Multi-Task Deep Learning for Head and Neck Cancer: Segmentation, Survival Prediction, and HPV Classification in the HECKTOR 2025 Challenge

Jakob Dexl[1,2][0000−0002−0617−0460] and Michael Ingrisch[1,2][0000−0003−0268−9078]

[1] Department of Radiology, University Hospital, LMU Munich, Germany
[2] Munich Center for Machine Learning (MCML)
jakob.dexl@lmu.de

Abstract. This paper describes our submissions (team CDS) to the HECKTOR 2025 challenge, which addresses three tasks: (1) tumor and lymph node segmentation, (2) recurrence-free survival prediction, and (3) HPV status classification. For Task 1, we trained a baseline UNet and refined the final model using stochastic weight averaging and small lesion removal. For Task 2, we employed a lightweight 3D ResNet18 that combines PET, CT, segmentation masks, and clinical metadata, optimized with a Cox loss. For Task 3, we extended the segmentation model with a classification head and metadata integration. Cross-validation results were promising, performance on the preliminary validation set was however lower, underlining the challenges of generalization in multi-center cohorts. Code and trained models are available at github.com/JakobDexl/HECKTOR25.

Keywords: Head and neck cancer · PET/CT · Segmentation · FDG

1 Introduction

Head and neck cancers are a diverse group of tumors that affect anatomically complex regions essential for breathing, swallowing, and speech. Despite advances in radiotherapy, chemotherapy, and targeted agents, treatment failure and disease recurrence remain common, leading to poor survival outcomes, particularly in advanced stages [2]. Clinical decision-making is typically based on imaging, histopathology, and staging systems, but these approaches only partly capture the biological and clinical heterogeneity of the disease. Additional biomarkers such as HPV status provide valuable prognostic information but are not yet consistently integrated into practice.

Recent progress in machine learning has demonstrated the potential of multimodal approaches that combine imaging and clinical data for segmentation, risk prediction, and biomarker classification. However, most studies are limited by small and homogeneous cohorts, restricting reproducibility and generalization. The HECKTOR 2025 challenge addresses this gap by providing a large, multi-center dataset of PET/CT images and clinical annotations. It enables systematic benchmarking across three tasks: tumor and lymph node segmentation, recurrence-free survival prediction, and HPV status classification. In this paper, we describe our methods for each task and report results obtained on the provided datasets.

2 Methods

2.1 Data

The challenge training dataset consists of 680 patients collected across seven institutions. For Task 2 and Task 3, smaller subsets were used, containing 522 and 542 cases, respectively. In Task 2, survival outcomes included 417 uncensored and 105 censored events. In Task 3, HPV status was strongly imbalanced, with 484 HPV-positive cases (90%) and 58 HPV-negative cases (10%). The preliminary validation set contained 128 patients for Task 1, 276 for Task 2, and 123 for Task 3. Further details on data collection and distribution are available in [7].

2.2 Preprocessing

All scans were cropped to a window size of 288 using the official preprocessing scripts. Data augmentation was performed with TrivialAugment [6], applying a single transformation per sample. We used affine transformations (random rotations within $\pm 15^\circ$, scaling in $[0.9, 1.2]$) and intensity augmentations (gamma adjustment, Gaussian noise, blurring). After augmentation, volumes were cropped to $(192, 192, 192)$. With low probability, the crop center was sampled randomly; otherwise, it was aligned with a foreground mask. Multimodal inputs were handled by concatenating CT and PET volumes.

2.3 Training and Evaluation

Models were trained using the Adam optimizer with a learning rate of 1×10^{-4}. Training lasted 500 epochs, evaluation and model selection followed the challenge metrics described in [7]. For Task 1, we used a small validation set of 36 samples to maximize training data. For Tasks 2 and 3, we employed 5-fold cross-validation. Models for Task 1 were selected based on a combined score of average Dice Similarity Coefficient (DSC) for primary tumor (GTVp), aggregated DSC for lymph nodes (GTVn Agg), and F1-score for lymph nodes (GTVn Agg F1). Models for Task 2 were selected by concordance index (c-index), and models for Task 3 by balanced accuracy. All training runs were executed on two Nvidia A100 GPUs in parallel. Batch sizes are stated per GPU.

