Automatic lesion and lymph node segmentation from PET and CT scans of the Head and Neck region: a HECKTOR 2025 Challenge Report

Sébastien Quetin[1,2] [0009−0005−5391−7319] and Shirin A. Enger[1,2,3]

[1] Medical Physics Unit, Department of Oncology, McGill University, Montreal, QC, Canada
[2] Montreal Institute for Learning Algorithms, Mila, Montreal, QC, Canada
[3] Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, QC, Canada

Abstract. The HECKTOR 2025 challenge provides a platform to benchmark automatic segmentation methods for Head and Neck (H&N) primary tumors and lymph nodes in FDG-PET and CT scans (Task 1). This study presents a challenge submission based on the nnU-Net framework. PET scans were first resampled to the CT resolution, after which a pseudo brain mask was derived from the PET scan to guide the cropping of both modalities. A custom clipping-based normalization was applied to the PET scan, and the paired PET and CT volumes were then processed by a Residual Encoder U-Net. Training was performed using custom augmentations. The proposed solution, submitted under the user name sebquet, achieved a mean Dice Score of 75.09% for primary tumors and 77.04% for metastatic lymph nodes on the validation set. These results demonstrate competitive performance within the challenge and highlight the effectiveness of combining modality-specific preprocessing with residual encoder architectures for H&N tumor segmentation.

Keywords: Head and Neck tumor · Deep Learning segmentation · PET and CT scans.

1 Introduction

Head and Neck (H&N) cancer ranks among the most common cancers worldwide and remains a major cause of cancer-related morbidity and mortality. Accurate tumor delineation is essential for treatment planning, particularly in radiotherapy, where precise localization directly impacts clinical outcomes. Positron Emission Tomography (PET) has proven to be a valuable modality for guiding H&N cancer management, as it provides metabolic information that complements the anatomical detail of Computed Tomography (CT) [3]. However, PET suffers from lower spatial resolution compared to CT, making the integration of both modalities crucial for reliable tumor and lymph node segmentation.

Several recent works have explored radiomics approaches based on PET and CT, with the goal of predicting prognosis in a non-invasive manner while utilizing routinely available diagnostic and treatment-planning scans [4,12,2]. While

radiomics approaches based on PET and CT have shown considerable promise, most studies to date have been conducted on relatively small patient cohorts. To ensure robust generalization and avoid overestimating predictive performance, validation on large, multi-institutional datasets is essential [1]. The HECKTOR 2025 challenge addresses this need by providing more than 1,200 PET/CT scans collected across multiple centers, enabling both reliable benchmarking and the development of clinically relevant models [10]. Three different tasks are proposed to tackle in the Hecktor 2025 challenge. This study addresses the first task which consists in segmenting the primary tumor volume (GTVp) and the metastatic lymph nodes (GTVn).

Building upon the HECKTOR 2025 dataset, we propose an nnU-Net–based approach [5,6] that integrates customized preprocessing strategies and augmentations to enable accurate automatic segmentation of tumor regions in H&N cancer patients.

2 Materials and Methods

The following methods describe the approach submitted to the test phase of the challenge. Alternative solutions previously explored during the validation phase are reported in the Results and Ablation sections.

2.1 Dataset

Our work was based solely on the HECKTOR 2025 training dataset [10] composed of 680 examinations from seven different centers with CT and registered PET scans along with manual expert contours for the GTVp and GTVn. No external dataset was used to develop our solution. Figure 1 shows a CT and a registered PET scan for a patient from the training dataset.

2.2 Data preprocessing

Resampling For training, ground truth masks and PET scans were resampled to the CT original resolution using a nearest neighbor and a Bspline interpolation, respectively, with the SimpleITK toolkit [8].

Cropping To train the deep learning model with the most informative signal, a preprocessing method was designed to crop the volumes around the head and neck region. The procedure consisted of two main steps: identifying the brain using PET Standardized Uptake Values (SUVs) and CT Hounsfiled Units (HUs) to define cropping in the x and y axes, and refining the z-axis cropping based on anatomical priors derived from the training dataset. Brain localization was achieved by first studying PET SUVs within the patient body, defined as values of PET scan voxels registered with CT scan voxels whose HUs were between –500 and 1000. The 95th percentile of these SUVs was selected as a threshold, and

