This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

A Late-Fusion Framework for Joint Tumor Segmentation, Staging, and Recurrence Prediction in Head and Neck Cancer

Janos Szalma[1 [0000−0003−4167−1858]], Adam Juhasz[1 [0009−0001−2801−9956]], Joel Schaerer[1], Emily Ebert[1], Robert Scarimbolo[1], Anu Bansal[1], and Alessandro Delmonte[1 [0000−0001−8724−9535]]

Clario, a part of Thermo Fisher Scientific, Philadelphia, PA 19103, USA
{first}.{last}@clario.com

Abstract. Accurate assessment of tumor burden, disease staging at diagnosis, and treatment outcome prediction is critical for personalized management of head and neck cancer. Recent advances in artificial intelligence have shown promise in complementing traditional prognostic models. However, clinical adoption is limited by the lack of standardized benchmarks and unified end-to-end solutions. In this work, we present a unified framework for joint tumor segmentation, TN staging, and recurrence-free survival prediction from ^{18}F-FDG PET/CT imaging and clinical data. A 3D residual U-Net was first trained to segment primary tumors and adenopathies, providing a shared imaging backbone from which patient-level feature sets were derived for TN staging and outcome prediction. To improve precision in detecting metastatic lymph node involvement, nodal-specific optimization strategies were incorporated during training. Imaging and clinical features were subsequently combined through a late-fusion strategy, leveraging their complementary information to improve both staging and survival prediction. The proposed framework was evaluated as part of the MICCAI HECKTOR 2026 challenge.

Keywords: Head and Neck Cancer · Deep Learning · Tumor Segmentation · Risk Stratification · Nuclear Imaging

1 Introduction

Head and neck cancer (HNC) remains a major cause of cancer-related mortality worldwide [1], and its management requires a multidisciplinary approach that may involve surgery and radiation therapy.

Quantification of primary and nodal tumor burden, disease staging and clinical outcome prediction provide complementary information that can support risk stratification and personalized treatment pathways [2].

In clinical studies, ^{18}F-FDG PET/CT is a key imaging modality for HNC screening and evaluation [3,4]. Accurate delineation of primary tumors and metastatic lymph nodes is critical for radiation therapy planning, yet remains

challenging due to the complex anatomy and coexistence of malignant and benign uptake patterns [5].

Beyond tumor burden evaluation, predicting patient outcomes such as recurrence-free survival (RFS) remains difficult. While traditional prognostic models primarily rely on clinical and pathological variables, recent studies suggest that imaging biomarkers derived from PET/CT may provide complementary information for risk stratification and treatment response assessment [6,7].

Recently, multiple artificial intelligence (AI) approaches have been proposed for HNC risk stratification and outcome prediction [8,9,10]. However, clinical translation remains limited by the lack of large annotated cohorts and standardized benchmarking frameworks for evaluating reproducibility and robustness across diverse patient populations. Furthermore, clinical workflows require simultaneous assessment of tumor burden, radiological staging, and prognosis, whereas existing approaches typically address these tasks independently.

The HEad and neCK TumOR (HECKTOR) challenge [11], organized in conjunction with MICCAI, addresses these limitations by providing a large multi-institutional PET/CT dataset comprising 1,423 patients with corresponding clinical variables and imaging annotations.

Building on previous editions, HECKTOR 2026 introduces radiological TN staging, placing increased emphasis on nodal disease burden. By jointly evaluating three tasks including tumor segmentation, TN staging, and RFS prediction, the challenge promotes a common benchmark for end-to-end frameworks that closely reflect clinical workflows [12].

To this end, we propose an end-to-end multi-stage framework for segmentation, staging and prognosis, built around a shared PET/CT AI backbone. In this work, we introduce a multi-modal residual U-Net to segment primary tumors (GTVp) and metastatic lymph nodes (GTVn). The resulting imaging backbone is then used to derive imaging features that are combined with quantitative biomarkers and clinical variables for TN-stage classification and RFS prediction.

