Predicted Masks Throughout: A Cross-Center Generalizable Pipeline for Joint Head and Neck Tumor Segmentation, Staging, and Prognosis

Jin Ouyang[1] [0009−0004−9566−3612], Mianyong Ding[1,2] [0009−0000−4642−3138], Esther Smeets[1] [0000−0002−8350−2131], and Baoqiang Ma[1,✉] [0000−0002−6461−1244]

[1] Division Imaging & Oncology, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, the Netherlands
j.ouyang@umcutrecht.nl, b.ma-2@umcutrecht.nl
[2] Princess Máxima Center for Pediatric Oncology. Heidelberglaan 25, 3584 CS Utrecht, the Netherlands

Abstract. Segmentation, staging, and prognosis of head and neck cancer are traditionally treated as separate tasks, despite their clinical interdependence. We present an end-to-end inference pipeline for the HECKTOR 2026 challenge that jointly predicts primary tumor and nodal segmentation (GTVp/GTVn), TN stage, and recurrence-free survival from [^{18}F]FDG-PET/CT and clinical data spanning multiple centers, developed on 8 centers and evaluated on centers held out during development. Staging and prognosis combine deep learning and radiomics branches through a diversity-driven ensemble strategy: candidate sources are screened by out-of-fold strength and Spearman rank correlation, favoring complementary, cross-paradigm information over marginal gains in individual source strength, with the final ensemble validated by paired bootstrap testing. Both branches consistently use Task 1's predicted, rather than ground-truth, masks, matching training time and inference-time conditions. Model selection throughout relied on leave-one-center-out cross-validation, prioritizing honest cross-center generalization, though Task 1's five-fold splits were not enforced to be center-disjoint, over same-distribution point estimates. Two submitted configurations, differing only in their prognosis ensemble, achieved weighted scores of 0.7205 and 0.6720 on the official Validation Phase leaderboard; the primary configuration's advantage was confirmed on a larger internal out-of-fold cohort (n=676) with statistical significance ($p \approx 0.006$).
Code is available at https://github.com/heanny/HECKTOR2026-UMCU.
Team: UMCU.

Keywords: Medical Imaging · Segmentation · Tumor Staging · Prognosis · HECKTOR challenge · Head and Neck Cancer Tumor

1 Introduction

Head and neck cancers (HNC) present significant clinical challenges due to complex regional anatomy, low soft-tissue contrast, and high morphological heterogeneity [5] [8]. Accurate GTVp/GTVn delineation is critical for radiotherapy

planning [6], yet manual contouring remains time-consuming with substantial inter-observer variability. While previous benchmarks like the 4th edition of HEad and neCK TumOR (HECKTOR) Challenge treated segmentation, and survival outcomes as isolated tasks [8], real-world clinical decision-making relies on the deep interdependence of these diagnostic factors.

To bridge this gap, the 5th edition of the HECKTOR benchmark introduces an expanded multi-institutional dataset [7] of approximately 1423 patients across 11 international centers. By combining functional metabolic insights from fluorodeoxyglucose-positron emission tomography ([^{18}F]FDG-PET) with high-resolution anatomical data from computed tomography (CT), the challenge motivates the development of holistic, AI-driven pipelines [8]. A central challenge of this expanded, multi-institutional setting is generalization across centers with heterogeneous acquisition protocols and patient populations, directly relevant to jointly modeling tumor localization, staging, and survival for real-world clinical translation.

This article presents our end-to-end approaches for automated PET/CT segmentation, tumor/nodal (T/N) staging, and survival prognosis. Our key contributions are threefold: a diversity-driven, cross-paradigm ensemble strategy for prognosis prediction, validated through statistical testing rather than point estimates alone; consistent use of predicted, rather than ground-truth, segmentation masks across the staging and prognosis branches; and a systematic comparison of cross-validation and ensemble design choices quantifying their effect on cross-center generalization.

