Head and Neck Gross Tumor Volume Deep-Learning Autocontouring Model Using BigROC Collaborative Data

Justin Hink on Behalf of the BigROC Collaboration

2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

ABSTRACT

Purpose: Accurate delineation of Head and Neck (H&N) Gross Tumor Volumes (GTV) is a prerequisite for effective radiotherapy, represents a significant cognitive challenge, and may contribute to an observed outcomes deficit between high-volume and low-volume H&N radiotherapy providers. While deep learning offers a potential solution, models require diverse, multi-institutional datasets to achieve clinical generalizability. This study evaluates the initial implementation of the BigROC collaboration dataset, a cloud-based repository designed to overcome the limitations of single-institutional cohorts.
Methods: The BigROC dataset is comprised of 338 cases from five institutions (UCLA, UCSF, UCSD, UCI, UCD), split into 305 training and 33 held-out test cases. We trained a 4-fold nnU-Net model and evaluated a mixture-of-experts (MoE) variant designed to handle heterogeneous, missing-modality multi-institutional inputs. Final predictions were generated via 4-fold ensemble averaging, and performance was assessed using per-case Dice Similarity Coefficient (DSC) against ground-truth contours on both the training and held-out test sets.
Results: The ensemble model achieved a mean training DSC of 0.34 (median 0.22, SD 0.33, n=305) and a mean test DSC of 0.27 (median 0.11, SD, 0.30, n=33). Dice distributions for both sets were strongly bimodal: a large cluster of near-zero DSC cases (indicating localization failure or missing GTVs) coexisted with a substantial subset of cases achieving DSC > 0.6, and several test cases exceeding 0.7-0.8. This bimodality, consistent across training and test sets, suggests the model performs well when it successfully localizes the tumor but fails outright in a meaningful fraction of cases, rather than producing uniformly mediocre contours.
Conclusion: Performance remains sharply divided between cases the pipeline localizes well (DSC > 0.6-0.8) and a substantial subset with near-zero DSC, and closing this gap is the primary barrier to clinical-grade H&N GTV autocontouring with the current ensemble pipeline. Characterizing the clinical and imaging features distinguishing failed and successful cases, along with further refinement of the mixture-of-experts approach for handling missing modalities, remains a central focus for improving generalizable performance. Continued case accrual from each participating institution is also a key next step, as expanding the size and diversity of the training set is likely necessary to reduce the rate of localization failures.

CONTACT

Justin Hink
University of California – Los Angeles
jhink@mednet.ucla.edu
(520) 334-5909

INTRODUCTION

The Problem: Accurate delineation of Head and Neck (H&N) Gross Tumor Volumes (GTV) is essential for effective radiotherapy planning, yet it remains a labor-intensive and highly variable task. When contouring the same tumor, different radiation oncologists can produce meaningfully different volumes. This variability is a likely contributor to the outcomes gap observed between high-volume and low-volume H&N radiotherapy centers. Deep learning offers a path toward faster, more consistent autocontouring, but models trained on data from a single institution tend to overfit to that institution's scanners, protocols, and patient population, and fail to generalize elsewhere.

The Solution: BigROC is a cloud-based, multi-institutional data collaboration built to overcome this single-site limitation. By centralizing H&N imaging and contours from five institutions (UCLA, UCSF, UCSD, UCI, and UCD) on AWS, BigROC enables model training on a dataset that is both larger and more heterogeneous than any single center could provide, better reflecting the diversity of scanners, protocols, and patient anatomy seen in real-world practice.