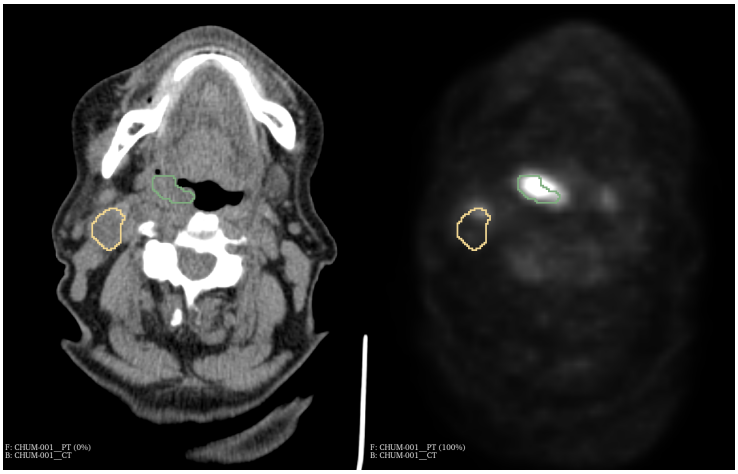

Fig. 1: CT scan and the corresponding registered PET scan from a patient in the HECKTOR 2025 training dataset, with manually annotated ground-truth masks overlaid. The GTVp contour is shown in green, and the GTVn contour is shown in yellow.

voxels above this value were masked in the resampled PET scan. The resulting high-SUV mask was then split into connected components; components smaller than 100 mm^3 were discarded. Among the remaining components, the one with the highest z-coordinate was selected as the brain, which avoided confusion with other high-uptake regions such as the bladder. Volumes were cropped above this brain component, and laterally around its x-y boundaries with a 50 mm margin.

To define the inferior boundary of the z-axis, we measured the distances between the top of the brain component and all ground-truth contours (GTVp and GTVn) across the training set. The maximum observed distance was used to crop along the z-axis below the brain, with an additional 10 mm margin for safety. The overall cropping process is illustrated in Figure 2.

2.3 Model Training

Normalization The CT scans were normalized using the nnU-Net CT normalization scheme. Foreground voxels, defined as voxels within regions where the ground-truth mask was non-zero, were used to compute the mean, standard deviation, and the 0.5^{th} and 99.5^{th} percentiles across the entire training dataset. Each CT scan was then normalized using these dataset-specific statistics by first clipping voxel intensities to the range defined by the 0.5th and 99.5th percentiles, followed by z-score normalization using the defined mean and standard deviation.

For the PET scans, the default nnU-Net image-based z-score normalization was not used. Instead, we applied the same dataset-specific approach as for the

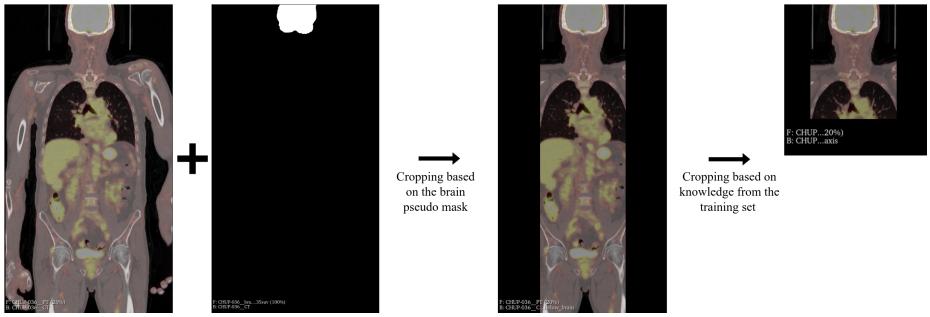

Fig. 2: Cropping process used in our pipeline. First, a pseudo mask is identified based on the PET SUVs and the volumes are cropped around a margin of this pseudo mask. Then the lower body is cropped based on the knowledge of the contour positions in the training set.

CT scans. Specifically, the mean and standard deviation of the foreground PET SUVs were estimated across the training dataset. However, instead of percentile-based clipping, custom thresholds were applied: SUV values were clipped to the range [0, 15] before z-score normalization.

Model architecture Although the benchmarking of different augmentations, clipping boundaries, and normalization strategies was conducted using a standard convolutional U-Net architecture (see Section 4), the final test-phase submission employed a Residual Encoder U-Net model. In both cases, the architectures were implemented using the nnU-Net framework with the 3D full-resolution configuration. For the Residual Encoder U-Net, the default parameters were modified by setting the GPU VRAM target to 16 GB, matching the hardware constraints of the HECKTOR 2025 challenge. Both U-Net variants were trained with six resolution levels comprising 32, 64, 128, 256, 320, and 320 feature maps.