2 Data and Preprocessing

2.1 Dataset

HECKTOR 2026 dataset is the largest multi-institutional, multimodal dataset for HNC imaging composed of PET and low-dose non-contrast-enhanced CT images and detailed clinical data from approximately 1423 cases from 11 centers. The training dataset has approximately 700 cases from 8 different centers. The validation dataset has approximately 50 unseen cases. The hidden leaderboard test dataset has approximately 400 cases from 3 centers.

2.2 Common Geometric Preprocessing

For all three tasks, all CT, PET, and (where available) label volumes were resampled to an isotropic $1 \times 1 \times 1$ mm^3 grid (B-spline for CT/PET, nearest-neighbor for labels), defined as the spatial intersection of the CT and PET bounding volumes. Each volume was then cropped to a fixed $200 \times 200 \times 310$-voxel neck region following the landmark-based strategy of Cai et al. [1]: the crop was centered on the centroid of the largest connected high-uptake PET component in the cranial 25% of the volume, exploiting brain [^{18}F]FDG uptake as a landmark independent of tumor location.

2.3 Task 1: Segmentation-Specific Preprocessing

Preprocessing followed nnU-Net's default automated planning pipeline (training: v2.1, packaged with STU-Net [3]; inference container: v2.5.1) [4],applied unmodified to the cropped CT/PET/label volumes from Section 2.2. Because the target spacing selected coincided with the existing 1 mm resolution, no additional resampling was applied. CT intensities were clipped to the dataset's $0.5th/99.5th$ percentile range ($[-150, 160]$ Hounsfield units (HU)) and standardized (mean 36.5, std 46.8); PET intensities used per-case z-score normalization, following nnU-Net's default per-channel scheme selection.

2.4 Task 2 & 3: TN Staging and Prognosis Prediction Preprocessing

Deep Learning Branch Inputs were resampled to $2\times2\times2$ mm^3 ($100\times100\times155$ voxels). CT was clipped to $[-200, 300]$ HU and scaled to $[0, 1]$; PET was normalized by each patient's 99th-percentile SUV and clipped to $[0, 1]$. Additional channels (CT, PET, `prob_T/N`, `hard_T/N`) were evaluated in combination; these were derived from Task 1's OOF predictions, using the same patient-level fold assignment as Task 1 to avoid leakage, with five-fold ensemble predictions used for the held-out CHUV cohort. Clinical variables were preprocessed per-fold (median imputation, z-score for age, binary/indicator encoding for gender and HPV status).

Radiomics Branch Quality control flagged 85/781 PET scans (10.9%) with suspiciously high SUV values (MDA: 66/443, 14.9%; CHUP: 15/75, 20.0%); 51 had confirmed unit inconsistencies and were corrected using organizer-provided factors, while the remaining 34 (elevated maximum only) were retained as-is. Features were extracted from the same out-of-fold predicted masks as the Deep Learning (DL) branch, using PyRadiomics [10] (bin width 25, $1 \times 1 \times 1$ mm^3, z-score normalization scaled to 100; shape, first-order, GLCM (Gray Level Co-occurrence Matrix), GLRLM (Gray Level Run Length Matrix), GLSZM (Gray Level Size Zone Matrix); no Wavelet/Laplacian of Gaussian). CT and PET shape features are numerically near-identical due to shared mask geometry. SUV_{mean}, SUV_{max}, and total lesion glycolysis (TLG) were additionally computed from the SUV-corrected PET; cases with empty predicted masks had all features zero-filled with an indicator flag retained.

3 Methods

Figure 1 provides an overview of the complete pipeline, detailed in the following subsections.

