Fine-Tuning and Evaluating AI Models for Organ-at-Risk Delineation in Radiotherapy Planning

Zhiyi Ding[1]

Supervisors:, Jamie McClelland[2], Clea Dronne[2]

[1] Department of Physics, Imperial College London, United Kingdom
[2] Hawkes Institute, Department of Medical Physics and Bioengineering, University College London, United Kingdom

Abstract :

This project investigates and evaluates the accuracy of a deep learning model for brainstem and spinal cord delineation on CT scans for head and neck radiotherapy to address clinical need for improved efficiency and consistency in treatment planning. A U-Net based model was trained on the RADCURE dataset, then hyperparameter configurations were explored and the best model was evaluated on an independent test set. Performance was quantified using the Dice Similarity Coefficient (DSC) and Normalised Surface Dice (NSD) at 1.0 mm tolerance, following a qualitative review of selected cases of high, median, and low performance. The model achieved an overall DSC of 0.85 ± 0.05 and an NSD of 0.97 ± 0.04, suggesting a good general performance. Spinal cord delineation was highly accurate (DSC 0.85 ± 0.07, NSD 0.96 ± 0.07), whereas brainstem delineation was less reliable (DSC 0.85 ± 0.04, NSD 0.99 ± 0.03). From the case reviews, the major segmentation failures were associated with mainly poor quality of ground-truth data, low contrast and image resolution. Hence, the model demonstrates a clear potential as an assistive tool to improve efficiency, accelerate clinical workflow, and reduces interobserver variability. These findings highlight that model reliabilty is fundamentally constrained by training data quality and a rigorous expert validation process for safe clinical implementation.

Index Terms: Organ-At-Risk, Delineation, Segmentation, Contouring, AI, Artificial Intelligence, Autocontouring, Radiotherapy, Head and Neck

1 Introduction

Radiotherapy is a common cancer treatment through physically destroying the tumours through doses of radiation. It is one of the three main treatments for cancer (along with surgery and drug therapy) and should account for $\tilde{5}0\%$ of patient treatments (RadiotherapyUK, 2023). However, access to radiotherapy varies significantly across the world. It is $\tilde{9}0\%$ in high-income countries, $\tilde{5}0$-60% in middle-income countries, and only $\tilde{1}0\%$ in low-income countries (VCDNP, 2022). In addition, 20 countries in Africa do not have radiotherapy machines at all (VCDNP, 2022). Improving access is not only about acquiring more machines. It also depends on making treatment planning more efficient.

Treatment planning is highly time-consuming, especially for head and neck cancer. On average, it takes about 5 to 7 hours to produce a plan (Guo et al., 2020). Contouring Organs-At-Risk (OAR), which are radiosensitive structures whose overdose may cause toxicity, is one of the main inefficiencies in the planning phase (Agazaryan et al., 2020). In head and neck Computed Tomography (CT), an X-ray scan used in radiotherapy planning, clinical scientists often struggle to delineate the brainstem, as its boundary appears similar to nearby soft tissues (low contrast). Moreover, manual contouring can be subjective and often demonstrates inter-observer variability. Reducing time and variability in OAR contouring may assist in streamlining planning and support the better use of limited radiotherapy resources.

Focusing on head and neck CT, this investigation evaluates the accuracy of AI-based auto-contouring for the brainstem and spinal cord. The evaluation targets geometric segmentation accuracy, assessed on an independent test set using the Dice Similarity Coefficient (DSC) and the Normalised Surface Dice (NSD), rather than dosimetric outcomes or clinical endpoints.

The primary model is a U-Net. Performance is reported on a test set using DSC and NSD. Images of high, median, and low performing cases are included to explain typical failure modes and boundary conditions.

Overall, the study provides a baseline evaluation for brainstem and spinal cord auto-contouring on head and neck CT and offers practical observations for clinical integration and model improvement, including guidance for staff without AI expertise to assess integrability. By improving the efficiency of treatment planning, the project aligns with the United Nations Sustainable Development Goals 3 (Good Health and Well-being) and 10 (Reduced Inequalities). The objective is to evaluate the extent to which AI may enhance the efficiency and accessibility of radiotherapy planning, thereby enabling more timely and effective treatment delivery.

2 Background

2.1 Convolution and Feature Maps

The main purpose of a convolution is to turn an image into a feature map. A feature map is a new version of the image where specific visual traits, such as edges, corners, or colors, are emphasised. This is done by moving a matrix of numbers, known as a kernel, across the image. At each step, the kernel interacts with a small patch of the image to produce a filtered output that highlights the desired features. This is analogous to filters scanning across a page to reveal different details depending on the lens.

2.2 Convolutional Neural Network (CNNs)

When many of these convolutions are stacked together, they form a Convolutional Neural Network (CNN). Each layer in the network applies several filters, allowing the model to gradually detect more complex patterns (e.g., from simple lines and shapes to complete

objects).

