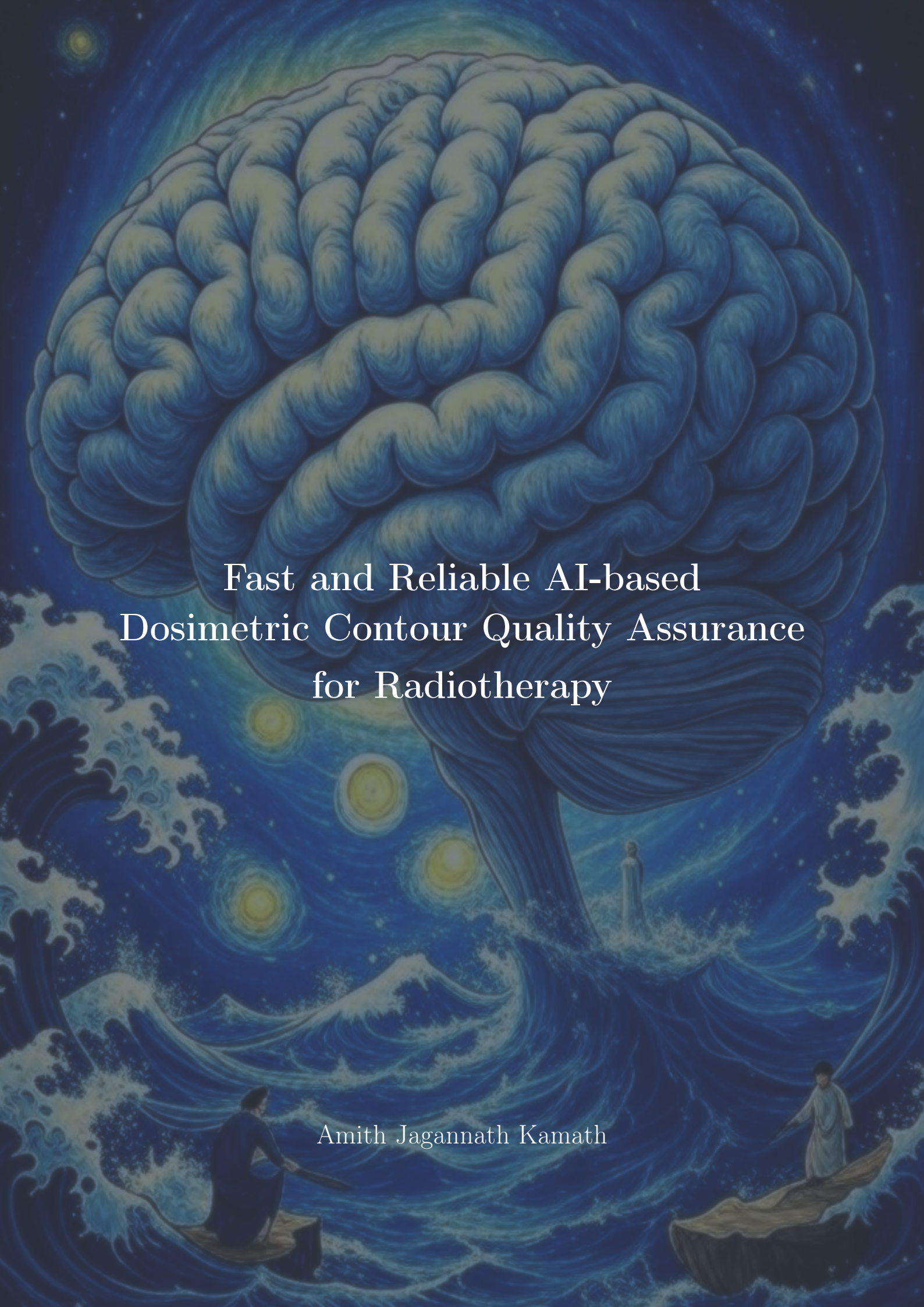


Fast and Reliable AI-based Dosimetric Contour Quality Assurance for Radiotherapy

Amith Jagannath Kamath

UNIVERSITÄT
BERN

Graduate School for Cellular and Biomedical Sciences
University of Bern

Fast and Reliable AI-based Dosimetric Contour Quality Assurance for Radiotherapy

PhD Thesis submitted by

Amith Jagannath Kamath

for the degree of

PhD in Biomedical Engineering

Supervisor
Prof. Dr. Mauricio Reyes
ARTORG Center for Biomedical Engineering Research
Faculty of Medicine of the University of Bern

Co-advisor
Prof. Dr. Nicolaus Andratschke
Klinik für Radio-Onkologie
Faculty of Medicine of the University of Zürich

"Tamaso mā jyotir gamaya / From darkness (lead me) to light."

Bṛhadāraṇyaka Upanishad (circa 7th–6th century BCE)

* * *

Declaration on the Use of Generative Artificial Intelligence (AI) and AI-Assisted Technologies: During the preparation of this work, Generative AI tools were used in order to improve grammar and language. After using this tool/service, the content was reviewed and edited as needed. Full responsibility for the accuracy of the content of this thesis is assumed by the author.

University of Bern - Faculty of Medicine - ARTORG Center for Biomedical Engineering Research

Abstract

Fast and Reliable AI-based Dosimetric Contour Quality Assurance for Radiotherapy

by Amith Jagannath Kamath

Radiotherapy planning relies heavily on accurate anatomical contouring and treatment planning. Contour quality assurance therefore assumes great importance for the efficacy of the overall treatment. Traditional geometric metrics have been shown to be uncorrelated with the dose distribution and hence patient outcomes. This thesis attempts to move the needle towards dosimetric contour quality assurance through three interconnected efforts: (1) validating the clinical need for quality assurance grounded in dosimetry, (2) assessing the reliability and speed of automated tooling to assist in such processes, and (3) exploring the viability of novel applications and methods impacting the workflow.

This work begins by investigating the perceptual and predictive capacities of clinical experts regarding the dosimetric consequences of segmentation inaccuracies. Through structured interviews and empirical studies, it demonstrates that deep learning based dose predictors could outperform human experts in triaging dosimetrically sub-optimal contours.

Then, it explores the sensitivity of dose prediction models to segmentation variability. Evaluations reveal that while these models accurately reflect dose changes induced by clinically relevant contour variations, their sensitivity varies across anatomical sites and expert-driven variabilities. Since such models commonly use the U-Net deep learning architecture, it examines its robustness under varying task complexities and domain shifts. Contrary to common assumptions, findings reveal that advanced U-Net variants are not universally more robust than simpler architectures. The research highlights a nuanced interplay between skip-connections and task complexity, showing that design choices must be context-aware to ensure robustness, especially in out-of-distribution scenarios.

Finally, it delves into two novel applications, starting with the development of ASTRA, a visualization tool for local sensitivity maps, enabling real-time, dosimetric contour evaluations, integrating both geometric and dosimetric uncertainties for comprehensive radiotherapy quality assurance. This is followed by another novel framework, AutoDoseRank, to automate segmentation prioritization based on clinical impact. This provides a pathway for dosimetric quality assurance in routine workflows.

These contributions collectively advocate for a paradigm shift from purely geometric to clinically meaningful, dosimetric contour evaluations. By attempting to bridge the gap between algorithmic performance and clinical utility, this work enhances the reliability, safety, and efficiency of radiotherapy planning, laying the groundwork for intelligent quality assurance for the next generation of treatment planning.

Contents

List of Figures

List of Tables

1 Introduction

"We make our world significant by the courage of our questions and the depth of our answers."

— Carl Sagan, in Cosmos (Chapter 12), 1980.

This chapter starts with the clinical background in Section 1.1, followed by the challenges in the treatment process in Section 1.2. Next, it discusses the unmet clinical needs in Section 1.3, and ends with the research hypothesis and the structure of the thesis in Section 1.4. New terms are explained in Chapter 2, which readers may skip to go directly to the clinical needs in Part I, technical contributions in Part II, and proof-of-concept experiments in Part III. A reading guide is shown in Figure 1.7.

1.1 Clinical Background

Glioblastoma, also known as Glioblastoma Multiforme (GBM), is the most common and aggressive primary malignant brain tumour in adults [1], with an incidence of about 3 per 100,000 [2, 3]. It was first recorded in 1800 and named in 1926 by Percival Bailey and Harvey Cushing, with "multiforme" describing its varied appearance due to necrosis (tissue death), bleeding, and cysts [4]. The World Health Organization (WHO) classifies GBM as a Grade IV astrocytoma, marked by fast growth, widespread invasion, and a tendency for necrosis and angiogenesis (new blood vessel formation to support tissue growth) [5]. GBM makes up 14.6% of all primary brain and Central Nervous System (CNS) tumours, 48.3% of primary malignant brain tumours, and 57.3% of all gliomas in adults [6]. It is more common in males than females and the mean age at diagnosis is between 59 and 62 years [2], with incidence increasing significantly to 15.24 per 100,000 between ages 75 to 84 [7].

GBMs are divided into two main types: primary (de novo) and secondary (progressive). Primary GBM appears suddenly without a prior lesion, usually affects older patients (average age of 62 years), and makes up about 95% of cases. Secondary GBMs develop from lower-grade gliomas, typically occur in younger patients (average age of 45 years), and represent around 5% of cases [8]. The Isocitrate Dehydrogenase (IDH) mutation status is an important molecular marker for GBM classification. According to the 2021 WHO classification [5], IDH-mutant astrocytomas (previously called secondary GBMs) are distinct from IDH-wildtype GBMs,

as they have different outcomes. Patients with IDH-mutant tumours generally live longer than those with IDH-wildtype GBMs [9].

1.1.1 Motivation: Glioblastoma Treatment

Managing GBM in clinical practice is difficult due to its complex nature. GBM shows significant differences within and between tumours at the genetic, molecular, and cellular levels [10]. These tumours have varied genetic mutations and molecular profiles, which result in different physical traits and responses to treatment [11]. This variation affects imaging features, such as differences in contrast enhancement, oedema, and necrosis seen on Magnetic Resonance Imaging (MRI) scans [12]. These differences make it hard to monitor tumour behaviour and growth patterns, which impacts treatment outcomes and complicates the development of effective therapies.

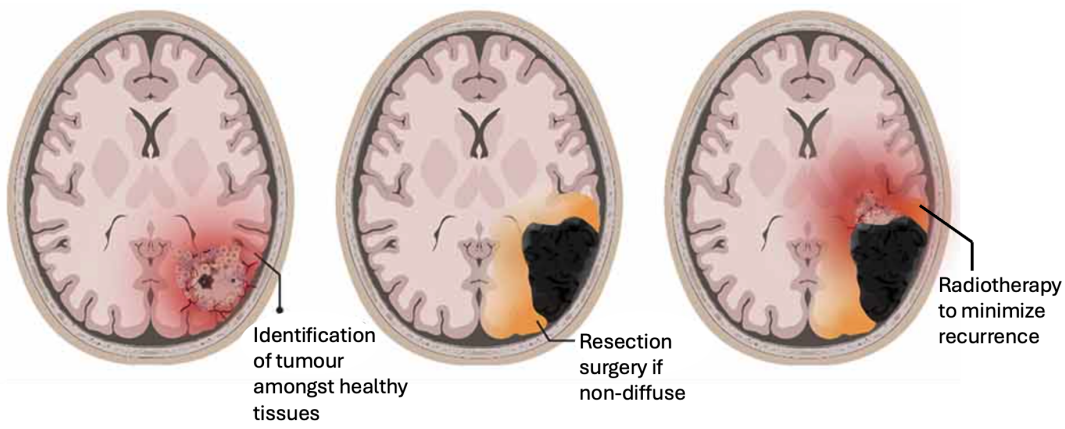


FIGURE 1.1: Standard-of-care treatment for GBM relies on identification of tumour boundaries, maximally resecting it, and then performing post-operative Radiotherapy (RT) to target residual microscopic extensions of the tumour cells at the boundaries of the resection area. Figure adapted from [13].

Standard-Of-Care: The standard treatment for GBM, known as the "Stupp protocol", involves a combination of methods. It includes maximal safe surgical removal of the tumour, followed by RT and chemotherapy with temozolomide [14]. This approach shown in Figure 1.1 aims to extend survival and enhance the quality of life for patients with GBM. The treatment consists of the following components:
Surgical Resection is the first and most important step in treating GBM. The main goals are to collect tissue for diagnosis and to reduce the tumour size while protecting brain function [15]. Research shows that removing more of the tumour improves overall survival and delays disease progression [16]. Advanced tools, such as intraoperative MRI [17], neuro-navigation systems [18], and fluorescence-guided surgery with 5-Aminolevulinic Acid (5-ALA) [19], help surgeons achieve the maximum safe removal of the tumour. However, complete removal is often not possible because the tumour spreads into surrounding brain tissue.
Postoperative RT is a key part of GBM treatment. Standard external beam RT provides a total dose of 60 Gray (an absorbed dose of 1 Joule/kilogram), given in 2 Gray doses over six weeks, targeting remaining tumour cells and areas with possible microscopic spread [14]. RT significantly increases average survival time compared to surgery alone [20]. Methods like Intensity-Modulated Radiotherapy (IMRT) and

Stereotactic Radiosurgery (SRS) improve precision, reducing harm to healthy brain tissue [21]. Shorter RT schedules are used for elderly patients or those in poor health to lessen treatment time and side effects [22]. Artificial Intelligence (AI) is a promising tool to improve RT by aiding in tumour contouring, refining treatment plans, and predicting how well treatment response [23].
Chemotherapy with temozolomide, an oral drug that damages tumour Deoxy-Ribonucleic Acid (DNA), is a key part of GBM treatment. The Stupp protocol [14] showed that using temozolomide alongside RT, followed by six cycles of temozolomide alone, greatly improves average survival and two-year survival rates compared to RT alone. The O6-Methyl Guanine-DNA Methyltransferase (MGMT) promoter's methylation status predicts how well temozolomide will work, with methylated tumours responding better [24]. However, resistance to temozolomide and tumour recurrence are major issues, leading to research on new drugs and combined treatments [25].

Dismal Prognosis and its Causes: Despite improvements in diagnosis, surgery, and treatment methods, GBM remains one of the most deadly cancers with a very poor outlook [26]. The average survival time after treatment is between 12–17 months [2, 14], but without treatment, it falls to around 6.1 months [27]. The two-year survival rate is approximately 26–33% [28], and the five-year survival rate for primary GBM is very low, reported at 2.2% from a single centre in France in 2024 [29]. Efforts to improve these outcomes have not led to significant progress [30], highlighting the need for large, multi-centre clinical trials to explore new treatment options.

GBM is highly aggressive due to its invasive growth, varied genetic makeup, and resistance to standard treatments [1]. The tumour cells spread widely into nearby healthy brain tissue, causing invasive growth and nearly always recurring, often within 2 centimetres of the original surgical site [31]. This makes GBM a "whole brain" or "whole CNS" disease, making complete tumour removal very difficult [32]. Despite strong multi-treatment approaches, almost all patients face tumour recurrence [33, 34]. Treating recurrent GBM is especially hard, as there are no clear standard guidelines [35].
Several factors contribute to the poor prognosis due to the ineffectiveness of current GBM treatment, including:

- **Diffuse Infiltration** of GBM cells into important brain areas makes complete surgical removal difficult. Maximal safe resection is a key part of GBM treatment, aiming to reduce tumour size and ease symptoms [14]. Research shows that removing more of the tumour improves survival, but the benefits lessen when there is a risk of causing neurological problems after surgery [36]. The cancer cells spread along white matter paths and blood vessels, often several centimetres from the primary tumour [37].

  These cells go beyond the resolution of current neuro-imaging tools, making it hard to identify exact tumour edges before or during surgery [38]. As a result, fully removing the tumour is nearly impossible without causing significant brain function damage, as the infiltrating cells are in critical brain areas [39].

- **Remarkable Resistance** is shown by GBM to standard treatments such as chemotherapy and radiation therapy [26]. This resistance is partly due to some GBM cells being located in areas of the tumour with low oxygen and tissue death [40].

  The Blood-Brain Barrier (BBB) and the natural resistance of glioma cells [41] create a major barrier to effective treatment, reducing the delivery and impact

of chemotherapy drugs [42–44]. Although the BBB is partly disrupted in GBM, the blood-tumour barrier still blocks many chemotherapy drugs from reaching the tumour in high enough amounts. Additionally, invasive tumour cells at the edges are protected by an intact BBB [27].

- **Therapeutic Challenges** arise from the wide variation in GBM at multiple levels, leading to treatment resistance and tumour recurrence [10]. This variation in physical traits adds to the extensive genetic differences, increasing the tumour's aggressiveness and creating significant barriers to effective treatment [45].

  Unknown underlying mechanisms complicates the development of targeted therapies. GBM cells show increased ability to move, driven by changes in cell adhesion molecules, extracellular matrix components [46], and signalling pathways like integrins, Matrix Metalloproteinases (MMPs)s, and the Phosphatidylinositol 3-Kinase (PI3K)/Akt pathway [47]. The tumour micro-environment, including low oxygen levels and interactions with surrounding cells, also encourages invasive behaviour [48].

1.1.2 Radiotherapy: an Essential Component

In the standard treatment protocol, RT is essential for managing tumour growth and improving patient survival. Complete surgical removal is not feasible due to the tumour's invasive characteristics [16]. Therefore, post-operative RT targets remaining tumour cells in the resection cavity and nearby brain tissue [49] to lower the chances of local tumour recurrence. RT employs high-energy ionizing radiation from the electromagnetic spectrum and can be delivered with four purposes: (i) curative: to eliminate cancer, often combined with chemotherapy; (ii) adjuvant: to support surgery and reduce the risk of tumour recurrence; (iii) palliative: to alleviate symptoms and improve patient comfort; and (iv) neo-adjuvant: to shrink tumours before surgery, such as in rectal cancer [50].

The core principle of RT lies in selectively targeting cancer cells while preserving healthy cells. Cancer cells often exhibit impaired DNA repair mechanisms, making them unable to recover from RT-induced DNA damage. In contrast, healthy cells can repair such damage to a certain extent. However, exceeding a tissue-specific threshold of ionizing radiation can overwhelm the repair capacity of normal cells, potentially leading to secondary cancers where healthy cells become malignant [51]. To mitigate the risks of normal tissue toxicity, a stringent radiation protection framework, encompassing laws, procedures, and regulations, is essential to minimize these hazards.

RT is typically administered using photons or other charged particles [52]. Photon-based therapy employs high-energy X-rays that penetrate deeply into body tissues while minimizing damage to the skin surface. These X-rays generate secondary electrons that disrupt DNA in both cancerous and healthy cells [53]. Both photon and electron therapies rely on a Linear Particle Accelerator (LINAC) for precise delivery. Proton therapy is increasingly utilized, particularly for paediatric cases and brain tumours, due to its ability to deliver radiation with minimal impact on surrounding healthy tissues [54]. However, challenges include its high cost, limited availability, and insufficient clinical evidence demonstrating clear advantages over other methods in many adult cancers [55]. Unlike photons, which travel through the body in near-linear paths, protons can be targeted to deposit energy locally, sparing tissues beyond the tumour site.

Treatment planning for GBM is complex due to the difficulty of defining Target Volume (TV) that include all infiltrative tumour cells while minimizing toxicity to healthy brain tissue [56]. To address microscopic disease, margins are typically extended beyond (called the Clinical Target Volume (CTV)) the visible tumour (called the Gross Tumour Volume (GTV)) on imaging studies, but this method lacks precision and may fail to encompass all infiltrative cells while exposing normal tissue to radiation [57]. The diffuse infiltration of GBM into surrounding brain tissue, often beyond what conventional imaging can detect [58], poses significant challenges for radiation oncologists in accurately delineating the tumour. This ambiguity complicates the use of automated tools, which may struggle to differentiate tumour tissue from normal brain tissue, potentially underestimating tumour volume [59]. Unclear margins reduce the precision of RT targeting. Both conventional approaches and emerging AI models designed to enhance treatment planning must account for the microscopic spread of tumour cells [60], a task made difficult by the lack of clear imaging markers.

Current imaging modalities, such as MRI and Positron Emission Tomography (PET), are limited in their ability to detect microscopic tumour infiltration [38]. Advanced techniques, including Diffusion Tensor Imaging (DTI) and Magnetic Resonance Spectroscopy (MRS) [61], provide greater sensitivity but are not commonly used in clinical settings due to economic, technical, and interpretive challenges. The difficulty in precisely identifying the full extent of tumour infiltration complicates effective surgical planning and accurate RT targeting.

1.2 Workflow Challenges

The foundation of effective radiation therapy is meticulous treatment planning. This section addresses two primary challenges in this process, which form the motivation of the research objectives explored throughout this thesis. Treatment planning commences with the acquisition of medical images, such as Computed Tomography (CT), PET, and MRI, capturing the patient's anatomy. These images form the backbone for contouring or delineation, where TV, encompassing the tumour and areas at risk for microscopic disease, and Organ at Risk (OAR) are carefully outlined [62].

Manual contouring is typically performed by radiation oncologists or dosimetrists using specialized software to delineate structures slice-by-slice on medical images. This task demands extensive anatomical knowledge, clinical expertise, and careful interpretation of imaging features. The resulting contours guide medical physicists in developing treatment plans that deliver the prescribed radiation dose to the TV while minimizing exposure to healthy OAR tissues [64]. Accurate contouring is critical, as it directly impacts treatment outcomes. Errors in contouring can result in geographic miss of the TV, leading to under-dosing of the tumour, or excessive irradiation of healthy OAR tissues, causing toxicity. Despite standardized guidelines and protocols [63] as shown in Figure 1.2, manual contouring remains heavily reliant on human judgment, leading to considerable inter- and intra-observer variability, which poses a significant challenge in RT planning [65]. Section 1.2.1 elaborates on these challenges and discusses current advancements in automation to address them.

The subsequent phase of treatment planning employs advanced techniques such as IMRT and Volumetric Modulated Arc Therapy (VMAT), which involve a highly iterative process heavily reliant on manual input. Treatment planners undertake multiple trial-and-error iterations to balance the competing goals of effective tumour

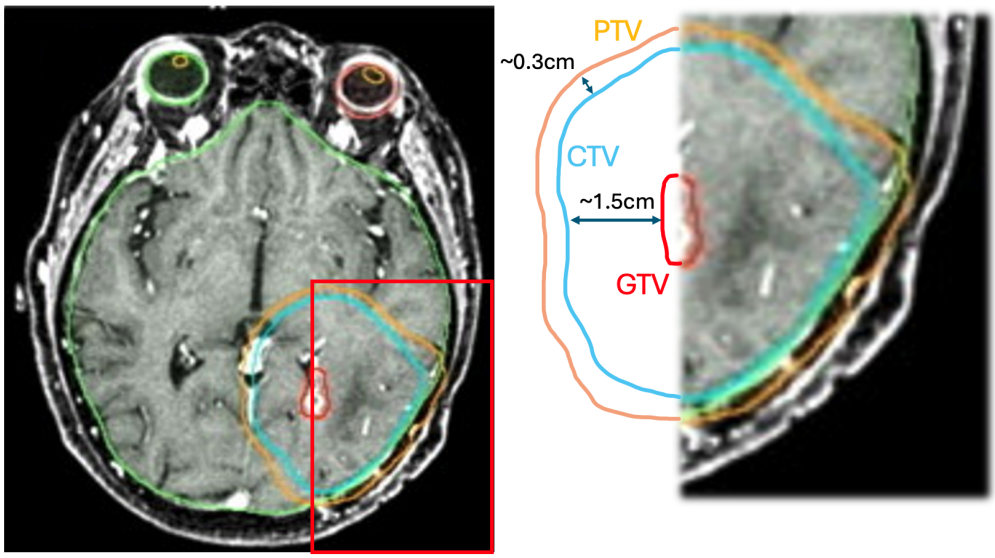


FIGURE 1.2: Typical situation in contouring, where the OARs like the eyes are delineated. Zoomed in section (in red) on the right shows the guidelines for TVs from [63], indicating typical extensions of the GTV to form the CTV and then the Planning Target Volume (PTV).

control and the preservation of healthy tissue [66]. This process requires repeated adjustments to planning parameters, dose constraints, and optimization objectives until a clinically acceptable plan is achieved [67, 68]. Section 1.2.2 provides a detailed examination of the challenges associated with this step, which leads to the core problem statement of this thesis, further discussed in Section 1.3.

1.2.1 Variability in Contouring Practice

Manual contouring in RT treatment planning is prone to significant variability, which presents in two main forms: *inter-observer variability*, arising from differences among clinicians, and *intra-observer variability*, stemming from inconsistencies in repeated contouring by the same clinician [69]. This issue is substantial, as inter-observer variability in TV contouring can exceed errors in other stages of the treatment planning and delivery process [70, 71]. Such variability may introduce systematic errors in dose delivery, potentially compromising local disease control [72]. The factors contributing to contour variability are illustrated in Figure 1.3.

Sources of variability: The causes of contour variability in RT treatment planning are complex and diverse. A primary factor is the *subjective interpretation* of anatomical boundaries, particularly in regions with limited tissue contrast [73]. This challenge is pronounced in tumours where soft tissue boundaries are obscured by motion or poor contrast with adjacent OARs [74]. For instance, studies on pancreatic tumours have demonstrated significant discrepancies, with ratios of the largest to smallest delineated GTV within the same patient reaching as high as 6.8 [75].

Image quality plays a critical role in contour accuracy, with artifacts, calcifications, and motion blur contributing to increased variability [76]. Certain anatomical structures pose greater challenges than others. Organs such as the lungs and bladder, which have well-defined boundaries, allow for consistent delineation. In contrast, structures like the parotid glands and small bowel, with less distinct borders, result

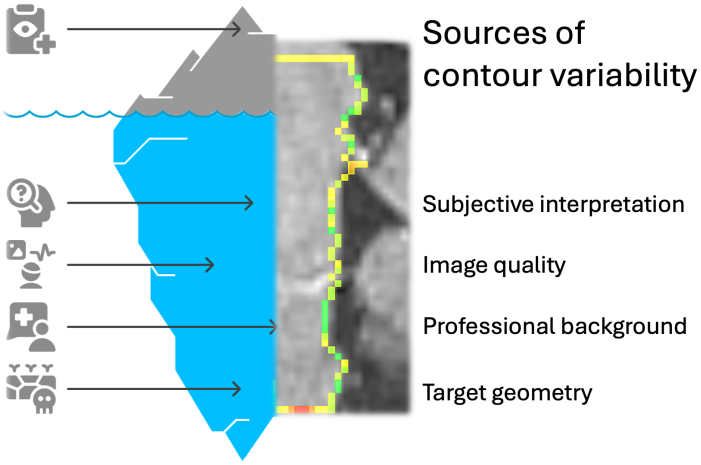


FIGURE 1.3: The various causes of contour variability are described in this iceberg representation. Right half of the figure represents an example of contouring variation (green: consistent, red: deviation) between three expert radiation oncologists for the brainstem contour, overlaid on the T1c image data, used to determine the boundaries.

in greater interpretive variability [77]. For example, studies have reported significant variations in contouring cardiac structures, particularly the Left Anterior Descending Artery (LAD), due to its poor visualization on non-contrast CT scans [78]. Additionally, the *background and level of experience* of the individual performing the contouring significantly influence variability [79]. Differences are observed not only among various healthcare professionals, such as radiation oncologists, dosimetrists, or radiologists, but also among clinicians with differing levels of training [80].

The *complexity of target geometry* significantly increases contouring challenges, particularly when tumours exhibit intensity levels similar to surrounding tissues [81]. Consistent difficulties arise in areas such as the longitudinal extent of head and neck cancers, tumours adjacent to consolidated lung tissue or potentially invading the mediastinum [82], regions near biliary stents or suspicious lymph nodes [75], and structures affected by motion, such as thoracic regions influenced by respiration [74]. The impact of this variability is substantial. As treatment techniques like IMRT become more conformal, variations in manual contouring can lead to critical errors, resulting in geographic miss of TV or excessive irradiation of OARs [83, 84]. Moreover, contouring inconsistencies hinder the standardization of patient treatments across different centres and clinicians [85].

Current Research Directions: To address contouring variability, the field has increasingly adopted automated contouring solutions, which offer notable advantages. A key strategy involves *developing and applying standardized anatomical guidelines* and contouring atlases [63, 86]. These resources provide clear delineation protocols to improve consistency among radiation oncologists. Furthermore, Quality Assurance (QA) sessions and specialized training have been identified as essential for reducing inter-observer variability and standardizing TV delineation across clinical settings [87]. Despite these efforts, guideline-based approaches alone have not fully resolved variability [88], prompting increased focus on automated contouring solutions.

Auto-contouring technologies have developed along two main pathways: traditional *atlas-based methods* and modern *AI approaches*, particularly Deep Learning (DL) [89]. Atlas-based auto-segmentation has demonstrated success in reducing

workload, but it often necessitates substantial manual adjustments to meet clinical standards [90]. Research indicates that less experienced physicians experience greater reductions in contouring time compared to their more experienced counterparts [91].

AI approaches have demonstrated significant time efficiency, reducing total contouring workflow time by up to 80 minutes (65%) in head and neck anatomy, with smaller time savings in other anatomical regions [92]. Multiple studies have shown that computer-assisted contouring methods substantially decrease inter-observer variability [93]. This enhanced consistency is particularly beneficial for standardizing treatments across various centres and clinicians, thereby improving the reliability of clinical trial data [94]. As these technologies advance, integrated strategies combining standardized guidelines, QA processes, and advanced auto-contouring tools offer the most promising approach to tackling the ongoing challenge of contouring variability in RT planning [95, 96].

Limitations of Current Approaches: Despite these advancements, most auto-contouring systems still require manual review and editing [97]. They are increasingly regarded as valuable tools that provide a starting point to reduce workload and enhance consistency, rather than fully replacing clinical expertise [92]. Evaluating the performance of these automated systems is a critical area of ongoing development. Common evaluation methods include geometric metrics such as the Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), and other measures of overlap and distance [98–100]. These metrics assess the accuracy of auto-generated contours against manually defined gold standards. However, research suggests that geometric metrics often do not strongly correlate with dosimetric outcomes, which are closer to estimating clinical impact [101].

1.2.2 Time-consuming Plan Optimization

The manual method used for treatment planning has several important limitations, as shown in Figure 1.4. One of the main issues is that it is *highly time-intensive*, often requiring planners to spend many hours or even several days on a single patient case [102, 103]. This challenge is even greater in cases involving complex areas of the body, such as head and neck cancer. In these situations, planners must consider multiple target dose levels and many nearby OARs, which makes the process especially difficult and time-consuming [104, 105].

The quality of a treatment plan depends greatly on the *expertise and experience of the planner* [106, 107]. Creating an effective plan requires a deep understanding of what the Treatment Planning System (TPS) can and cannot do, along with the ability to estimate what kind of dose distribution is realistic for each patient [108, 109]. Because this process relies so much on human judgement, there can be a lot of variation in the quality of plans, even among planners working in the same clinic [110, 111]. Treatment planning also involves *multiple stakeholders*, such as radiation oncologists, medical physicists, and dosimetrists [66]. Coordinating between these professionals adds complexity to the workflow. For example, treatment plans often go through several rounds of review and revision based on feedback from oncologists, which can take extra time [112]. This teamwork, although essential, can lead to delays, especially in urgent cases where time is limited [113].

For complex treatment sites, such as head and neck cancer, the planning process is particularly challenging due to the non-convex geometry of TVs and their close proximity to critical structures [104]. Advanced techniques, such as IMRT,

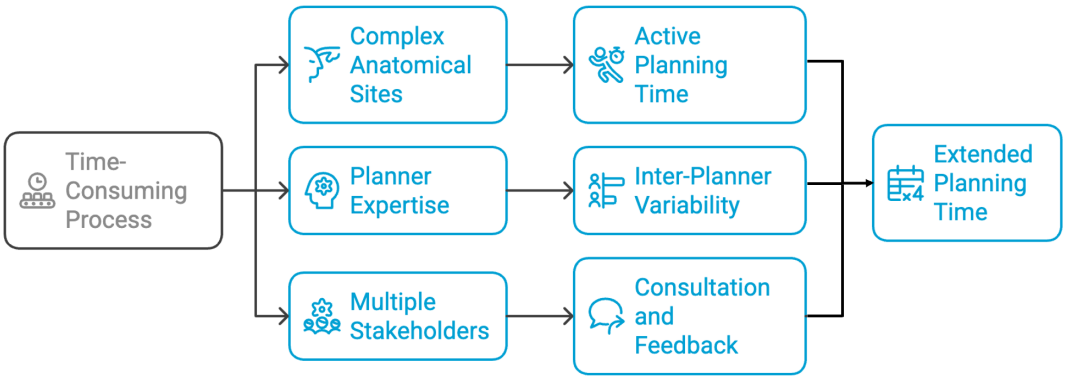


FIGURE 1.4: Why is the current workflow of treatment planning time consuming? Here are three common factors that lead to excessive delays in generating acceptable treatment plans.

enable highly conformal dose distributions but require significantly more complex and time-intensive planning [114–116]. The manual and iterative nature of treatment planning further reduces clinical efficiency and may affect treatment outcomes. Due to time constraints in clinical practice, compromises are sometimes necessary, which can lead to suboptimal plans being used in treatment [66, 117]. This underscores the conflict between achieving high-quality plans and maintaining efficiency in demanding clinical settings [118].

Planning Time Requirements: The time demands of RT treatment planning for various treatment techniques and anatomical sites have been the focus of many past research projects. For example, active planning time for ten cases of glioblastoma (GBM) and pancreatic cancer using standard IMRT was reported to average around 135 minutes per case [119], highlighting the need for automation. In brachytherapy, especially for cervical cancer, the contouring and planning phases are particularly time-consuming, often taking about 3 hours on average [120]. The planning process for techniques such as VMAT is also resource-intensive. A single optimization cycle can take up to 10 minutes using modern computing systems, and multiple iterations are usually needed to reach a clinically acceptable plan [121, 122]. For Naso-Pharyngeal Cancer (NPC) cases, even experienced planners using conventional algorithms may require several hours to complete a single plan [123]. In general, complex treatment sites often demand several hours of work per plan [124]. This extended planning time poses a significant barrier to the wider adoption of adaptive RT, where plans must be updated frequently based on changes in patient anatomy [125, 126].

Causes for Long-computation Time: The *complexity of RT treatment planning* is driven by several interconnected factors that can extend the time from initial simulation to the first treatment session to several days or even weeks [127]. Modern treatment techniques such as IMRT and VMAT are now widely used because they offer better target coverage and improved protection of healthy tissues. However, these methods also make the planning process much more complex and time-consuming [128]. These advanced techniques produce highly detailed dose distributions, with each small segment of the radiation beam (beamlet) having a different intensity. As a result, the planning process requires significant computational power and close attention from the planner [129]. Moreover, when IMRT plans become too complex,

they can lead to increased uncertainty in dose delivery, longer treatment times, and greater sensitivity to any changes in the patient's anatomy or position [130].

The *iterative trial-and-error approach* commonly used in treatment planning becomes especially difficult when dealing with complex anatomical sites. Planners often need to create and evaluate several versions of a plan before reaching one that meets clinical standards. This process can take a significant amount of time, even for experienced medical physicists [131, 132]. The workload is further increased by the growing need for detailed documentation, the introduction of new technologies, a wider range of available treatment options, and stricter regulatory QA requirements. Each of these factors adds additional steps to the planning process, increasing both the time and effort needed for each individual case [133].

The quality of treatment plans depends strongly on the *skill, experience, and available time* of the planner during the optimization process [134]. Because of this reliance on individual expertise, there can be significant variation in both the quality and efficiency of treatment plans, which may affect patient outcomes. As the demand for RT services continues to grow, many clinics face increasing pressure on their planning resources. This often leads to a conflict between the goal of delivering high-quality, personalized treatment plans and the practical limitations of a busy clinical workflow [135].

These planning challenges have important consequences in certain clinical situations. In palliative care, where the main goal is to relieve symptoms quickly, delays in treatment planning can be particularly harmful. Standard simulation and immobilization procedures may also be difficult to apply in patients who are medically fragile or have complex health conditions [136]. Furthermore, the time-consuming nature of current planning methods creates major bottlenecks that limit the progress of personalized and adaptive RT, especially techniques such as online treatment adaptation that require rapid plan generation [137].

The continued reliance on manual inputs and standardized protocols in the RT workflow, while generally effective, is increasingly challenged by patient-specific variability. As treatment cases become more complex and clinical demands grow, these traditional approaches may no longer provide optimal efficiency or consistency [138]. In response to these limitations, there has been growing interest in automated planning techniques. These methods are especially promising for complex treatments that involve advanced modalities and intricate target geometries, where manual planning can be particularly time-consuming and variable [139].

Impact on Care: The significant time required for RT planning presents major challenges in how clinical departments allocate resources, directly affecting patient care in several important ways.

Delays in planning can result in *longer intervals between simulation and the start of treatment*, which have been associated with worse survival outcomes [140, 141]. These delays also contribute to increased workload pressures, especially as cancer incidence rises, making manual scheduling more complex and time-consuming [142]. This issue is particularly critical in palliative cases, where rapid symptom relief is essential and patients may not tolerate standard simulation or immobilization techniques due to their medical condition [136]. In many cases, limited planning resources force a trade-off between delivering high-quality, *individualized treatments* and managing the day-to-day demands of a busy clinic [135]. When faced with high patient volumes and time constraints, clinicians may have no choice but to proceed

with suboptimal plans, which can negatively affect treatment quality. This is especially concerning given the well-documented link between adherence to RT protocols and improved treatment outcomes [143].

Efforts to Introduce Automation: The labour-intensive nature of current treatment planning processes often require several hours or even days per case, and has led to growing interest in automation and AI as potential solutions to reduce planning time [144]. These technologies offer the opportunity to streamline workflows and improve clinical efficiency without compromising plan quality. For instance, automatic VMAT planning could help overcome a major bottleneck in clinical practice. Recent advancements in planning technology have shown promising results, with some systems capable of generating high-quality prostate treatment plans in just 7–15 minutes. This level of speed could make real-time adaptive RT which was once considered impractical to be feasible in routine clinical workflows [125].
Personalized planning engines have achieved impressive reductions in overall planning time, including human inputs, optimization processes, and calculation times. Personalized approaches have reduced planning times to 60-80 minutes, approximately one-third of the time needed for manual planning [145].
Automated treatment planning pipelines are being developed to produce plans of comparable quality to those generated manually, but without requiring the several hours of labour per plan typically needed for conventional approaches [124]. These systems aim to improve efficiency while maintaining or enhancing plan quality consistency [146].
Online adaptive planning methods leverage automation to enable real-time plan adaptation while patients are on the treatment couch. The dramatic reduction in planning time achieved through automated approaches opens new possibilities for implementing adaptive RT strategies that were previously impractical due to time constraints [125, 137].
Knowledge-based planning models have demonstrated remarkable efficiency improvements, with some studies reporting treatment planning time reductions of up to 95% [147]. These approaches rely on data libraries built from prior treatments and planner-independent optimization algorithms to streamline the planning process, reducing reliance on manual trial-and-error iterations [123].
AI-based planning agents represent a future direction where planning can be efficient and effective with minimal human intervention [144]. These approaches have particular potential for time-sensitive scenarios such as adaptive RT that requires frequent and rapid planning [113].

1.3 Unmet Clinical Needs

The challenges listed in the previous section demonstrate the need for contour QA to be performed automatically and reliably. The widespread recognition of contouring variability has established *QA and peer review* as integral components of radiation oncology practice. Despite the existence of specific guidelines for OAR delineation in clinical practice and research, considerable inter- and intra-observer variations have been documented. These variations can lead to inconsistent dose evaluations, potentially compromising treatment efficacy and complicating toxicity analysis [148]. A study with four radiation oncologists and three radiologists delineating parotid glands showed that inter-observer variation in contour delineation is significant enough that nearly half of reviewed contours would have resulted in different

treatment plans if used clinically [149]. As auto-contouring becomes increasingly prevalent in clinical settings, QA remains crucial. Even as automated tools replace manual contouring in routine practice, quality checks still rely on clinicians [76]. AI-generated contours typically require review and adjustment by radiation oncologists before treatment planning, emphasizing the complementary relationship between automated systems and human expertise [150]. Automated approaches to contour review may decrease review time and improve consistency, making them valuable tools either stand alone or as assistants to human reviewers in identifying poor-quality contours [151]. This section elaborates on this unmet clinical need, motivating the research aims in this thesis.

1.3.1 Current Approaches to Contour QA

QA in general encompasses all procedures that ensure consistent and safe delivery of prescribed radiation doses to TVs while minimizing exposure to normal tissues and monitoring patient outcomes. This comprehensive process requires cooperation across all staff groups as the quality activities are interdependent throughout the RT workflow [152]. Contour assessment can be conducted both *visually and quantitatively*. While visual inspection remains the most common approach, it is time-consuming and subjective [153]. QA in RT contouring encompasses a range of methods and approaches, like shown in Figure. 1.5.

FIGURE 1.5: Current methods of contour QA spans the complexity spectrum from manual peer reviews and guideline-based adherence all the way up to machine learning methods. Figure generated using napkin.ai

Peer Review: is a formal review by another expert of the delineated contours used to produce a RT plan, has become a critical mechanism for quality improvement [154]. This process of traditional "chart rounds" involving contour review, radiation dose prescription scrutiny, and treatment plan evaluation has demonstrated significant clinical impact, with studies showing that peer review leads to changes in approximately one in nine radiation plans [155]. Weekly teleconferences with radiation oncology specialists for detailed target and OAR review prior to treatment plan creation, with changes classified as "major" (modifications to high-dose PTV or prescription) or "minor" (modifications to intermediate/low-dose PTV or OARs) [156]. From a survey among 115 radiation oncologists in the US, 44% reported performing a contour-specific peer review in their practice. Furthermore, 72% of these

respondents reported that contouring-related questions arise in at least half of cases in routine patient care [157]. Prospective contour and plan review with radiation oncologists and site specialists, along with secondary QA contour checks performed by trained dosimetrists or medical physicists during treatment has been reported [158]. Weekly contouring QA meetings where radiation oncologist contouring is peer reviewed prior to dosimetric planning, have served as both quality control and educational forums for trainees [159].

Standardization Approaches: Consensus contouring guidelines and atlases have been used to reduce TV delineation variability [160]. A survey indicated that 75% of radiation oncologists first consulted cooperative group guidelines and contouring atlases (e.g., RadioTherapy Oncology Group (RTOG)/NRG Oncology (NRG)) when contouring questions arose [157]. Institutional, national, or international protocols have also been shown to reduce intra- and inter-observer variation in TV delineation [159].

Quantitative Assessment Methods: Conformity Index (CI) used to mathematically quantify contour accuracy against gold standards, including: DSC, Jaccard Conformity Index (JCI), van't Riet Index (VRI), Geographical Miss Index (GMI), Discordance Index (DI), HD [153, 161], Protocol deviation scoring systems, commonly used in clinical trials, with standardized criteria: Score "1": OAR contours acceptable with no edits, Score "2": OAR contours acceptable with minor edits not likely to affect treatment plan, Score "3": OAR contours unacceptable, requiring major edits likely to affect treatment plan [94].

The quality of contour delineation has profound implications for treatment outcomes in RT. Inaccuracies in TV and OAR contour delineation directly impact both tumour control and normal tissue toxicities [162, 163]. The clinical significance of this impact has been quantitatively demonstrated in multiple studies. For instance, research examining parotid gland delineation found that the Dose-Volume Histogram (DVH)s of 46% of study contours were sufficiently different from those used clinically that they would have produced different intensity-modulated RT (IMRT) plans [148, 149].

1.3.2 Motivating Dosimetric Contour QA

The *dosimetric impact of contouring errors* is particularly significant with advanced RT techniques. The highly conformal nature of modern treatment modalities increases the dosimetric impact of delineation errors, making contour QA increasingly important [163]. This is especially true in anatomical regions like the head and neck, where the therapeutic window is narrow due to the proximity of PTV to OAR [162, 164]. Despite the critical importance of contouring accuracy, research has found that while 99.1% of studies presenting auto-contouring models report geometric agreement metrics, **only 23.1%** report the resulting dosimetric impact, which is essential for understanding the clinical consequences of contouring errors [165]. This disconnect between geometric evaluation and clinical impact assessment highlights a significant gap in this research space, leading to the motivations for this work.

Automated QUality Assurance (AQUA) systems have emerged as critical tools for RT, offering methods to systematically detect contouring errors that may impact treatment outcomes. These systems employ various approaches to evaluate contour quality without requiring extensive manual review. Machine learning-based error detection methods extract geometric features of OARs and apply classification algorithms to identify contours that deviate from expected parameters [166]. Image feature-based conditional random forest algorithms approaches have been

developed for different anatomical sites, with researchers using image features for thoracic structures, principal component analysis for pelvic structures, and surface-based metrics for female pelvis contours [167] tolerances of 1-3 mm for detecting contouring errors with accuracy higher than 0.9 for most targets and critical structures have been reported to be used for automating these checks [168].

DL techniques are increasingly being employed for automatically assessing contours, with recent work exploring deep active learning for OAR [148] and an intelligent DSC threshold based quality control system [169]. A secondary segmentation algorithm to help classify errors in the primary model has been developed with promising classification accuracy and recall [170]. Studies implementing DL-based contour models for OARs have shown significant time savings while maintaining or improving contour quality. In one study, total contouring time was reduced by 76% with radiation oncologist revision time specifically reduced by 35% compared to traditional workflows [171]. Novel approaches are being developed to not only detect but also correct contouring errors. An automated QA and adaptive optimization correction strategy has been proposed that can identify incorrect auto-contours, provide potential error reasons and locations through attention heat maps, and use vision-language representations with convex optimization algorithms to adaptively correct problematic contours [172].

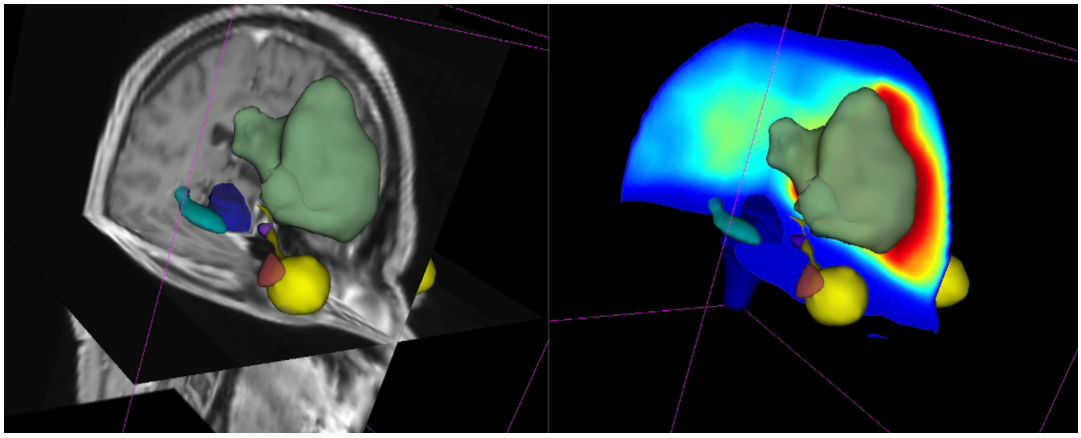


FIGURE 1.6: Motivating dosimetric QA: on the left: 3D slice plane and mask visualization for OARs and TV (in green) overlaid on T1c image data. On the right, the same overlaid on the dose distribution: hotter regions indicate higher dose, where accuracy of contours is arguably more important.

In summary, there is an unmet clinical need for developing dosimetric tools for contour QA to streamline the workflow and provide fast and reliable feedback to clinicians. However, external validation of these QA approaches is essential before deployment to clinics with different patient populations and clinical practices [173]. Automated approaches to contour review can decrease review time and improve consistency, making them valuable tools either on their own or as assistants to human reviewers in identifying poor-quality contours [151, 167, 168, 174–176]. The increasing adoption of AI tools in clinical practice has shifted the workflow paradigm. While auto-contouring is gradually replacing manual contouring in routine practice, quality checks still rely on clinicians [76], and reviewing them in a dose-aware manner is gaining importance.

1.4 Hypothesis, Contributions, and Structure

This work progresses from understanding human limitations in dosimetric interpretation, validating the reliability of dose distribution predictions under realistic variability, investigating the performance of underlying DL architectures across task complexities, and finally, to building proofs-of-concept form a comprehensive approach to achieving faster, clinically meaningful, and scalable QA.

1.4.1 Hypothesis and Aims

This leads to the central hypothesis of this work, that:

> *AI-based systems can standardize and automate fast and reliable dosimetric contour QA in RT planning.*

They are organized into research questions in three pillars:
A: Clinical Understanding and Needs Validation: Is there a need to standardize and automate dosimetric contour QA? with the sub-questions:
A1: What is the variability amongst radiation oncology professionals while performing contour quality assessments?
A2: How can dosimetric criteria be systematically formulated to develop an automated systems to replicate their behaviour?

If the answers to the group of questions in **A** indicates a clinical need, then, we investigate the following:
B: Technical Investigations and Analysis: How reliable and fast are AI-based systems for dosimetric contour QA? involving the following sub-questions:
B1: Do the predicted doses correlate better with the true dose differences compared to geometric metrics?
B2: How reliable are the core DL architectures under difficult conditions?

These explorations then lead to the following two proofs of concept, among many others listed in Chapter 13.
C: Proofs of Concept Studies: What clinical systems can such dosimetric contour QA be integrated into?
C1: Can such models assist in focusing contour review efforts on locations where segmentation variations are dosimetrically most critical?
C2: Can such models assist in dosimetrically ranking various auto-contour proposals?

This hypothesis and related questions are addressed in the following structure.

1.4.2 Research Contributions

The subsequent chapters begin by quantifying the gap between expert perception and dosimetric reality, establishing the need for this automation. It then leverages dose prediction models to fill the void of models that reliably automate dosimetric QA. From there, it explores how sensitive these models are to real-world input variations, providing evidence for their integration into planning workflows. Finally, the robustness of the segmentation architectures themselves is critically examined to ensure consistent performance under distribution shifts. This three-part structure

allows the thesis to systematically build and validate a case for practical, AI-assisted QA in RT. The research contributions related to this is included in Section C.

Clinical Understanding and Needs Validation: This pillar advocates for shifting QA from purely geometric to clinically meaningful dosimetric evaluations. The field has long relied on geometric metrics (e.g., DSC) for segmentation evaluation, yet these do not always correlate with clinical outcomes [101]. By embedding dose-awareness into contouring workflows, these works significantly improve the efficiency and safety of RT planning. The research contributions in this part include:

- Chapter 3 addresses **A1** and measures the *inter-expert variability in estimating the dosimetric impact* of TV contour changes on OAR toxicity through structured interviews across seven clinical experts. This work will appear in Radiotherapy and Oncology, mid-2025 (see J2 in C).
- Chapters 4 and 5 address **A2** and experimentally demonstrates that DL-based *dose predictors surpass human experts* in detecting dosimetrically sub-optimal contours, marking the first step towards automated dose-aware QA. Chapter 4 was presented at International Symposium for Biomedical Imaging (ISBI), 2023 (see C1 in C) and Chapter 5 was accepted as an oral talk (18% selection rate) at Medical Imaging with Deep Learning (MIDL) 2024 (see C4 in C).

Technical Investigations and Analysis: After validating the clinical need, this pillar analyses the behaviour of AI-based models to address the unmet needs discovered in the previous part. It shows that dose prediction models, while highly effective [177], require careful validation for robustness and sensitivity. Recognizing that U-Net based architectures are ubiquitous [178] in both auto-contouring and dose prediction model infrastructure, it shows a nuanced understanding of design choices relative to the complexity of data, guiding future architecture development with robustness in mind. The contributions include:

- Chapter 6 extends the work in Chapter 4 and addresses **B1**, providing an extensive *evaluation of sensitivity to OAR contour variations and robustness* against out-of-distribution cases, proposing data augmentation strategies to mitigate sensitivity gaps. They demonstrate that *model predictions align closely with inter-expert variability* in dose distribution, establishing strong evidence of model sensitivity. This work was published in Cancers, 2023 (see J1 in C).
- Chapters 7 extends the analysis in Chapter 8 to address **B2**, and critically examine *skip-connections across task complexities*, finding that their benefits are non-linear and context-dependent. Importantly, it shows that severing skip-connections in low-complexity tasks improves robustness, especially in out-of-domain data. Chapter 9 further addresses **B2** and analyses how *patch size affects segmentation performance and robustness*, revealing that while larger context benefits performance, attention-based and transformer-based models (UNETR, AGU-Net) are more sensitive to foreground ratio shifts than vanilla U-Nets. Chapter 7 is under review (see C). Chapter 8 was early-accepted as a conference paper at Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2023 (see C3 in C). Chapter 9 was accepted as an abstract at the Medical Imaging meets Conference on Neural Information Processing Systems (NeurIPS) workshop in 2022 (see W1 in C).

Proofs of Concept Experiments: Based on better clinical understanding and technical investigations, the final pillar of work groups proofs of concept experiments on novel tooling for dosimetric contour QA, for both the TV and OARs:

- Chapter 10 addresses **C1** and introduces *Atomic Surface Transformations for Radiotherapy quality Assurance (ASTRA), a method to visualize sensitivity maps* showing where local contour variations on OARs impact dose predictions, enabling dose-aware contour corrections. This work won the 2nd best student paper award at Conference on Engineering in Medicine and Biology (EMBC) 2023 (see C2 in C).
- Chapter 11 addresses **C2** and develops a *dosimetry-informed ranking framework called AutoDoseRank* that sorts TV contour candidates based on their clinical impact, outperforming 3 of 4 expert oncologists in triage tasks. This work was presented at the CaPTion workshop at MICCAI 2024 (see W2 in C).

1.4.3 Thesis Structure

The remainder of this thesis is structured as follows: Chapter 2 provides a background in AI and the RT workflow, which is followed by the research contributions that address the hypothesis and questions raised in this chapter. Chapter 12 revisits these questions and the hypothesis through a critical lens to include limitations of these studies. Finally, Chapter 13 provides some perspectives on future research in this domain. The topics are organized along validation of clinical needs (Part I), technical investigations and analysis (Part II) and proofs of concept experiments (Part III).

Each of these is also plotted along the TRL scale, as defined in the National Aeronautics and Space Administration (NASA) guideline, through a research to innovation pipeline, including a proposed structure to read the subsequent chapters, shown in Figure 1.7.

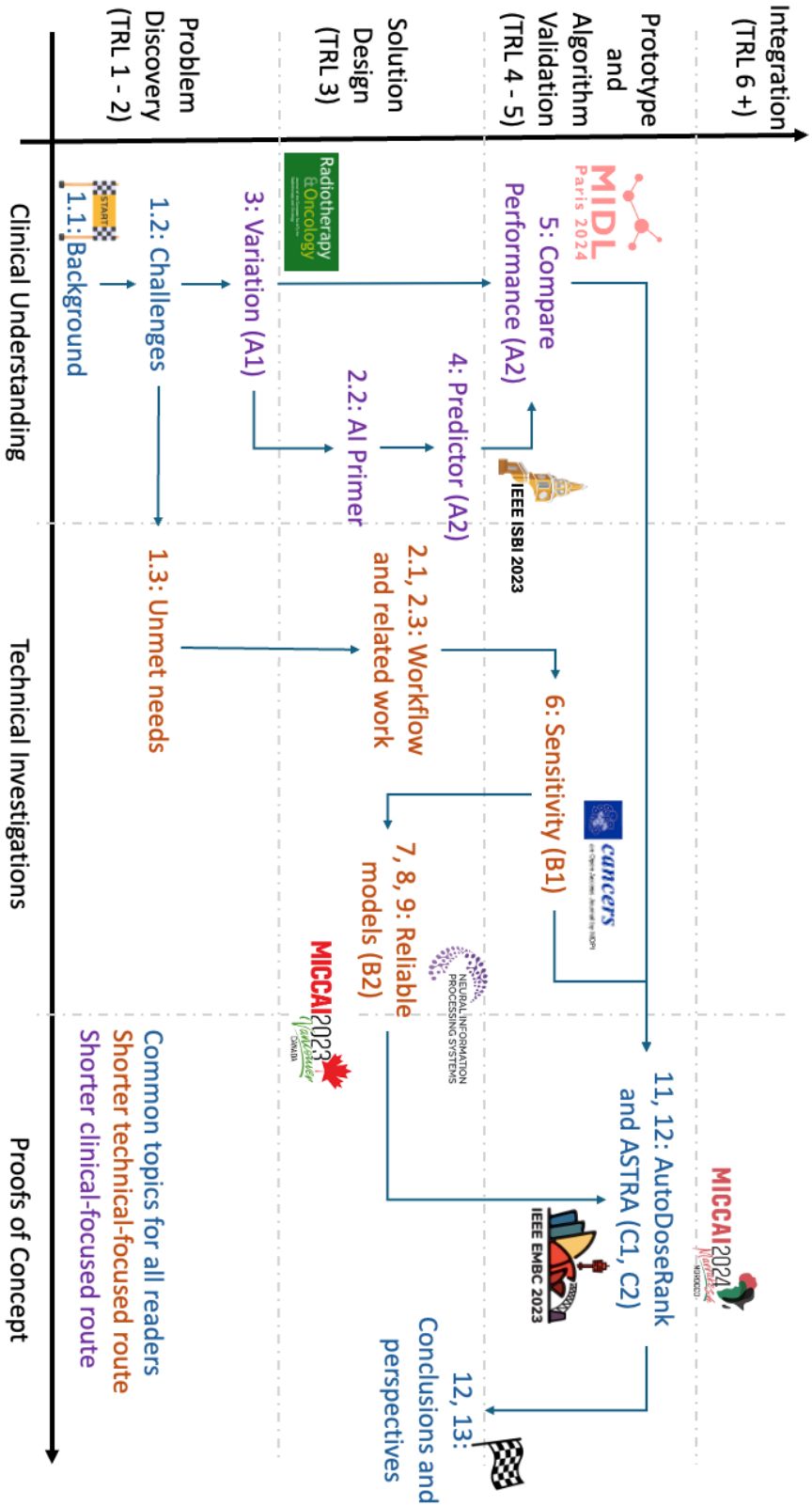


FIGURE 1.7: A proposed reading guide, organized into clinical, technical and proofs of concept buckets on the vertical axis, and the expected technology readiness level on the horizontal axis. The research to innovation pipeline aims to move the proofs of concept higher along the Technology Readiness Level (TRL) axis.

2

Background and Related Work

"In every surgical operation the position and size of each local "vital point" must first be considered, and the incision should be made so as not to affect that particular "vital point" even a slight injury near a "vital point" may prove fatal. (Śārīra-sthāna 6/86-87)"

- Sushruta, in Sushruta Saṃhitā ~600 BCE (1907 translation by Bhishagratna).

This chapter serves as the background by describing the RT workflow in Section 2.1, outlining the main steps involved and the key challenges for contour QA. Section 2.2 then introduces AI, and provides a brief overview of the development of models that are important for contouring and treatment planning tasks. Following this, Section 2.3 reviews related work to give further context for the research directions explored in this thesis. Finally, Section 2.4 presents the evaluation metrics used to assess the models and methods discussed in the following chapters.

2.1 The Radiotherapy Workflow

The RT workflow, shown in Figure 2.1, includes seven key stages designed to ensure safe and effective treatment. After initial consultation and consent, the process begins with simulation, where a planning CT volume is acquired using immobilization devices to maintain consistent positioning [136]. While CT is the standard for planning, additional imaging such as MRI or PET may be registered with the planning CT to improve target definition [179]. Next is contouring, where radiation oncologists outline the TV and OAR [180]. In the planning phase, dosimetrists or physicists define beam parameters such as energy, intensity, and angle, using Multi Criteria Optimization (MCO) to maximize tumour coverage while sparing healthy tissue [181]. The plan then undergoes QA to confirm its accuracy and adaptability to anatomical changes [182]. Treatment is delivered in multiple sessions, with image guidance used to align the patient before each fraction [179]. The workflow ends with patient discharge and long-term follow-up to monitor outcomes and late effects [183].The entire process from simulation to first treatment can take several days to weeks [136] depending on the complexity of treatment, pointing to the need for more efficient workflows. Each step in the process is elaborated upon next.

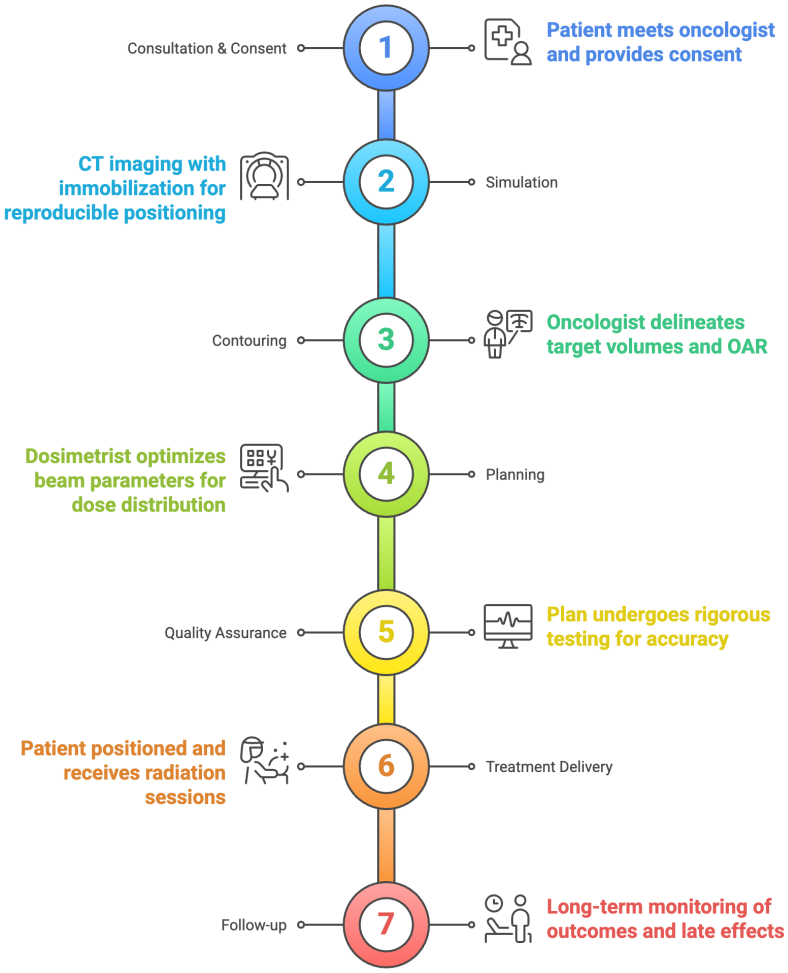


FIGURE 2.1: The seven stages of the RT workflow. Figure generated using napkin.ai

2.1.1 Steps in Current Workflow

Patient Consult: The RT workflow starts with a consultation, where the radiation oncologist reviews diagnostic scans, prescribes the dose and fractionation (how many sittings is the treatment delivered in; typically 30 fractions of 2 Gy each for GBM, with a cumulative prescription of 60 Gy to the TV), and discusses treatment options with the patient [179, 184]. Following this, the patient is scheduled for a CT simulation, the first technical step in planning [185]. Prior to imaging, MRI-safety screening is performed if MRI is part of planning [186]. Depending on tumour location, immobilization devices may be prepared, such as head and neck masks made in the mould room [185]. Setup instructions and dose details are recorded in the medical record [186]. Modern workflows increasingly adopt shorter treatment regimens, such as hypo- and single-fractionation, to enhance comfort and reduce visits [187].

Imaging and Simulation: The imaging and simulation stage forms the basis of RT planning, with the so-called planning CT as the primary modality for capturing anatomical detail and setup localization [179]. CT provides tissue density in Hounsfield Units (HU), essential for dose calculation. Additional imaging, such as MRI for soft tissue contrast and PET for metabolic data, enhances target delineation [188]. In advanced workflows like MRI-guided RT, both CT and MRI are acquired

and co-registered, with alignment verified by clinical staff [189]. Simulation is managed through electronic orders linking treatment systems, imaging devices, and the TPS [190]. The resulting images are used for contouring TV and OAR, establishing patient setup, and calculating radiation dose [182].

Contouring and Target Definitions: After imaging and simulation, the contouring phase is one of the most important steps in the RT workflow. In this step, radiation oncologists outline the target volumes, such as the GTV, and identify nearby OARs to spare during treatment [191]. Contouring is usually done by drawing 2D outlines on each relevant image slice. These 2D outlines together form the 3D volumes needed for treatment planning [191]. Contouring guidelines are used to help outline the TV and OARs correctly. These guidelines improve treatment accuracy and reduce the risk of toxicity [192]. Different guidelines exist for different parts of the body, which help keep results consistent across clinics. For example, the European Society for RadioTherapy and Oncology (ESTRO)-EANO guideline gives clear advice for contouring in cases of GBM, and recommends using imaging such as MRI to better define the tumour and nearby OARs [63]. These guidelines help radiation oncologists create high-quality treatment plans across different hospitals [193].

Modern RT often uses automatic contouring tools to save time and reduce variation. Systems like Radiation Planning Assistant (RPA) can draw outlines for normal tissues and some target volumes. Radiation oncologists then check and edit these outlines before drawing the GTV themselves [194, 195]. This method is useful for precise planning, where the automatic contours may still need some editing [196].

Treatment Planning: After contouring, the outlined TV and OAR are used in the next step called treatment planning. The CTV is drawn by expanding the GTV to account for microscopic extensions. The PTV is then created by expanding the CTV, while avoiding overlaps with OARs. Dose prescriptions and goals for OAR are also set at this stage [197]. The dose is measured in Gray (equal to 100 rad), which measures one joule of radiation energy absorbed per kilogram of tissue. This contour data is sent from the contouring software to the TPS, where the resulting images are used to calculate radiation dose [198, 199]. This step is critical because it converts anatomical information into a plan that can be used for delivering radiation. Planning starts with treatment setup, which includes defining the isocenter, choosing the beam arrangement, and setting the dose prescription [113]. In conventional external beam therapy, beams are usually placed to intersect at the isocenter, which is often the centre of the tumour. Parameters such as beam weights, wedge filters, and Multi-Leaf Collimator (MLC) positions are also defined at this stage [200].

The planning method depends on the complexity of the treatment. For simpler plans, such as three-dimensional conformal RT, forward planning is used. In this approach, clinicians can manually adjust beam settings to achieve an acceptable dose distribution [200]. For more complex techniques like IMRT or VMAT, inverse planning is required, where the specialized TPS calculates beam settings based on dose constraints provided by the physician [106]. They create custom treatment plans that deliver the prescribed dose to the outlined target volumes [201]. In many centres, staff rotate planning responsibilities. Most plans are completed within two to four days [199].

IMRT uses computer-controlled linear accelerators to adjust beam intensity across many small segments, known as beamlets. This allows for precise dose delivery that fits the shape of the tumour [202]. Inverse planning software is used to define the desired dose, and the system then finds the best beam intensities to meet these goals

[203]. MLCs help shape the radiation beam to match the tumour's outline during treatment [204]. IMRT is especially useful for tumours close to sensitive structures, such as those in the head and neck, prostate, or brain. It helps reduce side effects and improve tumour control [205]. The precision of IMRT also makes it possible to increase the dose to the tumour while protecting nearby healthy tissue [206].

VMAT is another advanced technique, which delivers radiation while the machine rotates around the patient. The beam intensity and shape change continuously during the rotation [207]. This method provides a highly conformal dose distribution, and can improve tumour coverage and offer higher protection to nearby normal tissues [208]. Compared to IMRT, VMAT can reduce treatment time and the number of Monitor Unit (MU) used, which improves patient comfort and clinic workflow [209].

New planning methods also use knowledge-based tools. For example, RapidPlan™ [1] can predict the expected Dose-Volume Histogram before final adjustments are made [210]. The DVH is a chart that shows how radiation dose is spread within different volumes. Another tool is the isodose line map, which connects points that receive the same dose using coloured lines. Systems like the RPA can create complete treatment plans with very little manual input for certain types of cancers [194, 195]. After optimization, the TPS calculates the dose distribution, using models such as the Anisotropic Analytical Algorithm (AAA) or Acuros XB [210]. The grid size for dose calculation is important. A spacing of 2.5 mm is often used because it keeps the dose error below one percent [210, 211].

Once the plan is complete, its quality is checked against the clinical goals. If the plan does not meet the goals, parameters are adjusted and the optimization is repeated [113]. In automated systems, this cycle continues in the background until the plan meets the required standards [113]. The final result is a treatment plan that delivers the prescribed dose to the tumour while protecting OARs, becoming the basis for delivering RT. Treatment planning requires advanced skills and is often the most complex and time-consuming part of the RT workflow [146, 197].

Plan Review and QA: After treatment planning, all RT plans go through the critical QA phase. This step checks the safety, accuracy, and deliverability of the plan [201]. Key metrics include the CI, which shows how well the dose matches the target shape, and the Homogeneity Index (HI), which measures how evenly the dose is spread. Other metrics include the dose to 95% of the target, mean dose, and how much of an organ receives a set dose.

The first review is done by the medical physicist who created the plan, and a second check is done by another qualified physicist who was not involved in the original planning [212]. Some centres also use automated plan-checking software. After the physics review/s, a radiation oncologist reviews the plan to confirm it matches the treatment intent [199].

For complex techniques like IMRT and VMAT, extra checks are done to confirm the machine can deliver the planned dose accurately [213]. This is usually done on the same day as the first treatment. The plan is delivered to a phantom, and the measured dose is compared to the calculated dose. Many centres also review plans in group meetings where doctors and physicists give feedback [201]. This team-based review adds another layer of safety and ensures consistent treatment quality.

[1] by Varian, a Siemens Healthineers Company. From: https://www.varian.com/products/radiotherapy/treatment-planning/rapidplan-knowledge-based-planning.

Studies have shown that human review is still essential, even in automated systems [196, 214]. The QA process involves several professionals, each checking different parts of the plan [199]. Clear communication between team members is important to avoid delays or mistakes [215]. Rushed reviews to meet deadlines can increase errors, so enough time must be given for proper checks [215].

After all reviews and approvals, the plan is sent from the TPS to the electronic medical record [179, 201]. The plan is then ready for treatment delivery, marking the end of the planning and QA process [216].

Treatment Delivery: After plan approval and QA, the patient enters the treatment delivery phase. At each session, radiation therapists position the patient using immobilization devices from simulation [179]. Image-guided RT, often using volumetric imaging such as Cone-Beam Computed Tomography (CBCT), is performed to verify patient and target alignment with the simulation reference [184]. The workflow follows a strict sequence to ensure treatment matches the approved plan and maintains safety [216]. Once alignment is confirmed, radiation is delivered, typically with photon beams of prescribed energy, coordinated through a treatment management system that tracks each step [190].

Advanced techniques may be used in complex cases. In online adaptive RT, real-time adjustments are made based on updated imaging, often using MRI, with auto-contouring and dose recalculation based on current anatomy [217]. For mobile targets like in the lung, respiratory gating aligns delivery with specific breathing phases. After treatment, verification may be performed to confirm accurate delivery [184].

2.1.2 Challenges in Contour QA

The *manual contour QA process* is known to be resource- and time-intensive. It requires strong anatomical knowledge, significant time, and financial support. Manual QA is also subject to inter-observer variability [149]. Most errors occur during planning, making pre-treatment contour review essential [72]. Time constraints are a major challenge, as evidenced by a survey which found that 58% of respondents cited limited time, and 22% cited lack of access to disease site experts, as barriers to effective contour review [157]. These challenges are worse in adaptive RT, where reviewing auto-contours increases clinician workload and reduces patient comfort [218]. Automated systems that flag poor-quality contours for manual review can improve efficiency in such cases [218].

