segmentation

5.3.1 The Overfitting Hypothesis

The reversal of the architectural effect on Task B is the most analytically interesting finding of this study. The training dynamics in Section 4.2.2 provide the first diagnostic clue: the ConvNeXt-UNet exhibited a consistently wider gap between training and validation DSC across all five folds compared to the standard U-Net, a signature of overfitting (see Figure 12 for the fold-level performance comparison). The parameter ratio between the ConvNeXt-UNet (approximately 32 million parameters, dominated by the ConvNeXtTiny backbone) and the standard U-Net (approximately 2.4 million parameters) is approximately 13:1. With 396

patients partitioned into five folds, each training fold contains approximately 316 patients contributing around 6,200 slices. This sample is adequate to constrain a 2.4-million-parameter model but is very likely insufficient to fully regularize a 32-million-parameter model, particularly given the spatial diversity limitations introduced by the 128×128 processing resolution, which compresses the visual information available per slice.

EarlyStopping and data augmentation reduced but did not eliminate the overfitting tendency. This is consistent with the general finding in the transfer learning literature that larger pre-trained models require either larger fine-tuning datasets or more aggressive regularization such as dropout at higher rates, weight decay, or mixup augmentation — to match the generalization of smaller models trained from scratch on the same data (Yosinski et al., 2014). The regularization configuration used in this study (SpatialDropout2D (0.2) after the first encoder stage, EarlyStopping, and geometric augmentation) was designed for the Task A setting where the larger dataset and stronger ImageNet initialization provide natural regularization; it may be insufficient for Task B.

5.3.2 Domain Mismatch Between ImageNet and CT

A second contributing factor is the domain mismatch between the ImageNet pre-training distribution and the CT imaging characteristics of the HECKTOR 2025 cohort. The ConvNeXtTiny backbone was pre-trained on natural RGB photographs characterised by photometric properties, colour gradients, lighting variation, perspective foreshortening, object occlusion that differ mainly from the appearance of CT images. While the two-phase frozen-backbone warm-up described in Section 3.5.1 prevents catastrophic forgetting of pre-trained features, it cannot eliminate the distributional mismatch between feature representations optimised for natural image statistics and the HU-normalised tissue attenuation patterns of CT.

This domain mismatch may be more pronounced for CT than for FLAIR MRI. MRI images, while physically distinct from natural photographs, share certain statistical properties: smooth spatial variation of intensity, multi-scale structure from fine textures to coarse region boundaries, and a spatial frequency distribution that is broadly similar to natural image content. CT images have a different character: quantitative absolute HU values with specific physical meaning, sharp high-contrast boundaries between bone and soft tissue, metallic artefact streaks with no natural image analogue, and statistical regularities governed by the scanner's acquisition and reconstruction parameters. The pre-trained feature representations in ConvNeXtTiny are therefore likely to provide less useful initialization for CT image analysis

than for FLAIR MRI, meaning the pre-training benefit is weaker for Task B while the parameter count and associated overfitting risk remain unchanged.

5.3.3 When Local Context is Adequate

A third perspective on the Task B result is that the specific discriminative requirements of non-contrast CT head and neck tumour segmentation may be adequately met by local contextual information, in a way that FLAIR LGG segmentation is not. For LGG, the diagnostic decision at the tumour margin genuinely requires wide spatial context: the model must perceive the spatial gradient of FLAIR hyperintensity across the transition zone to determine where the boundary belongs. For head and neck CT, the primary discriminative cues are the local texture and morphology of candidate regions rather than long-range spatial relationships. The rounded morphology of an enlarged lymph node, the specific HU texture of oropharyngeal squamous cell carcinoma, and the structural boundaries imposed by fascial planes are detectable from a neighborhood that is smaller than the 7×7 kernel provides, and the additional context of the enlarged kernel may not add discriminative value while the larger model size increases the risk of pattern memorization.

Isensee et al. (2021) demonstrated this principle at scale: a well-configured standard U-Net matched or exceeded architecturally sophisticated models across the 10 Medical Segmentation Decathlon tasks, which include CT-based targets where local features are the primary discriminators. The Task B result of this thesis provides a concrete example in the specific setting of HNSCC CT segmentation, suggesting that the standard U-Net's structural simplicity, its smaller parameter count, stronger local feature inductive bias, and lower overfitting risk constitutes an architectural advantage for this type of task.

