Improving Survival Prediction of Head-and-Neck Cancer with Medical Image, Foundation Models and Multi-modal Fusion

Haotian Zhang ZHANGHT@UMICH.EDU
Yiming Liu LIUYM@UMICH.EDU
Yile Sun SUNYYYL@UMICH.EDU
Liyue Shen LIYUES@UMICH.EDU
Lise Wei LISWEI@UMICH.EDU
University of Michigan, USA

Abstract

Accurate survival prediction in head and neck squamous cell carcinoma (HNSCC) is critical for guiding treatment stratification but remains difficult due to small cohort sizes and the suboptimal integration of multimodal medical data. In particular, imaging features are high-dimensional, while clinical covariates are low-dimensional, making effective multimodal fusion non-trivial. To address these limitations, we propose to leverage pretrained modality-specific medical image foundation models (FMs), including those for CT and PET, to improve image representation learning under small-sample constraints. These models can augment a multimodal fusion baseline to enhance the structured fusion of imaging and clinical features. Specifically, FM-derived embeddings are passed through compact linear projection heads and fused by direct concatenation immediately before the survival prediction head. We systematically study the projection dimensionality and modality composition (CT-only, PET-only, CT+PET) and evaluate the performance using the average concordance index (C-index) and time-dependent AUROC under five-fold cross-validation. Preliminary results demonstrate that image embeddings extracted from pretrained medical foundation models consistently improve C-index and time-dependent AUROC, and reduce their fold-to-fold variance, with the CT+PET setting providing the most robust gains. These findings suggest that FM-based multimodal survival models can enhance risk stratification and ultimately support personalized treatment adaptation in HNSCC.

Keywords: HNSCC, medical foundation models, PET/CT, multimodal fusion, survival prediction

Data and Code Availability We use the HECKTOR 2022 public training split (524 PET/CT cases from 7 centers)[1] with co-registered PET/CT, tumor masks in the CT frame, recurrence-free survival (days), and 9 clinical covariates; evaluation follows the challenge protocol (C-index) and our additional metric (time-dependent AUROC)(Andrearczyk et al., 2023). The code will be released to GitHub[2] soon.

Institutional Review Board (IRB) IRB is not required for this work since we only use public data and models for the study.

1. Introduction

Head and neck squamous cell carcinoma (HNSCC) outcome prediction is clinically important yet difficult due to some restrictions such as small dataset size and the challenge of multimodal data fusion (Li et al., 2024). Prior end-to-end imaging pipelines can degrade under label scarcity, modality imbalance, and variability of data acquisition, which limits their reliability in realistic applications (Adeoye et al., 2023). Foundation models (FMs) offer a potential alternative: large-scale pretraining produces transferable visual representations that can be reused without task-specific fine-tuning, potentially improving stability under limited and noisy data (Moor et al., 2023). Recent FM studies also report improved cross-site gener-

[1] https://hecktor.grand-challenge.org/
[2] https://github.com/ZZHT666/DeepMTS-FM

alization and robustness to scanner variation in small dataset (Pai et al., 2024, 2025; Zhang et al., 2025).

Motivated by these ideas, we study an integration of modality-specific FMs into a strong recurrence-free survival prediction baseline. Concretely, we adopt a DeepMTS-style architecture for HNSCC survival prediction—i.e., a PET/CT segmentation backbone whose tumor probability map is thresholded and used to gate a cascaded survival network (CSN)—and equip it with a DAFT block, a dynamic affine feature-map transform that predicts channel-wise scale/shift from the clinical vector to modulate the last CSN features for image–tabular fusion (Pölsterl et al., 2021); modality-specific encoders (CT-FM, PET-FM) are kept frozen to extract fixed embeddings without any fine-tuning, which are passed through compact linear projection heads and late-fused by direct concatenation immediately upstream of the survival head (Meng et al., 2021).

We evaluate this design with five-fold cross-validation and assess performance using the average validation concordance index (C-index) and 1-year time-dependent AUROC among all five folds. In order to better understand how the FM embeddings influence survival prediction, we conduct an ablation study examining (i) projection dimensionality and (ii) modality composition (CT-only, PET-only, CT+PET). Beyond accuracy, the frozen-FM late-concatenation setup reduces trainable parameters and overfitting risk in low-data regimes. This strategy not only benchmarks FM utility under survival endpoints but also propose a lightweight way for incorporating stronger pretrained encoders as they emerge, enabling a more scalable and generalizable translation into clinical applications.