METHODS AND MATERIALS

Pre-processing:

- Anonymized all imaging data (CT, T1/T2 MRI, PET) prior to upload
- Uploaded data to a HIPAA-compliant Amazon Web Services (AWS) server, centralizing access for all five collaborating institutions
- Standardized GTV contour labels
- Resampled each case to isotropic 1.5mm resolution
- Co-registered all modalities to a common CT-based grid
- Filled missing modalities with a placeholder so every case could be input as a fixed four-channel (CT, T1, T2, PET) volume

Model Training:
Using the preprocessed dataset, we trained an nnU-Net segmentation pipeline with 4-fold cross-validation. To handle cases with missing modalities, this pipeline incorporates a mixture-of-experts (MoE) architecture. Final predictions were generated via ensemble inference across these folds. All preprocessing, training, and inference ran on AWS GPU instances, so this infrastructure is already in place to support continued data growth. An earlier MedNeXt-based model tested on a smaller subset of this data achieved comparable performance, suggesting the segmentation difficulty stems from the task itself rather than model choice.

Performance was measured using the Dice Similarity Coefficient (DSC) between predicted and ground-truth GTV contours, on both the training and held-out test sets (Table 1).

Table 1. Case Distribution by institution and split.

Institution	Training	Test	Total
UCLA	154	17	171
UCSF	82	9	91
UCSD	41	3	44
UCI	24	3	27
UCD	4	1	5
Total	305	33	338

RESULTS

The ensemble nnU-Net pipeline was evaluated on both the training set (n=305) and the independent held-out test set (n=33) using per-case Dice Similarity Coefficient (DSC).

On the training set, the ensemble achieved a mean DSC of 0.344 (median 0.219, SD 0.327). On the test set, the ensemble achieved a mean DSC of 0.273 (median 0.105, SD 0.303). In both cohorts, the distribution of DSC scores was highly polarized (Figure 1, Figure 2): a large cluster of cases fell near DSC = 0, indicating the model failed to localize the tumor entirely, while a separate cluster of cases achieved DSC > 0.6, with several test cases reaching 0.7-0.8. This pattern shows that when the model successfully identifies the tumor region, contour accuracy is high, but a substantial subset of cases result in outright localization failure rather than partial or degraded segmentation.