2.4 Task-Specific Approaches

Task 1: Tumor Segmentation. We used a compact MONAI DynUNet [1]. The model was trained with DiceCE loss and a batch size of 2. The best checkpoint was refined with stochastic weight averaging (SWA) [4]. As post-processing, lesions smaller than 100 voxels were removed.

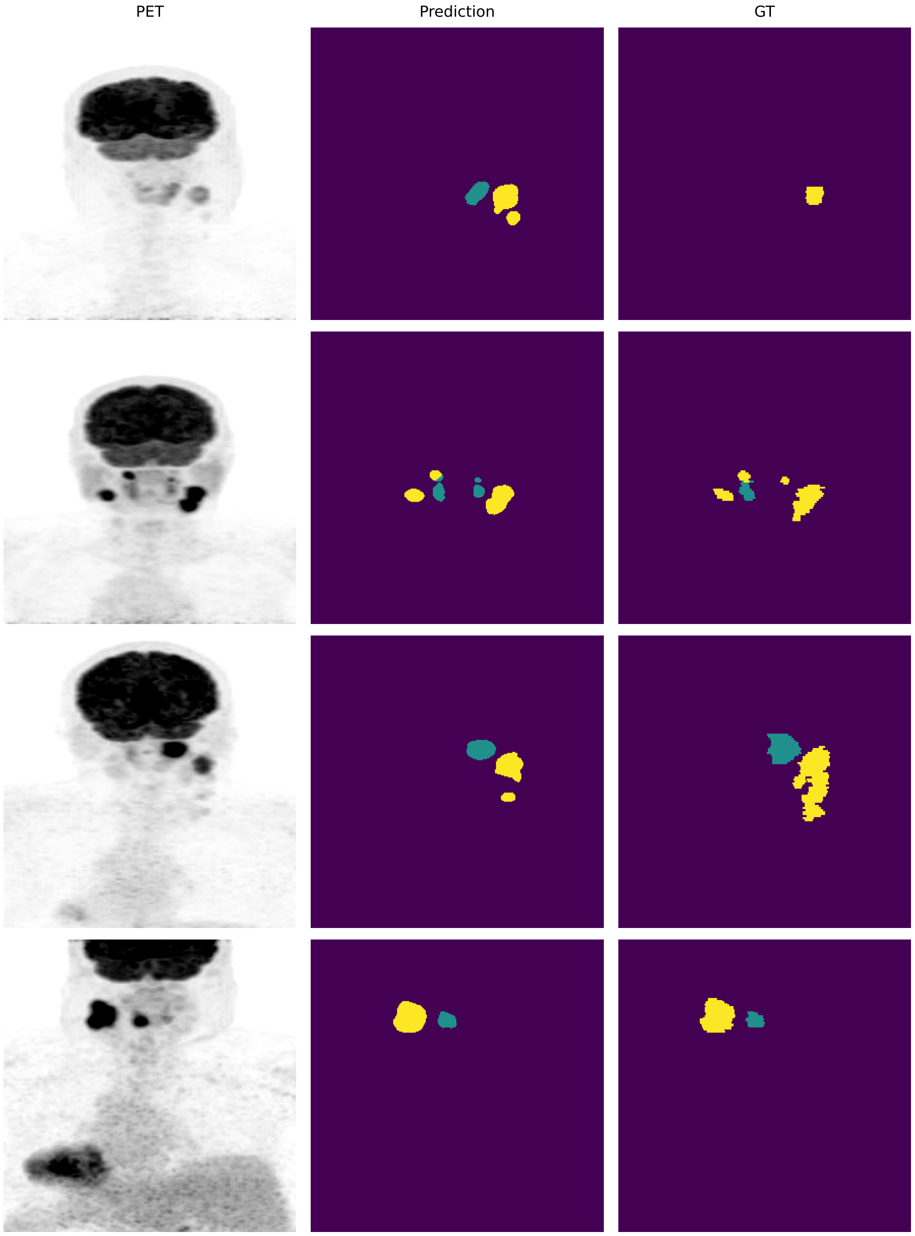

Fig. 1. Maximum intensity projections (MIP) of the PET, predicted mask, and ground truth. PET volumes are shown with center=3 and window=7. Primary tumors are colored in green, lymph node masks are yellow. Predictions generally exhibit smoother boundaries. Row one shows a case without primary tumor and the hallucinated prediction. Row two shows an error where a small lymph node is misclassified as primary tumor. Row three shows a larger lymph node divided into two smaller predicted nodes. The last row shows a good-performing case.

Task 2: Survival Prediction. We employed a lightweight 3D ResNet18 encoder [8] with PET, CT, ground-truth masks, and clinical metadata (HPV status, tobacco and alcohol consumption, age, gender, performance status, treatment, and M-stage) as input. Metadata were encoded with a two-layer MLP and fused with image features before classification. Missing values were set to -1, and age was z-normalized. The model was trained using the negative Cox partial log-likelihood, similar to DeepSurv [5]. For stability, predicted risks were centered, sorted by survival time, and aggregated via cumulative log-sum-exp. The final loss was the negative average contribution of observed events. A batch size of 16 ensured sufficient event signal per update. The final submission used the best fold model. Additionally, we used the Task 1 segmentation model for mask input and the inferred HPV status from a single-fold Task 3 model.

Task 3: Combined Segmentation and Classification. The segmentation model from Task 1 was extended with a wide linear bottleneck head for HPV prediction. Metadata (age, gender, tobacco and alcohol consumption) were incorporated as in Task 2. Training used a combination of DiceCE loss for segmentation and binary cross-entropy for classification, with a batch size of 4. The minority class was oversampled, and each batch included two HPV-positive and two HPV-negative cases. For the final submission, the segmentation decoder was removed for efficiency, and cross-validation models were ensembled by averaging prediction logits.