Compared with recent HNSCC outcome models, our contribution is a plug-and-play use of modality-specific frozen CT/PET FMs on a strong DeepMTS+DAFT survival baseline, yielding consistent gains with far fewer trainable parameters and lower fold-to-fold variance. For example, Ma et al. propose TransRP and show that PET/CT transformer features with clinical covariates improve multi-outcome prediction in OPSCC, but their pipeline is end-to-end and not FM-centric (Ma et al., 2024). Tian et al. develop a large-scale multimodal model (CT+WSI+clinical) with external validation for HNSCC prognosis/RT response, again using end-to-end learned features rather than frozen modality-specific FMs (Tian et al., 2025). Wang et al. integrate image and clinical text using cross-attention for survival analysis in HNC, showing the value of multimodal fusion but not leveraging medical FMs for CT/PET (Wang et al., 2024). Meneghetti et al. study FM-based multiple-instance learning primarily on histopathology (WSI) for HNSCC outcomes (Meneghetti et al., 2025); in contrast, we systematically map projection size and modality composition for radiology CT/PET FMs and quantify the stability benefits under small-sample regimes.

2. Method

2.1. Baseline Model

DeepMTS. We adopt DeepMTS as the survival baseline due to its hybrid multi-task design that jointly learns tumor segmentation and survival risk from 3D PET/CT. Concretely, a segmentation backbone (3D U-Net style with residual blocks and skip connections) acts as a hard-sharing encoder to focus feature extraction around primary tumors. The Cascaded Survival Network (CSN; a modified 3D DenseNet) receives PET/CT together with the predicted tumor probability map to capture prognostic context around the tumor. Deep features from the hard-sharing path and from the CSN are globally averaged and passed to fully-connected layers (with Cox partial log-likelihood for survival and Dice for segmentation), concatenated with clinical covariates, following the original formulation. This architecture balances local tumor cues and global context and is a strong baseline for PET/CT survival modeling (Meng et al., 2021).

DAFT block and integration. To condition image features on tabular clinical information for better data fusion, we employ the Dynamic Affine Feature Map Transform (DAFT). DAFT takes a high-level convolutional feature tensor and a clinical covariate vector, and learns per-channel scale/shift (α, β) via a lightweight auxiliary MLP. The MLP pools the feature maps, concatenates the clinical inputs, and produces an affine modulation $F'_c = \alpha_c F_c + \beta_c$. Here, $F \in \mathbb{R}^{C\times H\times W\times D}$ is the feature tensor, $c \in \{1, \ldots, C\}$ indexes channels, and F_c is the c-th channel map; α_c, β_c are scalars applied per channel and broadcast over spatial dimensions.

We insert a single DAFT module inside DeepMTS's CSN, placed after the last Dense block and before the global average pooling (GAP). This location allows DAFT to modulate the most semantic CSN representations using clinical context. The GAP-pooled CSN vector is then concatenated with the segmentation-backbone vector (and, optionally, clinical inputs) before being fed to the survival head. This plug-in preserves the original training scheme of DeepMTS while providing a principled image–tabular interaction with minimal parameters (Pölsterl et al., 2021).

2.2. Foundation Model

CT foundation model (CT-FM). We use the public CT-FM encoder (Pai et al., 2025) (SegRes-style 3D residual backbone) strictly in frozen mode. Each CT is preprocessed with intensities in Hounsfield units clipped to $[-1000, 2000]$ and min–max normalized to $[0, 1]$. A single forward pass yields a high-level volumetric feature map, which we global-average-pool over (D', H', W') to obtain a fixed 512-D CT embedding. This vector is then passed through a linear projection and concatenated immediately before the survival head.

PET foundation model (PET-FM). For PET, we adopt the SegAnyPET image encoder (Zhang et al., 2025) as a frozen 3D ViT: a 16^3 convolutional patch embed followed by multi-head self-attention blocks. PET volumes are percentile-clipped (0.5–99.5) and min–max normalized to $[0, 1]$, then padded/cropped to $128\times128\times128$. The encoder outputs a token sequence $[N \times C]$; mean-token pooling gives a global PET embedding ($C = 768$), which we linearly project and concatenate with the CT-FM projection upstream of the survival head.

The whole network structure of the model can be viewed in Figure 1. In this figure, the detailed implementation of CSN, segmentation backbone and DAFT block is omitted. Please follow the work of DeepMTS (Meng et al., 2021) and DAFT (Pölsterl et al., 2021) for more details.

