HECKTOR 2025 Challenge: Hierarchical Multi-modal Vision Network for Head-and-Neck Tumor Segmentation and Survival Prediction

Beining Wu[1⋆], Enyu Bao[1⋆], Zhiling Li[1], Yilei Chen[1], Feiwei Qin[1⋆⋆], and Yifei Chen[2⋆⋆]

[1] Hangzhou Dianzi University, 310018, Hangzhou, China
[2] Tsinghua University, 100084, Beijing, China
qinfeiwei@hdu.edu.cn justlfc03@gmail.com

Abstract. Head and neck cancer poses major challenges for precision oncology, where accurate tumor segmentation and reliable survival prediction are essential yet remain difficult due to heterogeneous morphology and multi-center variability in PET/CT. We introduce HM-VNet, a unified multimodal framework designed to address these challenges through end-to-end learning. For segmentation, HM-VNet integrates hierarchical Transformer-based encoding with multimodal fusion to achieve robust delineation of both primary tumors and metastatic lymph nodes. For survival prediction, a deep cross-modal fusion network combines imaging, clinical, and radiomic features, further guided by anatomical priors derived from segmentation results. Comprehensive evaluation on the HECKTOR 2025 Challenge confirms that HM-VNet consistently outperforms state-of-the-art approaches, demonstrating strong effectiveness and promising clinical relevance in advancing automated multimodal intelligence for head and neck cancer management. Our source code is available at https://github.com/Wu-beining/HM-VNet. (Team: HDUMedAI)

Keywords: Head and Neck Cancer · Tumor Segmentation · Survival Prediction · Multimodal Deep Learning · PET/CT · Vision Transformer

1 Introduction

Head and neck cancer (HNC) is one of the most prevalent malignancies globally, encompassing various subtypes, including nasopharyngeal carcinoma, laryngeal carcinoma, and oral squamous cell carcinoma. Each year, hundreds of thousands of new cases are diagnosed. The high incidence and mortality not only pose significant threats to patients' health and quality of life but also place considerable strain on public healthcare systems. In recent years, with the rapid advancement of molecular imaging, positron emission tomography/computed tomography (PET/CT) using fluorodeoxyglucose (FDG) has become a cornerstone in

⋆ Co-first authors.
⋆⋆ Corresponding authors.

clinical practice. This multimodal imaging technique provides high-resolution anatomical data from CT while simultaneously revealing tumor metabolic activity through PET, combining structural and functional information [1].

Despite the rich complementary anatomical and functional information provided by PET/CT, its clinical application still faces two critical challenges. First, accurate target delineation is essential for radiotherapy and personalized treatment, but it remains heavily reliant on manual annotation. The indistinct and heterogeneous boundaries of primary tumors and metastatic lymph nodes make manual contouring both labor-intensive and time-consuming, with significant inter-observer variability [2]. This issue is further exacerbated in multi-center studies, where differences in acquisition protocols and image quality increase inconsistencies . Second, survival prediction plays a crucial role in guiding personalized treatment strategies, but it is inherently complex and multifactorial. Imaging biomarkers, clinical indices, and demographic variables all influence recurrence-free survival (RFS). Extracting meaningful prognostic features from noisy, heterogeneous data and constructing robust risk prediction models remains a significant challenge in both clinical practice and research.

To advance progress in this field, the International Conference on Medical Image Computing and Computer Assisted Intervention organized the HECKTOR challenge [10], using a large-scale, multi-center, and standardized PET/CT dataset to define two core tasks: (1) fully automated segmentation of head and neck tumors and lymph nodes, and (2) RFS prediction based on multimodal data. Building upon this foundation, this study addresses the two core tasks of HECKTOR 2025 by proposing a series of advanced automated analysis methods, which are systematically evaluated on the challenge dataset, thereby validating their effectiveness and translational potential in real-world clinical applications.