The *quality of RT* directly affects treatment outcomes, especially in clinical trials. Many trial centres do not fully review contours after initial checks [163]. Poor-quality RT can lead to lower survival and more treatment failures [219]. A major trial in head and neck cancer showed worse outcomes in patients with protocol violations. Two-year survival was 50% vs 70%, and local control was 54% vs 78%, for non-compliant vs compliant plans, respectively [140]. Studies also report contouring errors during pre-trial QA in lung cancer trials [220]. Benchmark assessments help identify protocol misinterpretations. These can be addressed by QA teams and trial coordinators [220–222].

The *integration of contour QA* into clinical practice remains uneven. It is more common in academic centres and institutions with more staff, reflecting resource gaps [157]. Automated contour QA requires checks for each case and for system-wide changes like imaging updates [223]. Risk analysis studies using failure mode

and effects analysis show that human error remains a key risk in automated workflows. This supports the continued need for manual review by clinicians and physicists [224].

Studies show that contour quality is *strongly linked to institutional experience*. Centres treating fewer than five patients had a 29.8% rate of major contour issues, compared to 5.4% in centres treating 20 or more patients ($p<0.001$) [140, 225]. This shows that experience and specialization improve contour quality. Poor contours can also affect research. Errors in contour data can distort machine learning models used to predict treatment outcomes [226]. Contour QA is therefore essential for both patient care and data reliability in treatment development.

2.2 An Artificial Intelligence (AI) Primer

Artificial Intelligence (AI) has a large impact on the field of radiation oncology by supporting the entire treatment workflow, from diagnosis to post-treatment care [227, 228]. Applications include image contouring, treatment planning, outcome prediction, QA, and adaptive re-planning [229]. AI tools often perform at the level of human experts but in much less time [230] and can help meet rising demand for RT and support more equal access to care [144].

The modern field of AI has its roots in the 1950s, though related ideas existed earlier. The 1956 Dartmouth Conference is seen as the start of AI, where researchers like McCarthy and Minsky discussed creating what they called "thinking machines" [231]. More modestly, machine learning aims developing algorithms that can learn patterns from data without being explicitly programmed [232]. Early progress was followed by setbacks known as "AI winters," when funding and results slowed [233]. Early systems were rule-based to model human knowledge but struggled with real-world problems [234, 235]. AI has advanced through several key shifts from rule-based systems in the 1970s–80s, evolving to statistical models in the 1990s [236]. The current wave began around 2012 with Deep Learning (DL) and major improvements in image recognition [237]. This progress is reflected in radiation oncology, which also moved from rule-based tools to statistical models, and now to DL-based systems for image analysis and treatment planning [238].

2.2.1 Learning From Data

The field of AI is typically divided into several learning paradigms, as shown in Figure. 2.2.

Supervised Learning uses labelled data to train models. Each input has a known output, and the model learns to match inputs to outputs by reducing the error between its predictions and the true labels [234]. This method is useful in radiation oncology for tasks like tumour classification, where past cases with confirmed diagnoses can be used for training. The success of supervised learning depends on both the quality and quantity of the labelled data. In medicine, creating these labels often needs expert input, which takes time and can be costly. In radiation oncology, inter-expert variability in tasks like contouring can lead to inconsistent training data [238]. Common tasks include *classification* (e.g., identifying cancerous tissue) and *regression* (e.g., predicting radiation dose) [240].

Unsupervised Learning works with unlabelled data. It finds patterns or groups in the data without knowing the correct outputs [241]. One main method is *clustering*, where similar data points are grouped based on distance or similarity. In

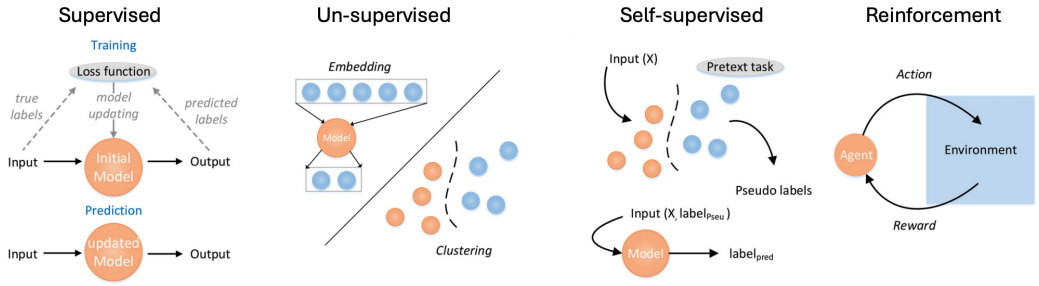


FIGURE 2.2: Categories of machine learning, adapted from [239]: Supervised learning: training process is shown above and prediction process is shown below. Unsupervised learning: the main two applications, embedding and clustering. Reinforcement learning: learn from environment and update from feedback. Self-supervised learning: a pretext task in self-supervised learning is a task designed to train a neural network to learn useful representations of input data without explicit supervision. The network is trained to solve the pretext task using the input data as the only source of supervision, and the learned representations can be transferred to downstream tasks where explicit supervision is available.

radiation oncology, clustering can help identify patient groups with similar treatment responses or uncover trends in treatment plans [242]. *Dimensionality reduction* techniques like Principal Components Analysis (PCA) and t-SNE [243] help simplify complex data and make it easier to understand. These are useful for visualizing the high-dimensional data common in radiation planning. Unsupervised methods can also detect outliers, which may indicate rare anatomy, equipment errors, or issues in treatment plans [244].

Semi- or Self-Supervised Learning (SSL) combine supervised and unsupervised methods by using a small amount of labelled data and a larger set of unlabelled data. This is helpful in medical imaging, where unlabelled scans are common but expert annotations are limited [245]. In self-supervised learning, the model creates its own learning tasks using unlabelled data. For example, it may learn to predict missing parts of an image using the full image as a reference [246]. SSL frameworks such as SimCLR [247], DIstillation with NO labels (DINO) [248], and masked autoencoders [249] are used to pre-train models, which can help with tasks like contouring. Other models like Mean Teacher [250] and FixMatch [251] use pseudo-labelling and consistency regularization, which are important when only a few labelled samples are available.

Reinforcement Learning teaches models to make decisions by using rewards. There are no labelled outputs. Instead, the model learns from trial and error by receiving rewards for good actions and penalties for bad ones.

Limitations of Traditional Machine Learning: While traditional machine learning algorithms have proven valuable in many applications, they have several limitations that DL addresses:

- Feature engineering dependency: Traditional algorithms rely heavily on manual feature engineering, which requires domain expertise and can miss complex patterns that aren't explicitly encoded [252].

- Difficulty with unstructured data: Images, text, and other unstructured data types are challenging for traditional algorithms without extensive preprocessing [253].
- Limited representation capacity: Many traditional algorithms struggle to capture complex, hierarchical patterns in data [254].
- Fixed model complexity: The complexity of traditional models is often fixed or limited by design, constraining their ability to scale with data size [255].
- Separate learning stages: Traditional pipelines often involve separate stages for feature extraction and model training, preventing end-to-end optimization [256].

DL addresses these limitations through its ability to automatically learn hierarchical representations from raw data, scale with data and computational resources, and enable end-to-end training [256]. However, traditional methods retain advantages in scenarios with limited data, when interpretability is crucial, or when computational resources are constrained [257].

2.2.2 Evolution of DL for Computer Vision

Deep Learning (DL) refers to a group of machine learning models inspired by how the human brain works [256]. These models use many layers of connected units, called artificial neurons, to process and learn from data [258]. The term "deep" comes from the presence of multiple hidden layers between the input and output [255].

Each artificial neuron receives inputs, multiplies them by weights, applies a non-linear function, and produces an output. When combined in layers and trained on large datasets, these neurons can detect patterns and make predictions. The layered structure allows the model to learn features at different levels of complexity, reducing the need for manual feature design [254].

DL has driven major progress in AI since the early 2010s [259]. This progress is due to larger datasets, better computing power (especially from Graphics Processing Unit (GPU)s), and improvements in training methods [258].

Theoretical Foundations: DL is effective because of two key capabilities: performing region-wise computations with non-linear transformations over large receptive fields, and using efficient gradient-descent training through back-propagation [260]. These strengths have allowed DL to outperform older models that relied on handcrafted features.

The Convolutional Neural Network (CNN) is the most widely used deep learning model for visual tasks. Developed in the late 1980s by Yann LeCun, CNNs are designed to learn spatial patterns in grid-based data such as images [261]. They use convolution filters that scan across the image, sharing weights to reduce the number of parameters and improve efficiency [262]. A typical CNN has several layers stacked together to extract spatial features. During training, the model adjusts its parameters to reduce the difference between its predictions and the ground truth [262], which allows CNNs to learn features directly from the data, removing the need for manual feature engineering [263]. Feature learning in DL is hierarchical, where early layers learn basic patterns like edges or colours, while deeper layers capture more specific and complex features [264]. This makes transfer learning possible, where models trained on large datasets can be adapted to new tasks with less data [261].

Modern DL models often use a modular design. A "backbone" network extracts general features, and a "head" network performs a specific task such as classification or segmentation [263]. This design supports the development of task-specific architectures like U-Net, which is widely used for biomedical image segmentation [265]. When trained on enough data, these models can generalize well and achieve strong performance on tasks like classification and segmentation [266].

Architectural Innovations: The development of DL architectures has transformed computer vision, evolving from general classifiers to specialized segmentation models (Figure 2.3). Before this shift, segmentation relied on edge- or region-based methods [267]. Early CNNs like AlexNet, VGG-16, GoogLeNet, and ResNet introduced key components later reused in segmentation networks [237, 268, 269]. A major breakthrough came with Fully Convolutional Network (FCN), which enabled end-to-end pixel-level predictions by removing fully connected layers [270]. This approach adapted classification networks for segmentation through fine-tuning and influenced many later methods [271].

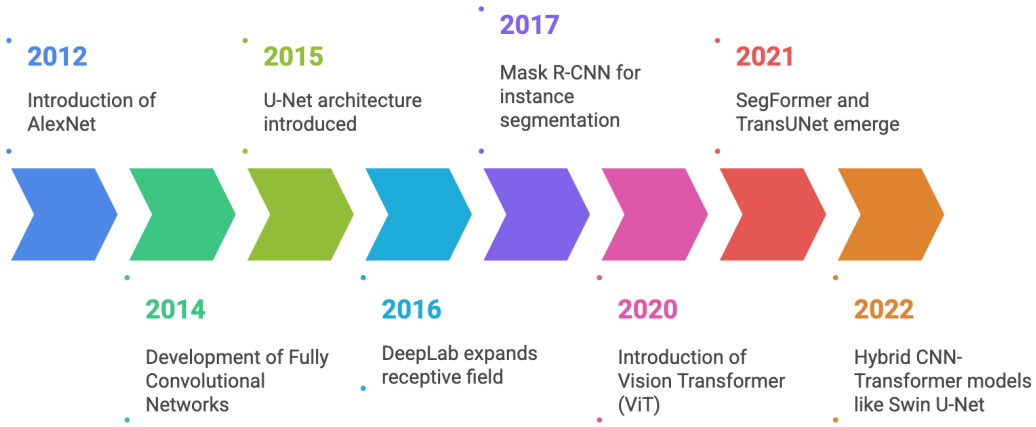


FIGURE 2.3: Evolution of DL architectures for image segmentation: from FCNs to transformer based models and multi-modal architectures. Figure generated using napkin.ai

Encoder-decoder architectures, such as U-Net and SegNet, became widely used for their ability to combine contextual understanding with precise localization [272]. U-Net, in particular, proved effective in medical imaging with limited data [265]. Dilated convolution networks like DeepLab expanded the receptive field without increasing parameters and introduced techniques such as Atrous Spatial Pyramid Pooling (ASPP) and Continuous Random Field (CRF)s for better boundary detection [273]. Other models, including deconvolution networks and large-kernel approaches, further improved segmentation detail and performance [274, 275]. Instance segmentation progressed through top-down methods like Mask R-CNN [276] and bottom-up approaches like Deep Watershed Transform [277].

Attention mechanisms improved global context modelling. Architectures such as Attention Gated U-Net (AGU-Net) enhanced U-Net by highlighting important regions, aiding detection of small or subtle structures [278]. Transformers, introduced through Vision Transformer (ViT) [279], were later adapted for segmentation with models like SegFormer, TransUNet, and U-Net + Transformer Hybrid (UNETR),

offering strong performance in complex medical tasks [280–282]. Hybrid CNN-Transformer models, including Swin U-Net and CoTr, combined local feature extraction with global attention, proving effective for 3D data [283, 284]. 3D and multimodal fusion architectures like nnU-Net and V-Net are now standard in volumetric segmentation tasks using MRI and CT, often incorporating multi-modal inputs for better accuracy [285, 286].

Modern segmentation models deliver high accuracy and efficiency, turning segmentation into a robust, end-to-end process with wide applications [287].

U-Net Architecture: U-Net, introduced by Ronneberger et al. in 2015, is a widely used architecture for semantic segmentation, especially in medical imaging [265]. It features a symmetric encoder-decoder design, with skip connections that link encoder and decoder layers to retain spatial details lost during downsampling. These connections also help with gradient flow and improve training stability [288]. Originally developed for biomedical images, U-Net showed strong performance with limited training data, aided by extensive data augmentation. It is efficient, segmenting high-resolution images in under a second on modern GPUs [265], and has become a standard in many domains [289].

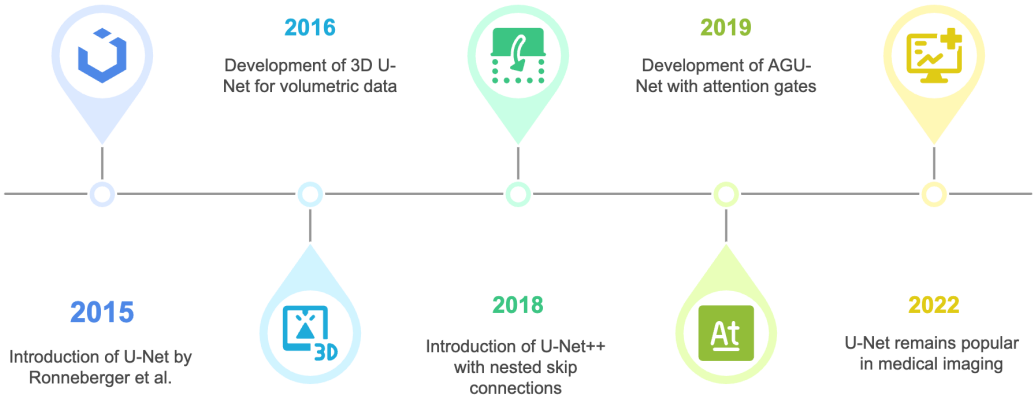


FIGURE 2.4: The U-Net architecture has evolved several variations based on backbone, skip-connection, and data-flow enhancements. Figure generated using napkin.ai

3D U-Net extends this approach to volumetric data by using 3D operations and online elastic deformations for augmentation [290]. Variants like U-Net with multiple skip-connections (UNet++) improve feature alignment between encoder and decoder paths using nested skip connections [288]. U-Net has also been adapted to use different encoder backbones, such as ResNet, to leverage transfer learning [291]. Lightweight versions reduce computation by using fewer layers and simpler upsampling, though often at the cost of detail around object boundaries [292].

The evolution of U-Nets across the past decade has been illustrated in Figure 2.4. Further developments include AGU-Net, which uses attention gates to focus on important regions [278], and V-Net, which uses a DSC-based loss to address class imbalance in 3D segmentation [286]. U-Net remains popular due to its balance of accuracy, speed, and adaptability. Its success with small datasets makes it ideal for medical imaging, and it continues to serve as the foundation for newer segmentation architectures [178].

2.3 Related Work

2.3.1 AI in Contouring

In recent years, AI, and particularly DL, has significantly advanced auto-contouring in radiation therapy (RT). These approaches have consistently outperformed conventional atlas-based methods [92]. A notable milestone was the 2017 American Association of Physicists in Medicine (AAPM) Grand Challenge, where DL models surpassed the previous gold standard model-based methods for thoracic anatomy segmentation on CT scans [293, 294]. Other studies have similarly shown that DL-based segmentation yields more accurate results than atlas-based techniques [92, 295]. While atlas-based methods rely on registering annotated CT datasets to new patient scans, DL models use CNNs trained to detect spatial patterns and variations of anatomical structures [295]. This shift marks a major advancement in auto-contouring technology [296].

The implementation of AI-based auto-contouring brings several benefits to clinical practice. It can reduce clinician workload, improve workflow efficiency, and enhance standardization across institutions and users [297]. These advantages are particularly relevant in time-sensitive procedures such as brachytherapy, where implanted needles or applicators increase the complexity of contouring [298]. Moreover, rapid and accurate contour generation supports the development of online adaptive RT [296]. The commercial landscape for auto-contouring has also expanded, with both established vendors and new companies offering AI-driven solutions [299]. This growth is driven by the ongoing need to address challenges in manual contouring, including inter-observer variability, time constraints, and the increasing complexity of modalities such as stereotactic radiosurgery (SRS) and stereotactic body radiotherapy (Stereotactic Body Radiation Therapy (SBRT)) [295].

The accuracy and clinical performance of AI-based auto-contouring tools have been widely evaluated. Studies often report results using quantitative metrics such as the Dice Similarity Coefficient (DSC), Mean Surface Distance (MSD), and Hausdorff Distance (HD) to compare AI-generated contours with those created manually, consistently showing that AI-based tools can match or even exceed expert-level performance.

Evolution of Auto-contouring Systems: The evolution of auto-contouring techniques in radiation therapy has followed several distinct phases, each aimed at overcoming the limitations of earlier methods. Initial efforts relied on simple techniques such as image value thresholding to automate segmentation tasks [92]. These early approaches were eventually replaced by more advanced atlas-based auto-segmentation (Atlas-Based Auto-Segmentation (ABAS)) methods, which became widely used in clinical practice over the past decade [300].

A major shift occurred with the introduction of AI, particularly Deep Learning (DL) methods. DL algorithms, based on CNNs, are trained to identify spatial features and anatomical variations in medical images [295]. This transition from atlas-based to DL-based auto-contouring has accelerated in recent years, supported by rapid advancements in algorithm development and increasing clinical adoption of AI-based segmentation tools [299].

Over the past several decades, various methodologies have been developed for auto-contouring in RT, including the following:
Atlas-Based Segmentation: This method involves registering previously contoured CT datasets to new patient scans and transferring contours accordingly. Although

atlas-based approaches represented a significant advancement over manual contouring, they are limited by their sensitivity to differences in image quality between reference and target CT scans, as well as by the accuracy of registration algorithms and the effectiveness of post-processing steps [295]. These systems typically require substantial computational resources and several minutes per case. Despite widespread adoption, optimal atlas selection remains a challenge, with evidence suggesting that even state-of-the-art selection algorithms perform suboptimally compared to ideal selection benchmarks [301].

Machine Learning Approaches: These approaches leverage algorithms capable of detecting patterns through learning processes, enabling more adaptable integration of prior knowledge for anatomical structure labelling [93, 302]. While the majority of prior research has focused on improving model architecture with fixed datasets (a model-centric approach), recent interest has shifted toward data-centric strategies, where the focus is on enhancing data quality while using fixed models [303, 304].

Deep Learning (DL): has become the most widely used framework for medical image segmentation. Architectures such as U-Net have gained popularity due to their ability to balance global context with precise spatial localization, while also performing well with limited training data [265]. Similarly, 3D U-Net architectures [290] have demonstrated strong generalization across international datasets for head and neck OAR segmentation, performing comparably to expert clinicians [305]. More recently, ViT-based models [282] have been competing against traditional CNNs [285], with various research articles comparing the two [306], while Segment Anything Model (SAM)-based models have been modified for medical imaging workflows to demonstrate significant time-savings in the auto-contouring process [307].

Deep Learning for GBM Target Delineation: Auto-segmentation of glioblastoma (GBM) target volumes has the potential to significantly reduce inter-observer variability and accelerate the planning process. While most previous DL-based methods have focused on pre-operative tumour segmentation without incorporating surrounding OARs or post-surgical cavities [308–310], recent models have been developed to address these limitations [311]. Comparative studies consistently report that DL-based approaches outperform atlas-based methods in target delineation accuracy across various anatomical regions [312–314]. For example, combining deep learning-based auto-contouring with manual corrections has been shown to improve accuracy and efficiency in post-operative lung cancer CTV delineation [315], although no equivalent studies currently exist for GBM.

Commercial Implementations: The transition to AI-driven segmentation has led to rapid commercial development. Numerous vendors now offer pre-trained, clinically deployable auto-contouring tools [296, 299]. These include products from Manteia Medical Technologies, Mirada Medical, and Carina Medical, among others [312, 316]. Such solutions have demonstrated measurable benefits in improving the efficiency of OAR segmentation and reducing inter-observer variability across clinical workflows [299].

Anatomical Sites: Comparative studies across multiple anatomical regions using seven different AI systems show that no single platform consistently performs best across all sites, highlighting the need to tailor tool selection to specific clinical contexts [317–319]. Anatomy-specific models have shown strong performance in various cancer types.

In head and neck cancer, 3D U-Net architectures generalize well across international datasets, achieving expert-level accuracy. Clinical adoption is growing, with up to 50% of auto-contours used without modification and time savings of up to 112

minutes per case [302]. For glioblastoma, CNN-based segmentation has addressed the variability and effort of manual contouring [311]. In nasopharyngeal cancer, AI tools achieved 79% accuracy and reduced both inter-observer variability (by 54.5%) and contouring time (by 39.4%) [320, 321].

DL models in breast cancer significantly reduce contouring time, in some cases from 40 to 10 minutes [297]. For lung cancer, DL-assisted methods improve segmentation accuracy and reduce planning time by 35%, with multi-centre trials reporting reduced inter-observer variability by about 50% [320, 322, 323].

In cervical cancer, VB-Net models achieved DSC up to 0.88, with 63.5% of contours requiring only minor edits. Performance was comparable to senior clinicians and superior to junior ones [298, 324]. In prostate cancer, 45% of surveyed oncologists already use AI tools clinically, reporting time savings up to 72% [318, 325]. For colorectal cancer, recent studies have demonstrated growing applicability of AI-based methods [318, 320].

Benefits and Challenges of using AI-contouring: AI-based auto-contouring significantly reduces inter-observer variability. In nasopharyngeal carcinoma, it lowered variation by 54.5% [320, 321], and in lung cancer, variation decreased by approximately 50% [323]. This consistency supports more standardized RT planning. Efficiency gains are also notable. Time savings range from 35% in lung cancer [320, 322] to 39.4% in nasopharyngeal carcinoma [320, 321], and up to 72% and 84% in prostate and head and neck cancers, respectively [150].

Despite these benefits, contour revisions can be time-consuming. In some cases, correcting AI-generated contours takes nearly as long as manual contouring [326], especially in adaptive workflows like proton therapy [198]. Human oversight remains essential. While AI improves consistency, clinical review ensures accuracy and safety [150]. Implementation also requires adherence to standardized contouring protocols, as clinician-AI discrepancies may affect planning outcomes [327]. AI tools often perform inconsistently across anatomical regions. Accuracy drops for small or complex structures, and no single platform outperforms others across all sites [319], requiring careful tool selection per clinical context.

The "black box" nature of AI presents challenges. Limited model interpretability and reliance on training data affect clinician trust and integration into clinical practice [296]. User training also influences outcomes. Less experienced users may spend more time adjusting contours, underscoring the need for proper training [328, 329]. Effective clinical use depends on strong QA protocols. Human review is still essential to ensure reliability and mitigate errors [150]. Model performance depends on the quality and diversity of training data. Lack of representative datasets can limit generalizability to real-world clinical scenarios [296].

Finally, adaptable models are needed to address diverse 3D structures and imaging modalities [330]. While AI tools are not yet perfect, they provide major gains in efficiency and standardization, enhancing modern RT workflows [331].

2.3.2 Robust Auto-contour Models

The growing reliance on DL models for contouring and treatment planning makes robustness essential for building trust among clinicians and patients. Reliable performance across anatomical and imaging variations is critical, especially in volumetric models that, despite strong results in organ and tumour segmentation, remain vulnerable to adversarial attacks [332, 333]. Poor RT quality has been linked to treatment failure, lower survival, and increased toxicity in clinical trials [140].

In oncology, robust contouring is vital for accurate tumour analysis and planning. Radiomics depends on precise contours to extract imaging biomarkers that guide treatment. Even small contouring errors can impact therapy decisions or lead to missed diagnoses [334]. Robust features also help identify tumour subtypes with distinct molecular traits and outcomes [335]. The use of multiple imaging modalities such as CT, MRI, and ultrasound further highlights the need for consistent auto-contouring. Models capable of stable performance across modalities support standardized treatment planning across diverse clinical environments. However, even state-of-the-art models like SAM show limitations in multi-modal, multi-object segmentation tasks [336].

Necessity for Robustness of State-of-the-art Solutions: The reliability of DL systems depends on both accuracy and robustness to input perturbations. Recent studies suggest that medical image segmentation models may be more vulnerable to adversarial attacks than previously believed, raising concerns for clinical use [337]. Sensitivity analysis is crucial in evaluating how input variations affect predictions, offering insights into model reliability in diverse anatomical and imaging scenarios [338–340].

Robust contouring models are also vital for building trust between clinicians and patients. Consistent model outputs across imaging conditions and patient anatomy help clinicians feel confident in AI-generated contours [332], a key factor for adoption. However, improving robustness typically requires extensive expert annotations, posing challenges for deployment in resource-constrained settings [341]. Models that achieve robustness with fewer labelled examples are more likely to succeed in real-world clinical workflows [342].

Types of Robustness: Auto-contouring models must exhibit several forms of robustness to be clinically reliable. One key type is *domain robustness or generalization*, the ability to generalize across data from different scanners, protocols, or institutions [343, 344]. *Texture robustness* addresses the sensitivity of DL models to textural bias. Simulating textural noise during training can improve invariance and performance on scans affected by unseen noise, especially in 3D data [345]. *Image quality robustness* ensures stable performance under conditions like blur or noise, which are common in intraoperative or real-time settings where training data is often cleaner [336]. *Modality robustness* is also essential. Even foundation models such as SAM show inconsistent performance across imaging modalities, highlighting the difficulty of achieving zero-shot generalization on multi-modal medical datasets [336].

Challenges in Building Robust Models: Despite progress, vulnerabilities in clinical deployment remain a concern, highlighting the need for robust solutions where model failure can have serious consequences [338]. A key challenge is *intensity inhomogeneity*, where varying pixel intensities within the same tissue lead to inaccurate contours [346]. This is worsened by common noise types (e.g., Gaussian, speckle) in medical imaging [347].

Weak boundaries between structures further complicate segmentation, especially in low-contrast or complex regions, often causing clinically relevant errors [348, 349]. Poor-quality annotations in datasets also introduce label noise that limits model performance [350]. *Domain shift* is another major barrier. Variability in scanners, protocols, and modalities creates distribution mismatches, and DL models often overfit to texture and style cues that do not generalize well [351, 352]. Moreover, *adversarial*

attacks, which are imperceptible perturbations crafted to mislead models, pose serious risks to segmentation models [353]. Such attacks are especially concerning in volumetric applications like OAR and TV contouring, where current defences lack theoretical guarantees [337, 354, 355].

Current Solution Approaches: To address these challenges, local statistics-based models have been used to mitigate intensity inhomogeneity and noise by leveraging localized image features [356]. *Uncertainty estimation* is another strategy to flag low-confidence predictions [348, 357, 358], although models remain well-calibrated at the dataset level but not at the subject level [359]. Simulated perturbations via data augmentation or adversarial training have improved robustness across domains and textures [360].

For modality robustness, uncertainty-based methods help address boundary ambiguities, including probabilistic contouring frameworks [336]. Some foundation models have shown improved domain generalization after fine-tuning [361]. Addressing these issues requires robust evaluation practices: assessing performance under varied conditions, validating across diverse patient groups, and documenting assumptions about data inputs that may affect clinical outcomes [340, 362–364].

2.3.3 AI for Fast Dose Predictions and QA

AI is well-suited for radiation oncology for analysing large datasets to improve treatment planning. In IMRT and VMAT, AI models can predict key variables at each planning stage, either by learning from prior clinical plans or by mimicking human decision-making [365]. Integrating AI into RT workflows helps automate repetitive tasks, reduce costs, and enhance QA and patient care [228]. These tools also support clinicians in delivering efficient, personalized treatment aligned with individual patient needs [120].

AI for Treatment Planning: Treatment planning in radiation therapy is complex and time-intensive, demanding manual expertise. AI offers a promising solution to automate and standardize this process, enhancing both efficiency and consistency [102]. AI applications fall into two categories: one predicts optimal DVH, 3D dose distributions, or fluence maps based on prior plans; the other mimics human decision-making during optimization [365]. Both reduce trial-and-error, improving plan quality.

Knowledge-Based Planning (KBP) trains models on curated datasets of clinical plans, enabling automatic generation of high-quality plans while reducing variability due to planner experience [366, 367]. DL-based models, especially those using CNNs, can directly predict fluence maps from anatomy in seconds, with performance comparable to clinician plans [368]. Recent developments include Large Language Model (LLM)-based agents for planning [369] and systems that auto-generate large numbers of plans to support AI training [124], helping standardize workflows and enhance quality [370].

AI has proven effective in complex sites like head and neck. The RPA system from M. D. Anderson Cancer Center automates contouring and planning for multiple cancer types, often matching or exceeding manual plan quality while reducing time [144, 371]. For prostate SBRT, AI plans rival manual ones on clinical LINACs [372], while whole breast plans can be optimized in under 20 seconds [373]. Despite these advances, clinical implementation demands rigorous validation. Concerns about robustness in diverse scenarios remain, and outcomes must be verified

by clinical physicists [374]. Human oversight remains essential for ensuring plan safety and customization for individual anatomy [366, 375].

AI for QA: QA is vital in RT, but traditional processes are time-consuming and prone to uncertainty. AI offers automation and accuracy improvements for both machine-specific and patient-specific QA [227, 229]. In patient-specific QA, AI models predict gamma passing rates for IMRT and VMAT using plan complexity metrics, achieving accuracies within 3% of measured values via tree-based models like AdaBoost and XGBoost [376, 377]. DL models such as CNNs can also classify dose errors (e.g., MLC, monitor units, setup) from dose maps, outperforming basic gamma analysis by identifying root causes [378–381]. In proton therapy, where traditional systems lack built-in monitor unit calculations, AI models using Gaussian Process Regression (GPR) and neural networks predict output factors with mean errors under 2% [382–384]. Virtual QA using architectures like UNet++ can predict dose distributions and gamma rates without requiring physical measurements, offering faster alternatives [385–387].

Medical physicists validate AI models through comparisons with standard dose calculations and phantom tests [388]. This is especially important in adaptive RT, where AI can reduce QA workload during daily replanning [389, 390]. AI models can monitor machine components like MLCs and imaging systems over time, helping predict failures and reduce downtime [391]. Bayesian network tools have been used for automated plan review [392], while AI-based contouring systems now undergo QA checks for both geometric and dosimetric validity before use [150]. Challenges remain, as most studies are retrospective or simulated [393]. Safe deployment, model maintenance, and workflow adaptation are key issues. AI is poised to improve contouring, registration, and real-time adaptive treatment, including MR-based synthetic CT generation and QA integration [394, 395].

2.4 Evaluation Metrics

Performance evaluation must be tailored to the specific nature of the predictive task, whether classification, regression, segmentation, or treatment planning, and should be grounded in clinical relevance, particularly the potential implications of false positives and false negatives [396]. In the context of biomedical imaging and radiation oncology, a comprehensive set of evaluation metrics is essential for reliable model validation and clinical translation. This section provides an overview of the classification 2.4.1, segmentation 2.4.2 and radiation oncology-specific 2.4.3 metrics used in the research works that follow.

2.4.1 Classification Tasks

For classification tasks, the following metrics are commonly used:

- **Accuracy**:
$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$
This represents the proportion of correctly classified instances (true positives and true negatives) out of all predictions. It is most informative in datasets with balanced class distributions but can be misleading in imbalanced datasets, which are common in medical imaging.

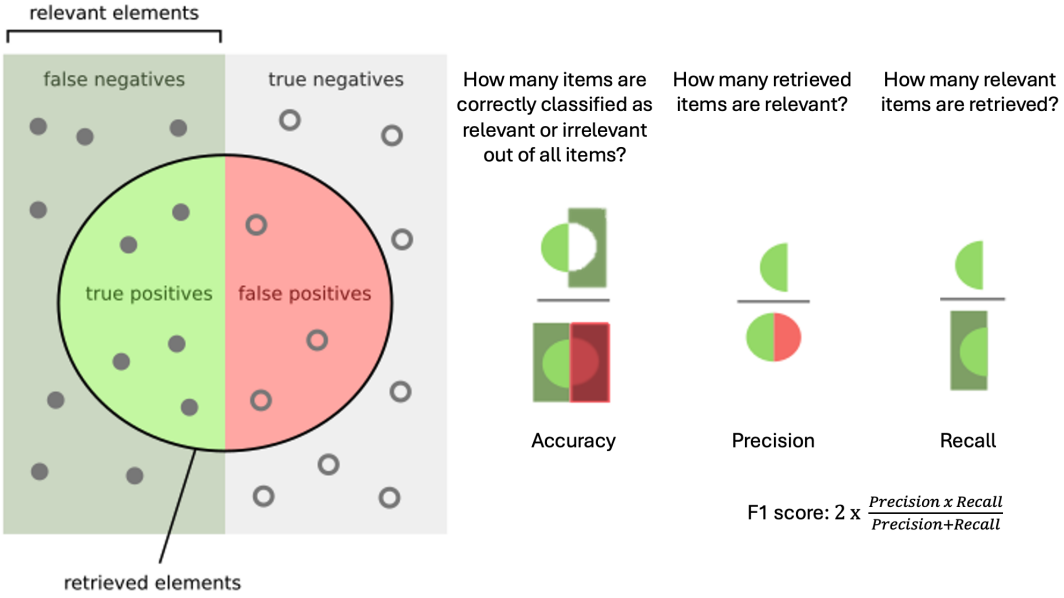


FIGURE 2.5: Visualization of typical classification metrics in a binary scenario. Figure adapted and modified from Wikipedia.

- **Precision (Positive Predictive Value)**:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision measures the proportion of predicted positive cases that are actually positive. It is particularly critical when the cost of a false positive is high, such as unnecessary biopsies or treatments.

- **Recall (Sensitivity or True Positive Rate)**:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall captures the ability of the model to correctly identify all actual positive instances. In diagnostic imaging, high sensitivity is often prioritized to minimize the risk of missed detections.

- **F1 Score**:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score provides a harmonic mean of precision and recall, offering a balanced measure when both false positives and false negatives are clinically consequential.

- **Area Under the Receiver Operating Characteristic Curve**: The AUC-ROC (Area Under the Receiver Operating Characteristic Curve) quantifies the model's discriminative capacity across all classification thresholds. It plots the true positive rate against the false positive rate and is particularly useful for comparing classifiers independent of class imbalance [397].

2.4.2 Segmentation/Contouring Tasks

In segmentation tasks, especially those used in RT for delineating OAR and TV, voxel-wise classification metrics are often insufficient. Instead, spatial agreement metrics are employed to assess the geometric concordance of predicted contours to ground truth annotations [100].

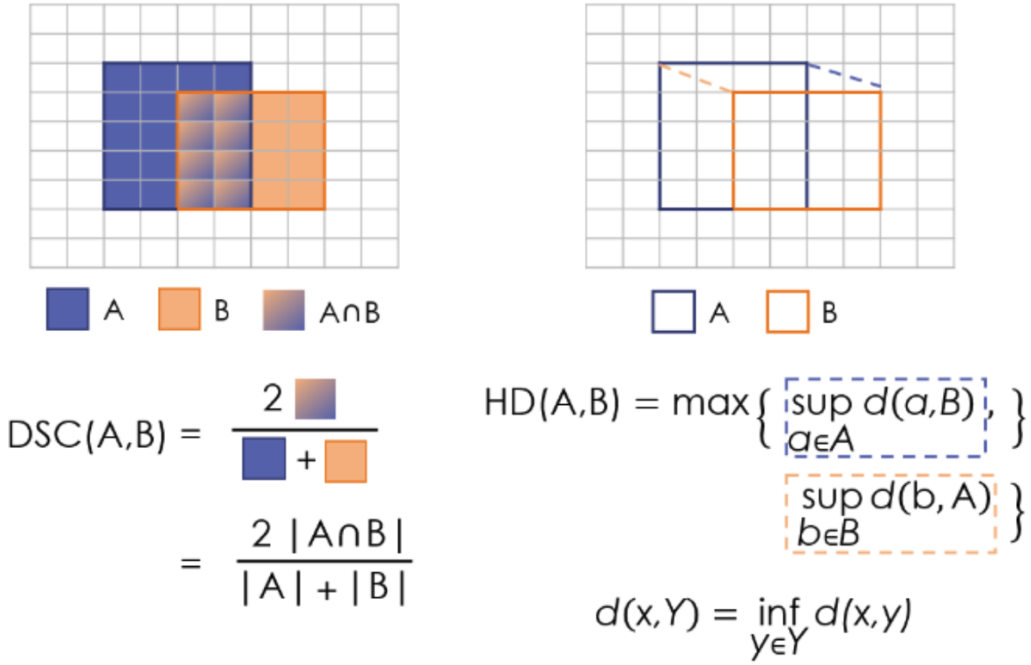


FIGURE 2.6: Visualization of the two types of segmentation metrics: overlap based (DSC) and distance based (HD). Surface DSC is a variation of DSC only on the borders, and Average Symmetric Surface Distance (ASSD) is a variation of HD. Figure adapted from [398].

- **Dice Similarity Coefficient**:

$$\mathrm{DSC} = \frac{2|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A}| + |\mathrm{B}|}$$

The DSC [399] quantifies volumetric overlap between the predicted segmentation A and the ground truth B. It ranges from 0 (no overlap) to 1 (perfect overlap) and is widely used for evaluating organ and tumour contours.

- **Hausdorff Distance**:

$$\mathrm{HD}(A, B) = \max \left\{ \sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(b, a) \right\}$$

The HD [400] measures the largest boundary deviation between two contours. It highlights worst-case errors, which are critical in high-precision applications like stereotactic radiotherapy.

- **ASSD**:

$$\text{ASSD}(A,B) = \frac{1}{|A|+|B|}\left(\sum_{a\in A}\min_{b\in B} d(a,b) + \sum_{b\in B}\min_{a\in A} d(b,a)\right)$$

ASSD computes the average boundary distance between the predicted and ground truth segmentations, offering a smoother measure than HD and less sensitivity to outliers.

- **Surface Dice**: A variant of DSC computed on the surface voxels within a tolerance (e.g., 1–2 mm), Surface Dice assesses the fraction of boundary points from one surface that are within a specified distance of the other. This is especially important in clinical settings where small boundary errors can significantly impact dosimetric outcomes [401].

2.4.3 Radiation Oncology-specific Metrics

In RT, evaluation must extend beyond geometric accuracy to include dose distribution and biological impact. Dose distributions are measured typically using a DVH, as shown in Figure 2.7.

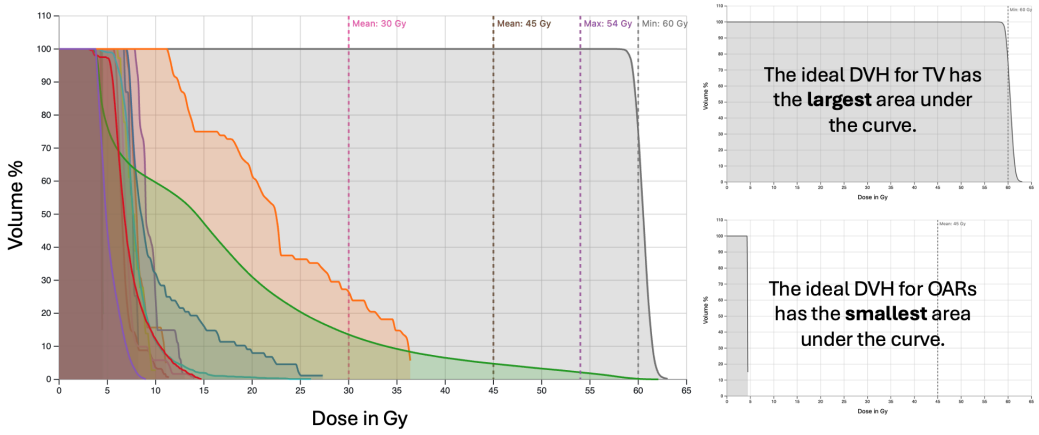


FIGURE 2.7: Typical DVH curves: representing 3D doses in a 2D histogram, where percentage of volume receiving levels of doses up to the prescribed treatment dosage is plotted.

It is a two dimensional graph that shows how much radiation dose is received by different volumes of tissue (both tumour TV and OARs). The x-axis represents the radiation dose (in Gray), and the y-axis shows the percentage of the tissue volume receiving at least that dose. By analysing the DVH, clinicians can assess whether a treatment plan is safe and effective. The following domain-specific metrics, some of which are based on points on the DVH plane are used to assess treatment planning and dose prediction quality [402, 403]:

- **OAR Sparing Metrics (DVH Parameters)**: DVH derived metrics for OARs quantify exposure levels:
  - D_2: Dose received by the hottest 2% of the volume.
  - D_{95}: Dose received by at least 95% of the volume, often used as a target coverage benchmark.

  - V_{20}, V_{30}, etc.: Volume fraction receiving at least 20Gy, 30Gy, etc.

  These are key indicators of potential toxicities to healthy tissues.

- **Target Coverage**:

$$\text{Coverage} = \frac{V_{Rx} \cap V_{Target}}{V_{Target}}$$

  This quantifies the proportion of the TV receiving at least the prescription dose. Ideal values approach 1.0 for conformal treatments.

- **Conformity Index**:

$$\text{CI} = \frac{V_{Rx}}{V_{Target}}$$

  A measure of how well the prescribed isodose volume conforms to the target. A CI close to 1.0 indicates optimal conformity.

- **Homogeneity Index**:

$$\text{HI} = \frac{D_{2\%} - D_{98\%}}{D_{50\%}}$$

  HI reflects the uniformity of dose distribution within the target volume. Lower values indicate more homogeneous dose delivery.

Part I

Validation of Clinical Needs

This part is focused on validating the clinical needs. The articles included have been modified slightly from the published versions so as to not expand abbreviations already used.

Chapter 3 has a shared first-authorship with Dr. Jonas Willmann, where he contributed to the design of the experiments and writing and reviewing for clinical correctness, and my contribution was organizing and running the experiments, analysing the results, generating the figures, thematic analysis, and writing and reviewing the manuscript. It has additionally been extended with unpublished work demonstrating early results in contour quality variations across a multi-centre clinical trial across several centres enrolled through European Organization for Research and Treatment of Cancer (EORTC) across Europe.

The work in Chapter 4 is an initial evaluation with a single OAR, which was then extended in Chapter 6, in Part II. Between the duration of time of publishing these papers and the time of writing this thesis, several advancements have been made in DL-based dose prediction (also called dose proposer) models, especially through the AAPM challenge in 2025 [124]. This field is under active development with full RT planning automation explored [404], including estimating the deliverable machine parameters [134].

The evaluation experiments were common to Chapters 3, 5 and 11, but the methodology of evaluating dose differences develops progressively into more complex and clinically relevant metrics. This explains the differences between the comparative results between clinicians and DL-based dose prediction models.

3

Predicting the Impact of Target Volume Contouring Variations on the Organ at Risk Dose: Results of a Qualitative Survey

3.1 Introduction

The accurate delineation of TV and OARs is a cornerstone of RT planning, directly impacting the balance between tumour control and the preservation of healthy tissue. Despite its importance, contouring remains a critical vulnerability in the RT workflow, subject to significant inter-observer variability and prone to error [405–407]. These inconsistencies have been shown to influence dosimetric outcomes, potentially compromising treatment efficacy and patient safety [140, 219, 408–411].

Advancements in auto-contouring have introduced opportunities to reduce variability and improve efficiency in clinical workflows [412]. Automated tools have demonstrated efficacy in delineating OARs and are increasingly being improved for TVs [413]. However, most automatically generated contours require manual review and adjustments before clinical application [414].

With the increasing reliance on auto-contouring systems, the role of radiation oncologists is shifting from manual contouring to reviewing and refining automatically generated contours. As time saving compared to manual delineation is a key goal of auto-segmentation, keeping the contour adjustment time at the necessary minimum is crucial [414]. This shift highlights the necessity of understanding inter-evaluator variability and its impact on dosimetric assessments in auto-contouring.

This study addresses these challenges by systematically investigating the subjective estimation of dosimetric impacts on OARs for GBM CTV contour variations and the related inter-evaluator variability. These findings are intended to support the development of standardized practices, ultimately enhancing the consistency and reliability of contour QA and thereby RT planning.

3.2 Methods

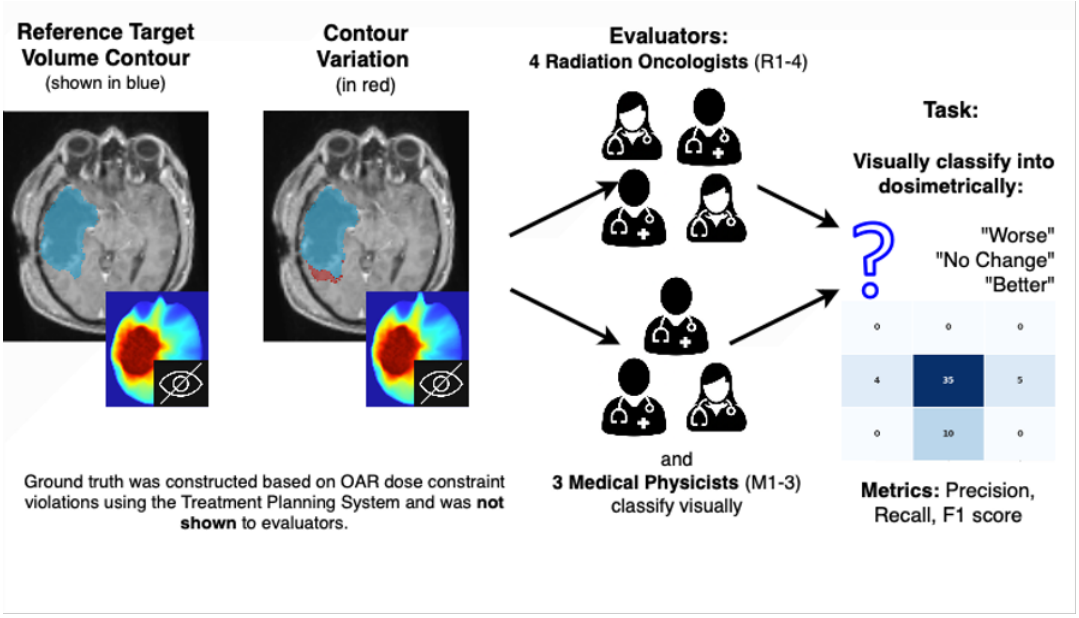


FIGURE 3.1: Schematic of the study design: 54 reference-variation pairs of target contours and variations along with OAR contours are presented to four radiation oncologists (R-1,2,3,4) and three medical physicists (M-1,2,3) for visually evaluating if the variations have negative ("Worse"), neutral ("No change") or positive impact ("Better") on the dose to the OARs. Ground truth categories are computed using dose distributions from treatment plans generated for each condition, not shown to the evaluators.

3.2.1 Data and Study Design

This study utilized imaging and contour data from 14 GBM patients treated at the Bern University Hospital. Data included CT and T1 contrast-enhanced MRI, contours for 13 OARs, and CTV. OARs included the brainstem, optic chiasm, cochleae, eyes, hippocampi, lacrimal glands, optic nerves, and pituitary gland, and were contoured based on clinical guidelines [415]. This is shown in Figure 3.1.

54 contour variations were generated across 14 patients by introducing up to four localized modifications to the reference CTV per patient (See Figures A.1 and A.2 in the appendix for examples). These variations were additive and designed by an experienced practitioner based on the spatial relationship to nearby OARs and their impact on dosimetric outcomes to mimic realistic changes a radiation oncologist would consider making to auto-segmented contours.

The contour variations varied in size with a median of 13.2 cc (range: 4.03 cc and 50.18 cc) and had a DSC with a median of 0.981 (range: 0.96 to 0.99), HD with a median of 7.06 mm (range: 4.88 mm to 12 mm) compared to the reference. The differences in dose variations for the high priority OARs (see Table A.1 in the appendix) were between +8.1 and -1.86 Gy in the brainstem (maximum dose), +2.84 to -4.32 Gy for chiasm (maximum dose), +2.81 to -2.11 Gy for the lacrimal glands (mean dose) and +2.22 to -3.05 Gy for the optic nerves (maximum dose).

3.2.2 Ground Truth Classification

Treatment plans for each CTV variation were created to assess the dosimetric impact on OARs. Standardized treatment plans were created using a double-arc co-planar VMAT protocol, optimized in Eclipse (Varian Medical Systems) with the photon optimizer and using the AAA dose calculation algorithm. Dose constraints for each OAR were based on clinical guidelines [415]. Plans were standardized by using identical prescription doses and objective prioritization for all patients, without individual adaptation beyond anatomical variation.

Ground truth labels ("better," "no change," or "worse") were assigned to each contour variation based on dosimetric analysis of the associated treatment plans. The same set of objectives was used to optimize a new plan on the new structure set that contains the variation. This results in new OAR doses for each variation. The evaluation was conducted using predefined dose constraints for the 13 OARs, as detailed in Table A.1 in the appendix. For each OAR, a point system was applied: +1 was assigned if the dose after the CTV variation exceeded the organ-specific threshold constraint as compared to the dose with the reference CTV contour, and -1 when the variation reduced the dose below the constraint if the original exceeded it. The points were then summed across all 13 OARs. Variations with a cumulative positive score were classified as "worse," those with a cumulative score of zero as "no change," and those with a cumulative negative score as "better."

3.2.3 Assessing Evaluator Performance Against the Defined Ground Truth

Clinicians' and physicists' performance in reviewing the contour variations and predicting if they would influence OAR dose was subsequently assessed by comparing them to the ground truth classifications. Four radiation oncologists and three medical physicists were presented with the reference contours and variations overlaid on the T1 contrast image using Slicer 3D as the visualization platform on a per-patient basis. A table describing the background and experience of each evaluator is included in Table A.2 in the appendix. Importantly, evaluators were instructed to assess only the impact of contour variations on OAR dose, without considering the quality or adequacy of the TV coverage. They were not shown the dose distribution and were only instructed that the treatment plan was VMAT with two coplanar arcs.

Results were summarized using confusion matrices to illustrate agreement and disagreement across classes and quantified with precision, recall, and F1 scores to comprehensively assess classification accuracy and reliability. Weighted averages are reported to account for class imbalance, offering a more clinically relevant evaluation of performance.

To qualitatively assess the evaluators' thought processes, they were recorded while verbally describing factors influencing their decision-making during the assessment tasks. Transcripts of these recordings were analysed based on the thematic analysis approach [416], using different LLMs (Grok version 3.0 and ChatGPT 4o) for text analysis. The output of one LLM (Grok) was chosen as the preferred themes, due to their perceived greater clinical relevance. Themes influencing decision-making were not pre-defined and emerged from the text analysis.

3.2.4 Inter-Evaluator Variability

Cohen's Kappa was used to evaluate agreement between evaluators, with thresholds of $\leq$ 0.20 indicating slight agreement, 0.21–0.39 fair, 0.40–0.59 moderate, 0.60–0.79 strong, and 0.80–1.00 almost perfect agreement [417].

3.3 Results

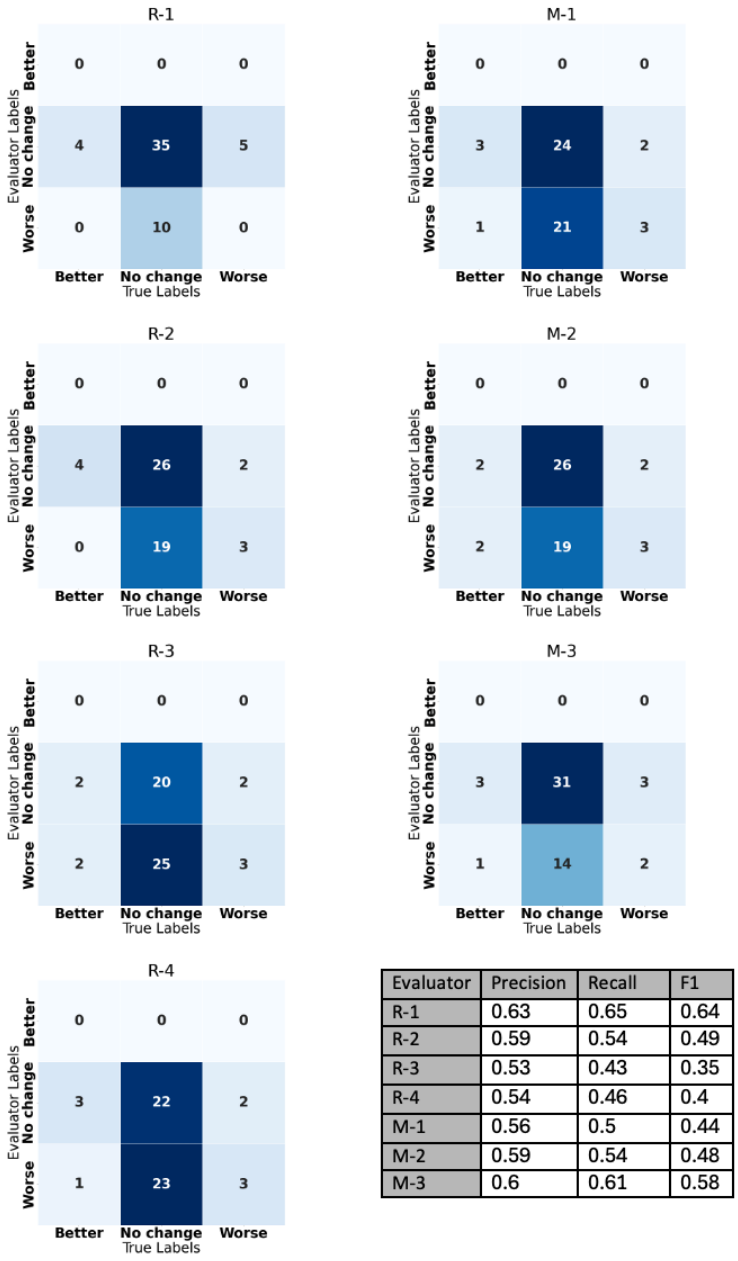
TABLE 3.1: Cohen's Kappa values (between -1 and 1; -1 indicating complete negative correlation and 1 indicating perfect positive correlation) between the seven evaluators. Pairwise Kappa values ranged from 0.33 (minimal agreement) to 0.73 (moderate agreement).

	R-1	R-2	R-3	R-4	M-1	M-2	M-3
R-1	1	0.47	0.33	0.47	0.41	0.34	0.60
R-2	0.47	1	0.66	0.60	0.74	0.62	0.65
R-3	0.33	0.66	1	0.67	0.68	0.72	0.60
R-4	0.47	0.60	0.67	1	0.63	0.59	0.67
M-1	0.41	0.74	0.68	0.63	1	0.66	0.73
M-2	0.34	0.62	0.72	0.59	0.66	1	0.59
M-3	0.60	0.65	0.60	0.67	0.73	0.59	1

The inter-evaluator agreement for estimating the impact of CTV variations on dose to OARs, assessed with Cohen's Kappa values, revealed a wide range of variability across participants (Table 3.1). Pairwise Kappa values had a median of 0.62 with an inter-quartile range of 0.08. They ranged from 0.33 (fair agreement) to 0.74 (strong agreement). Amongst radiation oncologists, the range was 0.33 to 0.67, while among the medical physicists, the range was between 0.59 and 0.73. Only 11 (52%) of 21 evaluator pairs demonstrated strong agreement (Kappa > 0.6).

Performance in evaluating the 54 contour variations revealed notable discrepancies compared to ground truth labels. The ground truth categories included 44 "no change", 5 "worse", and 4 "better" classifications. On average, 18 (42%) variations that resulted in no change of OAR dose were misclassified as "worse," i.e., clinicians expecting incorrectly that OAR dose constraints might be exceeded. None of the evaluators classified any variation as "better," i.e., reducing the OAR dose below constraints, despite ground truth indicating this category in 4 (7.4%) cases. Confusion matrices demonstrated variability in performance, with precision, recall, and F1 scores summarized in Figure 3.2. Precision is the ratio of true positive predictions to the total number of positive predictions made by the evaluator, reflecting the accuracy of positive classifications. Recall is the ratio of true positive predictions to all actual positive instances, measuring the evaluators' ability to identify all relevant cases. The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both concerns, especially useful when the distribution is uneven. Among the radiation oncologists, R1 demonstrated the highest classification performance, with a precision of 0.63, a recall of 0.65, and an F1 score of 0.64. In contrast, R3 showed the lowest performance, with a precision of 0.53, a recall of 0.43, and an F1 score of 0.35. Among the medical physicists, M3 achieved the best performance, recording a precision of 0.60, a recall of 0.61, and an F1 score of 0.58. None of the evaluators classified any case as "better." Instead, 32 (59%) variations were labelled as "no change" and 22 (41%) were labelled "worse" on average across all evaluators, contributing to an inflated "worse" category and a systematic underestimation of the other two.

Results of the thematic analysis include the top three themes behind the evaluators' thought process of decision making. The three themes are (i) proximity to OARs, (ii) size of the contour change, and (iii) shape of the contour change and



Evaluator	Precision	Recall	F1
R-1	0.63	0.65	0.64
R-2	0.59	0.54	0.49
R-3	0.53	0.43	0.35
R-4	0.54	0.46	0.4
M-1	0.56	0.5	0.44
M-2	0.59	0.54	0.48
M-3	0.6	0.61	0.58

FIGURE 3.2: Confusion matrices and precision, recall, and F1 metrics of each evaluator compared to the ground truth labels for each of the three categories: "Better," "No change," and "Worse." Weighted averages are reported to account for class imbalance, offering a more clinically relevant evaluation of performance.

beam geometry considerations. More details and quotes from the evaluators are included in Table A.3, and the distances between CTV variations and OARs are listed in Figure A.4 in the appendix.

3.4 Discussion

The variability observed among evaluators underscores a critical challenge in contour evaluation. The median Cohen's Kappa value pointed to generally strong agreement. However, 10 (48%) evaluator pairs achieved only weak to moderate concordance. The most notable variability occurred in the classification of "no change" cases, which were frequently misclassified as "worse." These findings highlight systematic overestimation of negative impacts, underscoring the need for standardized methods to reduce subjectivity, improve contour quality assessments, and avoid time-consuming yet unnecessary adjustments.

Evaluators demonstrated systematic biases in their evaluations, particularly a tendency to overestimate the negative impact of contour variations. The frequent misclassification of "no change" cases as "worse" reflects a cautious approach, likely motivated by the imperative to minimize potential harm to OARs. This conservative bias, while understandable, underscores the importance of establishing clearer criteria for assessing dosimetric impacts. The absence of any "better" classifications suggests a hesitancy to attribute positive impacts to contour modifications, even when ground truth indicates improvements. The "better" category was included in the evaluation because some modifications might reduce dose to nearby OARs without compromising target coverage. For instance, when a target volume is very close to an OAR, some under-coverage near the OAR may be necessary. However, if the target is modified by adding a region farther from the OAR that can be easily covered by the prescription dose, the overall target coverage may improve, as the relative impact of the under-covered area becomes smaller. The average dose reduction in the four "better" cases was 0.67 Gy, which is a subtle improvement, difficult to identify visually without direct dose visualization. Standardized guidelines or automated tools could assist in identifying clinically relevant improvements, thereby reducing this bias.

R1 has the highest precision and recall (0.63 and 0.65) compared to all the evaluators, while also having the lowest pairwise Cohen's Kappa scores compared to the rest of the evaluators. A possible explanation for this difference is that R1 has extensive experience in quality assurance of contours and associated dose distributions through prior training and experience with clinical trials. Analysis of the transcripts of the recordings during the experiment indicates that R1 was more decisive and used a more structured mental model, with a generally conservative scoring tendency, contributing to differences compared to the other evaluators.

The skewed distribution of ground truth classifications further complicated the evaluation process. With 83.3% (45) variations classified as "no change" and only 4 (7.4%) as "better," evaluators faced an inherently imbalanced task. The disproportionate assignment of cases to "no change" or "worse" categories suggests a lack of confidence in distinguishing meaningful improvements from minor or negligible changes. This highlights the importance of incorporating quantitative, dose-aware metrics into the evaluation process to aid objective decision-making.

The thematic analysis of decision-making processes uncovers several relevant insights. Intuitively, when a contour change brings the target closer to these OARs, more dose might need to be deposited in the OAR to maintain target coverage. Furthermore, the size of a contour change influences the volume of tissue irradiated. Therefore, larger variations might have a greater impact on OAR dose. Complex shapes (e.g., U-shaped targets) do create treatment plan optimization challenges, potentially increasing dose to OARs or reducing target coverage.

The cumulative point system employed in this study offers a structured and pragmatic framework for quantifying the dosimetric impact of contour variations, addressing the current lack of consensus methodology. By assigning positive, negative, or neutral scores to each OAR based on predefined dose constraints, the system enables an objective and repeatable approach to classification. Its simplicity ensures clarity across evaluators, minimizing ambiguity in interpretation and capturing genuine inter-evaluator variability. This approach provides an initial method for evaluating how 3D dose distributions are affected by contour changes, as illustrated in Figure A.3 in the supplementary material, showing dose levels within each OAR due to CTV contour modifications. Further refinement and validation are, however, necessary to enhance its clinical relevance.

Overall, these findings emphasize the urgent need for standardized contour evaluation protocols in radiotherapy. Clear guidelines on measuring dosimetric impact on OARs, the likelihood of toxicity, and guidance on dose constraint interpretations would provide a consistent framework for assessing the dosimetric impact of contour variations, reducing subjectivity and variability among evaluators. Additionally, integrating automated tools into the workflow could enhance the efficiency and accuracy of contour evaluations. For instance, AI-based algorithms to predict dose distributions could provide quantitative assessments of dose distributions [418], serving as an objective benchmark for evaluating contour modifications during the contour evaluation workflow of clinicians.

Several limitations of this study warrant consideration. First, the system assumes equal weighting for all OARs, which may not reflect their varying clinical significance. Critical structures like the brainstem could require greater emphasis than less vital OARs, such as the lacrimal glands. Second, the reliance on fixed dose constraints may, in some cases, miss relevant changes (e.g., when the absolute dose difference is large but does not exceed dose constraints) or overestimate effects (e.g., when a dose constraint is breached despite small absolute dose differences). Additionally, simple aggregation of scores across multiple OARs risks diluting the impact of significant variations in any high-priority structure. In cases with relatively large dose changes, higher-priority OARs like the brainstem and chiasm were typically more impacted than lower-priority structures. A substantial negative impact on a high-priority OAR could be masked by minimal changes in other low-priority regions. Future work should focus on integrating OAR-specific weighting schemes, Normalized Tissue Complication Probability (NTCP) modelling to guide the assessment of clinically meaningful contour variations. Validation of the system's classifications against clinical outcomes, such as toxicity rates or treatment efficacy, will be crucial to ensure its utility. This study was only based on glioblastoma cases and thus might not necessarily be applicable to other tumour types and anatomical regions. The relatively small number of evaluators from two hospitals in the same country may not capture the full extent of inter-observer variability in clinical practice. Additionally, it remains unclear if the contour variations produced in this study accurately reflect those observed in real-world scenarios, particularly variations generated by auto-contouring tools. Expanding this research to include a broader range of clinical scenarios and larger datasets would provide a more comprehensive understanding of variability in contour evaluations.

3.5 Conclusion

TV contour variations were frequently judged as having a negative impact on the dose to OARs, while they did not have. This overestimation of dosimetric impact represents a critical challenge for clinical practice, where evaluation of automatically generated contours is increasingly common, and unnecessary adjustments decrease workflow efficiency. Standardized guidelines and automated tools to guide the evaluation are needed to mitigate subjective biases.

3.6 Unpublished extension to a multi-centre EORTC Study

Based on the results from this study, we are actively working on extending this analysis to a large multi-centre clinical trial called MIRAGE, where we gained access to data including "first" contours and treatment plans, and "last" contours and treatment plans, with which we aim to conduct analysis in the following manner:

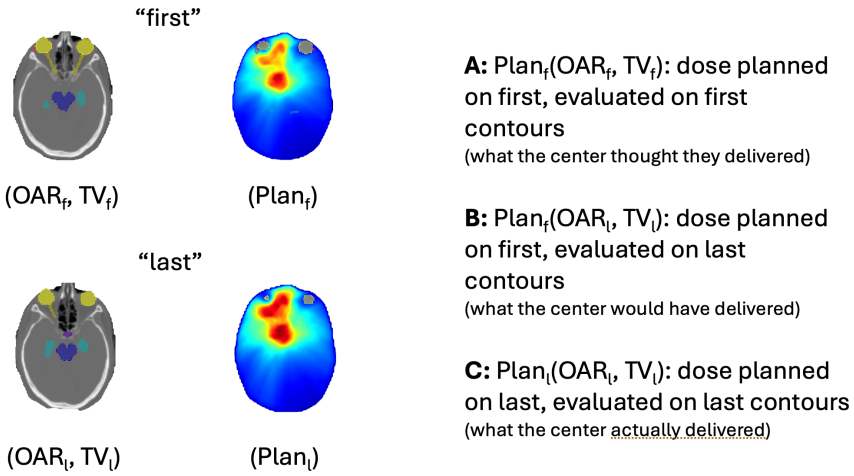


FIGURE 3.3: Study design for analysing the impact of contour QA as well as RT replanning efforts dosimetrically: given a set of "first" and "last" contour and corresponding treatment plan pairs, what is the actual dosimetric impact of re-contouring and re-planning?

(a) indicates the situation where the "first" contours and the "first" treatment plan dose distributions are used to compute dosimetric compliance according to predecided constraints for OARs and TV doses. This is done on a set of 41 subjects, in a prospective manner, at the point of inclusion of the centre into the clinical trial. (b) represents the situation where the "last" contours are evaluated over the "first" dose distribution, which provides a measure of what the centre would have delivered to the "true" OARs and TVs. Finally, (c) is the situation where the "last" contours are evaluated on the "last" dose distribution, demonstrating what happens after the contour review and replan. Some initial results in our analysis with this data is included in Figures A.5 and A.6.

4

How sensitive are Deep Learning based Radiotherapy Dose Prediction Models to Variability in Organs at Risk Segmentation?

4.1 Introduction

Aggressive tumours like glioblastoma account for 45% of all malignant primary brain tumours [419]. Current treatment is a combination of surgery, adjuvant RT, and concomitant and adjuvant chemotherapy [14]. The aim of RT planning is to conform dose to the TV (i.e., tumour or resection cavity, with adjacent areas of potential microscopic spread) while sparing OAR. This limits normal tissue toxicity while ensuring optimal tumour control [420].

It is hence critical to have an accurate segmentation of the anatomy to achieve this objective. Radiation oncologists draw contours around OAR and TV, either manually or semi-automatically. This process however can take up to seven hours per patient [421]. In a multi-institutional delineation study among radiation oncologists, incorrect TV segmentation has been reported to have caused 25% of non-compliant treatment plans [140]. TV and OAR segmentation are hence amongst the most time-consuming yet error-prone steps in the RT process. Efforts have hence been made to create segmentation standards and develop RT QAsystems [86].

Impact of segmentation variability: Fig. 4.1 (left) shows overlapping cyan lines representing potential choices of brainstem contours due to inter-expert variability. The PTV is represented in orange, and the CTV in yellow. A heat map indicating RT dose distribution (colour wash) for a treatment plan in Gray (red: high, blue: low dose) is overlaid for the dose context. Variations in these contours are most critical in the border of the higher dose area, where the dose gradient is most steep, while less critical elsewhere. Fig. 4.1 (right) shows 10 examples of plausible optic nerve (left) contours. Over-contouring, where volumes are larger than the 'true' extent result in (i) overestimating the OAR dose since less area of the OAR lies within high-dose region, and (ii) potentially under-dosing the tumour TV, to spare the OAR

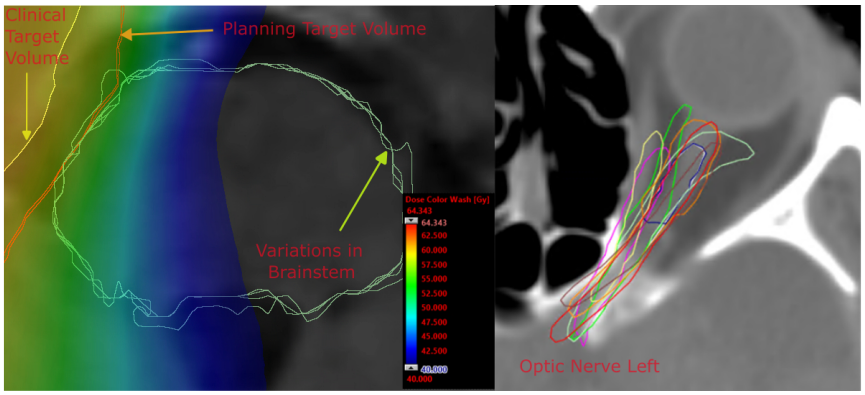


FIGURE 4.1: Visualizing inter-expert variability in OAR contours of brainstem in cyan (left). Orange and yellow contours are around the tumour TV. Overlaid heat map indicates dose. Various plausible left optic nerve segmentation (right) lead to changes in dose delivered.

from excess dose. This negatively impacting tumour control. Conversely, under-contouring i.e. missing 'true' areas of the OAR, would result in a under-estimation of the actual dose. This leads to excess toxicity.

Treatment dose plan computation is currently done independently after the contouring step, by medical physics experts. If this dose is not protocol compliant, reviewing any potentially incorrect contours and re-computing adds on to delays. This time between image acquisition and RT planning completion is reportedly 9.63 days on average [422]. This has motivated the use of DL for online and accurate RT dose predictions [423]. When such models are reliable, it could prevent re-planning by evaluating contour quality prior to or simultaneously with planning. However, to the best of our knowledge, the sensitivity of these DL models to local contour variations has not previously been analysed.

Hypothesis and Contributions: Our main hypothesis is that a DL dose prediction model that provides near-instant dosimetry is also sensitive to local contour changes, thereby being an efficient means of segmentation QA. The main benefit of using a DL dose prediction model is that it is near-instant (inference time of 15 seconds on a GPU). This enables an interactive segmentation process guided by dose estimates where contours are edited immediately based on dose compliance, as opposed to relying on post-facto dose evaluations leading to delays. To make this feasible in clinical practice, reliable sensitivity of such models to local changes in contours is essential.