3. Experiments and Results

3.1. Data Processing

We apply a concise, center-aligned pipeline to standardize PET/CT and masks:

- **Spacing normalization.** Resample CT, PET, and mask to a common voxel size of $1\times1\times3$ mm (linear interpolation for CT/PET; nearest-neighbor for masks).

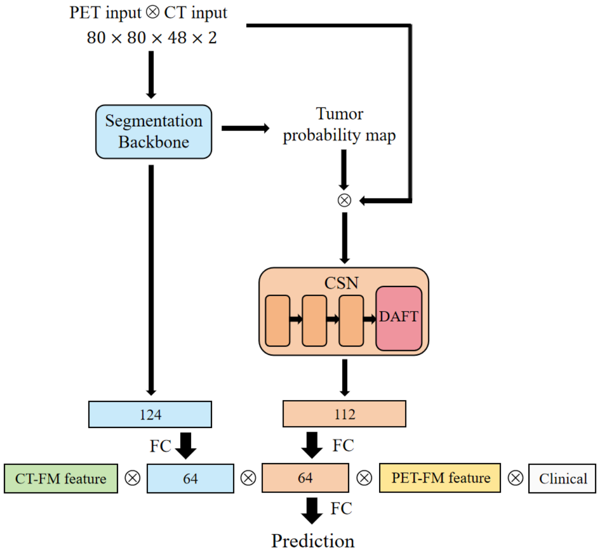


Figure 1: Network structure of the overall model.

- **Global crop sizing.** On the resampled masks, compute each case's tumor bounding box around the mask centroid. Aggregate per-axis maxima across all cases to obtain a global minimal crop size **G**. This crop size ensures that all tumor area are included after cropping without losing important information.
- **Center-aligned cropping.** For each case, align the output grid to the case-specific center, then extract a centered crop of size **G**. Apply constant padding when needed (image volumes padded with the per-volume minimum intensity; masks with 0). Image direction/origin are preserved.
- **Task-specific resizing.** From the centered crop, resize (size-only resampling; linear for CT/PET, nearest-neighbor for masks) to certain sizes: $80\times80\times48$ for DeepMTS input, $128\times128\times48$ for CT-FM input and $128\times128\times128$ for PET-FM input.

3.2. Experimental Setup

We use patient-level 5-fold stratified cross-validation (stratified by event indicator and Recurrence-Free Survival (RFS) quartiles). The objective function is $L = L_{\text{dice}} + 2\, L_{\text{cox}}$, and the primary metric is C-index and 1-year time-dependent AUROC on the validation split. Optimization uses Adam (lr 10^{-3}) with 5-epoch warm-up and step decay (×0.3 every 25 epochs) for at most 100 epochs. Regularization includes Exponential Moving Average (EMA; τ=0.999) applied at evaluation to stabilize the training curve (Morales-Brotons et al., 2024). Within each fold we select the checkpoint with the highest validation C-index and highest 1-year time-dependent AUROC, the we report mean±SD across folds.

3.3. Projection Dimensionality

We evaluate projecting the frozen FM embeddings with a single linear head to lengths $d\in\{64, 128, 256\}$ and compare against each FM's original size (CT: 512; PET: 768). Figure 2 shows 5-fold mean±SD C-index and Figure 3 shows 5-fold mean±SD 1-year AUROC .

CT-FM. For C-index, compressing CT feature to: d=64/128/256 obtain 0.599/0.647/0.621 (SD 0.066/0.047/0.158), while the original 512-d achieves **0.715±0.075**. This suggests that CT encodes prognostic detail that is lost under aggressive bottlenecks. For 1-year AUROC, the native 512-d CT likewise performs best at **0.710±0.083**, while aggressive bottlenecks underperform, mirroring the C-index pattern.

PET-FM. We can observe from C-index metric that PET benefits from a moderate projection: d=128 reaches **0.736±0.060**, surpassing both d=64/256 (0.685/0.694) and the original 768-d (0.700±0.058). The gain with d=128 is consistent with projection acting as a regularizer that attenuates PET noise/heterogeneity. The 1-year AUROC shows the same preference for a mid-sized bottleneck: d=128 attains **0.712±0.100**, exceeding both smaller/larger d and the native 768-d, consistent with the C-index trend.