By proposing a unified framework based on a common PET/CT feature extractor, we investigate whether a shared segmentation representation can support both staging and prognosis estimation for HNC screening.

2 Methods

The proposed framework consists of a shared PET/CT representation obtained through deeply supervised tumor segmentation. This representation is subsequently reused for TN staging and RFS prediction via three lightweight task-specific predictive models (Fig. 1), enabling joint modeling of tumor burden and treatment outcome.

First, a 3D residual U-Net was trained on co-registered PET/CT volumes to segment primary tumors and metastatic lymph nodes. Multi-scale encoder features were extracted and combined with quantitative radiomics descriptors derived from the estimated tumor masks and available clinical variables. The resulting feature vectors were used to train independent predictors for T stage, N

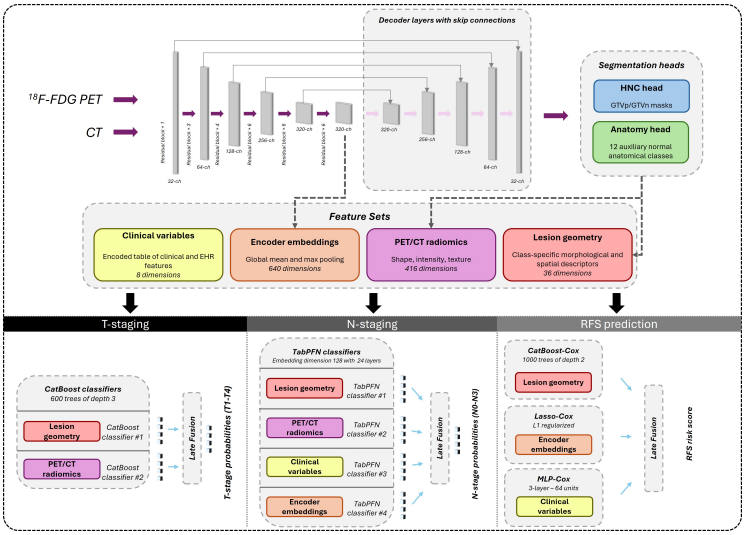

Fig. 1. Overview of the proposed framework. A shared segmentation backbone is used to segment primary tumors and metastatic lymph nodes and to extract imaging features. Feature sets derived from the backbone and clinical variables are combined through late-fusion for TN staging and RFS prediction.

stage and RFS, combining imaging embeddings with structured clinical variables, using dedicated predictive models supporting both categorical and continuous tabular inputs.

2.1 Tumor Segmentation Network

The segmentation backbone follows a 3D U-Net architecture [13] trained on co-registered ^{18}F-FDG PET/CT volumes. We used a residual encoder design inspired by nnU-Net [14], in which each encoder stage consists of stacked residual blocks with convolution, batch normalization, and ReLU activations. Residual identity mappings facilitate feature propagation across scales and improve optimization stability when learning deeper volumetric representations.

Accurate delineation of primary tumors and metastatic lymph nodes in HNC is challenging due to complex regional anatomy and highly heterogeneous physiological tracer uptake. To provide additional anatomical context, we introduced an auxiliary segmentation objective for normal head and neck structures. Twelve anatomical regions were considered, including the larynx, mandible, skull base, carotid complex, nasopharynx, cervical spine, internal jugular veins, parotid glands, sub-mandibular glands, hard palate, and epiglottic complex. These structures were selected based on their proximity to primary or nodal disease and their potential overlap with regions of physiological tracer uptake.

Pseudo ground-truth labels for the auxiliary structures were generated automatically using a multimodal promptable 3D segmentation foundation model [15].

The model was prompted using textual descriptions of each anatomical region of interest. To improve label reliability, prompts were evaluated across the full cohort, and structures detected in fewer than 80% of cases were excluded. Qualitative visual inspection confirmed the anatomical plausibility of the generated labels, although no manual corrections were performed. The final segmentation model was trained by jointly optimizing tumor and auxiliary anatomical reference segmentation.