We test this sensitivity by constructing simulated expert variations in contours and evaluating the similarity of dose predictions from such models to a reference plan. Our contributions in this paper are therefore threefold:

- Based on a data set of 100 clinical cases, we show that a Cascaded Three-Dimensional (C3D) U-Net dose prediction model [177] achieves a mean dose score of 0.906 and mean DVH score of 1.942, indicating strong potential usage in treatment planning. To the best of our knowledge, this is the first such analysis on dose predictions for glioblastoma.

- Based on a per-OAR (a total of 13) analysis of dose and DVH scores, we find that model performance depends on both size (larger is worse) and proximity to tumour TV (closer is worse).

- We further analyse the sensitivity of dose predictions to small yet realistic contour changes of the left optic nerve, selected due to its clinical relevance and sensitivity to radiation. We show a strong correlation of 0.921 between predicted dose versus reference dose differences.

4.2 Materials and Methods

Data: Our data set included imaging and contour data from 100 subjects who were diagnosed with GBM. This included CT imaging data, along with associated binary segmentation masks of 13 OARs (see full list in Table. 4.1) as well as the PTV. Each of these subjects also had a reference dose plan, calculated using a standardized clinical protocol with Eclipse (Varian Medical Systems Inc., Palo Alto, USA). This reference was a double arc co-planar VMAT plan with 6 mega volt flattening filter free beams, optimized (Varian photon optimizer version 15.6.05) to deliver 30 times 2 Gray while maximally sparing the OARs. The dose was calculated with the AAA algorithm [424], normalized so that 100% of the prescribed dose covers 50% of the TV. Sixty randomly chosen subjects formed the training set, 15 were used as validation (five samples excluded due to missing contours) and the rest of the 20 were used as the test set.

Model: We used a two-level C3D U-Net [177] as the dose prediction network (i.e, the input to the second U-Net is the output of the first concatenated with the input to the first U-Net). The model input was a normalized CT volume and binary segmentation masks for each of the 13 OARs and TV, and predicted a continuous valued dose volume (up-scaled from [0, 1] to 0 to 70 Gray) of the same dimension as the input. The loss was computed as

$$Loss = 0.5 * L1(reference, A) + L1(reference, B) \tag{4.1}$$

where A and B were the outputs of the first and second U-Nets respectively, $reference$ was the reference dose and $L1$ refers to the L1 loss. All volumes were resampled to 128^3 voxels, due to GPU memory constraints. The hyper parameters for training the C3D model were unchanged from the original implementation [177], except the number of input binary masks was updated to 14, to match the number of OARs in our data set. The weights were randomly initialized using the 'He' method. Training ran for 80000 iterations and the model with the best validation dose score was saved. All experiments were run with PyTorch 1.12 on an NVIDIA RTX A5000 GPU, and each training run took 24 hours. We trained the model five times with the same hyper parameter set but different random seed initialization to ensure reliable convergence.

Metrics: We adopted the dose and DVH score as evaluation metrics, from openKBP [423], an international challenge designed for head and neck tumours. Dose scores indicate the Mean Absolute Error (MAE) of predicted versus reference dose within a mask (either body, brain, or an OAR). DVH scores are the average of the MAE between prediction and reference for mean and 0.1cc dose of OARs, and the average of MAE for $1st$, $95th$ and $99th$ percentile of the dose for tumour TV, also computed within their masks.

Sensitivity experiments: To analyse the sensitivity of the trained model to inter-expert variability in segmentation, we manually modified the left optic nerve OAR for a single subject in the test set, to create ten variations (as shown in Fig. 4.1 (right)), validated by radiation oncologists for plausibility. For each of these, a new reference dose was computed using the same settings as used for the training data. The

left optic nerve was chosen because of its proximity to the tumour TV, resulting in large dose changes even for small contour changes. We then compared the predicted (P_i) and reference dose (R_i) qualitatively with DVH curves, and quantitatively by analysing the difference in mean OAR dose between reference and predicted dose, for nine variations against a reference (index 0 without loss of generality).

4.3 Results

Over five training runs, we report a mean dose score of 0.906 (stdev. 0.009), and a mean DVH score of 1.942 (stdev. 0.041) on 20 test subjects, which was in the same range as the winning entry [177] of the openKBP challenge. For subsequent analysis, we used the best performing model (out of five) with a dose score of 0.891 and DVH score of 1.919.

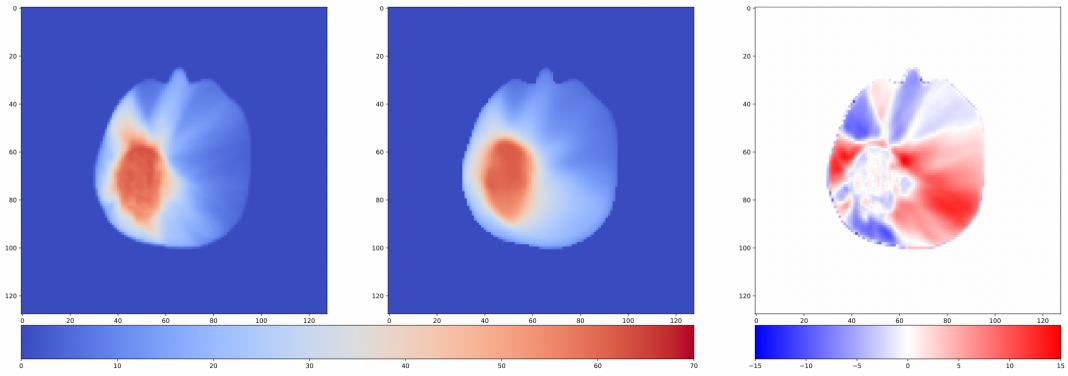


FIGURE 4.2: Comparison of dose predictions: Reference (left), model prediction (middle) (range of these values are from 0 to 70 Gray), and differences (right) (range of these differences is −15 to 15 Gray). For the difference image (right), darker blue regions are underestimates, and darker red are overestimates.

Fig. 4.2 demonstrates the dose prediction in the axial plane. The model tracked the reference well and avoided higher dose in the eye region (the two blue streaks at the top in Fig. 4.2), while also effectively enveloping the shape of the tumour TV. The difference between the two (right panel) consist mostly of radial streaks which were hardware specific artifacts that were not clinically pertinent to our assessment. Across the 20 test set subjects, the dose score varied between 0.470 and 2.167, where higher scores are typically due to larger tumour TV, because of higher overall dose to the entire anatomy. The DVH score varied between 0.451 and 4.203.

Table 4.1 shows the per-OAR results as mean (standard deviation) of dose and DVH scores. Larger OARs like the brainstem yielded better metrics, while smaller e.g., lacrimal glands are worse, leading to more over/under-estimates. Proximity to tumour TV was also an important factor in the dose score, where closer OARs had higher scores. Dose differences within tumour TV were nonetheless always under 2.5 Gray, which was less than 5% of prescribed dose.

Sensitivity analysis: Table. 4.2 shows the difference in the mean dose for the reference plans (second column), predicted plans (third column) and the corresponding DSC (fourth column) between a reference contour and nine variations of the left optic nerve. The contours are indexed in ascending order of the difference in mean reference dose (column two). The difference in the predicted dose tracked the difference in reference well, while the DSC trends were harder to use for making contour

TABLE 4.1: Mean (stdev.) of dose and DVH scores for 13 OARs in 20 test dose predictions. Lower values are better.

OAR	Dose Score	DVH Score
Brainstem	1.399 (1.392)	2.025 (1.746)
Chiasm	2.985 (2.418)	2.798 (2.469)
Cochlea L	1.856 (4.728)	1.036 (2.347)
Cochlea R	2.433 (5.109)	1.406 (2.673)
Eye L	1.487 (2.194)	1.707 (2.517)
Eye R	2.210 (3.939)	2.836 (4.832)
Hippocampus L	2.101 (1.743)	1.976 (1.618)
Hippocampus R	2.601 (2.945)	2.381 (2.166)
Lacrimal Gland L	1.448 (1.320)	1.617 (1.404)
Lacrimal Gland R	1.938 (2.011)	1.912 (2.069)
Optic Nerve L	2.121 (2.464)	2.475 (3.122)
Optic Nerve R	2.266 (2.342)	2.072 (2.135)
Pituitary	1.889 (1.780)	1.932 (1.689)
Overall	0.891 (0.376)	1.919 (1.216)

TABLE 4.2: Sensitivity analysis: R_i is the reference mean dose and P_i is the predicted mean dose for index 'i', both for optic nerve left. DSC(i) is the DSC between index 'i' and '0'. Dose difference (ΔD) reported in Gray.

Index (i)	ΔD: $\lvert R_i - R_0 \rvert \downarrow$	ΔD: $\lvert P_i - P_0 \rvert$	DSC(i)
1	0.145	0.418	0.325
2	0.283	0.222	0.627
3	0.357	1.032	0.783
4	0.435	0.519	0.363
5	2.089	3.171	0.590
6	2.402	2.487	0.509
7	3.027	1.483	0.197
8	4.815	5.436	0.612
9	7.591	5.625	0.229
Mean	2.115	2.039	0.523

quality decisions. For example, the contour with index 8 could be considered reasonable when looking only at DSC. But, the predicted dose showed that it is not as good dosimetrically, as index 1 having a lower DSC.

The average difference in the mean reference dose (ΔD: $\lvert R_i - R_0 \rvert$) was 2.115, while the same for predicted dose (ΔD: $\lvert P_i - P_0 \rvert$) was 2.039. The correlation coefficient between reference and predicted dose differences across these contour variations was a strong 0.926, while that with the DSC was -0.471. Furthermore, Fig. 4.3 shows the similarity in DVH curves to qualitatively compare reference and predicted doses for four representative variations. These evaluations confirm that the prediction model reliably tracked dose changes across contour variations. This demonstrates the utility of dose prediction models during contouring so that edits are based on clinically relevant dosimetry rather than the current practice of using just image-based anatomy and geometry.

Discussion: In this paper, we showed with experiments on a data set of 100 clinical GBM cases that our dose prediction model has a mean dose error of less than 1

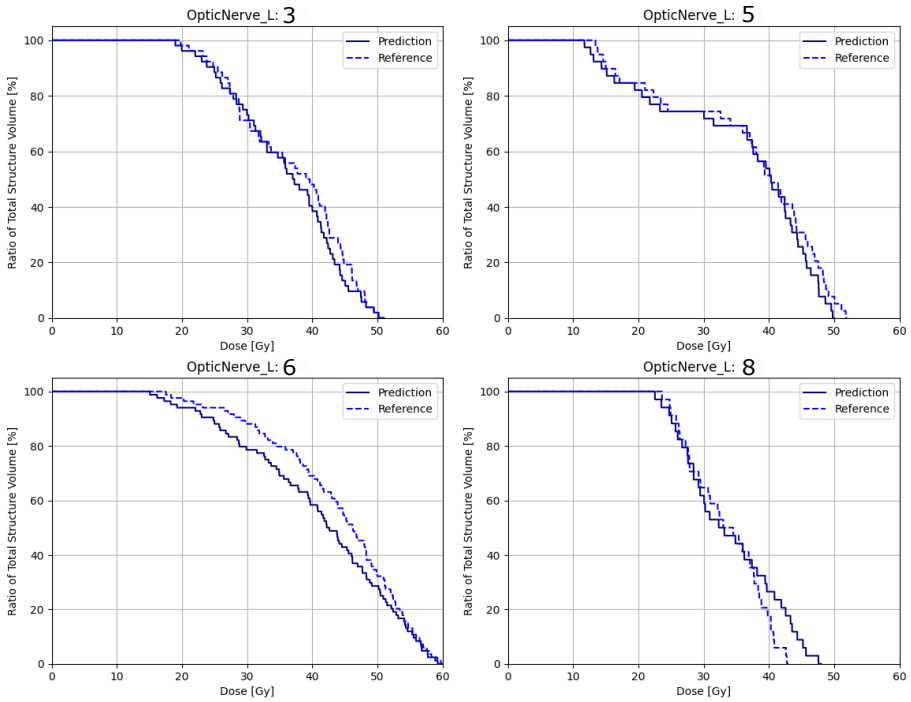


FIGURE 4.3: Comparison of DVH for Optic Nerve Left - for four representative realistic contour variations (index matches those in Table. 4.2). A smaller gap between the two curves indicates better results.

Gray on a test set of 20 clinical cases. This model was sufficiently sensitive to contour changes with a strong correlation while tracking dose changes, helping make better-informed contour editing decisions. However, a limitation is that separate models need to be trained for every tumour location, delivery machine, and planning software. We plan to build on this initial result to devise further experiments focusing on sensitivity.

Compliance with ethical standards: This study was conducted on retrospective human subject data from Inselspital (University Hospital Bern). Ethical approval was obtained from the regional ethics committee of the Canton of Bern.

5

Comparing the Performance of Radiation Oncologists versus a Deep Learning Dose Predictor to Estimate Dosimetric Impact of Segmentation Variations for Radiotherapy

5.1 Introduction

GBM, accounting for about 45% of brain tumours [419], is an aggressive malignant tumour treated with surgery, RT, and chemotherapy [14]. RT aims to target the tumour while minimizing dose to healthy OAR. The planning involves a trade-off between tumour control and tissue toxicity [420]. A critical step is the segmentation of structures, which is time-consuming when done manually, can take up to seven hours per patient in the head and neck anatomy [421].

With advancements in DL-based auto-segmentation, the role of radiation oncologists is shifting from manually drawing to monitoring and correcting these automated segmentations [425]. QA are hence crucial since it has been reported that incorrect tumour TVs cause 25% of non-compliant treatment plans, leading to untreated tumours or harmful RT doses [140]. While geometric metrics like DSC and HD are currently the de-facto metrics to evaluate segmentation quality, it has been reported that they do not correlate with dosimetric effects of contouring errors [101, 426]. In RT, it has been postulated that auto-segmentation methods must be evaluated using a diverse range of performance metrics, including impact on delivered dose [427], which ultimately impacts clinical outcome.

Related work: The clinical RT community has developed standards for target contouring [63], which mainly includes geometrical and anatomical considerations. Dosimetry-based considerations require dose plan calculations, which are time-consuming and necessitate iterations between the radiation oncologist and dosimetrists or medical

physicists. Hence, due to its time-consuming nature, dosimetric assessment of segmentation quality has not been conventionally employed in the clinics. Nonetheless, as recently pointed out by [425], dosimetry considerations are urgently needed in the quality control process of tumour and OAR segmentations.

Previously proposed approaches to evaluate the quality of automated segmentations include methods that predict segmentation metrics, such as DSC, [428] or use Graph Neural Networks to identify segmentation errors in RT [429]. These approaches assume that these geometric metrics reflect the quality of dosimetry, which is not the case [101, 426]. Furthermore, models predicting DSC perform poorly with low-quality segmentations (the main target of such QA system) due to a lack of representative training data for this performance range.

Other approaches have explored the use of uncertainty estimation in OAR segmentation in head and neck cancer [430], under the premise that high uncertainty is linked with potentially low-quality segmentations. However, geometric variability and uncertainty estimates may not imply dosimetric effects (e.g., high uncertainty of a contour located in a non-dosimetrically relevant area), and uncertainties based on imaging information alone may not sufficiently guide QA.

In line with dosimetry-focused QA, a DL-based dose prediction model is utilized in [431] to guide radiation oncologists on which volume slices require manual adjustments. This segmentation editing tool demonstrates potential for time efficiency while maintaining dosimetric equivalence with distribution maps produced without its assistance. In [432] we introduced a method that uses a DL-based dose predictor to assess the impact of local segmentation changes on dosimetric outcomes. However, this work focused on OAR segmentation, and not on tumour lesions, which is clinically more important due to the higher complexity of this segmentation task.

Hypothesis and Contributions: Beyond the state of the art, we postulate in this study that a DL-based dose predictor can be employed within a quality control framework to detect dosimetrically worse segmentations, with levels of performance superior to human experts. We substantiate this by comparing the performance of our DL-based QA method with that of three expert radiation oncologists, using a test dataset comprising 54 segmentation variations from brain tumour patients, and reference dose plans produced by a clinically validated TPS (Varian Medical Systems Inc., Palo Alto, USA).

To the best of our knowledge, this is the first study comparing the levels of dosimetric awareness on contour modifications between human experts and a DL dose predictor model.

5.2 Methods

Our study design, depicted in Figure 5.1, involves a set of reference tumour segmentations (ground-truth) and corresponding expert-derived contour variations (four per reference segmentation). For each reference and contour variation (n=54 pairs), dose plans are computed. Each contour is classified as *"Worse"*, *"No change"*, or *"Better"* based on dosimetric variations relative to the reference segmentation. This classification scheme is used to evaluate the ability of three experienced radiation oncologists and the proposed DL dose predictor model to accurately classify each contour variation.

We report classification metrics and time taken to perform the task.

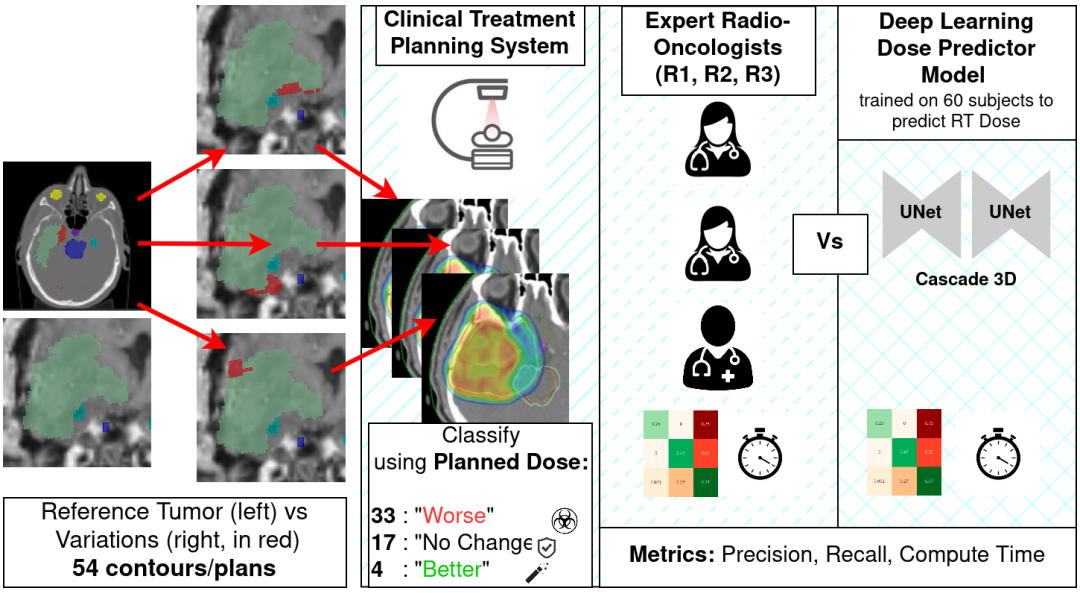


FIGURE 5.1: Is a DL dose prediction model able to ascertain dosimetric impact of tumour TV contour changes when compared to radiation oncologists? An experimental study is run with 54 contour variations which are individually re-planned to generate three categories of results: "Worse", "No Change" and "Better".

5.2.1 Data

Our data set includes imaging and segmentation data from 100 patients diagnosed with GBM. This includes CT and MRI T1 contrast-enhanced images, and binary segmentation masks of 13 OARs as well as the tumour TV. The OARs include brainstem, optic chiasm, cochlea (left and right), eye (left and right), hippocampus (left and right), lacrimal gland (left and right), optic nerve (left and right), and the pituitary gland. Each of these subjects also has a reference dose plan, calculated using a standardized clinical protocol with Eclipse (Varian Medical Systems Inc., Palo Alto, USA). This reference is a double arc co-planar VMAT plan with 6 mega volt flattening filter free beams, optimized (Varian photon optimizer version 15.6.05) to deliver 30 times 2 Gray while maximally sparing the OARs. The dose is calculated with the AAA algorithm [424], normalized so that 100% of the prescribed dose covers 50% of the TV. All the volumes are resampled to an isometric 2x2x2 millimeter grid of size 128^3 voxels using PyRaDiSe [433] and converted to Neuroimaging Informatics Technology Initiative (NIfTI) files to use for training and evaluation.

5.2.2 Experimental Setup

We train a C3D U-Net dose prediction model [177], which has been previously evaluated to show a mean prediction error of 1.38 Gray [418, 434] on a subset of n=60 cases (from the original 100 cases). The inputs are the CT volume, the OAR and tumour binary segmentation masks (14 volumes), and the output is the dose prediction. Ten cases are used as validation, and 14 are used as the test set for this study (Implementation details below). We save the rest of the 16 cases for future evaluations.

Contour modifications and replanning: For 14 test subjects, between three to four variations to the tumour TV are made independently by an expert using the same

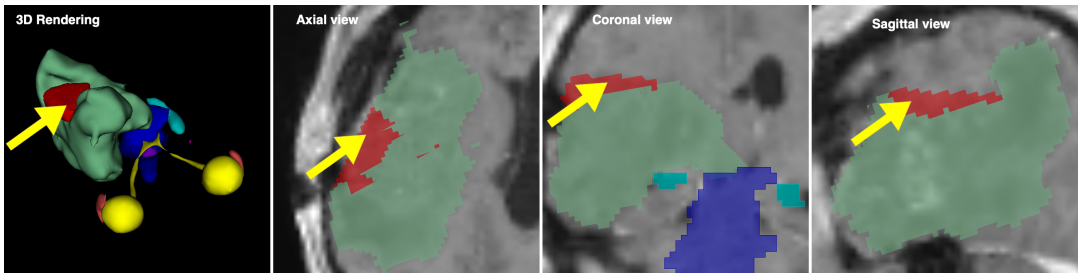


FIGURE 5.2: Example showing a tumour TV contour change that is "Worse" - negatively impactful. The green overlay is the reference contour, red overlay indicates change, marked with a yellow arrow. OAR contours are shown for anatomic reference.

rationale as in [101]. The reference segmentation follows current delineation standards. The dose is then re-planned with the same settings as earlier to construct a ground truth dose plan for each of the variations. This leads to 54 test scenarios, which are categorized into "Worse", "No Change" and "Better" from a dosimetry perspective. Figure. 5.2 shows how a variation (in red) looks compared to the reference (in green).

Ground-truth dosimetric categorization of contour variations: We define three categories of dosimetric impact based on a 10% change on the maximum dose recorded within each of the 13 OARs with respect to each reference contour as computed using the commercial TPS. A contour variation is considered as dosimetrically impactful if at least one OAR crosses this threshold. We define three categories: "Worse", "No Change" or "Better" - indicating if such a change to the tumour target volume leads to higher OAR dose, no change, or lower OAR dose as compared to the reference segmentation, respectively. On the set of 54 cases, this definition leads to 33 Worse, 17 No Change, and 4 Better scenarios based on the ground truth reference dose plans.

DL-based dosimetric categorization of contour variations: To automate the categorization of contours, we use the trained dose prediction model on the reference contour and each analysed contour variation, yielding dose maps D_{ref} and D_{cv}, respectively. Each contour variation is then classified following algorithm 1, which follows the class definitions presented above for ground-truth generation.

Algorithm 1 examines each OAR to identify those exceeding a specified percentage dose change, determined by threshold T. If the count of such OARs surpasses the hyper parameter $nOAR$, the case is deemed "Worse". In the absence of dose violations, the algorithm investigates potential dose improvements (line 5). If none are found, the contour variation is labelled as "No-Impact". The hyper parameter α serves as a calibration parameter, akin to the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) threshold in classification models. The hyper parameter $nOAR$ sets the model's overall sensitivity to dose violations.

Implementation details: Each of the two U-Nets in the cascade [177] has a depth of five sets of convolution layers in the encoder and decoder with $16, 32, 64, 128, 256$ channels in the first level and twice this number in the second. The model input is a normalized CT volume and binary segmentation masks for each of the 13 OARs and TV, and predicts a continuous-valued dose volume (up-scaled from [0, 1] to 0 to 70 Gray) of the same dimension as the input. The loss is computed as a weighted sum of L1 losses between outputs of the first and second U-Nets versus the reference dose: $Loss = 0.5 * L1(reference, A) + L1(reference, B)$, where A and B are the outputs of the first and second U-Nets respectively, $reference$ is the reference dose and $L1$ refers

Algorithm 1 Contour Variation Quality Classification - $Q(cv)$

for each OAR **do**
  if $max(D_{cv}) > max(D_{ref})(1 + \alpha T)$ **then**
    increment counter for Worse
  else
    if $max(D_{cv}) < max(D_{ref})(1 - \alpha T)$ **then**
      increment counter for Better
    end if
  end if
end for
if counter for Worse $\geq$ nOAR **then**
  $Q(cv)$ = Worse
else if counter for Better $\geq$ nOAR **then**
  $Q(cv)$ = Better
else
  $Q(cv)$ = No Change
end if

to the L1 loss. The weights are randomly initialized using the 'He' method. Training runs for 80000 iterations and the model with the best validation dose score is saved. The training batch size is set to 2 and the learning rate to $1e-3$ with a weight decay of $1e-4$. Data augmentation is done with random flipping and random rotation along the z-axis (in the axial plane). T in Algorithm. 1 is set to 0.1. All experiments are run with PyTorch 1.12 on an NVIDIA RTX A5000 GPU, and each training run takes approximately 24 hours.

Comparing against human expert baselines: For each test case, three expert radiation oncologists evaluated contour variations and were asked to classify them using the defined dosimetric categorization. We use 3D Slicer (version 5.6.0) for this evaluation and show all three slice planes (see Figure. A.7 in the Appendix). We also include a 3D rendering of the geometric relationship between the OARs and the tumour TV, highlighting where the contour change is made. We time their responses, and show all the variations for each subject simultaneously so that they can make visual comparisons against the reference for each contour variation.

As classification metrics, we report precision and recall, and the confusion matrices. Average time to evaluate each variant is also presented to compare performance.

5.3 Results

Table. 5.1 shows the weighted average (across the three categories) precision and recall as well as the average time taken to evaluate each of the 54 variants by the three radiation oncologists (in rows marked R1, R2 and R3) as compared to the DL dose prediction model in the last row. On both precision and recall, the model outperforms all three experts. Notably, we underline the high inter-rater variability in performance among the three experts. Expert R3, being the most meticulous and expert rater, used significantly more time than other experts. While the proposed model tends to classify more "No-Impact" contours as "Worse", we view this as a beneficial trade-off. In practice, it would lead to additional checks, which is preferable to potentially overlooking increased toxicity to the patient. The time taken by the model is dominated by two inference runs through the reference as well as the

variant contours. Additionally, the range of time taken varies broadly between experts, from 19 to 138 seconds per variation, while the DL predictor always takes the same quantum of time irrespective of the difficulty in geometry.

	Precision	**Recall**	**Time Taken (per variant)**
Radiation Oncologist #1	0.41	0.35	48 [19 - 64]s
Radiation Oncologist #2	0.48	0.46	50 [28 - 100]s
Radiation Oncologist #3	0.55	**0.57**	71 [29 - 138]s
Deep Learning Dose Predictor	**0.57**	**0.57**	30 s

TABLE 5.1: Precision and recall (weighted average) for each of the three expert radiation oncologists compared with model predictions. Average (max - min) time taken per variant evaluated is indicated in the last column in seconds.

FIGURE 5.3: Confusion matrices for the classifier using the dose predictor model versus the performance of three expert radiation oncologists. Sensitive predictions imply more entries in the upper triangular region, leading to further manual checks, while still saving clinician time for correctly classified variations (on the diagonal).

Figure. 5.3 shows the confusion matrices (normalized by true category) for the three expert radiation oncologists (R1, R2 and R3) and the dose predictor model (right-most panel marked “Prediction”). Darker green on the diagonal is better, while darker red on the off-diagonal is not. The model outperforms all three radiation oncologists in the “Worse" category. None of the experts mark any variant as “Better".

Figure. 5.4 shows the sensitivity of (weighted average) precision and recall to the hyper parameters α and $nOAR$ used to classify dosimetric impact. α (horizontal axis) ranges from 0.005 to 0.15 in each of the two heat maps. Specifically, $\alpha = 0.1$ means that the percentage change threshold for the model is 0.1 times that used for the reference (in this case, 1%). The vertical axis is $nOAR$, where smaller values make the model more sensitive and strict, and larger values increase model robustness while trading off sensitivity. Both the precision and recall metrics show a reasonably smooth variation, except that the precision values drop significantly for small α and $nOAR$. As good trade-off we chose $\alpha = 0.1$ and $nOAR = 3$ for the results presented in Figure. 5.3 and Table. 5.1.

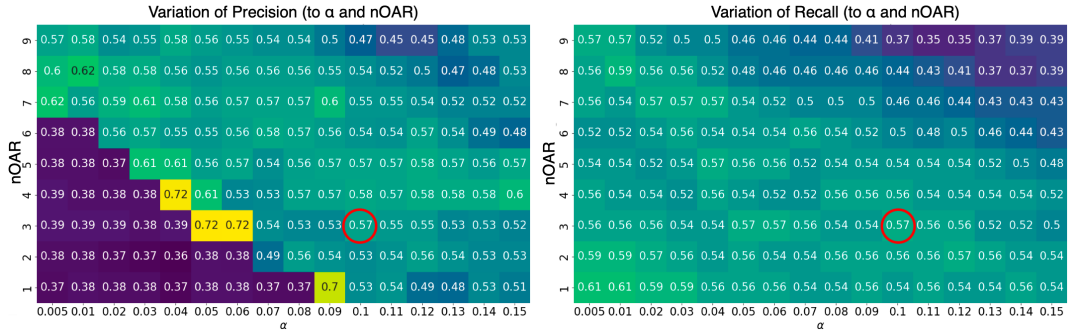


FIGURE 5.4: Performance of dose predictor model on variation of α and number of OARs crossing the threshold based on precision and recall. We prefer models with reasonable precision and higher recall - as we want the classification to be more sensitive in catching "Worse" plans as opposed to missing out on those that may have "No change". Red circles indicate values chosen for comparing with experts.

5.4 Discussion and Conclusion

Radiation oncologists indicate that their mental model emphasizes proximity to OARs (closer and in the line of sight between the tumour TV and OARs are more impactful) and the size of the variation (larger causes higher residual dose to OARs).

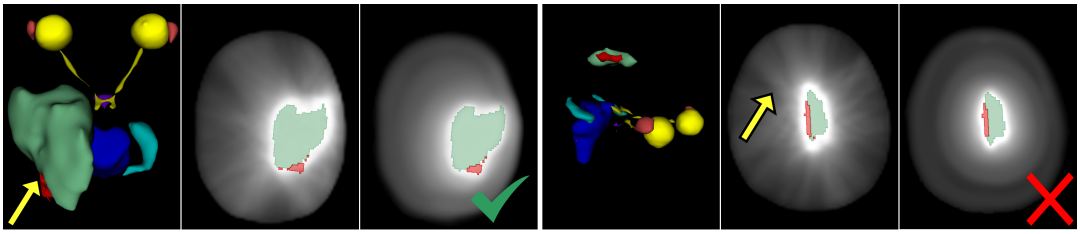


FIGURE 5.5: Two exemplar situations. Each set is shown as a 3D render, reference dose plan (axial) and predicted dose (axial). Left half: all three experts mark "No Change" due to its posterior nature (yellow arrow) away from the OARs, while the model predicts correctly. Right half: experts mark correctly as "No Change" while the model incorrectly flagged as "Worse". The yellow arrow shows the beam artifacts in the reference dose plan which are not replicated by our model.

Figure. 5.5 demonstrates two exemplar situations that we use to showcase the strengths as well as weaknesses of our proposed idea. The left half shows one such condition where the model correctly classifies a condition that conventionally would be considered to be not impactful. Conversely, the right half shows a situation where the model overestimates the severity. This can be attributed to the predicted dose being a smooth proxy to the actual beam structure. Recent advances [435] aim to account for this effect.

We present a novel dosimetry-driven quality control framework, where our dose predictor model outperforms human experts, indicates a promising baseline on which to build on improvements. This work demonstrates human clinician baseline, upon which we plan to work on the next set of evaluations where radiation oncologists

are shown assistive maps like [432] to measure if their performance improves both in time and accuracy.

Part II

Technical Investigations and Analysis

This part is focused on technical investigations and analysis. The articles included have been modified slightly from the published versions so as to not expand abbreviations already used.

Chapter 6 is an extended version of Chapter 4, with more extensive data variability in TV shapes and serves as a more comprehensive clinical validation for the C3D model [177] for the purpose of dose predictions to assist in contour QA. This article has a shared first-authorship with Dr. Robert Poel, where he contributed to the design of the experiments, generating variations of TV contours, and writing and reviewing for clinical correctness, and my contribution was extending the model presented in Chapter 4 for these experiments, organizing the data, re-training the models, generating figures, and reviewing the manuscript.

Chapter 7 is an extension of Chapter 8, and is currently under review. It includes more extensive experiments over six model architectures and four data sets with four segmentation metrics. Since the publication of these papers, the auto-contouring field has advanced significantly through several architectural modifications based on ViTs [279], "promptable" medical image versions of SAM [349], and most recently, the nnInteractive [307] set of models for both fully-automated, and semi-automated (with a rich set of user-interactivity tools) open-set automatic segmentation tools.

6

Deep-Learning-Based Dose Predictor for Glioblastoma–Assessing the Sensitivity and Robustness for Dose Awareness in Contouring

6.1 Introduction

Many cancers are currently treated by a combination of local and systemic therapy. Local therapy often consists of a combination of surgery and RT. The latter requires a sophisticated planning process to guarantee successful treatment. To improve the efficiency and quality of treatment planning in RT, methods for predicting possible dose distributions or DVH curves have been introduced in previous years [436–442]. Since 2012, KBP methods have been used to predict what is achievable in treatment planning. This was not only used for QA in treatment planning [443–445], but also to speed up the planning process by initializing the treatment plan based on the prediction [66]. By reducing the number of inputs from the user, treatment planning becomes much more consistent, more resource-efficient, and potentially beneficial for treatment quality.

In recent years, three-dimensional dose prediction through neural networks has been shown to be a viable method for this purpose. In 2016, the first study using neural networks for dose predictions was published by Shiraishi et al. [446]. In the following years, approximately 30 more studies that tried to predict the 3D dose distribution with DL were published. Most of these were based on treatments with a relatively standard target orientation, with minor anatomical variations from patient to patient, such as prostate and oro/nasopharyngeal cancers [447–450]. However, there are also promising results for models in the brain, breast, and lungs [436, 438, 451]. In 2020, the Open-Access KBP (OpenKBP) Challenge was organized, providing an open-access dataset of head and neck treatment plans to train prediction models and evaluate them on a set of standardized metrics [423]. A total of 195 participants competed in this challenge, where the best-ranked team scored a MAE of 2.43 and

1.48 for the dose and DVH scores, respectively (see Section 6.2.6). Their methodology is publicly available and described as a technical note [177].

Dose prediction models are mainly used for treatment planning. This means that, in practice, in addition to the dose prediction, a second model is required to convert the predicted dose into an actual plan that is executable for the specific treatment technique. In this latter step, a final optimization incorporates individual case properties, physical constraints, and dose delivery hardware [452]. In our case, we want to use the dose prediction model for another purpose, namely contour evaluation.

The contouring of TVs and OARs, the step that takes place prior to planning, is also subject to automation to improve efficiency and consistency with respect to the current manual process. With the implementation of AI, it is important to have specific tools for QA [238]. To ensure quality, an assessment of the contours is required. Usually, visual inspection is the go-to method; however, this is a time-consuming task. For each TV and OAR, every image slice needs to be visually inspected. If necessary, manual adjustments need to be made if a contour is deemed incorrect. Especially for DL-based auto-segmentation models, a lack of robustness could result in unpredictable errors that can happen anywhere within the image volume [453]. There are, thus, good reasons to automate this QA step as well. There have been several attempts to provide an automatic assessment of automatic segmentations. Most of these were based on geometrical prior knowledge [454–456] , but they lacked certainty for proper QA [457]. Moreover, more recent DL-based approaches based on uncertainty maps seem to have issues [458]. To complement the current work performed on QA for auto-segmentation, we postulate that it would be beneficial to know the possible clinical impact of a specific segmentation. A DL model that can give an accurate prediction of the dose received by an OAR instantly could provide the required information to assess the clinical impact of contour variations. Such dose awareness, for the evaluation of contours, can improve the efficiency of the process. Moreover, it provides clinically relevant feedback to the evaluator, who can then focus on potential segmentation errors that will have a larger impact on the treatment.

To assess the feasibility of a near-instant dose prediction model in providing dose awareness for the evaluation of auto-segmented contours, we made use of a DL model to predict the dose for GBM cases. Based on the network of Liu et al. [177] that was used in the OpenKBP challenge, a model was trained on a set of curated GBM cases. Unlike with current dose prediction algorithms, we wanted to verify the model's performance for contouring QA. This means that, beyond specific accuracy and sensitivity, robust predictions for a broad range of situations are required. To do so, we tested our trained model on specific sets of contour alterations to assess its sensitivity. Furthermore, we stress-tested the model by using a specific worst-case test set, including rare cases where we expected it to fail. This enabled us to determine the robustness of the model and understand where further improvements are required. Subsequently, based on the outcome observed on the worst-case test set, we improved the robustness of the model by augmenting the training set with synthetically generated cases characterizing the observed failure patterns.

6.2 Materials and Methods

6.2.1 Data Collection and Preparation

Imaging data from a cohort of 125 GBM patients treated with RT at the Inselspital University Hospital (Bern, Switzerland) were available. For all patients, the PTV and

the OARs were curated by a mutual agreement between three radiation oncology professionals. A plan was constructed using a strict dose prescription and standard templates for planning setup and dose optimization initiation for all cases. Of the first 95 cases, 60 were selected for model training, 15 were chosen as validation, and 20 were used as a test set. Of the remaining 30 cases, 10 were used to construct a worst-case test set manually, and the other 20 for improving the training by adding specific out-of-distribution cases. Section 6.2.5 further details how the worst-case test set and the out-of-distribution cases were designed.

6.2.2 Dose Planning

All cases were planned according to the clinical dose prescription of 60 Gy in 30 fractions based on the ESTRO-EANO guidelines [63] in the Eclipse TPS V15.06.05 (Varian Medical Systems, Palo Alto). All OARs were subject to a dose constraint, which, according to a priority list, could or could not be compromised (Table 6.1). All plans used a VMAT with a double full co-planar arc with 6 mega volt beams containing a flattening filter. The plans were optimized with the photon optimizer, and doses were calculated with the AAA [424]. After dose calculation, the dose was normalized so that 50% of the PTV was covered by 100% of the prescribed dose, according to the institutional clinical guidelines.

6.2.3 Training

The planning CT and the structures were available in Digital Imaging and Communications in Medicine (DICOM) format for each case. All data were converted from DICOM to NIfTI files, using the PyRaDise package [433]. The Radio Therapy Structure Set (RTSS) files containing the PTV and the OARs were divided into 14 separate 3D binary masks, each containing a single structure. The input files consisted of 16 3D volumes per case: the planning CT, the dose distribution, the PTV binary mask, and 13 OAR binary masks (Figure 6.1).

We trained a two-level C3D U-Net [177] as the dose prediction network (i.e., the input to the second U-Net is the output of the first, concatenated with the input to the first U-Net). The U-Net is the most commonly used DL neural network for dose prediction. It was first used in this context as a 2D network by Nguyen et al. [459]. Since then, many advances have been made, notably an extension to a 3D U-Net [460]. The C3D model was proposed to incorporate both global and local anatomical features for dose prediction and showed the best results among all the competing model architectures in the OpenKBP challenge [177, 423].

The model input was a normalized CT volume and binary segmentation masks for each of the 13 OARs and TV. As output, the model predicted a continuous-valued 3D dose (upscaled from $[0,1]$ to $[0,70]$Gy to normalize to the full range of the dose within the cohort) of the same dimension as the input. The loss was computed as follows: $Loss = 0.5 * L1(reference, A) + L1(reference, B)$ where A and B are the outputs of the first and second U-Nets, respectively. In this equation, reference indicates the reference dose, and L1 refers to the L1 loss. All volumes were resampled to 128^3 voxels due to GPU memory constraints. The hyper parameters for training the C3D model were unchanged from the original implementation [177], except that the number of input binary masks was updated to 14 to match the number of structures in our dataset. The model's weights were randomly initialized using the "He" method [461]. The training process ran for 80,000 iterations, and the model with the best validation dose score was saved. All experiments were run with PyTorch 1.12 on an

TABLE 6.1: Clinical dose-planning guidelines for GBM treatment.

OAR	Constraint	Priority
Brain-PTV	• $V_{60\,\mathrm{Gy}} \leq 3\,\mathrm{cc}$	2
Brainstem	• $D_{0.03cc} \leq$ 60 Gy (hard constraint)	1
	• $D_{0.03cc} < 54\,\mathrm{Gy}$	4
Chiasm	• $D_{0.03cc} \leq$ 54 Gy (hard constraint)	1
	• $D_{0.03cc} \leq 50\,\mathrm{Gy}$	3
Cochlea (Ipsi-lat)	• $D_{\mathrm{mean}} \leq 45\,\mathrm{Gy}$ (< 30% hearing loss)	5
	• $D_{\mathrm{mean}} \leq 32\,\mathrm{Gy}$ (< 20% tinnitus)	9
Cochlea (Bi-lat)	• $D_{\mathrm{mean}} \leq 45\,\mathrm{Gy}$ (< 30% hearing loss)	7
	• $D_{\mathrm{mean}} \leq 32\,\mathrm{Gy}$ (< 20% tinnitus)	9
Hippocampus	• $D_{\mathrm{mean}} \leq 30\,\mathrm{Gy}$ (< 30% IQ loss)	8
	• $D_{0.03cc} \leq 30\,\mathrm{Gy}$	14
	• $D_{40\%} \leq$ 7.3 Gy (long-term NCF)	11
Lacrimal Gland	• $D_{\mathrm{mean}} \leq$ 25 Gy (clinic) (hard constraint)	1
Lens	• $D_{0.03cc} \leq 7\,\mathrm{Gy}$ (< 25% cataract)	12
Optic nerves (Ipsi-lat)	• $D_{0.03cc} \leq$ 54 Gy (hard constraint)	1
	• $D_{0.03cc} \leq 50\,\mathrm{Gy}$	3
Optic nerve (Bi-lat)	• $D_{0.03cc} \leq$ 54 Gy (hard constraint)	1
	• $D_{0.03cc} \leq 50\,\mathrm{Gy}$	6
Pituitary	• $D_{\mathrm{mean}} \leq$ 45 Gy (panhypopituitarism)	10
	• $D_{\mathrm{mean}} \leq$ 20 Gy (growth hormone deficiency)	13
Retina	• $D_{0.03cc} \leq$ 45 Gy (hard constraint)	1
Target	**Objective**	**Priority**
PTV	• $D_{90\%} > 57\,\mathrm{Gy}$ (95%)	1
CTV	• $D_{100\%} > 60\,\mathrm{Gy}$ (100%)	2
PTV	• $D_{0.03cc} <$ 64 Gy (107%)	3

NVIDIA RTX A5000 GPU. We trained the model five times with the same hyper parameter set but a different random seed initialization to ensure reliable convergence. Each training run took 24 hours. A single inference to run the dose prediction model takes about 1 to 2 minutes on a standard consumer computer and about 15 seconds

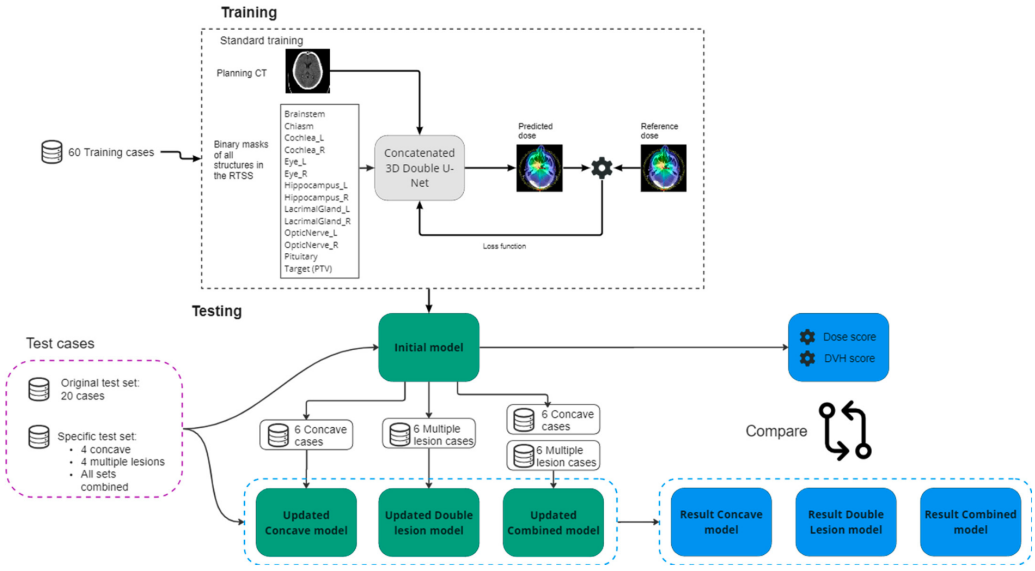


FIGURE 6.1: Schematic overview of the training and testing process. The upper block represents the training procedure of the initial model with its inputs and outputs. The initial model is tested on the training procedure of the initial model with its inputs and outputs. The initial model is tested on the test cases, resulting in dose and DVH scores for each test set. The initial model (green block) is updated threefold with concave cases, multiple lesion cases, and a combination of the two. The updated models are tested on the same test sets. The results are then compared (blue blocks).

with a GPU.

6.2.4 Assessing the Model's Sensitivity

One of the goals of the dose predictor is to provide realistic dosimetry information based on the input contours. It should additionally be able to predict realistic dose changes produced by small and realistic changes to the contours of an organ (i.e., inter-expert variability [462]). To analyse the sensitivity of the dose prediction model to these changes, a specific case was chosen where the GBM TV is near the Optic Nerve Left (ONL). In practice, the optic nerves are prone to variability in contours due to the intra- and inter-fractional movement of the eyes which also affect the optic nerves. In this case, small changes in the contour of the ONL would lead to significant dose changes. To simulate this situation, ten alternative contours of the ONL were manually drawn. The dose was re optimized and recalculated on the TPS for each alternative contour in order to serve as a reference dose. The reference doses were then compared qualitatively and quantitatively to the doses predicted by the model. The DSC for the alternatives were calculated to correlate the dose differences to the geometric discrepancy of the ONL contours.

6.2.5 Improving the Model–Worst-Case Test Set

To assess the robustness of the model, a worst-case test set was selected. Based on the analysis and evaluation of the dose score results on the standard test set and the statistical analysis of the normal distribution of the training set, a number of test cases were defined where we expected the model to fail or have difficulties. The PTVs of

these cases were manually manipulated to simulate rare cases not described by the training dataset (Out Of Distribution (OOD) cases), as well as to present a challenge in terms of the physical limitations of obtaining perfect dose conformity. Among these 10 cases, we included (i) TV of larger and smaller size than those present in the training set; (ii) TV consisting of multiple lesions; (iii) irregular shapes, such as elongated or concave TV; and (iv) TV that present an overlap with the OARs.

According to the worst-case test set results, we gained insight into which situations the model performs poorly in and where it could benefit from additional training. Our observations showed that the prediction model struggled mostly with the physical limitations of conforming the dose according to the TVs outlined for specific shapes. Where conformity of the actual dose prediction decreases with concave shapes or multiple TVs close to each other, the dose predictor overestimates the dose fall-off in these regions. To increase the overall robustness of the model while also improving the model for these situations, we updated the trained model by including a set of concave-shaped TV cases and a set of cases where the TV consists of multiple lesions.

The respective new training sets were constructed by manually adjusting the TV (Figure 6.2). The ten cases used for both sets are from different patients and were used in any previous model training. All new cases included dose planning according to the same protocol described above in order to serve as the reference dose.

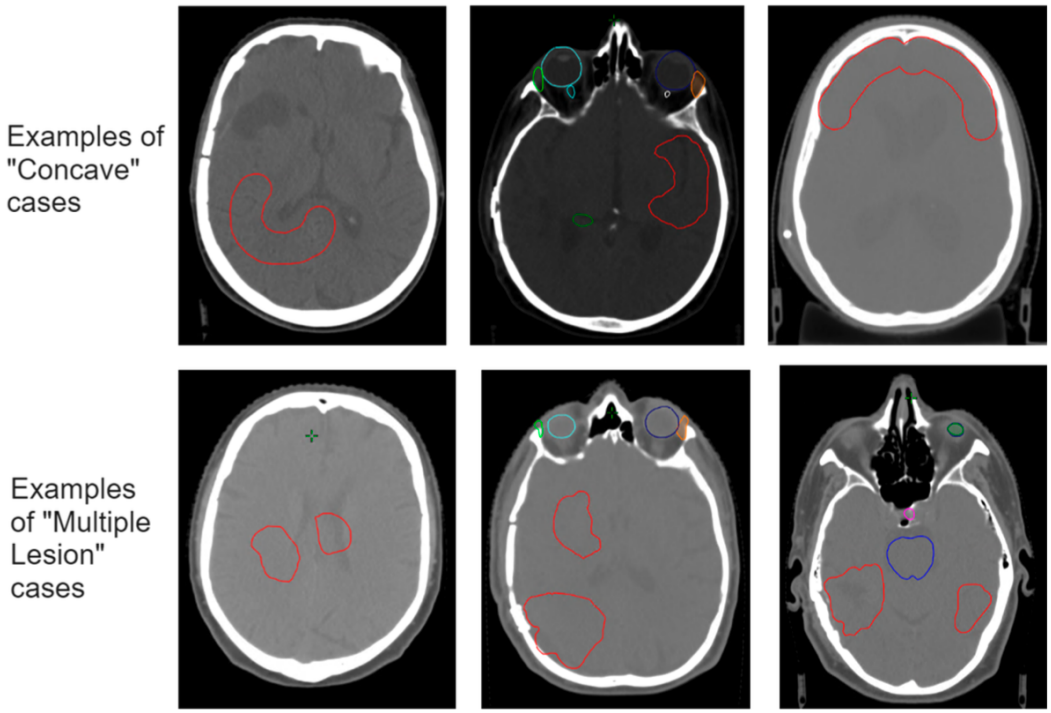


FIGURE 6.2: Examples of the additional training cases for the concave TV (above) and the multiple TVs (below). The TVs are drawn manually in red and do not represent actual tumour situations. The structures in other colours represent OARs.

First, we trained an updated "Concave Model" with 60 standard cases + 6 concave cases. The remaining 4 concave cases were used as the test set. Second, we trained an updated "Multiple Lesion Model" with the 60 standard cases + 6 multiple lesion cases. Again, 4 cases were used as a test set. Finally, we retrained the

initial model with 60 standard cases + 6 concave cases + 6 multiple lesion cases. We tested these models on the standard test set, as well as these specific additional test cases, and compared this with the results of the initial model. An overview of the experimental setup is given in Figure 6.1.

6.2.6 Evaluation

The trained models were evaluated on the test set, and the prediction of the dose was compared to the actual planned dose by means of the standardized metrics used by the OpenKBP challenge [423]: the dose score and the DVH score. The dose score measures the mean error over all the voxels between the two 3D volumes. In this case, we used the whole brain to measure the dose score instead of the whole CT volume or whole body volume. Taking a larger volume dilutes the results to a more positive outcome. The DVH score is the mean error over a set of criteria specific to the given volume. For OARs, these criteria are the mean dose (Dmean) and the maximum dose to 0.1cc (D0.1cc). For the TV, the criteria are the dose received by 1%, 95%, and 99% of the voxels within the volume (D1, D95, and D99). The DVH score is calculated for all OARs used in training (lens and retina are combined within the eye, since overlapping masks were not possible). We report the mean results for the five trained models and use one of them for a subsequent sensitivity analysis.

Additionally, the initial trained model and the updated models were tested on a set of concave TV cases, a set of multiple lesion cases, and a combined test set that included both plus the standard test set.

6.3 Results

Based on the initial training set of 60 cases, the performance of the model was determined based on the standard test set of 20 cases. The mean result over five independently trained models showed a dose score, which was measured over the whole brain volume of 0.94 (stdev = 0.36). The mean DVH score over all OARs and the TV was 1.95 (stdev = 0.95).

6.3.1 Results for Sensitivity

An overview of the nine alternative left optic nerve contours is shown in Figure 6.3. The mean dose to the ONL, based on the TPS, and the mean dose based on the prediction model for the reference and the nine alternative contours are shown in Table 6.2. There is a reasonable variation in the mean dose among the different alternative contours with respect to the reference contour. In some cases, only minor changes to the mean dose occur, even though the DSC metric shows a significant difference in contour similarity. In other cases, the mean dose change with respect to the reference contour can be as high as 5 to 7 Gy. The difference between the calculated dose and the predicted dose seems to follow a trend and varies under a maximum of 3.50 Gy, with a mean of 1.38 Gy. This shows that the predicted dose is more often overestimated.

The average difference of the calculated mean dose for the alternatives with respect to the calculated reference mean dose was 2.44 Gy (i.e., the difference between the alternatives to the reference dose). For the predicted dose, this difference was 2.32 Gy. This means that the correlation coefficient between reference and predicted dose differences across the contour alternatives was 0.89, while the correlation coefficient with the DSC was -0.42 [101].

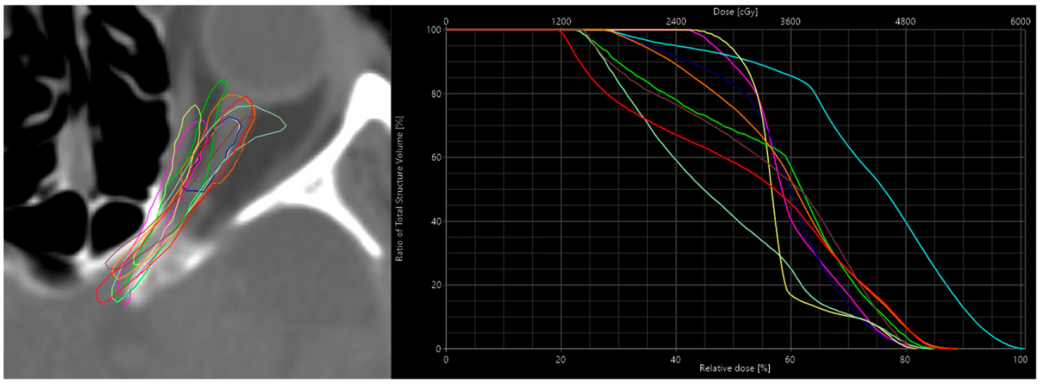


FIGURE 6.3: On the left is an overview of all the 9 alternative contours of the ONL. On the right, the dose's respective DVH curves are calculated with the TPS, which shows the dose's respective DVH curves are calculated with the TPS, which shows the variation in the dose these contours have. The colours in the DVH curve correspond to the colours of the contours on the left.

TABLE 6.2: Predicted mean doses in Gy for the different ONL contours.

ONL Contour	Calc. Dose	Pred. Dose	Δ Calc-Pred	DSC	Δ to Calc-Ref	Δ to Pred-Ref
Reference	34.7	35.5	-0.8	n.a.	n.a.	n.a.
Alternative-1	32.2	35.7	-3.5	0.31	-2.5	0.2
Alternative-2	30.7	32.4	-1.7	0.26	-4	-3.1
Alternative-3	34.2	34.5	-0.3	0.63	-0.5	-1
Alternative-4	31.8	34.1	-2.3	0.59	-2.9	-1.4
Alternative-5	26.9	30.1	-3.2	0.51	-7.8	-5.4
Alternative-6	32.8	36	-3.2	0.20	-1.9	0.5
Alternative-7	41.8	41.2	0.6	0.16	7.1	5.7
Alternative-8	35.3	33.1	2.2	0.58	0.6	-2.4
Alternative-9	34.5	36.1	-1.6	0.05	-0.2	0.6
Mean	33.49	34.87	-1.38	0.37	Corr. Coeff.: 0.89	

6.3.2 Improving the Model–Worst-Case Test Set

While analysing the results of the worst-case test set, we saw flaws, particularly in cases where TVs have concave shapes and consist of multiple lesions (see Figure 6.3). In such cases, the prediction model overestimated the ability to conform the dose to the TVs. This is mainly reflected in higher dose scores and less so in the DVH scores of the TV since the dose discrepancy occurs just outside of the TV structure. We updated our training data with six concave TV cases and six multiple lesion TV cases and a combination of both. The results for the different test sets are given for the dose score, the DVH score for the OARs, and the DVH score for the TV separately in Table 6.3.

Based on the standard test set, the updated models scored similarly to the initial model on the dose score and the DVH score for TVs. The DVH score for OARs improved for all updated models. The updated model with the concave TV cases shows, by far, the best results of all trained models.

Focusing on the concave updated model, we observed improvements in the dose and DVH score on the concave test set, as well as improved results on the multiple test set. This means that, for this this small set of specific cases, the concave updated model scores significantly better than the initial model. For the multiple-lesions updated model, we also observed an improvement, but to a lesser extent. The combined updated model, containing both concave and multiple TV lesions, scored the worst of all updated models. For the standard test set, the combined updated model did not show improved scores with respect to the initial model.

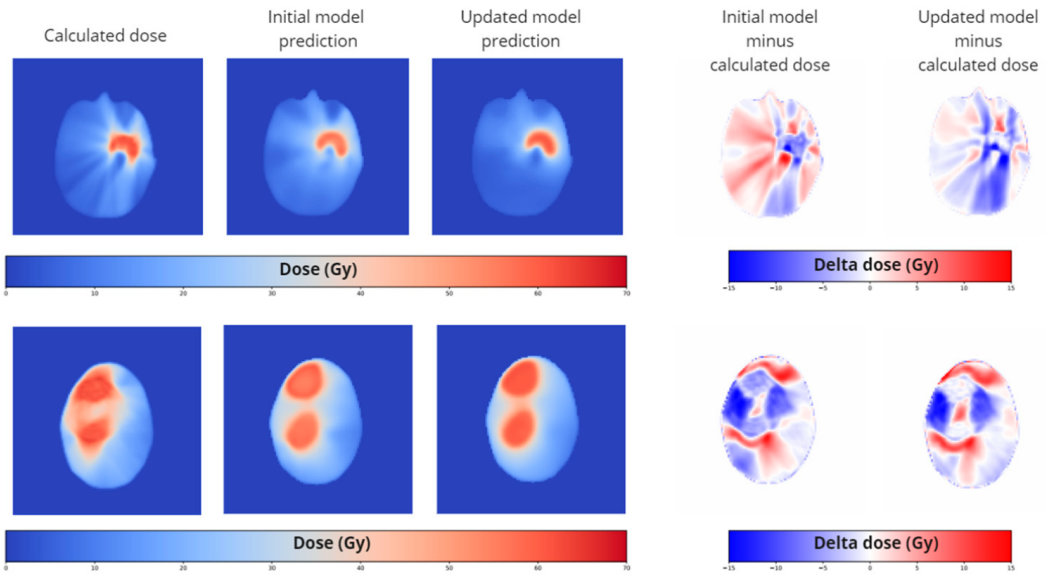


FIGURE 6.4: Dosimetric comparison of the calculated dose, the initial prediction model, and the updated model for a concave (above) and a multiple lesion case (below). The images represent a single axial slice. On the right, the dose difference maps of the corresponding axial slice are given. The difference between the latter shows improvements in the dose prediction. The depicted cases were not used in the training of the initial model or the updated model.

For the combined test set, which is a combination of the three previous test sets, the updated model scored consistently better than the initial model.

Qualitatively, we can see an improvement in the spatial distribution of the predicted dose in the updated models at exactly the locations of concern, the concave parts of the TV, and the space between multiple lesions, especially in the axial direction.

TABLE 6.3: Results of the dose score and DVH scores of the initial and the updated dose prediction models. Lower values represent better scores.

Test Set	Initial Model	Concave Updated Model	Multiple Lesion Updated Model	Combined Updated Model
Dose scores' whole brain volume				
Standard test set	0.94	0.94	0.92	0.98
Concave test set	0.87	0.81	0.81	0.87
Multiple test set	1.30	0.84	1.24	1.02
Combined test set	0.98	0.90	0.95	0.97
DVH scores' OARs				
Standard test set	2.01	1.73	1.85	1.89
Concave test set	2.11	1.67	1.99	2.08
Multiple test set	3.05	1.86	3.05	2.67
Combined test set	2.18	1.74	2.04	2.03
DVH scores' TV				
Standard test set	1.19	1.12	1.20	1.26
Concave test set	1.72	1.67	1.51	1.66
Multiple test set	3.62	1.92	3.18	2.91
Combined test set	1.61	1.31	1.53	1.55

6.4 Discussion

By means of an existing dose prediction model that was trained for head and neck cases, we obtained good results for translating the dose prediction model to GBM cases in the brain. For these initial results, only 60 cases were used for training. Compared to the results of the OpenKBP challenge, using their proposed metrics, our initial trained model scored better values than the top ranked participants did. However, we are aware that the anatomies of the evaluated treatment sites are different. It must be noted that different 3D volumes are used, as well as different OAR structures, which will have an influence on the used scores. We do not yet have a benchmark for these metrics in the brain region. On the other hand, whereas head and neck TVs are much more similar from case to case, GBM TVs vary much more in size, shape, and location. Nonetheless, the model was able to achieve good results on a relatively small training set. The overall dose score was less than 1 Gy. Although this might be relevant in a final dose distribution, for the dose predictor model and the purposes we aim to use it for, we believe this to be of sufficient accuracy. Although the prediction model might have struggled in reproducing the exact streaking patterns of the actual calculated dose, this did not have an impact on the relevant metrics, such as the mean and maximum dose received by the specific OARs. We conclude from this that a C3D U-Net is capable of predicting the dose when provided with high-quality and well-curated data to train it.

Although some other groups have published on dose prediction in the brain [442, 446, 451], for both VMAT and conventional IMRT, they did not specify the nature of the brain tumours. Furthermore, different treatment prescriptions were included in these works, certain tumour locations were excluded, and additional planning parameters were used in the training model. This is the first dose prediction model

specified for GBM treatment.

It seems that 3D dose distribution prediction is a great application for deep-learning models. Although the predictions are not perfect, they are useful in the initialization processes of automatic planning. It also provides a near-instant estimation of the dose distribution that outperforms any currently available analytical or mathematical prediction method. Our work shows that a dose prediction model for a specific purpose, i.e., treatment area and modality, is relatively easy to obtain. For our specific purpose, we wanted to predict the dose to OARs. We were able to improve the accuracy of the DVH score for the OARs with some minor additions to the training set.

Considering it is a 3D U-Net, the training times can be relatively long. In our scaled-down version (128^3 voxels), a single training course lasted around 24 hours. Adding more training data, for instance, doubling or tripling the amount, would be feasible. In addition, increasing the resolution to 256 voxels would improve dose prediction resolutions at the cost of longer training and inference times. On the other hand, the fast inference time of only seconds on a high-end GPU enables multiple applications where the inference is run in a high-throughput manner. Some examples where a multitude of dose predictions are used are to show the dose effect of atomic surface transformations in OARs in the brain [432] or for a clinically relevant guidance of the loss function in automatic segmentation of TVs for GBM [463].

Furthermore, we did not analyse the interpretability of the model. It is, however, feasible to obtain saliency maps for the outcomes of the dose prediction model since it is based on the standard U-Net architecture [339].

To obtain the training data, we used plans that were made according to a strict and standardized protocol. This makes the resulting dose distribution better to predict. It might therefore be one of the reasons for the good results. On the other hand, using such a strict protocol makes the model only valid for treatments following this strict protocol. However, in clinical practice, different approaches are used depending not only on case specifics but also on the individual preference of doctors, planners, and the availability of specific hardware. In this case, we used a VMAT technique. The dose distribution of such plans is more predictable since, in every case, two full 360-degree co-planar arcs are used. In other techniques, such as conventional IMRT, using a set number of beam angles, or in more sophisticated VMAT techniques making use of non-coplanar arcs, the dose might be more difficult to predict. To solve this issue, one needs to make a dose prediction model for every different treatment strategy. Although this seems cumbersome, in clinical reality, this comes down to a few treatment strategies per treatment site for any department. Given our model's results, only a limited number of data are required to obtain a viable model.

The main contribution of this paper was to show the feasibility of training an accurate deep- learning-based dose predictor for GBM treatment. If data are limited for a particular scenario but a demarcated treatment protocol exists, even for non-homogeneous anatomies (i.e., not prostate or head and neck, for which most of the dose prediction methods are proposed), satisfactory results can be obtained compared to currently reported outcomes. Although the main drive behind dose prediction models is the purpose of automation in dose planning, dose prediction models can also be important for many other purposes. Our hypothesis is that they can be useful in the quality management of the RT steps that take place prior to planning. There are possible positive consequences for the effectiveness of QA, the accuracy of segmentation, and the optimization of treatment planning. By introducing clinically relevant objectives early in the radiotherapy planning process, we can

improve automation and enable automated QA. This makes the RT process faster and also improves the quality.

We tested the obtained dose prediction model for specific criteria important for a dose predictor in quality management, sensitivity, and robustness. We showed that the initially trained model is sensitive enough to detect dose trends on realistic contour variability in a critical OAR such as the ONL. We also tested the initial model against robustness. Although we found that the model did lack a certain accuracy in specific situations, we showed that with a simple strategy of adding specific cases to the training set, the robustness and the overall accuracy of the model increased. We anticipate that dose prediction models can be more accurate when using larger datasets of carefully curated data. In addition, the models can be tailored to have specific characteristics to fulfil the needs of different tasks in RT management.

6.5 Conclusions

This manuscript showed that obtaining a dose prediction model for GBM VMAT treatment that is both robust and sensitive to realistic segmentation variations is possible with less than 100 cases. Although the model is only valid for one specific treatment strategy, the relative ease of implementation and the near-instant inference time make it a useful technique to incorporate dose awareness into the automation processes in RT prior to planning.

7

The impact of U-Net architecture choices and skip connections on the robustness of segmentation across texture variations

7.1 Introduction

The U-Net, introduced by Ronneberger et al. [265], has become a widely used architecture for image analysis, particularly in medical image segmentation. Since its introduction, research on its applications and improvements has grown rapidly, with nearly 3,000 articles published in 2022 alone [464]. A key feature of the U-Net is its "skip connections," which play a crucial role by directly connecting corresponding layers between the encoder and decoder, providing an alternative path for data flow instead of going through the backbone containing convolution layers. This bypass of additional processing allows the network to combine high-resolution details with broader contextual features, with the intent of more accurate segmentation by preserving spatial information lost during down sampling and enhancing the model's ability to capture fine-grained details. It has been reported that skip connections can improve training speed and model performance by enabling better gradient flow. However, it was also observed that these connections can sometimes reduce performance in shallow networks, where all layers are already well-optimized [465].

Over the years, more than 100 variations of the U-Net architecture have been developed, mainly improving skip connections, backbones, and bottlenecks [464]. Among the skip connection improvements, U-Net++ [288] stands out for using dense skip connections, which help blend features more smoothly between the encoder and decoder. Another example is the AGU-Net [278], which applies an attention mechanism to highlight only the most important features before passing them to the decoder, improving segmentation accuracy. For backbone changes, V-Net [286] replaces max-pooling with strided convolutions, reducing memory usage while still capturing important details during training.

While skip connections are known to preserve spatial detail and improve training convergence, their influence on model robustness has not been sufficiently explored. From an information theory perspective, they act as information highways, preserving both useful signal and noise. The information bottleneck principle [466] suggests that effective learning requires compressing input to retain only task-relevant information. Under noisy input conditions, dense skip connections may conflict with this by reintroducing these redundant high-entropy, high frequency features to re-enter the decoder. By bypassing deeper layers, they may limit the network's ability to abstract away noise and other irrelevant features, especially in tasks where Foreground (FG) and Background (BG) image textures are similar. As a result, such architectures trade-off the benefits of detail preservation to degraded robustness due to higher sensitivity to Texture Similarity (TS) or domain shifts. In contrast, models without skip connections rely on deeper processing paths that could naturally attenuate such artifacts.

In the original U-Net, skip connections pass feature maps from the encoder to the decoder by concatenation, meaning the decoder has more feature channels than the encoder. In contrast, V-Net merges skip connections using element-wise addition, which keeps memory use lower while still preserving important information from earlier layers. Despite these architectural choices, Isensee et al. [285] showed through their nnU-Net study that the success of U-Net often depends more on a well-designed data pipeline, including proper data normalization, class balancing, and preprocessing, than on changes to the network itself, providing further evidence of this in Isensee et al. [467]. This raises questions about whether modifying skip connections and other network features is truly the key to progress, especially for medical imaging applications where accuracy and reliability are crucial.

Currently, the most widely used image segmentation models are based on the U-Net architecture, achieving accuracy levels comparable to the variability between experts on unseen tasks [468]. As a result, the focus of future developments has shifted towards improving the robustness and reliability of these models. Galati et al. define robustness as "the degree to which a system can function correctly in the presence of invalid inputs" [469]. In clinical settings, factors like patient movement, disease- or hardware-related image artifacts, and other disruptions can challenge the robustness of deep-learning segmentation models. Additionally, domain shifts caused by vendor choice, imaging protocols, or operator techniques can negatively affect the performance of models trained on well-curated datasets. For clinical applications, ensuring robust and consistent performance is essential [470, 471]. For safety-critical uses like medical imaging, one way to evaluate robustness is by testing models trained on clean data against corrupted data, reflecting the diverse challenges of real-world conditions.

Kamann and Rother [472] reported that the robustness of image segmentation models varies significantly depending on the type of corruption in the test data. They suggested this could be due to specialized modules within the model architecture. In medical imaging, such distribution shifts and corruptions are often specific to the imaging acquisition process or modality [473]. While the average accuracy of U-Net-based models has steadily improved over time, their robustness and reliability have not kept pace [309, 474]. It remains unclear how the choice of U-Net architecture, mainly skip connections, influences model robustness. Unlike previous studies that focus on the effect of data perturbations on robustness, our work examines the impact of architectural choices and skip connections.

Human performance in image classification tasks is primarily shape-driven, but

Geirhos et al. [475] showed that convolutional neural networks tend to rely on texture information. This texture bias also applies to segmentation tasks, as demonstrated by Sheikh and Schultz [476], who improved performance by augmenting training images with feature-preserving smoothing, while Chen et al. [477] propose a novel framework that separates shape from texture. You and Reyes [339] found that texture and contrast perturbations significantly affect the performance and robustness of U-Net models, and Ouyang et al. [478] resorted to augmentation approaches synthesizing texturally diverse samples to improve robustness. Building on this prior art, we focus on investigating how texture influences the robustness of image segmentation models. Specifically, we develop a novel framework to evaluate the robustness of U-Net-like models to TS applied to the FG and BG of segmentation masks.

We hypothesize that although skip connections provide the benefit of passing high-resolution detail directly from the encoder to the decoder with the intent of improving performance, they could also transfer noise, negatively impacting robustness. Architectural innovations like attention gates in the skip connections could amplify not only the detail but also the noise from the encoder branch, especially in regions of high entropy. We parametrize the amount of necessary detail versus noise by selectively analysing TS between FG and BG, showing large variations in medical images. Given that U-Nets are often used as "Swiss-army knife" architectures across various image modalities and quality levels, we analyse how the skip connection density relates to different texture similarities. Our study examines this by exploring three levels of skip connection density: enhanced, standard, and complete removal. This is done through controlled experiments on synthetic images with varying textures and clinical datasets, including Ultra Sound (US), histology, CT, and MRI.