Augmentations The default nnU-Net augmentation settings were strengthened by modifying both the augmentation probabilities and their application ranges. Specifically, the probability of applying rotations and scalings was increased from 0.2 to 0.5, while their parameter ranges remained unchanged. For Gaussian noise, the probability was increased from 0.1 to 0.25 for CT scans and to 0.5 for PET scans, with the variance left unchanged. The probability of the multiplicative brightness transform was raised from 0.15 to 0.75 for both modalities. For PET scans, the multiplier range was expanded from (0.75, 1.25) to (0.6, 1.4), while the default range was retained for CT scans. The contrast transform underwent the same changes as the multiplicative brightness transform.

Optimization For training, all volumes were resampled to the median CT voxel spacing of 0.98mm $\times$ 0.98mm $\times$ 3.27mm, as computed from the training dataset.

The plain convolutional U-Net model was trained with a patch size of 160 × 192 × 64 voxels, while the Residual Encoder U-Net used a larger patch size of 192 × 224 × 80 voxels. Default nnU-Net training hyperparameters were applied: a learning rate of 0.01, batch size of two, and a total of 1000 training epochs. Models were trained with five-fold cross-validation to minimize the Dice–Cross Entropy loss, each fold using 80% of the training set for optimization and 20% as an internal validation set. For each fold, the model checkpoint that maximized the exponential moving average of the pseudo-Dice score, computed by the nnU-Net framework on the internal validation set during training, was selected and used for inference.

2.4 Inference

At inference time, the PET scan was resampled to the CT scan as described in Section 2.2 followed by application of the cropping strategy outlined in Section 2.2. The distance to crop the inferior portion of the body in the z-axis was the one defined used in the training dataset.

Ensembling was performed using the five cross-validation models trained within the nnU-Net framework. For each fold, test-time augmentation was applied by flipping the input volumes along all three axes. The corresponding model outputs (logits) were then realigned by resetting the original axes and subsequently averaged. This procedure was repeated for the best-performing model from each fold. The mean logits across all five models were ultimately averaged and converted into segmentation masks using the argmax operation.

Finally, the predicted mask volume was padded back to match the original CT scan dimensions. No further post-processing was applied.

2.5 Evaluation

Challenge evaluation Two earlier versions of this work were submitted to the HECKTOR 2025 challenge platform for evaluation during the validation phase. These versions differed slightly from the methodology described in the previous sections. Specifically, both validation submissions employed the default nnU-Net augmentations rather than the customized augmentations described in Section 2.3, used a batch size of eight instead of two, and were based on the plain convolutional U-Net architecture.

Additionally, both submissions employed a different PET normalization strategy than that described in Section 2.3. In the first submission, PET scans were normalized using the nnU-Net CT normalization scheme, with clipping performed at the 0.5th and 99.5th percentiles of foreground voxels rather than using fixed boundaries. The second submission applied the default nnU-Net PET normalization, consisting of standard image-specific z-score normalization without any clipping.

The HECKTOR 2025 challenge platform evaluated our submission on a withheld testing set of 50 patients using the Dice score for the GTVp, an aggregated Dice score for the GTVn adapted from the Aggregated Jaccard Index [7], and a

custom aggregated F1-score for GTVn. For the latter, true positives (TP), false negatives (FN), and false positives (FP) were accumulated across correctly detected lesions, with a detection considered correct if the intersection-over-union (IoU) exceeded 30%.

Local evaluation To optimize the segmentation pipeline, the training dataset, already partitioned into five folds using an 80–20% split, was evaluated fold by fold on each respective internal validation set (n=136). Performance metrics were subsequently averaged across folds. Unlike the challenge submission, no ensemble predictions were performed.

To benchmark the trained models, the HECKTOR 2025 challenge evaluation metrics were replicated to the best of our ability; however, they may differ from the metrics computed by the official challenge platform. Segmentation performance for GTVp was assessed using the Dice score. For GTVn, both an aggregated Dice score and a custom aggregated F1-score were computed. For the aggregated Dice score, TP, FP, and FN were summed across all patients in the internal validation set of each fold, and the final metric was calculated using the standard Dice formula: $AggregatedDice = 2 * TP/(2 * TP + FP + FN)$.