CT-FM+PET-FM. In C-index metric, late concatenation is strongest with a mid-sized bottleneck: d=256 yields **0.743±0.067**, outperforming d=64/128 (0.735/0.712) and the original variant (0.691). Larger d improves complementarity without overfitting, whereas using original dimension appears over-parameterized for our dataset size. Across all settings, the ablation confirms that projection helps PET and multimodal fusion but not CT alone. Consistently on 1-year AUROC, the late-fused d=256 variant achieves **0.744±0.094**, aligning with the C-index winner.

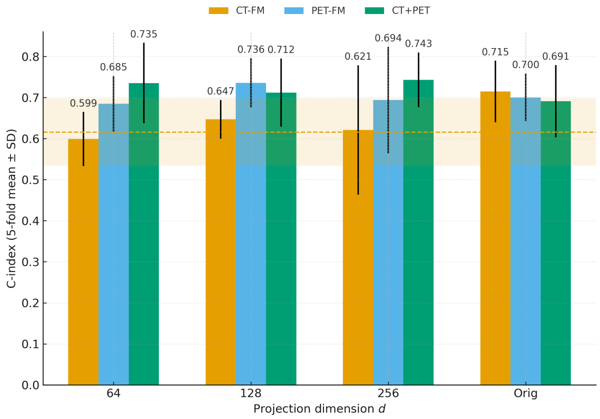


Figure 2: Projection dimension ablation. Grouped bars show 5-fold mean±SD C-index across $d \in \{64, 128, 256, \mathrm{Orig}\}$ for CT-FM, PET-FM, and late-fused CT+PET. The dashed line/band denote the baseline mean and ±SD.

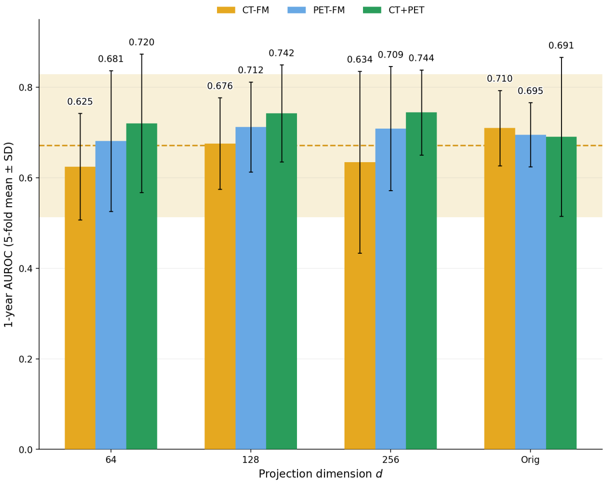


Figure 3: Projection dimension ablation. Grouped bars show 5-fold mean±SD 1-year AUROC across $d \in \{64, 128, 256, \mathrm{Orig}\}$ for CT-FM, PET-FM, and late-fused CT+PET. The dashed line/band denote the baseline mean and ±SD.

Table 1: Overall comparison across two metrics (5-fold). Best projection per configuration is used (CT: d=512; PET: d=128; CT+PET: d=256).

Method	C-index ↑		1-year AUROC ↑	
	mean±SD	Δ vs. Base	mean±SD	Δ vs. Base
DeepMTS+DAFT (baseline)	0.616 ± 0.082	–	0.671 ± 0.158	–
+ CT-FM (d=512)	0.715 ± 0.075	+0.099	0.710 ± 0.083	+0.039
+ PET-FM (d=128)	0.736 ± 0.060	+0.120	0.712 ± 0.100	+0.041
+ CT+PET (d=256)	$\mathbf{0.743 \pm 0.067}$	**+0.127**	$\mathbf{0.744 \pm 0.094}$	**+0.073**

3.4. Overall Comparison

Using the best d per configuration (Table 1), every FM-augmented model improves upon the DeepMTS+DAFT baseline (0.616±0.082). CT-FM (512-d) already delivers a sizable gain (+0.099) with reduced variance (0.075). PET-FM at d=128 further improves (+0.120; SD 0.060), indicating that frozen PET embeddings provide noise-tolerant signal once modestly regularized. The late-fused **CT+PET** at d=256 attains the best overall C-index (**0.743±0.067**, **+0.127**), and the lowest variance among multimodal variants. Because encoders are frozen, training cost and hyperparameter footprint remain minimal—the only added learnables are the lightweight projection layers (when used) and the survival head—while robustness improves under heterogeneous PET/CT quality. On 1-year AUROC, the ordering is the same (CT+PET d=256 > PET d=128 > CT 512-d > baseline), corroborating the C-index ranking.