2.3 Supervised Learning: Training, Validation, Testing

To determine the best numbers for these kernels, the network learns through a process called supervised learning, where it is trained using examples that already have correct answers (the ground truth). In this project, the ground truth comprised of CT scans with expert-drawn contours of the brainstem and spinal cord. Supervised learning involves three phases: **Training, Validation, and Testing.**

During training, the network compares its predictions to the ground truth. The difference between the two is called **loss**, which is a measure of error. A higher loss score means that the prediction is poor. The network then adjusts its filters to reduce this loss through a process called **backpropagation**.

However, undergoing training solely on the network presents the risk of it starting to 'remember' the answers instead of learning general rules. This problem is called **overfitting**, where the model works well on known data but fails on new cases. To prevent this, validation data are used to check that learning remains general and balanced.

Finally, once the model performs well on the validation data, it is tested on unseen data. Its segmentation accuracy is then evaluated using two types of metrics: **overlap-based metrics and boundary metrics.**

2.4 Evaluation Metrics

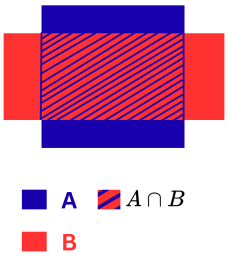

Figure 1. Visual representation of the components for the Dice Similarity Coefficient (DSC). The metric quantifies the overlap between two sets, such as a ground truth (A) and a prediction (B), by comparing the size of their intersection ($A \cap B$) to the total size of both sets.

2.4.1 Dice Similarity Coefficient (DSC): The Dice Similarity Coefficient (DSC) was chosen as the overlap-based metric to account for the high variability in Organ-at-Risk (OAR) sizes across the patient cohort (Maier-Hein et al., 2024). It can be calculated as

$$DSC(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{1}$$

where A stands for the set of pixels within A, and B stands for the set of pixels within B.

2.4.2 Normalised Surface Dice (NSD): For the boundary-based evaluation, the Normalised Surface Distance (NSD) was chosen, specifically focusing on the precision of the structure boundaries (Maier-Hein et al., 2024).

It can be calculated as:

$$NSD(A, B)^{(\tau)} = \frac{|S_A \cap B_B^{(\tau)}| + |S_B \cap B_A^{(\tau)}|}{|S_A| + |S_B|} \tag{2}$$

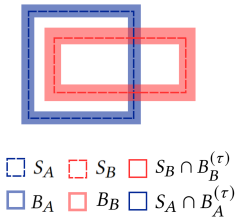

Figure 2. Visual representation of the Normalised Surface Distance (NSD) metric at a given tolerance (τ = 1mm). The NSD evaluates boundary agreement by calculating the percentage of the predicted surface (S_B) that falls within a tolerance band around the ground truth surface ($B_A^{(\tau)}$), and vice versa. S_A represents the set of boundary points of A, S_B represents the set of boundary points of B, $B_A^{(\tau)}$ is the set of all points within tolerance τ of S_A , and $B_B^{(\tau)}$ is the set of all points within tolerance τ of S_B .

Variable representations are identical to the caption in fig.2.

2.5 U-Net Architecture

In this research, a U-Net developed by (Ronneberger et al., 2015) was used to train the segmentation model. The U-Net has two main parts: **an encoder and a decoder** (Endemann, 2024). The **encoder** works like a CNN, down-sampling the image while understanding its key characteristics, but losing some fine spatial detail. The **decoder** then rebuilds the compressed image by up-sampling, creating a segmentation mask to re-establish the spatial information.

To recover fine-grained information lost during encoding, **skip connections** are used. These directly copy the features from a layer in the encoder and directly merge with the corresponding layer in the decoder. This design allowed the network to create segmentation masks that are both contextually accurate and spatially precise, giving the U-Net its U-shaped structure.

2.6 nnU-Net

Alternatively, the project also tested segmentation performance on a fine-tuned nnU-Net. In essence, a nnU-Net is a U-Net-based model that automatically adapts to new datasets (Isensee et al., 2020). This deviates from U-Net fine-tuning, which adapts to new datasets manually. However, rather than the full application of nnU-Net, this project incorporated a building block from the MONAI (Medical open network for AI) library that uses the same principles called Dyn-Unet (MONAI, 2025).

2.7 Definitions of Hyperparameters

Hyperparameters are configuration variables set before model training to define its learning process. Adjusting these hyperparameters is a critical step in optimising model performance. Key hyperparameters in this project include the learning rate, batch size, patience, dropout, and model architecture. A complete list and detailed definitions of all hyperparameters used in this study are available in Appendix A.0.1.

