Enhancing Survival Outcomes in Head and Neck Cancer through Joint HPV Classification and Tumor Segmentation

Dalal Chamseddine[1,2*], Shamimeh Ahrari[1*], Mira Rizkallah[2], Thomas Carlier[1,3], and Diana Mateus[2]

[1] Nuclear Medicine Department, University Hospital, Nantes, France
[2] Nantes Université, Centrale Nantes, CNRS, LS2N, UMR 6004, France
[3] Nantes Université, Inserm, CNRS, Université d'Angers, CRCI2NA, Nantes, France

Abstract. Head and Neck Cancer (HNC) is a broad term for cancers that develop in the head and neck region. Accurate survival prediction is critical for guiding patient management and treatment planning. Traditional survival models, such as Kaplan–Meier curves and Cox Proportional Hazards models, are limited by their dependence on linearity and proportional hazards assumptions. Recently, deep learning–based survival models have demonstrated promising results for risk prediction. However, current approaches still struggle to fully integrate multimodal data and to capture region-specific features effectively. In this work, we employed a Multitask Learning framework that simultaneously performs Human Papillomavirus (HPV) classification, tumor segmentation, and survival prediction based on Positron Emission Tomography, Computed Tomography imaging, and clinical features. Integrating these three tasks into a unified model enables the use of shared feature representations. The segmentation module facilitates the extraction of tumor-specific features, while the classification branch incorporates HPV status prediction as a critical prognostic factor in HNC.
This approach was developed and evaluated as part of our participation in the MICCAI 2025 HEad and neCK TumOR segmentation and outcome prediction challenge as the SIMS-LIFE team.

Keywords: Multitask learning · Survival prediction · PET/CT · Head and neck cancer.

1 Introduction

Head and Neck Cancer (HNC) is the seventh most common cancer worldwide. It typically arises from squamous cells covering the mucosal surfaces of the lip, oral cavity, pharynx, larynx, and paranasal sinuses. The prognosis for HNC varies greatly, with an overall 5-year survival rate of 50–60% [11]. The global incidence of HNC continues to rise, with 946,456 new cases and 482,001 deaths reported

* These authors contributed equally to this work.

worldwide in 2025 [2], emphasizing the importance of accurately and promptly diagnosing high-risk patients. In this context, survival analysis is a powerful tool for estimating risk, guiding early prognosis, and personalizing treatment plans.

^{18}F-FluoroDeoxyGlucose Positron Emission Tomography (PET) and Computed Tomography (CT) imaging play a crucial role in tumor characterization at both initial staging and follow-up of HNC, thanks to the complementary nature of both anatomical and functional information these medical images provide [10]. Traditional survival prediction methods are usually based on clinical and tabular information collected by clinicians or handcrafted radiomics features [7] extracted from PET/CT images, which are then modeled using statistical survival models such as the Cox Proportional Hazards (CoxPH) [5]. In this regard, several studies investigating the impact of radiomics features in patients with HNC have shown promising results in improving survival predictions. For example, Haidar et al. [8] demonstrated that such features can provide complementary prognostic information, enabling more accurate outcome estimations. Other studies have further shown that choices within the radiomics pipeline, such as harmonization strategies, sub-volume feature extraction, and segmentation methods, can significantly influence survival predictions for HNC, thereby improving the performance of the CoxPH models in risk stratification [23].

More recently, deep survival models [13, 15, 6] based on deep learning have been introduced to overcome the limitations of traditional approaches, which often assume linear effects and proportional risks. Deep survival models are designed to capture non-linear feature interactions and integrate heterogeneous data sources. In HNC, such approaches have shown promise in more accurately predicting patient outcomes [19] and have the potential to outperform traditional survival models [1]. Moreover, other studies have leveraged graph-convolutional networks to build survival prediction models [18].

Building on deep models, Multitask Learning (MTL) has been explored to further improve survival prediction. MTL lies in learning shared representations across related tasks, which reduces the prediction error and improves event occurrence estimation within specific time intervals [16]. Leveraging the multiple ground-truth tasks provided in the HEad and neCK TumOR (HECKTOR) Grand Challenge 2025, we incorporate two auxiliary tasks alongside our main task of survival prediction within a unified network to enhance risk estimation.