Figure 1: Ensemble Dice distribution, Training Set

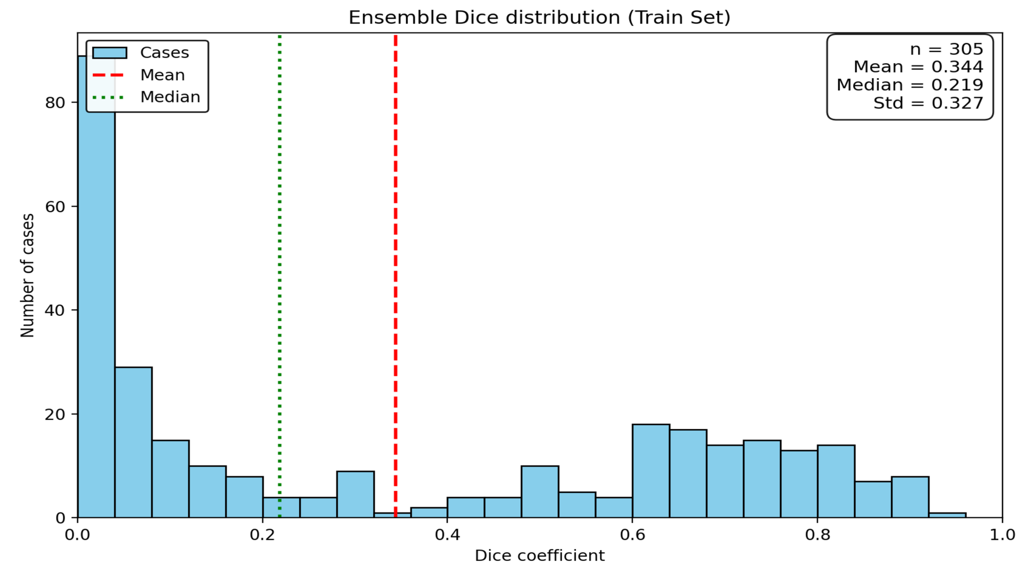
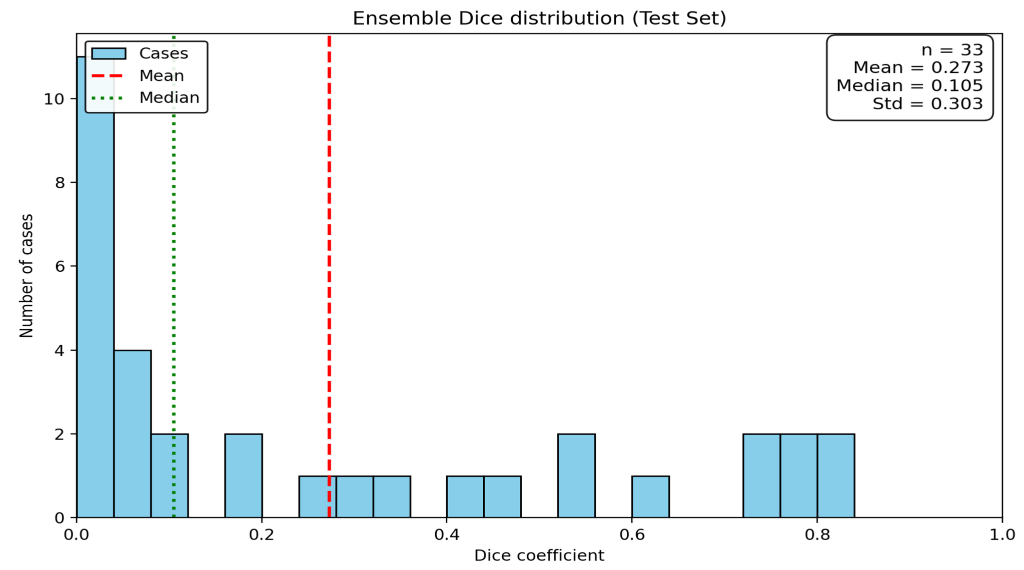
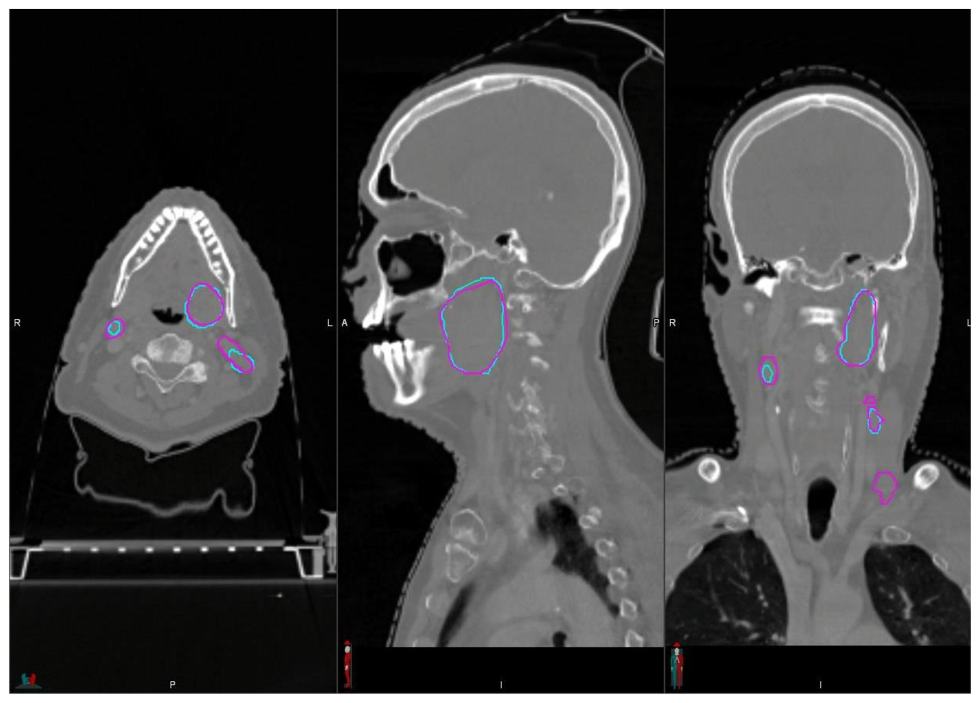


Figure 2: Ensemble Dice distribution, Test Set



To illustrate this best-case performance, Figure 3 shows a representative case from the test set where the model achieved strong overlap with the ground-truth contour (DSC = 0.826), demonstrating that the underlying architecture is capable of accurate GTV delineation when localization succeeds.