3 Methods

3.1 Training: Experimental Set Up and Data Preparation

This study delineates the brainstem and spinal cord on head and neck CT scans using the RADCURE dataset (Welch et al., 2024), which contains manual structural contours for each patient. CT images and segmentation masks are converted into NIfTI format, aligned by case ID, and filtered to include only cases containing both organs. Data processing and model training are implemented in PyTorch and MONAI, written in Python. The pipeline converts CT volumes to tensors, preserving the original CT intensity and spacing.

In terms of variables, the main independent variable is the model type, with a 3D U-Net (Ronneberger et al., 2015) as the primary network and Dyn-Net (MONAI, 2025) as a comparative model, along with hyperparameters (i.e., learning rate, model architecture, patience number, dropout). Dependent variables are the Dice Similarity Coefficient (DSC) and the Normalised Surface Dice (NSD) with a 1 mm tolerance. Data are split into 70% training, 10% validation, and 20% testing using a fixed random seed (42). The model is trained with the Dice loss and the Adam optimiser (learning rate $1 * 10^{-5}$), running up to 1000 epochs. Each run logs all configuration details, including model type and hyperparameters. Training was conducted on an NVIDIA RTX 5090 GPU after initial trials on the University College London HPC system.

3.2 Validation

In validation, the model generalises effectively and prevents overfitting by monitoring validation dice loss during training. The validation split (10%) remains constant across runs and follows the same preprocessing as the training set. Early stopping halts the training after 30/50 epochs without improvement and restores the best-performing checkpoint. The Dyn-Net model follows the same validation process for comparison only.

3.3 Testing

The final model is evaluated on an independent 20% test set to measure generalisation. Results are compared with previous clinical segmentation standards using the DSC and NSD. Per-case and mean DSC and NSD values are calculated for both organs, serving as benchmarks for performance analysis.

Six representative cases (best, median, and worst for each metric) are selected for qualitative review. Sagittal slices are taken at the spinal cord's center of mass (COM). Consistent anatomical orientation is maintained using DICOM header information.

3.4 Figure Generation

For each case, two visualisations are created for the brainstem and spinal cord, respectively. A colour-coded error map was overlaid onto the original CT image, identifying the spatial locations of True Positives (TP), False Positives (FP), and False Negatives (FN).

4 Results

In this project, the optimal set of parameters selected for the final evaluation was:

Model Type:	U-Net
Learning Rate:	0.0001
Batch Size:	10
Patience:	30
Architecture:	(32,64,128,256)
Stride:	(2,2,2)
Dropout:	0.1
Surface to Total mm:	1.0

A summary of previously tested model architecture configurations is available in Table. 3 in Appendix A.0.2.

The quantitative performance of this optimal model on the test set is summarised in Table. 1. The metrics are presented for the overall segmentation task, as well as for each OAR independently, reported as the mean and standard deviation across all cases in the test set.

Table 1
Segmentation performance measured by mean DSC and NSD: overall and by class. NSD tolerance 1.0 mm.

Category	Mean DSC	Mean NSD
Overall	0.85 ± 0.05	0.97 ± 0.04
Brainstem	0.85 ± 0.07	0.96 ± 0.07
Spinal cord	0.85 ± 0.04	0.99 ± 0.03

A set of metrics for the selected cases is compiled in Figure. 2.

Table 2
DSC and NSD results for the selected set of cases

Case ID	Selection criterion	DSC	NSD (1.0mm)
A	DSC minimum	0.54	0.72
B	DSC medium	0.86	1.00
C	DSC maximum	0.92	1.00
D	NSD minimum	0.60	0.70
E	NSD medium	0.89	0.99
F	NSD maximum	0.89	1.00

4.1 Visualisation for selected cases

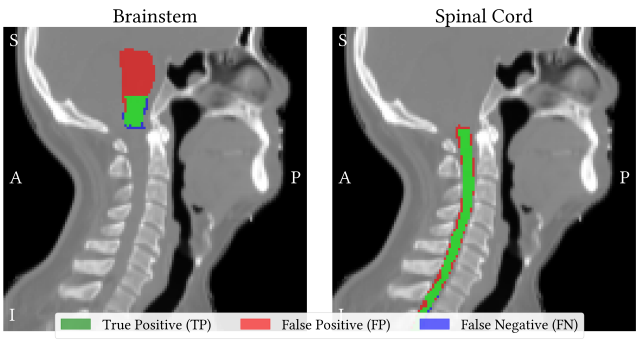

Figure 3. Case **A**: Segmentation results (sagittal CT view) for a low DSC example. (Left) Brainstem results. (Right) Spinal Cord results. Overlays illustrate the spatial distribution of True Positives (TP, green), False Positives (FP, red), and False Negatives (FN, blue). S/A/I/P stands for superior/anterior/inferior/posterior on a right sagittal view.

In Fig. 3, a large FP region surrounds the brainstem, and a posterior FP band extends along the inferior spinal cord. Short FN segments

are also present superiorly on the brainstem.