To improve precision on small metastatic lymph nodes, which are central to TN staging while contributing only marginally to overall tumor burden, we incorporated a supplementary focal Tversky loss term [16] (Eq. 1) into the optimization objective.

$$\mathcal{L}_{Tversky} = \frac{1}{|\mathcal{C}_{fg}|} \sum_{c \in \mathcal{C}_{fg}} \left(1 - \frac{TP_c}{TP_c + \alpha_c FP_c + \beta_c FN_c}\right)^{\gamma} \tag{1}$$

where α_c and β_c control the relative penalties assigned to false positives and false negatives, respectively, γ denotes the focal exponent and $C_{fg} = \{\text{GTVp}, \text{GTVn}\}$ denotes the set of foreground classes.

Model training was initialized from the winning solution of the AutoPET III challenge [17], which was pretrained on 1,611 whole-body ^{18}F-FDG PET/CT scans from lymphoma subjects.

Finally, to improve robustness to PET/CT registration variability, we incorporated an inter-modality misalignment augmentation strategy [18]. During training, random spatial perturbations were applied independently to the PET and CT volumes, simulating imperfect registration caused by acquisition variability or subject motion.

2.2 TN Staging and Recurrence-Free Survival Prediction

Latent features from the segmentation backbone, together with radiomics descriptors and available clinical variables, were integrated for TN staging and RFS prediction. Multiple patient-level feature sets were considered (Table 1) and evaluated independently to quantify their individual contribution to the final prediction tasks.

Different predictive models were evaluated for TN staging and RFS prediction. To leverage the complementary information captured by the different feature sets, a late-fusion strategy was adopted throughout. Independent predictors were trained on each selected feature set and their outputs were subsequently aggregated by averaging class probabilities.

For TN staging, CatBoost [19], MLPs, and TabPFN [20] were evaluated. The selected T-stage model was a CatBoost classifier using lesion geometry and PET/CT radiomics, whereas the selected N-stage model was an ensemble of four TabPFN classifiers trained on encoder embeddings, PET/CT radiomics, lesion geometry, and clinical variable feature sets.

For RFS prediction, a CatBoost-Cox classifier was trained on lesion geometry, an MLP-Cox model was trained on clinical variables while a Lasso-Cox [21] model

Feature Set	D	Description
PET/CT radiomics	416	Features extracted from the predicted tumor masks including GLCM, GLRLM, GLSZM and NGTDM statistics
Encoder embeddings	640	Concatenation of 320 features from global average pooling + 320 features from global max pooling at inner-most (5^{th}) layer.
Lesion geometry	36	For both GTVp and GTVn: binary presence; volume (vox and mm^3); count; largest lesion size and fraction; bounding-box size (x,y,z), diagonal, centroid (x,y,z) and maximum axis length. Cross-lesion features: GTVn/GTVp volume ratio; centroid distance and offsets (x,y,z); nodal fraction of total tumor burden, ≥ 2 and ≥ 3 nodal lesion.
Clinical variables	8	Age, gender, tobacco use, alcohol use, ECOG performance status, treatment, HPV status, and recruiting center.

Table 1. Input feature sets considered for patient-level TN staging and outcome prediction. D denotes feature vector dimensionality.

was trained on the encoder embeddings. The final model combined predictors by averaging their z-score normalized risk estimates.

3 Experimental Settings

3.1 Dataset

Experiments were conducted on the HECKTOR 2026 challenge dataset [22], comprising 782 patients from eight institutions with co-registered ^{18}F-FDG PET/CT scans and associated clinical data. An additional 50 and 200 subjects were provided for independent validation and testing, respectively.

Ground-truth segmentation of primary tumors (GTVp) and metastatic lymph nodes (GTVn) were provided, together with demographic, clinical, staging, and outcome variables.