Table 1. Results for Task 1 on the Local Validation data, Preliminary Validation, and Final Test.

	Local Validation	Preliminary Validation	Final Test
GTVp DSC	0.7057	0.7431	0.7300
GTVn Agg DSC	0.7726	0.7219	0.7608
GTVn Agg F1	0.7520	0.6502	0.6239

3 Results

The results of our segmentation model are shown in Table 1.The model achieved a GTVp of 0.7057, a GTVn Agg of 0.7726, and a GTVn Agg F1-score of 0.7520 on the local validation set. SWA improved the average DSC for the primary tumor by approximately 0.030 and the F1-score by 0.060, at the cost of a slight decrease in aggregated DSC for lymph nodes (–0.005). On the preliminary validation and final test sets, performance on the primary tumor class increased,

Table 2. Results for Task 3 cross-validation and preliminary validation and final test of the ensemble.

	Folds					Preliminary Validation	Final Test
	F1	F2	F3	F4	F5		
Balanced Accuracy	0.8443	0.8753	0.8622	0.8681	0.9201	0.7085	0.4908

whereas performance on the lymph nodes decreased, particularly for the GTVn Agg F1. Figure 1 shows qualitative segmentation results. Predictions were generally solid, though boundary accuracy was often imprecise and the predicted contours appeared smoother than in the ground truth. Larger lymph nodes were occasionally split into smaller components. In two test cases without primary tumors, the model hallucinated small primary tumors adjacent to lymph nodes.

The best-performing Task 2 fold achieved a c-index of 0.7442, but performance dropped to 0.4458 on the preliminary validation set and further to 0.3791 on the final test set.

Task 3 results are presented in Table 2. Performance dropped substantially on the preliminary validation set. We also submitted only the best fold (F5), which achieved a balanced accuracy of 0.6158, demonstrating that ensembling improved performance. To assess the contribution of the clinical variables, we conducted a small ablation study in which each variable was sequentially set to unknown (Table 3). Age had the largest impact on performance, whereas sex contributed almost no additional information. Alcohol consumption and smoking status had only minor effects.

