Unified vs modular methods for head-neck tumor and lymph node segmentation, staging, and survival prediction

Kai Geissler[0009−0008−9709−0335]

Fraunhofer Institute for Digital Medicine MEVIS, Max-von-Laue-Str. 2, 28359 Bremen, Germany
kai.geissler@mevis.fraunhofer.de

Abstract. This work compares unified and modular approaches for the HECKTOR 2026 challenge on joint primary-tumor and lymph-node segmentation, TN staging, and recurrence-free survival prediction from PET/CT data. The unified approach builds on RadYOLO with task-specific heads for segmentation, tumor and node staging, and survival prediction. The modular approach combines nnU-Net for segmentation, RadYOLO-based tumor staging and survival prediction, and decision-tree-based node staging from predicted lymph-node component volumes. On the 400 hidden final test cases, the unified method achieved a mean Dice score of 0.578, T- and N-stage balanced accuracies of 0.527 and 0.526, respectively, and a recurrence-free-survival C-index of 0.535. The modular method improved the mean Dice score to 0.674 and the T- and N-stage balanced accuracies to 0.542 and 0.565, respectively, while achieving the same C-index because both methods used the same survival-prediction component. These results indicate that task-specific segmentation and component-volume-based node-stage prediction improved the segmentation and staging components on the hidden test cohort and highlight the difficulty of developing a single unified method.
Team name on Grand Challenge: FME.

Keywords: Head and neck cancer · PET/CT imaging · Unified and modular deep learning · Tumor and lymph node segmentation · TN staging · Recurrence-free survival prediction

1 Introduction

Head and neck cancers represent a significant global health burden, complicated by complex regional anatomy, heterogeneous tumor appearance on imaging, and the difficulty of predicting outcomes [5]. Progress on its automated imaging analysis has long been limited by the scarcity of large, multi-institutional datasets, motivating the HECKTOR (HEad and neCK TumOR) challenge series. In its 2025 edition, HECKTOR combined PET/CT and electronic health records from over 1,100 patients across ten centers into a benchmark spanning segmentation of the primary tumor and lymph nodes, recurrence-free survival (RFS) prediction,

and human papilloma virus status (HPV) classification, attracting 35 teams and 15 submissions, with top methods reaching a mean Dice score of 0.75, a concordance index of 0.66 for RFS prediction, and a balanced accuracy of 0.56 for HPV status classification [6].

The 2026 edition expands the dataset to over 1,200 PET/CT studies from 11 centers ($\sim$ 700 training, $\sim$ 400 test cases). Each case includes a co-registered PET/CT volume, clinical variables such as age and sex, and the prediction targets: expert tumor and lymph node metastasis contours, tumor and node stages, and recurrence-free survival times with recurrence flags. Rather than treating tasks in isolation, HECKTOR 2026 unifies segmentation, TN staging, and prognosis into a single end-to-end pipeline mirroring clinical workflow, with performance ranked per subtask (Dice, balanced accuracy, C-index) and combined via a weighted score emphasizing the downstream clinical tasks. The challenge's overall weighted score S is computed as

$$S = 0.25 \times \frac{\mathrm{Dice_T} + \mathrm{Dice_N}}{2} + 0.35 \times \frac{\mathrm{BalAcc_T} + \mathrm{BalAcc_N}}{2} + 0.40 \times \text{C-index}, \quad (1)$$

assigning 25 % weight to segmentation, 35 % to the mean balanced accuracy of T- and N-stage prediction, and 40 % to recurrence-free survival prediction.

In this paper, we describe our challenge submission from team FME. We submitted two methods to the final test phase:

1. Unified model based on RadYOLO [1], which uses the CT and PET images as inputs and jointly predicts the tumor and node segmentations, the tumor and node staging as well as the recurrence-free survival time.
2. Modular method, which uses nnU-Net [2] for segmentation (with the CT and PET images as input), reuses the previous RadYOLO model for tumor staging and recurrence-free survival prediction, and uses a decision tree for node staging, based on the node segmentation masks.

2 Methods

2.1 Data

The HECKTOR 2026 dataset comprises more than 1,200 FDG-PET/CT studies collected from 11 international centers, split into approximately 700 training cases from 8 centers and 400 test cases from 3 centers (new, previously unseen data). Each case consists of a co-registered 3D PET and CT volume of the head and neck region, with PET intensities converted to standardized uptake value (SUV). The SUV conversion was missing for about 30 cases in the official dataset and added manually before model training. For training cases, expert-delineated segmentation masks are provided for the primary tumor and involved lymph nodes, with multiple nodal lesions sharing a single label. Accompanying clinical metadata includes recruiting center, age, sex, tobacco and alcohol consumption, and HPV status, while training cases additionally include T- and N-stage, recurrence status, and recurrence-free survival. Some clinical variables are missing

for a subset of patients. Test data were not distributed directly to participants. The evaluation is instead performed exclusively through Docker container submissions on the Grand Challenge platform.

