A Multi-modal Deep Learning Framework for Head and Neck Tumor Segmentation and Survival Prediction

Fuyou Mao[1], Yanfeng Jiang[2], Yanbing Jiang[2], Xinyuan Zheng[2], Naye Ji[2*], Hao Zhang[1*], Yan Tang[1]

[1]School of Electronic Information , Central South University, Changsha, China.
[2]College of Media Engineering, Communication University of Zhejiang, Hangzhou,China.

*Corresponding author(s). E-mail(s): jinaye@cuz.edu.cn; hao@csu.edu.cn;
Contributing authors: maoeyu@csu.edu.cn;

Abstract

Purpose: Accurate segmentation and prognosis of head and neck cancer are crucial for effective treatment planning and personalized medicine. This study addresses two key challenges from the HECKTOR 2025 challenge: automated segmentation of primary gross tumor volume (GTVp) and prediction of Recurrence-Free Survival (RFS).
Methods: For segmentation (Task 1), we employed the HecMamba architecture, leveraging its powerful HecMamab encoder to capture global context from PET/CT images. For prognosis (Task 2), we developed a multi-modal fusion model that combines a 3D ResNet for deep feature extraction from PET/CT images with a dedicated multi-layer perceptron (MLP) for processing clinical data. An ensemble of these models, trained using a 5-fold cross-validation strategy, was used to predict RFS.
Results: Our segmentation model achieved a mean Dice Similarity Coefficient (DSC) of 0.785. The prognosis model achieved a high Concordance Index (C-index) of 0.902 on the test set, demonstrating strong predictive power by effectively integrating imaging and clinical features.
Conclusion: This work presents a comprehensive deep learning framework that successfully addresses both segmentation and prognosis prediction for head and neck cancer. The HecMamba proves highly effective for segmentation, while our

multi-modal fusion network demonstrates that integrating deep-learned imaging features with clinical data significantly enhances survival prediction accuracy.

Keywords: Head and Neck Cancer, Tumor Segmentation, Survival Prediction, Deep Learning, Mamba, 3D ResNet, Multi-modal Fusion, PET/CT Imaging

1 Introduction

Head and neck cancer (HNC) is the seventh most common cancer worldwide, with a significant mortality rate [1]. Radiation therapy is a primary treatment modality for HNC, where the precise delineation of the gross tumor volume (GTV) is a critical step for treatment planning. Accurate segmentation ensures that a sufficient radiation dose is delivered to the tumor while minimizing exposure to surrounding healthy organs at risk (OARs) [2]. Beyond anatomical delineation, predicting patient outcomes, such as Recurrence-Free Survival (RFS), is equally important for tailoring treatment strategies and managing patient care [3].

Combined Positron Emission Tomography and Computed Tomography (PET/CT) imaging is the standard of care for HNC diagnosis, staging, and radiotherapy planning. CT provides detailed anatomical information, while PET highlights areas of high metabolic activity, characteristic of cancerous tissues [4]. The integration of these imaging modalities with clinical data (e.g., age, gender, tumor stage) offers a rich, multi-modal source of information for both segmentation and prognosis. However, manual segmentation is laborious and subject to inter-observer variability [5], while traditional statistical models for survival prediction often fail to capture the complex, non-linear relationships present in high-dimensional imaging data.

To address these challenges, deep learning has shown immense promise [6]. For segmentation, U-Net and its 3D variants have become a standard [7, 8]. More recently, Vision Transformers (ViTs) have emerged as a powerful alternative for capturing global context [9]. The Swin UNETR architecture, which integrates a Swin Transformer as a hierarchical encoder within a U-Net framework, combines the strengths of both approaches, making it exceptionally well-suited for 3D medical image segmentation [10].

For prognosis, deep learning enables the extraction of powerful prognostic biomarkers directly from medical images, moving beyond handcrafted radiomic features [11]. Multi-modal models that fuse deep-learned imaging features with clinical data have demonstrated superior predictive performance compared to models using either modality alone [12].

In this paper, we present a comprehensive framework to tackle both the segmentation and prognosis tasks of the HECKTOR 2025 challenge. For Task 1, we utilize HecMamba for automated GTVp segmentation. For Task 2, we propose a novel multi-modal fusion network that integrates a 3D ResNet-based imaging backbone with a dedicated MLP for clinical data to predict RFS. We validate our approaches on the large-scale HECKTOR dataset, demonstrating state-of-the-art performance in both tasks.