This study builds on the work of Kamath et al. [479], which introduced a novel analysis pipeline to evaluate the robustness of image segmentation models based on texture differences between FG and BG. The initial findings suggested that removing skip connections could improve model robustness, with examples of specific failure modes. In this work, we expand upon these contributions by:

1. **Broader Architectural Evaluation:** In addition to U-Net, AGU-Net, and NoSkipU-Net, we evaluate UNet++ [288], V-Net [286], and a "No-Skip" V-Net to understand better how skip connection enhancements and backbone changes affect robustness. Our findings consistently show that reducing the number of skip connections results in more robust and reliable models when tested on perturbed data.

2. **Extended Synthetic Experiments:** We explore generalized cases of BG blending into the FG and FG blending into the BG, revealing asymmetric behaviour across different architectures. These experiments show that BG blending into the FG is more challenging, with notable changes at varying levels of texture difficulty.

3. **Inclusion of Histology Images:** We expand our datasets to include histology images as a new medical dataset and report training durations for each model dataset combination. Results reveal up to 100x differences in training times between model architectures using standard implementations from Medical Open Network for AI (MONAI) [480].

4. **Expanded Medical Image Testing:** We test five levels of image perturbations, resulting in 120 testing scenarios. These experiments confirm that models with

fewer skip connections consistently perform more robustly and reliably across various perturbation levels.

The rest of this paper is structured as follows: Section 7.2 describes the methods and experiments, Section 7.3 presents the results, and Section 7.4 provides a discussion.

7.2 Materials and Methods

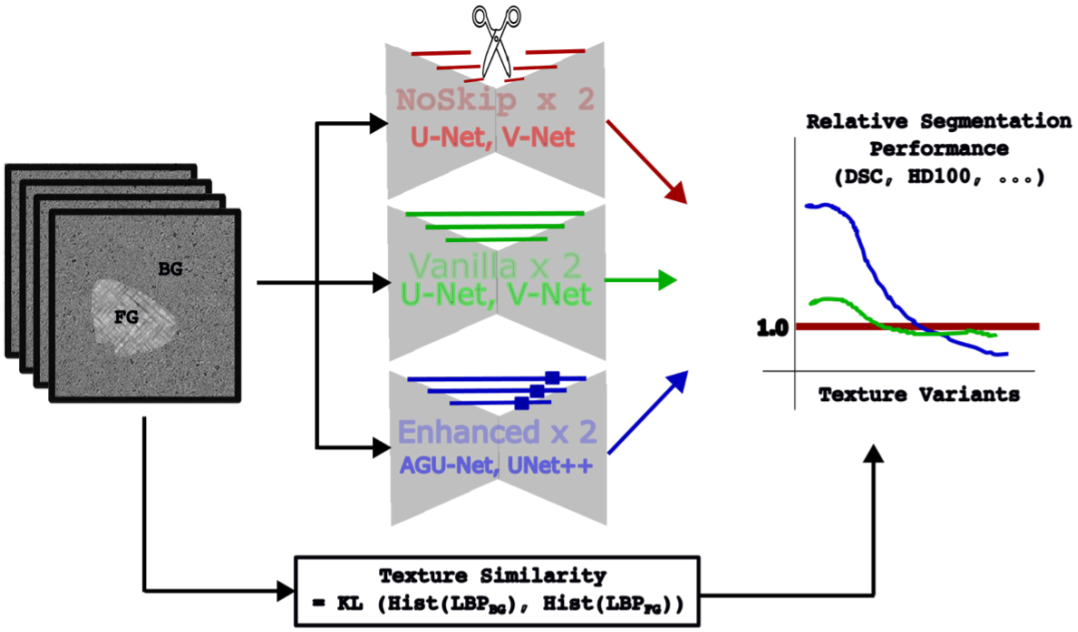


FIGURE 7.1: To evaluate the role of U-Net skip connections under varying levels of textures, defined by the similarity between FG and BG textures, we trained six U-Net variants. These included models without skip connections (NoSkipU-Net, NoSkipV-Net), standard architectures (U-Net [265], and V-Net [286], and enhanced models (AGU-Net [278], and UNet++ [288]). Each architecture represents a different strategy: no information transfer via skips, direct information transfer (identity transform), and selective filtering of skip information. Training datasets were created with controllable FG and BG TS, measured using the Kullback–Leibler (KL) divergence of Local Binary Pattern (LBP) histograms. Models were trained across multiple levels of FG-to-BG TS and evaluated on unperturbed textures and four levels of perturbed textures. For each condition, the performance and robustness of the models were assessed, with segmentation metrics calculated both absolutely and relative to the NoSkipU-Net. This analysis allowed us to compare how different skip connection strategies impact model behaviour across texture complexities and perturbations.

7.2.1 Experiment Design

Figure 7.1 describes our experimental setup to assess the impact of skip connections in U-Net-like architectures under varying levels of TS. Given a set of N pairs of

labelled training images $\{(I,S)_i : 1 \leq i \leq N\}$, $I \in \mathbb{R}^{H\times W}$ and $S \in \{0,1\}^{H\times W}$, corresponding ground truth segmentation, a DL segmentation model $M(I) \mapsto S$ is commonly updated by minimizing a standard loss term, such as the binary cross entropy or dice loss. To evaluate how the models behave at varying TS, we construct training data sets where each training sample is subjected to a linear transformation (i.e., perturbation) where its FG is blended with the BG: $I(x|Z(x) = 1) = \alpha I(x|Z(x) = 1) + (1-\alpha)I(x|Z(x) = 0)$. To investigate any asymmetry in behaviour, we also inverted the direction of the blending, with the BG progressively blended with the FG.

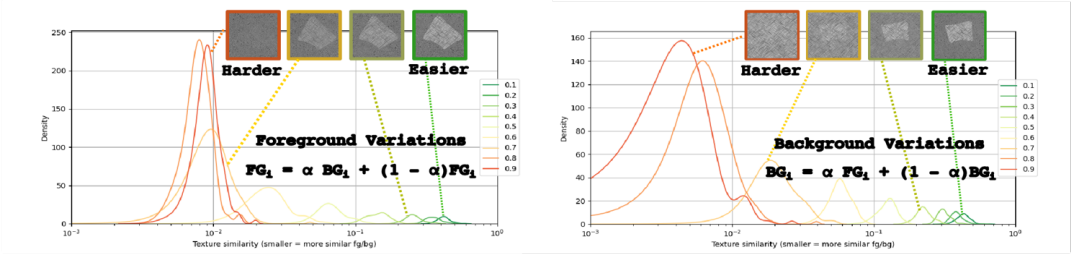


FIGURE 7.2: Generation of synthetic data samples as a function of blending BG texture into the FG (left) and FG into the BG (right). Numbers in the legend indicate the proportion of blending, ranging between 0.1 to 0.9 in steps of 0.1.

Increasing the value of α from 0 to 1 progressively adds more BG texture to the FG mask (see Figure 7.2, left) while the BG texture remains unchanged. Conversely, Figure 7.2, right, illustrates the inverse scenario, where the FG texture remains unaffected, but the BG increasingly resembles the FG texture. We measured the KL divergence between their LBP histograms to quantify the similarity between FG and BG regions. LBP was chosen as it is a well-established and widely used texture descriptor in machine-learning applications [481–483].

$$\mathrm{TS} = \mathrm{KL}(H(L(I)_{\mathrm{BG}})||H(L(I)_{\mathrm{FG}})) \tag{7.1}$$
$$L(I)_{\mathrm{BG}} = \mathrm{LBP}(I(x|Z(x) = 0)) \tag{7.2}$$
$$L(I)_{\mathrm{FG}} = \mathrm{LBP}(I(x|Z(x) = 1)) \tag{7.3}$$

Where TS refers to the level of texture similarity, $H()$ corresponds to the histogram, and $L(I)_{\{\mathrm{BG,FG}\}}$ refers to LBP calculated for BG or FG in image I. The LBP was computed using a 3×3 neighbourhood to generate the histogram with an eight-bit resolution of texture patterns. TS captures local structural entropy, effectively quantifying the similarity in the information content between FG and BG textures.

In our experiments, we evaluated six U-Net-derived architectures: NoSkipU-Net, NoSkipV-Net, U-Net [265], V-Net [286], AGU-Net [278], and UNet++ [288]. These models represent varying approaches to skip connections: no skip connections, unaltered information transfer (identity transform), and selective filtering via attention mechanisms.

Synthetic Data Experiments: We trained models using synthetic data at varying TS levels between FG and BG, calculated using the KL divergence of LBP histograms (Eq. 7.2 and 7.3). For each value of α ($\in \{0.1, 0.2, \ldots, 0.9\}$) used to create a training set, models were trained and then evaluated on synthetic test sets with the same unperturbed α to measure baseline performance. The robustness of these models

was further evaluated by testing them across the same range of α values but with TS applied to the test sets.

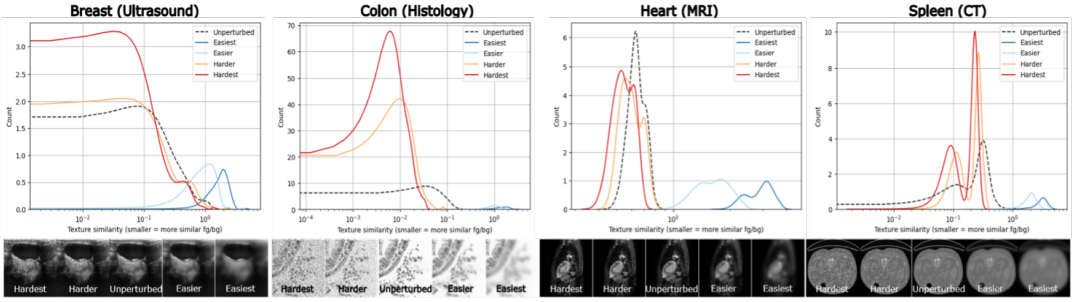


FIGURE 7.3: Selected medical data test sets and histograms of TS at different levels of perturbations. TS for the unperturbed (dashed grey), easier task (light and dark blue, low similarity), and harder task (orange and red, high similarity) distributions. Four modalities tested include US (Breast), Histology (Colon Cancer), CT (Spleen), and MRI (Heart), whose TS are in the same range as the synthetic data in Fig. 7.3.

Medical Data Experiments: Using Eq. 7.1 and ground truth segmentation labels, we calculated the TS for test images from the medical datasets. Corruptions, such as speckle noise or blurring, were applied to either increase or decrease TS, depending on the imaging modality. Details on how these perturbations were generated, along with additional examples and segmentation results, are provided in the Appendix. Figure 7.3 (bottom) illustrates example images with the applied perturbations for each modality. We assessed the robustness of the models to texture changes across four levels of TS: Easiest (highest TS), Easier, Harder, and Hardest (lowest TS). These evaluations were conducted on perturbed test sets with texture similarities differing from the original TS, allowing us to study model behaviour under diverse conditions.

7.2.2 Description of Data

Synthetic Textures: We used two representative greyscale textures from the synthetic images described in Hoyer et al. [484] as BG and FG patterns. These textures were selected to ensure their TS values aligned with the range of TS values observed in the medical datasets (between $1e^{-2}$ and 1). Synthetic segmentation masks were created using Bézier curves to mimic the curvature and size characteristics of clinical image segmentation tasks. Examples of these synthetic images are shown, where Figure 7.2 (left) depicts the scenario where the BG blends into the FG using an alpha-blending scheme. In contrast, Figure 7.2 (right) illustrates the opposite case, where the FG blends into the BG texture. One hundred image-mask pairs were generated at nine levels of alpha ($\alpha \in \{0.1, 0.2, \ldots, 0.9\}$) to produce training datasets with varying TS. Each greyscale texture was cropped randomly to 256×256 pixels to generate these images. The dataset was split into 70 images for training, 10 for validation, and 20 for testing, consistent across all levels of TS. Given the simplicity of the task, this amount of data was sufficient to train the segmentation models.

Figure 7.2 also includes kernel density estimates for the nine levels of TS between FG and BG. For instance, the green curve ($\alpha = 0.1$) represents a dataset where the FG mask contains only 10% of the BG texture and 90% of the original FG texture. In contrast, the red curve ($\alpha = 0.9$) represents the reverse, 90% BG texture in the FG

region. This creates a significantly more challenging segmentation task for humans and models, exemplifying a highly texturally complex scenario.

Both FG and BG blending pose significant challenges in medical image analysis, where the goal is to distinguish pathological structures from surrounding normal tissue. BG Blending occurs when a structure of interest has a texture like its surroundings. For example, brain or liver metastases on non-contrast-enhanced CT or MRI scans may closely resemble the normal brain or liver tissue, as they lack the enhanced contrast typically provided by intravenous agents. This similarity makes it difficult to detect these structures. FG Blending happens when the surrounding tissue has a texture like the structure of interest. For instance, a lung tumour might be adjacent to areas of pneumonia (lung inflammation) or atelectasis (a collapsed part of the lung), which can resemble the tumour's texture on CT images. Similarly, a brain tumour, such as a glioma, may be surrounded by oedema or haemorrhage, which can appear visually similar on specific MRI sequences. These situations make it challenging to accurately define the tumour's borders, particularly when the tumour and adjacent pathology are contiguous. The label map remains unchanged despite the texture variations in all these cases. This requires the model to learn how to handle intermediate textures of varying similarity, making the segmentation task particularly demanding.

Medical Datasets: We evaluated U-Net architecture variants on four binary medical segmentation datasets: Breast US [485], Colon Histology [486], Spleen CT, and Heart MRI [468].

- **Breast US Dataset:** 647 images, with 400 for training, 100 for validation, and 147 for testing.
- **Colon Histology Dataset:** 660 images, with 340 for training, 240 for validation, and 80 for testing.
- **Heart MRI Dataset:** 829 images, with 563 for training, 146 for validation, and 120 for testing.
- **Spleen CT Dataset:** 899 images, with 601 for training, 82 for validation, and 216 for testing.

For the Breast data set, images from the benign and malignant categories were used, excluding the "normal" category (no FG to segment). For the Spleen and Heart datasets, we selected 2D axial slices (Spleen) and sagittal slices (Heart) where at least one pixel corresponded to the FG. Care was taken to ensure that slices were selected at the patient level to prevent cross-contamination between training and testing splits.

To assess the robustness of U-Net variants, we manipulated the TS of medical test images by applying perturbations:

- **Increasing TS (Harder and Hardest):** Speckle noise with variances of 0.1 and 0.3 was added to both the FG and BG. This reduced TS by making their textures more similar, creating more complex segmentation challenges. These scenarios are labelled as "harder" and "hardest" in Figure 7.3.
- **Decreasing TS (Easier and Easiest):** To create a test set with more distinct textures, the BG was blurred using a Gaussian kernel with variances of 3.0 and 7.0, while the FG pixels remained unblurred. This increased TS by making the FG and BG more distinguishable, resulting in easier segmentation tasks. These scenarios are labelled as "easier" and "easiest" in Figure 7.3.

These perturbations allowed us to evaluate how the U-Net variants perform under varying levels of TS.

7.2.3 Model Architecture and Training Settings

The network architectures were implemented using MONAI [480] v1.1, with random weight initialization and fixed random seeds for reproducibility across all experiments. The U-Net is configured with one input and one output channel, using a 256×256 input image size. The model had five levels for synthetic experiments with channels increasing to 16, 32, 64, 128, and 256; an additional level with 512 channels was included for medical image experiments. Intermediate layers used a stride of 2 and Rectified Linear Unit (ReLU) activation. The V-Net is implemented with MONAI's default parameters and ReLU activation. The NoSkipU-Net and NoSkipV-Net are identical to U-Net and V-Net, respectively, but with skip connections removed. The AGU-Net is based on U-Net but includes attention-gating in the skip connections. Finally, the UNet++ is configured with six levels (16, 32, 64, 128, 256, and 512 channels) due to MONAI's implementation requirements. ReLU activation and batch normalization were applied. No dropout was used in any model to reduce stochasticity. Training parameters were standardized across all models to ensure a fair comparison.

Training Setup: The models were trained with a learning rate $1e^{-3}$ using the Adam optimizer [487] and a cosine annealing scheduler for synthetic data. A constant learning rate of $1e^{-2}$ for medical data was used without a scheduler. These models were trained for 100 epochs without early stopping, optimizing for dice loss. Based on validation DSC (evaluated every two epochs), the best model was saved for testing.

Data Augmentation and Hyper parameter Settings: Only random 90-degree rotations (with a 50% probability) were applied during training to avoid altering texture characteristics. No other augmentations, such as scaling, were used to prevent introducing confounding effects. Hyper parameter tuning and ensembling were deliberately avoided, as the focus was on avoiding experimental confounders while comparing robustness rather than achieving the best absolute performance based on standard metrics.

Experiment Protocol: Each model was trained three times using the same random seeds, and statistics were reported for various evaluation metrics. Training and testing were conducted on an NVIDIA A5000 GPU with 24 GB RAM and CUDA version 11.4. This setup ensured minimal confounding factors, providing a robust basis for comparing performance across architecture variants.

7.2.4 Evaluation Metrics

We used four metrics to evaluate model performance and an additional measure of robustness. Performance metrics were implemented using the metrics package from MONAI. Results are reported over a single random seed for synthetic experiments and averaged over three random seeds for medical datasets.

- **DSC:** As introduced by Zou et al. [399], DSC is defined as $\text{DSC} = 2 \cdot |X \cap Y| / (|X| + |Y|)$ where $|X|$ and $|Y|$ represent the ground truth segmentation and predicted mask, while $|X \cap Y|$ indicates their intersection. DSC ranges from 0 (no match, worst) to 1 (exact match, best).

- **HD 100:** As described by Huttenlocher et al. [400], this measures the extent to which points in one set (e.g., model prediction) lie near points in another

set (e.g., ground truth) and vice versa. The "100" refers to recording the maximum pixel-wise distance between the two sets. A value of 0 indicates a perfect match, while higher values (unbounded) indicate worse performance.

- **Surface Distance:** The Average Symmetric Surface Distance (ASSD) [488] measures the average distance between the boundaries of the predicted mask and the ground truth, computed symmetrically. Lower values indicate better performance, with zero being optimal.
- **Surface DSC:** Surface DSC [401] measures the normalized overlap of two surfaces within a predefined margin set at 5 pixels for our experiments. It ranges from 0 (no overlap, worst) to 1 (perfect match, best).

Robustness Metric: To evaluate robustness, we used the Coefficient of Variation (CV), defined as the ratio of the standard deviation to the mean [489]. This metric was computed for the above performance metrics across five texture conditions: easiest, easier, unperturbed, harder, and hardest. Smaller CV values indicate lower dispersion, reflecting greater robustness to texture variations in the test set. CV focuses on consistency across texture variations, identifying the most robust model architecture rather than the one with the best absolute performance for a given metric. This ensures a robust comparison across architectures in varying test conditions.

7.3 Results

We first report results on the synthetic texture variations grouped by the evaluation metric and then report results for the medical image data sets.

7.3.1 On Synthetic Texture Variants

For the synthetic experiments, we used the NoSkipU-Net, U-Net, and AGU-Net as representative architectures from each of the three categories of models to demonstrate our results. Our experiments generate nine levels of blending of FG and BG textures (i.e., $\alpha = [0.1, 0.2, \ldots, 0.9]$), which are used for training and testing with all 81 combinations evaluated (i.e., nine combinations for training and testing, respectively). When the training and testing texture levels match, this represents the unperturbed situation where the performance is expected to be the highest. When the texture levels do not match, the spread of results indicates how robust these models are for texture-perturbed testing images.

For example, Figure 7.4 (top row) indicates the DSC for three models: AGU-Net, U-Net, and NoSkipU-Net, with the rows in each 9×9 grid indicating the α values used to train the model and the columns indicating the α values used to construct the test set. The α value is used as a proxy for TS, as shown in Figure 7.2. The unperturbed performance is shown on the primary diagonal - left-top to right-bottom, and the off-diagonal scores represent a measure of the robustness of the performance in other TS scenarios (both when training textures and testing textures are perturbed). We also report the relative performance of the AGU-Net and U-Net compared to the NoSkipU-Net to better visualize the improvements or degradations they may cause. For example, Figure 7.4 (bottom row) is a relative performance measure of the AGU-Net and U-Net when using the NoSkipU-Net as the baseline. Red patches indicate better performance than the NoSkipU-Net, whereas blue patches indicate where the NoSkipU-Net model is better.

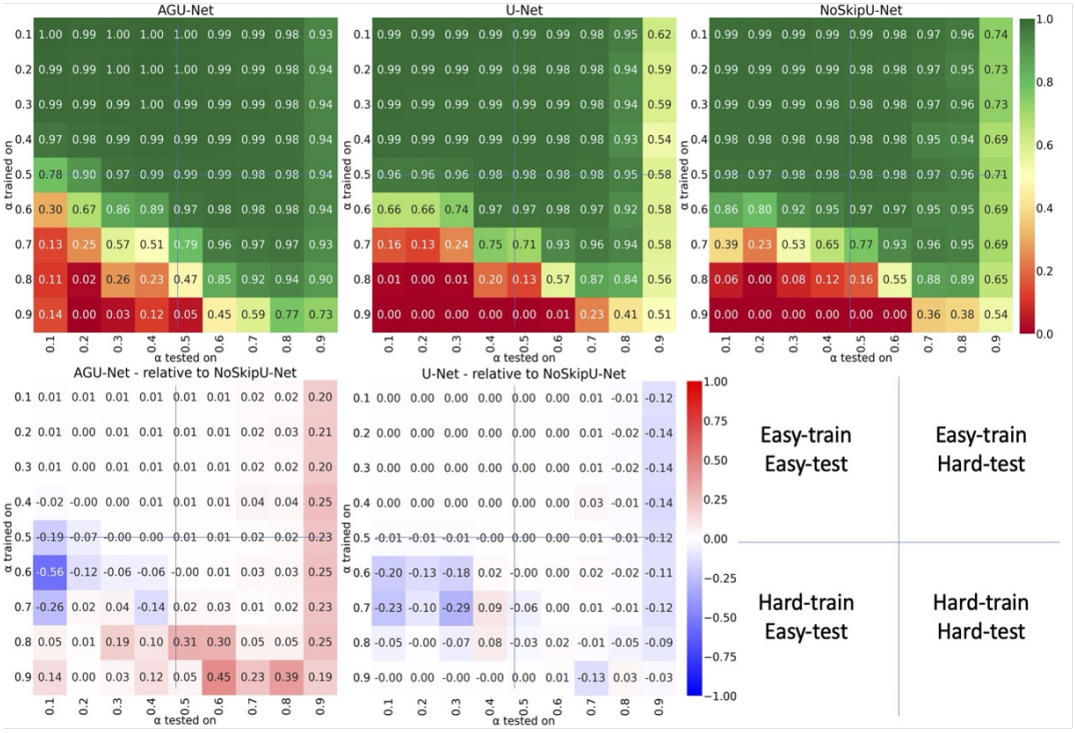


FIGURE 7.4: Model robustness measured using DSC for U-Net-like architectures when the BG blends into the FG texture. For the first row, showing absolute DSC values, higher/green is better. For the second row, showing relative DSC values compared to the NoSkipU-Net model, blue indicates that NoSkipU-Net is better.

To further facilitate the analysis of these results, on each figure presenting results on the grid of 9×9 results, we also present a visual aid describing four quadrants for: (i) "Easy-train/Easy-test," where minimal perturbations were applied during training and testing; (ii) "Easy-train/Hard-test," where minimal perturbations were applied on the training data, and increased levels of perturbations were applied to the testing data; (iii) "Hard-train/Easy-test," where increased levels of perturbations were applied to the training data, and minimal perturbations were applied to the testing data, and (iv) "Hard-train/Hard-test," where increased levels of perturbations were applied to both training and testing data.

DSC Results on Synthetic Experiments

Based on the results in Figure 7.4, for training and testing α values < 0.5 ("Easy-train/Easy-test"), there is negligible difference between the DSC performances of all the U-Net variants, indicating their ability with or without the skip-connections to learn the distributions of the FG and BG textures at that level of TS. This represents the canonical case of high-quality training and high-quality testing data, and our findings here indicate that all the architectures perform reasonably well on the synthetic data set.

In the case of both training and testing $\alpha > 0.5$ ("Hard-train/Hard-test"), differences in performance appear. The AGU-Net outperforms both the U-Net and NoSkipU-Net models. This indicates that the benefit of attention-gating as a function of TS is non-linear: models do not benefit from skip connections at lower ranges

of TS, but filtering the information flowing through the skip connections is vital at larger ones.

Arguably, the more clinically relevant quadrants of the evaluation are where the training data is of good quality (i.e., "Easy-Train/Easy-Test" and "Easy-Train/Hard-Test") for models that have been trained using appropriate data curation and quality control workflows. In these two scenarios, we observed in the synthetic experiments that the AGU-Net model outperforms the U-Net and NoSkipU-Net only at very high levels of test set perturbation ($\alpha = [0.8, 0.9]$). In contrast, the performance levels are similar at lower regimes of perturbation. For the quadrant "Hard-Train/Easy-Test," it is evident from the more significant differences in behaviour across the models that these networks find it hard to learn the task when the training data set has been perturbed such that FG and BG are similar. This situation is included for completeness as we believe that models trained for real-life clinical applications should be trained with good data quality and under a data-centric scheme.

While referring to the relative comparison of DSC metrics (i.e., bottom rows in Figure 7.4), we note that there are combinations of training α and testing α where the NoSkipU-Net outperforms the other models, indicating that the best choice of network architecture seems to depend on the TS. Enhancements to the skip connections do not necessarily always lead to improved performance.

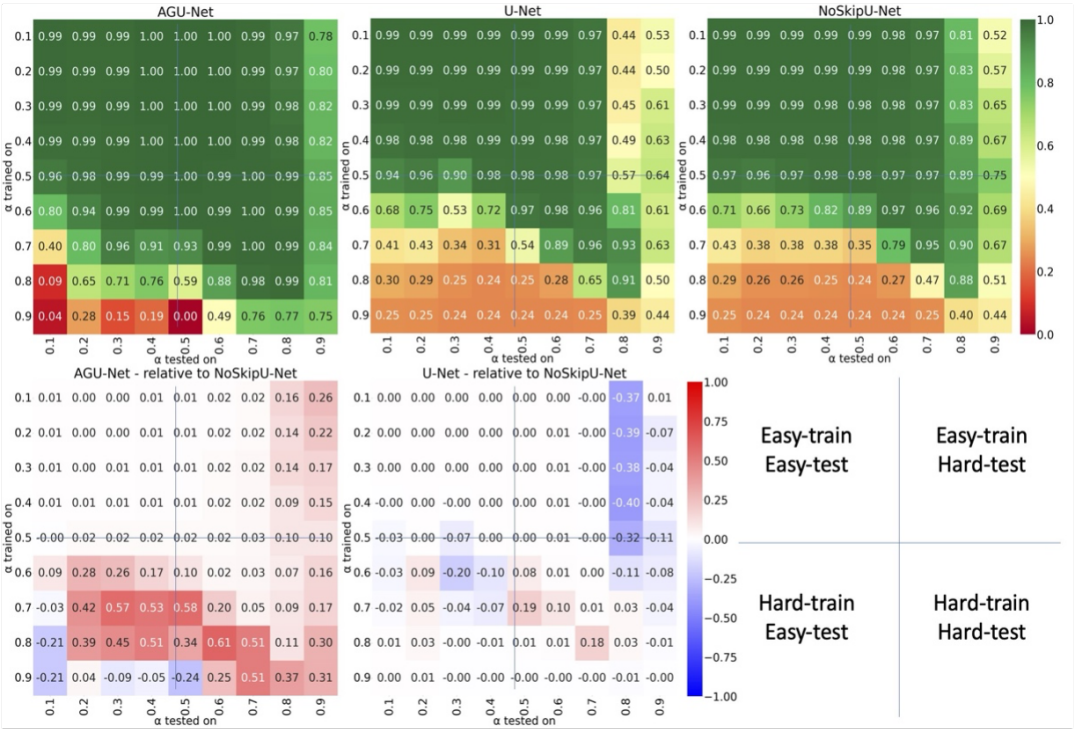


FIGURE 7.5: Model robustness measured using DSC for U-Net-like architectures when the FG blends into the BG texture. For the first row showing absolute DSC values, higher/green is better. For the second row showing relative values compared to the NoSkipU-Net, blue indicates the NoSkipU-Net model is better.

Figure 7.5 is identical to Figure 7.4, with the difference being that the FG blends into the BG texture instead when α increases. Interestingly, the trends for the easy-training, easy-testing case remain the same as in Figure 7.4: the choice of architecture does not impact the performance appreciably. However, the "Hard-Train/Easy-Test" case shows better DSC scores than the BG, which blends into the FG. In this

case, AGU-Net appears to be the most robust model architecture, while NoSkipU-Net generally performs as well or better than the standard U-Net across α values.

HD Results on Synthetic Experiments

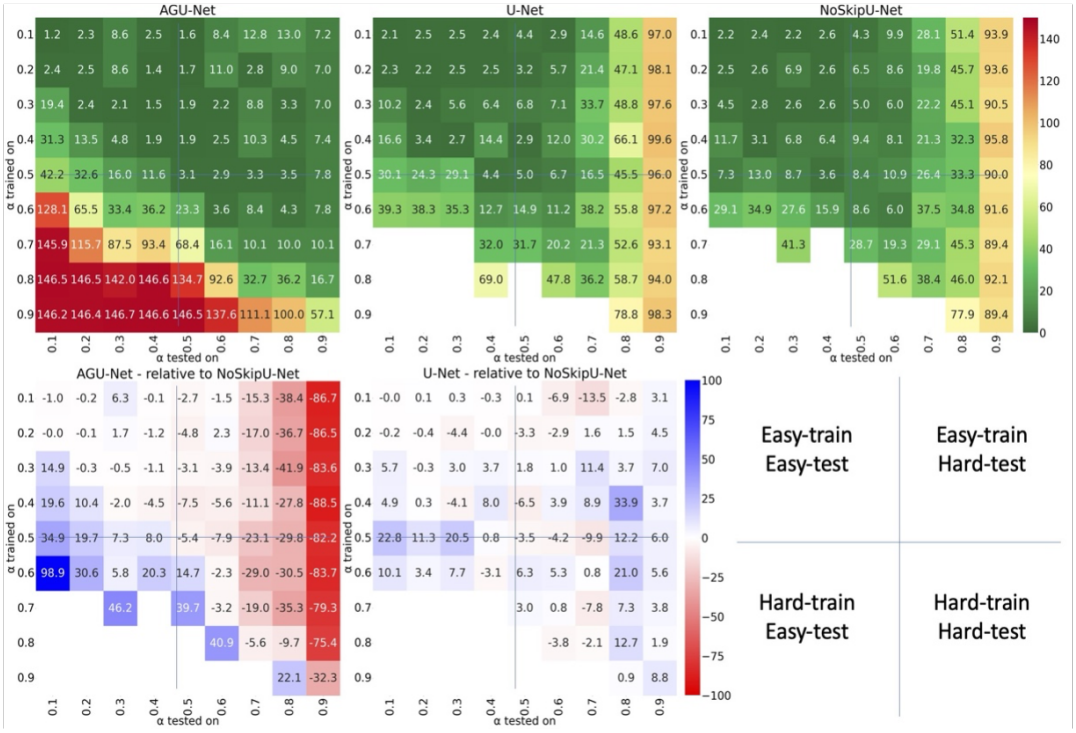


FIGURE 7.6: Model robustness measured using HD100 for U-Net-like architectures when the BG blends into the FG texture. Locations where the BG is white indicate undefined values of HD. For the first row, higher/green is better. For the second row, blue indicates that NoSkipU-Net is better.

Figures 7.6 and 7.7 show the behaviour of the AGU-Net, U-Net, and NoSkipU-Net on the FG-blending and BG-blending cases, respectively. This metric shows more significant differences between the U-Net and NoSkipU-Net robustness than DSC, as indicated by the more significant concentration of cells with values larger than 5 mm along the diagonal in Figure 7.7 (second row, second column). There is considerable variation in the relative performance of the AGU-Net compared to the NoSkipU-Net, especially at higher levels of perturbations (second row, first column). The blue boxes indicate situations where NoSkipU-Net performs better, further supporting our claim that skip connections are not always beneficial.

Comparing figures 7.6 and 7.7, significant differences appear in the hard-train, easy-test scenarios; otherwise, the variations are similar. It is worth noting that the NoSkipU-Net is mostly on par or better than the standard U-Net on both polarities. The AGU-Net fares the best in the easy-train, hard-test scenario and in both blending directions. Beyond DSC and HD, we also calculated surface distance and surface DSC (See Appendix), where similar trends are observed, further supporting our claims.

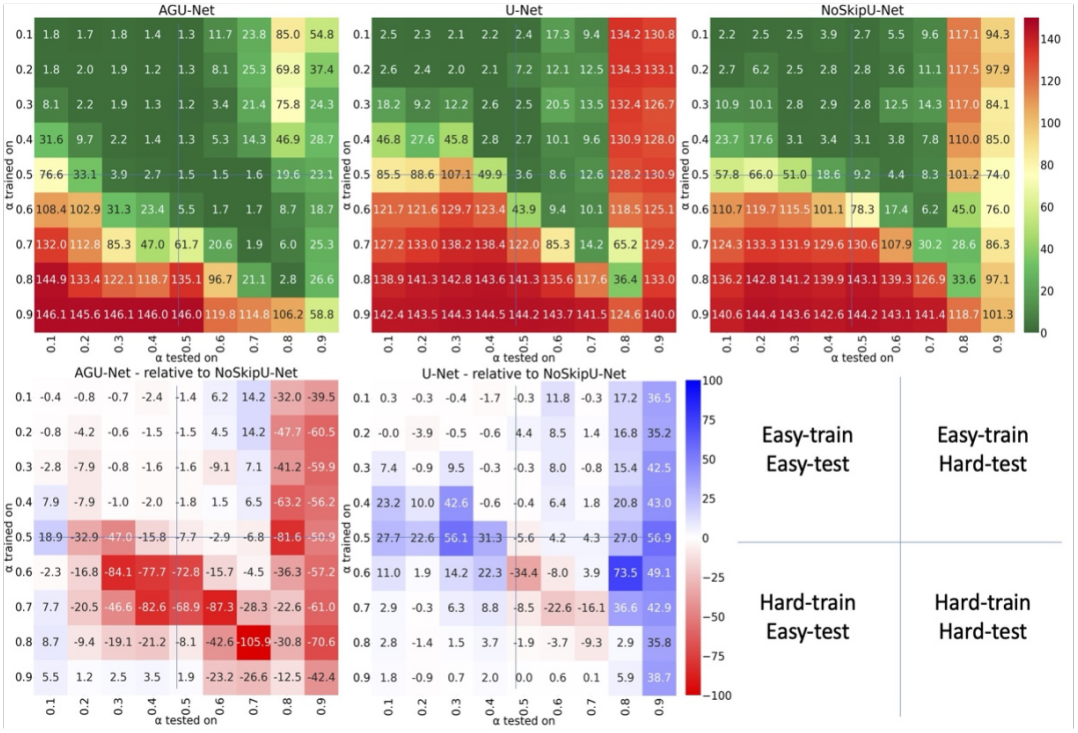


FIGURE 7.7: Model robustness measured using HD 100 for U-Net-like architectures when the FG blends into the BG texture. For the first row, higher/green is better. For the second row, blue indicates that NoSkipU-Net is better.

7.3.2 Results on Medical Image Datasets

DSC Results on Medical Image Datasets

Figure 7.8 shows how the DSC scores change for six model architectures across four medical image datasets. The black bar in the middle of each group represents the model's performance on the test set when the textures are like those in the training set. The orange and red bars on the left show results for more challenging textures (where the TS is lower), while the blue bars on the right show results for easier textures (where TS is higher, meaning the FG and BG textures are more distinct). As expected, the black bars (unperturbed textures) generally show the best performance, while the other bars are lower, indicating that changes in texture reduce performance. A steeper score drop on either side means the model is less robust, as it struggles more when textures are altered, according to the DSC metric. These ranges of TS are chosen to span the breadth of real-world perturbations caused by scanner settings, noise in acquisition, and imaging protocols. Figures A.18 and A.19 in the appendix demonstrate that the ranges of TS using such perturbations are well within the tested ranges in this section.

For the Breast US dataset, the V-Net model shows the most minor performance changes across TS based on the CV. This makes it the most robust model, as indicated by the trophy symbol in Figure 7.8. The enhanced set of model architectures shows the largest coefficient of variation (least robustness) amongst all the architecture classes, indicating that more skip connections are likely adversely impacting robustness even if the training set has a large TS range.

For the Colon Histology dataset, performance is more consistent across all models, likely because the TS spread in the unperturbed data set (which the model was

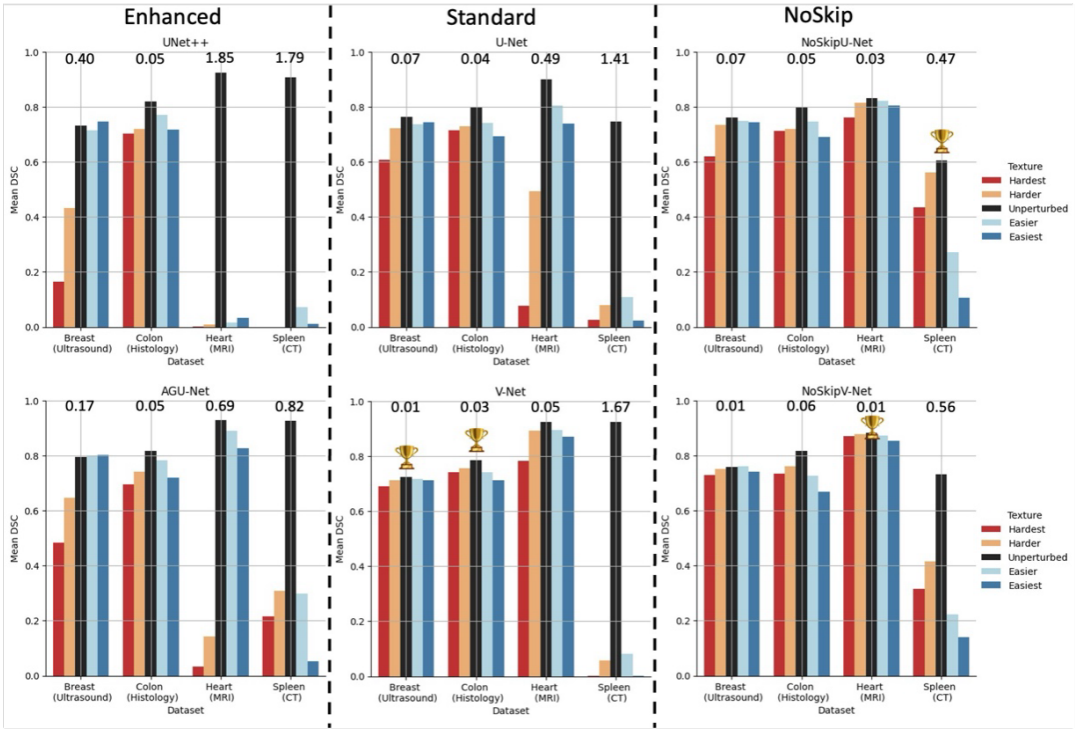


FIGURE 7.8: DSC variations across model types - UNet++, U-Net, NoSkipU-Net, AGU-Net, V-Net, and NoSkipV-Net over five levels of TS. The columns indicate the category of models: Enhanced, Standard, and NoSkip. Numbers at the top of each group of bars indicate the CV, lower ones indicate flatter profiles and, hence, more robustness in performance, and the trophy icon suggests the best model per data set based on the coefficient of variation. Higher bars are better, and flatter profiles across bars indicate more robustness to TS.

trained on) is the widest amongst all other data sets, and the overlap between unperturbed and the harder conditions is subsequently higher. Although most models (except the V-Net) achieve an "unperturbed" score above 0.8, all models show a noticeable drop in performance with texture changes. V-Net remains the most robust model in this case.

The Heart MRI dataset has the narrowest range of TS in its training data (unperturbed). The UNet++, AGU-Net, and U-Net models experience significant performance drops with harder textures, showing that a sharper concentration of training data in a narrow TS band impacts robustness more significantly in the presence of more skip connections. In contrast, the V-Net and NoSkip variations perform better. Among them, the NoSkipV-Net is the most robust, showing the least variation in performance.

The Spleen CT dataset's performance varies the most, as shown by the largest coefficient of variation, with significant drops across all models under perturbed texture scenarios. This follows the general observation that a narrower TS range while training (unperturbed) leads to the largest robustness impact while testing. The NoSkipU-Net performs the best in terms of robustness here, even though its mean DSC is the lowest. Particularly in the harder texture cases, it outperforms all other models significantly as the only one with a drop in DSC < 0.2.

Table 7.1 ranks the six model architectures by identifying which model performs

best on each test image compared to the others. This is known as case-based aggregation [490]. In the harder texture scenarios, NoSkip variants perform best 5 out of 8 times. In the easier texture scenarios, models with more skip connections (enhanced category) outperform others 6 out of 8 times. All three model categories are represented across the four datasets in the unperturbed scenario.

TABLE 7.1: The best-performing model (Ranking on the test set: model with the highest metric per image for the most images in the test set wins) uses DSC for all four data sets. The numbers next to each model name indicate the proportion of the test set for which this model wins.

	Breast US	Colon Histology	Heart MRI	Spleen CT
Easiest	AGU-Net (0.340)	V-Net (0.300)	AGU-Net (0.425)	NoSkipV-Net (0.629)
Easier	AGU-Net (0.346)	AGU-Net (0.362)	AGU-Net (0.491)	AGU-Net (0.421)
Unperturbed	NoSkipU-Net (0.251)	UNet++ (0.275)	V-Net, AGU-Net (0.383)	V-Net (0.384)
Harder	NoSkipV-Net, NoSkipU-Net (0.265)	AGU-Net (0.325)	V-Net (0.616)	NoSkipU-Net (0.750)
Hardest	NoSkipV-Net (0.469)	V-Net (0.325)	NoSkipV-Net (0.783)	NoSkipU-Net (0.731)

HD Results on Medical Image Datasets

Figure 7.9 shows the variations in HD values in the same format as Figure 7.8. The V-Net is the most robust model architecture for the Breast US, the NoSkipV-Net on the Heart MRI data set, and the NoSkipU-Net is the most robust on the Spleen CT data set. The AGU-Net architecture is the most robust for the Colon histology data set, even though the mean distances are uniformly high.

Table 7.2 shows the ranking of the best model architecture in the same format as Table 7.1. Using HD, the NoSkip variants of architectures are still the best performing in the harder situations (winning in 7 out of 8 cases), and the enhanced category of models is the best for the easier situations (winning in 7 out of 8 cases). The "unperturbed" case has representation again from all the model architecture categories.

7.3.3 Training Durations Across Model Architectures and Data Sets

Table 7.3 reports the mean and standard deviation of training time in seconds per epoch for all four data sets and all six model architectures. Training with the UNet++ architecture took the longest time, about a hundred times that of the U-Net and NoSkipU-Net. Such variations are also reported by Gut et al. [491], where the UNet++ takes more than four times the training time as compared to the U-Net. The difference in performance could likely come from the choice of implementation. Training the V-Net variants (both with and without skip connections) took about ten times the time for U-Net variants, and the AGU-Net took about six times the time compared to the U-Net variants.

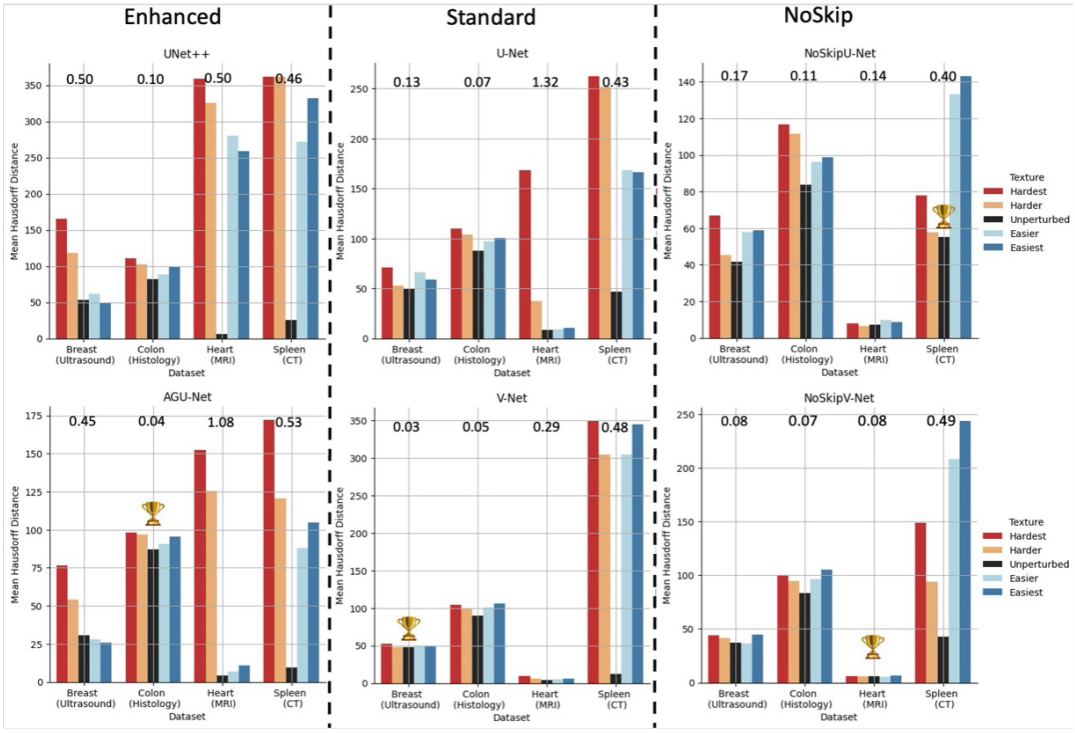


FIGURE 7.9: HD variations across model types - UNet++, U-Net, NoSkipU-Net, AGU-Net, V-Net, and NoSkipV-Net over five levels of TS, in the same format as Figure. 7.8. Lower bars are better, and flatter profiles across bars indicate more robustness to perturbations.

TABLE 7.2: The best-performing model using HD for all four data sets, in the same format as Table 7.1.

	Breast Ultrasound	Colon Histology	Heart MRI	Spleen CT
Easiest	AGU-Net (0.387)	AGU-Net (0.337)	NoSkipV-Net (0.475)	AGU-Net (0.643)
Easier	AGU-Net (0.421)	UNet++ (0.362)	AGU-Net (0.325)	AGU-Net (0.587)
Unperturbed	AGU-Net (0.292)	UNet++, NoSkipV-Net (0.225)	V-Net, AGU-Net (0.308)	V-Net (0.467)
Harder	NoSkipV-Net (0.319)	AGU-Net, NoSkipV-Net (0.275)	NoSkipV-Net (0.483)	NoSkipU-Net (0.708)
Hardest	NoSkipV-Net (0.496)	AGU-Net (0.537)	NoSkipV-Net (0.791)	NoSkipU-Net (0.824)

TABLE 7.3: Mean (standard deviation) of training time in seconds per epoch for Breast (US), Histopathology, Spleen (CT), and Heart (MRI) data sets on various alternatives of the U-Net architecture.

	UNet++	AGU-Net	U-Net	V-Net	NoSkipU-Net	NoSkipV-Net
Spleen (CT)	816.20(1.47)	53.76(0.15)	8.32(0.11)	81.11(0.26)	7.12(0.02)	81.71(0.19)
Colon (Histology)	508.31(2.43)	34.48(0.66)	6.45(0.06)	49.55(2.04)	5.80(0.08)	53.84(0.19)
Heart (MRI)	773.19(6.69)	58.40(11.47)	9.48(1.89)	77.51(0.23)	8.34(1.84)	77.87(0.15)
Breast (US)	556.26(2.90)	38.20(0.16)	6.11(0.01)	58.51(0.45)	5.42(0.04)	59.45(0.04)

7.4 Discussion

Through systematic experiments using synthetic texture images at various levels of complexity and validating these findings on medical image data sets from four different modalities, we show in this work that skip connections demonstrate a complex relationship with robustness dependent on image texture features within the FG and BG labels. The asymmetric results in the synthetic experiments show that situations where the training set has larger FG-to-BG textural differences (visually easier to distinguish, called 'easy-train, hard-test' in Figure 7.4) lead to better performance on a wider range of textures during the test compared to training with smaller textural differences, which are evidently also harder for humans to deal with (called 'hard-train, easy-test'). This indicates that prior knowledge of texture similarities and their range can inform the robustness of model architectures with and without skip connections. A possible explanation for why the FG blending results (i.e., FG blends into the BG) are worse than the BG blending results in the 'hard-train, easy-test' (bottom left quadrant in Figure 7.5) case is that the model is overexposed to the FG texture all across the image, both within and outside the label mask, leading to inferior learning as compared to the BG blends into the FG case, where the location of the label mask could still provide some information to learn where the BG texture must not exist [492].

On the medical image data sets, the NoSkip-category of model architecture outperformed the other architectures in 7 out of 16 tasks, comparing favourably to the enhanced category of architectures, which only outperformed others in 3 tasks. This supports our claim that skip connections are not always beneficial, especially in texturally complex situations with similar FG and BG textures. Moreover, we observe for the breast US data set that the AGU-Net improves in the easier tasks but declines in performance in the harder ones. Amongst the U-Net and V-Net and its NoSkip variants, the V-Net variants show a smaller CV, indicating better robustness across TS. The performance of all the six model architectures on the unperturbed version of the colon histology data set is comparable, at around DSC of 0.8, following similar results demonstrated on data set number 8 [491], and between the U-Net and its variants [493]. However, the drops in performance for the easier and harder situations of the colon histology dataset are consistent across all model architectures. For the heart data set, UNet++ showed a severe robustness drop in easier and harder texture situations, which we believe is because it has the narrowest range of TS in the 'unperturbed' training set. Interestingly, the U-Net and AGU-Net only showed a more pronounced drop in the harder texture situation. The V-Net and both the NoSkip variants were amongst the most robust. In the spleen data set, all the architecture variants showed poor robustness to texture variations, which we again believe is due to narrower TS ranges in the training set. The NoSkip group performs marginally better than the others, further supporting our claims.

These results, both on synthetic and medical image data, suggest that skip connections not only facilitate the transfer of high-resolution details from the encoder to the decoder but also introduce sensitivity to noise, depending on the range of textures encountered during training. Specifically, architectures with denser skip connections (the Enhanced group) achieve superior performance in the unperturbed texture case, where training and testing textures are highly similar. However, they exhibit diminished robustness when tested on textures that deviate from their training distribution. This vulnerability seems to arise because skip connections reinforce high-frequency details, amplifying noise when encountering previously unseen texture variations. For example, in models such as AGU-Net, where attention gates

modulate skip connection pathways, noise in the encoder features can be selectively amplified rather than suppressed, leading to a detrimental effect on robustness.

Additionally, our findings indicate that the choice of operation in skip connections might play a key role in how robustness is affected. In standard U-Net architectures, skip connections rely on concatenation, which preserves all feature information from the encoder without modification, potentially reinforcing undesirable texture variations. In contrast, in architectures such as V-Net, skip connections employ element-wise addition, which forces an implicit alignment of feature representations between the encoder and decoder. This regularization effect appears to enhance robustness, suggesting that addition-based skip connections may mitigate the excessive amplification of texture-specific features.

In contrast, architectures without skip connections (the NoSkip group) sacrifice performance gains from direct encoder-to-decoder information transfer but gain robustness against OOD textures. Most evaluations of these architectures focus solely on performance in standard conditions, yet our study highlights that architectural design choices also influence robustness characteristics, offering a broader perspective on their strengths and weaknesses. Based on these findings, future work includes the development of TS-driven adaptive skip connection mechanisms, where attention based on texture variations regulates the information flow leading to a more optimal trade-off between robustness and performance.

Limitations: While our study employs a 2D TS metric to analyse the relationship between texture and segmentation robustness, we acknowledge that a 3D metric could further refine this analysis. However, existing literature suggests that robustness degradation patterns remain consistent between 2D and 3D models [494], indicating that our findings likely generalize to volumetric settings. Furthermore, due to the computational demands associated with training 3D models at the scale of our study (120 experiments across multiple datasets and metrics), we prioritized a 2D approach to enable a broader and more exhaustive evaluation of texture-based robustness effects. Future work should explore hybrid approaches that integrate 2D and 3D texture descriptors to further refine robustness-aware segmentation strategies. We chose two representative greyscale textures [484] so that the synthetic experiments spanned the range of TS observed in medical image data (between $1e-2$ and 1). Exploring performance outside this range of TS with more texture pairs (see Figure A.20 in the appendix) would be computationally too expensive, while not providing any further useful insights. Noise levels in our experiments were setup based on sweeping parameter ranges with TorchIO. This was designed to cover real world texture variations in medical image data (across eight noise types for MRI and CT imaging scenarios, see Fig A.18 and A.19 in the appendix). Therefore, we expect our findings to hold in such scenarios. Although this is a good first approximate model of the ranges of variations, we acknowledge that other noise types and modalities need to be investigated. We believe that a better approach to model noise levels could be to employ publicly available real world medical image data sets characterizing realistic noise ranges observed in practice. To the best of our knowledge, such data sets do not currently exist.

We intentionally focus on varying only the density of skip connections within the model architecture and do not perform extensive hyper parameter tuning, as this could introduce confounding factors that obscure the direct impact of skip connections on performance and robustness. While hyper parameter tuning is critical for optimizing absolute performance in settings such as model benchmarking challenges, our focus is on analysing the relative impact of texture variations under different skip connection configurations. To ensure fair comparisons, we use default

settings from MONAI, follow well-accepted hyper parameter choices from prior work, and fix random seeds across all experiments. This allows us to attribute variations in performance and robustness specifically to skip connection density rather than to differences in optimization strategies. The coefficient of variation metric used to measure the robustness of results across texture variants may only partially match the properties we desire from a robust system. It prioritizes consistency in results and does not necessarily focus on the optimal value of the metric in question. CV was chosen as a simple yet effective metric, and the overall merits of architecture choice may depend on a larger set of metrics [495].

Outlook: The main finding of this study is that the robustness of models depends on the density of skip connections and the textures the model is trained on. More specifically, narrower TS ranges and more skip connections lead to more brittle models than fewer skip connections and larger training TS ranges. This indicates that developing a texture-aware adaptive U-Net architecture that matches the TS range in training with real-world conditions is an interesting research avenue. Furthermore, combining multiple models in a weighted manner dependent on the TS of the resulting segmentation could further enhance performance while maintaining robustness. For example, if the TS is outside the range of the training set TS, one would prefer a NoSkip variant, while if it is well represented in the training set, one can more safely assume that the performance of a Standard or Enhanced model would be better. We would also like to explore region-specific metrics by examining how these model architectures perform at edges, corners, and texturally flat regions as measured by LBP. Finally, we would like to focus on interpretability by computing saliency maps for each experimental setting of model architecture, texture variant, and medical image modality [496] to analyse the variability of gradient information across the different networks upon different perturbation levels and across different layers of the models. We believe these findings can direct future work on architectural innovations around the U-Net to be more focused on robustness, so they are safer to use in critical application areas like medical image segmentation.

8

Do We Really Need that Skip-Connection? Understanding Its Interplay with Task Complexity

8.1 Introduction

Due to the broad success of U-Nets [265] for image segmentation, it has become the go-to architecture in the medical image computing community. Since its creation in 2015, much research has been dedicated to explore variants and improvements over the standard base model[464]. However, Isensee et al. [285] showed with their not-new-U-Net (nnUNet) that the success of the U-Net relies on a well-prepared data pipeline incorporating appropriate data normalization, class balancing checks, and preprocessing, rather than on architecture changes. Arguably the two most important challenges at present for medical image segmentation are generalization and robustness. A lack of generalization decreases the performance levels of a model on data sets not well characterized by the training data set, while poor robustness appears when models under-perform on data sets presenting noise or other corruptions [472]. Modern neural networks have been shown to be highly susceptible to distribution shifts and corruptions that are modality-specific [473]. While the average accuracy of U-Net-based models has increased over the years, it is evident from literature that their robustness level has not improved at the same rate [309, 469, 474].

One of the key elements of the U-Net are the skip-connections, which propagate information directly (i.e, without further processing) from the encoding to the decoding branch at different scales. Azad et al. [464] mention that this novel design propagates essential high-resolution contextual information along the network, which encourages the network to re-use the low-level representation along with the high-context representation for accurate localization. Nonetheless, there is no clear evidence supporting this intuition and moreover, there is limited knowledge in the literature describing to what extent skip-connections of the U-Net are necessary, and what their interplay is in terms of model robustness when they are subjected to different levels of task complexity.

Currently, the U-Net is used more as a "Swiss-army knife" architecture across different image modalities and image quality ranges. In this paper, we describe the interplay between skip-connections and their effective role of "transferring information" into the decoding branch of the U-Net for different degrees of task complexity, based on controlled experiments conducted on synthetic images of varying textures as well as on clinical data comprising US, CT, and MRI. In this regard, the work of [475] showed that neural networks are biased toward texture information. Recently, [339, 476] similarly showed the impact of texture modifications on the performance and robustness of trained U-Net models. Contrary to these prior works analysing the impact of data perturbation to model performance (e.g. [472, 473]), in this study we focus on analysing the role of skip-connections to model performance and its robustness. We hypothesize therefore that skip-connections may not always lead to beneficial effects across varying task complexities as measured with texture modifications. Our major contributions through this paper are:

(i) We describe a novel analysis pipeline to evaluate robustness of image segmentation models as a function of the difference in texture between FG and BG.

(ii) We confirm the hypothesis that severing these skip-connections could lead to more robust models, especially in case of OOD test data. Furthermore, we show that severing skip-connections could work better than filtering feature maps from the encoder with attention-gating.

(iii) Finally we also demonstrate failure modes of using skip-connections, where robustness across texture variations appear to be sacrificed in the pursuit of improvements within domain.

8.2 Materials and Methods

8.2.1 Experiment design

Figure 8.1 describes our experimental setup to assess the impact of skip-connections in U-Net-like architectures under varying levels of task complexity.

Given a set of N pairs of labelled training images $\{(I,S)_i : 1 \leq i \leq N\}, I \in \mathbb{R}^{H\times W}$ and $S \in \mathbb{Z} : \{0,1\}^{H\times W}$, corresponding ground-truth segmentation, a DL segmentation model $M(I) \mapsto S$ is commonly updated by minimizing a standard loss term, such as the binary cross entropy or dice loss. To evaluate how the model behaves at varying task complexities, we construct training data sets where each training sample is subjected to a linear transformation where its FG is blended with the BG: $I(x \mid Z(x) = 1) = \alpha I(x \mid Z(x) = 1) + (1-\alpha)I(x \mid Z(x) = 0)$.

By increasing α from zero to one, more of the FG texture is added in the FG mask, which otherwise is made up of the BG texture (See Figure 8.2), while the BG itself is not impacted. We then quantify the similarity between FG and BG regions by measuring the KL divergence between their LBP [481] histograms. We selected LBP since it is a commonly used and benchmarked texture descriptor in machine learning applications [482, 483].

$$\begin{aligned} \mathcal{TS} &= KL(\mathcal{H}(\mathcal{L}(I)_{BG}) || \mathcal{H}(\mathcal{L}(I)_{FG})) \\ \mathcal{L}(I)_{BG} &= LBP(I(x \mid Z(x) = 0)) \\ \mathcal{L}(I)_{FG} &= LBP(I(x \mid Z(x) = 1)) \end{aligned} \tag{8.1}$$

where $\mathcal{TS}$ refers to the level of TS, $\mathcal{H}()$ corresponds to histogram, and $\mathcal{L}(\mathcal{I})_{\{BG,FG\}}$ refers to LBP calculated for BG or FG. The LBP histogram was computed using a 3×3 neighbourhood with 8 points around each pixel in the image.

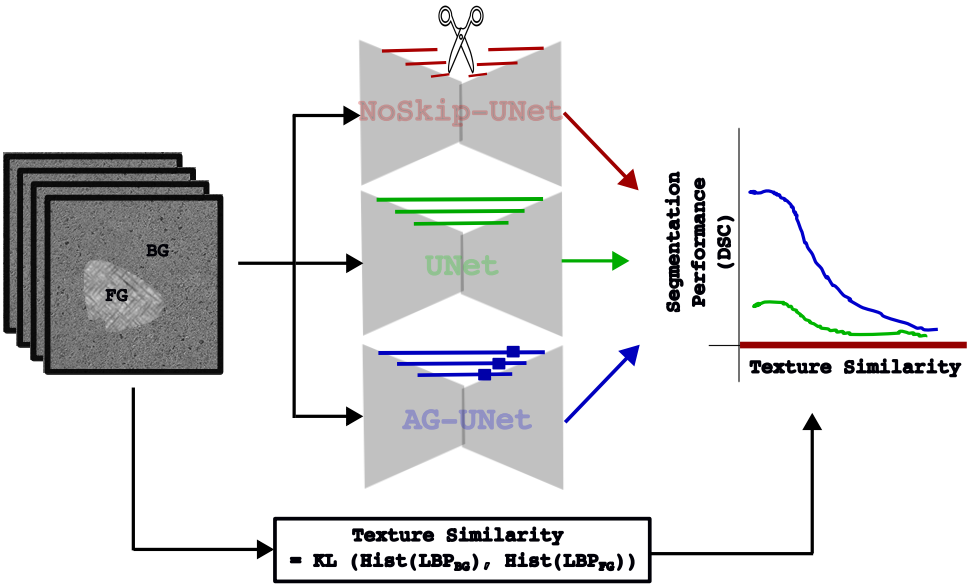


FIGURE 8.1: Experimental design to evaluate the role of U-Net's skip-connections under different levels of task complexity. Given training images with controllable BG and FG textures, three variants of the U-Net were trained featuring no skip-connections (NoSkip-U-Net), standard U-Net (U-Net)[265], and Attention-Gated U-Net (AGU-Net)[278], each characterizing a different strategy (zeroing information through skips, identity transform and filtering information through skips, respectively). Each model was trained with different levels of TS between BG and FG, based on the KL divergence of LBP histograms for FG and BG regions. For each level of FG-to-BG TS, the performance for each model was recorded in-domain, and robustness measured with out-of-domain texture similarities.

Three U-Net models were trained featuring three different skip-connection strategies: NoSkip-U-Net, U-Net, and AGU-Net, representing the absence of skip-connections, the use of an identity transform (i.e, information through skips is kept as is), and filtering information via attention through skip-connections, respectively. Models were trained at different levels of $\mathcal{TS}$ between the FG and BG regions, determined based on the KL divergence of LBP histograms, Eq. 8.1. For each level of α used to create a training set, we trained a model to be evaluated on a synthetic test set using the same α to measure within-domain performance, and across a range of α, to measure their OOD robustness.

Next, using Eq. 8.1 and ground truth labels, we computed the $\mathcal{TS}$ of images from the test set of the medical data sets, and applied corruptions by way of noise or blurring in order to increase and decrease $\mathcal{TS}$ depending on the imaging modality being analysed. Then we evaluated the robustness of these models to texture changes in these data sets. We did this at two levels of task complexity (easier - where $\mathcal{TS}$ is higher, and harder, where $\mathcal{TS}$ is lower) and different from the original $\mathcal{TS}$. We report all model performances using DSC.

8.2.2 Description of data

Synthetic textures: We took two representative greyscale textures from the synthetic toy data set described in [484], and used them as the BG and FG patterns.

These patterns were chosen such that the $\mathcal{TS}$ values matched the range of medical data sets described next. We also generated synthetic segmentation masks using bezier curves setup such that the curvature and size of the FG simulate clinical image segmentation problems. Examples of such images is shown in Figure 8.2.

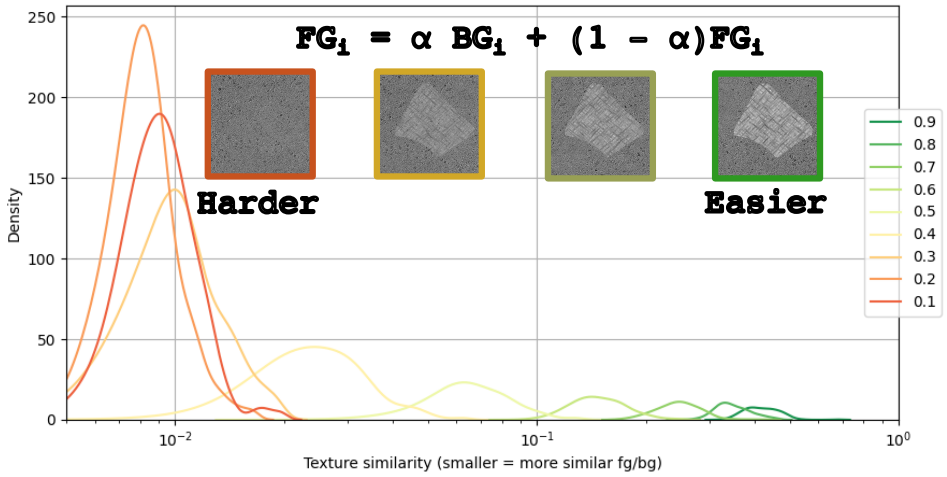


FIGURE 8.2: Generation of synthetic data samples as a function of blending FG texture into the BG. Numbers in the legend indicate proportion of FG blended within the FG mask.

We generated 100 such image-mask pairs at 9 levels (for $\alpha \in \{0.1, 0.2, ..., 0.9\}$), so that we create training data sets at various task complexities. These images are generated by randomly cropping the greyscale textures to 256×256 pixels. 70 of these were used as the training set, 10 were reserved for validation, and the rest of the 20 formed the test set, identically split for all task complexities. Figure 8.2 show kernel density estimates of each of these 9 data sets along the TS axis. The curve in orange ($\alpha = 0.1$) indicates that the FG mask in this set contains only 10% of the actual FG texture and 90% of the BG texture blended together. This represents a situation where it is texturally hard for humans as well as for segmentation models. The data set in green ($\alpha = 0.9$) shows the reverse ratio - the FG region now contains 90% of the FG texture, thereby making it an easier task to segment.

Medical data sets: We tested the three variants of the U-Net architecture on three medical binary segmentation data sets: a Breast US [485], a spleen CT and a heart MRI data set [468]. The breast US data set contained 647 images, 400 of which were used as training, 100 as validation and 147 as the test set. We used the benign and malignant categories in the breast US data and excluded images with no FG to segment (i.e. the "normal" category). The spleen data set contained 899 images, 601 of which were used as training, 82 as validation and 216 as test set images. The heart data set contained 829 images, 563 of which were used as training, 146 as validation, and 120 as test set images. We selected 2D axial slices from the spleen, and sagittal slices from the heart data sets, both of which were originally 3D volumes, such that there is at least one pixel corresponding to the FG. Care was taken to ensure that 2D slices were selected from 3D volumes and split at the patient level to avoid cross-contamination of images across training/test splits.

To vary $\mathcal{TS}$ of images in the test set, and to evaluate the robustness of the U-Net variants, speckle noise with variance 0.1 was added to both the FG and BG. This made the textures more similar, hence lowered $\mathcal{TS}$, and essentially rendered

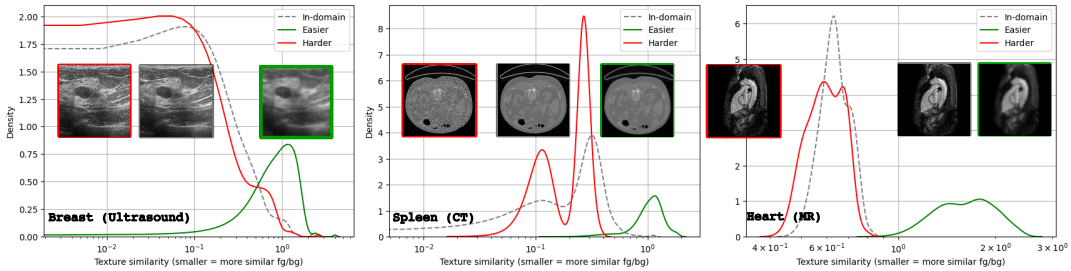


FIGURE 8.3: Medical data test sets on the TS ($\mathcal{TS}$) axis with in-domain (dashed gray), easier task (green, low similarity) and harder task (red, high similarity) distributions. Three modalities tested include US, CT, and MR, whose $\mathcal{TS}$ are in the same range as synthetic data in Fig. 8.2.

them harder to segment. This is shown in the red boxes in Figure 8.3. We also created another test set with textures that are less similar by blurring the BG using a Gaussian kernel of variance 3.0 while not blurring the FG pixels. These are shown in the green boxes in Figure 8.3, where it can be seen they are easier to segment.

8.2.3 Model architecture and training settings

The network architectures were created using MONAI [480] v1.1 and were all trained with random weight initialization. The U-Net was implemented with one input and output channel, with input image size set to 256×256 pixels across all experiments. The model had five levels with $16, 32, 64, 128, 256$ channels each for synthetic experiments and six levels (an additional level with 512 channels) for medical image experiments, all intermediate channels with a stride of 2. The ReLU activation was used, and no residual units were included. To reduce stochasticity, no dropout was used in any variant.

The NoSkip-U-Net was identical to the U-Net except for severed skip-connections. This led to the number of channels in the decoder to be smaller as there is no concatenation from the corresponding encoder level. The AGU-Net was setup to be the same as the U-Net, except with attention gating through the skip-connections.

The training parameters were kept constant across compared models for fair comparison. Our experiments[1] were implemented in Python 3.10.4 using the PyTorch implementation of the adam [487] optimizer. We set the learning rate to be $1e^{-3}$ for synthetic experiments (and $1e^{-2}$ for medical image experiments), maintaining it constant without using a learning rate scheduler. No early stopping criteria were used while training, and all models were allowed to train to 100 epochs (up to 300 epochs for medical image experiments). We trained our models to optimize the dice loss, and saved the model with the best validation DSC (evaluated once every two epochs) for inference on the test set.

We did not perform any data augmentation that could change the scale of the image content, thereby also changing the texture characteristics. Therefore, we only do a random rotation by 90 degrees with a probability of 0.5 for training, and no other augmentations. We also refrained from fine-tuning hyper parameters and did not perform any ensembling as our study design is not meant to achieve the best possible performance metric as much as it attempts to reliably compare performance across architecture variants while keeping confounding factors to a minimum. We

[1] Code made available at GitHub: amithjkamath/to_skip_or_not

therefore trained each model using the same random seeds (three times) and report the DSC statistics. Training and testing were performed on an NVIDIA A5000 GPU with 24 GB RAM and CUDA version 11.4.

8.3 Results

8.3.1 On synthetic texture variants

In-domain (performance): Figure 8.4 (left) indicates the relative improvement in DSC scores between the three U-Net variants using the NoSkip-U-Net as the baseline. To make the interpretation easier, the α value is used as a proxy for $\mathcal{TS}$ on the horizontal axis. It is worth noting that for α values > 0.3, there is negligible difference between the DSC performances of all the U-Net variants, indicating their ability with or without the skip-connections to learn the distributions of the FG and BG textures at that level of task complexity. Below α values of 0.3, the benefits of using attention gating in the skip-connections start to appear. This indicates that the benefit of attention-gating as a function of complexity is non-linear: models do not benefit from skip-connections at lower ranges of task complexity, but at larger ones, filtering the information flowing through the skip connections is important. What is interesting is also how the standard U-Net performance is noisy compared to NoSkip-U-Net, indicating that passing through the entire encoder feature map to be concatenated with the decoder feature maps may not always be beneficial.