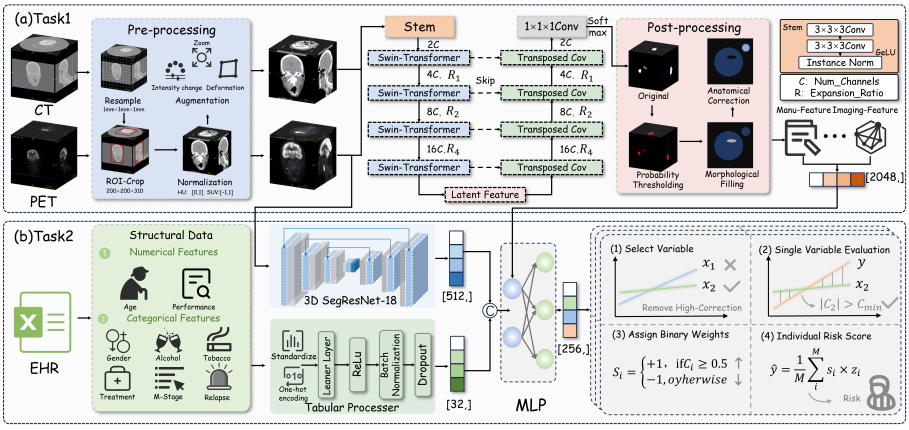

Fig. 1: The proposed end-to-end multimodal analysis framework. (a) Task 1: 3D Swin Transformer-based PET/CT tumor segmentation. (b) Task 2: Survival risk prediction by fusing radiomic and clinical features.

2 Method for Task 1

2.1 Task 1 Overall Architecture

To address the morphological diversity, indistinct boundaries, and inter-center heterogeneity of primary tumors (GTVp) and metastatic lymph nodes (GTVn) in PET/CT, we propose the Hierarchical Multimodal Vision Network (HM-VNet), an end-to-end 3D segmentation framework. As illustrated in Fig. 1(a), HM-VNet is composed of three key components: **(1)Multimodal feature fusion module**, which jointly encodes PET and CT images at the input stage; **(2)Hierarchical Transformer encoder**, which extracts deep semantic representations across multiple resolutions; **(3)Convolutional decoder**, which restores spatial details through skip connections and generates the final segmentation masks.

2.2 Data Preprocessing and Augmentation

To harmonize data from multiple centers and optimize network inputs, we developed a multi-stage preprocessing pipeline. PET, CT, and label images were resampled to an isotropic resolution of 1 mm^3 using SimpleITK, with spline interpolation applied to PET and CT images, and nearest-neighbor interpolation used for label maps. An automated PET-based ROI localization strategy identified the head-and-neck center, from which a fixed volume of 200 $\times$ 200 $\times$ 310 voxels was cropped to minimize computational overhead. The cropped volumes were then reoriented to the standard RAS coordinate system, with CT Hounsfield Units clipped to the range [-250, 250] and normalized to [0, 1], while PET images were standardized using Z-score normalization. To enhance model generalization and robustness against multi-center domain shifts, data augmentation techniques, including random scaling, elastic deformations, and intensity perturbations, were applied during training.

2.3 Multimodal Feature Fusion and Embedding

PET and CT images provide highly complementary information: PET reflects the metabolic activity of tumors, whereas CT offerFs a clear depiction of surrounding anatomical structures. To fully exploit this complementarity, we designed a shallow feature fusion module. Concretely, the input PET and CT volumes are first processed independently through modality-specific convolutional stems. Each stem consists of two $3 \times 3 \times 3$ convolutional layers, followed by instance normalization and GELU activation, enabling preliminary feature extraction and distribution adaptation while capturing low-level edges and texture patterns characteristic of each modality. Subsequently, the modality-specific feature maps are concatenated along the channel dimension, yielding a fused feature tensor that serves as the input to the downstream Transformer encoder. In this way, the module achieves an effective integration of metabolic and anatomical representations, laying the foundation for robust multimodal learning:

$$\mathbf{f}_0^{CT} = \mathrm{Stem}_{CT}(\tilde{\mathbf{X}}^{CT}), \quad \mathbf{f}_0^{PET} = \mathrm{Stem}_{PET}(\tilde{\mathbf{X}}^{PET}), \quad \mathbf{f}_0 = \mathrm{Concat}[\mathbf{f}_0^{CT}, \mathbf{f}_0^{PET}], \tag{1}$$

where $\tilde{\mathbf{X}}^{CT}$ and $\tilde{\mathbf{X}}^{PET}$ denote the normalized CT and PET volumes.