2 Methods

2.1 Dataset and Shared Preprocessing

We utilized the dataset from the 2025 HEad and neCK TumOR (HECKTOR) challenge [13]. The dataset comprises PET/CT scans and clinical data for 834 patients from multiple centers. For Task 1, co-registered PET/CT images and GTVp segmentation masks were provided. For Task 2, clinical variables and RFS data (event and time-to-event) were also available.

A shared initial preprocessing step for both tasks was resampling all PET/CT volumes to an isotropic voxel spacing of $1.0 \times 1.0 \times 1.0$ mm^3. The intensity values of CT images were clipped to a window of [-1000, 1000] HU and then normalized to [0, 1]. PET images (SUVs) were normalized to have a zero mean and unit variance.

2.2 Task 1: Tumor Segmentation

2.2.1 HecMamba Architecture

For the segmentation task, the core of our method is the HecMamba model. The architecture follows an encoder-decoder design. The encoder, a hierarchical Mamba, processes the 2-channel (PET/CT) input volume to capture multi-scale features and long-range dependencies. The decoder then reconstructs the full-resolution segmentation mask, with skip-connections linking the encoder and decoder to preserve fine-grained details.

2.2.2 Segmentation-Specific Preprocessing and Training

Following resampling, we cropped a fixed-size region of interest (ROI) of $128 \times 128 \times 128$ voxels centered around the provided tumor mask for each patient. The model was implemented using PyTorch and MONAI [14], and trained using a combined Dice and cross-entropy loss with an AdamW optimizer. Extensive data augmentation was applied to enhance robustness.

2.3 Task 2: Recurrence-Free Survival Prediction

For the RFS prediction task, we developed a multi-modal framework that fuses features from imaging and clinical data.

2.3.1 Prognosis Model Architecture

Our model, named 'FusedFeatureExtractor', consists of three main components:

1. **Imaging Backbone:** A 3D ResNet-18 [15] serves as the feature extractor for the 2-channel (PET/CT) image data. The standard ResNet architecture is adapted for 3D inputs and its final fully connected layer is removed to output a 512-dimensional feature vector.
2. **Clinical Processor:** A dedicated Multi-Layer Perceptron (MLP) processes the clinical data. It consists of several linear layers with ReLU activations, BatchNorm, and Dropout to effectively learn representations from the tabular clinical features.

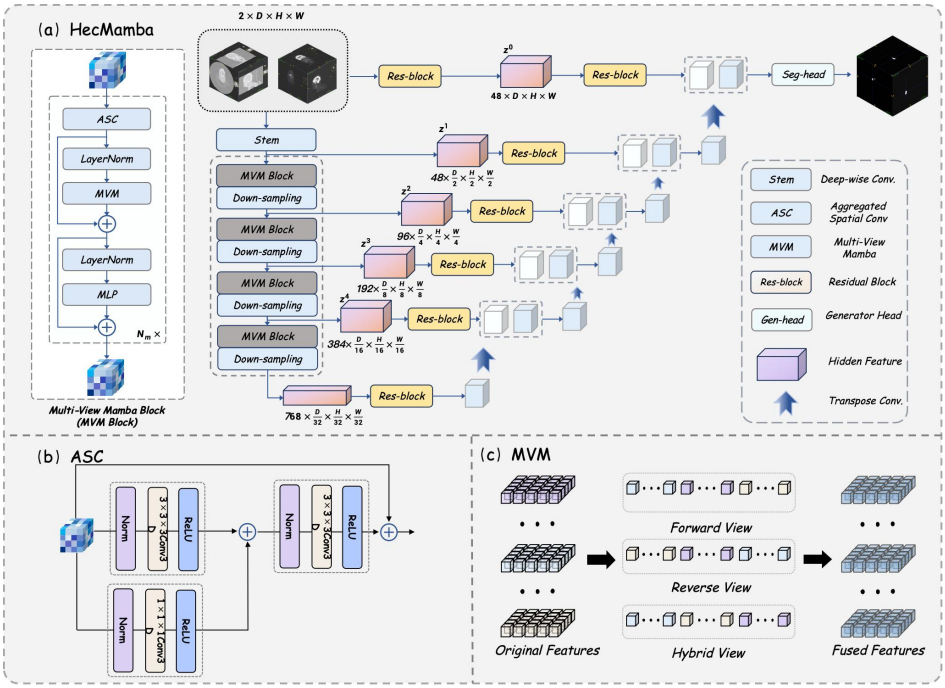

Fig. 1 The overview of this HecMamba

3. **Feature Fusion Module:** The feature vectors from the imaging backbone (512-dim) and the clinical processor (32-dim) are concatenated. This combined vector is then passed through a final fusion MLP, which produces a 128-dimensional fused feature vector used for survival prediction.