2.2 Image cropping

To reduce the size of imaging data processed by the prediction models, we first cropped the PET/CT volumes to the head-neck region. To this end, we trained a RadYOLO model to detect a bounding box on the CT volume comprising the head and spine down to vertebra C7. RadYOLO is a computationally efficient architecture developed for object detection and segmentation in 3D medical images. To create training data for the cropping model, we used TotalSegmentator V2 [7] to segment the HECKTOR challenge data and extract the respective bounding boxes. In addition, we added the cases from the TotalSegmentator dataset to the training data which covered a sufficient number of slices and included the head. The RadYOLO model for cropping used a fixed resolution of $256 \times 256 \times 128$ voxels. It clipped the image intensities using the 0.05^{th} and 99.5^{th} percentiles and performed a z-score normalization with mean and standard deviation. Percentiles, mean, and standard deviation were precomputed within the foreground boxes of the training dataset. Because the cropping was a bit too tight in some cases, we added a margin of 10, 10, and 5 on both sides in the x, y, and z dimensions when feeding the cropped images to the downstream prediction RadYOLO and margins of 15, 15, and 30 when feeding the cropped images to nnU-Net.

2.3 Unified model: RadYOLO

The HECKTOR 2026 challenge strongly encouraged unified approaches which tackle all three subtasks with one joint model. We used RadYOLO and added additional prediction heads to predict image-level labels like the tumor and node stages as well as recurrence-free survival time.

The standard RadYOLO detection and segmentation branches produced the primary-tumor and lymph-node masks. For the image-level tasks, we extracted the neck features from the RadYOLO feature pyramid on each detection level right before the detection heads. The feature channels on each level were projected down to fewer feature channels using a $1 \times 1 \times 1$ convolution, and we applied global average pooling and global max pooling to each reduced feature tensor. The pooled max and average features from all pyramid levels were concatenated and passed through a shared MLP hidden layer. Separate output layers then produced T-stage logits, N-stage logits, and the survival-prediction activations. Thus, the segmentation and image-level tasks shared the backbone features, while their final predictions were task-specific.

T-stage and N-stage prediction used categorical cross-entropy with class weights inversely proportional to class frequency. Survival prediction used a weighted combination of a multi-task logistic regression (MTLR) loss [8] and

a Cox partial likelihood loss [4]. The total training objective was the weighted sum of the RadYOLO detection, segmentation, staging, and survival losses.

PET and CT were provided as two input channels. We resampled the cropped PET/CT volumes to a fixed resolution of $192 \times 224 \times 96$ voxels. For the CT image we clipped the intensities using the 0.05^{th} and 99.5^{th} percentiles and performed a z-score normalization with precomputed mean and standard deviation. The percentiles, mean and standard deviation were computed within the foreground masks of all cases in the training dataset. For the PET image an image-wise z-score normalization was used. The clinical variables were not used as model input (neither for the unified model nor the modular method). Table 1 and 2 list the hyperparameter settings used for the RadYOLO model training.

We trained an ensemble of five models by splitting the training data into five folds and using four folds for training and one for validation, i.e. best model selection. During inference we employed test-time augmentation by flipping each input image once along each axis (eight flipping combinations in total) for each model and then aggregated the results by averaging them.

2.4 Modular method: nnU-Net, RadYOLO and Decision Tree

The modular pipeline was designed to test whether components with task-specific inductive biases could improve the segmentation and node-staging parts of the workflow. This was motivated by our performance on the validation leaderboard where we observed that our unified model was competitive for tumor staging, but significantly worse in the segmentation task, node staging and recurrence-free survival prediction than the top submissions on the validation leaderboard. Therefore, we developed another method which is modular and solves different tasks with different models.

For tumor and node segmentation we trained an ensemble of five nnU-Nets [2] using the default 3D nnU-Net plans for the full-resolution model and a random data split. For the CT we selected nnU-Net's CT normalization and for the PET image we selected its z-score normalization. nnU-Net selected a resampling voxel size of $1.93 \times 1.93 \times 2.96$ mm^3 and patch size of $160 \times 160 \times 96$ voxels. During inference we applied it with test-time augmentation (flipping along each axis, eight flips in total).