Table 3. Ablation impact of removing clinical variables (Age, Sex, Alcohol consumption, and Smoking status) in Task 3. Variables in red were mask with the unknown identifier.

				Balanced Accuracy
Age	Sex	Alc	Smo	0.8790
🟥 Age	Sex	Alc	Smo	0.8153
Age	🟥 Sex	Alc	Smo	0.8790
Age	Sex	🟥 Alc	Smo	0.8636
Age	Sex	Alc	🟥 Smo	0.8736
🟥 Age	🟥 Sex	🟥 Alc	🟥 Smo	0.7957

4 Discussion

The presented methods for Task 1 and Task 3 achieved promising results on preliminary validation. Our overall strategy was to keep the models simple. Initially, we aimed to train a single multitask model, but stability issues prevented this. The multitask model resembled the Task 3 architecture with an additional survival head. However, the large model size limited the batch size, destabilizing the Cox loss. Consequently, we opted for task-specific models, such as a lightweight ResNet18 for Task 2. It is quite surprising that performance varied substantially between the local, preliminary, and final evaluations, indicating a potential domain or distribution shift.

Task 2 performed considerably worse on preliminary validation. We attribute this to the reliance on perfect ground-truth masks and HPV status during training, whereas the final submission used predicted masks and HPV status. Moreover, HPV inference relied on a single fold model instead of an ensemble, further degrading performance. This error propagation likely explains the large performance drop.

For Task 3, performance differed between local cross-validation and preliminary and final validation. Two possible factors may explain this. First, a label shift may exist between the training and validation sets, as suggested by dataset distribution differences [7]. Second, metadata stratification was not considered, despite known correlations (e.g., higher HPV incidence in men than in women) [3].

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

References

1. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Darestani, M.Z., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B.S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P.F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L.A.D., Roth, H.R., Xu, D., Bericat, D., Floca, R., Zhou, S.K., Shuaib, H., Farahani, K., Maier-Hein, K.H., Aylward, S., Dogra, P., Ourselin, S., Feng, A.: MONAI: An open-source framework for deep learning in healthcare (Nov 2022). https://doi.org/10.48550/arXiv.2211.02701, mONAI
2. Chajon, E., Lafond, C., Louvel, G., Castelli, J., Williaume, D., Henry, O., Jégoux, F., Vauléon, E., Manens, J.P., Le Prisé, E., de Crevoisier, R.: Salivary gland-sparing other than parotid-sparing in definitive head-and-neck intensity-modulated radiotherapy does not seem to jeopardize local control. Radiation Oncology **8**(1), 132 (May 2013). https://doi.org/10.1186/1748-717X-8-132

3. D'Souza, G., Westra, W.H., Wang, S.J., van Zante, A., Wentz, A., Kluz, N., Rettig, E., Ryan, W.R., Ha, P.K., Kang, H., Bishop, J., Quon, H., Kiess, A.P., Richmon, J.D., Eisele, D.W., Fakhry, C.: Differences in the Prevalence of Human Papillomavirus (HPV) in Head and Neck Squamous Cell Cancers by Sex, Race, Anatomic Tumor Site, and HPV Detection Method. JAMA Oncology **3**(2), 169–177 (Feb 2017). https://doi.org/10.1001/jamaoncol.2016.3067
4. Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., Wilson, A.G.: Averaging Weights Leads to Wider Optima and Better Generalization (Feb 2019). https://doi.org/10.48550/arXiv.1803.05407
5. 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 (Feb 2018). https://doi.org/10.1186/s12874-018-0482-1
6. Müller, S.G., Hutter, F.: TrivialAugment: Tuning-Free Yet State-of-the-Art Data Augmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 774–782 (2021)
7. 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
8. Solovyev, R., Kalinin, A.A., Gabruseva, T.: 3D convolutional neural networks for stalled brain capillary detection. Computers in Biology and Medicine **141**, 105089 (Feb 2022). https://doi.org/10.1016/j.compbiomed.2021.105089