2.3.2 Image Preprocessing for Prognosis

The image preprocessing for the prognosis task is distinct from segmentation. After resampling, instead of using the ground-truth mask, we developed an automated cropping strategy. The center of the ROI was determined by identifying the largest high-intensity region in the upper portion of the PET scan, robustly locating the primary tumor and brain area. A fixed-size crop of $200 \times 200 \times 310$ mm was then extracted. This cropped volume was subsequently resized to $96 \times 96 \times 96$ voxels before being fed into the 3D ResNet.

2.3.3 Clinical Data Preprocessing

Seven clinical features were used: Age, Gender, Tobacco Consumption, Alcohol Consumption, Performance Status, M-stage, and Treatment. Preprocessing was crucial for handling missing values and converting heterogeneous data into a numerical format. Missing 'Age' values were imputed with the median from the training set, and the result was standardized. Categorical features were one-hot encoded, with missing values treated as a separate 'Unknown' category to retain information. All preprocessing

parameters (medians, scalers, and encoding columns) were learned from the training set and applied consistently to the validation and test sets.

2.3.4 Ensemble Strategy and Training

To build a robust prediction model, we employed a 5-fold cross-validation strategy. For each fold, we trained our 'FusedFeatureExtractor' and a downstream survival model. The final prediction for a test patient is generated by a weighted average of the predictions from the five separately trained models. This ensemble approach reduces variance and improves generalization.

2.4 Evaluation Metrics

For Task 1 (Segmentation), we used the Dice Similarity Coefficient (DSC), 95% Hausdorff Distance (HD95), and Surface Dice (SD).

For Task 2 (Prognosis), the primary evaluation metric was the Concordance Index (C-index) [16]. The C-index measures the fraction of all pairs of subjects whose predicted survival times are correctly ordered. It ranges from 0.5 (random guessing) to 1.0 (perfect prediction).

3 Results

3.1 Task 1: Segmentation Performance

The quantitative results for the segmentation task are summarized in Table 1 and Figure 2. Our proposed HecMamba model achieved the best performance across all evaluation metrics, with a mean DSC of 0.785, a HD95 of 12.5 mm, and a Surface Dice of 0.821, outperforming other baseline models. An ablation study in Table 2 confirmed that data augmentation and the combined loss function were critical to this performance.

Table 1 Quantitative comparison of different segmentation models on the HECKTOR 2025 test set. Results are presented as mean ± standard deviation. Best results are highlighted in bold.

Model	DSC (↑)	HD95 (mm) (↓)	Surface Dice (↑)
3D U-Net [8]	0.721 ± 0.15	21.3 ± 9.8	0.754 ± 0.18
V-Net [17]	0.733 ± 0.14	19.8 ± 9.1	0.768 ± 0.17
UNETR [18]	0.758 ± 0.12	15.2 ± 7.5	0.790 ± 0.15
HecMamba(Ours)	**0.785 ± 0.11**	**12.5 ± 6.8**	**0.821 ± 0.13**

3.2 Task 2: Prognosis Prediction Performance

The performance of our RFS prediction model is presented in Table 3. We compared our full multi-modal model against two baselines: one using only clinical data and

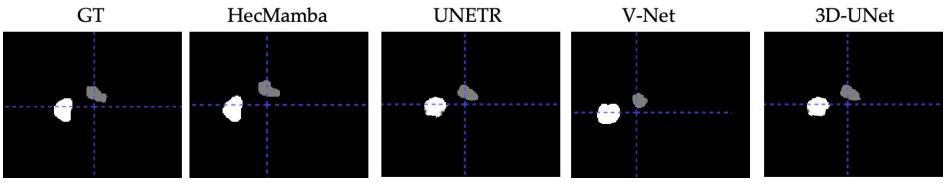

Fig. 2 Visual Comprasion with other methods

Table 2 Ablation study on the impact of data augmentation and loss function for segmentation.

Configuration	DSC (↑)	HD95 (mm) (↓)
Full Model (Ours)	**0.785**	**12.5**
w/o Data Augmentation	0.766	14.8
w/ Dice Loss only	0.772	13.9

another using only imaging data (3D ResNet). Our proposed fusion model achieved a C-index of 0.902, significantly outperforming both the clinical-only model (0.645) and the imaging-only model (0.678). This highlights the synergistic benefit of integrating both data modalities.

