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

Author: Una Zhiyi Ding[1]
Supervisors: Jamie McClelland[2], Clea Dronne[2]
[1]Blackett Laboratory, Department of Physics, Imperial College London, UK.
[2]Hawkes Institute, Department of Medical Physics and Bioengineering, University College London, UK.

Background

- Radiotherapy is a common cancer treatment, but the treatment planning phase is a significant bottleneck, often taking 5-7 hours per patient.
- **Problem:** On head & neck CTs, the brainstem and spinal cord are hard to see (low contrast), and manual contouring is subjective and inconsistent.
- **Project Goal:** Evaluate an AI model to auto-contour these OARs, and to improve efficiency, consistency, and accuracy.

Methods

- **Model: U-Net** [1] based deep learning architecture.
- **Data:** RADCURE dataset (head & neck CT scans, [2]), split 70% train / 10% validation/ 20% test.
- **Implementation:** PyTorch & MONAI pipeline. Trained on an NVIDIA RTX 5090 GPU.
- **Evaluation Metrics:** Model performance analysed on test sets using:
  Dice Similarity Coefficient (DSC): Measures the volumetric overlap.
  Normalised Surface Dice (NSD): Measures the accuracy of the boundary surface (1.0 mm tolerance).

DSC Segmentation Examples

Qualitative Error Analysis

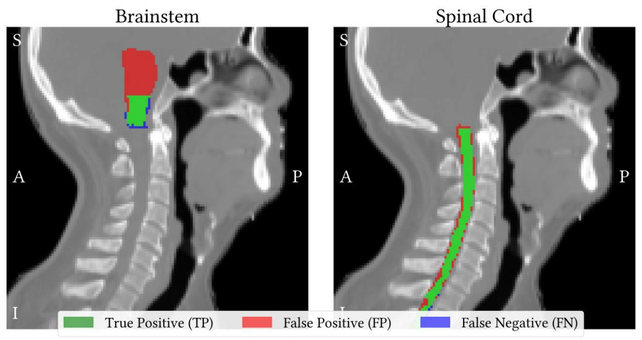

Figure A: DSC Minimum (0.54)

Case A (Lowest DSC)
Error: Dominant over-segmentation
Detail: A large FP region encases the brainstem, and a long FP band runs along the spinal cord

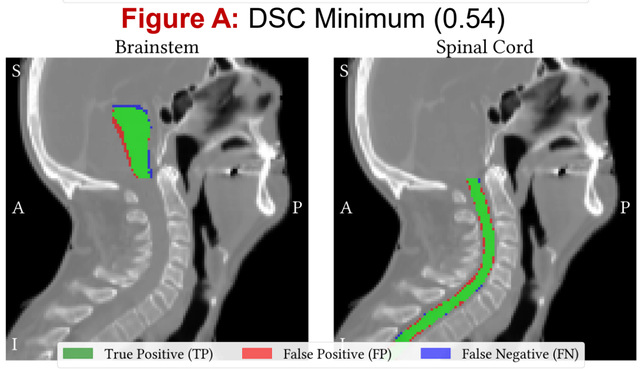

Figure B: DSC Median (0.86)

Case B (Median DSC)
Error: Good alignment with minor, systematic errors.
Detail: A thin, posterior FP band is visible on the spinal cord, with a small mixed FP/FN patch at the junction

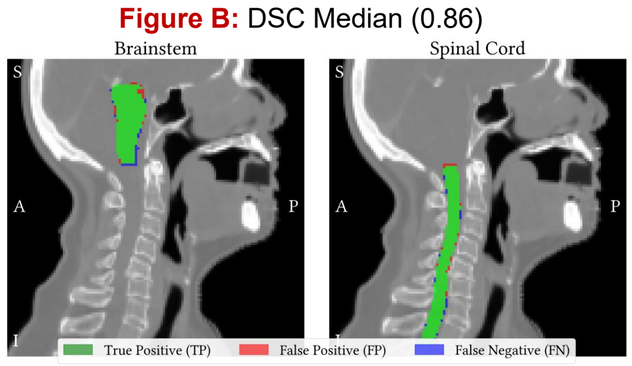

Figure C: DSC Maximum (0.92)

Case C (Highest DSC)
Error: Near-perfect alignment.
Detail: Only isolated, minor FP and FN patches at the boundaries, likely due to random image noise.