5.3.4 Interpreting the Absolute DSC Values

The absolute DSC values for Task B (U-Net: 0.3000, ConvNeXt-UNet: 0.2558) are notably lower than published results from the HECKTOR challenge, where the best-performing teams achieved GTVp DSC values of 0.75 to 0.82 (Andrearczyk, Oreiller, Abobakr, et al., 2023). This gap reflects the combined effect of three structural constraints that distinguish the experimental conditions of this study from the published state-of-the-art. First, the absence of PET information: Guo et al. (2019) quantified DSC reductions of 0.15 to 0.20 when PET was withheld from multi-modal models, accounting for a notable fraction of the gap. Second, the 2D processing limitation, which removes inter-slice context that 3D volumetric models exploit for spatially consistent predictions and for detecting the elongated three-dimensional shape of

lymph node chains. Third, the 128×128 resolution, which reduces sub-centimeter structures to two or three pixels per side, making reliable detection effectively impossible.

The clinically important observation is therefore not the absolute values but the relative performance of the two architectures under identical constraints. The U-Net's consistent advantage across all five folds, confirmed by a statistically significant paired t-test ($p = 0.0212$), identifies it as the better-matched architecture for this specific task under the specific conditions of CT-only 2D processing, and this finding is reliable to the absolute performance level.

5.3.5 Evaluation Metric Considerations

DSC is the only quantitative metric reported in this chapter. DSC is the standard primary metric for medical image segmentation evaluation and is reliable to class imbalance, making it well-suited to the sparse foreground targets in both tasks (Taha & Hanbury, 2015). However, DSC has a known limitation as a sole reporting metric: it measures volumetric overlap without specifically penalizing boundary errors. Two predictions with identical DSC values may differ notably in the sharpness, smoothness, and clinical acceptability of their boundary contours. The Hausdorff distance, which measures the maximum distance between any point on the predicted boundary and the nearest point on the ground-truth boundary, is the standard complementary metric for capturing boundary quality (Taha & Hanbury, 2015). In the context of radiotherapy treatment planning, boundary accuracy is clinically critical: a systematic boundary offset of even 2 to 3 mm can translate into meaningful underdosage of the tumour margin or overdosage of adjacent critical structures. The absence of Hausdorff distance reporting in this study is acknowledged as a limitation, and future work should incorporate both DSC and Hausdorff distance as co-primary metrics to provide a more complete characterization of segmentation quality.

5.4 The Cross-Task Contrast: A Novel Contribution to Literature

5.4.1 Challenging the Assumption of Modality-Agnostic Superiority

The central theoretical implication of the cross-task contrast is that imaging modality should be treated as an explicit variable in architecture selection, rather than as an irrelevant background context. The published literature on ConvNeXt-based medical segmentation including MedNeXt (Roy et al., 2023), 3D UX-Net (Lee, 2022), CI-UNet (Dong et al., 2024), and others reviewed in Chapter 2, consistently report improvements over U-Net baselines in

single-modality evaluations and presents these improvements as evidence of general architectural superiority. The results of this thesis demonstrate that the implicit assumption of modality-agnosticism is empirically incorrect: an architectural modification that provides a 69.7% relative DSC improvement in FLAIR MRI brain tumour segmentation results in a statistically significant performance disadvantage in CT head and neck cancer segmentation under otherwise identical conditions.

This finding is not easily explained by differences in dataset size, training procedure, or evaluation protocol, because these were held identical across the two tasks. It is also not explained by a failure of the ConvNeXt-UNet to converge on Task B — the model converged stably and achieved competitive validation DSC during Phase 1 warm-up — but rather by a systematic tendency to underperform the simpler U-Net on validation patients across all five folds. The reversal is therefore genuinely attributable to an interaction between the architectural properties of the ConvNeXt-UNet and the specific imaging and task characteristics of the CT head and neck setting.

5.4.2 A Framework for Architecture-Modality Matching

Based on the theoretical analysis of both tasks, it is possible to identify three properties that jointly predict whether the ConvNeXt-UNet will outperform the standard U-Net for a given medical image segmentation task.

Property 1: Spatial diffuseness of the segmentation target. Tasks where the boundary between foreground and background is spatially gradual, requiring contextual information from a wide neighborhood for accurate delineation, benefit from the enlarged receptive field of the 7×7 depthwise convolution. Tasks where the boundary is locally detectable from intensity or texture cues do not require this enlarged context and may not benefit architecturally.