2.4 Hierarchical Vision Transformer Encoder

The core of HM-VNet is a hierarchical encoder that combines the strengths of Swin Transformer and V-Net for efficient and effective 3D spatial modeling. Organized into four stages, the encoder progressively halves spatial resolution while doubling channel dimensionality. This process constructs multi-scale feature representations, capturing both fine-grained local details and global semantic contexts. Within each stage, conventional convolutional operations are replaced by 3D Swin Transformer modules, whose key mechanisms are as follows:

1. **Window-based Multi-head Self-Attention (W-MSA).** Attention computation is restricted to non-overlapping local 3D windows, which substantially reduces computational complexity. This design allows the model to efficiently capture fine structural details within high-resolution 3D volumes, such as tumor interiors and boundary variations. For each 3D window w, the input token matrix $\mathbf{X}_w \in \mathbb{R}^{M^3 \times d}$ undergoes self-attention:

$$\begin{aligned} \mathbf{Q} = \mathbf{X}_w \mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}_w \mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}_w \mathbf{W}_V, \\ \mathrm{Attn}(\mathbf{X}_w) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{B}_{\mathrm{rel}} \right) \mathbf{V}. \end{aligned} \tag{2}$$

2. **Shifted Window-based Multi-head Self-Attention (SW-MSA).** To facilitate information exchange across adjacent windows, the window partitioning is cyclically shifted between consecutive Transformer layers. This mechanism enables the aggregation of broader contextual information, extending the receptive field beyond local windows to effectively capture global dependencies. Such capability is particularly critical in distinguishing metabolically active tumor regions from inflammatory or physiologically active tissues with similar uptake patterns.

Through this hierarchical design, the encoder simultaneously captures global semantics,which include spatial relationships between tumors and adjacent organs in the deeper, low-resolution feature maps, while preserving precise boundary details in the shallower, high-resolution representations. This synergy equips HM-VNet with robust and fine-grained segmentation performance when dealing with complex multimodal head-and-neck imaging data.

2.5 Post-processing Strategy

To refine the raw segmentation outputs and enforce clinical plausibility, we implemented a sequential post-processing pipeline. First, the network's softmax probabilities were converted into a binary mask using a threshold of 0.7. A 3D morphological closing operation was then applied to fill small internal holes and smooth the lesion contours. Subsequently, we performed a connected component

analysis to remove noise. Any component with a volume smaller than 50 mm^3 was discarded as an artifact. Based on clinical priors, we retained only the three largest components, designating the largest as the GTVp. Finally, we applied an anatomical correction rule: any remaining lesion component whose centroid was more than 150 mm away from the GTVp's centroid was removed, thereby eliminating anatomically implausible false positives.

2.6 Segmentation Decoder and Loss Function

The decoder is designed symmetrically to the encoder and employs transposed convolutions to restore the spatial resolution of feature maps progressively. To fully integrate multi-scale information and compensate for detail loss introduced during downsampling, skip connections are incorporated at each stage, concatenating high-frequency encoder features with the corresponding upsampled decoder features. This strategy substantially improves boundary precision in the segmentation outputs. The final prediction layer consists of a $1 \times 1 \times 1$ convolution, which compresses channel dimensions and maps the features into three classes, background, GTVp, and GTVn. A Softmax activation is then applied to produce voxel-wise class probability distributions.

For model optimization, we adopt a weighted combination of Dice loss and cross-entropy loss as the training objective:

$$\begin{aligned} \mathrm{Dice}_c = \frac{2\sum_{i\in\Omega} p_{i,c} y_{i,c} + \varepsilon}{\sum_{i\in\Omega} p_{i,c} + \sum_{i\in\Omega} y_{i,c} + \varepsilon}, \mathcal{L}_{\mathrm{Dice}} = 1 - \frac{1}{3}\sum_{c=0}^{2} \mathrm{Dice}_c, \\ \mathcal{L}_{\mathrm{CE}} = -\frac{1}{|\Omega|}\sum_{i\in\Omega}\sum_{c=0}^{2} y_{i,c}\log p_{i,c}, \mathcal{L}_{\mathrm{total}} = \lambda\mathcal{L}_{\mathrm{Dice}} + (1-\lambda)\mathcal{L}_{\mathrm{CE}}, \quad \lambda = 0.5. \end{aligned} \tag{3}$$

where $\mathcal{L}_{\mathrm{Dice}}$ encourages accurate volumetric overlap, $\mathcal{L}_{\mathrm{CE}}$ penalizes voxel-level misclassification, and λ is a balancing coefficient. This composite loss leverages the complementary strengths of both terms, enabling robust optimization in the presence of class imbalance and heterogeneous tumor morphology.