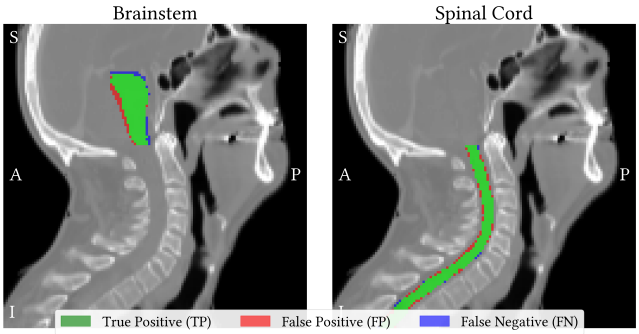

Figure 4. Case **B**: Segmentation results (sagittal CT view) for a moderate DSC example. (Left) Brainstem results. (Right) Spinal Cord results. Overlays illustrate the spatial distribution of True Positives (TP, green), False Positives (FP, red), and False Negatives (FN, blue) S/A/I/P stands for superior/anterior/inferior/posterior on a right sagittal view. .

In Fig. 4, the predicted outline is closely aligned with the GT contours across most of the brainstem and spinal cord. A thin, posterior band of FP appears along the spinal cord mid-to-inferior segment; a small mixed FP/FN pattern is visible at the cervicomedullary junction.

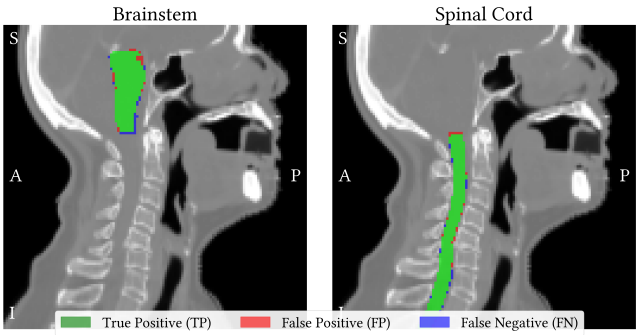

Figure 5. Case **C**: Segmentation results (sagittal CT view) for a high DSC example. (Left) Brainstem results. (Right) Spinal Cord results. Overlays illustrate the spatial distribution of True Positives (TP, green), False Positives (FP, red), and False Negatives (FN, blue). S/A/I/P stands for superior/anterior/inferior/posterior on a right sagittal view.

In Fig. 5, there is a mix of FN and FP segments around the boundaries of the brainstem and spinal cord.

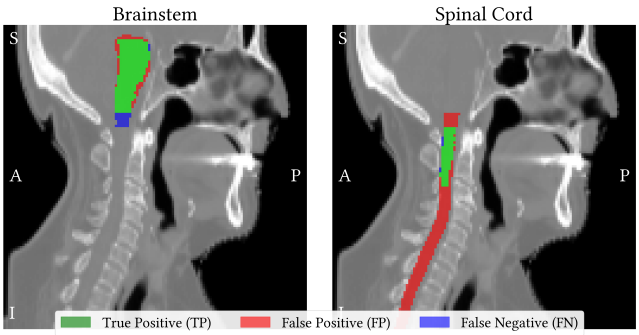

Figure 6. Case **D**: Segmentation results (sagittal CT view) for a low NSD example. (Left) Brainstem results. (Right) Spinal Cord results. Overlays illustrate the spatial distribution of True Positives (TP, green), False Positives (FP, red), and False Negatives (FN, blue). S/A/I/P stands for superior/anterior/inferior/posterior on a right sagittal view.

In Fig. 6, a long, wide, posterior FP band runs from superior to inferior along the spinal cord, and the upper brainstem contains a contiguous FP region.

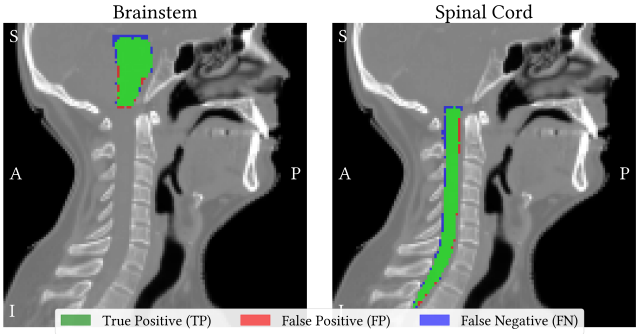

Figure 7. Case **E**: Segmentation results (sagittal CT view) for a moderate NSD example. (Left) Brainstem results. (Right) Spinal Cord results. Overlays illustrate the spatial distribution of True Positives (TP, green), False Positives (FP, red), and False Negatives (FN, blue). S/A/I/P stands for superior/anterior/inferior/posterior on a right sagittal view.