Property 2: Statistical proximity of the imaging modality to the pre-training distribution. Modalities whose images share statistical properties with the natural photographs on which ImageNet encoders are trained, smooth spatial variation, multi-scale structure, and moderate dynamic range, benefit more from the pre-trained ConvNeXt encoder's initialization. Modalities with mainly different statistics, quantitative absolute intensity scales, domain-specific artefacts, modality-specific noise patterns, benefit less and may be penalized by the larger parameter count relative to the smaller U-Net.

Property 3: Adequacy of the training dataset relative to model capacity. The ConvNeXt-UNet's 13:1 parameter advantage over the standard U-Net requires a proportionally larger training set to achieve comparable generalization. When the training data is adequate for a standard U-Net but borderline for the ConvNeXt-UNet, the U-Net's smaller capacity provides a regularization advantage.

For Task A, all three properties favour the ConvNeXt-UNet: LGG has a diffuse boundary, FLAIR MRI is statistically proximal to natural images, and the 10,929-slice dataset is sufficient for the ConvNeXtTiny encoder. For Task B, none of the three properties favour the ConvNeXt-UNet: HNSCC on CT has relatively local discriminative cues, CT is statistically distant from natural images, and the 7,774-slice single-centre dataset is borderline for a 32-million-parameter model. The observed performance pattern is entirely consistent with this framework.

This framework is preliminary and based on two tasks. Its validation requires testing additional modality-task combinations in which the three properties take different values, which is the most important direction for future work arising from this thesis.

5.5 Novelty and Scientific Contribution

The contributions of this study are presented in full in Section 6.3. In brief, the primary empirical contribution is the first controlled cross-modal paired evaluation of a ConvNeXt-UNet against a standard U-Net on both FLAIR MRI and CT segmentation under identical conditions — a design that produced the unexpected Task B reversal documented in Chapter 4. The secondary contribution is the experimental methodology itself: patient-wise GroupKFold cross-validation with matched fold assignments and pre-specified statistical tests. The third contribution is the documented experimental record including configurations that failed. The fourth contribution is the architecture-modality matching framework proposed in Section 5.4.2.

5.6 Limitations of the Study

5.6.1 Two-Dimensional Slice-Based Processing

The most consequential methodological limitation is the exclusive use of 2D axial slice processing. Three-dimensional volumetric processing is the standard approach for tasks where the segmentation target has meaningful three-dimensional extent, and it consistently improves

performance over 2D approaches by exploiting inter-slice context and enforcing spatial consistency between adjacent slices (Isensee et al., 2021; Roy et al., 2023). For Task A, 3D processing would improve boundary consistency at the superior and inferior poles of the tumour. For Task B, 3D processing would specifically improve GTVn detection by enabling the model to perceive the elongated superior-inferior shape of lymph node chains, a cue that is entirely invisible to a 2D model. The adoption of 2D processing was imposed by the 13 GB VRAM available on the Kaggle Tesla T4 GPU, and extending to 3D is the highest-priority direction for future work.

5.6.2 Spatial Resolution

All images were resampled to 128×128 pixels. At native CT resolution (typically 512×512 with 0.8–1.0 mm in-plane pixel spacing), a 10 mm lymph node spans approximately 10–12 pixels per side; after resampling to 128×128 it spans approximately 2–3 pixels, making reliable detection and delineation effectively impossible. The resolution constraint disproportionately affects Task B relative to Task A, because LGG tumours are typically larger relative to the image field of view at 128×128 brain MRI resolution. Higher-resolution processing (256×256 or 512×512) would improve boundary precision for both tasks but requires proportionally more GPU memory.

5.6.3 Single Evaluation Metric

The exclusive use of DSC as the quantitative evaluation metric means that boundary quality cannot be fully characterised from the results in Chapter 4. The Hausdorff distance, specifically the 95th percentile HD95, is the standard complementary metric for boundary quality evaluation in medical image segmentation and is required by the reporting standards of most major segmentation venues (Maier-Hein et al., 2024). In radiotherapy treatment planning, a systematic contour offset of 2–3 mm has direct clinical consequences, and DSC alone cannot detect such offsets if they affect both false-positive and false-negative regions symmetrically. Future work should report both DSC and HD95 as co-primary metrics.