3 Method for Task 2

3.1 Task 2 Overall Architecture

For Task 2, we propose a deep multimodal fusion network to enable precise prediction of patient survival risk. As illustrated in Fig. 1(b), The framework consists of three parallel branches that separately encode CT/PET imaging data, electronic health records (EHR), and radiomic features derived from Task 1 segmentation results. The outputs of these branches are integrated within a cross-modal fusion module and subsequently passed into the ICARE survival prediction model, thereby forming an end-to-end analysis pipeline.

3.2 Imaging and Clinical Branches

The imaging branch takes registered CT and PET volumes as inputs, $\mathbf{X}_{\mathrm{CT}} \in \mathbb{R}^{H\times W\times D}$ and $\mathbf{X}_{\mathrm{PET}} \in \mathbb{R}^{H\times W\times D}$, which are concatenated along the channel dimension to form dual-channel 3D data $\mathbf{X}_{\mathrm{img}} = \mathrm{Concat}(\mathbf{X}_{\mathrm{CT}}, \mathbf{X}_{\mathrm{PET}}) \in \mathbb{R}^{2\times H\times W\times D}$. A 3D ResNet-18 serves as the feature encoder, learning hierarchical representations ranging from low-level textures to high-level semantics while alleviating gradient vanishing issues through residual connections. To obtain a global imaging representation, the classification head of ResNet-18 is removed, and the output of its final average pooling layer is the feature vector:

$$\boldsymbol{f}_{\mathrm{img}} = \mathbf{F}_{\text{ResNet-18}}(\mathbf{X}_{\mathrm{img}}) \in \mathbb{R}^{512}. \tag{4}$$

The clinical branch encodes structured patient variables. Continuous variables are normalized to yield $\mathbf{X}^{\mathrm{norm}}_{\text{clin-cont}} = \frac{\mathbf{X}_{\text{clin-cont}}-\mu}{\sigma} \in \mathbb{R}^{M}$, while categorical variables are transformed into one-hot encodings to obtain $\mathbf{X}^{\mathrm{onehot}}_{\text{clin-disc}} = \mathrm{OneHot}(\mathbf{X}_{\text{clin-disc}}) \in \mathbb{R}^{P}$, where P is the total dimension after one-hot encoding. These two representations are concatenated into a unified vector $\mathbf{X}_{\mathrm{clin}} = \mathrm{Concat}(\mathbf{X}^{\mathrm{norm}}_{\text{clin-cont}}, \mathbf{X}^{\mathrm{onehot}}_{\text{clin-disc}}) \in \mathbb{R}^{M+P}$, which is then passed through a multi-layer perceptron (MLP) consisting of fully connected layers, ReLU activation, batch normalization, and dropout.This operation can be symbolized as:

$$\boldsymbol{f}_{\mathrm{clin}} = \mathrm{MLP}_{L}(\mathbf{X}_{\mathrm{clin}}) \in \mathbb{R}^{32}, \tag{5}$$

where MLP_L represents a multi-layer perceptron with L-layer hidden layers. The final output is a 32-dimensional embedding vector that unifies heterogeneous clinical factors while capturing their latent associations:

3.3 Multi-mask Generation Module

To better characterize tumor heterogeneity and mitigate segmentation errors, we construct a multi-scale mask ensemble based on Task 1 lesion masks. The ensemble includes original lesions and bounding boxes, SUV-threshold-refined variants, as well as morphologically expanded or shell-derived masks, yielding a total of 11 variants. For each mask, 93 radiomic features $\mathbf{F}_{\mathrm{rad}}(\mathbf{M}_i, \mathbf{X}_{\mathrm{modal}}) \in \mathbb{R}^{93}$ are extracted from both CT and PET using *pyradiomics*, supplemented with handcrafted indicators such as the number of tumors and lymph nodes. This process produces a 2048-dimensional feature vector capturing diverse radiomic and morphological descriptors,which can be expressed as:

$$\boldsymbol{f}_{\mathrm{rad}} = \mathrm{Concat}\left(\bigcup_{i=1}^{11} \mathrm{Concat}\left(\boldsymbol{F}_{\mathrm{rad}}(\boldsymbol{M}_i, \boldsymbol{X}_{\mathrm{CT}}), \boldsymbol{F}_{\mathrm{rad}}(\boldsymbol{M}_i, \boldsymbol{X}_{\mathrm{PET}})\right), \boldsymbol{f}_{\mathrm{manual}}\right) \in \mathbb{R}^{2048}. \tag{6}$$

3.4 Cross-modal Feature Fusion Module

The fusion module concatenates the 512-dimensional imaging representation with the 32-dimensional clinical embedding to form a 544-dimensional vector,

which is then combined with the 2048-dimensional radiomic features. This composite feature vector is passed through a deep MLP that learns nonlinear, high-order cross-modal interactions. The output is a 256-dimensional integrated representation, serving as the unified input for survival prediction. The specific formula can be expressed as follows:

$$\begin{aligned} \boldsymbol{f}_{\text{img-clin}} &= \text{Concat}(\boldsymbol{f}_{\text{img}}, \boldsymbol{f}_{\text{clin}}) \in \mathbb{R}^{512+32} = \mathbb{R}^{544}, \\ \boldsymbol{f}_{\text{combined}} &= \text{Concat}(\boldsymbol{f}_{\text{img-clin}}, \boldsymbol{f}_{\text{rad}}) \in \mathbb{R}^{544+2048} = \mathbb{R}^{2592}, \\ \boldsymbol{g}_t = \text{ReLU}(\boldsymbol{W}_t \boldsymbol{g}_{t-1} + \boldsymbol{b}_t), &\quad \boldsymbol{g}_0 = \boldsymbol{f}_{\text{combined}}, \boldsymbol{f}_{\text{fused}} = \boldsymbol{W}_{\text{fuse-out}} \boldsymbol{g}_T + \boldsymbol{b}_{\text{fuse-out}}, \end{aligned} \tag{7}$$

where $\boldsymbol{W}_t \in \mathbb{R}^{E_t \times E_{t-1}}$, $\boldsymbol{b}_t \in \mathbb{R}^{E_t}$ are the weight and bias of the t-th layer, respectively, and $\boldsymbol{W}_{\text{fuse-out}} \in \mathbb{R}^{256 \times E_T}$, $\boldsymbol{b}_{\text{fuse-out}} \in \mathbb{R}^{256}$ are the weight and bias of the fuse-layer, respectively.

3.5 Prediction Model

For risk prediction, we construct a binary-weighted ICARE model designed to mitigate overfitting under limited sample conditions. Unlike conventional models that assign continuous weights, our approach restricts each feature weight to either $+1$ or -1, indicating whether the feature promotes or suppresses risk. The modeling process consists of four steps:

1. **Feature preprocessing.** Pearson correlation coefficients are computed to detect highly correlated variables. If correlation exceeds a threshold, one redundant variable is randomly discarded to reduce collinearity.
2. **Univariate consistency evaluation.** For each retained feature xi, Harrell's Concordance Index (C-index) is used to assess its concordance with survival outcomes y. Features with predictive power exceeding a minimum threshold $|c_i| > c_{\min}$ are preserved, where $|c_i| = \max\{1 - c_i, c_i\}$.
3. **Binary weighting.** Based on the concordance direction, each feature is assigned a binary weight ($+1$ or -1), where $s_i = +1$ if $c_i \geq 0.5$, and $s_i = -1$ otherwise, ensuring both interpretability and robustness.
4. **Risk scoring.** At prediction time, all input features are standardized via Z-score normalization. The weighted average is then computed to yield the individual risk score for each patient as follows:

$$\hat{y} = \frac{1}{M} \sum_i s_i \times z_i, \tag{8}$$

where z_i denotes the standardized value of the input feature.