3.1 Segmentation

Data from seven of the eight training centers (731 cases) were used for model development, while 51 patients from CHUV formed a held-out internal cohort

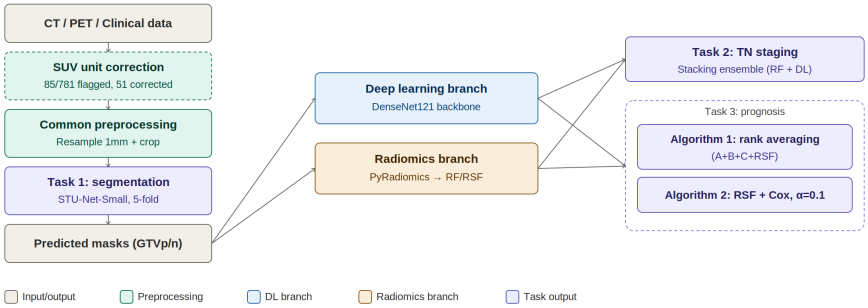

Fig. 1. Overview of the HECKTOR 2026 pipeline. All tasks share common geometric preprocessing and consistently use Task 1's predicted masks in the deep learning and radiomics branches. Algorithm 1 and Algorithm 2 differ only in the Task 3 ensemble strategy; quantitative results appear in Section 4.

used for model selection, distinct from the official HECKTOR Validation Phase cohort. Five-fold cross-validation was performed on the development cohort, with model selection based primarily on five-fold cross-validation and CHUV test performance.

We used STU-Net-S [3], implemented within the nnU-Net framework [4], as the baseline architecture. This choice follows the setting as the winning HECKTOR 2025 segmentation solution from Team MEDAI (Cai et al.) [1], but we used a five-fold ensemble rather than a ten-fold ensemble. STU-Net-S consists of a six-stage encoder–decoder architecture with residual convolutional blocks and feature dimensions of 16, 32, 64, 128, 256, and 256. Preprocessed PET and CT volumes were provided as separate input channels. Models were trained from scratch for 1000 epochs with an initial learning rate of 10^{-4}, and a polynomial learning-rate decay schedule. Preprocessing, network configuration, and training hyperparameters otherwise followed default nnU-Net settings; full plans and configuration details are available in our GitHub repository.

Based on these results, STU-Net-S was retained as the final architecture, and the five cross-validation models were ensembled by averaging their predictions.

3.2 TN Staging

Individual Models *Radiomics Random Forest.* Separate random forest (RF) classifiers were trained for T-stage and N-stage prediction using the radiomics features described in Section 2.4. The T-stage model used shape features only (56 features, effectively 28 distinct descriptors given CT/PET redundancy; Section 2.4). The N-stage model additionally included clinical variables (age, gender, HPV status), the TLG proxy, and empty-mask indicator flags (63 features total). Five RF classifiers (500 trees, max depth 6, balanced class weighting) were

trained per target with different random seeds; final class probabilities were averaged across seeds before arg-max selection.

Deep Learning Classifier. A standalone three-dimensional DenseNet121 (MONAI [2]) was used as the shared imaging backbone across all deep learning models. The classification layer was removed; global average pooling produced a 1024-dimensional feature vector, concatenated with a 64-dimensional clinical encoding (age, gender, HPV-negative/unknown status) and processed by fully connected layers (512, 256 units) into two four-class heads (T1–T4, N0–N3). Training used ordinally weighted cross-entropy (errors weighted by distance from the reference stage), optimized with Adam (see GitHub repository for full hyperparameters); model selection used mean balanced accuracy. The configuration combining clinical, CT, PET, and predicted masks (`hard_T,hard_N`) was selected for the final ensemble.

Ensemble Strategy Radiomics and standalone deep learning predictions were combined via stacking ensemble on their aligned OOF coverage ($n = 724$; 7 cases lacking complete DL input channels were excluded, while the radiomics branch retained them via zero-filled features). For each of T-stage and N-stage, OOF class probabilities from the RF (5 classes) and DL classifier were concatenated and fit to a logistic regression meta-learner ($C = 1.0$, balanced class weighting). Because the DL classifier did not predict T0, T-stage stacking used only T-stage $\geq$ T1 cases (9-dimensional input: 5 RF + 4 DL class probabilities), though the RF's T0 probability was retained as an input feature.