Segmentation and N-stage labels were available for all patients, T-stage labels for 778 patients, and valid RFS data for 727 patients, including 148 recurrence events (20.4%). Following the challenge protocol, T0 and T1 categories were merged, resulting in 4 classes for both T and N staging.

The training dataset was partitioned into five folds for cross-validation, with each iteration using 80% of the data for training and 20% for validation.

3.2 Image pre-processing

Prior to training, PET images were resampled to the corresponding CT using third-order B-spline interpolation.

The head-and-neck region was automatically localized using anatomical landmarks detection. Landmarks were extracted from a region of interest comprising

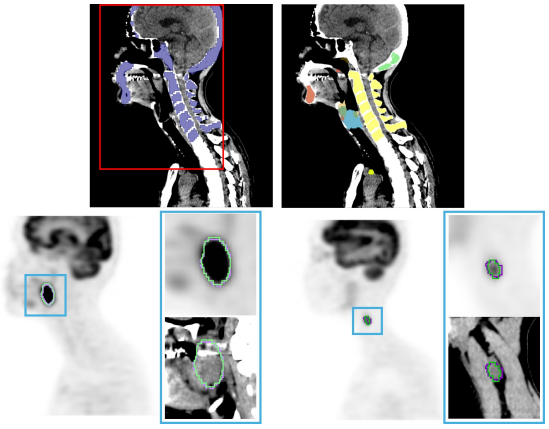

Fig. 2. Representative pre-processing and segmentation results from a HNC subject. Top row: (left) skeletal segmentation and derived ROI used for cropping; (right) auxiliary anatomical structures used during training. Bottom row: primary tumor (left) and metastatic lymph node (right) segmentation results. Ground-truth contours are shown in purple and predictions in green.

the skull and cervical vertebrae, which was automatically defined using a previously trained segmentation network [14,23]. A bounding box around these structures was expanded by 50 mm inferiorly, 30 mm anteriorly, and 50 mm laterally, and subsequently used to crop both modalities (Fig. 2).

CT intensities were clipped to $[-300, 300]$ HU and PET volumes were z-score normalized. The cropped images were resampled to a resolution of $1.0 \times 1.0 \times 3.3\,\mathrm{mm}^3$ and processed using patches of size $64 \times 224 \times 192$. Due to inference memory constraints, a small subset of acquisitions ($< 5\%$) exceeding 64×10^6 voxels underwent additional center cropping. As cropping was applied after head-and-neck localization, the region of interest was preserved.

3.3 Implementation Details

The model training was performed using PyTorch 2.13 on a 24 GB NVIDIA RTX 4090.

Segmentation network The segmentation network employed stage widths of [32, 64, 128, 256, 320, 320] with [1, 3, 4, 6, 6, 6] residual blocks per stage.

The model was optimized using SGD with Nesterov momentum ($m = 0.99$), weight decay ($3\times^{-5}$) and polynomial learning-rate decay. Training was performed for 200 epochs with a batch size of 8. The final objective function combined Dice, cross-entropy, and Tversky losses (Eq. 2)

$$\mathcal{L}_{total} = \mathcal{L}_{Dice} + \mathcal{L}_{CE} + \lambda_{anat}\,\mathcal{L}_{anat} + \lambda_{tv}\,\mathcal{L}_{Tversky} \tag{2}$$

where $\mathcal{L}_{Dice}$ and $\mathcal{L}_{CE}$ are the soft Dice and cross-entropy terms on GTVp/GTVn, $\mathcal{L}_{anat}$ is a combined Dice and cross-entropy loss on the auxiliary normal anatomical labels ($\lambda_{anat} = 0.2$). The Tversky factor ($\lambda_{tv} = 0.05$) was introduced to increase precision for nodal lesions ($\alpha = 0.7$, $\beta = 0.3$, $\gamma = 1.33$).

During inference, predictions from the five cross-validation folds were ensembled by averaging following sliding-window inference and test-time augmentation.