To further improve stability and generalization, we employ a bagging strategy in which multiple independent binary-weighted models are trained on bootstrap samples of the original dataset. In inference, their outputs are aggregated by median voting. This ensemble method reduces uncertainty from data sampling and feature selection, substantially enhancing prediction robustness.

3.6 Loss Function

The overall training objective combines two complementary components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DeepHit}} + \alpha \cdot \mathcal{L}_{\text{Contrastive}}, \quad \alpha = 0.1. \tag{9}$$

The DeepHit loss serves as the primary optimization objective, enhancing the model's ability to rank patients by survival risk. It consists of two parts: (1) a likelihood term, inspired by the Cox model, which maximizes the probability that patients with events are assigned higher risks than those still at risk, and (2) a ranking term, which penalizes incorrect patient pair orderings, such as when a patient with shorter survival is predicted to have a lower risk. This loss supervises the outputs of the risk prediction head. In contrast, the Survival Contrastive loss acts as an auxiliary regularizer, refining the feature space by using contrastive learning to bring together patients with similar survival outcomes (e.g., comparable survival times with recurrence) and separate those with markedly different outcomes. This encourages the network to learn more discriminative multimodal representations, improving the stability and robustness of inputs for the downstream ICARE survival prediction model.

4 Experiments

4.1 Dataset

We utilized the HECKTOR 2025 Challenge dataset, which comprises registered PET/CT scans from 700 head and neck cancer patients across 10 centers. Following the official 8:2 training-testing split, our model was trained using the provided expert annotations and evaluated on the hidden test set by the challenge organizers. The dataset provides ground-truth masks for GTVp and GTVn segmentation, alongside comprehensive clinical metadata including TNM staging, demographic information, and RFS outcomes for the prediction task.

4.2 Experimental Setup

All experiments were conducted on a server equipped with an NVIDIA 5090 GPU (32 GB memory). The models were implemented in PyTorch, with data preprocessing, augmentation, and network construction facilitated by MONAI.

For Task 1: HM-VNet was trained end-to-end using the AdamW optimizer for 1000 epochs, with an initial learning rate of 1×10^{-4}. A cosine annealing schedule was employed for dynamic adjustment, and the batch size was set to 2.

For Task 2: AdamW was also used, training the feature extractor in 5 iterations of 10 epochs each, with an initial learning rate of 1×10^{-3} adjusted via ReduceLROnPlateau. The batch size was set to 8. To ensure reproducibility, a global random seed of 42 was fixed, and five-fold stratified cross-validation was performed on the training set.

4.3 Evaluation Metrics

Following the official guidelines, model performance was evaluated for both tasks. For Task1, we used the mean Dice Similarity Coefficient (DSC) for GTVp, alongside the Aggregated Dice (DSC_{agg}) and Aggregated F1-Score for GTVn. For Task2, the Concordance Index (C-index) served as the sole evaluation metric.

4.4 Experimental Results

Segmentation Task. We benchmarked HM-VNet against a wide range of state-of-the-art 3D segmentation models from distinct architectural families. The comparison included CNN-based approaches such as SegResNet [6], Transformer-based architectures like SwinUNETR [5], and a variety of advanced hybrid models, namely Restormer [12], H-Denseformer [11], MMCA-Net [13], MICFormer [4], H2ASeg [8], and AIMERS [7].

Table 1: Performance comparison of different advanced segmentation models on the HECKTOR 2025 validation set. Best: highlighted. Second-best: underlined.

Method	GTVp	GTVn	GTVn
	Mean Dice	Aggregated Dice	Aggregated F1-Score
SegResNet [6]	0.6621	<u>0.7181</u>	**0.5297**
SwinUNETR [5]	<u>0.6673</u>	0.7053	0.1713
Restormer [12]	0.6589	0.7024	0.3439
MICFormer [4]	0.6521	0.6802	0.2157
H-Denseformer [11]	0.6453	0.6826	0.4050
MMCA-Net [13]	0.6437	0.6671	0.2507
H2ASeg [8]	0.5824	0.6855	<u>0.4756</u>
AIMERS [7]	0.5250	0.5110	0.3849
Ours	**0.6833**	**0.7452**	0.4544