5.6.4 Single-Centre Task B Data

Task B used the MDACC cohort from HECKTOR 2025 exclusively. CT-based segmentation performance is sensitive to scanner-specific acquisition parameters, tube voltage, current modulation, reconstruction kernel, and slice thickness that vary notably across institutions. A model trained and evaluated on MDACC data may not generalize to the imaging characteristics

of other institutions, and external validation on data from other participating centres would be required to establish cross-institutional generalizability. The Task B performance figures should therefore be interpreted as institution-specific estimates rather than general performance bounds.

5.6.5 No Clinical Validation

The study evaluates performance against radiologist-annotated ground truth. This is the standard evaluation paradigm for segmentation research but does not constitute clinical validation in the sense of demonstrating safety and effectiveness for actual treatment planning use. Clinical validation requires prospective studies measuring time saved, contour consistency after expert editing, and ultimately treatment outcomes. The results of this thesis justify investing in such studies for the Task A system, but they do not substitute for clinical evaluation.

5.7 Directions for Future Work

5.7.1 Three-Dimensional Extension

The highest-priority next step is implementing the ConvNeXt-UNet with 3D volumetric processing following the MedNeXt design (Roy et al., 2023) which applies 7×7×7 depthwise convolutions and volumetric skip connections. For Task A, this would improve boundary consistency at the tumour poles. For Task B, it would specifically enable GTVn detection through the superior-inferior shape cue. The principal engineering challenge is GPU memory: training a 3D ConvNeXtTiny-based model at 128×128×128 voxel resolution requires approximately 24–48 GB VRAM, accessible through cloud platforms with higher GPU tiers or institutional HPC clusters.

5.7.2 PET-CT Fusion for Task B

Incorporating the PET channel from the HECKTOR 2025 dataset would address the mossignificant information deficit of the Task B experiments. A dual-encoder ConvNeXt-UNet accepting CT and PET in parallel and fusing them at the bottleneck through attention-weighted integration could leverage FDG-PET's metabolic discriminability to complement the CT anatomical detail, consistent with the 0.15–0.20 DSC gains reported when PET is added to CT-only models (Guo et al., 2019). This extension would also allow testing whether the ConvNeXt-UNet advantage re-emerges when the FDG-PET channel introduces the spatial diffuseness of metabolic uptake, potentially re-satisfying Property 1 of the architecture-modality matching framework.

5.7.3 Domain-Specific Self-Supervised Pre-training

The domain mismatch identified in Section 5.3.2 motivates investigation of CT-specific pre-training for the ConvNeXt encoder. Woo et al. (2023) demonstrated that ConvNeXt V2 pre-trained with a fully convolutional masked autoencoder on ImageNet notably outperformed supervised ImageNet pre-training for downstream dense prediction. Applying an equivalent framework to large collections of unlabelled CT images, the full HECKTOR archive or the TCIA CT collection would produce feature representations learned from CT-specific statistics, potentially eliminating the domain mismatch while retaining the architectural advantages of ConvNeXt. This direction is particularly promising for Task B and represents a natural extension given the results presented here.

5.7.4 Validating the Architecture-Modality Matching Framework

The three-property framework proposed in Section 5.4.2 generates testable predictions for task-modality combinations beyond those examined in this thesis. Specifically, it predicts that ConvNeXt-UNet should outperform U-Net on diffuse, spatially extensive MRI targets (such as white matter lesion segmentation in multiple sclerosis) and on any MRI target where large-scale spatial context is clinically relevant, while U-Net should remain competitive or superior on compact, locally discriminable CT targets (such as vertebra or rib segmentation). Testing these predictions across a systematic benchmark spanning multiple modalities and target morphologies would provide the empirical foundation needed to elevate the framework from a preliminary interpretive account to a validated architectural selection guideline.

6. Conclusion

6.1 Returning to the Research Questions

This thesis began with two questions that existing literature had not addressed together. The first was whether replacing the standard 3×3 convolutional building block of a U-Net with a ConvNeXt-style block incorporating a 7×7 depthwise separable convolution, an inverted bottleneck structure, Layer Normalization, and GELU activation, produces measurable and statistically significant improvements in automated medical image segmentation. The second was whether any such improvements are consistent across imaging modalities, or whether they are specific to the particular combination of architecture and imaging characteristics under

which they were first observed. The experimental results answered both questions, but not in the way that a reading of the prior literature would have predicted.