4. Conclusion and Discussion

Conclusion. We introduced a plug-and-play way to strengthen HNSCC survival prediction by injecting frozen PET/CT foundation-model (FM) embeddings into a DeepMTS+DAFT baseline (Meng et al., 2021; Pölsterl et al., 2021). With simple linear projections and late concatenation, all FM variants improved or matched the baseline while reducing fold-to-fold variance. PET-FM delivered a larger standalone improvements than CT-FM, and the best, most stable performance was obtained when fusing PET and CT features together (Pai et al., 2025; Zhang et al., 2025). This indicates that foundation model have the ability to improve the average model performance while increasing robustness.

Discussion. Following these results, we explore possible explanations for why PET-FM features outperform CT-FM features. One possible explanation is that the amount of information carried varies between modalities. PET carries metabolic signal that is directly related to tumor activity and outcome, while CT mainly encodes morphology. This view is supported by a head-to-head radiomics comparison in HNSCC showing that PET-derived features—particularly intensity/heterogeneity descriptors of intratumoral metabolism—provide more stable and generalizable prediction of local tumor control than CT-derived features; CT models often lost discrimination and tended to be optimistic on validation, whereas PET models retained performance (Bogowicz et al., 2017). After resampling, interpolation, and normalization, high-frequency CT textures are weakened; in our ablation study, projecting to d=256 regularizes CT-FM more than PET-FM, which is consistent with CT having less diagnostic information and ability. Finally, other factors, such as train–validation split and preprocessing methods may also affect the quality of the extracted CT features.

Future work. We will probe three axes—models, data, and evaluation. On the model side, we plan to swap in diverse CT/PET foundation backbones and vary insertion points (with light fine-tuning), and to compare fusion schemes beyond late concatenation, including gated addition, cross-attention, and DAFT-conditioned adapters at different CSN depths. On the data side, we aim to enlarge the cohort and test robustness with center/site–stratified and leave-one-site-out splits, as well as controlled noise/blur and slice-thickness perturbations. Finally, we will broaden evaluation to integrated Brier score, calibration, and decision-curve analysis to better assess clinical utility.

References

John Adeoye, Liuling Hui, and Yu-Xiong Su. Data-centric artificial intelligence in oncology: a systematic review assessing data quality in machine learning models for head and neck cancer. *Journal of Big Data*, 10:28, 2023. doi: 10.1186/s40537-023-00703-w.

Vincent Andrearczyk, Valentin Oreiller, Moamen Abobakr, Azadeh Akhavanallaf, Panagiotis Balermpas, Sarah Boughdad, Leo Capriotti, Joel Castelli, Catherine Chéze Le Rest, Pierre Decazes, Ricardo Correia, Dina El-Habashy, Hesham Elhalawani, Clifton D. Fuller, Mario Jreige, Yornna Khamis, Agustina La Greca, Abdallah Mohamed, Mohamed Naser, John O. Prior, Su Ruan, Stephanie Tanadini-Lang, Olena Tankyevych, Yazdan Salimi, Martin Vallières, Pierre Vera, Dimitris Visvikis, Kareem Wahid, Habib Zaidi, Mathieu Hatt, and Adrien Depeursinge. Overview of the hecktor challenge at miccai 2022: Automatic head and neck tumor segmentation and outcome prediction in pet/ct. In *Head and Neck Tumor Segmentation and Outcome Prediction. HECKTOR 2022*, Lecture Notes in Computer Science, pages 1–30. 2023. doi: 10.1007/978-3-031-27420-6_1.

Marta Bogowicz, Oliver Riesterer, Laura S. Stark, Gabriela Studer, Jan Unkelbach, Matthias Guckenberger, and Sabina Tanadini-Lang. Comparison of pet and CT radiomics for prediction of local tumor control in head and neck squamous cell carcinoma. *Acta Oncologica*, 56(11):1531–1536, 2017. doi: 10.1080/0284186X.2017.1346382.

Yihao Li, Mostafa El Habib Daho, Pierre-Henri Conze, and Others. A review of deep learning-based information fusion techniques for multimodal medical image classification. *Computers in Biology and Medicine*, 177:108635, 2024. doi: 10.1016/j.compbiomed.2024.108635.