1. **Comparison with baseline models:** For GTVp segmentation, HM-VNet achieved a mean Dice score of 0.6833, significantly outperforming both the CNN-based SegResNet (0.6621) and the Transformer-based SwinUNETR (0.6673). This demonstrates the effectiveness of our hierarchical Transformer architecture with multimodal fusion in capturing the complex morphology of primary tumors. In the more challenging task of GTVn segmentation, the advantage of HM-VNet is even more pronounced, reaching an aggregated Dice of 0.7452, which is substantially higher than SegResNet (0.7181) and SwinUNETR (0.7053). These results highlight our model's superior ability to localize and segment small, spatially dispersed metastatic lymph nodes.
2. **Comparison with advanced hybrid architectures:** For GTVp segmentation, HM-VNet again achieved the best performance, with a mean Dice of 0.6833, surpassing Restormer (0.6589) and H-Denseformer (0.6453). For the more challenging GTVn segmentation, HM-VNet attained an aggregated Dice of 0.7452, markedly outperforming Restormer (0.7024) and MICFormer (0.6802). Furthermore, in the lesion-level F1 score, HM-VNet achieved 0.4544,

slightly lower than H2ASeg (0.4756), but significantly higher than MICFormer (0.2157) and MMCA-Net (0.2507), demonstrating superior overall segmentation accuracy and robustness.
3. **Lesion-level detection performance analysis:** At the lesion level, SegResNet (0.5297) achieved the highest aggregated F1 score for GTVn, likely due to its strong capability in local feature extraction. Our HM-VNet (0.4544) and H2ASeg (0.4756) followed closely, both performing competitively and clearly outperforming other Transformer-based methods such as SwinUNETR (0.1713) and MICFormer (0.2157). These results confirm that HM-VNet maintains strong competitiveness in accurately detecting lymph node lesions.

Survival Prediction Task. For Task 2, our proposed method was systematically benchmarked against a spectrum of established survival prediction models, which represent two distinct methodological paradigms. The first paradigm encompasses traditional statistical and machine learning baselines, featuring the seminal Cox Proportional-Hazards (CoxPH [10]) model, which is the field's gold standard and ICARE, [10]a robust model for feature selection in high-dimensional data. The second paradigm consists of advanced deep learning approaches, including DeepHit [10] and MTLR [10], which respectively reformulate survival analysis as discrete-time and multi-task problems, alongside contemporary multimodal frameworks including HMT [3] and Lyn's [9].

Table 2: Performance comparison of different survival prediction models on the HECKTOR 2025 validation set. Best: highlighted. Second-best: underlined.

Method	**C-index**
CoxPH [10]	0.6073 ± 0.0637
DeepHit [10]	0.5457 ± 0.0674
MTLR [10]	0.5877 ± 0.0900
ICARE [10]	0.6705 ± 0.0608
HMT [3]	0.5688 ± 0.0840
Lyn's [9]	0.6032 ± 0.0869
W/O Task1	0.6826 ± 0.0578
Ours	$\mathbf{0.7045 \pm 0.0568}$

1. **Comparison with traditional baselines:** Our model achieved a C-index of 0.6826, significantly outperforming the classical CoxPH model (0.6073) and surpassing the strong modern machine learning method ICARE (0.6705). This remarkable result vividly demonstrates the capability of our end-to-end deep learning framework to capture complex and nonlinear survival-related patterns from heterogeneous data, thereby breaking through the performance bottleneck of traditional approaches.
2. **Comparison with deep learning baselines:** When compared with a range of advanced deep learning methods, our model likewise exhibited a marked advantage. Its C-index of 0.6826 was substantially higher than DeepHit (0.5457), MTLR (0.5877), HMT (0.5688), and Lyn's model (0.6032).

These findings strongly validate the superiority of our unique model architecture in multimodal data fusion, representation learning, and downstream risk prediction, relative to other state-of-the-art deep learning approaches.

3. **Incorporating segmentation-derived features:** To further enhance survival prediction performance, we integrated segmentation results obtained from the Task 1 model into the Task 2 framework. Experimental results revealed that with this additional auxiliary information, the C-index improved further to 0.7045 ± 0.0568. This result clearly demonstrates that spatial localization and anatomical priors derived from segmentation can effectively regularize feature learning, thereby enhancing both the reliability of survival prediction and its overall clinical relevance.