NSD Segmentation Examples

Qualitative Error Analysis

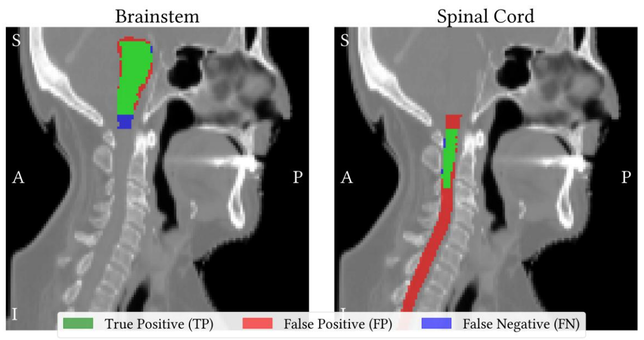

Figure D: NSD Minimum (0.70)

Case D (Lowest NSD):
Error: Major over-segmentation.
Detail: A large FP region on the superior brainstem and a long, wide FP band along the entire spinal cord.

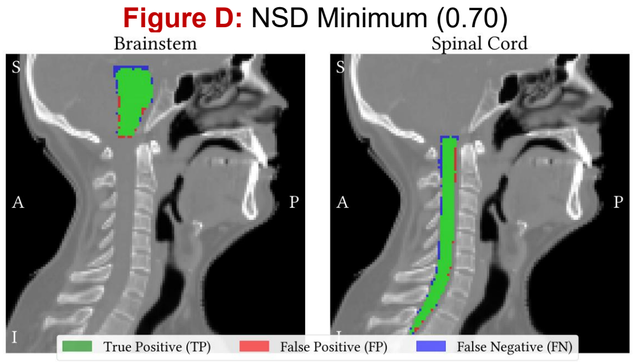

Figure E : NSD Median (0.99)

Case E (Median NSD):
Error: Localized under-segmentation.
Detail: A contiguous False Negative (FN) "cap" on the superior brainstem. The spinal cord is highly accurate with only sparse errors.

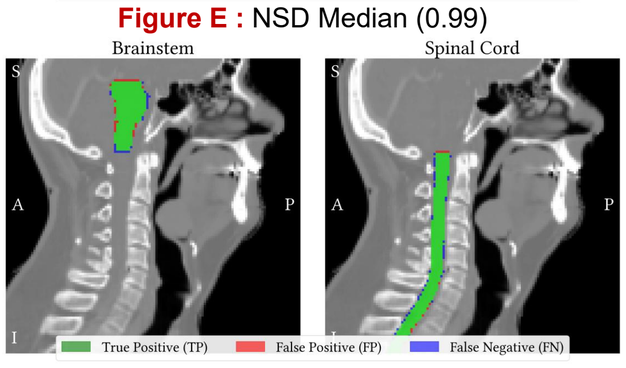

Figure F: NSD Maximum (1.00)

Case F (Highest NSD):
Error: Localized, thin errors.
Detail: A thin FN streak (under-segmentation) on the spinal cord and a small FP patch on the brainstem.

Overall Performance

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

Key Findings:
- The optimal model achieved good overall performance.
- The DSCs was similar, indicating model is equally good at identifying both organ regions.
- The NSDs was significantly higher spinal cord, likely due to the encasement of the spinal cord in vertebral canal, providing a clear, high-contrast boundary for the AI to learn.

Clinical Relevance

Assistive Tool Only: Mandatory expert review is critical to prevent severe toxicity in serial OARs.
Targeted Review: Uncertainty-guided review is needed to meet clinical efficiency standards (DTAC) [4].
Workflow Potential: Model can accelerate planning and reduce inter-observer variability.

Discussion & Limitations

Brainstem Errors: Low CT contrast creates ambiguous boundaries.
Spinal Cord Errors: Persistent posterior over-segmentation (False Positive) due to partial ground-truth labels in training data
Data Quality: RADCURE dataset (2000-2015) predates modern contouring standards [3], causing inconsistent ground truth and limiting model accuracy.
Clinical Risk: These are serial OARs. Any under-segmentation (False Negative) risks severe, irreversible toxicity.

Optimised Hyperparameters

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

Conclusions

- AI model shows strong but **inconsistent potential** for OAR delineation.
- Primary errors are from low CT contrast and **poor-quality/inconsistent ground-truth data**.
- Shows promise as an **assistive tool** to accelerate workflow and reduce variability.
- **Must NOT be used for autonomous segmentation** due to high clinical risk (especially False Negatives).
- **Mandatory expert review and modification** is required for all contours before clinical use.

Acknowledgments

With thanks to the Laidlaw Foundation for funding, and to Prof. McClelland, C. Dronne, and T. Zhang for their generous support!

References:
[1] Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation".
[2] Welch, M. L. et al. (2024). "RADCURE: An open-source head and neck cancer CT dataset for clinical radiation therapy insights".
[3] Brouwer,CharlotteL.etal.(Oct.2015).“CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCICCTG, NCRI, NRG Oncology and TROG consensus guidelines”. In: *Radiotherapy and* Oncology117, pp.83–90.doi:10.1016/j.radonc.2015.07.041.
[4] NICE (2023). Artificial intelligence technologies to aid contouring for radiotherapy treatment planning.