This configuration was selected as the final Task 2 model for both submitted algorithms based on its stability across development iterations; internal and leaderboard performance are reported in Section 4.1.

3.3 Prognosis Prediction

The two submitted algorithms shared identical Task 1 and Task 2 configurations (confirmed by identical validation Dice, T-stage, and N-stage scores; see Table 3) and differed only in the Task 3 ensemble strategy, described below as Algorithm 1 and Algorithm 2.

Individual Models *Clinical Random Survival Forest.* A Random Survival Forest (RSF) was trained on the full set of CT and PET radiomics features (Section 2.4) together with clinical variables (age, gender, HPV status), totaling 433 features. Five RSF models (300 trees, minimum leaf size 10, $\sqrt{p}$ features per split, where $p = 433$ is the total number of input features) were trained with different random seeds; at inference, risk scores were averaged across the five seed models. OOF risk scores were generated via a leave-one-center-out (LOCO) procedure, in which each center was held out in turn, and features were standardized using only the remaining centers' training data.

Deep Learning Survival Models. Two deep learning architectures, sharing the DenseNet121 imaging backbone described in Section 3.2, were used to produce candidate risk scores for Task 3.

The *standalone Task 3 model* used the shared backbone together with five clinical features (age, gender, HPV-negative status, unknown HPV status, and treatment); the fused imaging and clinical representation was processed by fully connected layers with 512 and 256 units, followed by a single output unit representing the predicted Cox log-risk score. The model was optimized by minimizing the negative Cox partial log-likelihood, with model selection based on the concordance index (C-index).

The *multi-task model* jointly predicted T-stage, N-stage, and recurrence-free survival from a shared DenseNet121 backbone, using separate clinical encoders for the staging branch (age, gender, HPV-negative/unknown status) and the survival branch (same four variables plus treatment). The staging branch followed the standalone Task 2 classifier structure (Section 3.2). The eight softmax staging probabilities were concatenated with the shared imaging features and survival-specific clinical features before recurrence-risk prediction; these were detached from the computational graph, so the survival loss did not update the staging heads. Training used a 15-epoch staging-only warm-up, after which the total loss summed T-stage, N-stage, and weighted Cox losses (weight 1); model selection used mean staging balanced accuracy plus C-index. The model was optimized with AdamW and cosine-annealing scheduling, using the same epoch budget, batch size, and early-stopping criteria as the standalone models. Input channel combinations were evaluated for both architectures, denoted `task3/*` (standalone) and `multitask/*` (multi-task).

Source Selection for Task 3 Ensemble Fourteen deep-learning risk-prediction configurations (differing in input channel combinations) together with the Clinical Random Survival Forest (RSF) model were evaluated as candidate sources for the final prognosis ensemble. Each candidate's OOF C-index was computed as a strength measure, and pairwise Spearman rank correlations were computed to assess redundancy between candidates.

Four DL configurations were selected using a mechanical screening rule with thresholds fixed prior to inspecting fusion results: OOF C-index above 0.650, and Spearman correlation with all selected members below 0.500 (candidates considered in descending order of strength). This selected A, B, C, and D (Table 1), excluding RSF (OOF C-index 0.639, below threshold).

The mechanical rule alone, therefore, selected A+B+C+D, excluding RSF for falling below the strength threshold. Correlation analysis (Table 2) showed D moderately correlated with selected members (D–A: 0.408, D–B: 0.491), while RSF showed the **lowest correlation** with DL sources overall (RSF–B: 0.190, RSF–D: 0.279, RSF–A: 0.458, RSF–C: 0.419), consistent with its distinct modelling paradigm. D was therefore replaced with RSF: despite lower individual strength, its low correlation motivated inclusion as a complementary, cross-paradigm source.