5 Conclusion

We proposed HM-VNet, a hierarchical multimodal framework that jointly tackles tumor segmentation and survival prediction in head and neck cancer. By coupling hierarchical multimodal encoding with cross-modal fusion, our approach achieves robust performance across heterogeneous clinical data. Results on the HECKTOR benchmark confirm its clear advantage over existing methods, underscoring strong potential for real-world clinical translation. This work moves a step closer to reliable multimodal intelligence for precision oncology.

Acknowledgements This work was supported by the Fundamental Research Funds for the Provincial Universities of Zhejiang (No. GK259909299001-006), and Anhui Provincial Joint Construction Key Laboratory of Intelligent Education Equipment and Technology (No. IEET202401).

Disclosure of Interests The authors declare no competing interests.

References

1. Chen, Y., Zhang, C., Ke, Y., Huang, Y., Dai, X., Qin, F., Zhang, Y., Zhang, X., Wang, C.: Semi-supervised medical image segmentation method based on cross-pseudo labeling leveraging strong and weak data augmentation strategies. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2024). https://doi.org/10.1109/ISBI56570.2024.10635443
2. Chen, Y., Zou, B., Guo, Z., Huang, Y., Huang, Y., Qin, F., Li, Q., Wang, C.: Scunet++: Swin-unet and cnn bottleneck hybrid architecture with multi-fusion dense skip connection for pulmonary embolism ct image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 7759–7767 (2024)
3. Cui, J., Xu, Y., Zheng, H., Wu, X., Zhou, J., Liu, Y., Wang, Y.: Hmt: A hybrid multimodal transformer with multitask learning for survival prediction in head and neck cancer. IEEE Transactions on Radiation and Plasma Medical Sciences **9**(7), 879–889 (2025). https://doi.org/10.1109/TRPMS.2025.3539739

4. Fan, X., Liu, L., Zhang, H.: Multimodal information interaction for medical image segmentation. arXiv preprint arXiv:2404.16371 (2024)
5. He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., Xu, D.: Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 416–426. Springer (2023)
6. Hsu, C., Chang, C., Chen, T.W., Tsai, H., Ma, S., Wang, W.: Brain tumor segmentation (brats) challenge short paper: Improving three-dimensional brain tumor segmentation using segresnet and hybrid boundary-dice loss. In: International MICCAI Brainlesion Workshop. pp. 334–344. Springer (2021)
7. Jain, A., Huang, J., Ravipati, Y., Cain, G., Boyd, A., Ye, Z., Kann, B.H.: Head and neck primary tumor and lymph node auto-segmentation for pet/ct scans. In: Andrearczyk, V., Oreiller, V., Hatt, M., Depeursinge, A. (eds.) Head and Neck Tumor Segmentation and Outcome Prediction. pp. 61–69. Springer Nature Switzerland, Cham (2023)
8. Lu, J., Chen, J., Cai, L., Jiang, S., Zhang, Y.: H2aseg: Hierarchical adaptive interaction and weighting network for tumor segmentation in pet/ct images. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 316–327. Springer Nature Switzerland, Cham (2024)
9. Lyu, Q.: Combining nnunet and automl for automatic head and neck tumor segmentation and recurrence-free survival prediction in pet/ct images. In: Andrearczyk, V., Oreiller, V., Hatt, M., Depeursinge, A. (eds.) Head and Neck Tumor Segmentation and Outcome Prediction. pp. 192–201. Springer Nature Switzerland, Cham (2023)
10. 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 (2025), `https://arxiv.org/abs/2509.00367`
11. Shi, J., Kan, H., Ruan, S., Zhu, Z., Zhao, M., Qiao, L., Wang, Z., An, H., Xue, X.: H-denseformer: An efficient hybrid densely connected transformer for multimodal tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 692–702. Springer (2023)
12. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5728–5739 (2022)
13. Zhao, W., Huang, Z., Tang, S., Li, W., Gao, Y., Hu, Y., Fan, W., Cheng, C., Yang, Y., Zheng, H., et al.: Mmca-net: a multimodal cross attention transformer network for nasopharyngeal carcinoma tumor segmentation based on a total-body pet/ct system. IEEE Journal of Biomedical and Health Informatics **28**(9), 5447–5458 (2024)