Out-of-domain (robustness): Rows in Figure 8.4 (right) includes a heat map to represent the α values that the model was trained on, and columns correspond to the α value it was tested on. The entries in the matrix are normalized DSC differences between AGU-Net and NoSkip-U-Net (comparisons between standard U-Net and NoSkip-U-Net show similar trends). The diagonal entries here correspond to the AGU-Net plot in Figure. 8.4 (left). For α values 0.3 and 0.4 in training and 0.9 on testing (corresponding to an OOD testing scenario), the NoSkip-U-Net performs better than the AGU-Net, indicating that there indeed are situations where skip-connections cause more harm than benefit.

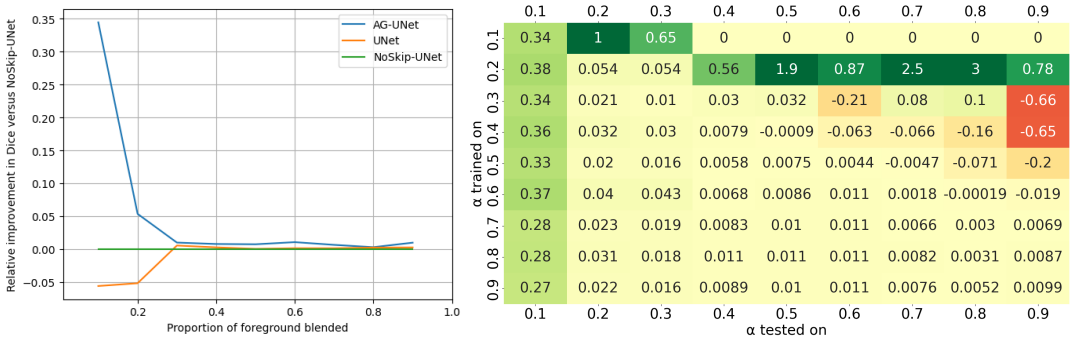


FIGURE 8.4: Relative performance in-domain (left) across U-Net variants, and out-of-domain robustness metrics (right) for AGU-Net versus NoSkip-U-Net.

8.3.2 On medical image textures

In-domain (performance): Looking at the "In-domain" rows in Table. 8.1, on all three data sets, the AGU-Net outperforms both the other variants. However, the relative improvements in performance vary across modalities, with the performance

differences on CT being the most stark. On the US data set, the NoSkip-U-Net outperforms the standard U-Net, supporting our hypothesis that skip-connections may not always be beneficial.

TABLE 8.1: Mean (standard deviation) of DSC for each of hard, in-domain and easy textures on the Breast (US), Spleen (CT) and Heart (MRI) data sets. Best performing model at each texture level is highlighted in bold.

Data set	Texture level	AGU-Net	U-Net	NoSkip-U-Net
Breast (US)	Harder	0.600 (0.304)	0.601 (0.323)	**0.635** (0.295)
	In-domain	**0.737** (0.260)	0.640 (0.304)	0.648 (0.281)
	Easier	**0.755** (0.244)	0.629 (0.280)	0.637 (0.267)
Spleen (CT)	Harder	0.310 (0.226)	0.074 (0.152)	**0.558** (0.265)
	In-domain	**0.927** (0.092)	0.745 (0.275)	0.606 (0.265)
	Easier	**0.809** (0.201)	0.394 (0.354)	0.486 (0.292)
Heart (MRI)	Harder	0.139 (0.242)	0.500 (0.316)	**0.815** (0.126)
	In-domain	**0.929** (0.055)	0.900 (0.080)	0.833 (0.111)
	Easier	**0.889** (0.073)	0.805 (0.129)	0.823 (0.103)

Out-of-domain (robustness): Focusing on the rows "Harder" and "Easier" in Table. 8.1, we observe for the US data set that the AGU-Net improves in the easier task, but declines in performance in the harder one. The drop in performance is most pronounced for the U-Net, but moderate for the NoSkip-U-Net. For the spleen data set, both the AGU-Net and the standard U-Net demonstrate severe drop in performance in the harder case. However, AGU-Net is better and the standard U-Net is worse than the NoSkip-U-Net in the easier texture situations. The heart data set shows the same trend as in the spleen data set.

8.4 Discussion & Conclusion

Through extensive experiments using synthetic texture images at various levels of complexity and validating these findings on medical image data sets from three different modalities, we show in this paper that the use of skip-connections can both be beneficial as well as harmful depending on what can be traded off: robustness or performance. A limitation of our work is that we vary only the FG in synthetic experiments but BG variations could demonstrate unexpected asymmetric behaviour. We envision the proposed analysis pipeline to be useful in quality assurance frameworks where U-Net variants could be compared to analyse potential failure modes.

9

How do 3D image segmentation networks behave across the context versus foreground ratio trade-off?

9.1 Introduction

Image segmentation is a ubiquitous task in medical image analysis, and recent advances using DL have yielded promising results [497]. The U-Net [265] and its variations have been in use since 2015, and is considered a reliable workhorse with wide ranging applications even beyond medical imaging. More recently, Attention-gated [278] and Transformer based models [282] have shown promising performance potential and are hence being considered for broader usage. These model architectures take as input a 3D volume containing anatomy to be labelled, and returns a volume of the same size with the corresponding voxels replaced by the estimated labels. An important parameter in all three models is the input volume dimensions, which we called patch size.

Case for smaller patch sizes: The Foreground to Background Ratio (FBR) of voxels is known to impact performance of segmentation networks [498]: this is of particular concern in medical image data sets. For example, the Spleen segmentation data set from the Medical Segmentation Decathlon (MSD) contains 41 training volumes, where the physical dimensions of the Spleen vary between 50000 mm^3 to 500,000 mm^3 (10x), while the entire volume imaged is typically 5,000,000 mm^3 (10x the largest Spleen) [468]. At best, 1 in 100 voxels (or less) is part of the foreground to be segmented, indicating why this is a challenging task. This makes the case for processing smaller patches, where the network is able to learn on data where this ratio is not so severely skewed.

Case for larger patch sizes: Smaller patches however mean that the model does not have as much global context of what it is attempting to segment, thereby losing out on the potential to learn these relative location-based details. Previous results [499, 500] indicate that larger patch sizes while training helps improve performance. Increasing the patch size causes the FBR to skew further, thereby intuitively making it harder for the network to identify the true foreground voxels. Furthermore, if the

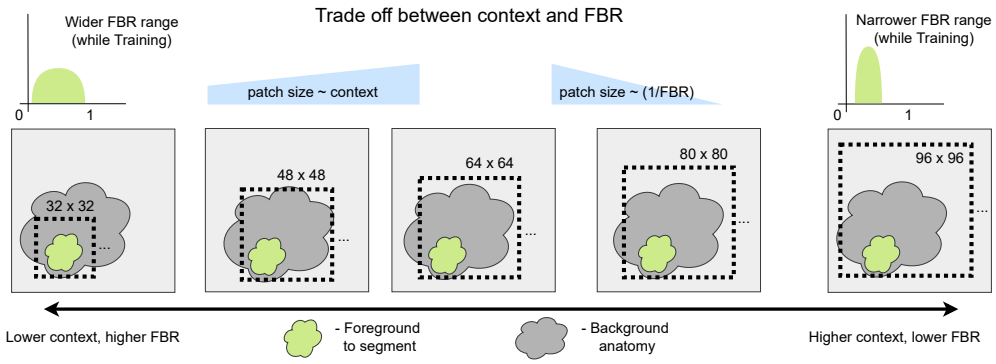


FIGURE 9.1: Do image segmentation networks need more context, or prefer a closer look at the foreground? We investigate the effects of varying the image/window size of the input to three families of segmentation network architectures.

range of sizes of the foreground has a large variance, this could make the networks' task even more challenging.

This leads to the central question behind this work. As shown in Fig. 9.1: how does the choice of patch-size (context versus FBR) impact the performance of three related but complementary model architectures: fully-convolutional (vanilla U-Net), attention gate-based: AGU-Net (with attention gates in the skip connections), and transformer-based: UNETR (where the encoder is replaced by a Transformer-based architecture)? We run our experiments on a controlled synthetic situation, and then observe if the trends occur as well on clinical data sets. We show that (1) larger patch sizes are preferred across the three network architectures, (2) UNETR and AGU-Net are more sensitive to patch size changes, and (3) ensuring a wide range in FBR during training is a prerequisite for robustness.

9.2 Materials and Methods

For vanilla U-Net, we use the implementation from MONAI with one input channel, two output channels, intermediate channel widths of $(16, 32, 64, 128, 256)$ and strides $(2, 2, 2, 2)$. We use two residual units, and batch normalization. For the AGU-Net, we use the same parameters as the vanilla U-Net for fair comparison. For UNETR, we use a feature size of 16, hidden size of 768, 3072 Multi-Layer Perceptron (MLP) dimensions, 12 heads, the perceptron position embedding and instance normalization. We also enable the residual blocks, and do not use any dropout.

Synthetic data experiments: we generate 100 volumes of size 96^3 voxels, with random noise generated with variance 0.8, constituting the background. The foreground consists of random spheres with radius between 25 to 35 voxels, with random centres. All the spheres are fully contained within the volume bounds. 70 of these are used for training and 30 as validation. We generate 100 more as the independent test set, where the foreground radius varies from 5 to 48 voxels (half the dimension of the volume), to evaluate how these networks perform across FBRs, including ones it has not encountered while training.

Clinical data experiments: we use the Spleen segmentation data set from the MSD. We choose this specific data set due to its large range of FBR. We use 26 volumes as training data, 5 as validation, and 10 volumes as the test set.

Training settings: the training batch size is 2 and we use Adam optimizer with an initial learning rate of $1e-4$ and cosine decay scheduling. For the clinical data, we train for 500 epochs and save the network with the best validation DSC. For the synthetic task, we train for 50 epochs, as it is a relatively easier problem. Each combination of network and patch size is tested thrice with random seeds and averaged results are presented to minimize noise from a single run.

We test five patch sizes $(32, 48, 64, 80, 96)$ symmetric in 3D, by using a random crop augmentation (implemented via MONAI transforms), with an equal proportion of volumes where the center voxel is foreground or background. 96 is the largest possible size due to memory and image data size constraints. We further record the FBR that the model sees while training, to track how segmentation metrics vary between within-range (called 'in-train') and outside. We do not employ any other augmentation or post-processing to avoid confounding our inferences due to them.

9.3 Results

From Table 9.1 we observe that larger patch sizes lead to better DSC for the synthetic experiment. Vanilla U-Nets outperform the other two in the 'in-train' range. Outside this range (see colour coordinated histograms at the bottom of Figure. 9.2), the performance drops considerably for UNETR and AGU-Net. This points to a lack of robustness to such drifts. Second row in Figure. 9.2 shows how the DSC vary for the Spleen data, where all the test samples are within range of training FBR. Table. 9.2 show that the trends we describe follow from synthetic to clinical data sets. Larger patch sizes are still preferred, and vanilla U-Nets outperform the other two architectures in this case as well. We believe the drop in robustness could be due to UNETR and AGU-Net having more generic inductive biases, hence needing more training data and larger context.

TABLE 9.1: Mean (stdev.) DSC for the synthetic data set over various patch sizes, within training foreground ratios and corresponding reduction outside the range.

Network	Patch Size: 32	Patch Size: 48	Patch Size: 64	Patch Size: 80	Patch Size: 96
U-Net (in-train)	**0.98** (0.02)	**0.99** (0.02)	**0.99** (0.00)	**0.99** (0.00)	0.99 (0.00)
drop outside	–	–	$\Delta = -0.02$	$\Delta = -0.01$	$\Delta = -0.02$
UNETR (in-train)	0.68 (0.35)	0.64 (0.39)	0.96 (0.12)	0.99 (0.01)	**0.99** (0.00)
drop outside	–	–	$\Delta = -0.73$	$\Delta = \mathbf{-0.37}$	$\Delta = \mathbf{-0.28}$
AGU-Net (in-train)	0.63 (0.34)	0.66 (0.37)	0.95 (0.13)	0.99 (0.00)	0.97 (0.02)
drop outside	–	–	$\Delta = \mathbf{-0.84}$	$\Delta = -0.30$	$\Delta = -0.26$

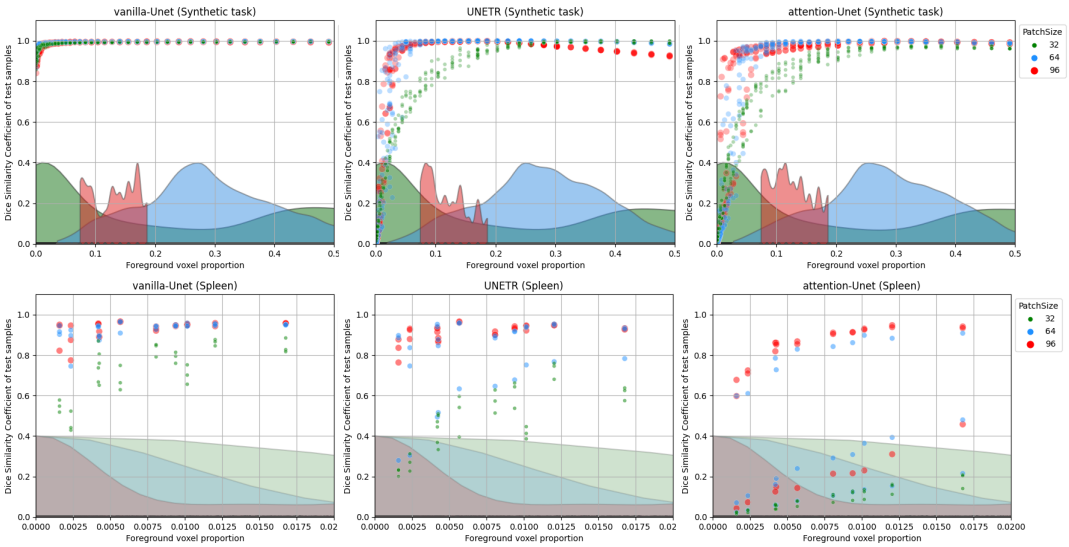


FIGURE 9.2: DSC metrics for the synthetic task (top row) and Spleen (bottom row): using U-Net (left), UNETR (middle), and AGU-Net (right). Distributions at the bottom indicate proportion of training samples with that FBR during training. Only patch sizes 32, 64, 96 shown for clarity.

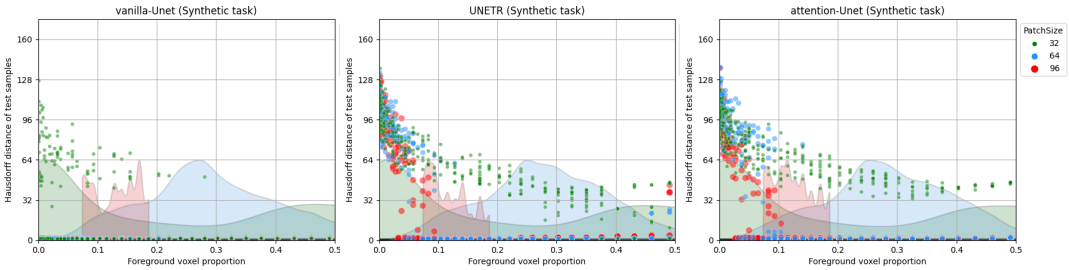


FIGURE 9.3: HD for the synthetic task using U-Net (left), UNETR (middle), and AGU-Net (right). Only patch sizes 32, 64, 96 shown for clarity.

TABLE 9.2: Mean (stdev.) DSC for the spleen segmentation over patch size variations.

Network	PatchSize: 32	PatchSize: 48	PatchSize: 64	PatchSize: 80	PatchSize: 96
U-Net	**0.721** (0.130)	**0.907** (0.045)	**0.928** (0.040)	**0.922** (0.047)	**0.932** (0.042)
UNETR	0.481 (0.158)	0.766 (0.178)	0.799 (0.186)	0.852 (0.116)	0.915 (0.042)
AGU-Net	0.086 (0.052)	0.102 (0.060)	0.384 (0.313)	0.582 (0.326)	0.634 (0.327)

Part III

Proofs of Concept Experiments

This part is focused on proofs of concept experiments. The articles included have been modified slightly from the published versions so as to not expand abbreviations already used.

Chapter 10 uses the C3D model [177] and generates a brute-force heat map of all possible local "atomic" perturbations on the surface of OARs for the purpose of dosimetric QA.

Chapter 11 is an extension of the qualitative survey in Chapters 3 and 5. This is joint work with Zahira Mercado, who was a master student advisee from August 2023 to February 2024. She is the first author of this work, and my contribution as the second author are at an advisory level along with assistance in extending the work done previously to creating the ranking scheme, organizing the code in GitHub: amithjkamath/autodoserank, and reviewing and editing the manuscript.

10

ASTRA: Atomic Surface Transformations for Radiotherapy Quality Assurance

10.1 Introduction

Approximately 45% of all malignant primary brain tumours are accounted for by aggressive tumours such as GBM [419]. Treatment consists of surgery, adjuvant RT, and concomitant and adjuvant chemotherapy [14]. RT planning aims to conform the dose to the TV (i.e., tumour or resection cavity, with adjacent areas of potential microscopic spread) while sparing OAR, thereby ensuring optimal tumour control and limiting normal tissue toxicity [420]. Accurate segmentation of the tumour TV and OARs is critical to achieving this objective. Radiation oncologists or dosimetrists draw contours around OAR and tumour TV, either manually or semi-automatically. This is time-consuming and can take up to seven hours per patient in the head and neck anatomy [421]. In a multi-institutional delineation study among radiation oncologists, incorrect TV segmentation has been reported to have caused 25% of non-compliant treatment plans [140]. Tumour TV and OAR segmentation are amongst the most error-prone and time-consuming steps in the RT planning process. This has led to efforts to create segmentation standards and develop RT QA systems [86].

Motivation for visualizing impact of local variations: Fig. 10.1 shows multiple cyan lines representing different brainstem segmentations performed by several expert radiation oncologists. The clinical target volume is shown in yellow, and the planning target volume in orange. The underlying heat map indicates the dose distribution (colour wash) for a treatment plan in units of Gray (red: high, blue: low dose) and provides the dose context. If OAR volumes are drawn larger than their actual extent, it leads to overestimating the OAR dose and potentially under-dosing the tumour TV to spare the OAR from excess dose. This negatively impacts tumour control. Conversely, missing the actual extent of the OAR (under-segmentation) would result in an underestimation of the true dose. This leads to excess toxicity to the OAR.

Treatment dose plan computation is currently done by medical physics experts and is a separate step in the RT process. Treatment delays occur if this dose is not

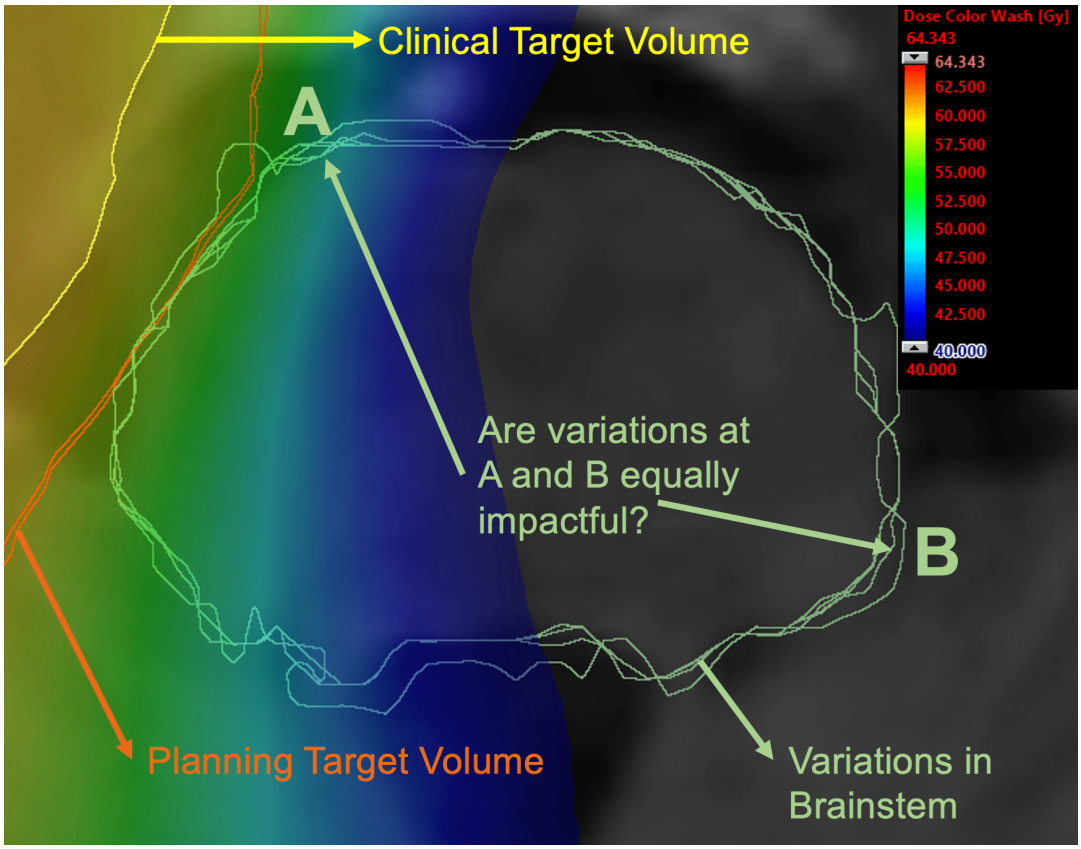


FIGURE 10.1: Visualizing variations in segmentation: are variations at locations A and B equally impactful from an RT perspective? To date, radiation oncologists review and correct segmentations without information on how potential corrections might affect radiation dose distributions, leading to an ineffective and suboptimal segmentation correction workflow.

protocol compliant and necessitates re-computation and reviewing any potentially incorrect segmentation. This time between image acquisition and RT planning completion is reported to be 9.63 days on average [422]. It is unknown beforehand if local corrections made at a specific region of the segmentation lead to better or worse dosimetric outcomes. As re-planning and dosimetric assessment are both time-consuming, they demand faster workflows. To address this, DL-based dose prediction models were proposed in a head & neck cancer challenge [423]. Recently, in [434], we presented the first such model for GBM. In this paper, we introduce an automated DL-based method called 'atomic surface transformations' for RT QA (ASTRA) that predicts the potential impact of local segmentation variations on RT dose predictions in the form of a sensitivity map for faster review of dosimetric effects of local segmentation changes. We believe this can help prioritize review efforts on the most critical areas.

Hypothesis and Contributions: Our hypothesis is that a DL dose prediction model offering near-instant dosimetry can aid radiation oncologists to focus their review efforts on locations where segmentation variations are dosimetrically most critical.

The proposed approach identifies locations of the OARs that most contribute to variations in dose impact, presenting radiation oncologists with an assessment that highlights the relative significance of local segmentation accuracies.

We demonstrate this using a novel method called ASTRA, which performs voxel-wise local transformations in a high-throughput manner across the OAR's surface, to then compute the mean absolute difference between the dose predicted with and without transformations to estimate the dosimetric impact of local segmentation changes. Our contributions in this paper are threefold:

- First, we introduce and show representative visualizations with ASTRA - modelled as spheres added onto the surface of the OAR segmentation, to evaluate how these local variations impact radiation dose computations.
- Then, we analyse the sensitivity of dose predictions to atomic surface transformations (with over 2000 inference predictions made on 10 test subjects each) and compute correlations with the smallest distance-to-TV and local dose gradient. We observe stronger correlation between dose changes and distance-to-target (on larger OARs), and a weaker correlation with local dose gradient magnitude (across most OARs).
- Finally, we show how the dose changes are impacted by the size of the transformation - using three different sizes of spheres to simulate segmentation changes of varying magnitudes. We show that the sensitivity map is robust to changes in this parameter.

10.2 Materials and Methods

Data: Our data set included imaging and segmentation data from 100 patients diagnosed with GBM. This included CT imaging data, and binary segmentation masks of 13 OARs as well as the PTV. Each of these subjects also had a reference dose plan, calculated using a standardized clinical protocol with Eclipse (Varian Medical Systems Inc., Palo Alto, USA). This reference was a double arc co-planar VMAT plan with 6 mega volt flattening filter free beams, optimized (Varian photon optimizer version 15.6.05) to deliver 30 times 2 Gray while maximally sparing the OARs. Sixty randomly chosen subjects formed the training set, 15 were used as validation (five samples excluded due to missing segmentations) and the rest of the 20 were used as the test set.

Model: We used a two-level C3D U-Net [177] as the dose prediction network (i.e, the input to the second U-Net is the output of the first concatenated with the input to the first U-Net). The model input was a normalized CT volume and binary segmentation masks for each of the 13 OARs and TV, and predicted a continuous-valued dose volume (up-scaled from [0, 1] to 0 to 70 Gray) of the same dimension as the input. The loss was computed as a weighted sum of L1 losses between outputs of the first and second U-Nets versus the reference dose. All volumes were resampled to 128^3 voxels, due to GPU memory constraints. As indicated in [434], this model had a MAE of 0.906 Gray, making it suitable for this work.

What are Atomic Surface Transformations: Fig. 10.2 describes the process of generating 'atomic surface transformations' to estimate the relative impact of local variations in segmentation to overall dose predictions. From the baseline segmentation an initial dose prediction is generated. Then, a sphere with a radius of 3 (equivalent to approximately 7.5 mm in the axial plane) is added to a single point on the surface of a single OAR, yielding a perturbed segmentation, on which a new dose prediction is generated. The mean absolute difference between these dose predictions is computed and recorded at the centre of the corresponding sphere. This is

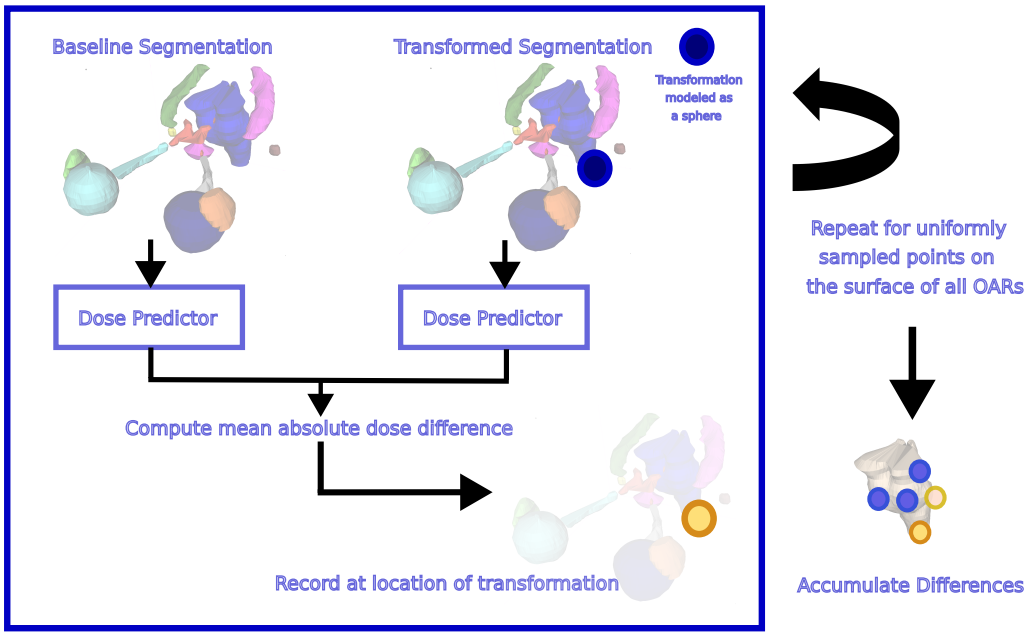


FIGURE 10.2: How are sensitivity maps constructed from 'atomic surface transformations'? We use a DL based dose prediction inference model sampled uniformly at each perturbed point on the surface of each OAR.

repeated over a uniformly sampled set of points on the OAR's surface, yielding a sensitivity map, as shown in Fig. 10.3.

Sensitivity experiments: Beyond constructing the sensitivity maps using 'atomic surface transformations', we run two experiments to evaluate the ability of sensitivity maps to describe dose-sensitive areas of segmentation changes. We first look at the correlation between the mean average dose difference due to an atomic surface transformation with the minimum distance of the transformation location to the tumour TV. We expect that points closer to the tumour TV will be more highly impacted by dose differences, indicating their relative importance to the segmentation quality. The second experiment is to compute the correlation of dose differences against the local magnitude of the dose gradient. The hypothesis here is that steeper dose gradients indicate regions where dose estimates have to be sensitive to segmentation changes, leading to a positive correlation.

Transformation size experiments: Next, we investigate what happens when the size of the atomic surface transformations change - and how, if at all, would it impact dose predictions. The size in physical units of each voxel is generally in the range of 2.7 mm in the axial plane and 2 mm in the transverse direction. We experiment with transformations of radii $3, 5, 7$, on the brainstem of two test set subjects to determine how the dose predictions respond to larger changes.

10.3 Results

Visualizing Atomic Surface Transformations: Fig. 10.3 includes four examples of visualizing the relative importance of local regions on the surface of segmentation for its impact on dose prediction. The tumour target volume (represented in red) and is displayed as an isosurface. The sensitivity map is overlaid on the surfaces of five OARs: brainstem, eye (left and right) and hippocampus (left and right). The colours

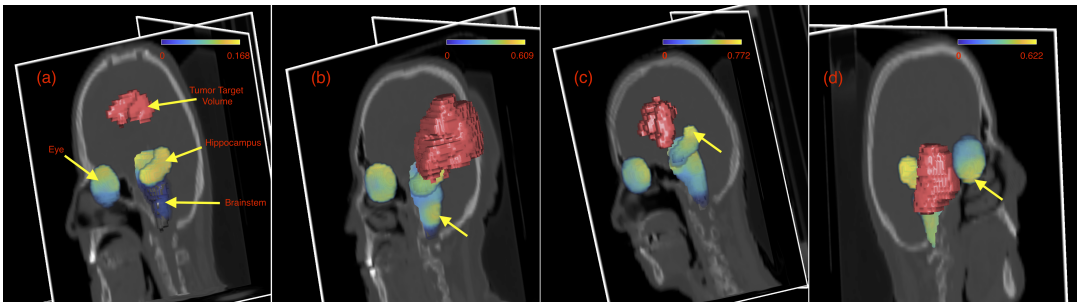


FIGURE 10.3: Visualizing atomic surface transformations: demonstrated using selected OARs. Tumour TV is shown in red; brighter yellow regions overlaid on OAR segmentations are most impactful on dose predictions, while darker blue regions describe the lowest impact. (a) to (d) demonstrate increasingly interesting situations: simple; large tumour; complex tumour shape; and tumour close to hippocampus and brainstem.

TABLE 10.1: Sensitivity to segmentation changes per OAR measured as mean absolute dose difference (in Gray) in the brain, averaged over all the transformation points on its surface.

ID	Brainstem	Eye L	Eye R	Hipp. L	Hipp. R
(a)	0.0069	0.0115	0.0092	0.1041	0.0819
(b)	0.0163	0.0485	0.0593	0.3737	0.2710
(c)	0.0097	0.0419	0.0286	0.3085	0.2626
(d)	0.0191	0.0613	0.0514	0.3673	0.3151

indicate the mean absolute dose difference in the brain region due to a transformation at that location, and are measured in Gray. This is done using the MATLAB® Medical Imaging Toolbox. Table. 10.1 indicates the average mean absolute dose difference per OAR for the same four cases in Fig. 10.3.

Fig. 10.3 (a) shows a relatively small tumour TV far from the entire OAR (about 3 cm). The most impactful points on the OARs are indeed the ones that are closer to the tumour TV, while the magnitude of these changes are small (maximum of 0.168). Fig. 10.3 (b) shows a larger tumour TV, with a wider distribution of dose differences across the surfaces of the OARs. The maximum mean absolute dose difference is also higher at 0.609 on the left hippocampus. The yellow arrow indicates a region on the surface of the brainstem that is away from the tumour TV, yet has more impact on dose than points above it closer to the tumour TV.

Fig. 10.3 (c) has a tumour TV with a complex shape, leading to higher overall mean absolute dose differences. The yellow arrow indicates a region at the edge of the left hippocampus to be the most impactful, as it likely lies on critical beam angles. Fig. 10.3 (d) shows how the map changes when the tumour is closer to the OARs. Note the yellow arrow indicating the inferior parts of the eye more sensitive than superior, which differs from other cases. Another interesting point is both the hippocampi have a higher average mean absolute dose difference (see Table. 10.1) even though the tumour TV being close to only one of them. These observations are non-trivial and demonstrate the utility of this approach.

Sensitivity analysis: We demonstrate these results on ten test set subjects, featuring more than 2000 transformation points (more points on larger OARs like Brainstem and Eye) across all 13 OARs per subject. Table. 10.2 shows the correlation between the smallest distance to tumour TV from the point of transformation and the

TABLE 10.2: Correlation between dose difference and minimum distance to tumour TV (XCorr - dist), and local gradient of dose (XCorr - grad). Average size of the OAR (in voxel units) included for reference.

OAR	Size	XCorr (dist)	XCorr (grad)
Brainstem	2133.14	-0.39	0.22
Chiasm	48.71	-0.20	0.05
Cochlea L	8.14	-0.17	0.18
Cochlea R	8.42	-0.33	0.18
Eye L	637.42	-0.76	0.06
Eye R	628.14	-0.63	0.01
Hippocampus L	195.00	-0.35	0.41
Hippocampus R	206.85	-0.14	0.50
Lacrimal Gland L	78.57	0.58	0.39
Lacrimal Gland R	80.28	0.01	-0.22
Optic Nerve L	55.28	-0.03	0.05
Optic Nerve R	63.42	0.29	-0.15
Pituitary	61.00	-0.26	0.26

mean average dose difference caused due to atomic surface transformations. Note that for larger OARs, like brainstem and eyes, the correlation is more strongly negative, indicating that points that are closer to the tumour TV are more impacted by the transformations from a dose difference perspective. The correlation was found to be lower for smaller OARs like the cochlea or the lacrimal glands because the extent of these OARs does not vary much with respect to the distance from the tumour TV.

A similar analysis was done to compute the correlation between local dose gradient magnitude versus the mean average dose difference. Results show a lower correlation than correlations obtained for distance-to-targets. However, as expected, there is still a positive overall correlation, confirming that dose estimates are sensitive to segmentation changes in locations where dose gradients are higher in magnitude.

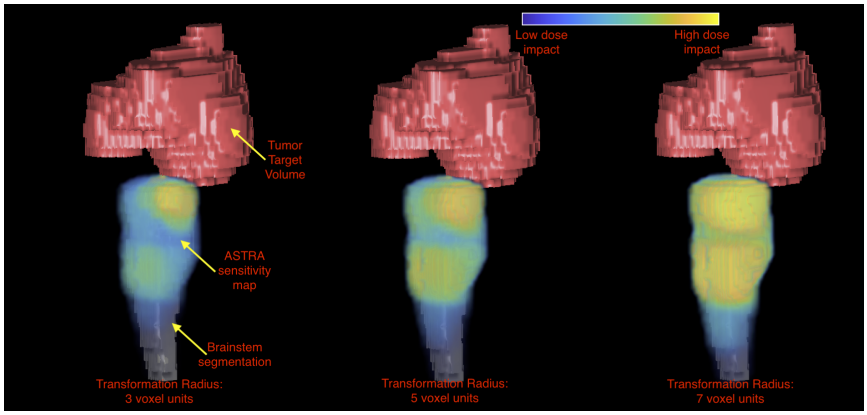


FIGURE 10.4: When radius of atomic surface transformation is varied, the dose impact scales appropriately: results on representative brainstem with more than 900 atomic surface transformations. Radius measured in voxel units.

Transformation size analysis: Fig. 10.4 shows the impact of changing the radius of the atomic surface transformation in the range $3, 5$ and 7. The dose differences correlate well by location (0.949 between radius 3 and 5, and 0.939 between 5 and 7), indicating the stability of finding important regions across this parameter.

Discussion: In this paper, we demonstrated with experiments on a test set of ten subjects with more than 20000 transformations at various points on the surfaces of OARs (each being an inference run of our deep learning-based dose prediction model) that the proposed 'atomic surface transformation' method can identify locations on segmentation boundaries that most impact dose computation. This would not have been practically feasible to compute with pre-DL methods. A limitation of this setup is that separate models need to be trained for every tumour location, delivery machine, and planning software. We deem this to be minor in practice, though, but also a target of potential future research.

Through this work, we hope to assist the clinical workflow of reviewing segmentations by drawing the attention of radiation oncologists to specific local regions of highest sensitivity to the RT plan. We believe clinicians would benefit from more detailed feedback in the form of sensitivity maps rather than a single numeric score. Next, we hope to improve the interpretation of the visualization and link it with dose constraints, such that locations, where transformations cause changes beyond the maximum dose limits for each OAR, could be highlighted. In the longer term, we plan to build on this further to create an automated dosimetry-aware segmentation QA system.

11

AutoDoseRank: Automated Dosimetry-Informed Segmentation Ranking for Radiotherapy

11.1 Introduction

GBM is a particularly aggressive brain tumour that constitutes approximately 45% of all brain tumours [419]. The conventional treatment strategy includes surgery, RT, and chemotherapy, with the objective of precisely targeting the tumour while sparing as much healthy tissue (also known as OAR) as feasible [14]. This necessitates a careful equilibrium between tumour coverage and tissue toxicity control. An integral component of this procedure is the segmentation of structures, which can be laborious when performed manually [421]. The entire process, from the patient's initial scan to the commencement of the RT treatment, can take more than a week on average, with segmentation and RT planning being the most time-intensive stages [422]. The advent of DL-based auto-segmentation is anticipated to significantly decrease this time.

This suggests that the role of medical professionals will transition from performing the segmentation task to supervising and rectifying the outcomes generated by automated systems. The task of ensuring the quality of an automated segmentation is crucial, as errors in tumour segmentation have been shown to cause 25% of treatment plans to be non-compliant, potentially leading to untreated tumours or harmful radiation doses [140]. It's crucial to emphasize that in RT, the impact of segmentation variations on the administered dose is a determining factor in clinical outcomes as it influences tumour control and toxicity to healthy tissue. There is an urgent need for dosimetric considerations in the process of assessing and measuring the effects of corrected or alternative automated segmentations [425].

In the context of computer vision and natural image scenarios, geometric measures such as the DSC and HD are frequently employed to assess the quality of segmentations. These measures have also been utilized to automate the evaluation of segmentation quality [428–430]. However, these metrics do not strongly correlate with either the dosimetric consequences of contouring errors [101] or the assessment

of clinicians [426]. Recent studies [501, 502] further underscore the importance of incorporating domain knowledge when selecting and designing metrics.

The integration of dosimetric effects into the evaluation process calls for the computation of dose plans. These computations are time-consuming and necessitate a collaborative effort between the radiation oncologist and dosimetrists or medical physicists. Due to its time-intensive nature, dose-guided evaluation of segmentation quality has not been commonly employed in clinical environments.

With recent advances in DL-based models for RT planning, a dose prediction model has been used to guide radiation oncologists on the volume slices that require manual modifications [431]. This segmentation editing tool demonstrates potential for time efficiency while maintaining dosimetric equivalence with dose distribution maps produced without its assistance. Furthermore, methods that utilizes a DL-based dose predictor to assess the impact of local segmentation alterations on dosimetric outcomes have also been described [432]. However, it mainly concentrates on OAR segmentation and not on tumour lesions, which hold more clinical significance due to the increased variability in the segmentation task. Moreover, these methods do not offer a technique for ranking different segmentation candidates or alternatives based on their dosimetric effects, which is of higher clinical interest to assist medical experts in the supervision and correction of automated segmentation results. These could potentially lead to more precise and efficient treatment plans.
Contributions: We propose the first (to the best of our knowledge) dose-informed framework for ranking a set of segmentations, taking advantage of recent advancements in DL-based dose prediction. This dosimetric triage of segmentations is based on (i) OAR-specific dose constraints and (ii) relative prioritization between OARs. This effectively brings forward into the workflow of evaluating segmentation quality, knowledge from the subsequent step of RT planning for brain tumour patients, thereby making the process more clinically relevant. We demonstrate the ability of the proposed approach, termed AutoDoseRank for **Auto**mated **Do**simetry-informed **Se**gmentation **Rank**ing, by comparing and analysing its performance against four radiation oncologists.

11.2 Methods

AutoDoseRank is made up of two primary elements: (i) a DL-based dose predictor to estimate dose distributions for a segmentation candidate; and (ii) a ranking module that examines a set of such segmentations based on clinically interpretable parameters, including OAR specific dose effects and planning prioritization. These components together help construct clinically relevant segmentation candidate rankings.
Formally, given an initial original segmentation S_{orig} and corresponding segmentation candidates $\{S_a\}_{a:1,..,n}$, corresponding dose predictions, D_{orig} and $\{D_a\}_{a:1,..,n}$ are computed using a DL-based dose predictor model $DP(S) \rightarrow D \in \mathbb{R}^{W \times H \times C}$, for the original and candidate segmentations, respectively. In practice, the candidate segmentations could be generated as part of an expert's online correction workflow or come as proposals from an auto-segmentation model. Each voxel in D corresponds to the local predicted dose in Gray (Gy). We utilize a previously reported model [434], with an average dose scores of 1.38 Gy [418]. The novelty in AutoDoseRank beyond using dose predictors to inform quality assessments to help rank and triage candidate segmentations is a new decomposable patient-level metric for estimating dose impact, DI. DI can be decomposed into per-OAR impact, DI_i, computed using

the predicted dose map $\tilde{D}_a = D_a \cdot m_i$ and the original dose map $\tilde{D}_{orig} = D_{orig} \cdot m_i$, where m_i is the mask for the i-th OAR.

$$DI_i = \frac{\Phi(\tilde{D}_a) - \Phi(\tilde{D}_{orig})}{C_i} \left(\frac{\Phi(\tilde{D}_{orig})}{C_i} \right)^{\gamma} \qquad \textit{OAR level} \tag{11.1}$$

$$DI = DI(S_a) = \frac{\sum_i \frac{1}{p_i} DI_i}{\sum_i \frac{1}{p_i}} \qquad \textit{patient level} \tag{11.2}$$

D_{orig} corresponds to the dose map of the original segmentation S_{orig}. $\Phi(D)$ represents either the maximum or mean dose computed from the dose map D within the boundaries of the OAR [418]. C_i corresponds to the dose constraint for the i-th OAR, complying with current clinical guidelines. The parameter γ enables regulating the penalization at the OAR level when the candidate alternative dose metric approaches the OAR-specific dose constraint. In our experiments, parameter $\gamma = 1$ for a linear penalization at the patient level. Finally, p_i corresponds to the priority of the i-th OAR, which is specified based on clinical recommendations and used within the TPS (e.g., p("eye")=8, p("optic nerve")=1, etc.).
Dose impact values, denoted as $DI(S_a)$, are obtained for each candidate segmentation. These values are then ranked in ascending order to establish the final ranking. The intuition behind equation (1) and (2) is to penalize dose variations caused by a candidate segmentation. This is relative to the OAR-specific constraint (i.e., Eq. (11.1) OAR level) further intensified when the candidate dose metric approaches the corresponding OAR specific dose constraint.

11.2.1 Data

Our evaluation data set comprises 65 segmentation candidates from 13 GBM patients. This dataset includes CT scans, T1 contrast-enhanced MRI, and binary segmentation masks of 13 OARs, as well as the tumour TV. The OARs consist of the brainstem, optic chiasm, cochleae, eyes, hippocampi, lacrimal glands, optic nerves, and the pituitary gland. In addition, each case also includes a ground truth dose plan, computed with Eclipse (Varian Medical Systems Inc., Palo Alto, USA), and a predicted dose plan using a C3D U-Net (trained separately on 60 independent patients) [434]. The ground truth is computed using a standard clinical protocol with double arc co-planar VMAT plan with 6 mega volt flattening filter-free beams. It is optimized (using Varian photon optimizer version 15.6.05) to deliver 30 fractions of 2 Gy each, while sparing the OARs as much as possible. The dose is calculated using the AAA algorithm [424], and normalized such that the prescribed dose fully covers 50% of the TV. All volumes are resampled to an isometric $2 \times 2 \times 2$ mm grid of size 128^3 voxels using PyRaDiSe [433] and converted to NIfTI files for training and evaluation purposes.

11.2.2 Experimental Setup

For each of the 13 patients, an experienced RT planner manually generated four distinct segmentation candidates, leading to a total of 65 segmentations (i.e, 13 original segmentations $+13 \times 4$ candidates $= 65$ segmentations), on which dose plans are calculated with Eclipse (Varian Medical Systems Inc., Palo Alto, USA). Each original segmentation is ranked per case using the dose impact formulation (Eq. 11.1) using Eclipse (clinical TPS) dose map distributions instead of the DL-based dose

predictions. Both the candidate modifications and the creation of RT plans are time-consuming and effort-intensive tasks. Figure 11.1 shows an example of a original and corresponding segmentation candidates.

Expert-based Ranking of Segmentation Candidates:

We perform evaluation sessions for each test case with four experienced radiation oncologists, who are asked to rank the four candidates based on their potential negative impact on the dose received by the OARs. We use 3D Slicer (v5.6.0) to display all three slice planes. We also provide a 3D view of the spatial relationship between the OARs and the tumour TV, highlighting the location of the candidate alteration. All variations for each subject are presented simultaneously, allowing for visual comparison against the original.

Evaluation:

We adopt the following metrics, commonly used to evaluate rankings: (1) The Normalized Distance-based Performance Measure (NDPM), a ranking metric that quantifies the accuracy of a ranking [503]. NDPM yields a value in the range of $[0, 1]$, where 0 indicates an ideal ranking and 1 indicates the maximum deviation from the ideal ranking. Essentially, NDPM imposes a penalty on a ranking proportional to its deviation from the perfect ranking (2). Kendall's Tau is a statistical measure used to quantify the ordinal association between two measured quantities [504]. It is a non-parametric test that measures rank correlation, or the similarity of the orderings of the data when ranked by each of the quantities. The value of Kendall's Tau is in the range of $[-1, 1]$, where a value of 1 indicates the ranking's complete agreement, -1 indicates the rankings are the exact reverse of each other, and a value of 0 indicates no relationship. Kendall's Tau can provide valuable insights into the relationship between two sets of rankings. We also perform a bootstrapping analysis of the Kendall Tau ranking correlations to obtain further insights into the rankings under sampling variability. We report results performing 1000 sampling with replacement and 90% Confidence Interval (CI).

11.3 Results

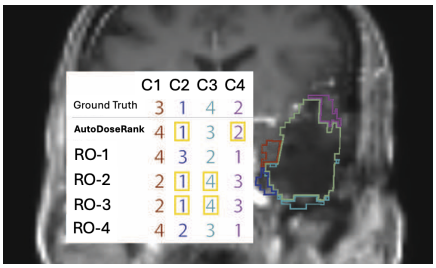


FIGURE 11.1: A representative example of original tumour TV segmentation (green outline) with four candidates denoted as: C1 (red), C2 (dark blue), C3 (turquoise) and C4 (purple). The table inlaid shows how Ground Truth (Eclipse), AutoDoseRank, and four experts (RO-1 to 4) rank these in order of dose impact. Yellow boxes indicate correct matches with the ground truth.

Figure 11.1 presents a typical instance of a case that includes the original segmentation, four alternative candidates, and the corresponding rankings given by AutoDoseRank and four experts. In general, the experts vary significantly in time taken to complete the ranking task, from 19 to 138 seconds per candidate. AutoDoseRank on the other hand takes a constant time quantum (of $\sim$ 30 seconds), regardless of geometrical complexity.

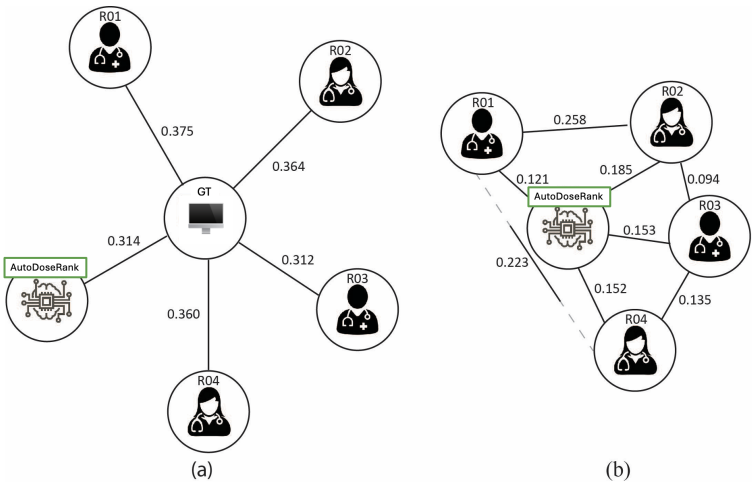


FIGURE 11.2: NDPM visualization with scaled distance lengths representing value: (a) NDPM compared to Eclipse (ground truth) with a scaled distance of 0.75 (b) Cross-references between AutoDoseRank and the radiation oncologists with scaled distance of 1. Note: RO1 vs RO4 shown with a dotted line; black continuous line indicates correct scale.

Figure 11.2 shows the NDPM as a graph. This metric compares Eclipse, AutoDoseRank, and each of the four experts. A lower NDPM value indicates a higher level of agreement. As shown in Figure 11.2(a), AutoDoseRank and expert "RO-3" exhibit similar ranking performance (two lowest NDPM), outperforming the other three experts. Figure 11.2(b) presents NDPM within the group of experts and AutoDoseRank, showing a higher level of agreement among them compared to their agreement with the ground truth (Eclipse). Furthermore, AutoDoseRank's agreement with the experts is more consistent than the inter-rater agreement, indicating that it provides more reliable rankings.

TABLE 11.1: Summary of Kendall's Tau ranking correlation performed with 1000 resamplings comparing AutoDoseRank and the four experts, denoted as RO-1 to -4, to the ground truth. RO-3 and AutoDoseRank yield higher correlations ranging from weak to moderate, outperforming the others. RO-3 is the most experienced and meticulous expert, who also took the longest time to perform the task.

Ground Truth (Eclipse) vs	**Mean Kendall's Tau**	**90% CI**
AutoDoseRank	0.129	[-0.097, 0.354]
RO-1	0.014	[-0.194, 0.231]
RO-2	0.038	[-0.163, 0.239]
RO-3	0.148	[-0.056, 0.347]
RO-4	0.041	[-0.161, 0.238]

Table. 11.1 shows Kendall's Tau ranking correlation summary performed with 1000 resamplings comparing AutoDoseRank and the four radiation oncologists to the ground truth (Eclipse). AutoDoseRank shows a stronger correlation (between weak to moderate) to the ground truth rankings and outperforms three out of four experts. Interestingly, AutoDoseRank performs similarly to RO-3, who is the most experienced and meticulous amongst the four.

To complement Table. 11.1, and obtain more insights into the distribution of Kendall's Tau ranking correlations, Figure 11.3 presents the Cumulative Distribution Function (CDF) of Kendall's Tau metric values, calculated over 1000 bootstrapping resamplings for AutoDoseRank and the four experts. The distribution of these values corroborates the NDPM findings, indicating that AutoDoseRank and expert 'RO-3' outperform the other three experts. This superior performance is highlighted by the noticeable gap between their CDF curves in Fig. 11.3.

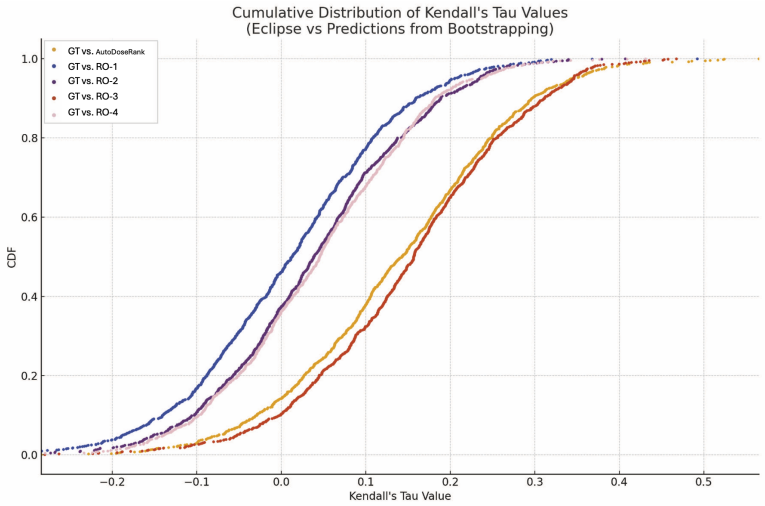


FIGURE 11.3: Comparative CDFs of Kendall's Tau Correlation Coefficients: Ground Truth versus AutoDoseRank and four experts.

Ablation experiment on priority: Figure 11.4 shows the Cumulative Distribution Function of Kendall's Tau Correlation coefficients of AutoDoseRank with and without applying different priorities to each OAR. This assesses the importance of prioritization used when computing the patient-level dose impact. We observe a clear worsening in the ranking with no prioritization, denoted as AutoDoseRank-No-Priority.

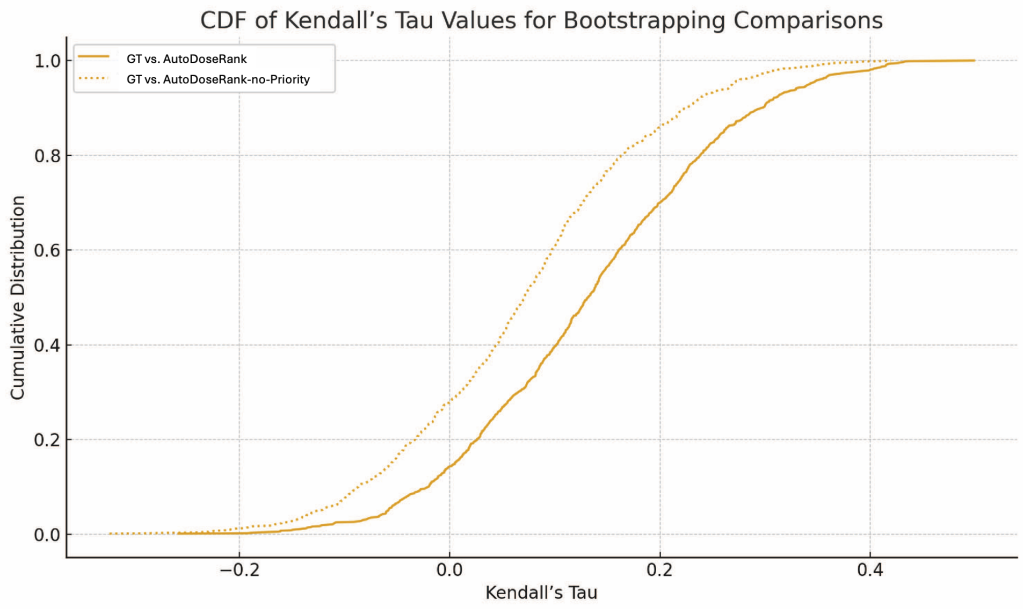


FIGURE 11.4: Ablation on OAR prioritization: Comparative CDFs of Kendall's Tau Correlation Coefficients: The ground truth versus AutoDoseRank with and without priority weighting.

11.4 Discussion and Conclusion

In this paper, we present AutoDoseRank, a dose-informed DL-based approach to rank segmentation candidates, aiming at aiding radiation oncologists in the task of monitoring segmentation quality. AutoDoseRank outperformed three out of four experts while being slightly outperformed by the most experienced and meticulous one. This highlights the capability of this framework to assist clinicians in monitoring segmentation quality semi-autonomously.

Bearing in mind that the clinician's time for running these human versus automated-model evaluations is not easily scalable, we emphasize the multi-expert evaluation presented, along with some limitations. Using more than 13 clinical cases to evaluate the approach could certainly improve the statistical robustness of our findings. The segmentation and planning preparation process consumes more than 70% of the total time from diagnosis to treatment, which typically spans over two weeks [422]. Consequently, producing meaningful segmentation candidates and re-planning corresponding dose maps using Eclipse (clinical TPS) is a significant time investment. We aim to mitigate this by building more accurate dose prediction models. Such models could also help us evaluate the sensitivity (by modifying γ) of the analysis to dose changes. Furthermore, we plan to run these evaluations with more radiation oncologists and dosimetrists to assess how experts vary amongst themselves versus future iterations of AutoDoseRank. Beyond the presented results, we believe this framework can be used in training radiation oncologists to improve their dosimetric awareness to segmentation variations.

Part IV

Conclusion and Perspectives

12

Conclusions

"I was born not knowing and have had only a little time to change that here and there."

— Richard Feynman, in What Do You Care What Other People Think?, 1988.

The findings presented in this thesis span three main areas: clinical validation (Part I), technical investigations (Part II), and proof-of-concept demonstrations (Part III). Together, they address key challenges in radiation therapy planning and contribute to progress in the field.

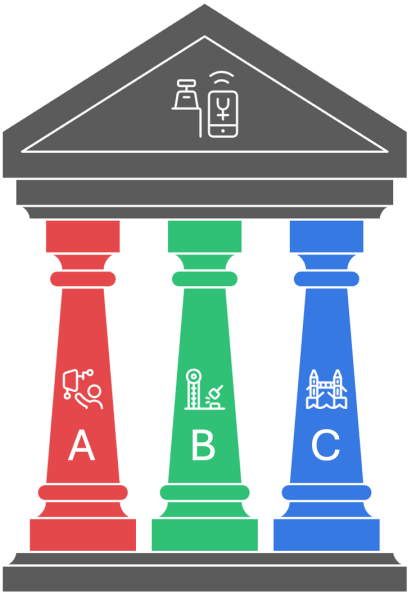
This chapter revisits the hypothesis and research questions introduced in Chapter 1. It provides a critical summary of the main findings, discusses their limitations, and reflects on their impact on clinical practice and future research. The guiding themes of this work are summarized in Figure 12.1.

AI-based systems can standardize and automate fast and reliable dosimetric contour QA in RT planning.

Chapters 3 to 11 provide strong support for the central hypothesis and advance the field by aiming to *improve the precision, efficiency, and reliability of RT using advanced AI tools*. Through the integration of **dose-aware QA, predictive dosimetry, and robust model design**, this work lays the foundation for more efficient RT workflows, with the potential to enhance patient outcomes.

The overarching impact of these results can be summarized into:
1. Standardizing Dosimetric Contour QA: shifting from geometry-based to clinically relevant metrics. A central finding of this work is the need to reduce inter-evaluator variability in contour assessment, as highlighted in Chapters 3 and 5. Manual contour errors contribute to up to 25% of non-compliant treatment plans [140]. DL models assessed in Chapters 6 through 11 show strong performance in estimating the dosimetric impact of realistic TV contour variations—matching or exceeding expert accuracy.

By focusing on dosimetric outcomes rather than geometric similarity, these tools support more objective and clinically aligned QA. This is important given the weak correlation between geometric metrics and treatment quality [101]. These results support the development of RT QA workflows that prioritize clinical relevance and



A: Clinical Understanding and Needs Validation
A1: Variability amongst radiation oncology professionals.
A2: Formulating sound dosimetric criteria.

B: Technical Investigations and Analysis
B1: Correlation between predicted and true dose differences.
B2: Reliability of core deep learning architecture choices.

C: Proofs of Concept Studies
C1: Focusing contour reviews on dosimetrically critical areas.
C2: Automate dosimetric ranking of auto-contour proposals.

FIGURE 12.1: The three pillars on which the experiments to test the hypothesis of this thesis stands: A: validating the clinical need, B: investigations into the reliability and speed of technical solutions, and finally C: proof-of-concepts of novel clinical applications. Sub-questions within each of these pillars are listed on the right.

consistency over subjective judgement.

2. Automating Workflows for Faster Planning: enabling real-time, dosimetry-informed QA that can exceed expert performance in specific tasks. Chapter 3 highlights the need for standardized, automated contour QA. Chapters 4 and 6 demonstrate that DL models can support this by reducing the burden of manual checks. These models enable near-instant predictions of dose changes from contour edits, helping clinicians assess trade-offs between tumour coverage and OAR sparing more efficiently.

This capability shortens the planning and review cycle—an important factor in reducing delays between diagnosis and treatment [187]. Chapter 10 introduces ASTRA, the first dose-aware sensitivity map that directs clinicians to high-impact areas needing correction [85]. Chapter 11 presents AutoDoseRank, which prioritizes urgent interventions, particularly relevant for aggressive cancers like glioblastoma [144].

Combining dose prediction with auto-contouring enhances planning precision and speed, supporting more personalized and efficient care. This is especially valuable in busy clinical settings, where planning delays can compromise outcomes [140].

3. Robustness Testing of Architectures: evaluating design choices for reliable performance under clinical variability. Chapters 7, 8, and 9 examine segmentation model design, focusing on robustness across varied clinical settings. A key finding is that skip connections, while improving accuracy in controlled conditions, reduce robustness in OOD scenarios. This trade-off is critical in tumour segmentation, where imaging varies across scanners and patient conditions [337].

While models like AGU-Net and UNet++ may perform well in ideal settings,

NoSkip architectures or those with addition-based skip connections (e.g., V-Net) offer better reliability. The results also highlight the role of dataset diversity—specifically, texture variation and foreground-background balance—in improving generalizability, as shown in frameworks like nnU-Net [285].

Computational efficiency is another factor; for example, UNet++ can be up to 100× slower to train. These insights support the development of adaptive, efficient, and texture-aware models that maintain accuracy while ensuring robustness in real-world clinical use.

The following sections, 12.1, 12.2, and 12.3, present the key conclusions from the three core parts of this thesis (Parts I, II, and III), and explain how they address the research questions outlined in Chapter 1.

12.1 Clinical Understanding and Needs Validation

Findings from Part I highlight the urgent need for automated and standardized dosimetric contour quality assessment. Chapter 3 quantifies inter-observer variability in this task, while Chapters 4 and 5 show that DL-based dose prediction models can support automation and improve consistency in evaluation.

Is there a need to standardize and automate dosimetric contour QA?

A1: What is the variability among radiation oncology professionals in contour quality assessment?

Summary: Chapter 3 explores this through a qualitative study involving four radiation oncologists and three medical physicists. They evaluated 54 glioblastoma target contour variations across 14 patients. The results showed **high variability** in assessments. Cohen's Kappa values ranged from 0.33 to 0.74, indicating only fair to moderate agreement. Evaluators misclassified 46% of "no change" cases as "worse" and failed to identify any of the four "better" contours. Performance varied widely: the best evaluator achieved a precision of 0.63, recall of 0.65, and F1 score of 0.64, while the lowest scored 0.53, 0.43, and 0.35. These results confirm limited reliability in manual dosimetric contour assessment.

Limitations: The scoring system used in this study gave equal weight to all OARs, without considering their clinical importance. This may reduce the relevance of findings where high-priority structures are involved. The study focused only on glioblastoma, with data from 14 patients and seven evaluators from a single country, limiting generalizability to other tumour types or clinical settings. The small sample of evaluators may not reflect the full range of clinical practice. Finally, the study did not link contour assessments with clinical outcomes such as toxicity or NTCP, which limits its ability to assess clinical impact.

A2: How can dosimetric criteria be systematically formulated to develop an automated systems to replicate their behaviour?

Summary: Chapter 5 explores this by using a DL-based dose prediction model, introduced in Chapter 4, to classify contour variations. The model was tested on 54 target volume variations from 14 brain tumour patients. **Dosimetric criteria were explicitly defined**: a variation was "Sub-optimal" if any OAR exceeded a 10% dose increase, "Improved" if the change was beneficial, and "No Impact" otherwise. These thresholds were tunable via parameters α (deviation threshold) and n_{OAR} (number of affected organs), ensuring flexibility and clinical relevance.

The model outperformed three radiation oncologists, achieving a precision and recall of 0.57, slightly higher than the best expert. Confusion matrices showed better sensitivity in identifying "Sub-optimal" cases. These results demonstrate that automated, dosimetry-based evaluation is not only feasible but can also enhance human performance. The approach provides a replicable and efficient framework for contour QA, aligned with clinical priorities.

Limitations: The use of a fixed 10% dose threshold simplifies classification but may miss subtler or context-specific dose effects. Experts were not given formal definitions or dose distributions, which may have limited their performance and affected comparisons. The test set was small (54 variations from 14 GBM cases), limiting generalizability. The sensitivity study in Chapter 4 was also limited to simulated variations of a single OAR and did not include real inter-observer segmentations. Broader validation across tumour types, anatomical sites, and clinical scenarios is needed to confirm robustness and clinical utility.

12.2 Technical Investigations and Analysis

Findings from Part II demonstrate the potential of DL-based dose prediction models to improve treatment planning for glioblastoma using VMAT. Chapters 6, 5, 10, and 11 show that these models enable real-time dosimetric contour evaluation due to their fast inference times. Chapters 7 and 9 further investigate the reliability of DL architectures, particularly for auto-contouring tasks.

How reliable and fast are AI-based systems for dosimetric contour QA?

B1: Do the predicted doses correlate better with the true dose differences compared to geometric metrics?

Summary: Chapters 4 and 6 provide strong quantitative evidence that DL-based dose predictors can accurately estimate dose changes from clinically relevant contour variations. Chapter 6 evaluated a cascaded 3D U-Net [177] trained on 60 cases, and tested on 10 OOD scenarios. It achieved a dose error of 0.98 Gy and a mean DVH error of 1.89 Gy. After fine-tuning on OOD data, dose error slightly improved to 0.97 Gy, with a small rise in DVH error to 2.03 Gy.

In Chapter 4, the same model was tested for sensitivity to expert variation in left optic nerve contours. It achieved a dose MAE of 2.039 Gy, closely matching the ground truth MAE of 2.115 Gy. **A high correlation coefficient of 0.926 confirmed the model's accuracy in capturing dose changes**. These results suggest such models could reduce RT planning times, which currently average 9.63 days [422].

Limitations: Both studies are limited to glioblastoma cases using a single VMAT protocol and data from one institution, restricting generalizability. Model performance in other tumour types, modalities, or multi-centre settings remains untested. As a result, re-training may be necessary for each new clinical context, and the potential for a universal "foundation model" remains unproven [124].

No correlation was made between predicted dose changes and clinical outcomes such as toxicity or treatment efficacy. This limits the clinical relevance of the findings and reflects the lack of standardized, publicly available datasets and benchmarks. Additionally, only a cascaded 3D U-Net was evaluated. Alternative architectures were not compared, limiting insight into possible improvements in accuracy, interpretability, or efficiency, which is especially important for real-time clinical deployment.

B2: How reliable are the core DL architectures under difficult conditions?

Summary: Chapters 7, 8, and 9 examine the role of skip connections in U-Net-based architectures. While skip connections help preserve spatial detail for complex segmentation tasks, the findings show that advanced variants like AGU-Net and UNETR are **not consistently more robust** than standard U-Net models, particularly under distribution shifts.

Enhanced models perform well in clean conditions (e.g., AGU-Net: 0.795 on Breast US) but degrade more under perturbations (e.g., AGU-Net drops to 0.645 vs. NoSkipU-Net at 0.735). No-Skip models like NoSkipV-Net show better stability in harder settings, winning 78.3% of the most difficult Heart MRI cases. Lower coefficient of variation scores (e.g., 0.20 for NoSkipV-Net vs. 0.35 for AGU-Net) further support this. Chapter 9 shows that UNETR and AGU-Net are highly sensitive to foreground-background ratio changes, with DSC drops of up to 0.369, compared to just 0.0139 for U-Net.

These results suggest that skip connections can amplify noise under domain shifts. However, larger patch sizes improve performance across architectures, indicating that global context remains valuable for complex segmentation tasks.

Limitations: The studies rely on synthetic or curated datasets with controlled perturbations, which do not fully reflect real-world clinical variability such as scanner differences, anatomical diversity, or multi-modal noise. The analysis is limited to U-Net variants, without testing broader architectures like transformer hybrids or alternative attention mechanisms. In addition, segmentation quality is measured using standard metrics (e.g., DSC, HD), without linking performance to clinical outcomes or decision-making. This limits understanding of how robustness improvements impact actual clinical utility.

12.3 Proofs of Concept Experiments

Part III, through Chapters 10 and 11, presents initial proof-of-concept demonstrations showing how the advances from earlier chapters could be integrated into clinical applications.

What clinical systems can such dosimetric contour QA be integrated into?

C1: Can such models assist in focusing contour review efforts on locations where segmentation variations are dosimetrically most critical?

Summary: Chapter 10 introduces the novel ASTRA method, which enhances QA by generating sensitivity maps that highlight contour regions with the greatest dosimetric impact [505]. By simulating over 2,000 local surface perturbations per subject, ASTRA quantifies the relationship between contour errors and dose differences (e.g., correlation of –0.76 for the left eye). Although based on synthetic perturbations rather than inter-expert variability, it represents the **first approach to produce local, dosimetry-informed heat maps** for guiding expert review.

Limitations: The study is limited to glioblastoma cases treated with VMAT using a single planning system, which restricts generalizability. Separate training is needed for different clinical setups, posing scalability challenges. The method does not assess whether sensitivity maps influence treatment decisions or clinical outcomes, such as toxicity or efficacy. ASTRA uses simplified, uniform perturbations that may not reflect real-world, anatomically constrained errors. Further validation is needed to compare estimated and true dose differences and to justify its clinical reliability.

C2: Can such models assist in dosimetrically ranking various auto-contour proposals?

Summary: Chapter 11 introduces AutoDoseRank, a method for ranking segmentations by dosimetric quality using a DL-based dose predictor and a novel dose impact metric. This metric incorporates OAR-specific constraints and priority weights (e.g., p("optic nerve") = 1, p("eye") = 8), and aggregates per-OAR scores into a patient-level ranking. Tested on 65 segmentations from 13 patients, AutoDoseRank outperformed three of four radiation oncologists, achieving a mean Kendall's Tau of 0.129, close to the best expert (0.148) and well above the worst (0.014). This approach provides **a practical, dosimetry-driven method for ranking contours**.

Limitations: The study is limited to a small dataset of 13 GBM patients, which affects the statistical strength of the results. The method depends on a DL-based dose predictor, and any errors in dose estimation could bias the rankings. The impact of prediction uncertainty on ranking reliability is not quantified. Additionally, the clinical utility of AutoDoseRank remains untested. Its effects on decision-making, correction time, or treatment quality are not evaluated, leaving its real-world benefits uncertain.

* * *

The research presented across Parts I, II, and III provides strong evidence supporting the hypothesis that AI-based systems can reliably and efficiently standardize and automate dosimetric contour QA in RT planning. Building on these findings, Chapter 13 outlines future directions to advance this field further.

13

Perspectives

"We can only see a short distance ahead, but we can see plenty there that needs to be done."

— Alan Turing, Computing Machinery and Intelligence, 1950.

This chapter is structured in two parts. Section 13.1 outlines specific future research directions, building on current findings from both clinical (13.1.1) and technical (13.1.2) perspectives. Section 13.2 places these directions within the broader context of how radiation oncology may evolve with advances in AI, considering the viewpoints of both clinicians (13.2.1) and patients (13.2.2).

13.1 Concrete Near-Term Directions

The directions outlined in this section aim to support the development of workflows that are more reliable, time-efficient, and standardized. These improvements are expected to benefit both clinical outcomes and computational efficiency. Section 13.1.1 discusses emerging clinical needs identified through this research, followed by proposed technical advancements in Section 13.1.2.