For the GTVn F1-score, both the ground-truth and predicted masks were separated into connected components. For each predicted component, if its IoU with a ground-truth GTVn component exceeded 30%, the corresponding TP, FP, and FN were accumulated. If the IoU did not meet this threshold, the sum of voxels in the predicted component was added to the FP count. If a ground-truth component had no overlapping predicted component, the sum of its voxels was added to the FN count. The final metric was then computed using the same formula as for the Aggregated Dice.

3 Results

3.1 Challenge evaluation on the validation set

Table 1 summarizes the results of our two submissions to the HECKTOR 2025 challenge validation phase using earlier versions of our pipeline.

It can be seen that the normalization strategy applied to PET scans significantly affects the pipeline's performance on new patients. Specifically, using the nnU-Net CT normalization scheme for PET scans improves the Dice score for GTVp (75.09 vs 73.72%), whereas an image-based z-score normalization yields a higher aggregated Dice score for GTVn (77.86 vs 77.04%).

3.2 Local evaluation on internal validation sets

The two pipelines submitted to the validation phase were also evaluated locally. Table 1 reports the performance of each pipeline using a single fold, averaged across folds. Metrics for each fold were computed on the corresponding internal validation set, which included 136 patients. Overall, the locally computed metrics

Table 1: Results from our two submissions during the validation phase obtained directly from the HECKTOR 2025 challenge platform. Challenge metrics are averaged over the patients of the challenge validation set. Local metrics were computed over all patients in the internal validation set of each fold and then averaged across folds. Results are reported as mean ± standard deviation across fold averages, with the minimum and maximum fold values indicated in brackets [minimum maximum].

Submission	Evaluation	GTVp Dice Score (%)	GTVn Aggregated Dice Score (%)	GTVn Aggregated F1-score (%)
1 (CT normalization for PET scan)	Challenge	**75.09**	77.04	62.07
2 (z-score normalization for PET scan)	Challenge	73.72	**77.86**	62.07
1 (CT normalization for PET scan)	Local	**72.59** ± 1.47 [70.77, 75.1]	76.70 ± 0.67 [75.43, 77.25]	61.54 ± 3.41 [57.76, 66.44]
2 (z-score normalization for PET scan)	Local	72.21 ± 1.15 [70.75, 73.66]	**77.03** ± 0.61 [76.44, 78.11]	**62.22** ± 3.58 [56.93, 67.46]

generally correlated with the results obtained during the validation phase. In particular, a higher average Dice score across folds for GTVp, 72.59% against 72.21%, corresponded to a higher GTVp Dice score in the challenge, with a similar trend observed for the aggregated GTVn Dice score. For the aggregated F1-score, although our local implementation produced an average value close to that obtained in the challenge, higher values for one submission did not translate to a higher score with the challenge platform evaluation.

4 Ablation Study

To optimize our pipeline prior to submission for the HECKTOR challenge test phase, we empirically evaluated and benchmarked its components, resulting in the final configuration presented in Section 2. Unless otherwise specified, all results reported below were obtained using the plain convolutional U-Net described in Section 2.3. Performance metrics were computed locally and follow the definitions in Section 2.5.

Batch size and Augmentations Data augmentation strategies were benchmarked using the pipeline from our first validation-phase submission, which employed the nnU-Net CT normalization scheme for PET scans and a batch size of two instead of eight. Table 2 presents the performance of the pipeline on the internal validation set of fold 0, comparing default nnU-Net augmentations with our custom augmentation scheme. The custom augmentations resulted in lower Dice scores for GTVp with 74.90% against 75.20%, but improved performance for both the aggregated GTVn Dice score and the aggregated F1-score.

Although evaluation on the remaining four folds is needed to confirm this trend, the substantially stronger performance on the aggregated F1-score, from 58.65% to 61.64%, combined with the challenge timeline, motivated the selection of the custom augmentation strategy.

Table 2: Performance results of the pipeline using nnU-Net augmentations or our custom augmentations for the training. Results are averaged over all patients in the internal valdiation set of fold 0.