Baoqiang Ma, Jiapan Guo, Alessia De Biase, Lisanne V. van Dijk, Peter M. A. van Ooijen, Johannes A. Langendijk, Stefan Both, and Nanna M. Sijtsema. Pet/ct based transformer model for multi-outcome prediction in oropharyngeal cancer. *Radiotherapy and Oncology*, 197:110368, 2024. doi: 10.1016/j.radonc.2024.110368.

André R. Meneghetti, Dennis Bontempi, Ibrahim Hadzic, Vasco Prudente, Matteo Sokac, Tafadzwa L. Chaunzwa, Simon Bernatz, Ahmed Hosny, Raymond H. Mak, Nicolai J. Birkbak, and Hugo J. W. L. Aerts. End-to-end prediction of clinical outcomes in head and neck squamous cell carcinoma with foundation model-based multiple instance learning. *BMC Artificial Intelligence*, 2 (1):3, 2025. doi: 10.1186/s44398-025-00003-8.

Mingyuan Meng, Bingxin Gu, Lei Bi, Shaoli Song, David Dagan Feng, and Jinman Kim. Deepmts: Deep multi-task learning for survival prediction in patients with advanced nasopharyngeal carcinoma using pretreatment pet/ct, 2021. URL https://arxiv.org/abs/2109.07711.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023. doi: 10.1038/s41586-023-05881-4.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits. *Transactions on Machine Learning Research*, 2024. doi: 10.48550/arXiv.2411.18704. URL https://arxiv.org/abs/2411.18704. arXiv:2411.18704.

Suraj Pai, Dennis Bontempi, Ibrahim Hadzic, Vasco Prudente, Mateo Sokač, Tafadzwa L. Chaunzwa, Simon Bernatz, Ahmed Hosny, Raymond H. Mak, Nicolai J. Birkbak, and Hugo J. W. L. Aerts. Foundation model for cancer imaging biomarkers. *Nature Machine Intelligence*, 6(3):354–367, 2024. doi: 10.1038/s42256-024-00807-9.

Suraj Pai, Ibrahim Hadzic, Dennis Bontempi, Keno Bressem, Benjamin H. Kann, Andriy Fedorov, Raymond H. Mak, and Hugo J. W. L. Aerts. Vision foundation models for computed tomography, 2025. URL https://arxiv.org/abs/2501.09001.

Sebastian Pölsterl, Tom Nuno Wolf, and Christian Wachinger. Combining 3d image and tabular data via the dynamic affine feature map transform, 2021. URL https://arxiv.org/abs/2107.05990.

Ruxian Tian, Feng Hou, Haicheng Zhang, Guohua Yu, Ping Yang, Jiaxuan Li, Ting Yuan, Xi Chen, Ying Chen, Yan Hao, Yisong Yao, Hongfei Zhao, Pengyi Yu, Han Fang, Liling Song, Anning Li, Zhonglu Liu, Huaiqing Lv, Dexin Yu, Hongxia Cheng, Ning Mao, and Xicheng Song. Multimodal fusion model for prognostic prediction and radiotherapy response assessment in head and neck squamous cell carcinoma. *npj Digital Medicine*, 8 (302), 2025. doi: 10.1038/s41746-025-01712-0.

Zhaonian Wang, Chundan Zheng, Xu Han, Wufan Chen, and Lijun Lu. An innovative and efficient diagnostic prediction flow for head and neck cancer: A deep learning approach for multi-modal survival analysis prediction based on text and multi-center pet/ct images. *Diagnostics*, 14(4):448, 2024. doi: 10.3390/diagnostics14040448.

Yichi Zhang, Le Xue, Wenbo Zhang, Lanlan Li, Yuchen Liu, Chen Jiang, Yuan Cheng, and Yuan Qi. Seganypet: Universal promptable segmentation from positron emission tomography images, 2025. URL https://arxiv.org/abs/2502.14351.

Acknowledgments

This work is supported by the Michigan Institute for Data Science (MIDAS) PODS Grant.

References

John Adeoye, Liuling Hui, and Yu-Xiong Su. Data-centric artificial intelligence in oncology: a systematic review assessing data quality in machine learning models for head and neck cancer. *Journal of Big Data*, 10:28, 2023. doi: 10.1186/s40537-023-00703-w.