To this end, in this work, we employed a modified version of DeepMTS [20]. While the original framework focused on segmentation and survival analysis, our adaptation integrates both classification and segmentation with survival analysis to maximize the extraction of relevant features while preserving prognostic information. In particular, the classification branch predicts Human Papillomavirus (HPV) status, a key prognostic factor strongly associated with patient outcomes. Multimodal PET/CT data served as input, and deep features derived from both the classification and segmentation backbones were combined with clinical features to generate the final Recurrence-Free Survival (RFS) prediction. In addition, a strategy was adopted to address incomplete data, handling cases with missing labels and/or missing categorical clinical features. Experiments on the

challenge dataset suggest that incorporating HPV classification can contribute to improved risk prediction performance.

2 Multitask Framework

This study focused on RFS prediction, defined as the time interval between the date of diagnosis and the date of disease recurrence. Using PET, CT, and clinical features as input, the model performed survival prediction as the primary task, with classification and segmentation serving as auxiliary tasks. Accordingly, the survival label, HPV status, and segmentation mask were used as ground-truth labels.

Fig. 1 illustrates the overall model architecture, which is a modified version of DeepMTS [20]. Within this multitask framework, training was performed using

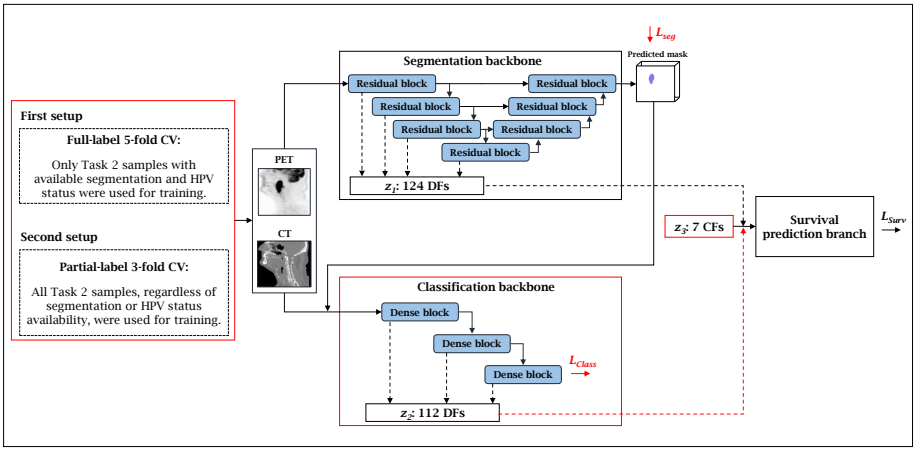


Fig. 1: Flowchart of the overall pipeline, with key modifications relative to the original DeepMTS highlighted in red. CV: Cross-Validation; PET: Positron Emission Tomography; CT: Computed Tomography; DF: Deep Feature; CF: Clinical Feature.

a combined multitask loss (L_{Total}) defined as follows:

$$L_{Total} = L_{Seg} + L_{Class} + L_{Surv} + 0.1L2_{Reg}, \tag{1}$$

Based on the original DeepMTS, the preprocessed PET and CT images were first concatenated on the channel axis and fed into the segmentation backbone, a modified version of 3D U-Net [4]. The predicted tumor probability map was

compared with the ground-truth segmentation label using the sum of Dice [21] and Focal [17] losses (L_{Seg}), instead of only the Dice used in DeepMTS [20], to emphasize hard tumor pixels that would otherwise be overwhelmed by the large background regions. Next we describe the details of the loss functions:

$$L_{Seg} = L_{Dice} + L_{Focal}, \tag{2}$$

$$L_{Dice} = 1 - \frac{2\sum_i^N p_i y_i}{\sum_i^N p_i^2 + \sum_i^N y_i^2}, \tag{3}$$

$$L_{Focal} = \frac{1}{N}\sum_{i=1}^{N} FL(p_i, y_i), \tag{4}$$

where

$$FL(p,y) = \begin{cases} -\alpha(1-p)^{\gamma} log(p) & y = 1 \\ -(1-\alpha)p^{\gamma} log(1-p) & y = 0 \end{cases} \tag{5}$$

Here N is the total number of voxels, $p \in \{0,1\}$ is the predicted tumor probability map, and $y \in \{0,1\}$ is the ground-truth segmentation label. In Eq. 5, α is for the voxel weighting factor and γ is a focusing parameter were set to 0.25 and 2, respectively.