Augmentations used for training and inference	**GTVp Dice Score (%)**	**GTVn Aggregated Dice Score (%)**	**GTVn Aggregated F1-score (%)**
nnU-Net default	**75.20**	78.46	58.65
Custom	74.90	**78.84**	**61.64**

Figure 3 presents the training and validation losses for pipelines trained with custom augmentations and different batch sizes. Training with a batch size of two, compared to eight, produced a noisier validation loss curve suggestive of greater model exploration and ultimately resulted in a lower validation loss. Moreover, a batch size of two reduced overfitting, as indicated by the closer alignment between training and validation losses.

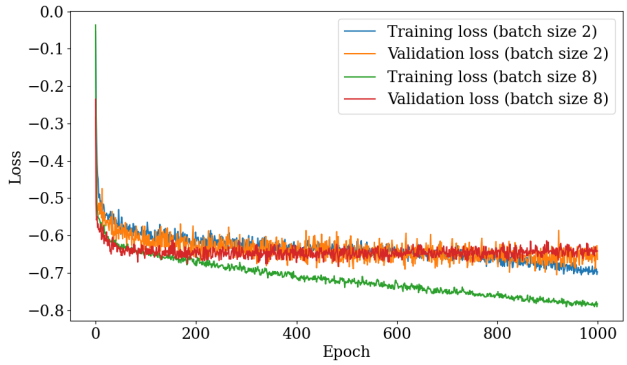

Fig. 3: Loss curves obtained when training our pipeline with different batch sizes.

Table 3 reports the pipeline performance averaged across the five folds, comparing training with nnU-Net default augmentations and a batch size of eight against training with custom augmentations and a batch size of two. The latter configuration produced substantially higher Dice scores for GTVp as well as improved aggregated Dice scores for GTVn, but at the cost of a markedly lower aggregated F1-score for GTVn. Based on these results, we selected the combina-

tion of custom augmentations and a batch size of two for subsequent experiments. All further ablation studies were conducted under this configuration.

Table 3: Performance results of a pipeline using different augmentations, batch sizes and spacings. Metrics were computed over all patients in the internal validation set of each fold and then averaged across folds. Results are reported as mean ± standard deviation across fold averages, with the minimum and maximum fold values indicated in brackets [minimum maximum].

Augmentations used for training and inference	**Batch size**	**Spacing**	**GTVp Dice Score (%)**	**GTVn Aggregated Dice Score (%)**	**GTVn Aggregated F1-score (%)**
nnU-Net default	8	Median	72.59 ± 1.47 [70.77, 75.1]	76.70 ± 0.67 [75.43, 77.25]	**61.54** ± 3.41 [57.76, 66.44]
Custom	2	Median	**74.70** ± 0.89 [73.77, 76.30]	**77.30** ± 1.00 [75.91, 78.84]	58.68 ± 5.90 [47.72, 64.95]
Custom	2	Isotropic 1mm	73.14 ± 2.0 [69.19, 74.47]	74.57 ± 4.4 [65.81, 77.59]	51.98 ± 15.78 [20.66, 63.48]

The effect of voxel spacing on model performance was benchmarked using a similar pipeline from our first validation-phase submission, which applied the nnU-Net CT normalization scheme to the PET scans, with a batch size of two instead of eight and with our custom augmentations. Motivated by previous work [9,11] reporting the use of an isotropic 1 mm resolution, we investigated resampling each CT scan to a 1 mm isotropic resolution, followed by resampling the corresponding PET and mask to this new CT resolution as a preprocessing. In this experiment, models were trained on isotropic 1 mm volumes, and at inference, the predicted masks were resampled back to the original CT resolution using nearest-neighbor interpolation. Table 3 summarizes the performance of the pipeline under both settings, averaged across folds. It can be seen that a model trained using the median CT spacing performs consistently better across the different metrics than a model whose inputs were resampled to a 1mm isotropic spacing.

Clipping As part of the ablation study, different empirically defined clipping boundaries for PET normalization were evaluated: [0–5], [0–8], [0–10], and [0–15]. These experiments were conducted using the custom augmentations described in Section 2.3 and a batch size of two. Table 4 reports the performance on the internal validation set of each fold, averaged across the five folds. A clear trend was observed for the GTVp Dice score: a stronger clipping consistently reduced the average performance. For the GTVn metrics, no consistent trend emerged; however, custom clipping with lower maximum boundaries than the 99.5th percentile appeared to improve the segmentation performance. Considering the trade-off between GTVp and GTVn performance, the [0–15] clipping range was selected for submission to the test phase.