Vincent Andrearczyk, Valentin Oreiller, Moamen Abobakr, Azadeh Akhavanallaf, Panagiotis Balermpas, Sarah Boughdad, Leo Capriotti, Joel Castelli, Catherine Chéze Le Rest, Pierre Decazes, Ricardo Correia, Dina El-Habashy, Hesham Elhalawani, Clifton D. Fuller, Mario Jreige, Yornna Khamis, Agustina La Greca, Abdallah Mohamed, Mohamed Naser, John O. Prior, Su Ruan, Stephanie Tanadini-Lang, Olena Tankyevych, Yazdan Salimi, Martin Vallières, Pierre Vera, Dimitris Visvikis, Kareem Wahid, Habib Zaidi, Mathieu Hatt, and Adrien Depeursinge. Overview of the hecktor challenge at miccai 2022: Automatic head and neck tumor segmentation and outcome prediction in pet/ct. In *Head and Neck Tumor Segmentation and Outcome Prediction. HECKTOR 2022*, Lecture Notes in Computer Science, pages 1–30. 2023. doi: 10.1007/978-3-031-27420-6_1.

Marta Bogowicz, Oliver Riesterer, Laura S. Stark, Gabriela Studer, Jan Unkelbach, Matthias Guckenberger, and Sabina Tanadini-Lang. Comparison of pet and CT radiomics for prediction of local tumor control in head and neck squamous cell carcinoma. *Acta Oncologica*, 56(11):1531–1536, 2017. doi: 10.1080/0284186X.2017.1346382.

Yihao Li, Mostafa El Habib Daho, Pierre-Henri Conze, and Others. A review of deep learning-based information fusion techniques for multimodal medical image classification. *Computers in Biology and Medicine*, 177:108635, 2024. doi: 10.1016/j.compbiomed.2024.108635.

Baoqiang Ma, Jiapan Guo, Alessia De Biase, Lisanne V. van Dijk, Peter M. A. van Ooijen, Johannes A. Langendijk, Stefan Both, and Nanna M. Sijtsema. Pet/ct based transformer model for multi-outcome prediction in oropharyngeal cancer. *Radiotherapy and Oncology*, 197:110368, 2024. doi: 10.1016/j.radonc.2024.110368.

André R. Meneghetti, Dennis Bontempi, Ibrahim Hadzic, Vasco Prudente, Matteo Sokac, Tafadzwa L. Chaunzwa, Simon Bernatz, Ahmed Hosny, Raymond H. Mak, Nicolai J. Birkbak, and Hugo J. W. L. Aerts. End-to-end prediction of clinical outcomes in head and neck squamous cell carcinoma with foundation model-based multiple instance learning. *BMC Artificial Intelligence*, 2 (1):3, 2025. doi: 10.1186/s44398-025-00003-8.

Mingyuan Meng, Bingxin Gu, Lei Bi, Shaoli Song, David Dagan Feng, and Jinman Kim. Deepmts: Deep multi-task learning for survival prediction in patients with advanced nasopharyngeal carcinoma using pretreatment pet/ct, 2021. URL https://arxiv.org/abs/2109.07711.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023. doi: 10.1038/s41586-023-05881-4.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits. *Transactions on Machine Learning Research*, 2024. doi: 10.48550/arXiv.2411.18704. URL https://arxiv.org/abs/2411.18704. arXiv:2411.18704.

Suraj Pai, Dennis Bontempi, Ibrahim Hadzic, Vasco Prudente, Mateo Sokač, Tafadzwa L. Chaunzwa, Simon Bernatz, Ahmed Hosny, Raymond H. Mak, Nicolai J. Birkbak, and Hugo J. W. L. Aerts. Foundation model for cancer imaging biomarkers. *Nature Machine Intelligence*, 6(3):354–367, 2024. doi: 10.1038/s42256-024-00807-9.

Suraj Pai, Ibrahim Hadzic, Dennis Bontempi, Keno Bressem, Benjamin H. Kann, Andriy Fedorov, Raymond H. Mak, and Hugo J. W. L. Aerts. Vision foundation models for computed tomography, 2025. URL https://arxiv.org/abs/2501.09001.

Sebastian Pölsterl, Tom Nuno Wolf, and Christian Wachinger. Combining 3d image and tabular data via the dynamic affine feature map transform, 2021. URL https://arxiv.org/abs/2107.05990.

Ruxian Tian, Feng Hou, Haicheng Zhang, Guohua Yu, Ping Yang, Jiaxuan Li, Ting Yuan, Xi Chen, Ying Chen, Yan Hao, Yisong Yao, Hongfei Zhao, Pengyi Yu, Han Fang, Liling Song, Anning Li, Zhonglu Liu, Huaiqing Lv, Dexin Yu, Hongxia Cheng, Ning Mao, and Xicheng Song. Multimodal fusion model for prognostic prediction and radiotherapy response assessment in head and neck squamous cell carcinoma. *npj Digital Medicine*, 8 (302), 2025. doi: 10.1038/s41746-025-01712-0.