Predicted masks were subsequently refined using class-specific post-processing rules. For primary tumors (GTVp), only the largest connected component was kept if its volume exceeded $50\,\text{mm}^3$ and its confidence was at least 0.3. For nodal disease (GTVn), up to six connected components larger than $20\,\text{mm}^3$ were retained.

T-stage CatBoost classifiers, consisting of 600 trees of depth 3 ($lr = 0.06$ with early stopping after 80 rounds), with tempered class weights ($\beta = 0.5$), were trained from geometry and radiomics features.

N-stage Four TabPFN classifiers with an embedding dimension of 128 and 24 in-context-learning layers were trained for 100 epochs on the combination of imaging and clinical features. All TabPFN hyperparameters were fixed according to the original implementation.

RFS CatBoost-Cox models ($depth = 2$, $N_{trees} = 1000$, $lr = 0.03$) were trained on lesion geometry. Lasso-Cox models were trained on features from the shared encoder bottleneck. Finally, a 3-layer MLP with 64 units was trained on the clinical features using Cox partial likelihood.

A late-fusion strategy was adopted for TN staging and RFS prediction. Independent predictors were trained on complementary feature groups and their outputs were aggregated by averaging class probabilities.

3.4 Evaluation Metrics

Each task was evaluated independently and aggregated following HECKTOR 2026 weighting. For segmentation, the Dice similarity coefficient (DSC) was selected for model evaluation using a 5-fold cross-validation strategy (Eq. 3).

$$\text{Dice}(y_{\text{true}}, y_{\text{pred}}) = \frac{2 \sum (y_{\text{true}} \cdot y_{\text{pred}})}{\sum y_{\text{true}} + \sum y_{\text{pred}}} \tag{3}$$

where y_{true} and y_{pred} denote the ground-truth and predicted segmentation masks, respectively. Following the HECKTOR 2026 evaluation protocol, only an aggregate DSC across GTVp and GTVn was available for the independent validation and test cohorts, precluding separate assessment of primary tumor and nodal segmentation performance.

Balanced accuracy (ACC) was implemented for TN staging, while C-index (Eq. 4) was evaluated for RFS prediction.

$$\text{C-index} = \frac{\sum_{i,j} \mathbf{1}[r_i > r_j] \cdot \mathbf{1}[t_i < t_j] \cdot e_i}{\sum_{i,j} \mathbf{1}[t_i < t_j] \cdot e_i} \tag{4}$$

where r_i and r_j denote the predicted recurrence risk scores for subjects i and j, respectively, t_i and t_j are the corresponding recurrence-free times, e_i is an event indicator equal to 1 if recurrence was observed and 0 otherwise. To account for the limited number of observations C-index was only computed after concatenating all 5-fold test set predictions.

4 Results

4.1 Task Ablation

Tumor Burden Segmentation For primary tumors and adenopathies segmentation, we evaluated the contribution of pre-training, loss design, and hyperparameter selection (Fig. 3).

Pre-training provided the largest individual improvement, increasing mean DSC from 66.8% to 67.9% with respect to the model trained from scratch.

The nodal Tversky term improved mean DSC to 68.6% and further increased performance to 69.0% when combined with larger-batch training. Fold-wise analysis showed that the Tversky loss increased GTVn precision from 67.1% to 69.8%, with a limited decrease in recall from 75.8% to 73.3%. This reduced the proportion of false-positive nodal detections from 79.3% to 67.8%, mitigating the major source of N-staging errors.

Increasing the batch size improved performance, raising the mean DSC from 67.96% to 68.69% when moving from gradient accumulation ($B = 2$) to a batch size of 8. In contrast, neither larger field-of-view ($96 \times 256 \times 224$) nor longer training schedules resulted in measurable improvements. The individual contribution of auxiliary anatomical supervision was not evaluated and should be investigated in future ablation studies.