For Task A, brain tumour segmentation from FLAIR MRI on the TCGA-LGG dataset, the ConvNeXt-UNet achieved a mean test-set DSC of 0.9328 ± 0.0125, compared to 0.5496 ± 0.0146 for the standard U-Net. This represents a relative improvement of 69.7%, confirmed with overwhelming statistical significance by a Wilcoxon signed-rank test applied to 2,186 paired per-image DSC scores ($p < 2.2\times10^{-308}$). The improvement was consistent across all five cross-validation folds without exception, and the ConvNeXt-UNet exceeded the 0.90 DSC threshold considered indicative of clinically acceptable automated segmentation in every fold, a level the standard U-Net did not approach in any fold.

For Task B, head and neck cancer segmentation from non-contrast CT on the HECKTOR 2025 MD Anderson cohort, the direction of the effect reversed entirely. The standard U-Net achieved a mean fold-level DSC of 0.3000 ± 0.0362, statistically significantly outperforming the ConvNeXt-UNet's 0.2558 ± 0.0214 (paired t-test: $t(4) = 3.6788$, $p = 0.0212$). The U-Net's advantage was again consistent across all five folds without exception. A Wilcoxon signed-rank test returned $p = 0.0625$, approaching but not reaching conventional significance at $\alpha = 0.05$ with only five-fold pairs, the parametric result, supported by the non-parametric trend, provides the primary statistical conclusion that the U-Net is the better-matched architecture for this specific task under these conditions.

The consistency of both effects across all folds in both tasks, combined with the use of an identical experimental framework and matched fold assignments for paired comparison, establishes these as genuine architectural effects rather than sampling artefacts.

The novelty of this study operates at a level that distinguishes it from most published architectural comparisons in medical image segmentation. It does not propose a new neural network component or a new training algorithm. Its original contribution is a controlled experimental design that for the first time places two architectures standard U-Net and ConvNeXt-UNet in direct paired comparison across two heterogeneous imaging modalities under identical conditions and reports the result honestly regardless of whether it confirms or contradicts the expectation. The finding that it contradicts the expectation that “modern" architecture is significantly worse on one of the two tasks is accurately what makes the contribution scientifically valuable. A study designed only to confirm known

advantages would not produce this result; a study designed to test the limits of those advantages can.

6.2 Summary of Principal Findings

Three principal findings emerge from the experimental programme and the discussion in Chapter 5.

Finding 1: The ConvNeXt-UNet provides a large and clinically meaningful performance advantage for diffuse MRI tumour segmentation. The 0.9328 mean DSC achieved on FLAIR LGG segmentation approaches the inter-observer agreement ceiling of approximately 0.85 reported between expert neuroradiologists (Bakas et al., 2017) and notably exceeds the published U-Net baseline of approximately 0.49 on the same dataset (López-Castaño, 2019). This result demonstrates that the architectural modernization principles of ConvNeXt (Liu et al., 2022) provide genuine functional advantages for segmentation tasks characterised by spatially diffuse boundaries where the enlarged effective receptive field of the 7×7 depthwise convolution enables contextual integration over the full extent of the transition zone that standard 3×3 kernels cannot achieve.

Finding 2: The standard U-Net outperforms the ConvNeXt-UNet on CT head and neck cancer segmentation. This reversal, statistically significant and directionally consistent across all folds, challenges the implicit assumption of modality-agnostic architectural superiority that characterizes virtually all published work on ConvNeXt-based medical segmentation. It arises from the interaction of three factors: the overfitting tendency of the larger ConvNeXt-UNet on the available CT training data, the domain mismatch between ImageNet pre-training and CT image statistics, and the adequacy of local contextual information for the discriminative requirements of non-contrast CT tumour boundary detection, conditions that collectively favour the simpler, more locally focused standard U-Net.

Finding 3: Architectural performance is modality-dependent, not modality-agnostic. The cross-task contrast is the most theoretically significant result of this study. The same architectural modification that produces a 69.7% relative improvement on FLAIR MRI produces a statistically significant performance disadvantage on CT under otherwise identical conditions. This finding has direct implications for how architectural comparisons in medical image segmentation should be conducted and reported: single-modality validation is

insufficient to establish general architectural superiority, and cross-modal evaluation is an empirically essential component of any rigorous comparative study.