Zhaonian Wang, Chundan Zheng, Xu Han, Wufan Chen, and Lijun Lu. An innovative and efficient diagnostic prediction flow for head and neck cancer: A deep learning approach for multi-modal survival analysis prediction based on text and multi-center pet/ct images. *Diagnostics*, 14(4):448, 2024. doi: 10.3390/diagnostics14040448.

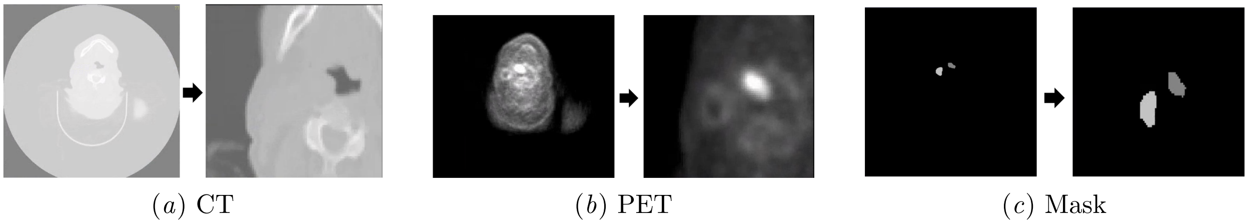
Yichi Zhang, Le Xue, Wenbo Zhang, Lanlan Li, Yuchen Liu, Chen Jiang, Yuan Cheng, and Yuan Qi. Seganypet: Universal promptable segmentation from positron emission tomography images, 2025. URL https://arxiv.org/abs/2502.14351.

Appendix A. Dataset and Processing Details

This appendix will indicates more details of the dataset and how we do the data processing.

Clinical variables and labels. HECKTOR’22 provides nine covariates used for conditioning: age, gender, weight, tobacco, alcohol, performance status, HPV status, chemotherapy and surgery. We standardize continuous variables within each training fold (z-score) and one-hot encode categorical ones; missing entries default to fold medians. Survival labels are recurrence-free time (days) with censoring flags, following the challenge protocol (C-index evaluation).

Details beyond the main pipeline. When determining the global minimal crop **G**, we enforce even side lengths per axis to keep the crop center voxel-aligned across all cases. The tumor center is computed from the mask centroid on the native grid with a fallback to the bounding-box center; after initial resampling to 1×1×3 mm, we refine the center on the resampled mask before extracting the **G**-sized, center-aligned crop. For FM inputs, we apply modality-specific intensity normalization that is not used elsewhere in the paper: CT volumes are


(*a*) CT

(*b*) PET

(*c*) Mask

Figure 4: One representative case. Each panel shows *before* (left) and *after* (right). Standardizing to 1×1×3 mm and a global tumor-centered crop preserves full tumor coverage, focuses the field of view on lesion context, and reduces input size for faster training/inference.

HU-clipped to $[-1000, 2000]$ then min–max scaled to $[0, 1]$; PET volumes use per-volume percentile clipping (0.5–99.5) followed by min–max scaling. A GPU path mirrors SimpleITK resampling via `grid_sample` (aligning output origins to the physical tumor center), which preserves geometry while reducing wall-clock time. This design preserves tumor coverage, focuses the field-of-view on disease, and reduces memory/compute.

Figure 3 shows the comparison a slice of CT, PET, and mask before and after processing for one of the patients.

Appendix B. Training Details

Below is the detailed settings and hyperparameters of the training.

Setting	**Value**
Cross-validation	5-fold stratified
Input volume size	80×80×48
Loss	$L_{\text{Dice}} + 2\,L_{\text{Cox}}$
Optimizer	Adam
Learning rate	$1{\times}10^{-3}$; 5-epoch warm-up
LR schedule	Step decay ×0.3/25 epochs
Epochs	100 epochs
Training Steps	50 steps
Validation Steps	8 steps
Batch sizes	Train 8; Val/Test 16
EMA	Decay 0.999
Weight decay	Classifier 0.1; DenseNet $1{\times}10^{-4}$
Early stopping	Patience 40 on validation C-index
Checkpoint	Best validation C-index
Seed	42