In Fig. 7, the superior brainstem shows a contiguous FN cap, while the spinal cord displays only sparse FP/FN speckles. The remainder of the boundary is well aligned.

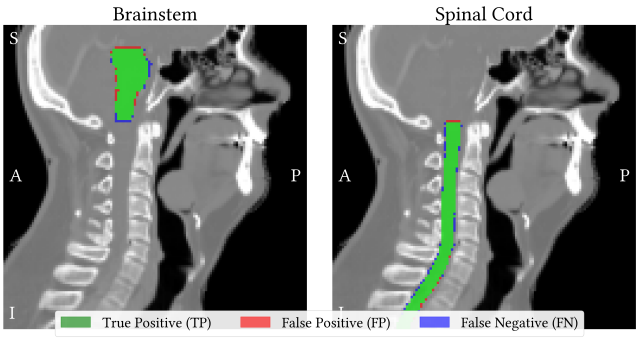

Figure 8. Case **F**: Segmentation results (sagittal CT view) for a high NSD example. (Left) Brainstem results. (Right) Spinal Cord results. Overlays illustrate the spatial distribution of True Positives (TP, green), False Positives (FP, red), and False Negatives (FN, blue). S/A/I/P stands for superior/anterior/inferior/posterior on a right sagittal view.

In Fig. 8, the spinal cord exhibits a predominantly FN streak running inferiorly, with a small FP patch at the superior brainstem. Anywhere else aligns closely with the boundaries.

5 Discussion

5.1 Metric Analysis

As shown in Table. 1, segmentation metrics reveal that the model performs at a level approaching clinical acceptability, though with notable variations. The average DSC was 0.85 ± 0.05 and the average NSD was 0.97 ± 0.04. These metrics are consistent with the visual appearance of the segmentation, indicating that the model produces plausible delineations. While the NSD meets expert-approved thresholds for radiotherapy segmentation accuracy (90–95%, NICE, 2023), the DSC falls slightly below this benchmark, indicating that although surface alignment is strong, volumetric overlap remains imperfect. However, the large standard deviations associated with both metrics indicate considerable performance variability across the dataset, highlighting a mix of successful and failed segmentation that is worth considering when evaluating model reliability. Model performance for different OAR showed variations. For brainstem segmentation, the model achieved an average DSC of 0.85 ± 0.07 and an NSD of 0.96 ± 0.07, indicating robust but imperfect performance. Spinal cord segmentation has a similar average DSC of 0.85 ± 0.04 and NSD of 0.99 ± 0.03. This similarity suggests that the model is equally proficient at identifying the general regions of the brainstem and spinal cord. However, the slight divergence in NSD values and uncertainties suggests that the model's contouring precision is dependent on anatomical context. In this case, this is demonstrated by the differences in the shape of the spinal cord and brainstem. The difference likely arises because the spinal cord being neatly encased by the vertebral canal. In contrast, the brainstem is surrounded by soft tissues, which makes the boundaries more ambiguous and challenging for the model to delineate with higher precision.

While the volumetric overlap achieved by the model was similar for both structures, its delineation of the surface contours was substantially more accurate and consistent for the spinal cord. This enhanced surface accuracy is likely attributable to the distinct anatomical context of the spinal cord, which is clearly defined by the surrounding vertebral canal, providing a clear boundary for the model.

5.2 Error Modes

5.2.1 Brainstem: Boundary errors concentrated at the cranio-cervical junction were most common. We observed thin FN bands at the superior brainstem and, in certain cases, FP and FN bands at the inferior border where the medulla transitions to the spinal cord. In Case **F**, the model slightly under-contoured the brainstem perimeter, while in Case **B**, a small mixed FP/FN region was present at the inferior brainstem, and in Case **D** , the superior brainstem exhibited a large FP region. In Case **A**, by contrast, the brainstem exhibited a large, contiguous FP patch, encasing much of the structure (dominant over-segmentation), consistent with its low DSC (0.54) and NSD (0.72).

5.2.2 Spinal Cord: The dominant error was a continuous FP band along the posterior cord, spanning a long superior-to-inferior segment. This pattern appears clearly in Case **B** and was most pronounced in Case **D**, where a long, high-amplitude posterior FP band extended across much of the spinal cord. Less frequently, short FN bands that tightened the contour inward and occasional short discontinuities at the junction were observed in cases of brainstem confusion.

5.3 Impact on Evaluation Metrics

5.3.1 Brainstem: Small filled FP regions around the brainstem inflate the volume and reduce DSC by a few percentage points, but leave NSD nearly unchanged because the erroneous surface lies within the 1 mm tolerance. This is consistent with Case **B**, which achieved DSC 0.91 and NSD 1.00. Thin FN bands can also have a limited effect on Dice if localized, as in Case **F** with DSC 0.89 and NSD 1.00. When the FN pattern forms a contiguous cap at the superior brainstem, both metrics may drop, as seen in Case **E** with DSC 0.89 and NSD 0.99. However, the decrease is only slightly, given the limited extent of the FN region in this case.