Subsequently, the tumor probability map and the preprocessed PET/CT images were concatenated and input into the HPV classification backbone, which is one of our main modifications. The classification branch was implemented as a cascaded network based on a modified 3D DenseNet [12], introducing it as a new branch added to the original DeepMTS [20]. The network parameters were optimized using a binary focal cross-entropy loss (L_{Class}) [17], which follows the same formulation as in Eq. 5, but this time comparing the predicted HPV status (p_c) of the whole image with the ground-truth classification labels (y_c).

Next, the clinical features (z_3) were concatenated with 124 deep features (z_1) derived from the encoder branch of the segmentation backbone and 112 deep features (z_2) extracted from the classification backbone, and fed into the survival branch, which consisted of three Fully Connected (FC) layers. As in the original DeepMTS, the predicted risk factor was compared with the ground-truth survival label using the Cox negative log partial likelihood loss (L_{Surv}) [13]:

$$L_{Surv} = -\frac{1}{N_{E=1}} \sum_{i:E_i=1} (h_i - \log \sum_{j \in R(T_i)} e^{h_j}) \tag{6}$$

where $N_{E=1}$ is the number of patients with disease recurrence, $E_i \in \{0,1\}$ is the event indicator (1 = event, 0 = censored), h denotes the predicted risk scores, T is the ground-truth observed times, and $R(T_i)$ is the risk set defined as $\{j : T_j \geq T_i\}$, representing the set of patients j who have survived up to time T_i.

Additionally, an L2 regularization ($L2_{Reg}$) term with a coefficient of 0.1 was applied to all FC layers of the survival branch, consistent with the original framework.

Since HPV and/or segmentation labels were not present for all samples, but we aimed to utilize all available data, we adopted a simple strategy to handle missing labels. When a segmentation mask was available, the primary tumor and metastatic lymph node delineations were merged into a single ground-truth segmentation label; otherwise, a zero-valued image was generated. Missing HPV status and categorical clinical features were assigned to a separate category to account for incomplete data. We considered an approach inspired by tree-based algorithms, such as XGBoost with a linear booster, to handle missing values [3]. Accordingly, all missing values were coded as 0, and the remaining categories were shifted forward by one to preserve their distinction.

3 Experimental Settings

3.1 Dataset

In this study, we used a publicly available dataset from the HECKTOR 2025 challenge [22], which was designed for developing and benchmarking automatic methods for HNC segmentation and survival prediction. For each sample, the CT image was paired with the corresponding registered PET image. The challenge organizers precomputed the Standardized Uptake Value (SUV) for the PET images, and all PET/CT files were provided in NIfTI format. The considered clinical features consisted of age, gender, tobacco and alcohol consumption, treatment (radiotherapy only or combined with chemotherapy and/or surgery), performance status, and distant metastasis stage.

For our main task of RFS prediction (Task 2), a training set of 678 samples was collected from seven centers, including CHUM (n=56), CHUP (n=44), CHUS (n=72), HGJ (n=55), HMR (n=18), MDA (n=422), and USZ (n=11). Six samples were excluded from the analysis, two due to negative SUV values and four due to significant extravasation at the injection area, which impairs the SUV computation. The HECKTOR challenge provided a validation set of 50 samples (HECKTOR validation set), while the test set contains 400 samples with withheld ground-truth labels. Further details of the dataset are available in [22].

3.2 Image Preprocessing

For early PET/CT fusion, we initially performed resampling onto a common voxel grid to address differences in voxel size between the two modalities. The CT images exhibited higher spatial resolution than the corresponding PET images, with CT voxel sizes ranging from $0.49 \times 0.49 \times 2\ mm^3$ to $2.73 \times 2.73 \times 3\ mm^3$, while PET voxel sizes ranged from $0.98 \times 0.98 \times 2.50\ mm^3$ to $5.47 \times 5.47 \times 3.27\ mm^3$. All images were resampled to isotropic voxels of $2 \times 2 \times 2\ mm^3$, using spline interpolation for the PET/CT images and nearest-neighbor interpolation for the corresponding segmentations. Subsequently, the images were cropped to $128 \times 128 \times 128$ voxels, centered on the tumor using PET intensity-based bounding boxes derived from SUV thresholding.

3.3 Metrics

Considering RFS prediction as the primary objective of this study, the following metrics were used to evaluate model performance for each task.