Table 1. Candidate sources selected by the mechanical screening rule for Task 3 fusion, with single-source OOF C-index. The RSF value in Table 1 uses the 676-case cohort aligned with sources (Section 4.2 reports 0.6403 on the full 683-case cohort).

Source	Configuration	OOF C-index
A	Multitask DL, clinical+CT+PET+hard_T+hard_N	0.672
B	Task-3 DL, clinical+CT+PET	0.668
C	Task-3 DL, CT+PET+prob_T+prob_N+hard_T+hard_N	0.660
D	Task-3 DL, CT+PET	0.660
RSF	Clinical Random Survival Forest	0.639

Table 2. Spearman rank correlation between candidate sources (n=676).

	A	**B**	**C**	**D**	**RSF**
A	1.000	0.495	0.454	0.408	0.458
B	0.495	1.000	0.237	0.491	**0.190**
C	0.454	0.237	1.000	0.395	**0.419**
D	0.408	0.491	0.395	1.000	**0.279**
RSF	0.458	**0.190**	**0.419**	**0.279**	1.000

The internal OOF evaluation cohort for the four-source fusion ($n = 676$) reflects the intersection of coverage across A, B, C, and RSF; RSF alone covered 683 cases, while source A lacked OOF predictions for 7 of these, reducing the aligned cohort to 676. The final submission in Algorithm 1 used equal-weight (0.25 each) rank averaging of sources A, B, C, and RSF, achieving an OOF C-index of **0.716** as described below.

Task 3 Algorithm 1: Four-Source Rank Averaging The primary submission Algorithm 1 combined sources A, B, C, and RSF (Section 3.3) via equal-weight rank averaging (performance reported in Section 4.1). At inference, each source's raw risk score was converted to a quantile rank by comparison against a pre-computed reference table of that source's OOF risk distribution on the full training cohort, and the four resulting quantile ranks were averaged with equal weight (0.25 each) to produce the final risk score. This deployment-time mapping (fit on the full cohort) differs from the LOCO mapping used for leakage-safe internal evaluation (Section 3.3).

Task 3 Algorithm 2: RSF + Cox Ensemble The backup submission Algorithm 2 combined the Clinical RSF risk score (Section 3.3) with a single deep-learning risk score, source A (Table 1), via a Cox proportional hazards model. Out-of-fold RSF and DL risk scores were each min-max normalized to $[0, 1]$ and used as the two covariates of a Cox proportional hazards (CoxPH) model regularized with $\alpha = 0.1$, fit on the training cohort. The fitted model assigned coefficients of 1.313 to the normalized RSF risk and 3.987 to the normalized DL

risk:

$$\text{risk}_{\text{final}} = 1.313 \cdot \tilde{r}_{\text{RSF}} + 3.987 \cdot \tilde{r}_{\text{DL}} \tag{1}$$

where $\tilde{r}_{\text{RSF}}$ and $\tilde{r}_{\text{DL}}$ denote the min-max normalized RSF and DL risk scores. A paired bootstrap (2000 resamples, n=676, 138 events) placed the RSF coefficient at 1.313 (95% CI $[0.480, 2.004]$) and the DL coefficient at 3.987 (95% CI $[2.222, 5.225]$); both intervals exclude zero and do not overlap, indicating that the deep-learning source contributes more strongly to the combined risk score. At inference, the same min-max scaling (fit on the training distribution) was applied to new RSF and DL risk scores before combination via the fitted CoxPH model.

The summary of submitted configurations is in Table 3. For both algorithms, Task 1 and Task 2 are identical; only Task 3 differs.

Table 3. Configuration of the two submitted algorithms.