Table 4: Performance results of a pipeline using a nnU-Net CT normalization for the PET scans with different clipping boundaries. Metrics were computed over all patients in the internal validation set of each fold and then averaged across folds. Results are reported as mean ± standard deviation across fold averages, with the minimum and maximum fold values indicated in brackets [minimum maximum].

Clipping boundaries	**GTVp Dice Score (%)**	**GTVn Aggregated Dice Score (%)**	**GTVn Aggregated F1-score (%)**
[0, 5]	74.43 ± 1.53 [72.67, 76.47]	**77.61** ± 1.05 [75.73, 78.66]	61.07 ± 6.27 [48.99, 66.89]
[0, 8]	74.69 ± 1.19 [72.79, 76.34]	77.14 ± 0.92 [75.5, 78.05]	59.36 ± 4.62 [51.51, 65.92]
[0, 10]	74.71 ± 1.33 [73.02, 76.06]	77.51 ± 0.73 [76.4, 78.46]	**61.88** ± 6.37 [50.72, 68.58]
[0, 15]	**74.80** ± 0.96 [73.26, 75.81]	77.54 ± 1.07 [75.62, 78.65]	61.49 ± 6.74 [49.02, 68.84]
[0.81, 22.66] (0.5, 99.5 percentile)	74.70 ± 0.89 [73.77, 76.3]	77.30 ± 1.00 [75.91, 78.84]	58.68 ± 5.9 [47.72, 64.95]

Model Architecture The final pipeline parameter evaluated was the model architecture. Both the Residual Encoder U-Net and the plain convolutional U-Net described in Section 2.3 were trained using the configuration detailed in Section 2. Table 5 summarizes the performance of the two architectures. Although the Residual Encoder U-Net resulted in a slight decrease in the average GTVp Dice score which went from 74.80% to 74.22%, it yielded improvements in the GTVn metrics. In particular, the aggregated F1-score increased from 61.49% to 63.59%, identifying the Residual Encoder U-Net as the stronger candidate for our test-phase submission.

5 Discussion

This study presents a pipeline submitted to the HECKTOR 2025 challenge for the automatic segmentation of primary tumor volumes and lymph nodes from CT and PET scans. We empirically evaluated various training strategies to identify those that improve segmentation performance. In particular, clipping, voxel spacing, batch size, and model architecture were found to have a substantial impact on the overall segmentation performance.

The learning rate was kept constant across experiments with different batch sizes, and the number of iterations per epoch was identical in all settings. Increasing the learning rate for a batch size of eight could potentially enhance the exploratory capacity of training and reduce the overfitting observed, thereby providing a more reliable comparison between batch sizes.

Table 5: Performance results of a pipeline using different model architecture. Results are averaged over all patients in the internal valdiation set of each fold. Metrics were computed over all patients in the internal validation set of each fold and then averaged across folds. Results are reported as mean ± standard deviation across fold averages, with the minimum and maximum fold values indicated in brackets [minimum maximum].

Model architecture	**GTVp Dice Score (%)**	**GTVn Aggregated Dice Score (%)**	**GTVn Aggregated F1-score (%)**
Plain Convolutional U-Net	**74.80** ± 0.96 [73.26, 75.81]	77.54 ± 1.07 [75.62, 78.65]	61.49 ± 6.74 [49.02, 68.84]
Residual Encoder U-Net	74.22 ± 1.55 [72.18, 76.56]	**78.18** ± 1.0 [76.94, 79.6]	**63.59** ± 4.83 [55.93, 69.83]

Our experiments indicate that resampling to 1 mm isotropic resolution does not improve performance, in contrast to the approach reported by [9]. A possible explanation is that the initial resampling excessively increased the PET resolution relative to the original spacing along the z-axis, which may have hindered effective feature learning during training.

The proposed segmentation task appears to exhibit high variability across different data splits, as indicated by the wide range of average metrics across folds. This variability may explain why ensembling numerous models, as performed by the HECKTOR 2022 winning team [9], is beneficial for improving performance.