6.3 Contributions of the Thesis

6.3.1 Empirical Contribution

This study provides what is, to the author's knowledge, the first controlled cross-modal comparison of a ConvNeXt-UNet against a standard U-Net on both FLAIR MRI brain tumour segmentation and CT head and neck cancer segmentation under identical experimental conditions, using patient-wise GroupKFold cross-validation with matched fold assignments for genuinely paired statistical comparison. The finding of a systematic performance reversal between modalities with the ConvNeXt-UNet excelling on MRI and the U-Net excelling on CT is empirically novel. It cannot be deduced from the existing single-modality literature, which consistently shows ConvNeXt advantages over U-Net baselines across MRI brain tumour, skin lesion, polyp, and retinal vessel tasks (Dong et al., 2024; Kamsari et al., 2025; Lee, 2022; Roy et al., 2023) and would, if taken at face value, predict ConvNeXt superiority on both tasks. The Task B result constitutes a necessary corrective to this generalization.

The Task A result also has a standalone empirical value. Achieving a mean DSC of 0.9328 on single-modality FLAIR segmentation of LGG, using 2D processing and a single freely available Kaggle GPU, without multi-modal input or volumetric processing, is a practically significant result for clinical research groups who lack access to multi-modal data or high-end computing infrastructure. The novelty of this empirical finding lies not in the magnitude of the Task A improvement, large as it is, but in the Task B reversal, a result that single-modality evaluations structurally cannot produce and that challenges the research community's current practice of generalizing from single-modality benchmark results to universal architectural recommendations. This thesis demonstrates that such generalization is empirically unsafe, and that cross-modal evaluation is a necessary scientific standard for architectural claims in medical image segmentation.

6.3.2 Methodological Contribution

The experimental design demonstrates several methodological practices that are important for valid architectural comparisons in medical image segmentation but are inconsistently applied in the published literature. Patient-wise cross-validation using GroupKFold, with the same fold

assignments applied to both competing models, prevents data leakage and ensures that performance differences reflect genuine differences in generalization to unseen patients rather than differences in the difficulty of the validation split (Zech et al., 2018). Pre-specification of the statistical analysis plan choosing the Wilcoxon signed-rank test for Task A and the paired t-test for Task B based on distributional properties before examining results, prevents post-hoc test selection bias (Demsar, 2006). The transparent documentation of the two-phase training strategy, the decision rationale for every preprocessing choice, and the experimental trajectory including approaches that did not work provides reproducibility value that most published architectural comparisons lack.

6.3.3 Conceptual Contribution

The architecture-modality matching framework proposed in Section 5.4.2 offers a principled account of when ConvNeXt-based designs are likely to outperform standard U-Net architectures, organizing the prediction around three properties: the spatial diffuseness of the segmentation target, the statistical proximity of the imaging modality to the ImageNet pre-training distribution, and the adequacy of the training dataset relative to the model's parameter count. This framework generates testable predictions for task-modality combinations beyond those examined in this thesis and represents a step toward a theoretical account of architectural-modality interaction that existing literature lacks.

6.4 Practical Implications

Beyond their scientific significance, the results of this thesis carry practical implications for clinical research groups considering the adoption of ConvNeXt-based segmentation architectures.

For groups working on brain tumour MRI segmentation, particularly on single-modality FLAIR data or other diffuse lesion types, the results strongly support adopting a ConvNeXt-UNet over a standard U-Net. The performance advantage is large, statistically significant, and achieved with modest computational resources. The two-phase frozen-backbone warm-up training strategy documented in this thesis provides a reproducible approach to stable fine-tuning that avoids the catastrophic forgetting problem that can arise when a large pre-trained encoder is trained end-to-end from the first epoch.

For groups working on head and neck cancer CT segmentation under resource constraints similar to those of this study, single modality, 2D processing, limited GPU memory, the results

suggest that the standard U-Net remains a stronger baseline than the ConvNeXt-UNet and that architectural investment would be better directed toward incorporating PET data, extending to 3D volumetric processing, or improving the spatial resolution of processing rather than replacing the encoder architecture. The relative simplicity of the standard U-Net — its smaller parameter count, lower GPU memory footprint, and faster training time — is a practical advantage.

More broadly, the results support a principle of architectural parsimony: adopting a more complex architecture is justified when the complexity provides a specific functional advantage matched to the task requirements, not as a general policy based on benchmark performance on unrelated tasks. The standard U-Net, properly trained with patient-wise cross-validation and appropriate loss functions, remains a strong and practically competitive baseline in settings where the conditions for ConvNeXt advantage are not met.

6.5 Limitations Revisited

The limitations of this study are addressed in detail in Section 5.6. They include the restriction to 2D slice processing, the 128×128 spatial resolution, the use of DSC as the sole evaluation metric, the single-centre Task B cohort, and the absence of prospective clinical validation. None of these invalidates the primary conclusion that architectural performance is modality-dependent, but each represents a dimension along which future work should extend these findings.