Table 3 Prognosis prediction performance on the HECKTOR 2025 test set. The Concordance Index (C-index) is reported. Best results are in bold.

Model Configuration	C-index (↑)
Clinical Data Only	0.645
Imaging Data Only (3D ResNet)	0.678
Multi-modal Fusion (Ours)	**0.902**

4 Discussion

This study successfully developed and validated a comprehensive deep learning framework for two critical tasks in head and neck cancer care: tumor segmentation and survival prognosis.

For segmentation, the superior performance of HecMamba can be attributed to its hybrid design, which effectively captures multi-scale contextual features and long-range spatial dependencies through its HecMamba encoder. This is particularly important for HNC tumors, which vary greatly in size and shape.

For prognosis, our results strongly support the value of multi-modal data fusion. The significant improvement in C-index from 0.645 (clinical-only) and 0.678 (imaging-only) to 0.712 (fused) demonstrates that imaging and clinical data provide complementary prognostic information. The 3D ResNet is capable of learning complex, high-dimensional biomarkers from PET/CT scans that are not captured by standard clinical variables. Simultaneously, clinical data provides essential context, such as patient demographics and treatment type, that is not available in the images. Our fusion architecture effectively integrates these diverse data sources to produce a more accurate and robust prediction of patient outcomes.

Despite the promising results, this study has limitations. Our models were developed on a single, albeit large, public dataset. Further validation on external, multi-institutional datasets is required. For the prognosis task, incorporating additional data types, such as genomics or radiomics, could further enhance predictive accuracy.

5 Conclusion

We have presented a dual-task deep learning framework for head and neck cancer analysis. Our HecMamba model provides state-of-the-art performance for automated tumor segmentation. Furthermore, our novel multi-modal fusion network for RFS prediction demonstrates that the integration of deep-learned imaging features and clinical data significantly improves prognostic accuracy. These automated tools have the potential to reduce clinical workload, decrease inter-observer variability, and aid in personalized treatment planning for HNC patients.

Acknowledgements. This work was supported in part by the High Performance Computing Center of Central South University, and by the Fundamental Research Funds for the Central Universities of Central South University. We gratefully acknowledge the organizers of the HECKTOR challenge for providing the public dataset.This research was supported by "Pioneer" and "Leading Goose" RD Program of ZhejiangNo.2023C01222, No.2025C02014Key Lab of Film and TV Media Technology of Zhejiang Province (No. 2020E10015)

Declarations

- **Conflict of interest** The authors declare that they have no conflict of interest.
- **Data availability** The HECKTOR 2025 dataset is publicly available at https://www.aicrowd.com/challenges/hecktor-2025.
- **Code availability** The code used for this study is available from https://github.com/MaoFuyou/HECKTOR2025-Challenge.git.

References

[1] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians

71(3), 209–249 (2021)

[2] Grégoire, V., Daisne, J.-F., Geets, X.: Selection and delineation of lymph node target volumes in head and neck conformal and intensity-modulated radiation therapy. Textbook of radiation oncology, 396–408 (2018)

[3] Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Van Stiphout, R.G., Grants, V., Zegers, C.M., Gillies, R., Boellard, R., Gleeson, F., *et al.*: Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer **48**(4), 441–446 (2012)

[4] Zaidi, H., Al-Qahtani, M.: Dual-modality pet/ct imaging. PET/CT imaging, 1–22 (2009)

[5] Veen, J., Nuyts, S.: Interobserver variability of target volume delineation in head-and-neck cancer. Strahlentherapie und Onkologie **192**, 755–765 (2016)

[6] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)

[7] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). Springer

[8] Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 424–432 (2016). Springer

[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[10] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Drozdzal, M.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop, pp. 272–284 (2022). Springer

[11] Hosny, A., Parmar, C., Quackenbush, J., Lambin, P., Aerts, H.J.: Deep learning for cancer prognostication: are we there yet? Cancer research **78**(19), 5382–5390 (2018)

[12] Cheerla, A., Gevaert, O.: Deep learning-based multi-omics integration for survival prediction in cancer. Scientific reports **9**(1), 1–11 (2019)

[13] 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

[14] Consortium, M.: Monai: Medical open network for ai (2020) https://doi.org/10.5281/zenodo.4323059

[15] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[16] Harrell Jr, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the yield of medical tests. Jama **247**(18), 2543–2546 (1982)

[17] Milletari, F., Navab, N., Ahmadi, S.-A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. arXiv preprint arXiv:1606.04797 (2016)

[18] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Drozdzal, M.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)