13.1.1 Addressing Clinical Needs

The following three ideas are based on clinical needs identified through discussions with clinicians during the experimental and evaluation phases. They illustrate how the methods developed in this study can support improvements in real-world clinical workflows.

Novel Dose-Aware Metrics: Developing new quantitative metrics is essential to compare DVH curves in a way that reflects clinically meaningful differences in dose distributions. Existing methods, such as calculating the area between curves, often ignore the histogram structure of DVHs and the non-linear clinical relevance of dose changes. For instance, a reduction from 61 Gy to 60 Gy may be more significant than from 2 Gy to 1 Gy, despite the same absolute difference.

A promising direction is to compare differential DVH curves using statistical divergence measures. While KL divergence is one option, its asymmetry may limit its usefulness. The Wasserstein distance offers a symmetric and more intuitive alternative that aligns better with clinical reasoning. Incorporating weighted difference

schemes can further enhance relevance by assigning higher penalties to deviations exceeding clinical thresholds. This would prioritize critical dose regions and support the development of AI models that are both quantitatively rigorous and clinically aligned.

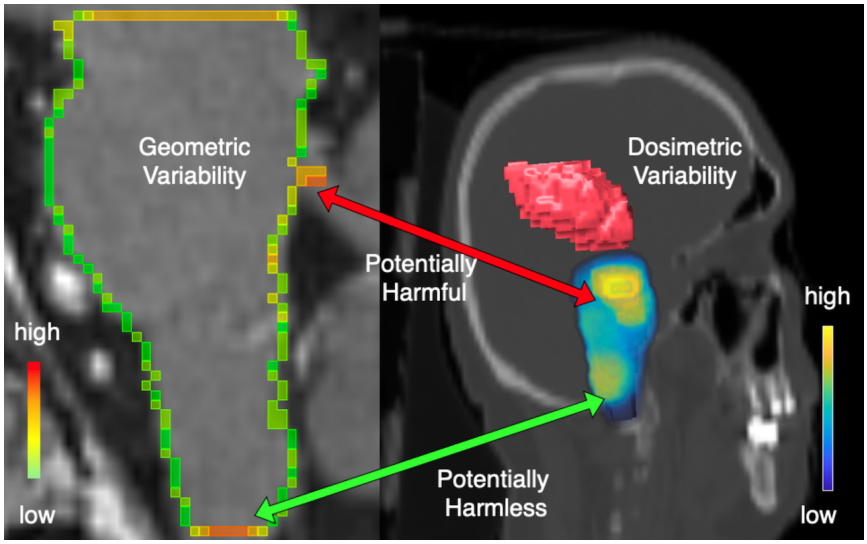


FIGURE 13.1: Geometric variability (left) feeds into dosimetric variability (right) for a holistic RT-QA system.

A second approach involves combining geometric and dosimetric variability to better understand how contour differences affect dose distributions. In the absence of large-scale inter-expert contour datasets, atlas-based heat maps can act as surrogates. A pilot study using contours from three radiation oncologists generated region-based heat maps, which were then combined with ASTRA-like dose sensitivity maps. Regions exceeding defined thresholds (e.g., geometric variation >2 mm and dose deviation >2 Gy) can be flagged for expert review. To ensure clinical flexibility, these thresholds should be user-configurable, avoiding unnecessary alerts when only one metric varies. An example is shown in Figure 13.1, where the green arrow highlights a region that does not require review. [1]

A third approach proposes generating a family of DVH curves to visualize how simulated contour variations affect dose metrics for specific OARs. For example, Figure 13.2 shows colour-coded DVH curves for the chiasm, where colour indicates DSC relative to the reference contour. The dose distribution is not recalculated but derived from the original plan. These visualizations highlight the range of potential outcomes if inaccurate contours are used and can support threshold-based assessment of clinical risk associated with contour uncertainty.

Personalized CTV Margin Optimization: In RT for GBM, microscopic tumour infiltration often extends beyond visible imaging boundaries, making CTV definition inherently uncertain. This infiltration is patient-specific, remains undetectable with current structural or functional imaging, and lacks reliable biomarkers. As a result, clinical practice typically applies a uniform margin (e.g., 1.5–2 cm) around the GTV, adjusted only by known anatomical barriers. This population-based approach does not reflect individual infiltration patterns and does not optimize the balance between tumour coverage and sparing of healthy tissue.

[1] Adapted from a one-page abstract presented at ISBI 2024.

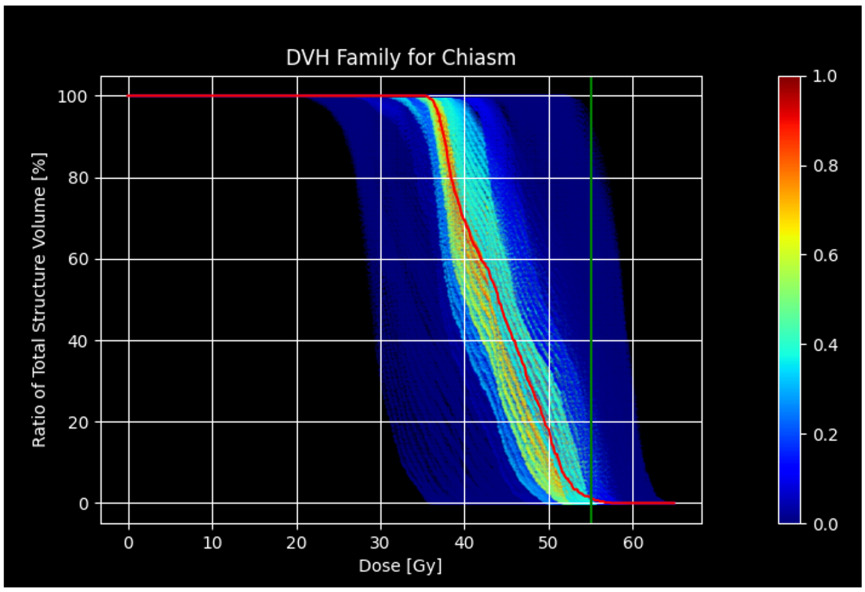


FIGURE 13.2: One visualization to rule them all: merging simulated geometric uncertainty into DVH curves to estimate overall robustness for risk assessment of contour quality. Green vertical line indicates dose constraint of 54 Gy maximum dose.

While expanding the CTV could theoretically improve local control, this is only beneficial if the added volume contains disease and does not violate dose constraints for nearby critical OARs such as the brainstem or optic nerves. The models developed in Chapters 4 and 6 provide a framework to support this decision-making. They enable the generation of personalized, anatomically and dosimetrically feasible CTV expansion envelopes.

This involves anisotropic expansion of the CTV in all directions, followed by dose prediction to assess whether the expansions remain within safety thresholds. A key advantage of the models is their near-real-time inference, allowing integration into clinical workflows. Tumour growth models could also be incorporated to simulate more aggressive strategies. The output includes both a safe maximum expansion volume and a map of high-risk regions where further expansion would exceed OAR constraints. A probability map of tumour presence aligned with dose feasibility may further guide clinical decisions.

These tools could support individualized margin definition in anatomically complex areas (e.g., periventricular or eloquent regions) and help identify patients who may benefit from more aggressive margins. They may also inform clinical trial design, such as evaluating dose-painting strategies or comparing personalized versus standard margins.

The clinical value of this approach is broad. Most recurrences occur just beyond standard margins; personalized expansions could help reduce this risk. Automated tools can also prevent excessive toxicity by ensuring larger CTVs remain within safe limits, improving cognitive, visual, or motor function preservation. Additionally, by reducing inter-clinician variability, these tools promote more consistent, data-driven treatment planning and can support standardized decision support systems. This opens new opportunities for clinical trials aimed at evaluating whether individualized, constraint-based expansions improve outcomes such as overall survival, thereby bridging the gap between conservative, guideline-driven planning and personalized, risk-adaptive care.

Dosimetry-Aware Learning Platform for Radiation Oncology Residents: Tools like AutoDoseRank can enhance both the quality and consistency of RT treatment plans. In addition to clinical use, they also have potential as interactive training platforms that help radiation oncologists develop a better understanding of how contour variability affects dose distributions. As automation becomes more integrated into treatment planning, the role of radiation oncologists is changing. Instead of manually contouring structures, they are increasingly responsible for reviewing and supervising AI-generated contours. Although automation reduces manual workload, it shifts the focus toward interpretation, prioritization, and quality assurance.

Current training practices rarely offer physicians quantitative, dose-aware feedback on how their contouring choices or modifications influence OAR doses or tumour coverage. This limits their ability to understand important trade-offs, particularly in anatomically constrained cases such as GBM. In these scenarios, even small geometric changes can lead to significant dosimetric consequences, which are often not captured by conventional geometric metrics.

To address this gap, AutoDoseRank could be developed into an interactive training tool. Trainees could be presented with multiple contour options, including initial AI-generated contours and manually corrected versions, and asked to rank them by expected dosimetric quality. The tool would then provide immediate feedback using predicted dose distributions and corresponding dosimetric rankings. It could also pose specific questions, such as "Which contour led to a higher brainstem dose?" or "Did expanding the target improve coverage without exceeding the optic chiasm limit?" Trainees could directly adjust contours, such as expanding or shrinking the CTV, and observe how these changes affect OAR doses in real time.

This creates a low-risk environment where users can learn to navigate clinical trade-offs and understand real-world planning constraints, such as mean dose limits to the hippocampi. As clinical roles evolve, training must also include decision-making skills and awareness of dose implications, not just contouring accuracy. AutoDoseRank could further be used to evaluate trainee performance by comparing their contour rankings to the tool's dosimetric output or expert-reviewed standards. This opens the possibility of using such platforms to assess readiness for advanced responsibilities, including independent QA review.

13.1.2 Technical Advances

This section outlines potential technical advancements aimed at addressing the clinical needs discussed earlier. The overarching goal is to support standardized treatment efficacy, reduce the risk of sub-optimal contouring due to limited model robustness, and shorten planning times through increased automation of the workflow.

Adaptive Robust Segmentation Models: Building on the results from Chapter 8 and the broader work in Part III, a texture-driven adaptive model architecture may offer a better balance between robustness and performance in high-stakes tasks such as tumour auto-contouring. One potential approach is a two-step iterative inference pipeline. In the first stage, a standard U-Net performs binary segmentation to estimate FG and BG regions. The TS between these regions is then computed and used to inform a second inference stage, where a model architecture optimized for the detected texture characteristics is selected. The hypothesis is that such an adaptive

pipeline would outperform static, single-pass models by offering greater robustness across variable imaging conditions.

An alternative direction involves exploring hybrid architectures that combine U-Net and V-Net-style skip connections with features such as attention gates (as in AGU-Net), dense skip pathways (as in UNet++), or transformer blocks. Investigating how these elements interact could reveal important trade-offs along the robustness–performance Pareto frontier.

To evaluate such trade-offs, a standard measure of robustness is needed. One possible metric is the degradation in mean Intersection over Union (mean Intersection over Union (mIoU)) between clean and corrupted inputs, averaged across a suite of perturbations with varying noise types and intensities. Mapping this robustness landscape, as illustrated in Figure 13.3, could provide valuable insights. To our knowledge, such a comprehensive analysis has not yet been conducted in the context of medical image analysis.

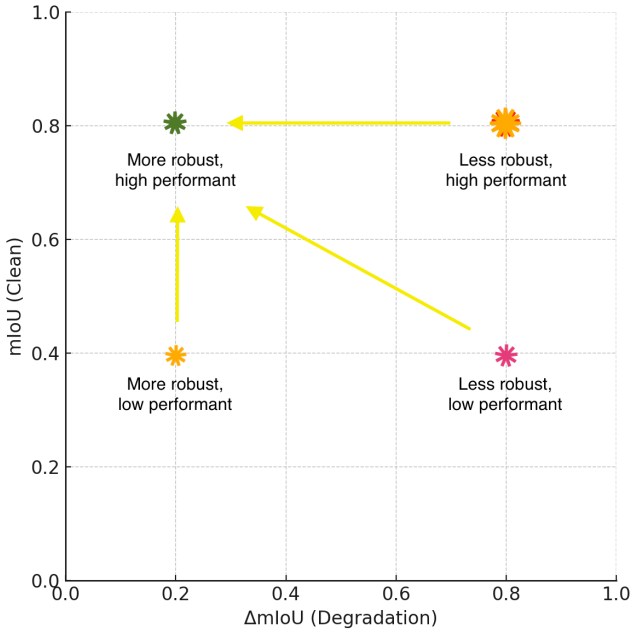


FIGURE 13.3: On the performance versus robustness scale, there is a need to move model behaviour towards the right top corner, where performance and robustness do not need to trade-off with each other. It is hypothesized that hybrid architecture choices can help break this barrier.

These architectural enhancements could contribute toward developing powerful foundation models for RT planning, as illustrated in Figure 13.4. Such models would account for the complexity of various planning techniques (e.g., IMRT, VMAT) and generate realistic dose distributions and treatment plans across diverse anatomical sites and planning parameters. They would also be adaptable to different vendors and imaging platforms, functioning as a digital twin for RT workflows.

Incorporating LLMs could further enhance these models by integrating clinical notes, patient history, and planning preferences. This would enable a natural language interface for oncologists and physicists, allowing interaction with these complex systems at a higher level of abstraction.

Dynamic Dose Prediction Engine: Predicting deliverable dose distributions using real-world planning parameters can help planners reach near-optimal solutions

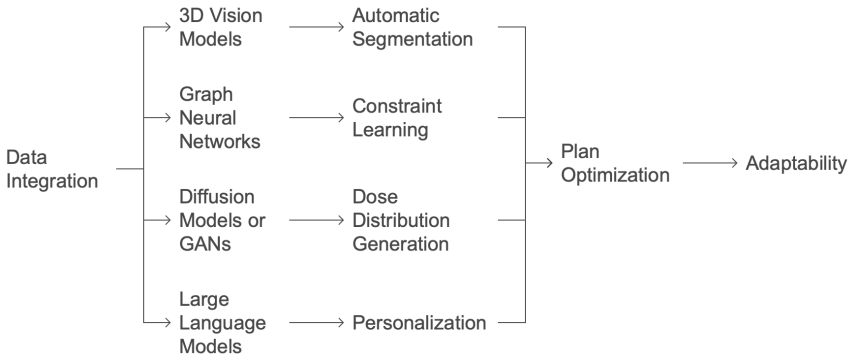


FIGURE 13.4: Components of a foundation model for radiation oncology - multimodal data with imaging, personalized constraints, clinical notes and machine settings to automatically generate contours and treatment plans.

more efficiently and safely, especially for complex cases such as GBM. In the current clinical workflow, radiation oncologists contour the TV and OARs. After this step, physicists or dosimetrists manually adjust planning parameters, including beam angles, arc configurations, and optimization constraints. This process often involves iterative adjustments to DVH trade-offs and depends heavily on the planner's experience. It is time-consuming and subjective, particularly when the geometry of the TV and proximity to OARs increase the complexity of planning.

The space of planning parameters is large and not easily navigable. As a result, many clinically acceptable plans may be overlooked because they are not easily found through manual trial and error. Based on the models developed in Chapters 4 and 6, it is possible to predict deliverable dose distributions and DVH curves before full plan optimization. These models can also estimate the machine parameter configurations that could produce such plans, as demonstrated in prostate planning by Heilemann et al. [134]. This predictive capability can support parameter selection during plan setup, help exclude infeasible plans, and suggest alternatives that improve conformity, homogeneity, or OAR sparing.

To enable this approach, it is necessary to curate datasets that include variation along two dimensions: the contours themselves and the treatment plan parameters. This concept is illustrated in Figure 13.5.

Building on prior studies exploring the latent space of machine parameters [121], a DL model could be trained to learn a latent representation where each point corresponds to a viable set of planning parameters. This approach could replace the need for manually adjusting complex and interdependent parameters in high-dimensional space. Such models could also support clustering of similar plan types and mapping them to expected dose outcomes.

However, training such a model would require access to thousands of prior RT plans that cover a wide range of parameter sets and dose distributions, including both optimal and sub-optimal cases. At present, such comprehensive datasets are not publicly available [124]. Once trained, the model could generate a small set of candidate plans for human review and refinement.

A DL-based latent space for planning parameters would simplify the large and complex planning space into a manageable set of high-quality options. It could also automate the exploration of trade-offs between competing clinical goals, helping

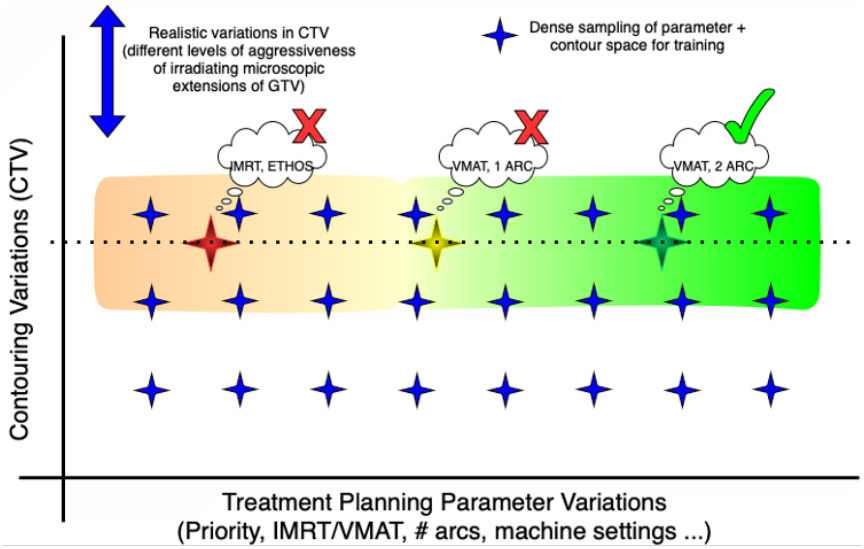


FIGURE 13.5: Curating a data set with contouring variations to learn the range of geometric differences and treatment plan parameter variations to learn the dose distribution ranges.

bridge the technical expertise of physicists with the treatment objectives of oncologists.

Towards Fully-Automated Treatment Planning: Recent studies have demonstrated promising results in progressing toward the final step of the RT workflow: generating complete treatment plans directly from contours [134]. These approaches derive machine parameters such as MLC positions and monitor units, enabling the integration of contouring, dose prediction, and treatment planning into a complete end-to-end pipeline for RT. Although earlier studies have focused on prostate anatomy [404], the same methodology can be extended to brain and head and neck cases. Preliminary systems using novel dose prediction techniques combined with dose mimicking have shown the ability to generate deliverable treatment plans in 12 to 13 minutes without user interaction [451].

Achieving higher-fidelity treatment planning will require large-scale datasets that cover a broad range of clinical conditions [124]. Future research can explore the use of advanced model architectures such as physics-informed neural networks [506] or new paradigms like the Segment Anything Model [507]. Training such systems will require access to plans that are not always clinically optimal, so the model can learn to differentiate between feasible and non-feasible plans. This becomes especially important for anatomically complex cases such as those involving the brain.

An additional area for improvement involves the use of clinically oriented loss functions. Geometric metrics do not reliably correlate with dose distributions, and dose itself has a complex and often indirect relationship with clinical outcomes such as toxicity or tumour control. Moving closer to outcome-focused loss formulations could help train models to prioritize accuracy in regions that have a greater clinical impact, particularly around the TV, and reduce focus on less critical areas. While dose-based loss functions have already been explored [463, 508], further work is needed. Ideally, this would involve combining dose data with models of toxicity or NTCP, helping better align training with therapeutic goals.

Both directions outlined here depend heavily on the availability of large, curated datasets. These would need to include multi-modal imaging (such as MRI

for contouring and CT for tissue density), expert-labelled contours for all relevant structures, associated dose distributions, and full treatment plan parameters such as MLC positions and monitor units. This would need to cover a range of planning types, including IMRT and VMAT, along with clinical data such as recurrence rates, toxicity outcomes, and survival estimates. Efforts to standardize the structure, format, quality, and variety of such data would be essential for building robust end-to-end treatment planning systems, moving beyond isolated optimization of individual components.

13.2 Future of Radiotherapy + Artificial Intelligence

The future of AI in RT is moving toward increasingly advanced systems that have the potential to transform how radiation therapy is conceived, planned, and delivered. The evolution of these systems has been described as progressing through three distinct generations: personalized treatment based on pretreatment data, response-driven RT, and eventually, dynamically optimized RT that can adapt in real time to patient-specific changes [509]. This section explores how the advances presented in this thesis align with a long-term vision for the future of RT shaped by AI. The discussion is structured into two parts: the impact on clinicians and providers (Section 13.2.1), followed by the impact on patient experience (Section 13.2.2).

13.2.1 Impact on Clinician Workflows

By 2040, it is anticipated that AI-assisted contouring and treatment planning will become routine, significantly improving the quality, consistency, and accessibility of RT plans. These advances are expected to enhance access to care and enable more frequent adaptive replanning [510]. A major area for future impact is the clinical adoption of auto-contouring technology. While 45% of surveyed radiation oncologists report using AI-based auto-contouring tools in practice, this use remains largely limited to OAR delineation [325]. As of 2025, tumour auto-contouring has not yet been adopted as a clinical standard [303]. However, with the development of robust QA mechanisms, broader clinical integration is likely. Improved QA tools are essential for ensuring the safe and effective use of AI-based systems, especially as the field advances toward second- and third-generation AI models focused on response-driven and dynamically optimized treatment planning.

According to the EORTC, two areas where AI is expected to have the most significant impact are tumour and OAR auto-contouring and associated QA workflows [511]. These systems could transform clinical practice by reducing contouring time while increasing standardization, accuracy, and consistency across institutions and practitioners [512]. As concerns around the reliability of AI-generated contours persist, future systems are expected to include built-in validation frameworks and quality metrics. These enhancements would support more consistent contouring, especially in the context of multi-institutional clinical trials [513].

For the multidisciplinary RT team, the growing role of AI introduces new responsibilities and skill requirements. Clinicians, physicists, and dosimetrists will need to understand the capabilities and limitations of AI methods to make informed decisions about their use. As AI tools evolve beyond automating individual tasks to supporting broader workflow integration, this understanding becomes critical. While these tools can improve consistency, productivity, and overall plan quality, they also raise concerns about skill degradation and insufficient training to maintain technical expertise [514]. These challenges highlight the importance of thoughtful

implementation strategies that maintain essential clinical skills while leveraging the strengths of automation. Workflow efficiency gains must not come at the cost of safety, creativity, or critical human oversight.

The future role of the RT professional may shift toward integrating and supervising AI systems rather than performing purely technical operations. Training tools that help radiation oncologists evaluate the dosimetric impact of different contouring decisions, and potentially treatment outcomes, could support the development of this evolving skill set. Although machine learning models may perform well in repetitive tasks such as delineating reproducible structures, clinical judgment remains indispensable. The collaboration between AI systems and clinicians must be designed to ensure that technology supports, rather than replaces, expert decision-making in complex clinical contexts [302].

The 2024 ESTRO congress report reflects continued scepticism about fully automating the RT care pathway, citing concerns over AI bias, accountability, and the essential role of human empathy in patient care [515]. These concerns suggest that clinicians will increasingly shift from technical tasks to patient-facing roles, dedicating more time to complex or nuanced cases [238]. As highlighted by Eric Topol in *Deep Medicine*, the greatest promise of AI lies not only in reducing errors or improving efficiency, but in restoring the human connection between doctors and patients [516]. This vision aligns with broader trends in other high-skill professions, where the most difficult 20% of cases increasingly receive 80% of clinical focus. This shift allows for simultaneous improvements in care quality and more meaningful patient–clinician interactions.

13.2.2 Impact on Patient Experience

AI has the potential to significantly enhance the patient experience in RT. One major factor influencing patient comfort is the need for fixed and rigid positioning relative to the scanner gantry. Future AI tools are expected to support clinical decision-making and QA aligning with the growing demand for personalized, patient-centred care [120, 510]. Third-generation AI systems may enable real-time adaptation to anatomical changes, allowing treatment to proceed without immobilization devices or positioning markers, and potentially enabling patients to self-position for treatment.

Another important aspect of patient experience is the duration and frequency of treatment. RT workflows are evolving to include accelerated regimens, such as hypo-fractionation and single-fraction treatments, which reduce the number of visits required [187]. In advanced cases, such as stereotactic ablative RT (Stereotactic Ablative Body Radiotherapy (SABR)) delivered using MRI-LINAC systems, some centres have implemented same-day workflows. These integrate consultation, simulation, planning, and treatment into a single visit, with a median completion time of approximately 6.6 hours [187]. What is currently known as online adaptive planning is likely to become the standard approach, enabled by the computational speed of AI systems that eliminate reliance on static imaging from the first day of treatment.

Perhaps the most transformative impact of AI will be in low- and middle-income countries, where access to RT is severely limited. By 2040, 70% of global cancer cases are projected to occur in these regions, yet fewer than 50% of patients in such settings currently have access to RT [195]. Furthermore, 80% of the global cancer population is served by only 5% of RT resources, with some countries lacking a single radiation oncology centre for the entire population. Automating parts of the treatment planning and delivery workflow in resource-rich settings could help extend simplified

and efficient models of care to underserved regions. Similar to the emergence of remote robotic surgery, the potential for remote RT planning and delivery may offer a scalable and cost-effective solution to expand access to cancer treatment worldwide.

* * *

The convergence of findings across the three parts of this thesis (Parts I, II, and III) highlights the ongoing transformation within radiation oncology. Through the integration of dose-aware QA, predictive dosimetry, and robust model design, this work contributes to addressing key challenges in treatment planning, especially around time inefficiencies and human variability. These studies demonstrate the potential to reduce planning timelines, improve treatment accuracy, and support clinician training, towards the broader goal of higher precision medicine.

However, realizing this potential requires further research into scalable and generalizable model architectures, well-curated data pipelines, and thorough clinical validation. Advancements in these areas will support the development of more efficient, consistent, and patient-centred approaches to care. Future work must prioritize interdisciplinary collaboration and the use of open-access datasets and standardized protocols to accelerate progress. By addressing these challenges and building upon the results presented in this thesis, the field would move toward a future where RT is not only more effective but also more equitable and accessible for patients globally.

A

Appendices from Research Contributions

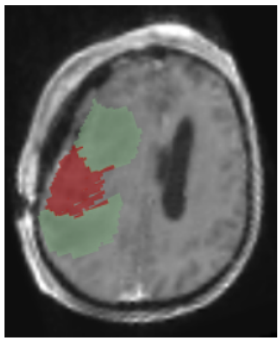
A.1 Chapter 3: Additional information about experiments.

TABLE A.1: Dose constraints used to construct ground truth classification. The eyes, with all their constituent parts, are considered a single OAR with a strict constraint of 10 Gray maximum dose.

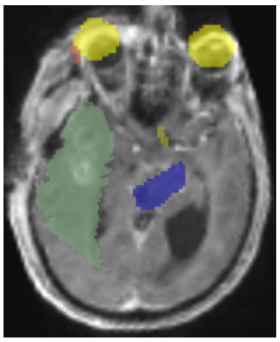
OAR	Constraint value (Gy)	Constraint Metric	Priority
Brainstem	54	Max	1
Chiasm	54	Max	1
Cochlea	45	Mean	7
Entire Eye	10	Max	8
Hippocampus	30	Mean	7
Lacrimal Gland	25	Mean	1
Optic Nerve	54	Max	1
Pituitary	45	Mean	10

TABLE A.2: Evaluator expertise and years of experience.

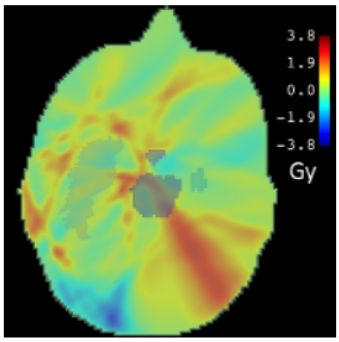
Evaluator ID	Years of Experience
R1 (Radiation Oncologist, Hospital 1)	07
R2 (Radiation Oncologist, Hospital 2)	15
R3 (Radiation Oncologist, Hospital 2)	08
R4 (Radiation Oncologist, Hospital 2)	02
M1 (Medical Physicist, Hospital 2)	11
M2 (Medical Physicist, Hospital 2)	10
M3 (Medical Physicist, Hospital 2)	19



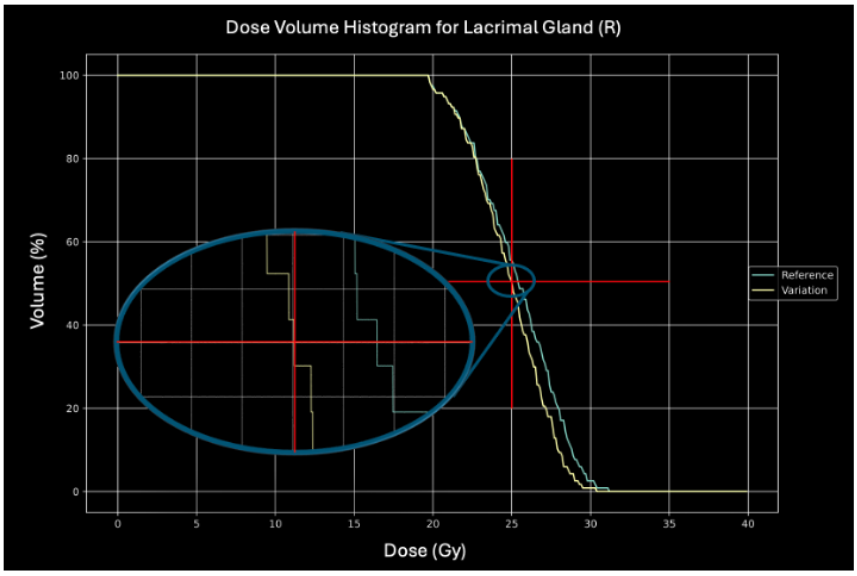
Axial slice with change: The target volume change (in red) connects two parts of the reference volume to make the shape more continuous/round. The reference target contour is shown in green.



Axial slice with OAR: The right lacrimal gland shown in orange has planar overlap with a part of the reference target volume (in green) but not with the changes. Brainstem is shown in blue, Chiasm and Eyes in yellow.

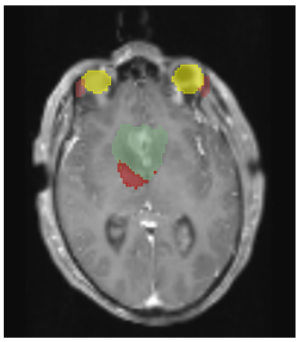


Dose difference: The dose difference (reference minus variation) heatmap shows the direction along which the dose is lower (in red) with the variation as compared to without, which passes through the lacrimal gland.

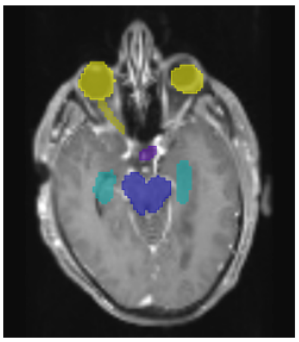


Dose Volume Histogram: showing the curves for the reference TV contour (marked reference) and TV contour variation (marked variation), showing that after the change, the right lacrimal gland constraint of 25 Gy mean dose is not violated, while in the reference contour case it is. Dose and volume constraints are marked in red, and the graph is zoomed in for clarity.

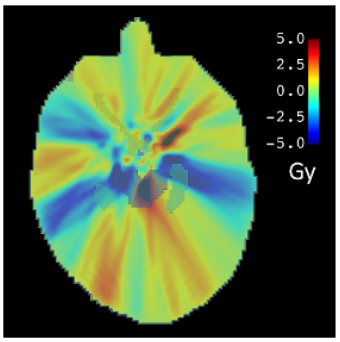
FIGURE A.1: Exemplar situation of when a contour variation led to a "Better" classification based on the right lacrimal gland (Subject 14, variation 3).



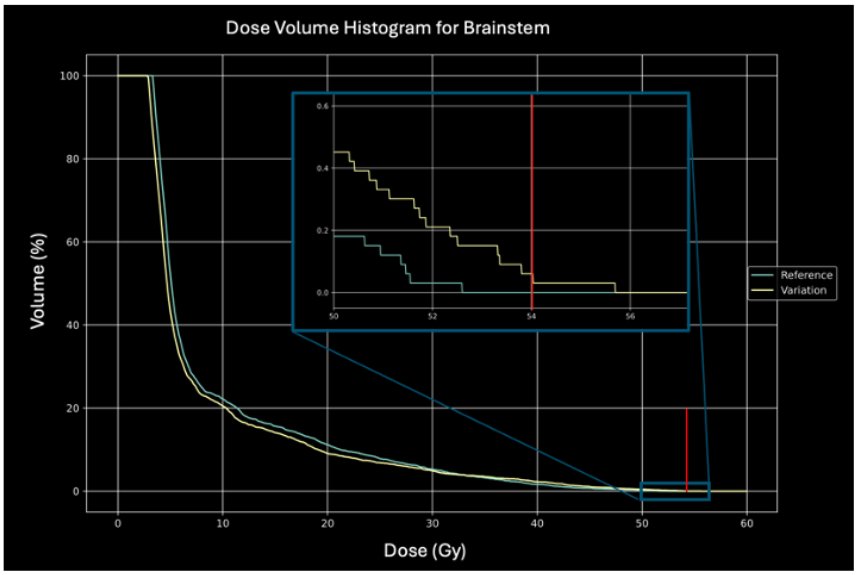
Axial slice with change: The target volume change (in red) extends the boundaries in the posterior direction. The reference target contour is shown in green.



Axial slice with OAR: The brainstem shown in blue has no planar overlap with the reference target volume or variations. Other OARs highlighted for anatomic reference.



Dose difference: The dose difference (reference minus variation) heatmap shows the location (in blue) in the brainstem where the maximum dose constraint is violated.



Dose Volume Histogram: showing the curves for the reference TV contour (marked reference) and TV contour variation (marked variation), showing that after the change, the brainstem constraint of 54 Gy maximum dose is violated, while in the reference contour case it is not. Dose and volume constraints are marked in red, and the graph is zoomed in for clarity.

FIGURE A.2: Exemplar situation of when a contour variation led to a "Worse" classification based on the brainstem dose (Subject 9, variation 1).

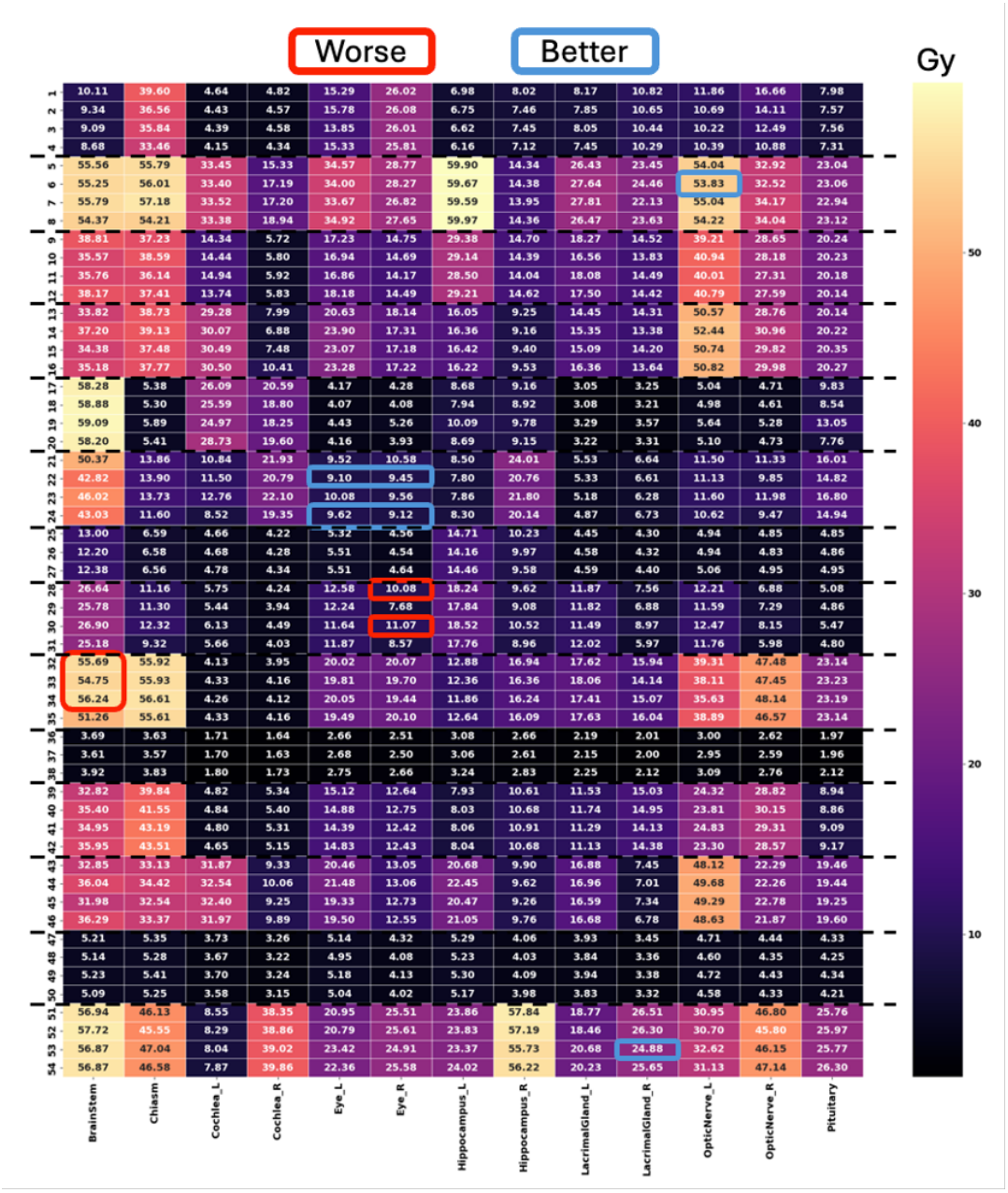


FIGURE A.3: Dose constraint values for each OAR for each of the 54 variations (Brainstem, Chiasm, Optic Nerve and Eyes are maximum dose, others are mean dose within structure). Horizontal dashed lines indicate groupings by patient, red (worse) and blue (better) borders highlight categories that are not "No change" based on crossing dose constraint thresholds.

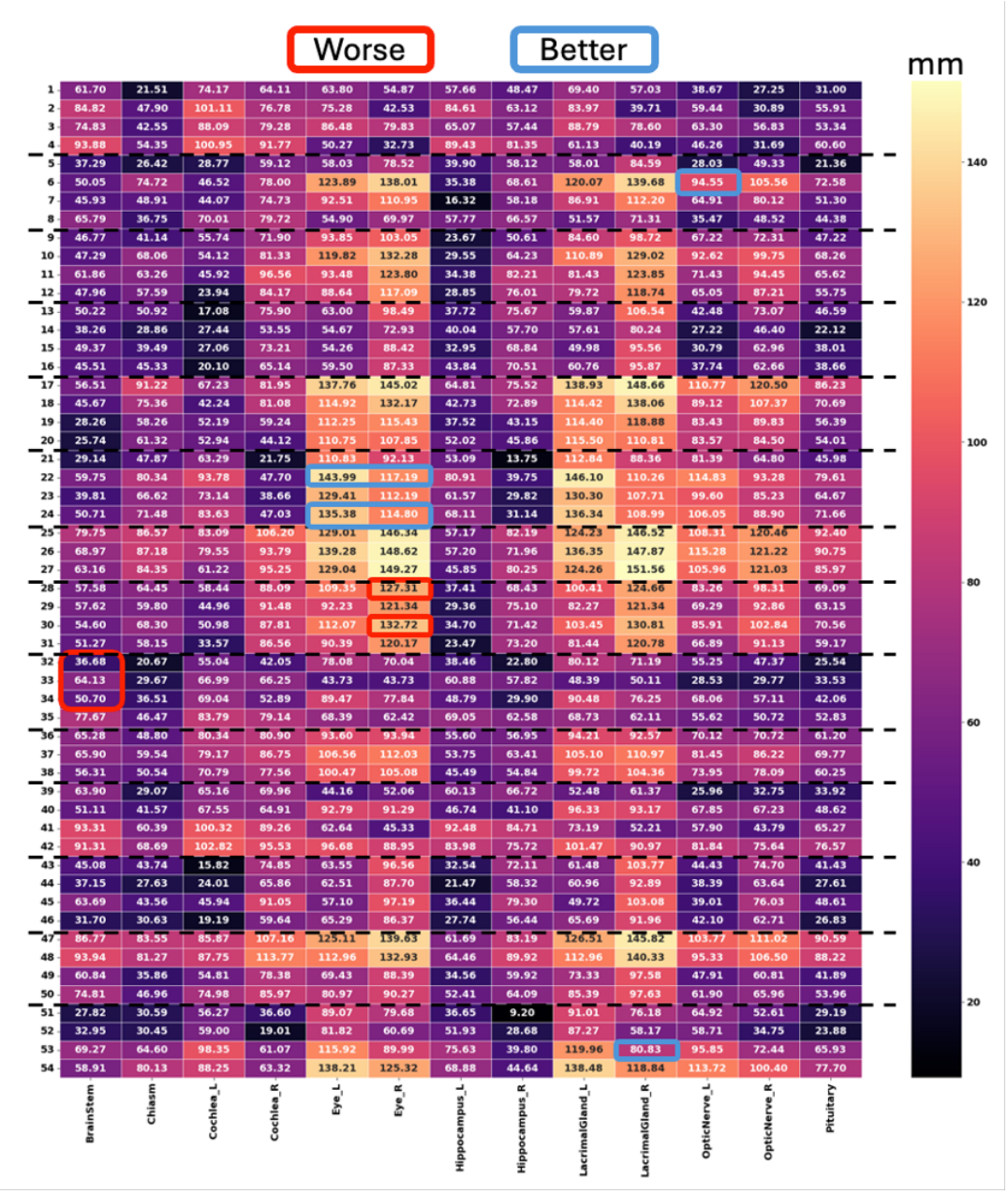


FIGURE A.4: Distances between the centroid of the TV modification made to the centroid of each OAR for each of the 54 variations. Horizontal dashed lines indicate groupings by patient, red (worse) and blue (better) borders highlight categories that are not "No change" based on crossing dose constraint thresholds.

TABLE A.3: Top three themes forming the basis of decision making amongst the evaluators.

Theme	Supporting Verbatim Quotes
Proximity to Organs at Risk (OARs)	M3: "So I think the most critical organs here close to the target are chiasm... so I would expect a negative impact to the optic nerves, but the distance is still so." R2: "It will impact, right... there won't be maybe an effect on the brainstem, but there will be for sure on the cochlea." R3: "It's on the level of the eyes... even small changes on the chiasm can have a relevant impact."
Size of Contour Change	R4: "Due to the size, 1 is almost irrelevant." R1: "It's really quite a large [change]... so I would also give it negative impact just because it's really quite large." M2: "It's a bit bigger than the one before, so close... I would go for low negative impact."
Shape and Geometry of the Contour Change	R2: "This part cannot be saved from irradiation... the beam shape, we could shape the beam... but not really like this." M2: "Addition looks a bit bigger to me than and also saying that... it's not super close to any of the organs at risk so I also go for [Worse]." M3: "It makes the shape a lot... yeah, that's..."

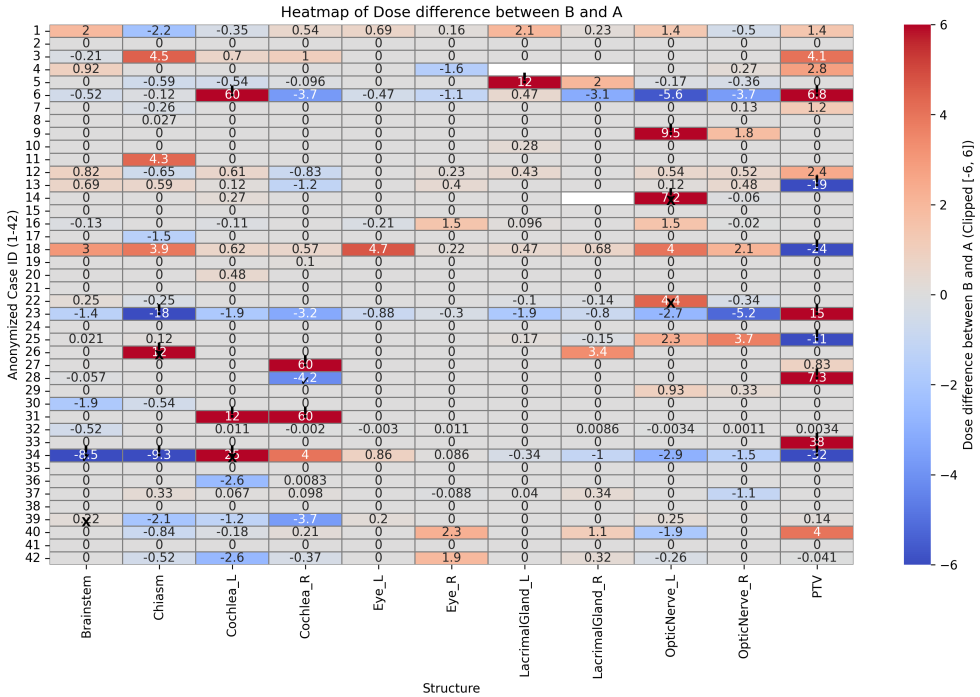


FIGURE A.5: What is the difference in doses when comparing situation (b) to (a)? This difference shows the residual dose each OAR and TV receives due to the error in contouring.

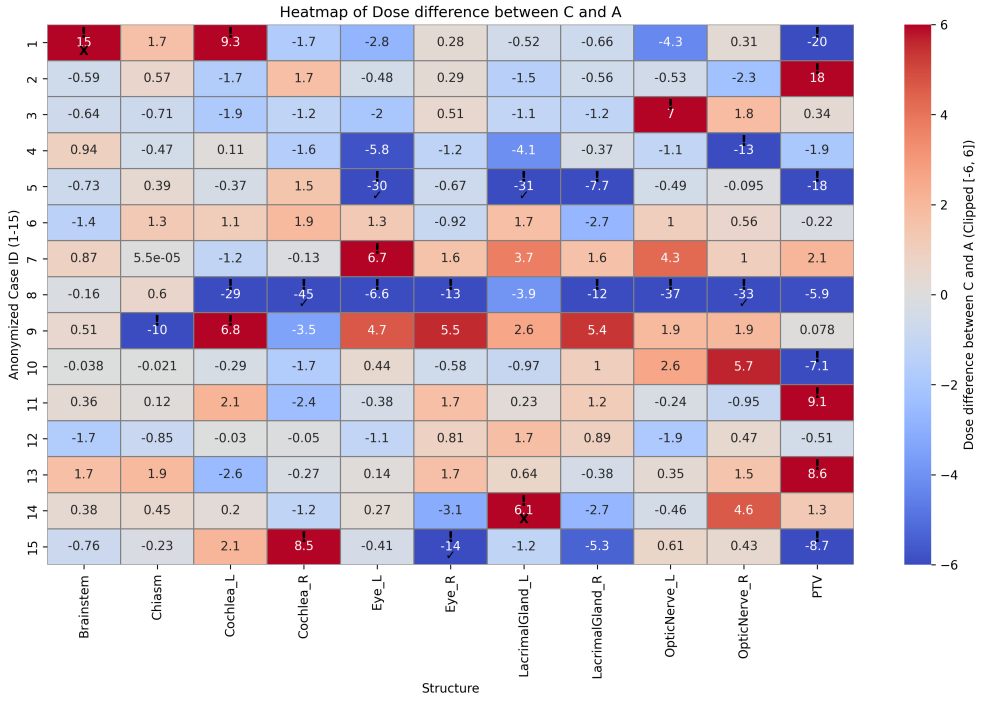


FIGURE A.6: What is the difference in doses when comparing situation (c) to (a)? This difference shows the residual dose each OAR and TV receives due to the cumulative errors in contouring and treatment planning.

A.2 Chapter 5: User Interface for evaluation experiments.

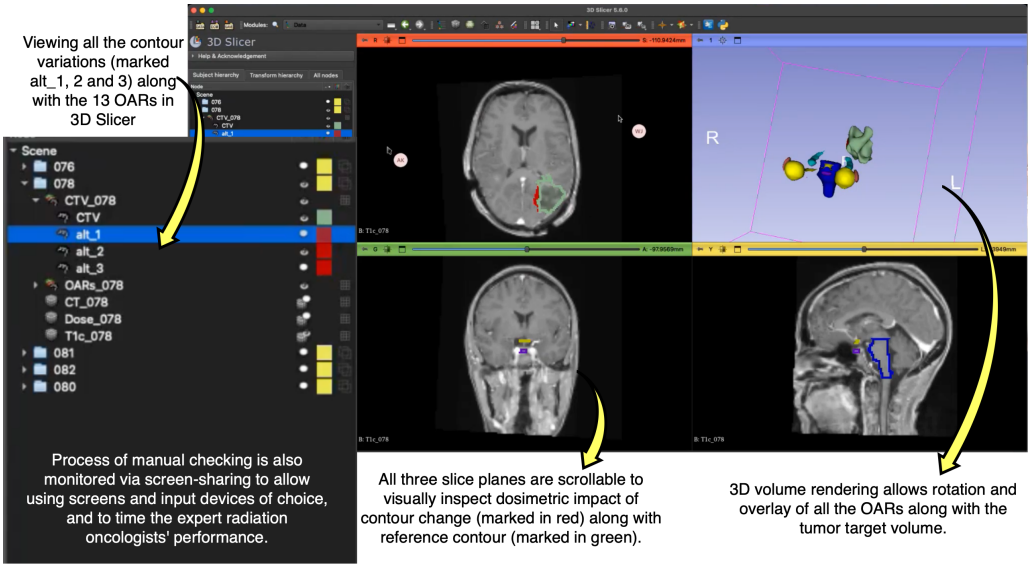


FIGURE A.7: User interface for radiation oncologists to review and classify contour changes (selecting variants via left panel) with three slice plane views and 3D volume rendering. 3D Slicer version 5.6.0.

A.3 Chapter 7: More evaluation metrics and results

Five variants of textures are generated using the test set in each of the four medical image data sets. Two of these levels are texturally "easier" than the original textures in the test set, also called the "unperturbed" case. These are artificially generated by selectively blurring the BG pixels using a symmetric Gaussian kernel with $\sigma = 3.0$ and $\sigma = 7.0$ in the "Easier" and "Easiest" cases. Two more are texturally "harder", which are also artificially generated by adding speckle noise to both the FG and BG pixels in the images, with variances of 0.1 and 0.3 for the "Harder" and "Hardest" cases, respectively. The TS variations caused by these changes are shown in Figure 7.3. Code to generate these variations is made publicly available [1].

A.3.1 Evaluation Metrics

In the following sections, results using the Surface DSC and surface distance metrics are demonstrated along the same lines as those using DSC and HDs described earlier.

Surface DSC

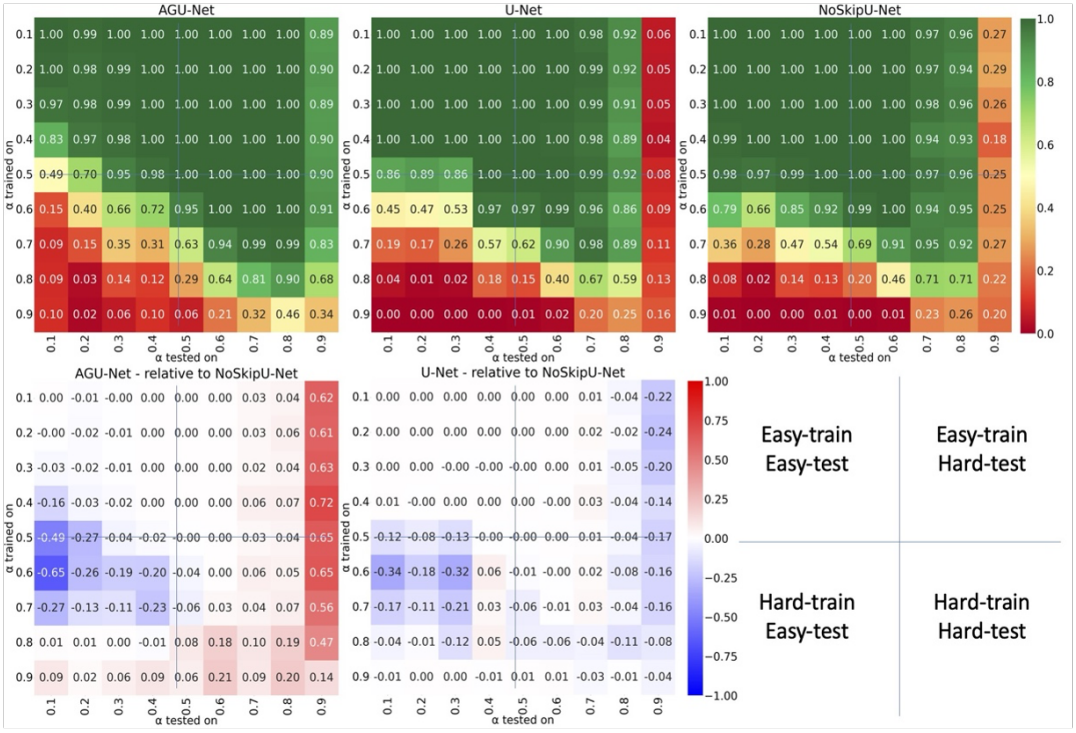


FIGURE A.8: Model robustness measured using Surface DSC for U-Net-like architectures when the BG blends into the FG texture. For the first row, higher/green is better. For the second row, blue indicates that NoSkipU-Net is better.

Surface DSC [401] can be between 0 (no match, worst) and 1 (matching entirely within the margin). Figure A.8 and A.9 show the Surface DSC variations for the BG and FG blending situations. These show more substantial differences than the standard DSC, especially in the hard-training texture scenarios. The general trend

[1] Code at GitHub: amithjkamath/to_skip_or_not

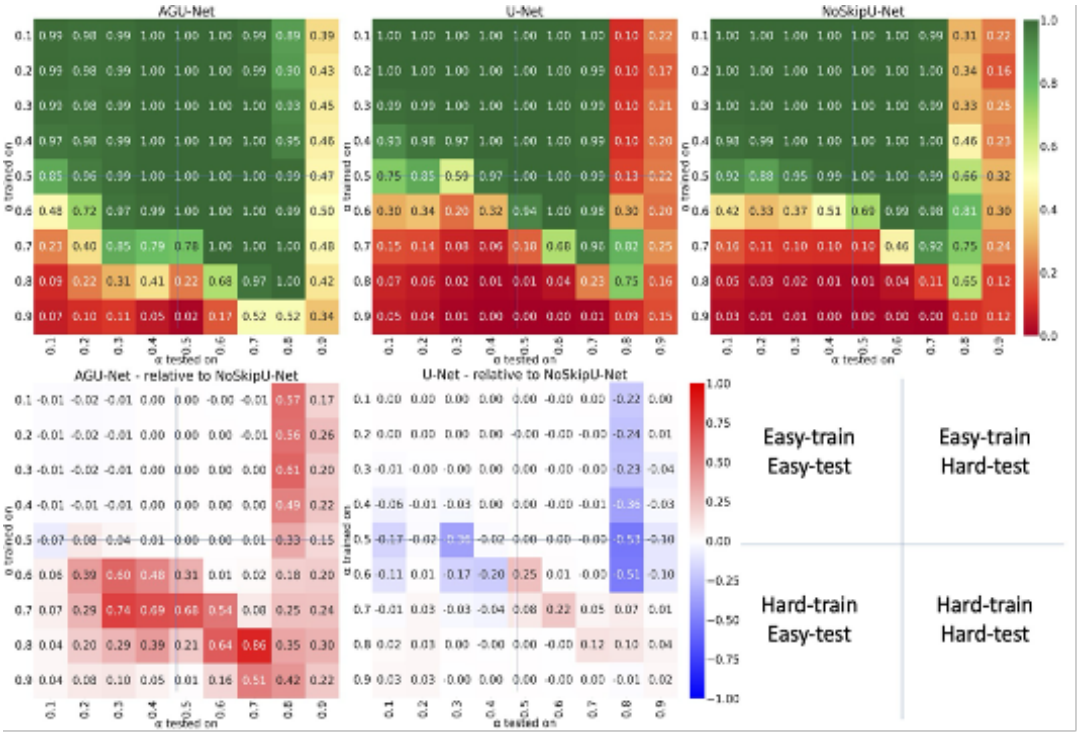


FIGURE A.9: Model robustness measured using Surface DSC for U-Net-like architectures when the FG blends into the BG texture. For the first row, higher/green is better. For the second row, blue indicates that NoSkipU-Net is better.

of low differences between the model architectures broadly and appreciably only in the hard-training texture scenario persists.

ASSD

The ASSD [488] can be between 0 (best) and unbounded at worst. Figure A.10 and A.11 demonstrate similar trends as DSC and HD and show that the differences between the three architectures are not significantly large. Figure A.10 indicates that all models trained with $\alpha > 0.5$ perform reasonably well across the range of textures in the test set. For the test set case of $\alpha = 0.9$, the AGU-Net performs better than the NoSkipU-Net, which is marginally better than the standard U-Net.

A.3.2 Results on Medical Image Datasets

This section includes additional results on the medical image data sets, specifically using the surfaceDSC and ASSD metrics.

SurfaceDSC

Figure A.12 demonstrates the performance of each of the six model architectures in the same format as Figure 7.8. Similarly, Table A.4 shows the best-performing model and the proportion of test cases where it wins against each of the other five architectures in the same format as Table 7.1. Note that the enhanced model architecture category performs better for the easier texture situations. In comparison, the NoSkip versions are generally better for the harder texture situations, following the same trend shown using the other evaluation metrics.

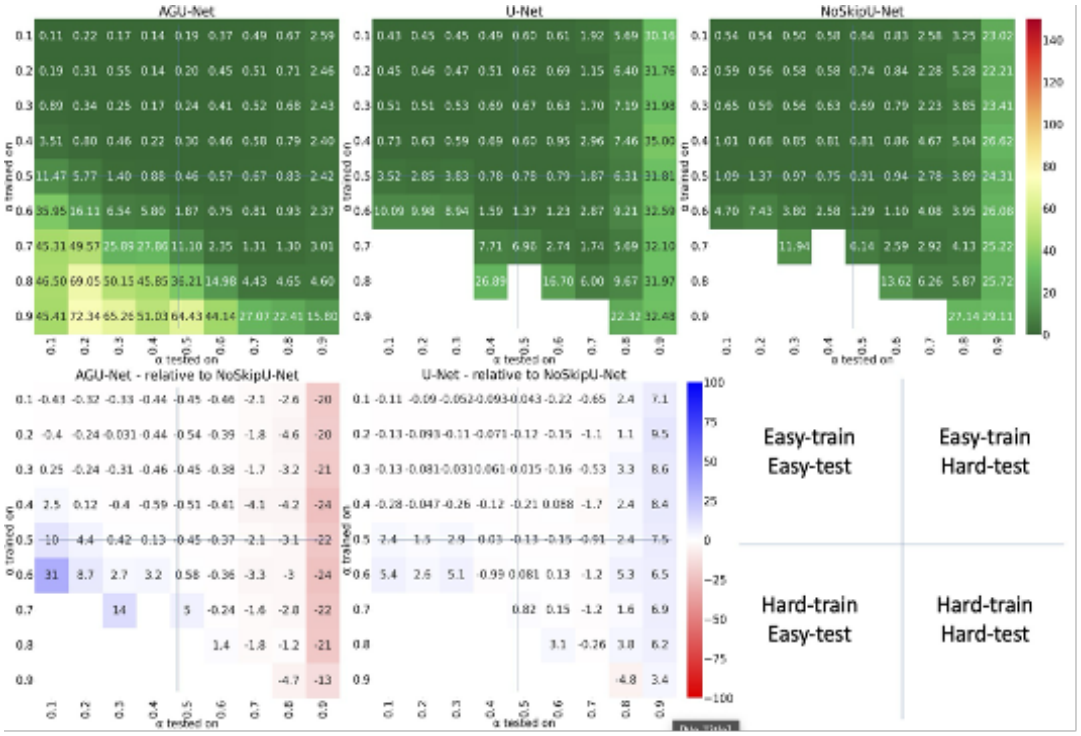


FIGURE A.10: Model robustness measured using ASSD for U-Net-like architectures when the BG blends into the FG texture. For the first row, higher/green is better. For the second row, blue indicates that NoSkipU-Net is better.

TABLE A.4: The best-performing model using surface DSC for all four data sets, in the same format as Table 7.1

	Breast US	Colon Histology	Heart MRI	Spleen CT
Easiest	AGU-Net (0.367)	V-Net (0.462)	NoSkipV-Net (0.316)	NoSkipU-Net (0.384)
Easier	AGU-Net (0.387)	UNet++ (0.362)	AGU-Net (0.441)	AGU-Net (0.444)
Unperturbed	UNet++ (0.272)	NoSkipV-Net, UNet++ (0.300)	UNet++ (0.533)	UNet++ (0.356)
Harder	NoSkipU-Net (0.265)	NoSkipV-Net (0.337)	V-Net (0.550)	NoSkipU-Net (0.759)
Hardest	NoSkipV-Net (0.394)	V-Net (0.412)	NoSkipV-Net (0.650)	NoSkipU-Net (0.736)

Surface Distance

Figure A.13 demonstrates the performance of each of the six model architectures in the same format as Figure 7.8. Similarly, Table A.5 shows the best-performing model and the proportion of test cases where it wins against each of the other five architectures in the same format as Table 7.1. The same trend as shown using surfaceDSC is seen in this case, where NoSkip versions perform better in the harder texture scenarios.

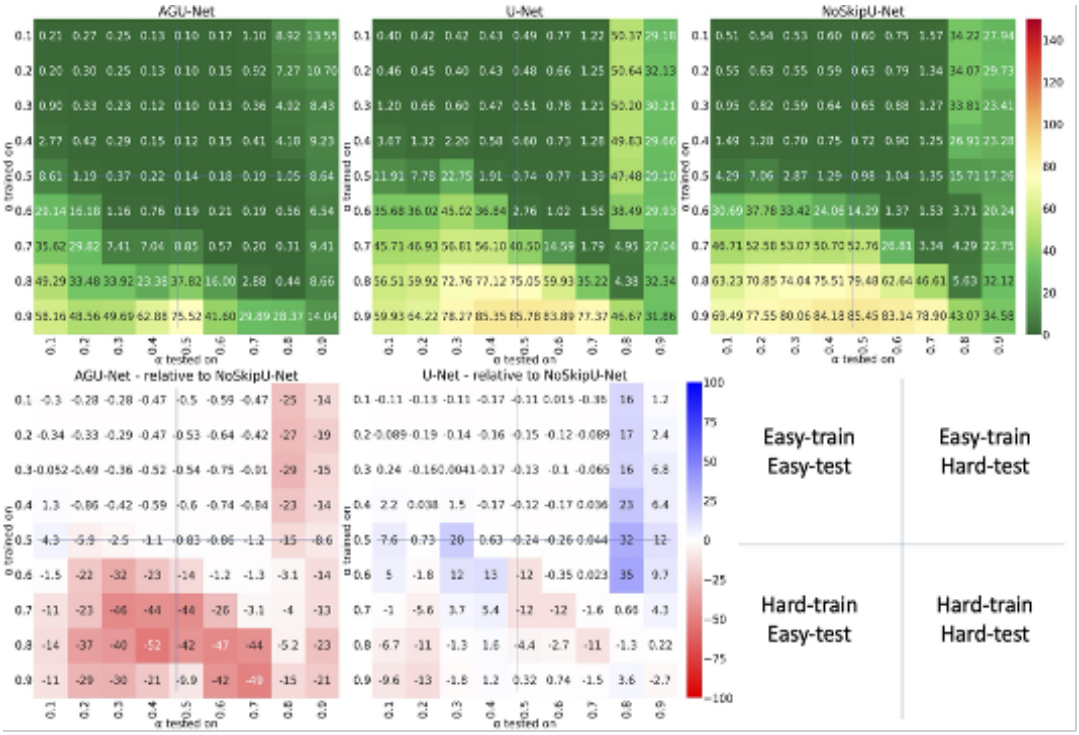


FIGURE A.11: Model robustness measured using ASSD for U-Net-like architectures when the FG blends into the BG texture. For the first row, higher/green is better. For the second row, blue indicates that NoSkipU-Net is better.

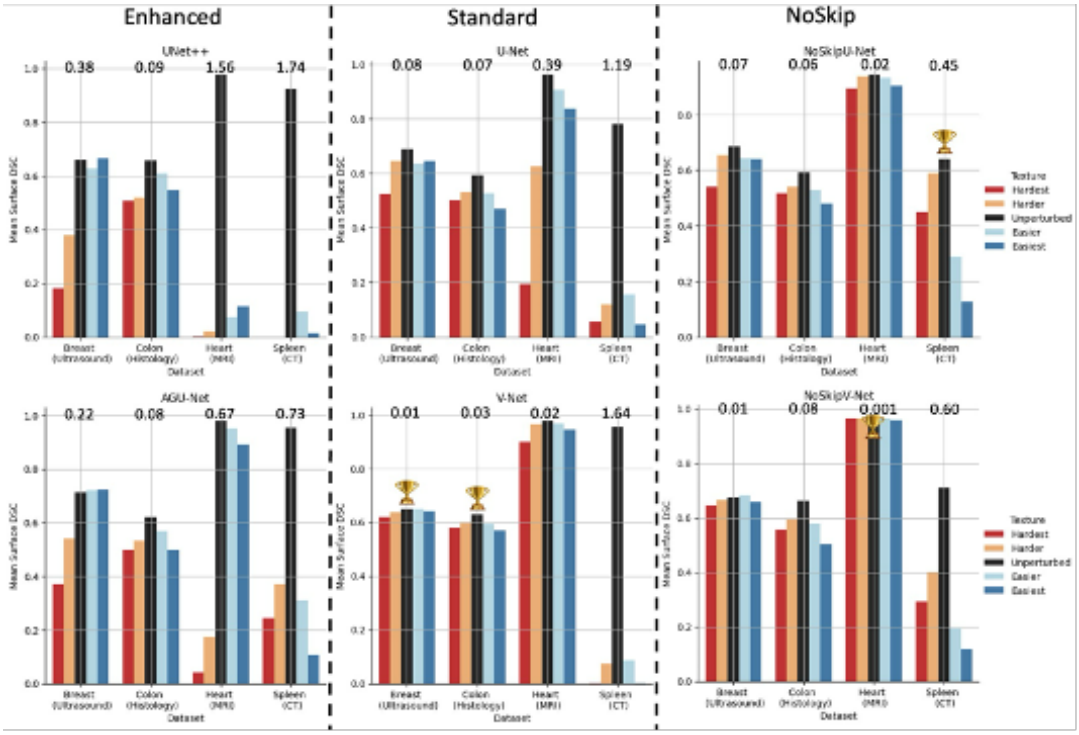


FIGURE A.12: SurfaceDSC variations across model types - UNet++, U-Net, V-Net, AGU-Net, NoSkipU-Net and NoSkipV-Net over five levels of TS, in the same format as Figure. 7.8. Lower bars are better.

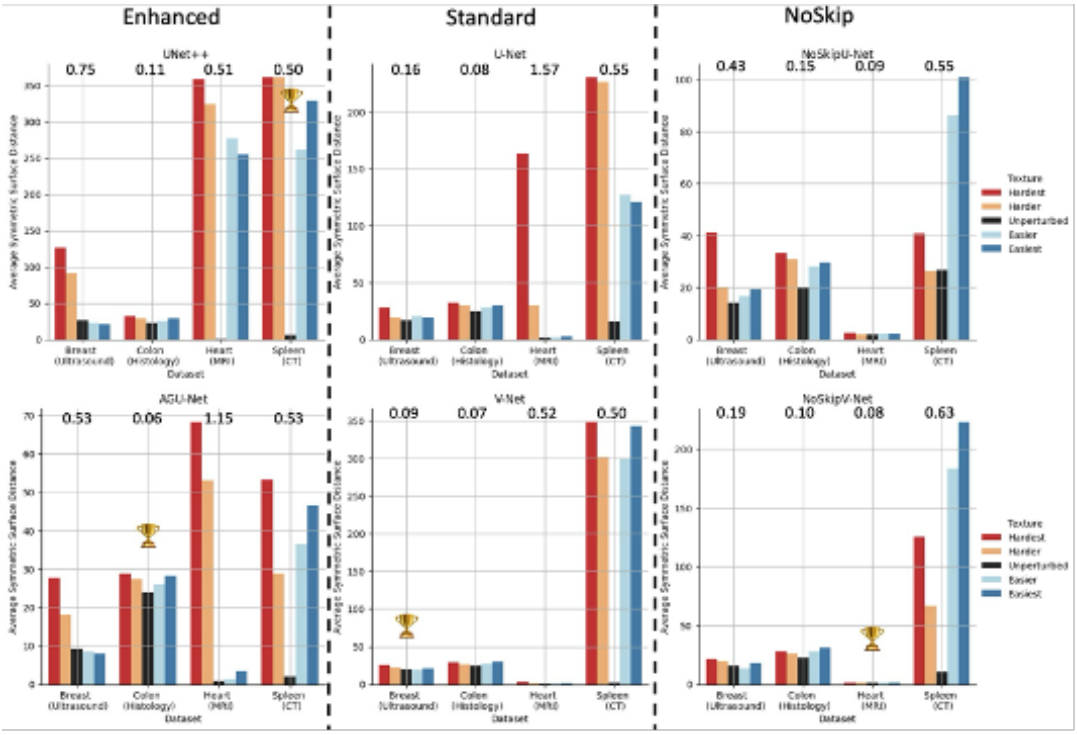


FIGURE A.13: ASSD variations across model types - UNet++, U-Net, V-Net, AGU-Net, NoSkipU-Net and NoSkipV-Net over five levels of TS, in the same format as Figure. 7.8. Lower bars are better.

TABLE A.5: The best-performing model using ASSD for all four data sets, in the same format as Table 7.1.

	Breast US	Colon Histology	Heart MRI	Spleen CT
Easiest	AGU-Net (0.428)	AGU-Net (0.287)	AGU-Net (0.408)	AGU-Net (0.740)
Easier	AGU-Net (0.469)	UNet++ (0.475)	AGU-Net (0.475)	AGU-Net (0.583)
Unperturbed	AGU-Net (0.285)	NoSkipV-Net (0.312)	AGU-Net (0.341)	V-Net (0.421)
Harder	NoSkipU-Net (0.265)	NoSkipV-Net (0.412)	V-Net (0.691)	NoSkipU-Net (0.754)
Hardest	NoSkipV-Net (0.448)	AGU-Net (0.350)	NoSkipV-Net (0.800)	NoSkipU-Net (0.796)

Representative Images from All Data Sets

Figures A.14, A.15, A.16, and A.17 show representative images from each of the four data sets we ran this analysis on for a qualitative demonstration of how different the results can be based on the architecture difference.