For tumor staging and recurrence-free survival prediction we employed the RadYOLO model from the previous section. For tumor staging we added a fall-back to predict tumor stage T0, if the nnU-Net tumor segmentation has a volume below $30\,\text{mm}^3$.

For lymph node staging we extracted the connected component volumes of the node masks that nnU-Net predicted on the validation data of each model. Based on the volumes of the four largest components we used scikit-learn (v1.6.1) to fit a decision tree with a maximum depth of three, balanced class weights and Gini impurity as splitting criterion. Fig. 1 displays the final decision tree. During development we also tested to use the total node mask volume and number of connected components as decision tree input, but found the volumes of the four largest components to perform better. We decided to use a small decision tree

Table 1. Training hyperparameters for unified RadYOLO model training.

Parameter	Value
Model scale	s (small)
Prototype Layer	Full resolution
Prototype Masks	16
Loss (bounding box)	L1 + CIoU loss [9]
Loss (objectness)	Binary cross entropy
Loss (separate object classification)	Categorical cross entropy
Loss (segmentation)	Dice + Binary cross entropy
Loss (image label)	Categorical cross entropy
Loss (survival prediction)	MTLR + Cox partial likelihood
Loss weight (CIoU)	7.5
Loss weight (L1)	100
Loss weight (objectness)	0.03
Loss weight (separate object classification)	0.05
Loss weight (segmentation)	5
Loss weight (image label)	0.025
Loss weight (survival prediction, MTLR)	0.05
Loss weight (survival prediction, Cox)	0.2
Optimizer	MuSGD (Muon + SGD)
Muon weight	0.5
Learning Rate Scheduler	Linear warmup + Linear decay
Total Epochs	1000
Warmup Epochs	4.4
Warmup Momentum	0.72
Learning rate (start, non-bias weights)	0.0
Learning rate (start, bias weights)	2.1×10^{-3}
Learning rate (base, after warmup epochs)	1.43×10^{-2}
Learning rate (final factor)	1.64×10^{-2}
Momentum	0.875
Weight Decay	5.45×10^{-6}
Image Size (x, y, z)	$192 \times 224 \times 96$
Batch Size	2
Nominal Batch Size (gradient accumulation)	14
Normalization CT	z-Score (precomputed)
Normalization PET	z-Score (per-image)
Task Aligned Assigner Top-k	3

Table 2. Data augmentation hyperparameters for unified RadYOLO model training. BG: `batchgenerators` library [3]

Parameter	Value
Mirror probability per axis	0.5
Rotation augmentation angle [min, max] per axis	[-15°, 15°]
Rotation augmentation probability	0.2
Cutout augmentation settings [size fraction, repeats]	[0.5, 1], [0.25, 2], [0.125, 4]
Cutout augmentation probability	0.1
Zoom augmentation [min, max]	[0.65, 1.5]
Zoom augmentation probability	0.5
Channel drop augmentation probabilities per channel	[0.05, 0.05]
BG Gaussian noise probability	0.2
BG Gaussian noise variance	[0.0, 0.1]
BG Gaussian blur probability	0.2
BG Gaussian blur sigma	[0.5, 1.0]
BG multiplicative brightness probability	0.2
BG multiplicative brightness range	[0.75, 1.25]
BG additive brightness probability	0.15
BG additive brightness mean	0.0
BG additive brightness sigma	0.3
BG contrast probability	0.15
BG contrast range	[0.75, 1.25]
BG low resolution probability	0.2
BG low resolution scale	[0.5, 1.0]
BG gamma probability	0.3
BG gamma range	[0.7, 1.5]
BG inverted gamma probability	0.1

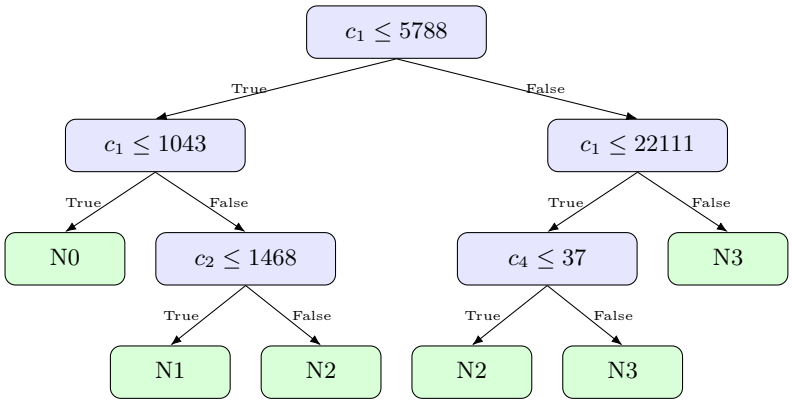

Fig. 1. Decision tree to predict node stage in the modular method. It is based on volumes $c_1, \ldots, c_4$ in mm^3 of the four largest connected components in the predicted lymph-node segmentation mask, ordered by volume.