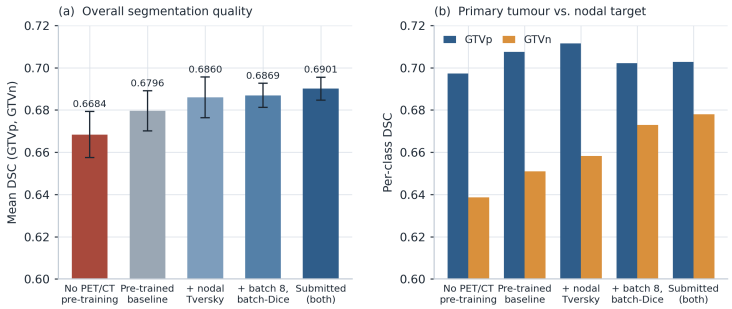

Fig. 3. Segmentation ablation study. Left: overall performance. Right: contribution of each configuration to nodal and non-nodal lesions.

TN staging and RFS prediction To quantify the contribution of each feature group, we evaluated individual feature sets independently across T-stage prediction, N-stage prediction, and RFS prediction (Fig. 4).

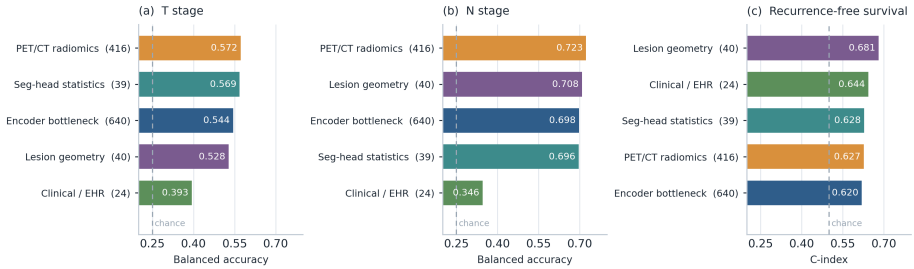

Fig. 4. Ablation of patient-level feature sets for T-staging (a), N-staging (b), and RFS prediction (c). Each result corresponds to a model trained on a single feature set.

For TN staging, PET/CT radiomics achieved the highest balanced accuracy among individual feature sets, reaching 57.2% and 72.3%, respectively. Clinical variables resulted in the lowest performance, obtaining accuracies of 39.3% for T staging and 34.6% for N staging (Fig. 4).

In contrast, for RFS, lesion geometry obtained the highest C-index (0.682), followed by clinical variables (0.644), PET/CT radiomics (0.628), and encoder embeddings (0.620).

Agreement between models trained on a single feature set averaged 57% for T-staging and 69% for N-staging, and RFS showed moderate correlation ($\rho = 0.35$). Clinical variables were the most independent source of variation.

To leverage this complementarity, we compared early- and late-fusion strategies. Early fusion concatenated all feature sets into a single vector used to train one predictive model. In contrast, late fusion was implemented by training an independent predictor for each feature set and then averaging class probabilities for TN staging or risk scores for RFS prediction.

For T-staging, balanced accuracy increased from 57.2% for the best individual feature group to 57.9% with early fusion and 59.4% with late fusion (Fig. 5). For N-staging, performance improved from 72.3% to 74.0% and 75.1%, respectively. The largest improvement was observed for RFS prediction, where late fusion increased the C-index from 0.682 for the best individual feature group and 0.680 for early fusion to 0.704 for late-fusion.

4.2 Final Configuration

The final configuration was selected based on five-fold OOF performance and subsequently evaluated on the independent validation and test sets (Table 2).