6.6 Future Research Agenda

The four most productive directions for future work are detailed in Section 5.7. In order of priority: 3D volumetric extension following the MedNeXt design; PET-CT fusion for Task B; CT-specific self-supervised pre-training using the ConvNeXt V2 masked autoencoder framework; and systematic multi-modal benchmarking to validate the architecture-modality matching framework proposed in Section 5.4.2.

6.7 Closing Reflection

A segmentation model with a DSC of 0.93 is not inherently useful. It becomes useful when it saves a radiologist thirty minutes per patient and does so consistently enough that clinicians learn to trust it, and its value is ultimately measured not by benchmark metrics but by whether it translates into faster, more consistent, and safer treatment for patients. The dramatic

improvement in LGG FLAIR segmentation demonstrated in this thesis from a mean DSC of 0.55 to 0.93 through an architectural change that requires no additional clinical data, no extra imaging sequences, and no specialist computing infrastructure represents exactly the kind of practically accessible advance that can narrow the gap between research demonstration and clinical adoption.

At the same time, the Task B result is a reminder that progress in deep learning for medical image segmentation is not monotonic, and that architectural innovations should be evaluated rigorously across the conditions of their intended deployment rather than adopted on the basis of performance in unrelated settings. The finding that the same architectural modification which dramatically improved MRI segmentation significantly harmed CT segmentation is not a failure of the research design; it is the research design working exactly as intended. Controlled comparative experiments that can detect both improvements and deteriorations provide more useful scientific knowledge than evaluations constructed to confirm a hypothesis.

The ConvNeXt-UNet is not the right architecture for every segmentation task. It works well for diffuse MRI targets with datasets large enough to support its parameter count. It works poorly on compact CT targets with limited data. Knowing this is more useful to the field than knowing only the cases where it succeeds. Architecture selection in medical AI should be driven by the properties of the clinical task, not by the results of benchmarks on unrelated data. That principle, illustrated through the contrasting outcomes of the two tasks examined in this thesis, is perhaps the most durable conclusion of this work.

7. References

1. Andrearczyk, V., Oreiller, V., Abobakr, M., Akhavanallaf, A., Balermpas, P., Boughdad, S., Capriotti, L., Castelli, J., Cheze Le Rest, C., Decazes, P., Correia, R., El-Habashy, D., Elhalawani, H., Fuller, C. D., Jreige, M., Khamis, Y., La Greca, A., Mohamed, A., Naser, M., … Depeursinge, A. (2023). *Overview of the HECKTOR Challenge at MICCAI 2022: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT* (pp. 1–30).
2. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., & Cardoso, M. J. (2022). The Medical Segmentation Decathlon. *Nature Communications*, *13*(4128).
3. Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). *Layer Normalization*.
4. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F. C., Pati, S., Prevedello, L. M., Rudie, J. D., Sako, C., Shinohara, R. T., Bergquist, T., Chai, R., Eddy, J., Elliott, J., Reade, W., … Bakas, S. (2021). *The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification*.
5. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., Freymann, J. B., Farahani, K., & Davatzikos, C. (2017). Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features.
6. Brouwer, C. L., & Steenbakkers, R. J. H. M. (2015). CT-based delineation of organs at risk in head and neck region. *Radiotherapy and Oncology*, *117*(1), 83–90.
7. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., & Zhou, Y. (2021). *TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation*.
8. Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.
9. Ding, X., Zhang, X., Han, J., & Ding, G. (2022). Scaling Up Your Kernels to 31×31: Revisiting Large Kernel Design in CNNs. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11953–11965.
10. Dong, C., Zheng, L., Liu, Y., Feng, G., Zhou, L., & Wen, Z. (2024). *CI-UNet: Melding ConvNeXt and cross-dimensional attention for robust medical image segmentation*.
11. Dosovitskiy, A. (2021). An image is worth 16×16 words. *ICLR*.
12. Goodfellow, I. (2016). *Deep Learning*. MIT Press.