Concordance Index (C-index) The C-index is an extension of the area under the curve that considers censored data. It refers to the model's ability to accurately produce a trustworthy ranking of survival durations according to the personal scores [9]. This metric is measured between 0 and 1, where any value less than 0.5 indicates random predictions. The equation for this metric is as follows:

$$C\text{-index} = \frac{\sum_{i,j} \mathbf{1}_{T_j<T_i} \cdot (\mathbf{1}_{r_j>r_i} + 0.5 \cdot \mathbf{1}_{r_j=r_i})}{\sum_{i,j} \mathbf{1}_{T_i<T_j} \cdot \delta_i} \tag{7}$$

Here T_i and T_j represent the time for a pair of patients, respectively, δ denotes the event indicator, and r represents the risk score.

Dice Coefficient It is a similarity metric calculated between the ground-truth and the predicted segmentations. Its value is between 0 (no overlap) and 1 (perfect prediction). This metric is defined as:

$$\text{Dice}(y_{\text{true}}, y_{\text{pred}}) = \frac{2\sum(y_{\text{true}} \cdot y_{\text{pred}}) + \epsilon}{\sum y_{\text{true}} + \sum y_{\text{pred}} + \epsilon} \tag{8}$$

where y_{true} and y_{pred} indicate the ground-truth and the predicted segmentations, respectively, and ϵ represents a constant of $1e-8$.

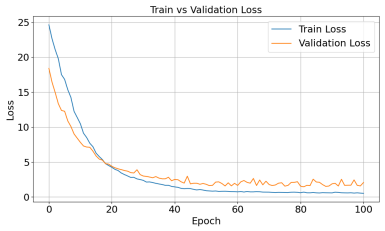
Balanced Accuracy It accounts for class imbalance by averaging the recall of each class and can be calculated according to the following equation:

$$Balanced\ \ Accuracy = \frac{1}{2} \cdot \Big(\frac{TP}{P} + \frac{TN}{N}\Big) \tag{9}$$

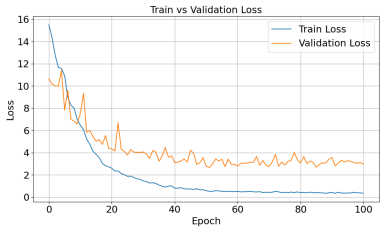
where TP represents true positives, TN indicates true negatives, P represents all positives, and N represents all negatives.

3.4 Implementation Details

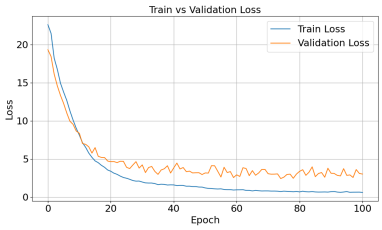
The model implementation was performed using TensorFlow 1.x on a 16 GB NVIDIA Quadro RTX 5000 for the first setup and TensorFlow 2.x on a 20 GB NVIDIA RTX 4000 Ada Generation for the second setup (see the description of setups in Section 4.2). Both models were trained for 500 epochs with a batch size of 4. Early stopping was applied based on the validation set C-index. The Adam optimizer [14] was initialized with a learning rate of 1×10^{-4} and subsequently reduced to 5×10^{-5}, 1×10^{-5}, and 1×10^{-6} at epochs 100, 200, and thereafter. The code is publicly available at `https://github.com/Dalalsh/HNC-Multi-task-Learning`.



(a) Seg+Surv loss



(b) Surv loss



(c) Class+Seg+Surv loss

Fig. 2: Total loss over 100 epochs for the ablation study.

4 Experimental Validation and Results

4.1 Task Ablation

For the ablation study, three model configurations were evaluated prior to selecting the final model for submission. The first model (Seg+Surv) followed the original DeepMTS framework [20], incorporating the segmentation and survival branches. The second model consisted solely of the survival branch (Surv). The last model was the modified version of DeepMTS (Fig. 1) that integrated three tasks: classification, segmentation, and survival (Class+Seg+Surv). In the ablation study, experiments were conducted using a single train/validation/test split across all patients, with 60% of the samples allocated to the training, 30% to the validation, and 10% to the test set. All models were trained for 100 epochs to assess performance with and without the multitask configuration.

As shown in Fig. 2, the validation loss curve for the Surv model exhibited some instability, whereas the curves for Seg+Surv and Class+Seg+Surv demonstrated better convergence. Comparing the validation C-index across the three model configurations, the Class+Seg+Surv configuration tended to provide a better performance. As reported in Table 1, incorporating HPV status classification improved overall network performance, with the multitask Class+Seg+Surv model achieving the higher validation C-index values than both the Surv and Seg+Surv configurations. This indicates that incorporating clinically relevant classification tasks can enhance the discriminative power of the model and contribute positively to survival prediction.