Task	**Configuration**
Task 1	STU-Net-Small, 5-fold ensemble
Task 2	Stacking ensemble (RF + standalone DL), $C = 1.0$
Task 3 (Algorithm 1)	Four-source rank averaging (A+B+C+RSF)
Task 3 (Algorithm 2)	RSF + Cox ensemble, $\alpha = 0.1$

4 Experiments and Results

4.1 Results Overview

Table 4 and Table 5 summarize performance across internal and HECKTOR validation cohorts (∼50 cases) for all three tasks. Algorithm 1 and Algorithm 2 share identical Task 1 and 2 configurations and thus identical Dice and staging BA on the leaderboard, differing only in Task 3 prognosis, which drives different Weighted Scores (**0.7205** vs. **0.6720**).

Table 4. Segmentation and TN staging performance across internal and leaderboard cohorts. TN staging $n = 724$ excludes 7 cases lacking DL OOF predictions. '–' indicates a metric not reported by that cohort's official leaderboard.

Cohort	**#**	**GTVp/n Dice**	**Mean Dice**	**T/N BA**	**Mean BA**
Internal OOF	731/724	0.686/0.688	0.687	0.583/0.681	0.632
Internal CHUV	51	0.754/0.658	0.706	0.515/0.745	0.630
Validation Leaderboard	∼50	–/–	0.671	0.489/0.800	–
Testing Leaderboard (Alg 1)	∼400	–/–	0.533	0.438/0.516	–
Testing Leaderboard (Alg 2)	∼400	–/–	0.514	0.437/0.519	–

Segmentation Dice was similar between the internal validation cohort and the leaderboard (0.687 vs. 0.671). TN staging mean balanced accuracy was also comparable across internal cohorts (0.632 OOF vs. 0.630 CHUV). In contrast, T-stage and N-stage balanced accuracy changed in opposite directions from the internal OOF cohort to the leaderboard, decreasing from 0.583 to 0.489 for T stage and increasing from 0.681 to 0.800 for N stage. These differences are discussed further in Section 5.

Table 5. Prognosis (C-index) across cohorts. Internal OOF includes 95% CI (paired bootstrap, 2000 resamples); CHUV/leaderboard are small-sample references.

Cohort	Algorithm 1	Algorithm 2	Difference
Internal OOF ($n = 676$, 138 events)	**0.716** [0.674, 0.760]	0.682 [0.637, 0.729]	+0.033* [0.011, 0.056]
Internal CHUV ($n = 44$, 8 events)	0.765	0.765	+0.000
Validation Leaderboard	**0.818**	0.696	+0.122
Testing Leaderboard	**0.575**	0.558	+0.017

*Two-sided paired bootstrap. $p \approx 0.006$; only comparison reaching significance.

As shown in Table 5, Algorithm 1 achieved a higher C-index than Algorithm 2 on the internal OOF cohort (**0.716** vs. 0.682; paired bootstrap $p \approx 0.006$). On the CHUV cohort, both algorithms achieved the same C-index (0.765), while on the leaderboard Algorithm 1 achieved **0.818** compared with 0.696 for Algorithm 2. These results motivated the selection of Algorithm 1 as the primary submission.

4.2 Task 2/3 Ensemble Design Choices

Three methodological decisions in the Task 2/3 pipeline prioritized honest cross-center generalization over same-distribution point estimates, quantified below on the Clinical RSF model's full OOF cohort ($n = 683$, 140 events; broader than the 676-case four-source-aligned cohort in Section 3.3, which requires alignment with the DL sources).

Cross-paradigm diversity via Spearman correlation. A mechanical strength threshold alone would have selected A, B, C, and D, excluding RSF (OOF C-index 0.639, below the 0.65 threshold; Section 3.3). Spearman correlation showed RSF substantially less correlated with the DL sources than the excluded candidate D (RSF–B: 0.190, RSF–D: 0.279, vs. D–A: 0.408, D–B: 0.491), motivating its inclusion as a complementary, cross-paradigm source.