5.3.2 Spinal Cord: Thin, consistent FP bands reasonably reduce DSC while keeping NSD high, as the predicted surface remains close to and parallel with GT. For instance, Case **B** illustrates this behavior with a DSC of 0.91 and NSD of 1.00. The lowest scoring examples in the selected set, Case **A** (DSC 0.54, NSD 0.72) and Case **D** (DSC 0.60, NSD 0.70), exhibited long FP bands, which reduced DSC due to amplified area penalties. Case **E** shows that when errors are sparse, metrics remain high, with DSC 0.89 and NSD 0.99.

5.4 Likely Contributing Factors and further improvements

5.4.1 Brainstem

1. **Low Soft Tissue Contrast:** As demonstrated in the CT scans, a primary challenge in CT imaging is the inherently low contrast between the brainstem and its adjacent soft tissues, such as the temporal lobe and cerebellum. This results in unclear anatomical boundaries, which complicate the delineation of ground truth contours by clinical experts and lead to high interobserver variability within the training data. Such inconsistencies can diminish model accuracy along these ambiguous boundaries, as demonstrated by FPs and FNs in Cases **A**, **B**, and **D**. To address this, multi-modal image fusion can be employed (Safari et al., 2023). Fusing CT scans, which demonstrate bone structures, with Mag-

netic Resonance Imaging, which enhances soft-tissue contrast, yields a composite image with more sharply defined anatomical structures. Another approach involves refining the head and neck OAR segmentation guidelines to provide clearer definitions of the brainstem's surrounding anatomy, hence reducing ambiguity and improving consistency in contouring.

2. **Limitations in Quality of the Training Set:** The limitations in the training dataset were a root cause of the AI segmentation failures in Cases **A** and **D**.It should be emphasised that the RADCURE dataset collected data from 2000 to 2015 (Welch et al., 2024). However, the first consensus international guidelines for head and neck OAR delineation were not published until 2017 by Brouwer et al., 2015. Therefore, the ground truth contours within this dataset may not adhere to current standards, potentially being incomplete or inconsistent across cases. Visually, although the model produced accurate contours of the OARs by averaging variations across the training data, the results appear unreliable because the ground truth data used for validation or testing lacks standardisation. To mitigate this, implement a systematic and rigorous data quality assurance protocol, with experts revising or redrawing them according to current international guidelines (e.g., DAHANCA and EORTC guidelines).
3. **Structure-Induced Imaging Artifacts:** Unlike any other region, the head and neck contain a bone-rich skull base, dense cortical bone, and may also include dental impacts. This presents a unique imaging challenge due to the prevalence of high-density structures, which can induce artifacts. For instance, beam hardening degrades image quality and obscures the boundaries. As mitigation, employ iterative metal artifact reduction (iMAR) (Bayerl et al., 2023) to reduce artifacts from dental hardware.
4. **Ground Truth Variability in Brainstem Shape:** Notably, the ground-truth segmentation for Case **A** was substantially smaller than that of most others, pointing to significant variations in brainstem shape and segmentation standards across the dataset.

5.4.2 Spinal Cord

1. **Variations in Ground Truth Segmentation:** Contrary to the brainstem, the spinal cord is not always segmented along its full extent, as it is not always routinely required in clinical workflows. This results in partial ground-truth annotations that vary across cases, introducing inconsistencies in both training and evaluation. Such variability can lead to apparent over-segmentation by the model in regions where no ground truth is available, complicating metric interpretation and potentially penalizing otherwise accurate predictions.
2. **Systematic and Random Errors:** Segmentation discrepancies also arose from small systematic differences and random noise. For example, regions of FP and FN in Cases **E** and **B** suggest consistent but minor deviations from the ground truth. In contrast, Cases **F** and **C** presented isolated FP and FN patches, which are likely attributed to random variations or imaging noise.

5.5 Clinical Relevance

Unlike the lungs and liver, the brainstem and spinal cord are serial OARs, where under-segmentation poses a significant clinical risk. Irradiation of even small, unprotected volumes can lead to severe, irreversible toxicities such as myelopathy (Xin et al., 2020). Therefore, despite acceptable overall DSC and NSD metrics, the presence of FNs along the OAR boundaries necessitates that this model functions solely as an assistive tool for manual contouring rather than as an autonomous solution. It should be closely monitored to avoid significant failures, such as Cases **A** and **D**. Currently, automated segmentation outputs are reviewed manually and corrected prior to finalising the treatment plan. In line with NHS guidelines (NICE, 2023), AI technologies may only be implemented after receiving approval through the Digital Technology Assessment Criteria (DTAC), which requires evidence of time savings, including the time needed for healthcare professionals to review and edit the segmentations . Uncertainty measures and error prediction can potentially help with this. This presents a significant hurdle, as a full review of every contour can be time-consuming. Uncertainty measures and error prediction tools integrated into this project help direct clinicians' focus toward regions with low confidence or potential inaccuracies. By streamlining the review process and minimising the need for manual corrections, these tools contribute to meeting the efficiency standards outlined in the DTAC approval criteria for emerging AI technologies not yet acknowledged by the NHS.