Table 1: Results of the ablation study. Higher C-index values across the validation set for each of the three model configurations at each epoch are highlighted in bold.

Epoch	Training C-index			Validation C-index		
	Seg+Surv	Surv	Class+Surv+Seg	Seg+Surv	Surv	Class+Surv+Seg
0	0.5445	0.4964	0.5265	0.4031	0.4687	**0.4901**
10	0.5018	0.5538	0.5603	0.4724	0.4444	**0.6070**
20	0.7216	0.7255	0.7717	0.5813	0.6604	**0.6939**
30	0.9217	0.8748	0.9322	0.6224	0.5999	**0.6580**
40	0.9323	0.9221	0.9338	0.6417	0.5442	**0.7206**
50	0.9625	0.9558	0.9406	0.5922	0.5791	**0.7986**
60	0.9772	0.9467	0.9634	0.5323	0.6065	**0.7899**
70	0.9769	0.9721	0.9658	0.6073	**0.7055**	0.6694
80	0.9772	0.9672	0.9835	0.6406	0.5520	**0.7198**
90	0.9818	0.9675	0.9679	0.5833	0.5480	**0.6326**
100	0.9823	0.9665	0.9778	0.5495	**0.6123**	0.5639

4.2 Final Submission

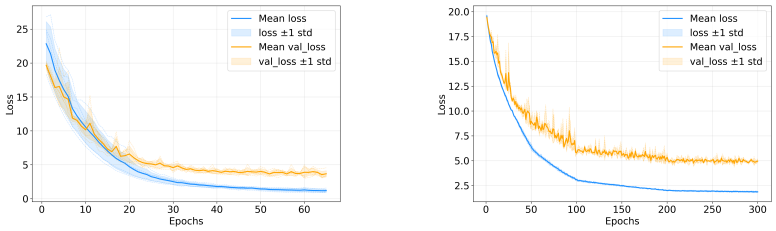
Based on observations from the ablation study (Table 1), the Class+Seg+Surv configuration was selected for our final multitask framework and applied Cross-Validation (CV) techniques to achieve a more generalized predictive model. Accordingly, two training and validation setups were evaluated (Fig. 1) using the Class+Seg+Surv model configuration.

Full-label 5-fold CV In the first setup, a 5-fold CV was performed using only samples from the Hecktor challenge Task 2 that had labels available for both segmentation masks and HPV status (n=518). Samples missing either label (n=154) were systematically added to the validation set of each fold.

Partial-label 3-fold CV The second setup employed a 3-fold CV that included all Task 2 samples in the training set (n=672) regardless of segmentation mask or HPV status availability. In this setup, samples with missing labels contributed to model optimization solely through their image data and were excluded from the corresponding loss calculations (L_{Class} and L_{Seg}), aiming to leverage the full image distribution to enhance feature representation despite partial labels.

For the final submission, we considered the average of the best model from each fold to obtain the final predicted risk score on the shared validation set. This approach was intended to reduce the influence of any outliers in the training data on the final predictions and to make predictions based on an ensemble of models. Fig. 3 shows the mean total loss across CV folds for both setups, with the networks exhibiting stable validation loss across the CV folds.

Table 2 report the C-index values for the considered setups. For the local validation sets, the results are presented as mean C-index values across CV folds,



(a) Full-label 5-fold CV (b) Partial-label 3-fold CV

Fig. 3: Mean total loss across CV folds for the Class+Seg+Surv multitask configuration with early stopping.

followed by the 95% confidence intervals. The C-index values on the HECKTOR validation and test sets are also provided, where our team achieved first place on the HECKTOR test set (0.658), with a clear margin over the second-place score (0.602).

Table 2: C-index results on the local validation set (from our CV), the HECKTOR validation and test sets (provided by the challenge), based on the Class+Seg+Surv multitask configuration.