Figure 3: Representative high-performing case: auto-segmented contour (Light Blue) overlaid on the clinical ground truth (Magenta).



DISCUSSION

Successes:

- The BigROC infrastructure successfully centralizes anonymized, multi-modal imaging data from five institutions on a HIPAA-compliant AWS platform.
- The ensemble nnU-Net pipeline produces well aligned contours in a subset of cases (Figure 3), with close agreement between auto-segmented and clinical ground-truth GTVs across primary tumor and nodal targets.
- The mixture-of-experts architecture and placeholder-based preprocessing allow the model to accept cases with inconsistent modality availability, an essential feature for a growing, heterogeneous multi-institution dataset.

Challenges:

- DSC distributions reveal both training and test sets (Figure 1-2) either localized the tumor well or failed to find it at all, rather than producing consistently mediocre contours.
- This pattern suggests the dominant limitation is tumor localization, likely driven by anatomic variability, imaging heterogeneity, and inconsistent modality availability across institutions.
- A comparably difficult result from an earlier MedNeXt model on a smaller cohort suggests this is a challenge rooted in the data and clinical task, not a limitation specific to one architecture.

Future Directions:

- Characterize the clinical and imaging features that distinguish failed cases from successful ones (e.g., tumor size, location, modality availability, institution)
- Continue case accrual from each participating institution to improve anatomic and imaging diversity.
- Leverage the AWS training infrastructure already established in this study, which is ready to scale as additional institutions and cases join the collaboration.

CONCLUSIONS

BigROC establishes a working infrastructure for multi-institutional H&N GTV segmentation. The ensemble nnU-Net pipeline achieved a mean test DSC of 0.27, with well-localized cases reaching 0.7-0.8, demonstrating that accurate contours are achievable when the model correctly identifies the tumor. While clinical deployment remains a distant goal, closing the gap between these successful cases and the substantial fraction of outright failures is a feasible path toward state-of-the-art performance, and is the direct focus of our ongoing work.

REFERENCES

1. Isensee, F., Jaeger, P.F., Kohl, S.A.A. *et al.* nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* **18**, 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z

Author List

Yasin Abdulkadir[1], Brian M. Anderson, PhD[2], Emma Fields[3], **Justin Brady Hink**[1], Dylan Hsu[4], Kelly Kisling, PhD, MS[5], Andrew Krupien[6], James M. Lamb[1], Jason Matney, PhD[7], Peter G. Maxim, PhD[8], Lance C Moore, MS[5], Jack Neylon, PhD[1], Timothy Ritter, PhD[9], Ricky R Savjani, MD, PhD[10], Gilmer Valdes, PhD[11], Nuno Vasconcelos, PhD[5], Yevgeniy Vinogradskiy, PhD[12] and Alon Witztum, DPhil, MS[11], (1)Department of Radiation Oncology, University of California, Los Angeles, Los Angeles, CA, (2)University of California, San Diego, San Diego, CA, (3)Virginia Commonwealth University, Richmond, VA, (4)Memorial Sloan Kettering Cancer Center, New York, (5)University of California, San Diego, La Jolla, CA, (6)University of California, Los Angeles, Los Angeles, (7)UC Davis Cancer Center, Davis, CA, (8)University of California Irvine, Irvine, CA, (9)VCU Health System, Richmond, VA, (10)UCLA Radiation Oncology, Los Angeles, CA, (11)University of California San Francisco, San Francisco, CA, (12)Thomas Jefferson University, Philadelphia, PA