5.6 Next Steps

Building on the challenges observed in head-and-neck OAR segmentation, this project could be extended to enhance delineation accuracy through multi-modal image fusion. A key next step is the development of variation-aware models that can flag high-risk regions for expert review, particularly when auto-segmentation deviates significantly from prior contours. This capability is essential for mitigating automation bias and supporting clinical adoption.

Furthermore, research should train the model using higher-quality datasets with consistent ground truth annotations, reducing noise and improving model reliability. Evaluating the accuracy of autosegmentations in other OARs beyond the head-and-neck region will help evaluate generalisability, while applying these robust methods to the more complex task of tumour segmentation could further demonstrate their clinical value.

6 Conclusion

This research evaluated an AI model for brainstem and spinal cord segmentation in head and neck CT scans, finding that while it can achieve a high degree of accuracy in many instances, its overall performance is inconsistent. The primary sources of error were concluded to be the inherently low soft-tissue contrast of CT imaging for the brainstem and interobserver variations between cases from the RADCURE training dataset, which were dated before the first proposed international contouring guidelines. Consequently, while the model shows good promise as an assistive tool to improve clinical efficiency and reduce inter-observer variability, it should not be used independently due to its potential for significant errors and the inherent risks in the profession of radiotherapy. Moreover, safer and more effective clinical integration can be achieved by improving model accuracy through the use of quality-

assured training datasets, enhanced image quality via multimodal fusion, and data augmentation. Crucially, any such integration requires a mandatory workflow wherein every automated contour is subjected to rigorous expert review and modification prior to clinical use.

Acknowledgements

This research is funded by the Laidlaw Foundation, hosted by Imperial College Business School, under the Laidlaw Scholars Leadership and Research Programme. I wish to express my sincere gratitude to my supervisor, Professor Jaime McClelland from the Hawkes Institute, University College London, for his generous commitment of time in arranging, attending, and hosting our project meetings, and for his foundational guidance in establishing this research. I would also like to extend my special thanks to PhD Student Clea Dronne, also from the Hawkes Institute, for her ongoing and patient support in navigating the technical challenges encountered throughout this project. I also extend my gratitude to Tianyuan Zhang, a peer in my course and year at Imperial College London, for his technical support in code execution using a local GPU and for generously lending his NVIDIA RTX 5090, which was used to train the model.

References

Agazaryan, Nzhde et al. (Sept. 2020). "The Timeliness Initiative: Continuous Process Improvement for Prompt Initiation of Radiation Therapy Treatment". In: *Advances in Radiation Oncology* 5, pp. 1014–1021. DOI: 10.1016/j.adro.2020.01.007.

Bayerl, Nadine et al. (Dec. 2023). "Iterative Metal Artifact Reduction in Head and Neck CT Facilitates Tumor Visualization of Oral and Oropharyngeal Cancer Obscured by Artifacts From Dental Hardware". In: *Academic radiology* 30, pp. 2962–2972. DOI: 10.1016/j.acra.2023.04.007. URL: https://pubmed.ncbi.nlm.nih.gov/37179206/ (visited on 10/07/2025).

Brouwer, Charlotte L. et al. (Oct. 2015). "CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines". In: *Radiotherapy and Oncology* 117, pp. 83–90. DOI: 10.1016/j.radonc.2015.07.041.

Endemann, Chris (Sept. 2024). *U-Net: Convolutional Networks for Biomedical Image Segmentation – Nexus: Crowdsourced ML Resources*. Github.io. URL: https://uw-madison-datascience.github.io/ML-X-Nexus/Toolbox/Models/UNET.html (visited on 10/08/2025).

Guo, Chenlei et al. (July 2020). "Accurate method for evaluating the duration of the entire radiotherapy process". In: *Journal of Applied Clinical Medical Physics* 21, pp. 252–258. DOI: 10.1002/acm2.12959. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7497908/ (visited on 10/05/2025).

Isensee, Fabian et al. (Dec. 2020). "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* 18, pp. 203–211. DOI: 10.1038/s41592-020-01008-z. URL: https://www.nature.com/articles/s41592-020-01008-z.

Maier-Hein, Lena et al. (Feb. 2024). "Metrics reloaded: recommendations for image analysis validation". In: *Nature Methods* 21, pp. 195–212. DOI: 10.1038/s41592-023-02151-z. URL: https://www.nature.com/articles/s41592-023-02151-z (visited on 10/07/2025).