Table 3. Validation leaderboard results. The modular method was not submitted during the validation phase. RFS: Recurrence-free survival, T/N Bal. Acc: Balanced accuracy for tumor/node stage classification.

Method	Dice	RFS C-Index	T Bal. Acc.	N Bal. Acc.	Weighted Score
Unified	0.535	0.729	0.555	0.560	0.620
Modular	-	-	-	-	-

Table 4. Final test results of submitted methods. RFS: Recurrence-free survival, T/N Bal. Acc: Balanced accuracy for tumor/node stage classification.

Method	Dice	RFS C-Index	T Bal. Acc.	N Bal. Acc.	Weighted Score
Unified	0.5778	0.5353	0.5265	0.5256	0.5427
Modular	0.6738	0.5353	0.5422	0.5651	0.5764

as model because it is less prone to overfitting and interpretable, which makes it straightforward to perform plausibility checks the resulting model.

3 Results

3.1 Validation phase

Table 3 depicts the performance of our unified model on the validation leaderboard. It achieved a mean Dice score of 0.535, balanced accuracies of 0.555 (tumor staging) and 0.560 (node staging) and a C-index of 0.729 for survival prediction. During the validation phase, our model was competitive with the top approaches with respect to tumor staging. However, it lagged behind the top approaches with respect to the segmentation Dice, node staging balanced accuracy, and survival prediction C-index. Overall the unified model achieved place 48/100 on the validation leaderboard.

We cannot report validation phase results for the modular method because it was not submitted in the validation phase, so we only report our internal results on the training dataset here. Using 5-fold cross-validation on the nnU-Net node segmentation of the training dataset, the decision tree for node stage classification achieved a balanced accuracy of ~ 0.74. The nnU-Net model for segmentation achieved a mean (standard deviation) Dice of 0.72 (0.02) for tumor segmentation and 0.64 (0.02) for node segmentation in internal 5-fold cross-validation of the individual models.

3.2 Test phase

Table 4 shows the final challenge results for both methods on the 400 hidden test cases. In the segmentation task, the unified model achieved a mean Dice of 0.578 while the nnU-Net-based approach of the modular method achieved a mean

Dice of 0.674. For recurrence-free survival prediction both methods achieved a C-index of 0.535 as they used the same model for their prediction. For tumor staging, the unified method achieved a balanced accuracy of 0.527 while the modular method achieved 0.542, where the increase stems only from the added fallback for T0 classification. The node staging balanced accuracy was 0.526 for the unified method and 0.565 for the decision tree approach from the modular method.

3.3 Things that did not work

During model development we tried to add the clinical variables (age, sex, tobacco/alcohol consumption, performance status, treatment and HPV status) to RadYOLO using a tabular transformer model. However, it did not provide a measurable improvement during internal validation, so we left this component out of the final models to make them more robust.

Our first unified model submitted to the validation phase used PET normalization with precomputed percentiles, mean and standard deviation analogous to the CT normalization. This model obtained very poor performance on the validation leaderboard with placements of 94/100 and 97/100, even though it performed well during internal validation. Therefore, we switched to per-image z-score normalization of PET images, as we suspected a data drift between PET images of the training dataset and the hidden test cohort of the challenge. Because the challenge test data is hidden we cannot validate this hypothesis.

4 Discussion

The final-test results provide a direct comparison of the unified and modular pipelines on the same 400 hidden test cases. The modular pipeline achieved a higher weighted challenge score than the unified model (0.576 versus 0.543). The largest improvement was observed for segmentation, where the mean Dice score increased from 0.578 to 0.674. The modular pipeline also improved T-stage balanced accuracy from 0.527 to 0.542 and N-stage balanced accuracy from 0.526 to 0.565. These results indicate that, in this challenge setting, replacing the unified segmentation component with nnU-Net and using a separate component-volume-based node staging component improved the overall segmentation and staging performance.

The task-specific improvements should nevertheless be interpreted carefully. The T-stage performance increase of 0.016 results from the added T0 fallback based on the predicted primary-tumor volume. It therefore does not demonstrate an advantage of a separate T-stage model. Likewise, the N-stage comparison adds two components simultaneously: the lymph-node segmentation model and the classifier that maps the segmentation to N stage. Consequently, the final-test results measure the benefit of the complete modular N-stage pipeline, but do not separately quantify the contribution of nnU-Net segmentation and the component-volume decision tree.