Setup	**Local validation**	**HECKTOR**	
		Validation set	**Test set**
Full-label 5-fold CV	0.666 [0.632, 0.705]	0.566	0.658
Partial-label 3-fold CV	0.682 [0.671, 0.703]	0.583	0.494

5 Conclusion and Limitation

This study focused on predicting survival outcomes using a multitask framework that simultaneously performs classification, segmentation, and survival prediction. A key observation was the critical role of incorporating HPV status as a clinical integration in survival prediction. However, implementing such a multimodal and multitask approach is challenging, as complete multimodal data are often not available for all samples. To address this, categorical features with missing values were assigned to a separate category. Additionally, to enable the model to learn image feature representations from all available samples, unlabeled data for classification and/or segmentation were included during training but excluded from loss computation. Finally, we considered an ensemble technique that uses a CV scheme to improve the generalizability of the model to

unseen samples. Although the improvements in the C-index were moderate, they were consistent across our local validation set and the HECKTOR validation set. This approach could potentially be extended by employing other losses beyond the survival loss to better handle these unlabeled samples. While the framework achieved promising performance, future work could explore more advanced survival prediction methods, such as DeepHit or attention-based networks, to strengthen the survival branch. In addition, refining training strategies for the classification branch and incorporating attention or uncertainty modeling can enhance interpretability in clinical settings.

Acknowledgments. This work has been partially funded by SIRIC ILIAD (INCa-DGOS-INSERM-ITMO Cancer 18011), the French "Program d'Investissement d'Avenir" (ANR-16-IDEX-0007), the region "Pays de la Loire" through their support to I-Site NExT, as well as by Siemens Healthineers, the industrial partner of the NExT research industrial chair IMRAM.

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

References

[1] Parnian Afshar et al. "From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities". In: *IEEE Signal Processing Magazine* 36.4 (2019), pp. 132–160.

[2] Karam El-Bayoumy et al. "Current challenges and potential opportunities for interception and prevention of head and neck cancer". In: *Carcinogenesis* 46.2 (2025), bgaf025.

[3] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[4] Özgün Çiçek et al. "3D U-Net: learning dense volumetric segmentation from sparse annotation". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2016, pp. 424–432.

[5] David R Cox. "Regression models and life-tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202.

[6] André Diamant et al. "Deep learning in head & neck cancer outcome prediction". In: *Scientific reports* 9.1 (2019), p. 2764.

[7] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. "Radiomics: images are more than pictures, they are data". In: *Radiology* 278.2 (2016), pp. 563–577.

[8] Stefan P Haider et al. "Potential added value of PET/CT radiomics for survival prognostication beyond AJCC 8th edition staging in oropharyngeal squamous cell carcinoma". In: *Cancers* 12.7 (2020), p. 1778.

[9] Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors". In: *Statistics in medicine* 15.4 (1996), pp. 361–387.

[10] Mathieu Hatt et al. "18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi–cancer site patient cohort". In: *Journal of nuclear medicine* 56.1 (2015), pp. 38–44.

[11] Anni Heinolainen et al. "Survival and data-driven phenotypes in head and neck cancer". In: *Scientific Reports* 15.1 (2025), p. 5985.

[12] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[13] Jared L Katzman et al. "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network". In: *BMC medical research methodology* 18.1 (2018), p. 24.

[14] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[15] Changhee Lee et al. "Deephit: A deep learning approach to survival analysis with competing risks". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.

[16] Yan Li et al. "A multi-task learning formulation for survival analysis". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1715–1724.

[17] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[18] Wenbing Lv et al. "Functional-structural sub-region graph convolutional network (FSGCN): application to the prognosis of head and neck cancer with PET/CT imaging". In: *Computer Methods and Programs in Biomedicine* 230 (2023), p. 107341.

[19] Saurav Mandal, Akshansh Gupta, and Waribam Pratibha Chanu. "Survival prediction of head and neck squamous cell carcinoma using machine learning models". In: *arXiv preprint arXiv:2105.07390* (2021).

[20] Mingyuan Meng et al. "DeepMTS: Deep multi-task learning for survival prediction in patients with advanced nasopharyngeal carcinoma using pretreatment PET/CT". In: *IEEE Journal of Biomedical and Health Informatics* 26.9 (2022), pp. 4497–4507.

[21] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation". In: *2016 fourth international conference on 3D vision (3DV)*. Ieee. 2016, pp. 565–571.

[22] Numan Saeed et al. *A Multimodal and Multi-centric Head and Neck Cancer Dataset for Segmentation, Diagnosis, and Outcome Prediction*. 2025. arXiv: 2509.00367 [cs.CV]. URL: https://arxiv.org/abs/2509.00367.

[23] Hui Xu et al. “Radiomics prognostic analysis of PET/CT images in a multicenter head and neck cancer cohort: investigating ComBat strategies, sub-volume characterization, and automatic segmentation”. In: *European Journal of Nuclear Medicine and Molecular Imaging* 50.6 (2023), pp. 1720–1734.