13. Gregoire, V., Jeraj, R., Lee, J. A., & O'Sullivan, B. (2018). *Radiological contouring for head and neck cancer: current concepts and future developments*.
14. Guo, Z., Guo, N., Gong, K., Zhong, S., & Li, Q. (2019). Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network. *Physics in Medicine & Biology*, *64*(20), 205015.
15. Hashemi, R. H., Bradley, W. G., & Lisanti, C. J. (2010). *MRI: The Basics*. Lippincott Williams and Wilkins.
16. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., & Xu, D. (2022). *SwinUNETR: Swin transformers for semantic segmentation of brain tumors in MRI images*.
17. He, K., Chen, X., Xie, S., Li, Y., Dollar, P., & Girshick, R. (2022). Masked Autoencoders Are Scalable Vision Learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15979–15988.
18. Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (GELUs). arXiv preprint. *ArXiv Preprint*.
19. Hounsfield, G. N. (1980). Computed medical imaging. *Journal of Computer Assisted Tomography*, *4*(5), 665–674.
20. Ioffe, S., & Szegedy, C. (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*.
21. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). *nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation*.
22. Kamsari, M., Sadeghi, S., de Oliveira, G. G., Alves, A. M., Sarshar, N. T., Anari, S., & Ranjbarzadeh, R. (2025). The role of data augmentation and attention mechanisms in UNet and ConvNeXt architectures for optimizing breast tumor segmentation. *Scientific Reports*, *15*(1), 45268.
23. Lee, H. H. (2022). *3D UX-Net*. https://doi.org/10.48550/arXiv.2209.15076
24. Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal Loss for Dense Object Detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007.
25. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88.
26. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002.
27. Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., & Xie, S. (2022). *A ConvNet for the 2020s*.
28. López-Castaño, M. A. (2019). U-Net glioblastoma segmentation. *CBMS*.

29. Luo, W., Urtasun, R., Li, Y., & Zemel, R. (2016). Effective receptive field. *NeurIPS*.
30. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M. A., Wiesenfarth, M., Kavur, A. E., Sudre, C. H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, T., … Jäger, P. F. (2024). Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, *21*(2), 195–212.
31. Menze, B. H. (2015). BRATS benchmark. *IEEE TMI*.
32. Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., … Van Leemput, K. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, *34*(10), 1993–2024.
33. Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *2016 Fourth International Conference on 3D Vision (3DV)*, 565–571.
34. Nikolov, S., Blackwell, S., Zverovitch, A., Mendes, R., Livne, M., De Fauw, J., Patel, Y., Meyer, C., Askham, H., Romera-Paredes, B., Kelly, C., Karthikesalingam, A., Chu, C., Carnell, D., Boon, C., D'Souza, D., Moinuddin, S. A., Garie, B., McQuinlan, Y., … Ronneberger, O. (2021). Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study. *Journal of Medical Internet Research*, *23*(7), e26151.
35. Oktay, O., Schlemper, J., Folgoc, L. Le, Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. (2018). *Attention U-Net: Learning Where to Look for the Pancreas*.
36. Patel, S. H. (2019). Low-grade glioma imaging. *Radiologic Clinics of North America*.
37. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2825–2830.
38. Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation* (pp. 234–241).
39. Roy, S., Koehler, G., Ulrich, C., & al., et. (2023). *MedNeXt: Transformer-driven scaling of ConvNets for medical image segmentation*.
40. Sanai, N., Chang, S., & Berger, M. S. (2011). Low-grade gliomas in adults. *Journal of Neurosurgery*, *115*(5), 948–965.

41. Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2023). Transformers in medical imaging: A survey. *Medical Image Analysis*, *88*, 102802.
42. Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, *52*(3/4), 591.
43. Siddique, N., Paheding, S., Elkin, C. P., & Devabhaktuni, V. (2021). U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications. *IEEE Access*, *9*, 82031–82057.
44. Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Jorge Cardoso, M. (2017). *Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations* (pp. 240–248).
45. Sung, H. (2021). Global cancer statistics 2020. *CA Cancer Journal*.
46. Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, *15*(1), 29.
47. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., & al., et. (2016). *Convolutional neural networks for medical image analysis*.
48. Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: methodological failures and recommendations for the future. *Npj Digital Medicine*, *5*(1), 48.
49. Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80.
50. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., & Xie, S. (2023). *ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders*.
51. Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). *How transferable are features in deep neural networks?*
52. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). *Variable generalization performance of a deep learning model to detect pneumonia*.
53. Zhang, R., Chen, L., & Wang, S. (2023). *Large kernel convolutions for brain tumour boundary detection in FLAIR MRI*.
54. Zhou, H.-Y. (2023). Self-supervised medical imaging. *ICCV*.