Recurrence-free survival prediction achieved the same C-index of 0.535 for both submissions because the modular pipeline reused the unified RadYOLO survival ensemble without modification. Thus, the present comparison does not establish whether unified or modular modeling is more effective for survival prediction. In addition, the achieved C-index of 0.535 is only slightly above chance-level concordance (0.5). On the reduced number of cases in the validation phase the same model achieved a C-index of 0.729. This underscores the need of large enough evaluation datasets to accurately measure algorithm performance.

The internal node-stage balanced accuracy of approximately 0.74 for the component-volume decision tree did not transfer to the same extent to the hidden test cohort, where the modular pipeline achieved a balanced accuracy of 0.565. Possible explanations include differences in cohort composition, the prevalence of N stages, segmentation errors, and sensitivity of the learned volume thresholds to data from unseen centers. This observation stresses the importance of evaluation with center-held-out validation splits to assess the generalizability of machine learning models.

Finally, the development experiments suggest that PET intensity normalization requires careful consideration in multi-center PET/CT data. Image-wise PET z-score normalization yielded substantially better validation-leaderboard performance than dataset-level normalization in our experiments. However, since the hidden data are unavailable, this observation cannot establish the source of the difference or confirm a specific domain shift. Similarly, adding clinical variables through a tabular transformer did not improve internal validation and was therefore not included in the final submissions.

5 Conclusion

We compared unified and modular approaches for joint primary-tumor and lymph-node segmentation, TN staging, and recurrence-free survival prediction in the HECKTOR 2026 challenge. On the 400 hidden final test cases, the modular pipeline improved the mean Dice score from 0.578 to 0.674, the T-stage balanced accuracy from 0.527 to 0.542, and the N-stage balanced accuracy from 0.526 to 0.565. These improvements increased the weighted challenge score from 0.543 for the unified RadYOLO model to 0.576 for the modular pipeline.

The recurrence-free-survival C-index was identical for both methods because the same RadYOLO survival-prediction component was used in both submissions. Therefore, the results support task-specific segmentation and component-volume-based node-stage prediction for this challenge setting, but do not establish an advantage of modular modeling for survival prediction.

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

References

1. Geissler, K., Müller-Groh, L., Meine, H.: RadYOLO: Computationally efficient 3D object detection and segmentation in CT and MRI (2026), https://arxiv.org/abs/2608.00508
2. 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)
3. Isensee, F., Jäger, P., Wasserthal, J., Zimmerer, D., Petersen, J., Kohl, S., Schock, J., Klein, A., Roß, T., Wirkert, S., Neher, P., Dinkelacker, S., Köhler, G., Maier-Hein, K.: batchgenerators - a python framework for data augmentation (Jan 2020). https://doi.org/10.5281/zenodo.3632567, https://doi.org/10.5281/zenodo.3632567
4. Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC medical research methodology **18**(1), 24 (2018)
5. Saeed, N., Hassan, S., Hardan, S., Aly, A., Taratynova, D., Nawaz, U., Khan, U., Ridzuan, M., Andrearczyk, V., Depeursinge, A., Xie, Y., 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., Bourhis, D., Leclere, J.C., Sambourg, A., Rahmim, A., Hatt, M., Yaqub, M.: A multimodal and multi-centric head and neck cancer dataset for segmentation, diagnosis and outcome prediction (2025), https://arxiv.org/abs/2509.00367
6. Saeed, N., Hassan, S., Hardan, S., Cai, L., Liang, X., Mazher, M., Qayyum, A., Bu, Y., Lyu, M., Lin, Y., Meng, M., Huang, C., Wang, L., Chamseddine, D., Ahrari, S., Wu, B., Chen, Y., Mao, F., Zhang, H., Zhao, B., Ray, S., Guo, M., Xiang, L., Dexl, J., Ingrisch, M., Depeursinge, A., Rahmim, A., Hatt, M., Andrearczyk, V., Yaqub, M.: HEad and neCK TumOR (HECKTOR) 2025: Benchmark of segmentation, diagnosis, and prognosis in multimodal PET/CT (2026), https://arxiv.org/abs/2606.20143
7. 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) (2023)
8. Yu, C.N., Greiner, R., Lin, H.C., Baracos, V.: Learning patient-specific cancer survival distributions as a sequence of dependent regressors. Advances in neural information processing systems **24** (2011)
9. Zheng, Z., Wang, P., Ren, D., Liu, W., Ye, R., Hu, Q., Zuo, W.: Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE transactions on cybernetics **52**(8), 8574–8586 (2021)