A.3.3 Representativeness of the Evaluated Perturbations on the TS Scale with Real-World Perturbations

This section justifies how real-world perturbations on MRI and CT images relate to the ranges of TS in which we evaluate the model architectures in the results. Figures A.18 and A.19 show real-world domain shifts during inference simulated through

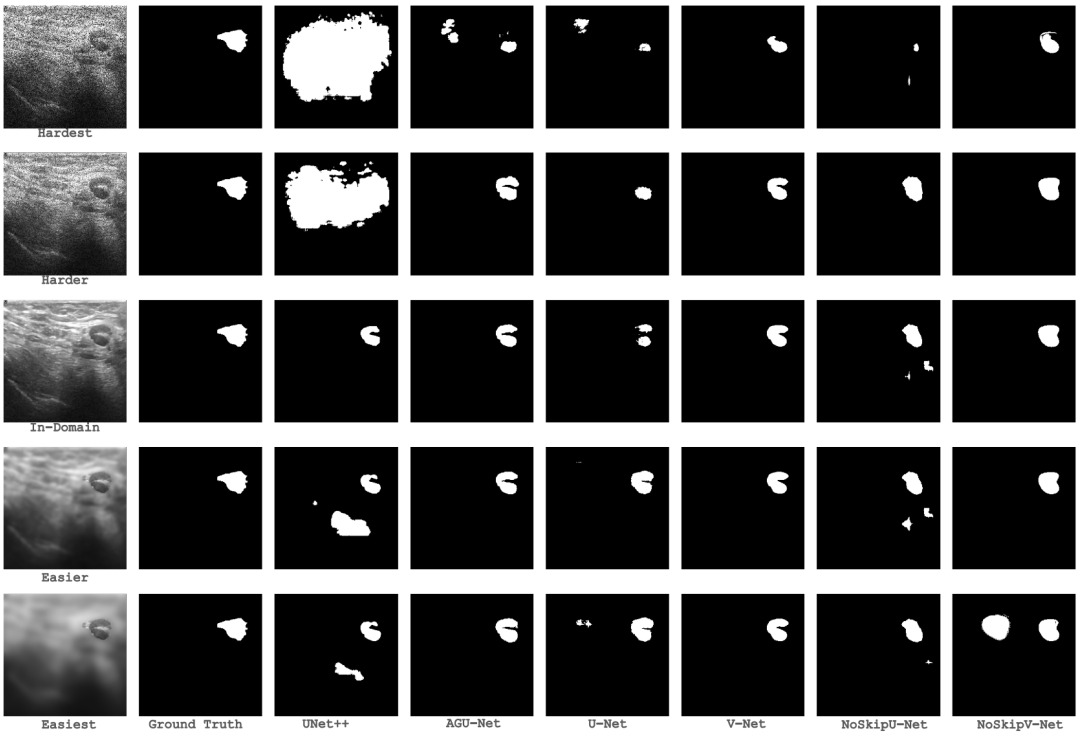


FIGURE A.14: Representative images, ground truth, and segmentation results for AGU-Net, U-Net and NoSkipU-Net for hardest, harder, unperturbed, easier, and easiest texture similarities for Breast US data set.

FIGURE A.15: Representative images in the same format as Figure. A.14 for Colon Histology data set.

various perturbations, with density distributions of the TS in these conditions. These

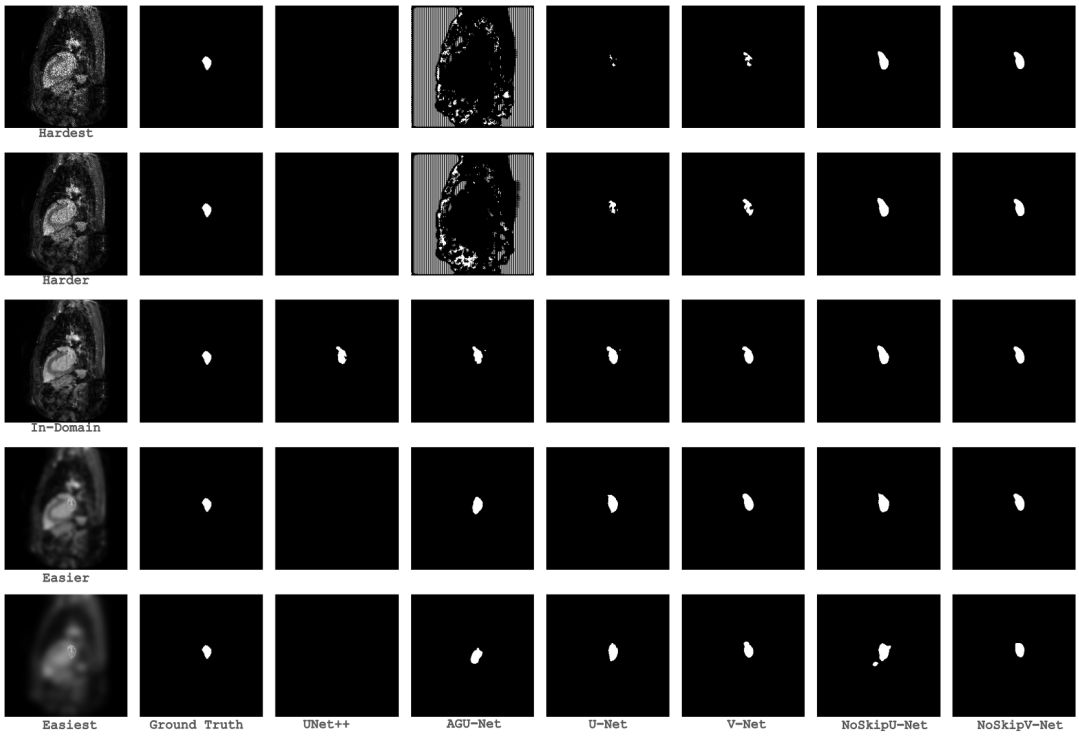


FIGURE A.16: Representative images in the same format as Figure. A.14 for Heart MRI data set.

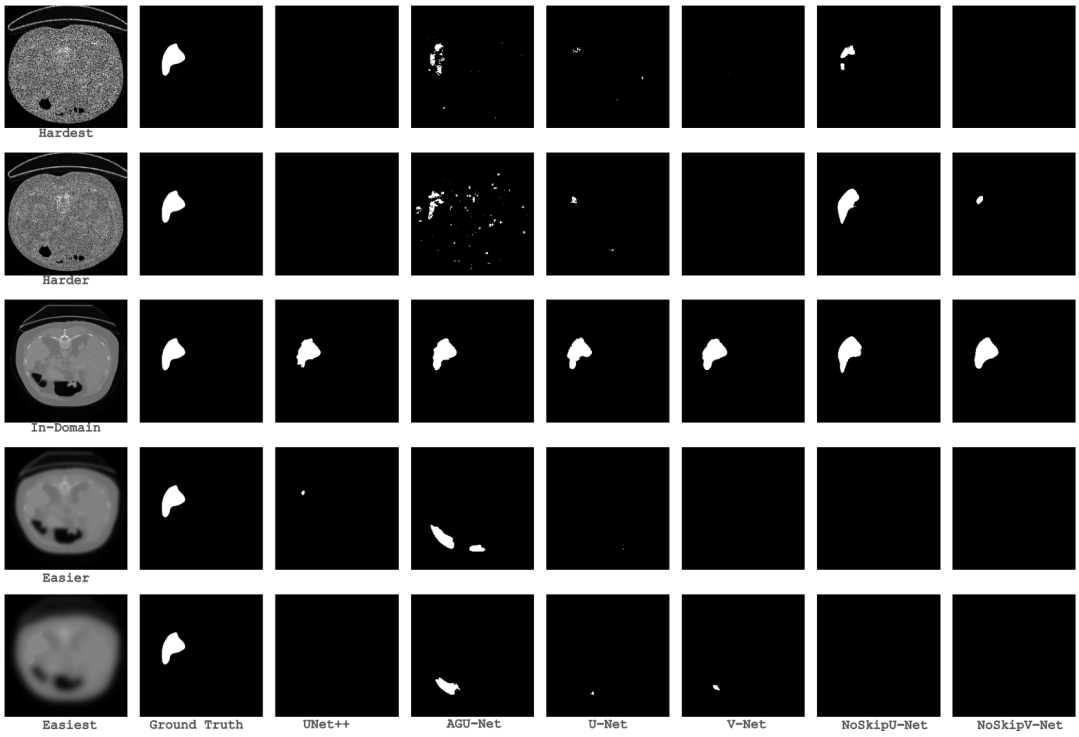


FIGURE A.17: Representative images in the same format as Figure. A.14 for Spleen CT data set.

are within the range of the five levels of perturbations reported in the results section, ranging from $1e^{-2}$ to 1.

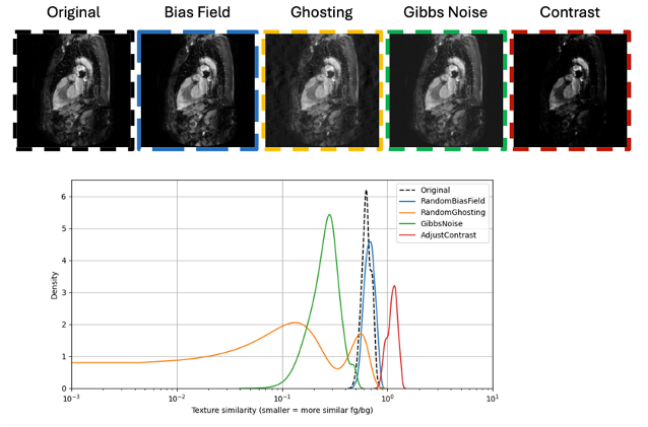


FIGURE A.18: Real-world domain shifts during inference are simulated through bias fields, ghosting, noise, and contrast adjustments on the heart MRI test set, with density distribution of the TS in these conditions. The boundary colours around each image on the top match the colour of the line used to represent the distribution of TS for the images in the category of domain shift. These are within the range of the five levels of perturbations reported in the results. They range from $1e{-}2$ to 1, being in a narrower range than the five levels in which we report our results.

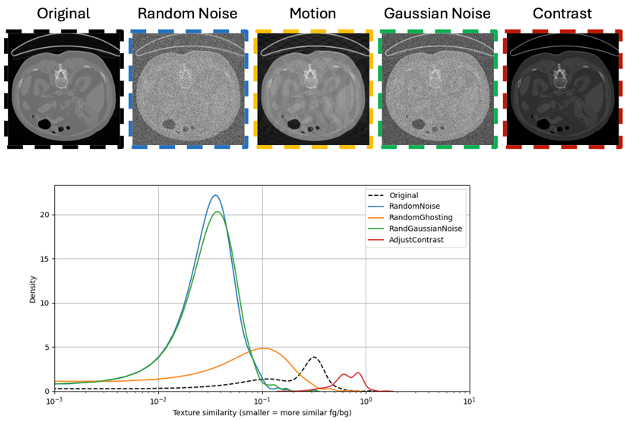


FIGURE A.19: Real-world domain shifts during inference are simulated through random noise, motion, Gaussian noise, and contrast adjustments on the Spleen CT test set, with density distribution of the TS in these conditions. These perturbations are generated using standard implementations from TorchIO and MONAI ([480], v1.1).

Figure A.20 shows three pairs of synthetic textures with nine levels of FG-blending-into-BG, indicating that the TS ranges are generally within $1e{-}2$ to 1, covering typical medical image TS ranges.

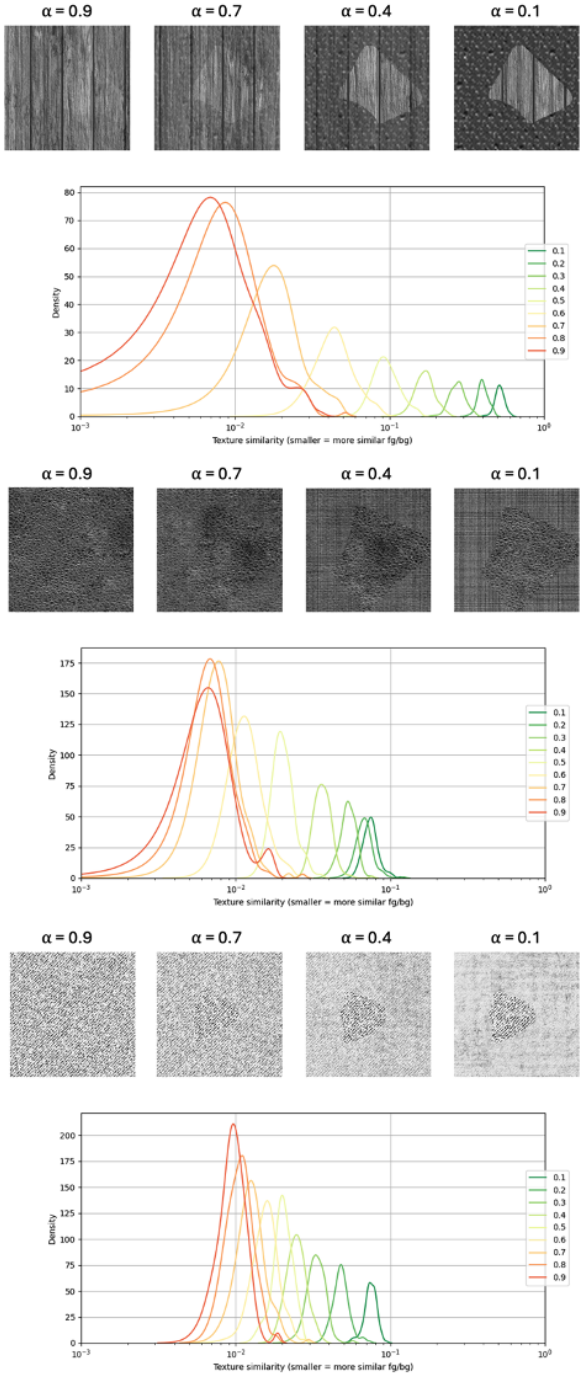


FIGURE A.20: Three pairs of synthetic textures with nine levels of FG-blending-into-BG, showing that the TS ranges are generally within $1e{-}2$ to 1, with varying bandwidths dependent on the textures chosen. All these ranges cover typical medical image TS ranges, indicating that the behaviour of the model architectures under test would be similar.

A.4 Chapter 8: More metrics and results

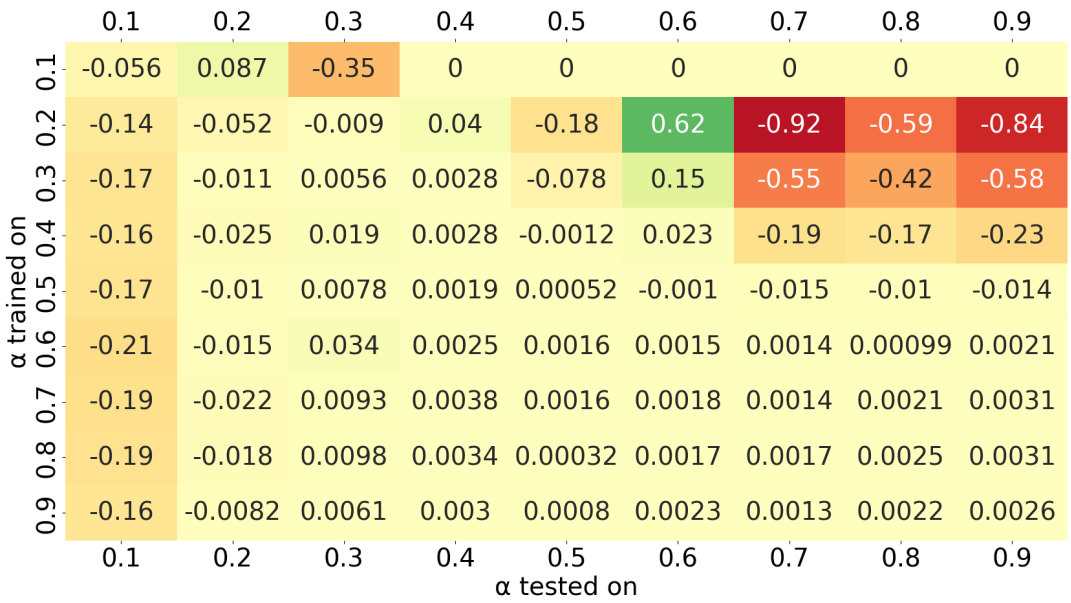


FIGURE A.21: OOD robustness metrics for U-Net versus NoSkip-U-Net, corresponding to Figure. 4 (right) in the main paper. Note similar texture combinations as AGU-Net where NoSkip-U-Net performs better than U-Net.

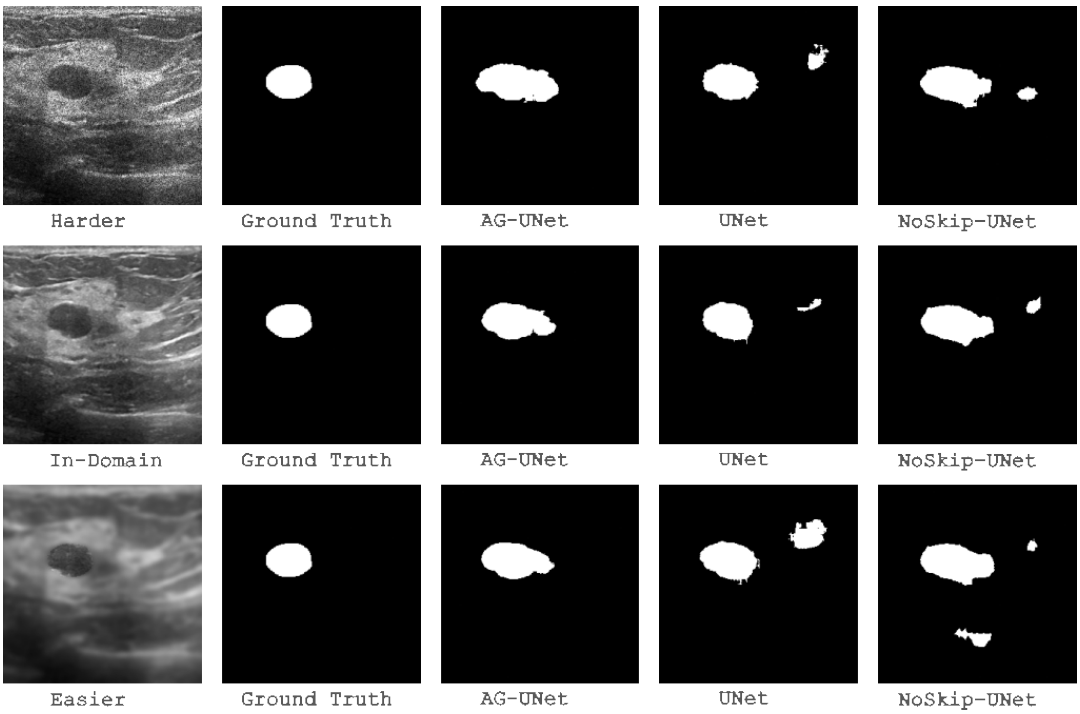


FIGURE A.22: Representative images, ground truth, and segmentation results for AGU-Net, U-Net and NoSkip-U-Net for harder, in-domain and easy texture similarities for Breast US data set.

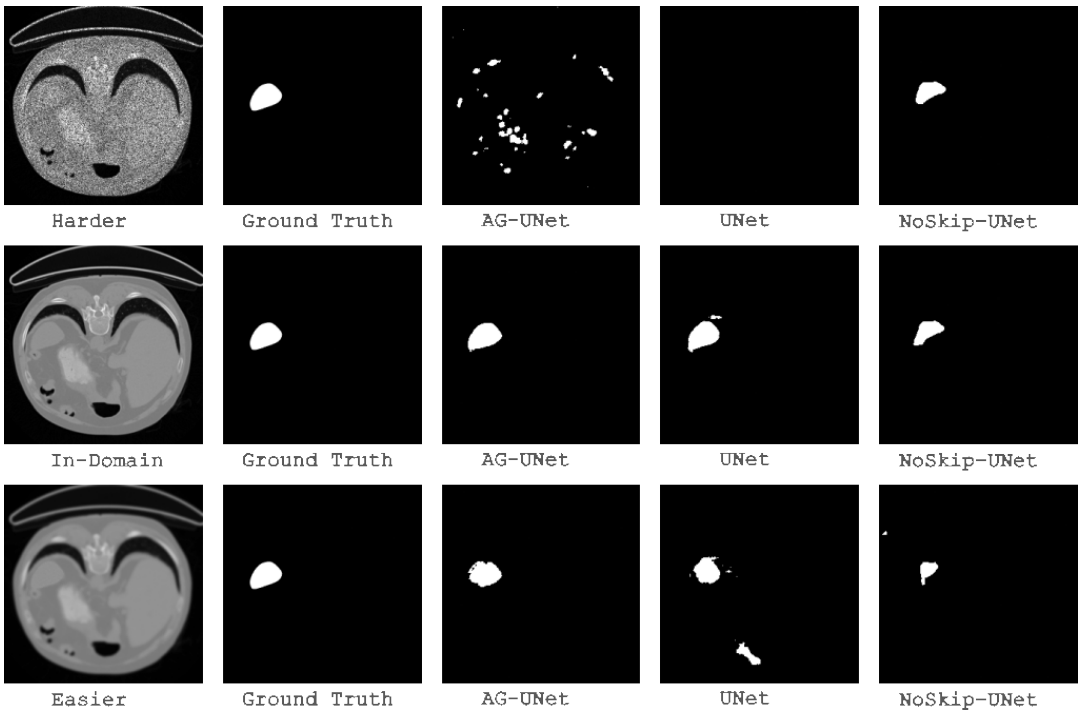


FIGURE A.23: Representative images in the same format as Figure. A.22 for Spleen CT.

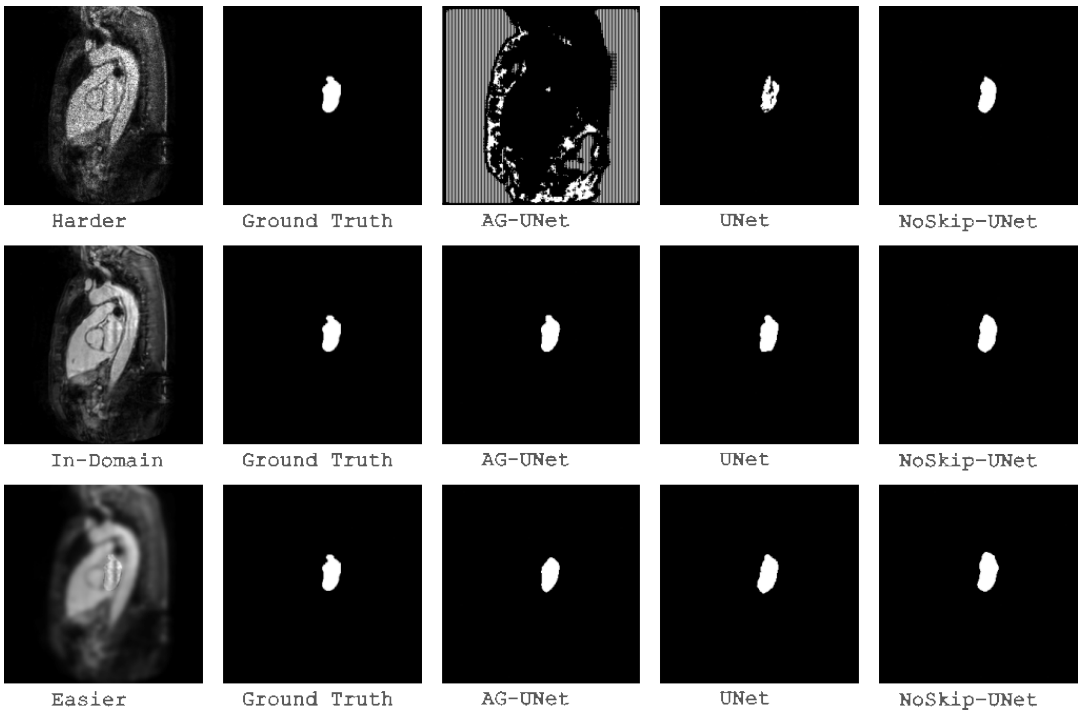


FIGURE A.24: Representative images in the same format as Figure. A.22 for Heart MRI.

TABLE A.6: Mean (standard deviation) of HDs for each of hard, in-domain and easy textures on the Breast (US), Spleen (CT) and Heart (MRI) data sets. Best performing model at each texture level is highlighted in bold. NaN and Inf values are replaced by diagonal of image size.

Data set	Texture level	AGU-Net	U-Net	NoSkip-U-Net
Breast (US)	Harder	**61.81** (49.93)	71.72 (75.92)	62.72 (58.44)
	In-domain	**41.31** (48.11)	66.51 (64.86)	64.61 (54.88)
	Easier	**38.30** (46.59)	75.77 (58.87)	79.69 (52.78)
Spleen (CT)	Harder	118.61 (58.45)	250.62 (138.83)	**57.96** (80.01)
	In-domain	**9.53** (21.33)	46.74 (71.86)	55.15 (82.64)
	Easier	**30.44** (44.53)	108.03 (125.93)	81.75 (102.90)
Heart (MRI)	Harder	126.49 (52.41)	36.92 (89.06)	**6.15** (4.92)
	In-domain	**4.31** (3.68)	8.46 (13.62)	7.24 (10.47)
	Easier	**6.64** (4.43)	9.19 (7.11)	9.57 (14.38)

B

Outreach Activities

B.1 Bern AI in Radiotherapy Symposium (March 2025)

The research for this project led to the organization of Bern AI in RadioTherapy (BART), a one-day academic event held in March 2025 at the University of Bern. It focused on the application of artificial intelligence in radiation oncology, bringing together over 100 in-person participants, including students, researchers, and industry representatives.

FIGURE B.1: Keynote talk at the BART Symposium.

BART featured keynote lectures from experts based in Munich, Cambridge, and Zurich, alongside a poster session with contributions from three countries. A technical panel discussion included both academic and industry stakeholders. The event successfully promoted interdisciplinary exchange and fostered connections among early-career researchers in biomedical engineering, medical physics, and oncology. A post-event survey showed high satisfaction, with many attendees expressing strong interest in a 2026 iteration of BART.

B.2 Science Pitches and Awards

FIGURE B.2: Receiving the Centre for AI in Medicine (CAIM) Young Researcher Award for Innovation, 2022.

This research was featured in an interview at the CAIM and won the 2022 CAIM Young Researcher Award for Innovation. Subsequently, the research leading to Chapter 11 also won this award in 2024 for Zahira Mercado, who was a master thesis advisee as part of this project.

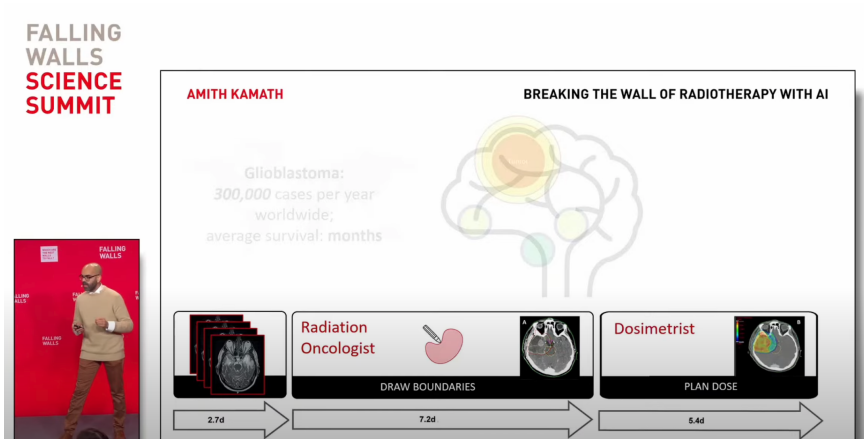


FIGURE B.3: At the Falling Walls Science Summit in Berlin, 2024.

This research was also selected amongst 100 "Lab" pitches for a 3 minute presentation at the 35th anniversary of the fall of the Berlin Wall, at the Science Summit 2024.

C

Software Artefacts and Repositories

[C4, Chapter 5, Research Question: A2]
"Comparing the Performance of Radiation Oncologists versus a Deep Learning Dose Predictor to Estimate Dosimetric Impact of Segmentation Variations for Radiotherapy." GitHub: amithjkamath/radonc-vs-dldp

[C3, Chapter 8, Research Question: B2]
"Do We Really Need that Skip-Connection? Understanding Its Interplay with Task Complexity." GitHub: amithjkamath/to_skip_or_not

[C2, Chapter 10, Research Question: C1]
"ASTRA: Atomic Surface Transformations for Radiotherapy Quality Assurance." GitHub: amithjkamath/astra

[C1, Chapter 4, Research Question: B1]
"How sensitive are Deep Learning based Radiotherapy Dose Prediction Models to Variability in Organs at Risk Segmentation?" GitHub: amithjkamath/deepdosesens

[W2, Chapter 11, Research Question: C2]
"AutoDoseRank: Automated Dosimetry-Informed Segmentation Ranking for Radiotherapy." GitHub: amithjkamath/autodoserank

[W1, Chapter 9, Research Question: B2]
"How do 3D image segmentation networks behave across the context versus foreground ratio trade-off?" GitHub: amithjkamath/context_vs_fbr

Additional contributions made to "PyRaDiSe: A Python package for DICOM-RT-based auto-segmentation pipeline construction and DICOM-RT data conversion" GitHub: ubern-mia/pyradise. These contributions specifically include being able to read and process RT Dose file formats, for conversion between NIfTI to DICOM for complete RT planning workflows.

List of Abbreviations

5-ALA 5-Aminolevulinic Acid. 2

AAA Anisotropic Analytical Algorithm. 22, 45, 53, 59, 71, 127

AAPM American Association of Physicists in Medicine. 29, 41

ABAS Atlas-Based Auto-Segmentation. 29

AGU-Net Attention Gated U-Net. xx–xxii, xxiv, xxv, 16, 27, 28, 81, 83–85, 88–92, 94–97, 103, 105–107, 110–112, 136, 139, 145, 159–163, 167, 169

AI Artificial Intelligence. vii, 3, 5, 7, 8, 11, 12, 14–17, 19, 24, 26, 29–31, 33, 34, 49, 70, 135, 138, 140–142, 144, 148, 149

AQUA Automated QUality Assurance. 13

ASPP Atrous Spatial Pyramid Pooling. 27

ASSD Average Symmetric Surface Distance. xviii, xxiv, xxviii, 36, 37, 89, 159–162

ASTRA Atomic Surface Transformations for Radiotherapy quality Assurance. 17, 118, 119, 136, 140, 142

AUC-ROC Area Under the Receiver Operating Characteristic Curve. 35, 60

BART Bern AI in RadioTherapy. xxv, 171

BBB Blood-Brain Barrier. 3, 4

BG Background. xx, xxi, xxiv, 82–87, 89–93, 97, 102–107, 144, 158–161, 165, 166

C3D Cascaded Three-Dimensional. 52, 53, 59, 67, 71, 78, 115, 119, 127

CAIM Centre for AI in Medicine. xxv, 172

CBCT Cone-Beam Computed Tomography. 23

CDF Cumulative Distribution Function. xxii, xxiii, 130, 131

CI Conformity Index. 13, 22, 38

CI Confidence Interval. 128

CNN Convolutional Neural Network. 26–31, 33, 34

CNS Central Nervous System. 1, 3

CRF Continuous Random Field. 27

CT Computed Tomography. xx, xxi, xxiv, xxv, xxvii, xxviii, 5, 7, 19, 20, 28–30, 32, 34, 53, 59, 60, 71, 75, 83, 86, 87, 94–96, 98, 102, 104, 105, 107, 119, 127, 148, 160, 162, 164, 165, 168, 169

CTV Clinical Target Volume. xvii, 5, 6, 21, 30, 43–47, 49, 51, 72, 142–144

CV Coefficient of Variation. xxi, 89, 93, 94, 97, 99

DI Discordance Index. 13

DICOM Digital Imaging and Communications in Medicine. 71, 173

DINO DIstillation with NO labels. 25

DL Deep Learning. xviii, xix, xxii, 7, 14–16, 24–27, 29–34, 41, 52, 57–62, 69–71, 85, 102, 109, 118, 120, 123, 125–127, 131, 135–140, 146

DNA Deoxy-Ribonucleic Acid. 3, 4

DSC Dice Similarity Coefficient. xviii, xx–xxii, xxiv, xxvii, xxviii, 8, 13, 14, 16, 28, 29, 31, 36, 37, 44, 54, 55, 57, 58, 73, 75, 76, 88–95, 103, 105–107, 111, 112, 125, 139, 142, 158–161

DTI Diffusion Tensor Imaging. 5

DVH Dose-Volume Histogram. xviii–xx, xxiii, xxvii, 13, 22, 33, 37, 52–56, 69, 70, 73, 75–79, 138, 141–143, 146

EMBC Conference on Engineering in Medicine and Biology. 17, 217

EORTC European Organization for Research and Treatment of Cancer. 41, 148

ESTRO European Society for RadioTherapy and Oncology. 21, 71, 149

FBR Foreground to Background Ratio. xxii, 109–112

FCN Fully Convolutional Network. xviii, 27

FG Foreground. xx, xxi, xxiv, 82–87, 89–93, 97, 102–107, 144, 158–161, 165, 166

GBM Glioblastoma Multiforme. xvii, xxvii, 1–5, 9, 20, 21, 30, 43, 44, 53, 55, 57, 59, 70, 72, 78–80, 117–119, 125, 127, 138, 140, 142, 144, 146

GMI Geographical Miss Index. 13

GPR Gaussian Process Regression. 34

GPU Graphics Processing Unit. 26, 28, 52, 53, 61, 71–73, 79, 88, 106, 119

GTV Gross Tumour Volume. xvii, 5, 6, 21, 142

HD Hausdorff Distance. xviii, xxi, xxii, xxvii, xxviii, 8, 13, 29, 36, 37, 44, 57, 88, 92, 93, 95, 96, 112, 125, 139, 158, 159, 169

HI Homogeneity Index. 22, 38

HU Hounsfield Units. 20

IDH Isocitrate Dehydrogenase. 1, 2

IEEE Institute of Electrical and Electronics Engineers. 217, 218

IMRT Intensity-Modulated Radiotherapy. 2, 5, 7–9, 13, 21, 22, 33, 34, 78, 79, 145, 148

ISBI International Symposium for Biomedical Imaging. 16, 142, 218

JCI Jaccard Conformity Index. 13

KBP Knowledge-Based Planning. 33, 53, 54, 69–71, 75, 78

KL Kullback–Leibler. xx, xxi, 84, 85, 102, 103, 141

LAD Left Anterior Descending Artery. 7

LBP Local Binary Pattern. xx, xxi, 84, 85, 99, 102, 103

LINAC Linear Particle Accelerator. 4, 33, 149

LLM Large Language Model. 33, 45, 145

MAE Mean Absolute Error. 53, 69, 119, 138

MCO Multi Criteria Optimization. 19

MGMT O6-Methyl Guanine-DNA Methyltransferase. 3

MICCAI Medical Image Computing and Computer-Assisted Intervention. 16, 17, 217, 218

MIDL Medical Imaging with Deep Learning. 16, 217

mIoU mean Intersection over Union. 145

MLC Multi-Leaf Collimator. 21, 22, 34, 148

MLP Multi-Layer Perceptron. 110

MMPs Matrix Metalloproteinases. 4

MONAI Medical Open Network for AI. 83, 88, 99, 105, 110, 111

MRI Magnetic Resonance Imaging. xx, xxiv, xxv, xxvii, xxviii, 2, 5, 19–21, 23, 28, 32, 44, 59, 83, 86, 87, 94–96, 98, 102, 104, 107, 127, 139, 147, 149, 160, 162, 164, 165, 168, 169

MRS Magnetic Resonance Spectroscopy. 5

MSD Mean Surface Distance. 29

MSD Medical Segmentation Decathlon. 109, 110

MU Monitor Unit. 22

NASA National Aeronautics and Space Administration. 17

NDPM Normalized Distance-based Performance Measure. xxii, 128–130

NeurIPS Conference on Neural Information Processing Systems. 16, 218

NIfTI Neuroimaging Informatics Technology Initiative. 59, 71, 127, 173

NPC Naso-Pharyngeal Cancer. 9

NRG NRG Oncology. 13

NTCP Normalized Tissue Complication Probability. 49, 137, 147

OAR Organ at Risk. xvii–xix, xxii, xxiii, xxvii, xxviii, 5–8, 11–14, 16, 17, 19, 21, 22, 30, 33, 36, 37, 41, 43–55, 57–61, 63, 70–72, 74–76, 78–80, 115, 117–123, 125–128, 130, 131, 137, 138, 140, 142–144, 146, 148, 151, 154–157

ONL Optic Nerve Left. xx, xxvii, 73, 75, 76, 80

OOD Out Of Distribution. xxv, 74, 98, 102, 103, 106, 136, 138, 167

PCA Principal Components Analysis. 25

PET Positron Emission Tomography. 5, 19, 20

PI3K Phosphatidylinositol 3-Kinase. 4

PTV Planning Target Volume. xvii, 6, 12, 13, 21, 51, 53, 70–73, 119

QA Quality Assurance. xvii, xviii, xxiii, 7, 8, 10–17, 19, 22–24, 31, 33, 34, 43, 50–52, 57, 58, 67, 69, 70, 79, 80, 115, 117, 118, 123, 135–140, 142, 144, 148–150

ReLU Rectified Linear Unit. 88, 105

RPA Radiation Planning Assistant. 21, 22, 33

RT Radiotherapy. xvii, xviii, xxii, xxiii, 2–6, 8–13, 15–17, 19–24, 29, 31, 33, 34, 36, 37, 41, 43, 50–52, 57, 58, 69, 70, 79, 80, 117, 118, 123, 125–128, 135, 138, 140, 142, 144–150, 173

RTOG RadioTherapy Oncology Group. 13

RTSS Radio Therapy Structure Set. 71

SABR Stereotactic Ablative Body Radiotherapy. 149

SAM Segment Anything Model. 30, 32, 67

SBRT Stereotactic Body Radiation Therapy. 29, 33

SRS Stereotactic Radiosurgery. 3, 29

SSL Semi- or Self-Supervised Learning. 25

TPS Treatment Planning System. xx, 8, 21–23, 58, 60, 71, 73, 75, 76, 127, 131

TRL Technology Readiness Level. xvii, 17, 18

TS Texture Similarity. xx, xxi, xxiv, 82–91, 93, 94, 96–99, 102–105, 144, 158, 161–163, 165, 166

TV Target Volume. xvii–xix, xxii, xxiii, xxviii, 5–8, 12–14, 16, 17, 19–21, 33, 36–38, 43, 45, 50–54, 57, 59–61, 63, 67, 70, 71, 73–79, 117, 119–122, 127, 128, 135, 146, 147, 155–157

UNet++ U-Net with multiple skip-connections. 28, 34, 136, 137, 145

UNETR U-Net + Transformer Hybrid. xxii, 16, 27, 110–112, 139

US Ultra Sound. xx, xxi, xxiv, xxv, xxvii, xxviii, 83, 86, 87, 93, 95–97, 102, 104, 105, 107, 160, 162, 163, 167, 169

VGG Visual Geometry Group. 27

ViT Vision Transformer. 27, 30, 67

VMAT Volumetric Modulated Arc Therapy. 5, 9, 11, 21, 22, 33, 34, 45, 53, 59, 71, 78–80, 119, 127, 138–140, 145, 148

VRI van't Riet Index. 13

WHO World Health Organization. 1

Bibliography

[1] Aaron C Tan et al. “Management of glioblastoma: State of the art and future directions”. In: *CA: a cancer journal for clinicians* 70.4 (2020), pp. 299–312.

[2] Dorothee Gramatzki et al. “Glioblastoma in the Canton of Zurich, Switzerland revisited: 2005 to 2009”. In: *Cancer* 122.14 (2016), pp. 2206–2215.

[3] Quinn T. Ostrom, Nirav Patil, Gino Cioffi, et al. “CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2013–2017”. In: *Neuro-Oncology* 22.Suppl 2 (2020), pp. iv1–iv96.

[4] George St Stoyanov and Deyan L Dzhenkov. “On the Concepts and History of Glioblastoma Multiforme - Morphology, Genetics and Epigenetics”. In: *Folia Medica (Plovdiv)* 60.1 (Mar. 2018), pp. 48–66. ISSN: 0204-8043. DOI: 10.1515/folmed-2017-0069.

[5] David N Louis et al. “The 2021 WHO classification of tumors of the central nervous system: a summary”. In: *Neuro-oncology* 23.8 (2021), pp. 1231–1251.

[6] Francesco Andreatta et al. “The Organoid Era Permits the Development of New Applications to Study Glioblastoma”. In: *Cancers* 12 (2020).

[7] Brian M. Alexander and Timothy F. Cloughesy. “Adult Glioblastoma”. In: *Journal of Clinical Oncology* 35.21 (2017), pp. 2402–2409.

[8] Sang Kyun Lim et al. “Glioblastoma multiforme: a perspective on recent findings in human cancer and mouse models”. In: *BMB reports* 44 3 (2011), pp. 158–64.

[9] Mustafa Emre Sarac et al. “Potential Biomarkers for IDH-Mutant and IDH-Wild-Type Glioblastomas: A Single-Center Retrospective Study”. In: *Journal of Clinical Medicine* 14.7 (2025), p. 2518.

[10] Andrea Sottoriva, Immaculada Spiteri, Sara G. M. Piccirillo, et al. “Intratumour Heterogeneity in Human Glioblastoma Reflects Cancer Evolutionary Dynamics”. In: *Proceedings of the National Academy of Sciences* 110.10 (2013), pp. 4009–4014.

[11] Roel G. Verhaak, Katherine A. Hoadley, Elizabeth Purdom, et al. “Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1”. In: *Cancer Cell* 17.1 (2010), pp. 98–110.

[12] Benjamin M. Ellingson, Martin Bendszus, Jerrold Boxerman, et al. “Consensus Recommendations for a Standardized Brain Tumour Imaging Protocol in Clinical Trials”. In: *Neuro-Oncology* 17.9 (2015), pp. 1188–1198.

[13] Kate E Hills, Kostas Kostarelos, and Robert C Wykes. "Converging mechanisms of epileptogenesis and their insight in glioblastoma". In: *Frontiers in molecular neuroscience* 15 (2022), p. 903115.

[14] Roger Stupp et al. "Radiotherapy plus Concomitant and Adjuvant Temozolomide for Glioblastoma". In: *The New England Journal of Medicine* 352.10 (Mar. 2005), pp. 987–996.

[15] Philip C. De Witt Hamer, Santiago G. Robles, Aeilko H. Zwinderman, et al. "Impact of Intraoperative Stimulation Brain Mapping on Glioma Surgery Outcome: A Meta-Analysis". In: *Journal of Clinical Oncology* 30.20 (2012), pp. 2559–2565.

[16] Nader Sanai, Mei Yin Polley, Michael W. McDermott, et al. "An Extent of Resection Threshold for Newly Diagnosed Glioblastomas". In: *Journal of Neurosurgery* 115.1 (2011), pp. 3–8.

[17] Christian Senft, Andrea Bink, Kea Franz, et al. "Intraoperative MRI Guidance and Extent of Resection in Glioma Surgery: A Randomised, Controlled Trial". In: *Lancet Oncology* 12.11 (2011), pp. 997–1003.

[18] Daniel A. Orringer, Alexandra Golby, and Ferenc A. Jolesz. "Neuronavigation in the Surgical Management of Brain Tumors: Current and Future Trends". In: *Expert Review of Medical Devices* 9.5 (2012), pp. 491–500.

[19] Walter Stummer, Uwe Pichlmeier, Thomas Meinel, et al. "Fluorescence-Guided Surgery With 5-Aminolevulinic Acid for Resection of Malignant Glioma: A Randomised Controlled Multicentre Phase III Trial". In: *Lancet Oncology* 7.5 (2006), pp. 392–401.

[20] Malcolm D. Walker, Thomas A. Strike, and George E. Sheline. "An Analysis of Dose-Effect Relationship in the Radiotherapy of Malignant Gliomas". In: *International Journal of Radiation Oncology, Biology, Physics* 5.10 (1979), pp. 1725–1731.

[21] Kwan Ho Cho et al. "Simultaneous integrated boost intensity-modulated radiotherapy in patients with high-grade gliomas". In: *International Journal of Radiation Oncology* Biology* Physics* 78.2 (2010), pp. 390–397.

[22] William Roa, Philip M. Brasher, Glenn Bauman, et al. "Abbreviated Course of Radiation Therapy in Older Patients With Glioblastoma Multiforme: A Prospective Randomized Clinical Trial". In: *Journal of Clinical Oncology* 22.9 (2004), pp. 1583–1588.

[23] Jean-Emmanuel Bibault, Philippe Giraud, and Anita Burgun. "Big Data and Machine Learning in Radiation Oncology: State of the Art and Future Prospects". In: *Cancer Letters* 382.1 (2016), pp. 110–117.

[24] Monika E. Hegi, Annie-Claire Diserens, Thierry Gorlia, et al. "MGMT Gene Silencing and Benefit From Temozolomide in Glioblastoma". In: *New England Journal of Medicine* 352.10 (2005), pp. 997–1003.

[25] Guido Frosina. "DNA Repair and Resistance of Gliomas to Chemotherapy and Radiotherapy". In: *Molecular Cancer Research* 7.7 (2009), pp. 989–999.

[26] Estefania Carrasco-Garcia, Miguel Saceda, and Isabel Martínez-Lacaci. "Role of Receptor Tyrosine Kinases and Their Ligands in Glioblastoma". In: *Cells* 3 (2014), pp. 199–235.

[27] James M. Markert. "Glioblastoma multiforme: introduction." In: *Cancer journal* 9 3 (2003), p. 148.

[28] Roger Stupp, Sophie Taillibert, Andrew A. Kanner, et al. "Effect of Tumor-Treating Fields Plus Maintenance Temozolomide vs Maintenance Temozolomide Alone on Survival in Patients With Glioblastoma". In: *JAMA* 314.23 (2015), pp. 2535–2543.

[29] Oumaima Aboubakr et al. "Long-term survivors in 976 supratentorial glioblastoma, IDH-wildtype patients". In: *Journal of neurosurgery* 142.1 (2024), pp. 174–186.

[30] Michael Weller et al. "Rindopepimut with temozolomide for patients with newly diagnosed, EGFRvIII-expressing glioblastoma (ACT IV): a randomised, double-blind, international phase 3 trial". In: *The Lancet Oncology* 18.10 (2017), pp. 1373–1385.

[31] Philip Huxham Jansen van Rensburg and Malan Nicolaas Jansen van Rensburg. "Glioblastoma multiforme has many faces". In: *South African Medical Journal* 16 (2012), pp. 94–99.

[32] Alexandre Ciuffi Faustino, Gustavo Arruda Viani, and Ana Carolina Hamamura. "Patterns of recurrence and outcomes of glioblastoma multiforme treated with chemoradiation and adjuvant temozolomide". In: *Clinics* 75 (2020), e1553.

[33] Brenda Auffinger et al. "New therapeutic approaches for malignant glioma: in search of the Rosetta stone". In: *F1000 Medicine Reports* 4 (2012).

[34] Amy Bradshaw et al. "Cancer Stem Cells in Glioblastoma Multiforme". In: *Frontiers in Surgery* 3 (2016).

[35] Jitin Bajaj et al. "Practice Patterns for Managing Recurrent Glioblastoma Multiforme". In: *Indian Journal of Neurosurgery* (2024).

[36] Michel Lacroix, Dima Abi-Said, David R. Fourney, et al. "A Multivariate Analysis of 416 Patients with Glioblastoma Multiforme: Prognosis, Extent of Resection, and Survival". In: *Journal of Neurosurgery* 95.2 (2001), pp. 190–198.

[37] Hans J. Scherer. "The Forms of Growth in Gliomas and Their Practical Significance". In: *Brain* 63.1 (1940), pp. 1–35.

[38] Norbert Galldiks et al. "Challenges, limitations, and pitfalls of PET and advanced MRI in patients with brain tumors: A report of the PET/RANO group". In: *Neuro-oncology* (2024), noae049.

[39] Nader Sanai and Mitchel S. Berger. "Glioma Extent of Resection and Its Impact on Patient Outcome". In: *Neurosurgery* 62.4 (2008), pp. 753–764.

[40] Franz Josef Klinz et al. "Genetic Diversity of Glioblastoma Multiforme: Impact on Future Therapies". In: 2011.

[41] D. Fortin. "The Blood-Brain Barrier: Its Influence in the Treatment of Brain Tumors Metastases". In: *Current Cancer Drug Targets* 12.3 (2012), pp. 247–259.

[42] Costas D. Arvanitis, Giovanni B. Ferraro, and Rakesh K. Jain. "The Blood–Brain Barrier and Blood–Tumor Barrier in Brain Tumors and Metastases". In: *Nature Reviews Cancer* 20.1 (2020), pp. 26–41.

[43] Jasmine L King and Soumya Rahima Benhabbour. "Glioblastoma Multiforme—A Look at the Past and a Glance at the Future". In: *Pharmaceutics* 13 (2021).

[44] Santosh Kesari. "Understanding glioblastoma tumor biology: the potential to improve current diagnosis and treatments." In: *Seminars in oncology* 38 Suppl 4 (2011), S2–10.

[45] Faye L. Robertson et al. “Experimental models and tools to tackle glioblastoma”. In: *Disease Models & Mechanisms* 12 (2019).

[46] Adolf Giese, Michelle A. Loo, Nhat Tran, et al. “Dichotomy of Astrocytoma Migration and Proliferation”. In: *International Journal of Cancer* 67.2 (1996), pp. 275–282.

[47] Mitsutoshi Nakada, Shunsuke Nakada, Ty Demuth, et al. “Molecular Targets of Glioma Invasion”. In: *Cellular and Molecular Life Sciences* 64.4 (2007), pp. 458–478.

[48] Nicola A. Charles, Eric C. Holland, Richard Gilbertson, et al. “The Brain Tumor Microenvironment”. In: *Glia* 59.8 (2011), pp. 1169–1180.

[49] Y. R. Lawrence, M. V. Mishra, M. Werner-Wasik, et al. “Improving Prognosis of Glioblastoma in the 21st Century: Who Has Benefited Most?” In: *Cancer* 118.17 (2012), pp. 4228–4234.

[50] Jayant S Vaidya. “Principles of cancer treatment by radiotherapy”. In: *Surgery (Oxford)* 39.4 (2021), pp. 193–201.

[51] Wayne D Newhauser and Marco Durante. “Assessing the risk of second malignancies after modern radiotherapy”. In: *Nature Reviews Cancer* 11.6 (2011), pp. 438–448.

[52] Marco Durante and Jay S Loeffler. “Charged particles in radiation oncology”. In: *Nature reviews Clinical oncology* 7.1 (2010), pp. 37–43.

[53] Burkhard Jakob et al. “Live cell microscopy analysis of radiation-induced DNA double-strand break motion”. In: *Proceedings of the National Academy of Sciences* 106.9 (2009), pp. 3172–3177.

[54] Marco Durante, Roberto Orecchia, and Jay S Loeffler. “Charged-particle therapy in cancer: clinical uses and future perspectives”. In: *Nature Reviews Clinical Oncology* 14.8 (2017), pp. 483–495.

[55] Jay S Loeffler and Marco Durante. “Charged particle therapy—optimization, challenges and future directions”. In: *Nature reviews Clinical oncology* 10.7 (2013), pp. 411–424.

[56] Anna-Karin Paulsson, Kyle P. McMullen, Ann-Marie Peiffer, et al. “Limited Margins Using Modern Radiotherapy Techniques Does Not Increase Marginal Failure Rate of Glioblastoma”. In: *American Journal of Clinical Oncology* 37.2 (2014), pp. 177–181.

[57] Giuseppe Minniti, Diego Amelio, Marco Amichetti, et al. “Patterns of Failure and Comparison of Different Target Volume Delineations in Patients with Glioblastoma Treated with Conformal Radiotherapy Plus Concomitant and Adjuvant Temozolomide”. In: *Radiotherapy and Oncology* 97.3 (2010), pp. 377–381.

[58] A. Giese et al. “Cost of Migration: Invasion of Malignant Gliomas and Implications for Treatment”. In: *Journal of Clinical Oncology* 21.8 (2003), pp. 1624–1636.

[59] Zeynettin Akkus et al. “Deep learning for brain MRI segmentation: state of the art and future directions”. In: *Journal of digital imaging* 30 (2017), pp. 449–459.

[60] Maximilian Niyazi et al. “Irradiation and bevacizumab in high-grade glioma retreatment settings”. In: *International Journal of Radiation Oncology* Biology* Physics* 82.1 (2012), pp. 67–76.

[61] Stephen J. Price, Neil G. Burnet, Tim Donovan, et al. "Diffusion Tensor Imaging of Brain Tumors at 3T: A Potential Tool for Assessing White Matter Tract Invasion?" In: *Clinical Radiology* 58.6 (2003), pp. 455–462.

[62] Laura A Dawson and Michael B Sharpe. "Image-guided radiotherapy: rationale, benefits, and limitations". In: *The lancet oncology* 7.10 (2006), pp. 848–858.

[63] Maximilian Niyazi et al. "ESTRO-EANO guideline on target delineation and radiotherapy details for glioblastoma". In: *Radiotherapy and Oncology* 184 (July 2023). ISSN: 18790887. DOI: 10.1016/j.radonc.2023.109663.

[64] X Wu et al. "Knowledge-based auto contouring for radiation therapy: Challenges in standardizing object definitions, ground truth delineations, object quality, and image quality". In: *International Journal of Radiation Oncology, Biology, Physics* 99.2 (2017), E740.

[65] Gašper Podobnik et al. "vOARiability: Interobserver and intermodality variability analysis in OAR contouring from head and neck CT and MR images". In: *Medical Physics* (2024). ISSN: 24734209. DOI: 10.1002/mp.16924.

[66] Chris McIntosh et al. "Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method". In: *Physics in Medicine & Biology* 62 (2016), pp. 5926–5944.

[67] Tomohiro Kajikawa et al. "A convolutional neural network approach for IMRT dose distribution prediction in prostate cancer patients". In: *Journal of Radiation Research* 60 (2019), pp. 685–693.

[68] David Djajaputra et al. "Algorithm and performance of a clinical IMRT beam-angle optimization system." In: *Physics in medicine and biology* 48 19 (2003), pp. 3191–212.

[69] Rita Simões et al. "Geometrical and dosimetric evaluation of breast target volume auto-contouring". In: *Physics and Imaging in Radiation Oncology* 12 (2019), pp. 38–43.

[70] Christopher F. Njeh. "Tumor delineation: The weakest link in the search for accuracy in radiotherapy". In: *Journal of Medical Physics / Association of Medical Physicists of India* 33 (2008), pp. 136–140.

[71] Barbara Segedin and Primoz Petric. "Uncertainties in target volume delineation in radiotherapy – are they relevant and what can we do about them?" In: *Radiology and Oncology* 50 (2016), pp. 254–262.

[72] Hana Baroudi et al. "Automated contouring and statistical process control for plan quality in a breast clinical trial". In: *Physics and Imaging in Radiation Oncology* 28 (2023).

[73] Jinhan Zhu et al. "Preliminary Clinical Study of the Differences Between Interobserver Evaluation and Deep Convolutional Neural Network-Based Segmentation of Multiple Organs at Risk in CT Images of Lung Cancer". In: *Frontiers in Oncology* 9 (2019).

[74] Suzanne Kirby et al. "Target Contour Consistency During Magnetic Resonance-Guided Online Adaptive Stereotactic Body Radiation Therapy". In: *Advances in Radiation Oncology* 10 (2025).

[75] Eva Versteijne et al. "Considerable interobserver variation in delineation of pancreatic cancer on 3DCT and 4DCT: a multi-institutional study". In: *Radiation Oncology (London, England)* 12 (2017).

[76] Christopher L. Nelson et al. “A real-time contouring feedback tool for consensus-based contour training”. In: *Frontiers in Oncology* 13 (2023).

[77] Susan Mercieca, José Belderbos, and Marcel B. van Herk. “Challenges in the target volume definition of lung cancer radiotherapy”. In: *Translational Lung Cancer Research* 10 (2020), pp. 1983–1998.

[78] Simona Arculeo et al. “The emerging role of radiation therapists in the contouring of organs at risk in radiotherapy: analysis of inter-observer variability with radiation oncologists for the chest and upper abdomen”. In: *ecancermedicalscience* 14 (2020).

[79] Antoine Attard and Susan Mercieca. “Beyond boundaries: exploring radiographers’ experiences and solutions in organ-at-risk delineation for radiotherapy planning”. In: *Journal of Radiotherapy in Practice* 23 (2024).

[80] Indra J. Das et al. “Intra- and inter-physician variability in target volume delineation in radiation therapy.” In: *Journal of radiation research* (2021).

[81] Venkata Veerendranadh Chebrolu et al. “Rapid Automated Target Segmentation and Tracking on 4D Data without Initial Contours”. In: *Radiology Research and Practice* 2014 (2014).

[82] Rudi Apolle et al. “Inter-observer variability in target delineation increases during adaptive treatment of head-and-neck and lung cancer”. In: *Acta Oncologica* 58 (2019), pp. 1378–1385.

[83] Vikram Velker et al. “Creation of RTOG compliant patient CT-atlases for automated atlas based contouring of local regional breast and high-risk prostate cancers”. In: *Radiation Oncology (London, England)* 8 (2013), pp. 188–188.

[84] Emin Emrah Özsavaş et al. “Automatic segmentation of anatomical structures from CT scans of thorax for RTP”. In: *Computational and Mathematical Methods in Medicine* 2014.1 (2014), p. 472890.

[85] Charlotte L. Brouwer et al. “3D Variation in delineation of head and neck organs at risk”. In: *Radiation Oncology (London, England)* 7 (2012), pp. 32–32.

[86] Maximilian Niyazi et al. “ESTRO-ACROP guideline “target delineation of glioblastomas””. In: *Radiotherapy and Oncology* 118.1 (2016), pp. 35–42.

[87] Carlos E. Cardenas et al. “Comprehensive Quantitative Evaluation of Inter-observer Delineation Performance of MR-guided Delineation of Oropharyngeal Gross Tumor Volumes and High-risk Clinical Target Volumes: An R-IDEAL Stage 0 Prospective Study”. In: *medRxiv*. 2022.

[88] Andrada Turcas et al. “Deep-learning magnetic resonance imaging-based automatic segmentation for organs-at-risk in the brain: Accuracy and impact on dose distribution”. In: *Physics and Imaging in Radiation Oncology* 27 (2023).

[89] Xiao Han. “TH-A-224-02: Atlas-Based Auto-Segmentation of CT Images for Radiotherapy Planning”. In: *Medical Physics* 38 (2011), pp. 3841–3841.

[90] Hasan Cavus et al. “Safety and efficiency of a fully automatic workflow for auto-segmentation in radiotherapy using three commercially available deep learning-based applications”. In: *Physics and Imaging in Radiation Oncology* 31 (2024).

[91] KS Clifford Chao et al. “Reduce in variation and improve efficiency of target volume delineation by a computer-assisted system using a deformable image registration approach”. In: *International Journal of Radiation Oncology* Biology* Physics* 68.5 (2007), pp. 1512–1521.

[92] Lorenzo Radici et al. "Implementation of a Commercial Deep Learning-Based Auto Segmentation Software in Radiotherapy: Evaluation of Effectiveness and Impact on Workflow". In: *Life* 12 (2022).

[93] Yunfei Hu et al. "Clinical assessment of a novel machine-learning automated contouring tool for radiotherapy planning". In: *Journal of Applied Clinical Medical Physics* 24 (2023).

[94] Helen S. Zhang et al. "Prospective Clinical Evaluation of Integrating a Radiation Anatomist for Contouring in Routine Radiation Treatment Planning". In: *Advances in Radiation Oncology* 7 (2022).

[95] Hui Lin et al. "Deep learning for automatic target volume segmentation in radiation therapy: a review." In: *Quantitative imaging in medicine and surgery* 11 12 (2021), pp. 4847–4858.

[96] Bruno A. G. da Silva, Álvaro Luiz Fazenda, and Fabiano Carlos Paixão. "Femural Head Autosegmentation for 3D Radiotherapy Planning: Preliminary Results". In: *ArXiv* abs/1812.04682 (2018).

[97] Mathis Ersted Rasmussen et al. "Potential of E-Learning Interventions and Artificial Intelligence–Assisted Contouring Skills in Radiotherapy: The ELAISA Study". In: *JCO Global Oncology* 10 (2024).

[98] Minsong Cao et al. "Analysis of Geometric Performance and Dosimetric Impact of Using Automatic Contour Segmentation for Radiotherapy Planning". In: *Frontiers in Oncology* 10 (2020).

[99] Gregory C. Sharp et al. "Vision 20/20: perspectives on automated image segmentation for radiotherapy." In: *Medical physics* 41 5 (2014), p. 050902.

[100] Abdel Aziz Taha and Allan Hanbury. "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool". In: *BMC Medical Imaging* 15 (2015).

[101] Robert Poel et al. "The predictive value of segmentation metrics on dosimetry in organs at risk of the brain". In: *Medical Image Analysis* 73 (2021), p. 102161. ISSN: 13618423. DOI: 10.1016/j.media.2021.102161. URL: https://doi.org/10.1016/j.media.2021.102161.

[102] Chunhao Wang et al. "Artificial Intelligence in Radiotherapy Treatment Planning: Present and Future". In: *Technology in Cancer Research & Treatment* 18 (2019).

[103] Maria Rago et al. "Evaluation of a generalized knowledge-based planning performance for VMAT irradiation of breast and locoregional lymph nodes—Internal mammary and/or supraclavicular regions". In: *PLoS ONE* 16 (2021).

[104] Xinyi Li et al. "An Artificial Intelligence-Driven Agent for Real-Time Head-and-Neck IMRT Plan Generation using Conditional Generative Adversarial Network (cGAN)." In: *Medical physics* (2020).

[105] Han Liu, Benjamin J. Sintay, and David B. Wiant. "A two-step treatment planning strategy incorporating knowledge-based planning for head-and-neck radiotherapy". In: *Journal of Applied Clinical Medical Physics* 24 (2023).

[106] Mingqing Wang et al. "A Review on Application of Deep Learning Algorithms in External Beam Radiotherapy Automated Treatment Planning". In: *Frontiers in Oncology* 10 (2020).

[107] Masoud Zarepisheh et al. "Automated and Clinically Optimal Treatment Planning for Cancer Radiotherapy". In: *INFORMS journal on applied analytics* 52 1 (2022), pp. 69–89.

[108] Jérôme Krayenbuehl et al. "Evaluation of an automated knowledge based treatment planning system for head and neck". In: *Radiation Oncology (London, England)* 10 (2015).

[109] Obioma Nwankwo et al. "Knowledge-based radiation therapy (KBRT) treatment planning versus planning by experts: validation of a KBRT algorithm for prostate cancer treatment planning". In: *Radiation Oncology (London, England)* 10 (2015).

[110] Reena Phurailatpam et al. "Can knowledge based treatment planning of VMAT for post-mastectomy locoregional radiotherapy involving internal mammary chain and supraclavicular fossa improve performance efficiency?" In: *Frontiers in Oncology* 13 (2023).

[111] Dawn Gintz et al. "Initial evaluation of automated treatment planning software". In: *Journal of Applied Clinical Medical Physics* 17 (2016), pp. 331–346.

[112] Huan Liu et al. "Interactive Treatment Planning in High Dose-Rate Brachytherapy for Gynecological Cancer". In: *Mathews Journal of Cancer Science* (2021).

[113] Yin Gao, Yang Kyun Park, and Xun Jia. "Human-like intelligent automatic treatment planning of head and neck cancer radiation therapy". In: *Physics in Medicine and Biology* 69 (2024).

[114] Gyanendra Bohara et al. "Using Deep Learning to Predict Beam-Tunable Pareto Optimal Dose Distribution for Intensity Modulated Radiation Therapy". In: *Medical physics* (2020).

[115] Leire Arbea et al. "Intensity-modulated radiation therapy (IMRT) vs. 3D conformal radiotherapy (3DCRT) in locally advanced rectal cancer (LARC): dosimetric comparison and clinical implications". In: *Radiation Oncology (London, England)* 5 (2010), pp. 17–17.

[116] Steve Webb. "The physical basis of IMRT and inverse planning." In: *The British journal of radiology* 76 910 (2003), pp. 678–689.

[117] George A. Sayre and Dan Ruan. "Automatic Treatment Planning with Convex Imputing". In: *Journal of Physics: Conference Series* 489 (2014), p. 012058.

[118] Stine Sofia Korreman, Jesper Grau Eriksen, and Cai Grau. "The changing role of radiation oncology professionals in a world of AI – Just jobs lost – Or a solution to the under-provision of radiotherapy?" In: *Clinical and Translational Radiation Oncology* 26 (2020), pp. 104–107.

[119] David L Craft et al. "Improved planning time and plan quality through multicriteria optimization for intensity-modulated radiotherapy". In: *International Journal of Radiation Oncology* Biology* Physics* 82.1 (2012), e83–e90.

[120] Bruno Fionda et al. "Artificial intelligence in interventional radiotherapy (brachytherapy): Enhancing patient-centered care and addressing patients' needs". In: *Clinical and Translational Radiation Oncology* 49 (2024).

[121] Noah Bice et al. "Latent space arc therapy optimization". In: *Physics in Medicine & Biology* 66 (2021).

[122] Zhen Tian et al. "Multi-GPU implementation of a VMAT treatment plan optimization algorithm." In: *Medical physics* 42 6 (2015), pp. 2841–52.

[123] Wai Tong Ng, Shao Hui Huang, and Hai-Qiang Mai. "Precision radiotherapy in nasopharyngeal carcinoma". In: *Annals of Nasopharynx Cancer* (2021).

[124] Riqiang Gao et al. "Automating High Quality RT Planning at Scale". In: *ArXiv* abs/2501.11803 (2025).

[125] Savino Cilla et al. "Personalized Treatment Planning Automation in Prostate Cancer Radiation Oncology: A Comprehensive Dosimetric Study". In: *Frontiers in Oncology* 11 (2021).

[126] Maria Antico et al. "Real-time adaptive planning method for radiotherapy treatment delivery for prostate cancer patients, based on a library of plans accounting for possible anatomy configuration changes". In: *PLoS ONE* 14 (2019).

[127] Lei Yu et al. "First implementation of full-workflow automation in radiotherapy: the All-in-One solution on rectal cancer". In: *Preprint* (2022).

[128] Kirsten van Gysen et al. "Rolling out RapidPlan: What we've learnt". In: *Journal of Medical Radiation Sciences* 67 (2020), pp. 310–317.

[129] Hwee Shin Soh et al. "Quantitative metrics for assessing IMRT plan conformity: A virtual phantom study". In: *Journal of Physics: Conference Series* 1248 (2019).

[130] Weiliang Du et al. "Quantification of beam complexity in intensity-modulated radiation therapy treatment plans." In: *Medical physics* 41 2 (2014), p. 021716.

[131] Xiang Xia et al. "An Artificial Intelligence-Based Full-Process Solution for Radiotherapy: A Proof of Concept Study on Rectal Cancer". In: *Frontiers in Oncology* 10 (2021).

[132] Jiawei Fan et al. "Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique". In: *Medical Physics* 46 (2018), pp. 370–381.

[133] Anis Ahmad et al. "Three discipline collaborative radiation therapy (3DCRT) special debate: Peer review in radiation oncology is more effective today than 20 years ago". In: *Journal of Applied Clinical Medical Physics* 21 (2020), pp. 7–13.

[134] Gerd Heilemann et al. "Ultra-fast, one-click radiotherapy treatment planning outside a treatment planning system". In: *Physics and Imaging in Radiation Oncology* 33 (2025).

[135] Luca Cozzi, Ben J.M. Heijmen, and Ludvig Paul Muren. "Advanced treatment planning strategies to enhance quality and efficiency of radiotherapy". In: *Physics and Imaging in Radiation Oncology* 11 (2019), pp. 69–70.

[136] Brandon A. Dyer et al. "Linear Accelerator-Based Radiotherapy Simulation Using On-Board Kilovoltage Cone-Beam Computed Tomography for 3-Dimensional Volumetric Planning and Rapid Treatment in the Palliative Setting". In: *Technology in Cancer Research & Treatment* 18 (2019).

[137] Zihang Qiu et al. "Online adaptive planning methods for intensity-modulated radiotherapy". In: *Physics in Medicine & Biology* 68 (2023).

[138] V. L. Wildman et al. "Recent advances in the clinical applications of machine learning in proton therapy". In: *medRxiv*. 2024.

[139] Livia Marrazzo et al. "Fully automated volumetric modulated arc therapy technique for radiation therapy of locally advanced breast cancer". In: *Radiation Oncology (London, England)* 18 (2023).

[140] Lester J Peters et al. "Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02". In: *Journal of Clinical Oncology* 28.18 (2010), pp. 2996–3001.

[141] Xinglei Shen et al. "Radiotherapy protocol deviations and clinical outcomes: A meta-analysis of cooperative group clinical trials." In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 30 34_suppl (2012), p. 181.

[142] Bruno Vieira et al. "Radiotherapy treatment scheduling: Implementing operations research into clinical practice". In: *PLoS ONE* 16 (2021).

[143] A. Anand et al. "Study design: Validation of clinical acceptability of deep-learning-based automated segmentation of organs-at-risk for head-and-neck radiotherapy treatment planning." In: *medRxiv*. 2021.

[144] Valerio Nardone et al. "The Role of Artificial Intelligence on Tumor Boards: Perspectives from Surgeons, Medical Oncologists and Radiation Oncologists". In: *Current Oncology* 31 (2024), pp. 4984–5007.

[145] Savino Cilla et al. "Template-based automation of treatment planning in advanced radiotherapy: a comprehensive dosimetric and clinical evaluation". In: *Scientific Reports* 10 (2020).

[146] Renato Bellotti et al. "Clinical utility of automatic treatment planning for proton therapy of head-and-neck cancer patients using JulianA". In: 2024.

[147] Francisco Roberto Cassetta Júnior and Felipe Orsolin Teixeira. "Artificial intelligence technology for radiation oncology understaff mitigation and cost-effective treatment planning". In: *Revista da Faculdade de Ciências Médicas de Sorocaba* (2023).

[148] Kuo Men et al. "Automated Quality Assurance of OAR Contouring for Lung Cancer Based on Segmentation With Deep Active Learning". In: *Frontiers in Oncology* 10 (2020).

[149] S W Loo et al. "Interobserver variation in parotid gland delineation: a study of its impact on intensity-modulated radiotherapy solutions with a systematic review of the literature." In: *The British journal of radiology* 85 1016 (2012), pp. 1070–7.

[150] Hasibul Hoque et al. "Clinical Use of a Commercial Artificial Intelligence-Based Software for Autocontouring in Radiation Therapy: Geometric Performance and Dosimetric Impact". In: *Cancers* 15 (2023).

[151] Jamison Brooks et al. "Knowledge-based quality assurance of a comprehensive set of organ at risk contours for head and neck radiotherapy". In: *Frontiers in Oncology* 14 (2024).

[152] Dler Khalid Ismael and Fatiheea Fatihalla Hassan. "3D-Conformal Radiation Therapy and IntensityModulated Radiation Therapy Techniques for Laryngeal Cancer Taking Parotid Glands as Organ at Risk". In: 2020.

[153] Samsara Terparia et al. "Automatic evaluation of contours in radiotherapy planning utilising conformity indices and machine learning". In: *Physics and Imaging in Radiation Oncology* 16 (2020), pp. 149–155.