MONAI (2025). *monai.networks.nets.dynunet — MONAI 0 Documentation*. Readthedocs.io. URL: https://monai-dev.readthedocs.io/en/fixes-sphinx/_modules/monai/networks/nets/dynunet.html#DynUNet (visited on 10/31/2025).

NICE (Sept. 2023). *Artificial intelligence technologies to aid contouring for radiotherapy treatment planning: early value assessment Health technology evaluation*. URL: https://www.nice.org.uk/guidance/hte11/resources/artificial-intelligence-technologies-to-aid-contouring-for-radiotherapy-treatment-planning-early-value-assessment-pdf-50261966238661 (visited on 10/07/2025).

RadiotherapyUK (June 2023). *RADIOTHERAPY: AN ANALYSIS OF HOW RADIOTHERAPY SERVICES IN THE UK COMPARE WITH OTHER COUNTRIES*. URL: https://radiotherapy.org.uk/wp-content/uploads/2023/06/International-comparisons-full-report-140623.pdf (visited on 10/04/2025).

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Lecture Notes in Computer Science* 9351, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

Safari, Mojtaba, Ali Fatemi, and Louis Archambault (Dec. 2023). "MedFusionGAN: multimodal medical image fusion using an unsupervised deep generative adversarial network". In: *BMC Medical Imaging* 23. DOI: 10.1186/s12880-023-01160-w.

VCDNP (June 2022). *Advancing Access to Radiotherapy in Middle-and Low-Income Countries*. Vienna Center for Disarmament and Non-Proliferation. URL: https://vcdnp.org/advancing-access-to-radiotherapy/ (visited on 10/04/2025).

Welch, Mattea L. et al. (Apr. 2024). "RADCURE: An open-source head and neck cancer CT dataset for clinical radiation therapy insights". In: *Medical Physics* 51, pp. 3101–3109. DOI: 10.1002/mp.16972. URL: https://pubmed.ncbi.nlm.nih.gov/38362943/ (visited on 10/05/2025).

Xin, Xin et al. (Dec. 2020). " Comparative Study of Auto Plan and Manual Plan for Nasopharyngeal Carcinoma Intensity-Modulated Radiation Therapy ". In: *Cancer management and research* Volume 12, pp. 12439–12445. DOI: 10.2147/cmar.s226495. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7719327/#:~:text=Organs%20are%20divided%20into%20serial (visited on 10/08/2025).

A Hyperparameter Details

A.0.1 Hyperparameter Definitions

- **Learning Rate**: A hyperparameter that controls how quickly model weights are updated during training.
- **Batch Size**: A hyperparameter that quantifies the number of samples processed before the model's weights are updated simultaneously.
- **Patience**: A hyperparameter in early stopping (an algorithm that prevents overfitting) specifying how many training epochs can pass without improvement in the validation score before training stops.
- **Architecture**: Specifies the design of the training model (i.e., U-Net or nnU-Net) in terms of the number of feature maps at each level of the encoder. For instance, an architecture defined as (16, 32, 64) means that the encoder has 16, 32, and 64 feature maps at successive levels.
- **Stride**: Indicates the down-sampling rate applied at the U-Net decoder. For example, a stride of (2, 2) reduces the number of feature maps from the architecture above to (8, 16, 32).
- **Dropout**: The probability of an artificial neuron being temporarily set to zero during a single training step.

A.0.2 Model Configuration Trials

Table 3
Hyperparameter settings for various model configurations.

Model Type	Learning Rate	Batch Size	Patience	Architecture	Stride	Dropout
U-Net	0.001	8	30	(16,32,64,128)	(2,2,2)	/
U-Net	0.0001	10	30	(16,32,64,128)	(2,2,2)	/
U-Net	0.0001	12	30	(16,32,64,128)	(2,2,2)	/
U-Net	0.0001	16	50	(16,32,64,128)	(2,2,2)	/
U-Net	0.0001	10	50	(16,32,64,128)	(2,2,2)	0.1
U-Net	0.0001	12	30	(32,64,128,256)	(2,2,2)	0.1
U-Net	0.0001	12	30	(64,128,256,512)	(2,2,2)	0.2
Dyn-Net	0.0001	4	30	(32,64,128,256)	(2,2,2)	0.1
Dyn-Net	0.0001	6	30	(32,64,128,256)	(2,2,2)	0.1
U-Net	0.0001	12	30	(64,128,256,512)	(2,2,2)	0.1
U-Net	0.00003	12	30	(64,128,256,512)	(2,2,2)	0.2
U-Net	0.00003	16	30	(64,128,256,512)	(2,2,2)	0.2
U-Net	0.00003	12	30	(32,64,128,256)	(2,2,2)	0.2