Multi-seed averaging and cross-validation scheme. The RSF OOF C-index varied across five seeds (range [0.6366, 0.6416], std 0.0017); averaging yielded a C-index nearly identical to the single-seed mean (0.6403 vs. 0.6400), so multi-seed averaging improved stability and reproducibility rather than the point estimate. Evaluating the same model under stratified 5-fold CV (mixing centers

across folds) yielded a higher C-index (0.6527) than LOCO (0.6403), a gap of 0.0124 – the optimism stratified CV would introduce given that training and testing cohorts here are drawn from different centers. LOCO was therefore used throughout Task 2 and Task 3 (Sections 3.2 and 3.3).

4.3 Segmentation Experiments

Segmentation with Background Class Decomposition We investigated subdividing the background class into anatomically meaningful structures during training [9], using TotalSegmentator [11] pseudo-labels for structures overlapping GTV_p/GTV_n annotations (`headneck_muscles`, `head_glands_cavities`, `headneck_bones_vessels`), with GTV labels retained at highest priority. We evaluated 11 and 4 auxiliary structures, trained on 3 of 5 cross-validation splits due to computational constraints. On the test set, GTVp DSC increased by 0.8–2.0 points, while GTVn results were inconsistent (4 structures: +1.5, 11 structures: −0.3), yielding overall gains of 0.8–1.1 points; however, validation DSC degraded by 0.7–2.0%. Given these marginal, inconsistent gains, the baseline model without auxiliary structures was retained (Table 6).

Table 6. Background class decomposition vs. baseline (ΔDSC, percentage points).

Configuration	Split	ΔGTVp	ΔGTVn	ΔAvg.
11 aux. structures	Val (439) / Test (51)	−1.3 / +2.0	−2.7 / −0.3	−2.0 / +0.8
4 aux. structures	Val (439) / Test (51)	−0.4 / +0.8	−1.0 / +1.5	−0.7 / +1.1

5 Discussion

This work presents an end-to-end inference pipeline for joint segmentation, TN staging, and recurrence-free survival prediction on the HECKTOR 2026 dataset, organized around a common principle: maximizing honest estimation of cross-center generalization rather than optimizing point estimates on same-distribution data.

Testing Phase results showed marked declines across all tasks (Tables 4 and 5), led by a segmentation Dice drop (0.687→0.510-0.530) that plausibly propagated into N-stage balanced accuracy (0.681→0.520 given the pipeline's predicted-mask reliance (Section 2.4). The prognosis C-index decline was consistent for both algorithms (Algorithm 1: 0.716→0.575; Algorithm 2: 0.682→0.558), consistent with the stratified-vs-LOCO optimism gap already quantified (Section 4.2) and with a Testing Phase center absent from development, a scenario LOCO cannot fully anticipate. Reporting worst-case rather than mean LOCO-fold performance would offer a more conservative generalization estimate.

Across all three subtasks, internal development-cohort performance tracked the official Validation Phase leaderboard reasonably closely. Segmentation Dice was nearly identical internally vs. leaderboard (**0.687** vs. **0.671**; Section 4.1), and TN staging mean balanced accuracy was similarly stable (**0.632** OOF vs. **0.630** CHUV). However, T-stage and N-stage balanced accuracy moved in opposite directions between the internal OOF cohort and the leaderboard, likely reflecting the markedly imbalanced, center-dependent staging label distributions (N2: 60.7% of the 731-case development cohort). This is precisely the heterogeneity LOCO cross-validation is designed to expose and stratified cross-validation would mask (Section 4.2). By contrast, T-stage labels were more evenly distributed (no class exceeded 37%), suggesting its lower balanced accuracy reflects genuine difficulty distinguishing intermediate stages (T2/T3) rather than imbalance, though the small size of both cohorts (51 and $\sim$50 cases) limits confidence in this observation.

Several of the design choices in Section 4.2 directly targeted this generalization gap: LOCO cross-validation produced C-index estimates 0.0124 points lower than stratified CV, quantifying the optimism a same-distribution split would introduce; including the Clinical RSF model despite its lower individual strength favored complementary cross-paradigm information over marginal strength gains; and consistent use of Task 1's OOF predicted masks across both branches matched training- and inference-time conditions, since ground-truth masks are unavailable at test time.