[154] Philippa Lewis et al. "Structure and Processes of Existing Practice in Radiotherapy Peer Review: A Systematic Review of the Literature." In: *Clinical oncology (Royal College of Radiologists (Great Britain))* (2020).

[155] Benjamin T. Cooper et al. "Development of a Comprehensive, Contour-Based, Peer Review Workflow at a Community Proton Center". In: *International Journal of Particle Therapy* 7 (2020), pp. 34–40.

[156] Ryan T. Hughes et al. "Head and neck radiotherapy quality assurance conference for dedicated review of delineated targets and organs at risk: results of a prospective study". In: *Journal of radiotherapy in practice* 22 (2022).

[157] Helen S. Zhang et al. "Radiotherapy contour quality improvement practices among United States radiation oncologists." In: *Journal of Clinical Oncology* (2021).

[158] Suliana Teoh et al. "Evaluation of hypofractionated adaptive radiotherapy using the MR Linac in localised pancreatic cancer: protocol summary of the Emerald-Pancreas phase 1/expansion study located at Oxford University Hospital, UK". In: *BMJ Open* 13 (2023).

[159] Luke Nicholls et al. "Maintaining prostate contouring consistency following an educational intervention". In: *Journal of Medical Radiation Sciences* 63 (2016), pp. 155–160.

[160] Amy Tien Yee Chang et al. "Challenges for Quality Assurance of Target Volume Delineation in Clinical Trials". In: *Frontiers in Oncology* 7 (2017).

[161] Xiaodong Li et al. "Evaluation of deformable image registration for contour propagation between CT and cone-beam CT images in adaptive head and neck radiotherapy." In: *Technology and health care : official journal of the European Society for Engineering and Medicine* 24 Suppl 2 (2016), S747–55.

[162] Barbara Marquez et al. "Analyzing the Relationship between Dose and Geometric Agreement Metrics for Auto-Contouring in Head and Neck Normal Tissues". In: *Diagnostics* 14 (2024).

[163] Hanne Nijhuis et al. "Investigating the potential of deep learning for patient-specific quality assurance of salivary gland contours using EORTC-1219-DAHANCA-29 clinical trial data". In: *Acta Oncologica* 60 (2021), pp. 575–581.

[164] Nikolett Buciuman and Loredana Gabriela Marcu. "Dosimetric justification for the use of volumetric modulated arc therapy in head and neck cancer—A systematic review of the literature". In: *Laryngoscope Investigative Otolaryngology* 6 (2021), pp. 999–1007.

[165] Katherine Mackay et al. "A review of the metrics used to assess auto-contouring systems in radiotherapy". In: *Clinical Oncology* 35.6 (2023), pp. 354–369.

[166] Hsin-Chen Chen et al. "Automated contouring error detection based on supervised geometric attribute distribution models for radiation therapy: a general strategy." In: *Medical physics* 42 2 (2015), pp. 1048–59.

[167] Dong Joo Rhee et al. "Automatic contouring QA method using a deep learning–based autocontouring system". In: *Journal of Applied Clinical Medical Physics* 23 (2022).

[168] Dong Joo Rhee et al. "Automatic detection of contouring errors using convolutional neural networks". In: *Medical Physics* 46 (2019), pp. 5086–5097.

[169] Shunyao Luan et al. "Machine Learning-Based Quality Assurance for Automatic Segmentation of Head-and-Neck Organs-at-Risk in Radiotherapy". In: *Technology in Cancer Research & Treatment* 22 (2023).

[170] Michaël Claessens et al. "Machine learning-based detection of aberrant deep learning segmentations of target and organs at risk for prostate radiotherapy using a secondary segmentation algorithm". In: *Physics in Medicine & Biology* 67.11 (2022), p. 115014.

[171] J. John Lucido et al. "Validation of clinical acceptability of deep-learning-based automated segmentation of organs-at-risk for head-and-neck radiotherapy treatment planning". In: *Frontiers in Oncology* 13 (2023).

[172] Shunyao Luan et al. "A multi-modal vision-language pipeline strategy for contour quality assurance and adaptive optimization". In: *Physics in Medicine & Biology* 69 (2024).

[173] Tucker J. Netherton et al. "External validation of an algorithm to detect vertebral level mislabeling and autocontouring errors". In: *Physics and Imaging in Radiation Oncology* 34 (2025).

[174] Jingwei Duan et al. "Contouring quality assurance methodology based on multiple geometric features against deep learning auto-segmentation." In: *Medical physics* (2023).

[175] Jordan Wong et al. "Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: a workflow study at two cancer centers". In: *Radiation Oncology (London, England)* 16 (2021).

[176] Julie van der Veen et al. "Interobserver variability in organ at risk delineation in head and neck cancer". In: *Radiation Oncology (London, England)* 16 (2020).

[177] Shuolin Liu et al. "Technical Note: A cascade 3D U-Net for dose prediction in radiotherapy". In: *Medical Physics* 48.9 (Sept. 2021), pp. 5574–5582. ISSN: 24734209. DOI: 10.1002/mp.15034.

[178] Reza Azad et al. "Medical Image Segmentation Review: The Success of U-Net". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46 (2022), pp. 10076–10095.

[179] Kathryn E. Mittauer et al. "STAT-ART: The Promise and Practice of a Rapid Palliative Single Session of MR-Guided Online Adaptive Radiotherapy (ART)". In: *Frontiers in Oncology* 9 (2019).

[180] Melek Cosar Yakar and Durmuş Etiz. "Artificial intelligence in radiation oncology". In: *Artificial Intelligence in Medical Imaging* (2021).

[181] Mohammad Hussein et al. "Automation in intensity modulated radiotherapy treatment planning-a review of recent innovations." In: *The British journal of radiology* 91 1092 (2018), p. 20180270.

[182] Oscar Pastor-Serrano and Zolt'an Perk'o. "Learning the Physics of Particle Transport via Transformers". In: *AAAI Conference on Artificial Intelligence*. 2021.

[183] Ralf Müller-Polyzou et al. "Assistance systems for patient positioning in radiotherapy practice". In: *Health Systems* 13 (2024), pp. 332–360.

[184] Joshua P. Schiff et al. "Simulation-Free Radiation Therapy: An Emerging Form of Treatment Planning to Expedite Plan Generation for Patients Receiving Palliative Radiation Therapy". In: *Advances in Radiation Oncology* 8 (2022).

[185] Douglas Moura Miranda and Mariana Pedrini Moura Miranda. "Discrete-event simulation applied to a radiotherapy process: a case study of a cancer center". In: *Brazilian journal of operations & production management* 18 (2021).

[186] Daniel Moore-Palhares et al. “Clinical implementation of magnetic resonance imaging simulation for radiation oncology planning: 5 year experience”. In: *Radiation Oncology (London, England)* 18 (2023).

[187] Luise A. Künzel and Daniela Thorwarth. “Towards real-time radiotherapy planning: The role of autonomous treatment strategies”. In: *Physics and Imaging in Radiation Oncology* 24 (2022), pp. 136–137.

[188] Gisèle Pereira, Melanie Traughber, and Raymond F. Muzic. “The Role of Imaging in Radiation Therapy Planning: Past, Present, and Future”. In: *BioMed Research International* 2014 (2014).

[189] Claudio Votta et al. “Evaluation of clinical parallel workflow in online adaptive MR-guided Radiotherapy: A detailed assessment of treatment session times”. In: *Technical Innovations & Patient Support in Radiation Oncology* 29 (2024).

[190] Shinichiro Mori et al. “Patient handling system for carbon ion beam scanning therapy”. In: *Journal of Applied Clinical Medical Physics* 13 (2012), pp. 226–240.

[191] Anet Aselmaa et al. “Using a contextualized sensemaking model for interaction design: A case study of tumor contouring”. In: *Journal of biomedical informatics* 65 (2017), pp. 145–158.

[192] Søren M Bentzen et al. “Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC): an introduction to the scientific issues”. In: *International Journal of Radiation Oncology* Biology* Physics* 76.3 (2010), S3–S9.

[193] EORTC DAHANCA et al. “CT-based delineation of organs at risk in the head and neck region”. In: *Radiotherapy and Oncology* (2015), p. 1.

[194] Laurence E. Court et al. “Radiation Planning Assistant - A Streamlined, Fully Automated Radiotherapy Treatment Planning System”. In: *Journal of Visualized Experiments : JoVE* (2018).

[195] Laurence E. Court et al. “Addressing the Global Expertise Gap in Radiation Oncology: The Radiation Planning Assistant”. In: *JCO Global Oncology* 9 (2023).

[196] Laurence E. Court et al. “Artificial Intelligence–Based Radiotherapy Contouring and Planning to Improve Global Access to Cancer Care”. In: *JCO Global Oncology* 10 (2024).

[197] Eric Aliotta et al. “An Automated Workflow to Improve Efficiency in Radiation Therapy Treatment Planning by Prioritizing Organs at Risk”. In: *Advances in Radiation Oncology* 5 (2020), pp. 1324–1333.

[198] Nicolas F. Chaves-de-Plaza et al. “Report on AI-Infused Contouring Workflows for Adaptive Proton Therapy in the Head and Neck”. In: 2022.

[199] F. A. Siebert et al. “Errors detected during physics plan review for external beam radiotherapy”. In: *Physics and Imaging in Radiation Oncology* 24 (2022), pp. 53–58.

[200] R. C. Fernandes et al. “A rule-based system proposal to aid in the evaluation and decision-making in external beam radiation treatment planning”. In: *ArXiv* abs/1811.12454 (2018).

[201] Reshma Munbodh et al. “Real-time analysis and display of quantitative measures to track and improve clinical workflow”. In: *Journal of Applied Clinical Medical Physics* 23 (2022).

[202] Byungchul Cho. "Intensity-modulated radiation therapy: a review with a physics perspective". In: *Radiation oncology journal* 36.1 (2018), p. 1.

[203] Gary A Ezzell et al. "Guidance document on delivery, treatment planning, and clinical implementation of IMRT: report of the IMRT Subcommittee of the AAPM Radiation Therapy Committee". In: *Medical physics* 30.8 (2003), pp. 2089–2115.

[204] Thomas LoSasso. "IMRT delivery performance with a varian multileaf collimator". In: *International Journal of Radiation Oncology* Biology* Physics* 71.1 (2008), S85–S88.

[205] Christopher M Nutting et al. "Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial". In: *The lancet oncology* 12.2 (2011), pp. 127–136.

[206] Michael J Zelefsky et al. "High-dose intensity modulated radiation therapy for prostate cancer: early toxicity and biochemical outcome in 772 patients". In: *International Journal of Radiation Oncology* Biology* Physics* 53.5 (2002), pp. 1111–1116.

[207] Karl Otto. "Volumetric modulated arc therapy: IMRT in a single gantry arc". In: *Medical physics* 35.1 (2008), pp. 310–317.

[208] Enzhuo M Quan et al. "A comprehensive comparison of IMRT and VMAT plan quality for prostate cancer treatment". In: *International Journal of Radiation Oncology* Biology* Physics* 83.4 (2012), pp. 1169–1178.

[209] David Palma et al. "Volumetric modulated arc therapy for delivery of prostate radiotherapy: comparison with intensity-modulated radiotherapy and three-dimensional conformal radiotherapy". In: *International Journal of Radiation Oncology* Biology* Physics* 72.4 (2008), pp. 996–1001.

[210] Dylan Callens et al. "Are offline ART decisions for NSCLC impacted by the type of dose calculation algorithm?" In: *Technical Innovations & Patient Support in Radiation Oncology* 29 (2024).

[211] James F. Dempsey et al. "A fourier analysis of the dose grid resolution required for accurate IMRT fluence map optimization." In: *Medical physics* 32 2 (2005), pp. 380–8.

[212] Kotaro Iijima et al. "Analysis of human errors in the operation of various treatment planning systems over a 10-year period". In: *Journal of Radiation Research* 65 (2024), pp. 603–618.

[213] Zilong Yuan et al. "Converting Treatment Plans From Helical Tomotherapy to L-Shape Linac: Clinical Workflow and Dosimetric Evaluation". In: *Technology in Cancer Research & Treatment* 17 (2018).

[214] Kelly A. Nealon et al. "Hazard testing to reduce risk in the development of automated planning tools". In: *Journal of Applied Clinical Medical Physics* 24 (2023).

[215] Ping Yan et al. "Design and implementation of automated notification systems and an electronic whiteboard for radiation therapy planning monitoring". In: *Journal of Applied Clinical Medical Physics* 25 (2024).

[216] Varun Kumar Chowdhry and Natalie E. Simpson. "Process Modeling a Radiation Oncology Clinic Workflow From Therapeutic Simulation to Treatment: Identifying Impending Strain and Possible Treatment Delays". In: *Advances in Radiation Oncology* 8 (2022).

[217] James Lamb et al. "Online Adaptive Radiation Therapy: Implementation of a New Process of Care". In: *Cureus* 9 (2017).

[218] Xinyuan Chen et al. "CNN-Based Quality Assurance for Automatic Segmentation of Breast Cancer in Radiotherapy". In: *Frontiers in Oncology* 10 (2020).

[219] Haoyu Zhong et al. "The Impact of Clinical Trial Quality Assurance on Outcome in Head and Neck Radiotherapy Treatment". In: *Frontiers in Oncology* 9 (2019).

[220] Yatman Tsang et al. "Assessment of contour variability in target volumes and organs at risk in lung cancer radiotherapy". In: *Technical Innovations & Patient Support in Radiation Oncology* 10 (2019), pp. 8–12.

[221] Yatman Tsang et al. "Clinical impact of IMPORT HIGH trial (CRUK/06/003) on breast radiotherapy practices in the United Kingdom." In: *The British journal of radiology* 88 1056 (2015), p. 20150453.

[222] Sarah Gwynne et al. "Improving radiotherapy quality assurance in clinical trials: assessment of target volume delineation of the pre-accrual benchmark case." In: *The British journal of radiology* 86 1024 (2013), p. 20120398.

[223] Xiyao Jin et al. "A quality assurance framework for real-time monitoring of deep learning segmentation models in radiotherapy". In: *ArXiv* abs/2305.11715 (2023).

[224] Kelly D. Kisling et al. "A risk assessment of automated treatment planning and recommendations for clinical deployment". In: *Medical Physics* 46 (2019), pp. 2567–2574.

[225] Tatiana Dragan et al. "Enhanced head and neck radiotherapy target definition through multidisciplinary delineation and peer review: A prospective single-center study". In: *Clinical and Translational Radiation Oncology* 48 (2024).

[226] Veeraj Shah et al. "Data integrity systems for organ contours in radiation therapy planning". In: *Journal of Applied Clinical Medical Physics* 19 (2018), pp. 58–67.

[227] Alexander F. I. Osman. "Radiation Oncology in the Era of Big Data and Machine Learning for Precision Medicine". In: *Artificial Intelligence - Applications in Medicine and Biology* (2019).

[228] Bruno Fionda et al. "Artificial intelligence (AI) and interventional radiotherapy (brachytherapy): state of art and future perspectives". In: *Journal of Contemporary Brachytherapy* 12 (2020), pp. 497–500.

[229] Kendall J. Kiser, Clifton David Fuller, and Valerie Klairisa Reed. "Artificial intelligence in radiation oncology treatment planning: a brief overview". In: *Journal of Medical Artificial Intelligence* (2019).

[230] Mariko Kawamura et al. "Revolutionizing radiation therapy: the role of AI in clinical practice". In: *Journal of Radiation Research* 65 (2023), pp. 1–9.

[231] John McCarthy et al. "A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955". In: *AI magazine* 27.4 (2006), pp. 12–12.

[232] Tom Michael Mitchell. *The discipline of machine learning*. Vol. 9. Carnegie Mellon University, School of Computer Science, Machine Learning . . ., 2006.

[233] Amirhosein Toosi et al. "A brief history of AI: how to prevent another winter (a critical review)". In: *PET clinics* 16.4 (2021), pp. 449–469.

[234] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. pearson, 2016.

[235] Hubert L Dreyfus. "A critique of artificial reason". In: *Thought: Fordham University Quarterly* 43.4 (1968), pp. 507–522.

[236] Vasant Dhar. "The paradigm shifts in artificial intelligence". In: *Communications of the ACM* 67.11 (2024), pp. 50–59.

[237] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[238] Elizabeth Huynh et al. "Artificial intelligence in radiation oncology". In: *Nature Reviews Clinical Oncology* 17.12 (2020), pp. 771–781.

[239] Yihao Liu and Minghua Wu. "Deep learning in precision medicine and focus on glioma". In: *Bioengineering & Translational Medicine* 8.5 (2023), e10553.

[240] Yi Luo, Shifeng Chen, and Gilmer Valdes. "Machine learning for radiation outcome modeling and prediction". In: *Medical Physics* 47.5 (2020), e178–e184.

[241] Teuvo Kohonen. "The self-organizing map". In: *Proceedings of the IEEE* 78.9 (1990), pp. 1464–1480.

[242] Rui Xu and Donald Wunsch. "Survey of clustering algorithms". In: *IEEE Transactions on neural networks* 16.3 (2005), pp. 645–678.

[243] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[244] Peng Huang et al. "Anomaly detection in radiotherapy plans using deep autoencoder networks". In: *Frontiers in Oncology* 13 (2023), p. 1142947.

[245] Veronika Cheplygina, Marleen De Bruijne, and Josien PW Pluim. "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis". In: *Medical image analysis* 54 (2019), pp. 280–296.

[246] Liang Chen et al. "Self-supervised learning for medical image analysis using image context restoration". In: *Medical image analysis* 58 (2019), p. 101539.

[247] Ting Chen et al. "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PmLR. 2020, pp. 1597–1607.

[248] Mathilde Caron et al. "Emerging properties in self-supervised vision transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–9660.

[249] Kaiming He et al. "Masked autoencoders are scalable vision learners". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16000–16009.

[250] Antti Tarvainen and Harri Valpola. "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results". In: *Advances in neural information processing systems* 30 (2017).

[251] Kihyuk Sohn et al. "Fixmatch: Simplifying semi-supervised learning with consistency and confidence". In: *Advances in neural information processing systems* 33 (2020), pp. 596–608.

[252] Yang Jiang et al. "Expert feature-engineering vs. deep neural networks: which is better for sensor-free affect detection?" In: *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part I 19*. Springer. 2018, pp. 198–211.

[253] Dinggang Shen, Guorong Wu, and Heung-Il Suk. "Deep learning in medical image analysis". In: *Annual review of biomedical engineering* 19.1 (2017), pp. 221–248.

[254] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[255] Ian Goodfellow et al. *Deep learning*. Vol. 1. 2. MIT press Cambridge, 2016.

[256] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[257] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature machine intelligence* 1.5 (2019), pp. 206–215.

[258] Jürgen Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural networks* 61 (2015), pp. 85–117.

[259] Michael I Jordan and Tom M Mitchell. "Machine learning: Trends, perspectives, and prospects". In: *Science* 349.6245 (2015), pp. 255–260.

[260] Catalin Ionescu, Orestis Vantzos, and Cristian Sminchisescu. "Training Deep Networks with Structured Layers by Matrix Backpropagation". In: *ArXiv* abs/1509.07838 (2015).

[261] Lucas Mohimont et al. "Computer Vision and Deep Learning for Precision Viticulture". In: *Agronomy* (2022).

[262] Gorana Gojić et al. "Comparing the Clinical Viability of Automated Fundus Image Segmentation Methods". In: *Sensors (Basel, Switzerland)* 22 (2022).

[263] Sharib Ali et al. "An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy". In: *Scientific Reports* 10 (2020).

[264] Matthew D. Zeiler and Rob Fergus. "Visualizing and Understanding Convolutional Networks". In: *ArXiv* abs/1311.2901 (2013).

[265] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[266] André Littek et al. "Automatic Segmentation of Osteonal Microstructure in Human Cortical Bone Using Deep Learning: A Proof of Concept". In: *Biology* 12 (2023).

[267] Dehui Xiong et al. "An End-To-End Bayesian Segmentation Network Based on a Generative Adversarial Network for Remote Sensing Images". In: *Remote. Sens.* 12 (2020), p. 216.

[268] Christian Szegedy et al. "Going deeper with convolutions". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 1–9.

[269] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 770–778.

[270] Evan Shelhamer, Jonathan Long, and Trevor Darrell. “Fully convolutional networks for semantic segmentation”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 3431–3440.

[271] Przemyslaw Polewski et al. “Segmenting objects with Bayesian fusion of active contour models and convnet priors”. In: *ArXiv* abs/2410.07421 (2024).

[272] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (2015), pp. 2481–2495.

[273] Liang-Chieh Chen et al. “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (2016), pp. 834–848.

[274] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. “Learning Deconvolution Network for Semantic Segmentation”. In: *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 1520–1528.

[275] Chao Peng et al. “Large Kernel Matters — Improve Semantic Segmentation by Global Convolutional Network”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 1743–1751.

[276] Kaiming He et al. “Mask r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[277] Min Bai and Raquel Urtasun. “Deep Watershed Transform for Instance Segmentation”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 2858–2866.

[278] Jo Schlemper et al. “Attention gated networks: Learning to leverage salient regions in medical images”. In: *Medical image analysis* 53 (2019), pp. 197–207.

[279] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[280] Enze Xie et al. “SegFormer: Simple and efficient design for semantic segmentation with transformers”. In: *Advances in neural information processing systems* 34 (2021), pp. 12077–12090.

[281] Jieneng Chen et al. “Transunet: Transformers make strong encoders for medical image segmentation”. In: *arXiv preprint arXiv:2102.04306* (2021).

[282] Ali Hatamizadeh et al. “Unetr: Transformers for 3d medical image segmentation”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 574–584.

[283] Hu Cao et al. “Swin-unet: Unet-like pure transformer for medical image segmentation”. In: *European conference on computer vision*. Springer. 2022, pp. 205–218.

[284] Yehao Li et al. “Contextual transformer networks for visual recognition”. In: *IEEE transactions on pattern analysis and machine intelligence* 45.2 (2022), pp. 1489–1500.

[285] Fabian Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature methods* 18.2 (2021), pp. 203–211.

[286] Fausto Milletarì, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation". In: *2016 Fourth International Conference on 3D Vision (3DV)* (2016), pp. 565–571.

[287] Haneen Alokasi and Muhammad Bilal Ahmad. "Deep Learning-Based Frameworks for Semantic Segmentation of Road Scenes". In: *Electronics* (2022).

[288] Zongwei Zhou et al. "UNet++: A Nested U-Net Architecture for Medical Image Segmentation". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support : 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain* 11045 (2018), pp. 3–11.

[289] Zhuangzhuang Zhang et al. "Weaving Attention U-net: A Novel Hybrid CNN and Attention-based Method for Organs-at-risk Segmentation in Head and Neck CT Images". In: *Medical physics* (2021).

[290] Özgün Çiçek et al. "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2016.

[291] Shahzad Ali et al. "Edge-Preserving Probabilistic Downsampling for Reliable Medical Segmentation in Resource-Constrained Environments". In: *IEEE Access* 13 (2025), pp. 21620–21634.

[292] Mengyu Liu and Hujun Yin. "Feature Pyramid Encoding Network for Real-time Semantic Segmentation". In: *British Machine Vision Conference*. 2019.

[293] Justin Roper, Mu-Han Lin, and Yi Rong. "Extensive upfront validation and testing are needed prior to the clinical implementation of AI-based auto-segmentation tools". In: *Journal of Applied Clinical Medical Physics* 24 (2022).

[294] Jinzhong Yang et al. "Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017". In: *Medical Physics* 45 (2018), pp. 4568–4581.

[295] Luis A Maduro Bustos et al. "Feasibility evaluation of novel AI-based deep-learning contouring algorithm for radiotherapy". In: *Journal of Applied Clinical Medical Physics* 24 (2023).

[296] Gerd Heilemann et al. "Clinical Implementation and Evaluation of Auto-Segmentation Tools for Multi-Site Contouring in Radiotherapy". In: *Physics and Imaging in Radiation Oncology* 28 (2023).

[297] Jin-qiang You et al. "Deep Learning-Aided Automatic Contouring of Clinical Target Volumes for Radiotherapy in Breast Cancer After Modified Radical Mastectomy". In: *Frontiers of Physics*. 2022.

[298] Abigayle C. Kraus et al. "Prospective Evaluation of Automated Contouring for CT-Based Brachytherapy for Gynecologic Malignancies". In: *Advances in Radiation Oncology* 9 (2023).

[299] Tingyu Wang et al. "Evaluation of AI-based auto-contouring tools in radiotherapy: A single-institution study". In: *Journal of Applied Clinical Medical Physics* 26 (2025).

[300] Tomaž Vrtovec et al. "Auto-segmentation of organs at risk for head and neck radiotherapy planning: from atlas-based to deep learning methods." In: *Medical physics* (2020).

[301] B W K Schipaanboord et al. "An Evaluation of Atlas Selection Methods for Atlas-Based Automatic Segmentation in Radiotherapy Treatment Planning". In: *IEEE Transactions on Medical Imaging* 38 (2019), pp. 2654–2664.

[302] Ian Boon, Tracy Au Yong, and Cheng Boon. "Assessing the Role of Artificial Intelligence (AI) in Clinical Oncology: Utility of Machine Learning in Radiotherapy Target Volume Delineation". In: *Medicines* 5.4 (Dec. 2018), p. 131. ISSN: 2305-6320. DOI: 10.3390/medicines5040131.

[303] Kareem A. Wahid et al. "Evolving Horizons in Radiation Therapy Auto-Contouring: Distilling Insights, Embracing Data-Centric Frameworks, and Moving Beyond Geometric Quantification". In: *Advances in Radiation Oncology* 9 (2023).

[304] Oussama H. Hamid. "From Model-Centric to Data-Centric AI: A Paradigm Shift or Rather a Complementary Approach?" In: *2022 8th International Conference on Information Technology Trends (ITT)* (2022), pp. 196–199.

[305] Stanislav Nikolov et al. "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy". In: *ArXiv* abs/1809.04430 (2018).

[306] Tassilo Wald et al. "Primus: Enforcing attention usage for 3d medical image segmentation". In: *arXiv preprint arXiv:2503.01835* (2025).

[307] Fabian Isensee et al. "nninteractive: Redefining 3d promptable segmentation". In: *arXiv preprint arXiv:2503.08373* (2025).

[308] Bjoern H Menze et al. "The multimodal brain tumor image segmentation benchmark (BRATS)". In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024.

[309] Spyridon Bakas et al. "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge". In: *arXiv preprint arXiv:1811.02629* (2018).

[310] Philipp Kickingereder et al. "Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study". In: *The Lancet Oncology* 20.5 (2019), pp. 728–740.

[311] Sogand Sadeghi, Mostafa Farzin, and Somayeh Gholami. "Fully automated clinical target volume segmentation for glioblastoma radiotherapy using a deep convolutional neural network". In: *Polish Journal of Radiology* 88 (2023), e31–e40.

[312] Curtise Kin Cheung Ng, Vincent W. S. Leung, and Rico Hing-Ming Hung. "Clinical Evaluation of Deep Learning and Atlas-Based Auto-Contouring for Head and Neck Radiation Therapy". In: *Applied Sciences* (2022).

[313] Jin-yuan Wang et al. "Evaluation Exploration of Atlas-Based and Deep Learning-Based Automatic Contouring for Nasopharyngeal Carcinoma". In: *Frontiers in Oncology* 12 (2022).

[314] Yimin Li et al. "Evaluating Automatic Segmentation for Swallowing-Related Organs for Head and Neck Cancer". In: *Technology in Cancer Research & Treatment* 21 (2022).

[315] Nan Bi et al. "Deep learning improved clinical target volume contouring quality and efficiency for postoperative radiation therapy in non-small cell lung cancer". In: *Frontiers in oncology* 9 (2019), p. 1192.

[316] Wen Chen et al. "Deep learning vs. atlas-based models for fast auto-segmentation of the masticatory muscles on head and neck CT images". In: *Radiation Oncology (London, England)* 15 (2020).

[317] Seongmin Choi et al. "Automated Organ Segmentation for Radiation Therapy: A Comparative Analysis of AI-Based Tools Versus Manual Contouring in Korean Cancer Patients". In: *Cancers* 16 (2024).

[318] Céline Meyer et al. "Artificial intelligence contouring in radiotherapy for organs-at-risk and lymph node areas". In: *Radiation Oncology (London, England)* 19 (2024).

[319] Young Woo Kim, Simon Biggs, and Elizabeth Ruth Claridge Mackonis. "Investigation on performance of multiple AI-based auto-contouring systems in organs at risks (OARs) delineation". In: *Physical and Engineering Sciences in Medicine* 47 (2024), pp. 1123–1140.

[320] Zi-Hang Chen et al. "Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine". In: *Cancer Communications* 41 (2021), pp. 1100–1115.

[321] Li Lin et al. "Deep Learning for Automated Contouring of Primary Tumor Volumes by MRI for Nasopharyngeal Carcinoma." In: *Radiology* 291 3 (2019), pp. 677–686.

[322] Nan Bi et al. "Deep Learning Improved Clinical Target Volume Contouring Quality and Efficiency for Postoperative Radiation Therapy in Non-small Cell Lung Cancer". In: *Frontiers in Oncology* 9 (2019).

[323] Ziming Han et al. "Artificial intelligence-assisted delineation for postoperative radiotherapy in patients with lung cancer: a prospective, multi-center, cohort study". In: *Frontiers in Oncology* 14 (2024).

[324] Chen Ma et al. "Deep learning-based auto-segmentation of clinical target volumes for radiotherapy treatment of cervical cancer". In: *Journal of Applied Clinical Medical Physics* 23 (2021).

[325] Sumeet Hindocha et al. "Artificial Intelligence for Radiotherapy Auto-Contouring: Current Use, Perceptions of and Barriers to Implementation." In: *Clinical oncology (Royal College of Radiologists (Great Britain))* (2023).

[326] Ti Bai et al. "A Proof-of-Concept Study of Artificial Intelligence Assisted Contour Revision". In: *ArXiv* abs/2107.13465 (2021).

[327] Jacob Adams et al. "Plan Quality Analysis of Automated Treatment Planning Workflow With Commercial Auto-Segmentation Tools and Clinical Knowledge-Based Planning Models for Prostate Cancer". In: *Cureus* 15 (2023).

[328] Lee C Goddard et al. "Evaluation of multiple-vendor AI autocontouring solutions". In: *Radiation Oncology (London, England)* 19 (2024).

[329] Yang Zhong et al. "A Preliminary Experience of Implementing Deep-Learning Based Auto-Segmentation in Head and Neck Cancer: A Study on Real-World Clinical Cases". In: *Frontiers in Oncology* 11 (2021).

[330] Zi Chen et al. "PAM: A Propagation-Based Model for Segmenting Any 3D Objects across Multi-Modal Medical Images". In: 2024.

[331] Nandan Maruti Shanbhag et al. "Integrating Artificial Intelligence Into Radiation Oncology: Can Humans Spot AI?" In: *Cureus* 15 (2023).

[332] Kevin Yingyin Zou et al. "Towards Reliable Medical Image Segmentation by utilizing Evidential Calibrated Uncertainty". In: 2023.

[333] Hashmat Shadab Malik et al. "On Evaluating Adversarial Robustness of Volumetric Medical Segmentation Models". In: *ArXiv* abs/2406.08486 (2024).

[334] Jia Wu et al. "Radiological tumor classification across imaging modality and histology". In: *Nature machine intelligence* 3 (2021), pp. 787–798.

[335] Yannick Suter et al. "Radiomics for glioblastoma survival analysis in pre-operative MRI: exploring feature robustness, class boundaries, and machine learning techniques". In: *Cancer Imaging* 20 (2020), pp. 1–13.

[336] Congzhen Shi et al. "A Survey on Trustworthiness in Foundation Models for Medical Image Analysis". In: *ArXiv* abs/2407.15851 (2024).

[337] Laura Alexandra Daza, Juan C. P'erez, and Pablo Arbel'aez. "Towards Robust General Medical Image Segmentation". In: *ArXiv* abs/2107.04263 (2021).

[338] Zhuo Kuang et al. "ROXSI: Robust Cross-Sequence Semantic Interaction for Brain Tumor Segmentation on Multi-Sequence MR Images". In: *IEEE Journal of Biomedical and Health Informatics* (2024).

[339] Suhang You and Mauricio Reyes. "Influence of contrast and texture based image modifications on the performance and attention shift of U-Net models for brain tissue segmentation". In: *Frontiers in neuroimaging* 1 (2022), p. 1012639.

[340] Farhad Maleki et al. "RIDGE: Reproducibility, Integrity, Dependability, Generalizability, and Efficiency Assessment of Medical Image Segmentation Models". In: *Journal of imaging informatics in medicine* (2024).

[341] Hayit Greenspan, Bram van Ginneken, and Ronald M. Summers. "Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique". In: *IEEE Trans. Medical Imaging* 35 (2016), pp. 1153–1159.

[342] Chenyu You et al. "Rethinking Semi-Supervised Medical Image Segmentation: A Variance-Reduction Perspective". In: *Advances in neural information processing systems* 36 (2023), pp. 9984–10021.

[343] Jaemoon Hwang and Sangheum Hwang. "Exploiting Global Structure Information to Improve Medical Image Segmentation". In: *Sensors (Basel, Switzerland)* 21 (2021).

[344] Chen Chen et al. "Cooperative Training and Latent Space Data Augmentation for Robust Medical Image Segmentation". In: *ArXiv* abs/2107.01079 (2021).

[345] Seoin Chai, Daniel Rueckert, and Ahmed E. Fetit. "Reducing Textural Bias Improves Robustness of Deep Segmentation Models". In: *Annual Conference on Medical Image Understanding and Analysis*. 2020.

[346] Mohammad Almasganj and Emad Fatemizadeh. "RHLS: A Robust Hybrid Level Set Model Using Global-Local Signed Energy-Based Pressure Force for Medical Image Segmentation". In: *IEEE Access* 13 (2025), pp. 2004–2017.

[347] Kuanquan Wang and Chao Ma. "A robust statistics driven volume-scalable active contour for segmenting anatomical structures in volumetric medical images with complex conditions". In: *BioMedical Engineering OnLine* 15 (2016).

[348] Yichi Zhang et al. "Enhancing the Reliability of Segment Anything Model for Auto-Prompting Medical Image Segmentation with Uncertainty Rectification". In: 2023.

[349] Yichi Zhang, Zhenrong Shen, and Rushi Jiao. "Segment Anything Model for Medical Image Segmentation: Current Applications and Future Directions". In: *Computers in biology and medicine* 171 (2024), p. 108238.

[350] Jialin Shi and Ji Wu. "Distilling effective supervision for robust medical image segmentation with noisy labels". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2021.

[351] Zhi Li et al. "Adaptive Interactive Segmentation for Multimodal Medical Imaging via Selection Engine". In: *ArXiv* abs/2411.19447 (2024).

[352] Kecheng Chen et al. "Learning Robust Shape Regularization for Generalizable Medical Image Segmentation". In: *IEEE Transactions on Medical Imaging* 43 (2024), pp. 2693–2706.

[353] Othmane Laousy et al. "Certification of Deep Learning Models for Medical Image Segmentation". In: *ArXiv* abs/2310.03664 (2023).

[354] Yingwei Li et al. "Volumetric Medical Image Segmentation: A 3D Deep Coarse-to-Fine Framework and Its Adversarial Examples". In: *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*. 2020.

[355] Anish Athalye, Nicholas Carlini, and David A. Wagner. "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples". In: *International Conference on Machine Learning*. 2018.

[356] Wenqi Zhao. "Robust image segmentation model based on binary level set". In: *ArXiv* abs/2403.13392 (2024).

[357] Ke Zou et al. "TBraTS: Trusted Brain Tumor Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022.

[358] Yichi Zhang et al. "Uncertainty-Guided Mutual Consistency Learning for Semi-Supervised Medical Image Segmentation". In: *Artificial intelligence in medicine* 138 (2021), p. 102476.

[359] Alain Jungo and Mauricio Reyes. "Assessing reliability and challenges of uncertainty estimations for medical image segmentation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer. 2019, pp. 48–56.

[360] Ainkaran Santhirasekaram et al. "Vector Quantisation for Robust Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022.

[361] Thi Hoang Ngan Le. "On the Out of Distribution Robustness of Foundation Models in Medical Image Segmentation". In: *Advances in Neural Information Processing Systems (NeurIPS), Workshop on robustness of zero/few-shot learning in foundation models*. NIPS. 2023.

[362] Ali S Tejani et al. "Updating the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) for reporting AI research". In: *Nature Machine Intelligence* 5.9 (2023), pp. 950–951.

[363] Seif Mzoughi, Mohamed Elshafeia, and Foutse Khomh. "Evaluating and Enhancing Segmentation Model Robustness with Metamorphic Testing". In: 2025.

[364] Giulio Rossolini et al. "On the Real-World Adversarial Robustness of Real-Time Semantic Segmentation Models for Autonomous Driving". In: *IEEE Transactions on Neural Networks and Learning Systems* 35 (2022), pp. 18328–18342.

[365] Yang Sheng et al. "Artificial intelligence applications in intensity modulated radiation treatment planning: an overview." In: *Quantitative imaging in medicine and surgery* 11 12 (2021), pp. 4859–4880.

[366] Michele Avanzo et al. "Artificial Intelligence and the Medical Physicist: Welcome to the Machine". In: *Applied Sciences* (2021).

[367] Du Wang et al. "Radiotherapy Plan Quality Assurance in NRG Oncology Trials for Brain and Head/Neck Cancers: An AI-Enhanced Knowledge-Based Approach". In: *Cancers* 16 (2024).

[368] Wentao Wang et al. "Fluence Map Prediction Using Deep Learning Models – Direct Plan Generation for Pancreas Stereotactic Body Radiation Therapy". In: *Frontiers in Artificial Intelligence* 3 (2020).

[369] Humza Nusrat et al. "Autonomous Radiotherapy Treatment Planning Using DOLA: A Privacy-Preserving, LLM-Based Optimization Agent". In: *ArXiv* abs/2503.17553 (2025).

[370] Vincent A Weidlich and Georg Weidlich. "Artificial Intelligence in Medicine and Radiation Oncology". In: *Cureus* 10 (2018).

[371] Tufia C. Haddad et al. "Impact of a cognitive computing clinical trial matching system in an ambulatory oncology practice." In: *Journal of Clinical Oncology* 36 (2018), pp. 6550–6550.

[372] Robin De Roover et al. "Automated treatment planning of prostate stereotactic body radiotherapy with focal boosting on a fast-rotating O-ring linac: Plan quality comparison with C-arm linacs". In: *Journal of Applied Clinical Medical Physics* 22 (2021), pp. 59–72.

[373] Yang Sheng et al. "Automatic Planning of Whole Breast Radiation Therapy Using Machine Learning Models". In: *Frontiers in Oncology* 9 (2019).

[374] Martin Hito et al. "Assessing the robustness of artificial intelligence powered planning tools in radiotherapy clinical settings-a phantom simulation approach." In: *Quantitative imaging in medicine and surgery* 11 12 (2021), pp. 4835–4846.

[375] Brian Wang and Gerald White. "The role of clinical medical physicists in the future: Quality, safety, technology implementation, and enhanced direct patient care". In: *Journal of Applied Clinical Medical Physics* 20 (2019), pp. 4–6.

[376] Dao Lam et al. "Predicting gamma passing rates for portal dosimetry based IMRT QA using machine learning." In: *Medical physics* (2019).

[377] Pawel Siciarz et al. "Machine learning for dose-volume histogram based clinical decision-making support system in radiation therapy plans for brain tumors". In: *Clinical and Translational Radiation Oncology* 31 (2021), pp. 50–57.

[378] Cecile J.A. Wolfs and Frank Verhaegen. "What is the optimal input information for deep learning-based pre-treatment error identification in radiotherapy?" In: *Physics and Imaging in Radiation Oncology* 24 (2022), pp. 14–20.

[379] Yuto Kimura et al. "Error detection model developed using a multi-task convolutional neural network in patient-specific quality assurance for volumetric-modulated arc therapy." In: *Medical physics* (2021).

[380] Matthew Nyflot et al. "Deep learning for patient-specific quality assurance: Identifying errors in radiotherapy delivery by radiomic analysis of gamma images with convolutional neural networks". In: *Medical Physics* 46 (2018), pp. 456–464.

[381] Nicholas J. Potter et al. "Error detection and classification in patient-specific IMRT QA with dual neural networks." In: *Medical physics* 47 10 (2020), pp. 4711–4720.

[382] Baozhou Sun et al. "A machine learning approach to the accurate prediction of monitor units for a compact proton machine." In: *Medical physics* 45 5 (2018), pp. 2243–2251.

[383] Hardev S. Grewal et al. "Prediction of the output factor using machine and deep learning approach in uniform scanning proton therapy". In: *Journal of Applied Clinical Medical Physics* 21 (2020), pp. 128–134.

[384] Hao Peng et al. "Recent Advancements of Artificial Intelligence in Particle Therapy". In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 7 (2022), pp. 213–224.

[385] Ying Huang et al. "Virtual Patient-Specific Quality Assurance of IMRT Using UNet++: Classification, Gamma Passing Rates Prediction, and Dose Difference Prediction". In: *Frontiers in Oncology* 11 (2021).

[386] Liyuan Chen et al. "Pretreatment patient-specific quality assurance prediction based on 1D complexity metrics and 3D planning dose: classification, gamma passing rates, and DVH metrics". In: *Radiation Oncology (London, England)* 18 (2023).

[387] Jia Deng et al. "AI-enhanced cancer radiotherapy quality assessment: utilizing daily linac performance, radiomics, dosimetrics, and planning complexity". In: *Frontiers in Oncology* 15 (2025).

[388] Eric Naab Manson et al. "Africa's readiness for artificial intelligence in clinical radiotherapy delivery: Medical physicists to lead the way." In: *Physica medica : PM : an international journal devoted to the applications of physics to medicine and biology : official journal of the Italian Association of Biomedical Physics* 113 (2023), p. 102653.

[389] Maria F. Chan, Alon Witztum, and Gilmer Valdes. "Integration of AI and Machine Learning in Radiotherapy QA". In: *Frontiers in Artificial Intelligence* 3 (2020).

[390] Olga Dona Lemus et al. "Adaptive Radiotherapy: Next-Generation Radiotherapy". In: *Cancers* 16 (2024).

[391] Wei Zhao et al. "Artificial intelligence in image-guided radiotherapy: a review of treatment target localization." In: *Quantitative imaging in medicine and surgery* 11 12 (2021), pp. 4881–4894.

[392] Petros Kalendralis et al. "Automatic quality assurance of radiotherapy treatment plans using Bayesian networks: A multi-institutional study". In: *Frontiers in Oncology* 13 (2023).

[393] Geert de Kerf et al. "A geometry and dose-volume based performance monitoring of artificial intelligence models in radiotherapy treatment planning for prostate cancer". In: *Physics and Imaging in Radiation Oncology* 28 (2023).

[394] Waleed Saeed Ali Al Hagawi et al. "Integrating Artificial Intelligence in Radiotherapy: Challenges and Opportunities in Clinical Workflows". In: *Journal of Ecohumanism* (2024).

[395] Serena Psoroulas et al. "MR-linac: role of artificial intelligence and automation". In: *Strahlentherapie Und Onkologie* 201 (2025), pp. 298–305.

[396] Steven A Hicks et al. "On evaluation metrics for medical applications of artificial intelligence". In: *Scientific reports* 12.1 (2022), p. 5979.

[397] Andrew P Bradley. "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern Recognition* 30 (1997), pp. 1145–1159.

[398] Annika Reinke et al. "Common limitations of performance metrics in biomedical image analysis". In: *Medical Imaging with Deep Learning*. 2021.

[399] Kelly H Zou et al. "Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports". In: *Academic radiology* 11.2 (2004), pp. 178–189.

[400] Daniel P. Huttenlocher, Gregory A. Klanderman, and William J. Rucklidge. "Comparing Images Using the Hausdorff Distance". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15.9 (1993), pp. 850–863. DOI: 10.1109/34.232073.

[401] Stanislav Nikolov et al. "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy". In: *arXiv preprint arXiv:1809.04430* (2018).

[402] Tejinder Kataria et al. "Homogeneity Index: An objective tool for assessment of conformal radiation treatments". In: *Journal of medical physics* 37.4 (2012), pp. 207–213.

[403] Charles S. et al. Mayo. "AAPM Report 263: Standardizing nomenclatures in radiation oncology". In: *Medical Physics* 47.7 (2020), e522–e529.

[404] Mark J Gooding et al. "Fully automated radiotherapy treatment planning: A scan to plan challenge". In: *Radiotherapy and Oncology* 200 (2024), p. 110513.

[405] Shalini K Vinod et al. "A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology". In: *Journal of medical imaging and radiation oncology* 60.3 (2016), pp. 393–406.

[406] Leslie Guzene et al. "Assessing interobserver variability in the delineation of structures in radiation oncology: a systematic review". In: *International Journal of Radiation Oncology* Biology* Physics* 115.5 (2023), pp. 1047–1060.

[407] Giovanni Mauro Cattaneo et al. "Target delineation in post-operative radiotherapy of brain gliomas: interobserver variability and impact of image registration of MR (pre-operative) images on treatment planning CT scans". In: *Radiotherapy and oncology* 75.2 (2005), pp. 217–223.

[408] Marcus Beck et al. "Adherence to contouring and treatment planning requirements within a multicentric trial: Results of the quality assurance of the SAKK 09/10 trial". In: *International Journal of Radiation Oncology* Biology* Physics* 113.1 (2022), pp. 80–91.

[409] Eleni Gkika et al. "The impact of radiotherapy protocol adherence on the outcome of patients with locally advanced NSCLC treated with concurrent chemoradiation: results from the radiotherapy quality assurance of the multicentre international randomized PET-Plan trial". In: *Radiotherapy & oncology* 163 (2021), pp. 32–38.

[410] Nitin Ohri et al. "Radiotherapy protocol deviations and clinical outcomes: A meta-analysis of cooperative group clinical trials". In: *Journal of the National Cancer Institute* 105.6 (2013), pp. 387–393.

[411] Ross A Abrams et al. "Failure to adhere to protocol specified radiation therapy guidelines was associated with decreased survival in RTOG 9704—a phase III trial of adjuvant chemotherapy and chemoradiotherapy for patients with resected adenocarcinoma of the pancreas". In: *International Journal of Radiation Oncology* Biology* Physics* 82.2 (2012), pp. 809–816.

[412] J Van der Veen et al. "Benefits of deep learning for delineation of organs at risk in head and neck cancer". In: *Radiotherapy and Oncology* 138 (2019), pp. 68–74.

[413] Aly H Abayazeed et al. "NS-HGlio: A generalizable and repeatable HGG segmentation and volumetric measurement AI algorithm for the longitudinal MRI assessment to inform RANO in trials and clinics". In: *Neuro-oncology advances* 5.1 (2023), vdac184.

[414] Femke Vaassen et al. "Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy". In: *Physics and Imaging in Radiation Oncology* 13 (2020), pp. 1–6.

[415] Silvia Scoccianti, Barbara Detti, Stefano Cipressi, et al. "Organs at Risk in the Brain and Their Dose Constraints in Adults and Children: A Review of the Literature". In: *Radiotherapy and Oncology* 108.2 (2013), pp. 165–173.

[416] Virginia Braun and Victoria Clarke. "Using thematic analysis in psychology". In: *Qualitative research in psychology* 3.2 (2006), pp. 77–101.

[417] J Richard Landis and Gary G Koch. "The measurement of observer agreement for categorical data". In: *biometrics* (1977), pp. 159–174.

[418] Robert Poel et al. "Deep-Learning-Based Dose Predictor for Glioblastoma–Assessing the Sensitivity and Robustness for Dose Awareness in Contouring". In: *Cancers* 15.17 (Sept. 2023). ISSN: 20726694. DOI: 10.3390/cancers15174226.

[419] J. Ricardo McFaline-Figueroa and Eudocia Q. Lee. "Brain Tumors". In: *The American Journal of Medicine* 131.8 (Aug. 2018), pp. 874–882. ISSN: 15557162. DOI: 10.1016/j.amjmed.2017.12.039.

[420] Claudia Scaringi, Linda Agolli, and Giuseppe Minniti. *Technical advances in radiation therapy for brain tumors*. Nov. 2018. DOI: 10.21873/anticanres.12954.

[421] Indra J. Das, Vadim Moskvin, and Peter A. Johnstone. "Analysis of Treatment Planning Time Among Systems and Planners for Intensity-Modulated Radiation Therapy". In: *Journal of the American College of Radiology* 6.7 (2009), pp. 514–517. ISSN: 15461440. DOI: 10.1016/j.jacr.2008.12.013.

[422] Chenlei Guo et al. "Accurate method for evaluating the duration of the entire radiotherapy process". In: *Journal of Applied Clinical Medical Physics* 21.9 (2020), pp. 252–258.

[423] Aaron Babier et al. "OpenKBP: The open-access knowledge-based planning grand challenge and dataset". In: *Medical Physics* 48.9 (Sept. 2021), pp. 5549–5561. ISSN: 24734209. DOI: 10.1002/mp.14845.

[424] Ann Van Esch et al. "Testing of the analytical anisotropic algorithm for photon dose calculation". In: *Medical Physics* 33.11 (2006), pp. 4130–4148. ISSN: 00942405. DOI: 10.1118/1.2358333.

[425] Michaël Claessens et al. *Quality Assurance for AI-Based Applications in Radiation Therapy*. Oct. 2022. DOI: 10.1016/j.semradonc.2022.06.011.

[426] Florian Kofler et al. "Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient". In: *Journal of Machine Learning for Biomedical Imaging* 2023 (2023), pp. 27–71. URL: https://melba-journal.org/2023:002.

[427] K. Harrison et al. "Machine Learning for Auto-Segmentation in Radiotherapy Planning". In: *Clinical Oncology* 34.2 (Feb. 2022), pp. 74–88. ISSN: 14332981. DOI: 10.1016/j.clon.2021.12.003.

[428] Vanya V Valindria et al. "Reverse classification accuracy: predicting segmentation performance in the absence of ground truth". In: *IEEE transactions on medical imaging* 36.8 (2017), pp. 1597–1606.

[429] Edward GA Henderson et al. "Automatic identification of segmentation errors for radiotherapy using geometric learning". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 319–329.

[430] L. Cubero et al. "Exploring Uncertainty for Clinical Acceptability in Head and Neck Deep Learning-Based OAR Segmentation". In: *Proceedings - International Symposium on Biomedical Imaging*. Vol. 2023-April. IEEE Computer Society, 2023. ISBN: 9781665473583. DOI: 10.1109/ISBI53787.2023.10230442.

[431] Benjamin Roberfroid et al. "DIVE-ART: A tool to guide clinicians towards dosimetrically informed volume editions of automatically segmented volumes in adaptive radiation therapy". In: *Radiotherapy and Oncology* 192 (Mar. 2024), p. 110108. ISSN: 01678140. DOI: 10.1016/j.radonc.2024.110108. URL: https://linkinghub.elsevier.com/retrieve/pii/S016781402400029X.

[432] Amith Kamath et al. "ASTRA: Atomic Surface Transformations for Radiotherapy Quality Assurance". In: *45th Annual International Conference of the IEEE Engineering in Medicine \& Biology Society (EMBC)*. Sydney, July 2023. DOI: 10.1109/EMBC40787.2023.10341062. URL: https://github.com/amithjkamath/astra.

[433] Elias Rüfenacht et al. "PyRaDiSe: A Python package for DICOM-RT-based auto-segmentation pipeline construction and DICOM-RT data conversion". In: *Computer Methods and Programs in Biomedicine* 231 (Apr. 2023). ISSN: 18727565. DOI: 10.1016/j.cmpb.2023.107374.

[434] Amith Kamath et al. "How sensitive are deep learning based radiotherapy dose prediction models to variability in Organs At Risk segmentation?" In: *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2023, pp. 1–4.

[435] Lin Teng et al. "Beam-wise dose composition learning for head and neck cancer dose prediction in radiotherapy". In: *Medical Image Analysis* 92 (Feb. 2024). ISSN: 13618423. DOI: 10.1016/j.media.2023.103045.

[436] Navdeep Dahiya et al. "Deep Learning 3D Dose Prediction for Conventional Lung IMRT Using Consistent/Unbiased Automated Plans". In: *arXiv preprint arXiv:2106.03705* (2021).

[437] C Kontaxis et al. "DeepDose: Towards a fast dose calculation engine for radiation therapy using deep learning". In: *Physics in Medicine & Biology* 65.7 (2020), p. 075013.

[438] Nienke Bakx et al. "Development and evaluation of radiotherapy deep learning dose prediction models for breast cancer". In: *Physics and imaging in radiation oncology* 17 (2021), pp. 65–70.

[439] Sang Hee Ahn et al. "Deep learning method for prediction of patient-specific dose distribution in breast cancer". In: *Radiation Oncology* 16 (2021), pp. 1–13.

[440] Yaoying Liu et al. "Dose prediction using a three-dimensional convolutional neural network for nasopharyngeal carcinoma with tomotherapy". In: *Frontiers in Oncology* 11 (2021), p. 752007.

[441] Meiyan Yue et al. "Dose prediction via distance-guided deep learning: Initial development for nasopharyngeal carcinoma radiotherapy". In: *Radiotherapy and Oncology* 170 (2022), pp. 198–204.

[442] Jinna Yang et al. "Deep learning architecture with transformer and semantic field alignment for voxel-level dose prediction on brain tumors". In: *Medical Physics* 50.2 (2023), pp. 1149–1161.

[443] Lindsey M Appenzoller et al. "Predicting dose-volume histograms for organs-at-risk in IMRT planning". In: *Medical physics* 39.12 (2012), pp. 7446–7461.

[444] Jim P Tol et al. "Can knowledge-based DVH predictions be used for automated, individualized quality assurance of radiotherapy treatment plans?" In: *Radiation Oncology* 10 (2015), pp. 1–14.

[445] Mary P Gronberg et al. "Deep learning–based dose prediction for automated, individualized quality assurance of head and neck radiation therapy plans". In: *Practical radiation oncology* 13.3 (2023), e282–e291.

[446] Satomi Shiraishi and Kevin L Moore. "Knowledge-based prediction of three-dimensional dose distributions for external beam radiotherapy". In: *Medical physics* 43.1 (2016), pp. 378–387.

[447] Jianhui Ma et al. "Individualized 3D dose distribution prediction using deep learning". In: *Artificial Intelligence in Radiation Therapy: First International Workshop, AIRT 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 1*. Springer. 2019, pp. 110–118.

[448] Siri Willems et al. "Feasibility of ct-only 3d dose prediction for vmat prostate plans using deep learning". In: *Artificial Intelligence in Radiation Therapy: First International Workshop, AIRT 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 1*. Springer. 2019, pp. 10–17.

[449] Xinyuan Chen et al. "A feasibility study on an automated method to generate patient-specific dose distributions for radiotherapy using deep learning". In: *Medical physics* 46.1 (2019), pp. 56–64.

[450] Xinyuan Chen et al. "Combining distance and anatomical information for deep-learning based dose distribution predictions for nasopharyngeal cancer radiotherapy planning". In: *Frontiers in Oncology* 13 (2023), p. 1041769.

[451] Chris McIntosh et al. "Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method". In: *Physics in Medicine & Biology* 62.15 (2017), p. 5926.

[452] Zihan Sun et al. "A hybrid optimization strategy for deliverable intensity-modulated radiotherapy plan generation using deep learning-based dose prediction". In: *Medical physics* 49.3 (2022), pp. 1344–1356.

[453] Robert Poel et al. "Impact of random outliers in auto-segmented targets on radiotherapy treatment plans for glioblastoma". In: *Radiation Oncology* 17.1 (2022), p. 170.

[454] MB Altman et al. "A framework for automated contour quality assurance in radiation therapy including adaptive techniques". In: *Physics in Medicine & Biology* 60.13 (2015), p. 5199.

[455] Hsin-Chen Chen et al. "Automated contouring error detection based on supervised geometric attribute distribution models for radiation therapy: a general strategy". In: *Medical physics* 42.2 (2015), pp. 1048–1059.

[456] J Zhang, O Ates, and A Li. "Implementation of a machine learning–based automatic contour quality assurance tool for online adaptive radiation therapy of prostate cancer". In: *International Journal of Radiation Oncology, Biology, Physics* 96.2 (2016), E668.

[457] Lars Johannes Isaksson et al. "Quality assurance for automatically generated contours with additional deep learning". In: *Insights into Imaging* 13.1 (2022), p. 137.

[458] Alain Jungo et al. "Uncertainty-driven sanity check: Application to postoperative brain tumor cavity segmentation". In: *arXiv preprint arXiv:1806.03106* (2018).

[459] Dan Nguyen et al. "A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning". In: *Scientific reports* 9.1 (2019), p. 1076.

[460] Vasant Kearney et al. "DoseNet: a volumetric dose prediction algorithm using 3D fully-convolutional neural networks". In: *Physics in Medicine & Biology* 63.23 (2018), p. 235022.

[461] Kaiming He et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[462] MA Deeley et al. "Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study". In: *Physics in Medicine & Biology* 56.14 (2011), p. 4557.

[463] Elias Rüfenacht et al. "Dose Guidance for Radiotherapy-oriented Deep Learning Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 525–534.

[464] Reza Azad et al. "Medical Image Segmentation Review: The success of U-Net". In: *arXiv preprint arXiv:2211.14830* (2022).

[465] Michal Drozdzal et al. "The importance of skip connections in biomedical image segmentation". In: *International Workshop on Deep Learning in Medical Image Analysis*. Springer. 2016, pp. 179–187.

[466] Naftali Tishby and Noga Zaslavsky. "Deep learning and the information bottleneck principle". In: *2015 ieee information theory workshop (itw)*. Ieee. 2015, pp. 1–5.

[467] Fabian Isensee et al. "nnu-net revisited: A call for rigorous validation in 3d medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 488–498.

[468] Michela Antonelli et al. "The medical segmentation decathlon". In: *Nature communications* 13.1 (2022), p. 4128.

[469] Francesco Galati, Sébastien Ourselin, and Maria A Zuluaga. "From accuracy to reliability and robustness in cardiac magnetic resonance image segmentation: a review". In: *Applied Sciences* 12.8 (2022), p. 3936.

[470] Gaël Varoquaux and Veronika Cheplygina. "Machine learning for medical imaging: methodological failures and recommendations for the future". In: *NPJ digital medicine* 5.1 (2022), p. 48.

[471] Lubomir Hadjiiski et al. "AAPM task group report 273: recommendations on best practices for AI and machine learning for computer-aided diagnosis in medical imaging". In: *Medical physics* 50.2 (2023), e1–e24.

[472] Christoph Kamann and Carsten Rother. "Benchmarking the robustness of semantic segmentation models with respect to common corruptions". In: *International Journal of Computer Vision* 129.2 (2021), pp. 462–483.

[473] Lyndon Boone et al. "ROOD-MRI: Benchmarking the robustness of deep learning segmentation models to out-of-distribution and corrupted data in MRI". In: *arXiv preprint arXiv:2203.06060* (2022).

[474] Olivier Bernard et al. "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?" In: *IEEE transactions on medical imaging* 37.11 (2018), pp. 2514–2525.

[475] Robert Geirhos et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness". In: *arXiv preprint arXiv:1811.12231* (2018).

[476] Rasha Sheikh and Thomas Schultz. "Feature preserving smoothing provides simple and effective data augmentation for medical image segmentation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. Springer. 2020, pp. 116–126.

[477] Kecheng Chen et al. "Learning robust shape regularization for generalizable medical image segmentation". In: *IEEE Transactions on Medical Imaging* (2024).

[478] Cheng Ouyang et al. "Causality-inspired single-source domain generalization for medical image segmentation". In: *IEEE Transactions on Medical Imaging* 42.4 (2022), pp. 1095–1106.

[479] Amith Kamath et al. "Do we really need that skip-connection? Understanding its interplay with task complexity". In: *MICCAI*. 2023.

[480] M Jorge Cardoso et al. "Monai: An open-source framework for deep learning in healthcare". In: *arXiv preprint arXiv:2211.02701* (2022).

[481] Timo Ojala, Matti Pietikäinen, and David Harwood. "A comparative study of texture measures with classification based on featured distributions". In: *Pattern recognition* 29.1 (1996), pp. 51–59.

[482] Niraj P Doshi and Gerald Schaefer. "A comprehensive benchmark of local binary pattern algorithms for texture retrieval". In: *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE. 2012, pp. 2760–2763.

[483] Li Liu et al. "Evaluation of LBP and deep texture descriptors with a new robustness benchmark". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer. 2016, pp. 69–86.

[484] Lukas Hoyer et al. "Grid saliency for context explanations of semantic segmentation". In: *Advances in neural information processing systems* 32 (2019).

[485] Walid Al-Dhabyani et al. "Dataset of breast ultrasound images". In: *Data in brief* 28 (2020), p. 104863.

[486] Korsuk Sirinukunwattana et al. "Gland segmentation in colon histology images: The glas challenge contest". In: *Medical image analysis* 35 (2017), pp. 489–502.

[487] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[488] T. Heimann et al. "Comparison and evaluation of methods for liver segmentation from ct datasets". In: *IEEE Transactions on Medical Imaging* 28.8 (2009), pp. 1251–1265. DOI: 10.1109/TMI.2009.2013851.

[489] B. Everitt. *The Cambridge Dictionary of Statistics*. Cambridge University Press, 1998, pp. 360–360.

[490] L. Maier-Hein et al. "Why rankings of biomedical image analysis competitions should be interpreted with care". In: *Nature Communications* 9 (2018), p. 5217. DOI: 10.1038/s41467-018-07619-7.

[491] D. Gut et al. "Benchmarking of deep architectures for segmentation of medical images". In: *IEEE Transactions on Medical Imaging* 41 (2022), pp. 3231–3241. DOI: 10.1109/TMI.2022.3185713.

[492] A. Kamath et al. "How do 3d image segmentation networks behave across the context versus foreground ratio trade-off?" In: *Medical Imaging meets NeurIPS Workshop*. 2022.

[493] Z. Zeng et al. "Ric-unet: An improved neural network based on unet for nuclei segmentation in histology images". In: *IEEE Access* 7 (2019), pp. 21420–21428. DOI: 10.1109/ACCESS.2019.2896924.

[494] Z. Liu et al. "Robustifying deep networks for medical image segmentation". In: *Journal of Digital Imaging* 34 (2021), pp. 1279–1293. DOI: 10.1007/s10278-021-00491-5.

[495] L. Maier-Hein, B. Menze, and et al. "Metrics reloaded: Pitfalls and recommendations for image analysis validation". In: *arXiv.org* (2022).

[496] M. Reyes et al. "On the interpretability of artificial intelligence in radiology: challenges and opportunities". In: *Radiology: Artificial Intelligence* 2.3 (2020), e190043. DOI: 10.1148/ryai.2020190043.

[497] Geert Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42 (2017), pp. 60–88. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2017.07.005.

[498] Zeju Li, Konstantinos Kamnitsas, and Ben Glocker. "Analyzing Overfitting under Class Imbalance in Neural Networks for Image Segmentation". In: *IEEE Transactions on Medical Imaging* (2020).

[499] Hongyi Wang et al. "Patch-Free 3D Medical Image Segmentation Driven by Super-Resolution Technique and Self-Supervised Guidance". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 131–141.

[500] Jared Hamwood et al. "Effect of patch size and network architecture on a convolutional neural network approach for automatic segmentation of OCT retinal layers". In: *Biomedical optics express* 9.7 (2018), pp. 3049–3066.

[501] Zhaobin Wang, E Wang, and Ying Zhu. "Image segmentation evaluation: a survey of methods". In: *Artificial Intelligence Review* 53 (2020), pp. 5637–5674.

[502] Annika Reinke et al. "Understanding metric-related pitfalls in image analysis validation". In: *Nature methods* (2024), pp. 1–13.

[503] Iman Avazpour et al. "Dimensions and metrics for evaluating recommendation systems". In: *Recommendation systems in software engineering* (2014), pp. 245–273.

[504] Maurice G Kendall. "A new measure of rank correlation". In: *Biometrika* 30.1/2 (1938), pp. 81–93.

[505] Charlotte L. Brouwer et al. "Assessment of manual adjustment performed in clinical practice following deep learning contouring for head and neck organs at risk in radiotherapy". In: *Physics and Imaging in Radiation Oncology* 16 (Oct. 2020), pp. 54–60. ISSN: 24056316. DOI: 10.1016/j.phro.2020.10.001.

[506] Zihan Sun et al. "A physics-informed deep learning model for predicting beam dose distribution of intensity-modulated radiation therapy treatment plans". In: *Physics and Imaging in Radiation Oncology* 34 (2025), p. 100779.

[507] Alexander Kirillov et al. "Segment anything". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 4015–4026.

[508] Sonia Martinot et al. "Differentiable Gamma Index-Based Loss Functions: Accelerating Monte-Carlo Radiotherapy Dose Simulation". In: *International Conference on Information Processing in Medical Imaging*. Springer. 2023, pp. 485–496.

[509] Alex Zwanenburg, Gareth Price, and Steffen Löck. "Artificial intelligence for response prediction and personalisation in radiation oncology". In: *Strahlentherapie und Onkologie* 201.3 (2025), pp. 266–273.

[510] Gabriel S Vidal and Julian C Hong. "Clinical Radiation Oncology in 2040: Vision for Future Radiation Oncology from the Clinical Perspective". In: *Artificial Intelligence In Radiation Oncology*. World Scientific, 2023, pp. 3–17.

[511] Gillian Thomas et al. "The European Organisation for Research and Treatment of Cancer, State of Science in radiation oncology and priorities for clinical trials meeting report". In: *European Journal of Cancer* 131 (2020), pp. 76–88.

[512] Ositomiwa O Osipitan, David Wiant, and Han Liu. "The implementation of knowledge-based planning with partial OAR contours for prostate radiotherapy". In: *Journal of Applied Clinical Medical Physics* (2025), e70004.

[513] Gabriele Palazzo et al. "Real-world validation of Artificial Intelligence-based Computed Tomography auto-contouring for prostate cancer radiotherapy planning". In: *Physics and Imaging in Radiation Oncology* 28 (2023), p. 100501.

[514] Miriam Santoro et al. "Recent applications of artificial intelligence in radiotherapy: where we are and beyond". In: *Applied Sciences* 12.7 (2022), p. 3223.

[515] Bárbara Barbosa. *Artificial Intelligence in Radiotherapy Practice: Highlights from ESTRO24*. URL: https://www.estro.org/About/Newsroom/Newsletter/RTT-Corner/Artificial-Intelligence-in-Radiotherapy-Practice-H.

[516] Eric Topol. *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK, 2019.

List of publications

Under Review :

[J3, Chapter 7, Research Question: B2]
Amith Kamath, Jonas Willmann, Nicolaus Andratschke, and Mauricio Reyes. "The impact of U-Net architecture choices and skip connections on the robustness of segmentation across texture variations", Under Review since November 2024.

Journal:

[J2, Chapter 3, Research Question A1, Joint first author, Link]
Jonas Willmann, **Amith Kamath**, Robert Poel, Elena Riggenbach, Lucas Mose, Jenny Bertholet, Silvan Muller, Daniel Schmidhalter, Nicolaus Andratschke, Ekin Ermiş, Mauricio Reyes. "Predicting the impact of target volume contouring variations on the organ at risk dose: results of a qualitative survey", Radiotherapy and Oncology, Volume: 210, 110999 (2025).

[J1, Chapter 6, Research Question: B1, Joint first author, Link]
Robert Poel, **Amith Kamath**, Jonas Willmann, Nicolaus Andratschke, Ekin Ermiş, Daniel M. Aebersold, Peter Manser, and Mauricio Reyes. "Deep-Learning-Based Dose Predictor for Glioblastoma–Assessing the Sensitivity and Robustness for Dose Awareness in Contouring." Cancers 15, no. 17 (2023): 4226.

Conference:

[C4, Chapter 5, Research Question: A2, Joint First Author, Link]
Amith Kamath, Zahira Mercado Auf der Maur, Robert Poel, Jonas Willmann, Ekin Ermiş, Elena Riggenbach, Nicolaus Andratschke, and Mauricio Reyes. "Comparing the Performance of Radiation Oncologists versus a Deep Learning Dose Predictor to Estimate Dosimetric Impact of Segmentation Variations for Radiotherapy." In Medical Imaging with Deep Learning. 2024.

[C3, Chapter 8, Research Question: B2, Link]
Amith Kamath, Jonas Willmann, Nicolaus Andratschke, and Mauricio Reyes. "Do We Really Need that Skip-Connection? Understanding Its Interplay with Task Complexity." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 302-311. Cham: Springer Nature Switzerland, 2023.

[C2, Chapter 10, Research Question: C1, Link, 2nd Best Student Paper Award]
Amith Kamath, Robert Poel, Jonas Willmann, Ekin Ermiş, Nicolaus Andratschke, and Mauricio Reyes. "ASTRA: Atomic Surface Transformations for Radiotherapy Quality Assurance." In the 45th *IEEE Conference on Engineering in Medicine and Biology*, pp. 1-4. IEEE, 2023.

[C1, Chapter 4, Research Question: B1, Link]
Amith Kamath, Robert Poel, Jonas Willmann, Nicolaus Andratschke, Mauricio Reyes,

2023. "How sensitive are Deep Learning based Radiotherapy Dose Prediction Models to Variability in Organs at Risk Segmentation?" *20th IEEE International Symposium for Biomedical Imaging* (2023).

Workshop:

[**W2, Chapter 11, Research Question: C2**, Master Thesis Supervision, Link]
Zahira Mercado Auf der Maur, **Amith Kamath**, Robert Poel, Jonas Willmann, Ekin Ermiş, Elena Riggenbach, Lucas Mose, Nicolaus Andratschke, and Mauricio Reyes. "AutoDoseRank: Automated Dosimetry-Informed Segmentation Ranking for Radiotherapy." In Medical Image Computing and Computer-Assisted Intervention Workshop on Cancer Prevention through Early Detection, pp. 221-230. Cham: Springer Nature Switzerland, 2024.

[**W1, Chapter 9, Research Question: B2**, Link]
Amith Kamath, Yannick Suter, Suhang You, Michael Mueller, Jonas Willmann, Nicolaus Andratschke, Mauricio Reyes, 2022. "How do 3D image segmentation networks behave across the context versus foreground ratio trade-off?" *In the Medical Imaging meets NeurIPS Workshop, Conference on Neural Information Processing Systems* (2022).

About this Work

This thesis advances *dosimetric contour quality assurance* in radiotherapy by emphasizing clinically meaningful evaluations over traditional geometric metrics. It validates the need for dosimetry-grounded quality assurance, demonstrates that deep learning-based dose predictors can outperform human experts in identifying sub-optimal contours, and assesses the sensitivity and robustness of such models to segmentation variability. The work introduces novel tools, including ASTRA for real-time visualization of local sensitivities and AutoDoseRank for automating segmentation prioritization based on clinical impact. These contributions promote a shift toward intelligent, efficient radiotherapy planning that enhances patient safety and outcomes.

About the Author

Amith Jagannath Kamath, born in Mangalore, India, is a researcher currently based in Bern, Switzerland. He has held various positions at MathWorks Inc., while also earning degrees from the University of Minnesota and Georgia Institute of Technology. His career includes contributions to medical imaging, recognized through multiple academic awards, and he enjoys exploring the intersection of research and practical applications in healthcare.

Cover Artwork

The front and back cover illustrations are computer generated, with stylistic inspirations drawn from the works of Vincent van Gogh and Katsushika Hokusai.

September 2025