The proposed framework achieved a validation DSC of 65.9%, TN staging ACC of 62.0% (T-stage: 44.1%, N-stage: 80.0%), and an RFS C-index of 0.69,

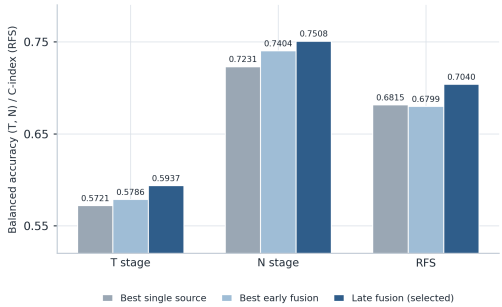

Fig. 5. Single-set, early- and late-fusion impact on TN staging and RFS prediction.

resulting in an overall score of 0.658. On the test set, the model achieved a DSC of 65.3%, TN staging ACC of 55.7%, and an RFS C-index of 0.6445, corresponding to an overall score of 0.6158. These results ranked among the top-performing solutions in the HECKTOR 2026 challenge across all tasks.

Component	Metric	5-fold OOF	Validation	Test	Weighting
Segmentation	$DSC_{GTVp,GTVn}$	0.6901	0.6586	0.6527	0.25
TN staging	Balanced accuracy				0.35
T stage		0.5718	0.4411	0.5058	
N stage		0.7508	0.7998	0.6074	
TN staging		0.6613	0.6204	0.5566	
RFS prediction	C-index	0.7003	0.6893	0.6445	0.40
Overall score		0.6841	0.6575	0.6158	

Table 2. Results of the selected model on the five-fold OOF, independent validation, and test sets. The overall score represents the weighted aggregation of individual tasks.

5 Conclusion

In this work, we presented a unified framework for joint tumor segmentation, TN staging, and recurrence-free survival prediction from PET/CT imaging and clinical data.

The proposed method achieved competitive performance across all tasks, qualifying in the top HECKTOR 2026 solutions and demonstrating that segmentation-derived representations, quantitative imaging biomarkers, and clinical variables provide complementary information for staging and outcome prediction.

Segmentation improvements were primarily driven by strategies targeting metastatic lymph node delineation rather than increased model capacity. While transfer learning provided the largest performance gain, additional improvements

were obtained through nodal-specific optimization, anatomical supervision, and PET/CT misalignment augmentation. These findings highlight the importance of accurate lymph node segmentation in head and neck cancer and support the inclusion of adenopathy assessment as an important component of tumor burden evaluation.

Using imaging features for staging and outcome prediction revealed strong differences in the information captured by individual feature sets. PET/CT radiomics, encoder embeddings, lesion geometry and clinical features showed moderate correlation, indicating that each set contributes complementary information. Consistent with these observations, a late fusion approach combining predictions from expert models consistently outperformed early feature concatenation by a single model.

T-stage prediction remained more challenging than N-stage prediction. This may reflect the greater complexity of T-staging, which depends on tumor location and invasion of adjacent anatomical structures, whereas N staging is determined by objective measures of nodal disease burden [24]. Consequently, T-staging may be more susceptible to data imbalance and overfitting.

Future work will further investigate weighting strategies for late fusion while maintaining a unified framework for HNC risk stratification.

Acknowledgements The authors thank the organizers of the HECKTOR 2026 challenge for providing the dataset and a collaborative framework. This work has been funded by Clario, a part of Thermo Fisher Scientific.

Disclosures of Interests The authors declare no competing interests relevant to the content of this article.

References

1. Heinolainen, A., Nguyen, B., Silén, S., Renkonen, R., Koskinen, M.: Survival and data-driven phenotypes in head and neck cancer. *Scientific Reports* **15**(1), 5985 (2025)
2. Anderson, G., Ebadi, M., Vo, K., Novak, J., Govindarajan, A., Amini, A.: An updated review on head and neck cancer treatment with radiation therapy. *Cancers* **13**(19), 4912 (2021)
3. Strohl, M.P., Ha, P.K., Flavell, R.R., Yom, S.S.: PET/CT in surgical planning for head and neck cancer. *Seminars in Nuclear Medicine* **51**(1), 50–58 (2021)
4. Beddok, A., Prendergast, C.M., Seban, R.D., et al.: PET imaging in head and neck squamous cell carcinoma: From staging to response assessment. *Current Oncology Reports* **28**(1), 13 (2026)
5. De Biase, A., Sijtsema, N.M., Janssen, T., et al.: Clinical adoption of deep learning target auto-segmentation for radiation therapy: Challenges, clinical risks, and mitigation strategies. *BJR Artificial Intelligence* **1**(1) (2024)
6. Fakhry, C., Zhang, Q., Nguyen-Tân, P.F., et al.: Development and validation of nomograms predictive of overall and progression-free survival in patients with oropharyngeal cancer. *Journal of Clinical Oncology* **35**(36), 4057–4065 (2017)