Acknowledgments. This work was supported by the Canada Research Chair Program (grant #252136) and partly supported by Mitacs through the Mitacs Accelerate program. This research was partly enabled by support provided by Calcul Quebec https://www.calculquebec.ca/en/ and the Digital Research Alliance of Canada https://alliancecan.ca/. Deep learning models were trained on the Narval cluster. We thank Yujing Zou for valuable discussions.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. HEad and neCK TumOR Lesion Segmentation, Diagnosis and Prognosis - Grand Challenge, https://hecktor25.grand-challenge.org/
2. Bogowicz, M., Riesterer, O., Stark, L.S., Studer, G., Unkelbach, J., Guckenberger, M., Tanadini-Lang, S.: Comparison of PET and CT radiomics for prediction of local tumor control in head and neck squamous cell carcinoma. Acta Oncologica **56**(11), 1531–1536 (Nov 2017). https://doi.org/10.1080/0284186X.2017.1346382, https://medicaljournalssweden.se/actaoncologica/article/view/25070

3. Caldarella, C., De Risi, M., Massaccesi, M., Micciché, F., Bussu, F., Galli, J., Rufini, V., Leccisotti, L.: Role of 18f-fdg pet/ct in head and neck squamous cell carcinoma: Current evidence and innovative applications. Cancers **16**(10) (2024). https://doi.org/10.3390/cancers16101905, https://www.mdpi.com/2072-6694/16/10/1905
4. Castelli, J., Depeursinge, A., Ndoh, V., Prior, J., Ozsahin, M., Devillers, A., Bouchaab, H., Chajon, E., De Crevoisier, R., Scher, N., Jegoux, F., Laguerre, B., De Bari, B., Bourhis, J.: A PET-based nomogram for oropharyngeal cancers. European Journal of Cancer **75**, 222–230 (Apr 2017). https://doi.org/10.1016/j.ejca.2017.01.018, https://linkinghub.elsevier.com/retrieve/pii/S095980491730076X
5. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (Feb 2021). https://doi.org/10.1038/s41592-020-01008-z, https://www.nature.com/articles/s41592-020-01008-z
6. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. arXiv preprint arXiv:2404.09556 (2024)
7. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Transactions on Medical Imaging **36**(7), 1550–1560 (2017). https://doi.org/10.1109/TMI.2017.2677499
8. Lowekamp, B.C., Chen, D.T., Ibáñez, L., Blezek, D.: The Design of SimpleITK. Frontiers in Neuroinformatics **7**, 45 (2013). https://doi.org/10.3389/fninf.2013.00045
9. Myronenko, A., Siddiquee, M.M.R., Yang, D., He, Y., Xu, D.: Automated head and neck tumor segmentation from 3d pet/ct hecktor 2022 challenge report. In: Andrearczyk, V., Oreiller, V., Hatt, M., Depeursinge, A. (eds.) Head and Neck Tumor Segmentation and Outcome Prediction. pp. 31–37. Springer Nature Switzerland, Cham (2023)
10. Saeed, N., Hassan, S., Hardan, S., Aly, A., Taratynova, D., Nawaz, U., Khan, U., Ridzuan, M., Andrearczyk, V., Depeursinge, A., Hatt, M., Eugene, T., Metz, R., Dore, M., Delpon, G., Papineni, V.R.K., Wahid, K., Dede, C., Ali, A.M.S., Sjogreen, C., Naser, M., Fuller, C.D., Oreiller, V., Jreige, M., Prior, J.O., Rest, C.C.L., Tankyevych, O., Decazes, P., Ruan, S., Tanadini-Lang, S., Vallières, M., Elhalawani, H., Abgral, R., Floch, R., Kerleguer, K., Schick, U., Mauguen, M., Rahmim, A., Yaqub, M.: A Multimodal and Multi-centric Head and Neck Cancer Dataset for Tumor Segmentation and Outcome Prediction (Sep 2025). https://doi.org/10.48550/arXiv.2509.00367, http://arxiv.org/abs/2509.00367, arXiv:2509.00367
11. Sun, X., An, C., Wang, L.: A coarse-to-fine ensembling framework for head and neck tumor and lymph segmentation in ct and pet images. In: Andrearczyk, V., Oreiller, V., Hatt, M., Depeursinge, A. (eds.) Head and Neck Tumor Segmentation and Outcome Prediction. pp. 38–46. Springer Nature Switzerland, Cham (2023)
12. Vallières, M., Kay-Rivest, E., Perrin, L.J., Liem, X., Furstoss, C., Aerts, H.J.W.L., Khaouam, N., Nguyen-Tan, P.F., Wang, C.S., Sultanem, K., Seuntjens, J., El Naqa, I.: Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Scientific Reports **7**(1), 10117 (Aug 2017). https://doi.org/10.1038/s41598-017-10371-5, https://www.nature.com/articles/s41598-017-10371-5