The comparison between the two submitted algorithms (Section 4.1) illustrates the value of relying on statistically grounded internal estimates rather than leaderboard results in isolation. While Algorithm 1's leaderboard C-index advantage over Algorithm 2 appeared substantial (**0.818** vs. **0.696**), the internal OOF comparison (n=676, 2000 resamples, paired bootstrap $p \approx 0.006$) showed a more modest but statistically significant difference (**0.716** vs. **0.682**).

The internal quality-control process also surfaced a data issue external to our modeling choices: 85/781 PET scans (10.9%) were flagged with high SUV values in two centers (Section 2.4), of which 51 were corrected using organizer-provided factors and 34 retained as-is. This uneven distribution is a further illustration of the center-level heterogeneity motivating this work's design choices.

A related limitation is that CT- and PET-derived shape features are numerically near-identical given the shared predicted mask (Section 2.4), halving the effective dimensionality of the shape feature set, though a post hoc check found this had negligible impact (LOCO BA change < 0.02).

6 Conclusion

We presented an end-to-end inference pipeline for the HECKTOR 2026 challenge, combining shared preprocessing, an nnU-Net-based segmentation model, and complementary radiomics and deep-learning branches for staging and prognosis. Two algorithms, differing only in their Task 3 ensemble strategy, achieved a Weighted Score of **0.7205** (Algorithm 1) on the official Validation Phase leader-

board, with internal OOF evaluation supporting its selection as the stronger configuration. However, performance declined substantially on the official Testing Phase, led by a drop in segmentation Dice that propagated into N-stage balanced accuracy and prognosis C-index.

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

References

1. Cai, L., Liang, X., Zhang, T., Huang, J., Tan, T., Yin, Y.: Less is more: Efficient pet/ct segmentation and multimodal prediction of recurrence-free survival and hpv status in head and neck cancer. In: Fourth Head and Neck Cancer Tumor Lesion Segmentation, Diagnosis and Prognosis
2. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
3. Huang, Z., Wang, H., Deng, Z., Ye, J., Su, Y., Sun, H., He, J., Gu, Y., Gu, L., Zhang, S., et al.: Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. arXiv preprint arXiv:2304.06716 (2023)
4. Isensee, F., Jaeger, P.F., Kohl, S.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 (2021)
5. Johnson, J.E., Burtness, B., Leemans, C.R., Lui, V.W.Y., Bauman, J.E., Grandis, J.R.: Head and neck squamous cell carcinoma. Nature Reviews Disease Primers **6**(1), 92 (2020)
6. Moe, Y.M., Groendahl, A.R., Tomic, O., Dale, E., Malinen, E., Futsaether, C.M.: Deep learning-based auto-delineation of gross tumour volumes and involved nodes in PET/CT images of head and neck cancer patients **48**(9), 2782–2792. https://doi.org/10.1007/s00259-020-05125-x
7. Saeed, N., Hassan, S., Hardan, S., Aly, A., Taratynova, D., Nawaz, U., et al.: A multimodal and multi-centric head and neck cancer dataset for segmentation, diagnosis and outcome prediction (2025), https://arxiv.org/abs/2509.00367
8. Saeed, N., Hassan, S., Hardan, S., Cai, L., Liang, X., Mazher, M., et al.: Head and neck tumor (hecktor) 2025: Benchmark of segmentation, diagnosis, and prognosis in multimodal pet/ct (2026), https://arxiv.org/abs/2606.20143
9. Saluja, R., Cihangir, A., Deng, R., Paetzold, J.C., Liu, F., Sabuncu, M.R.: Backsplit: The importance of sub-dividing the background in biomedical lesion segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8492–8502 (2026)
10. Van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.: Computational radiomics system to decode the radiographic phenotype. Cancer research **77**(21), e104 (2017)
11. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)