7. Van Dijk, L.V., Mohamed, A.S., Ahmed, S., et al.: Head and neck cancer predictive risk estimator to determine control and therapeutic outcomes of radiotherapy (HNC-PREDICTOR). *European Journal of Cancer* **178**, 150–161 (2023)
8. Turki, A., Alshabrawy, O., Woo, W.L.: Multimodal deep learning for stage classification of head and neck cancer using masked autoencoders and vision transformers with attention-based fusion. *Cancers* **17**(13), 2115 (2025)
9. Song, B., Leroy, A., Yang, K., et al.: Deep learning model of primary tumor and metastatic cervical lymph nodes from CT for outcome predictions in oropharyngeal cancer. *JAMA Network Open* **8**(5) (2025)
10. Diamant, A., Chatterjee, A., Vallières, M., Shenouda, G., Seuntjens, J.: Deep learning in head and neck cancer outcome prediction. *Scientific Reports* **9**(1), 2764 (2019)
11. Oreiller, V., Andrearczyk, V., Jreige, M., et al.: Head and neck tumor segmentation in PET/CT: The HECKTOR challenge. *Medical Image Analysis* **77**, 102336 (2022)
12. Saeed, N., Hassan, S., Hardan, S., et al.: HEad and neCK TumOR (HECKTOR) 2025: Benchmark of segmentation, diagnosis, and prognosis in multimodal PET/CT. arXiv:2606.20143 (2026)
13. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: *MICCAI 2016*, pp. 424–432. Springer (2016)
14. Isensee, F., Wald, T., Ulrich, C., et al.: nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. In: *MICCAI 2024*, pp. 488–498. Springer (2024)
15. Rokuss, M., Langenberg, M., Kirchhoff, Y., et al.: Voxtell: Free-text promptable universal 3D medical image segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37538–37557 (2026)
16. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: *Machine Learning in Medical Imaging*, pp. 379–387. Springer (2017)
17. Rokuss, M., Kovacs, B., Kirchhoff, Y., et al.: From FDG to PSMA: A hitchhiker's guide to multitracer, multicenter lesion segmentation in PET/CT imaging. arXiv:2409.09478 (2024)
18. Kovacs, B., Xiao, S., Rokuss, M., Ulrich, C., Isensee, F., Maier-Hein, K.H.: Data-centric strategies for overcoming PET/CT heterogeneity: Insights from the AutoPET III lesion segmentation challenge. arXiv:2409.10120 (2024)
19. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: CatBoost: Unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6639–6649 (2018)
20. Hollmann, N., Müller, S., Purucker, L., et al.: Accurate predictions on small data with a tabular foundation model. *Nature* **637**(8045), 319–326 (2025)
21. Tibshirani, R.: The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16**(4), 385–395 (1997)
22. Saeed, N., Hassan, S., Hardan, S., et al.: A multimodal and multi-centric head and neck cancer dataset for segmentation, diagnosis and outcome prediction. arXiv:2509.00367 (2025)
23. Wasserthal, J., Breit, H.C., Meyer, M.T., et al.: TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence* **5**(5) (2023)
24. Amin, M.B., Edge, S.B., Greene, F.L., et al. (eds.): *AJCC Cancer Staging Manual*. 8th edn. Springer, New York (2017)