TransRP: Transformer-based PET/CT feature extraction incorporating clinical data for recurrence-free survival prediction in oropharyngeal cancer

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

The growing number of subtypes and treatment options for oropharyngeal squamous cell carcinoma (OPSCC), a common type of head and neck cancer (HNC), highlights the need for personalized therapies. Prognostic outcome prediction models can identify different risk groups for investigation of intensified or de-escalated treatment strategies. Convolution neural networks (CNNs) have been shown to have improved predictive performance compared to traditional clinical and radiomics models by extracting comprehensive and representative features. However, CNNs are limited in their ability to learn global features within an entire volume. In this study, we propose a Transformer-based model for predicting recurrence-free survival (RFS) in OPSCC patients, called TransRP. TransRP consists of a CNN encoder to extract rich PET/CT image features, a Transformer encoder to learn global context features, and a fully connected network to incorporate clinical data for RFS prediction. We investigated three different methods for combining clinical features into TransRP. The experiments were conducted using the public HECKTOR 2022 challenge dataset, which includes pretreatment PET/CT scans, Gross Tumor Volume masks, clinical data, and RFS for OPSCC patients. The dataset was split into a test set (n = 120) and a training set (n = 362) for five-fold cross-validation. The results show that TransRP achieved the highest test concordance index of 0.698 (an improvement > 2%) in RFS prediction compared to several state-of-the-art clinical and CNN-based methods. In addition, we found that incorporating clinical features with image features obtained from the Transformer encoder performed better than using the Transformer encoder to extract features from both clinical and image features. The code for this study is available at (anonymized temporarily for review).

Keywords: Transformer, recurrence-free survival prediction, oropharyngeal cancer.

1. Introduction

The common treatment of head and neck cancer patients is radiotherapy with or without chemotherapy and/or surgery. Positron emission tomography (PET) and computed tomography (CT) scans are used in clinical practice to provide functional and structural visualization of tumors in head and neck regions, enabling oncologists to delineate tumor and organs-at-risk contours for precise radiation dose plans. Oropharyngeal squamous cell carcinoma (OPSCC) is a common type of HNC, and OPSCC patients with human papillomavirus (HPV) have significantly higher five-year overall survival rates (75%-80%) than those with tobacco- or alcohol-related OPSCC (45%-50%) (O'Sullivan et al., 2016). Prognostic outcome prediction models with good performance are important to optimize treatment strategies for individual patients.

© 2023 Anonymous.

The Tumor-Node-Metastasis (TNM) staging system, including the HPV status in the eighth edition of the American Joint Committee on Cancer (AJCC) guidelines (Amin et al., 2017), is widely used for prognostic outcome prediction in patients with OPSCC (Würdemann et al., 2017). Ang et al. (Ang et al., 2010) published a risk classification system for overall survival based on HPV status, smoking pack years, tumor stage and nodal stage. In addition to clinical models, radiomics features were found to have predictive value for outcomes such as overall survival (OS) and local control (LC) in OPSCC (Haider et al., 2020; Bos et al., 2021). While handcrafted radiomics features are more interpretable to oncologists due to their clear definitions, deep learning techniques can extract more comprehensive and representative image features, potentially obtaining better outcome prediction(Afshar et al., 2019).

Convolution neural networks (CNNs) have been successfully applied to various tasks, including image classification using ResNet (He et al., 2016), DenseNet (Huang et al., 2017) and Efficientnet (Tan and Le, 2019)architectures . A multi-center study on HNC demonstrated that a 2D CNN that inputting the central slice of the primary tumor from a CT image achieved higher area under the curve (AUC) values for 2-year distance metastasis (DM), loco-regional failure (LRF), and overall survival (OS) prediction compared to the radiomics approach (Diamant et al., 2019). With the same dataset, Pang et al. proposed a 3D multi-scale ResNet with a combination of three loss functions and achieved state-of-the-art AUCs of 0.91, 0.78, and 0.70 for 2-year DM, LRF, and OS prediction, respectively (Pang et al., 2022). However, these studies did not address time-to-event outcome prediction and only used pretreatment CT scans, without incorporating clinical predictors.

Several studies have demonstrated that deep learning is effective in time-to-event prediction for OPSCC or HNC. For example, Kim et al. (Kim et al., 2019) proposed to use DeepSurv (Katzman et al., 2018) model based on clinical parameters to predict overall survival of OPSCC patients, which outperformed random survival forest (RSF) and Cox proportional hazard (CPH) models. Additionally, other studies utilized imaging data only by CNNs. Cheng et al. (Cheng et al., 2021) proposed a fully automated CNN model for Gross Tumor Volume (GTV) segmentation and OS prediction in OPSCC patients using PET scans. Wang et al. (Wang et al., 2022b) compared the prediction performance of various 3D ResNet-based models trained on different combinations of PET/CT and GTV mask in HNC patients, and found that the PET-only model performed best in terms of DM and OS prediction.

The HECKTOR 2021 challenge (Andrearczyk et al., 2021) provided pretreatment PET/CT scans, clinical data, and GTV of primary tumors (GTVp) for three tasks: (1) GTVp segmentation, (2) progression-free survival (PFS) prediction without GTVp, and (3) PFS prediction using GTVp. The winning team in task 3 used a 3D DenseNet121 (Huang et al., 2017) model to input PET/CT/GTVp and clinical data, and demonstrated that clinical data and GTVp can improve PFS prediction(Wahid et al., 2021). The HECKTOR 2022 challenge (Andrearczyk et al., 2023) included two tasks: (1) GTV of primary tumors and lymph nodes segmentation, and (2) recurrence-free survival (RFS) prediction without GTV. High-ranking teams in task 2 either used clinical and radiomics features based on automatic GTV segmentation by CNNs (Rebaud et al.; Wang et al., 2022a) or a multi-task CNN-based model for GTV segmentation and RFS prediction (Meng et al., 2022). In particular, Meng et al. used residual and dense blocks in their RFS prediction model (Meng et al., 2022). These

studies suggest that CNN-based models with advanced blocks such as ResNet or DenseNet and utilizing clinical and CT/PET/GTV images together can achieve good outcome prediction in OPSCC patients. However, CNNs are limited in modeling explicit long-range relations due to the locality of convolution operations, and thus some researchers have used CNNs to extract rich image features and then employed a Vision Transformer (ViT) (Dosovitskiy et al., 2020) to learn global contexts for the segmentation task (Chen et al., 2021). Additionally, CNNs may be limited in combining multi-modality (clinical and image) information because CNN was designed only for image processing. As a result, the winner of the HECKTOR 2021 task 2 further used a ViT to learn the interaction between clinical data and image patches and obtained state-of-the-art performance in HECKTOR 2021 dataset (Saeed et al., 2022). Another study concatenated clinical data to CNN-extracted CT feature maps and used a ViT to learn global contexts, achieving state-of-the-art survival prediction in nasopharyngeal cancers (Zheng et al., 2022).

Our contributions in this study are summarized as follows: (1) We proposed a novel Transfomer-based model named TransRP for prediction of RFS in OPSCC patients based on PET/CT/GTV and clinical data. (2) TransRP combines a CNN-encoder for the extraction of rich image features and a Transformer-encoder for learning global contexts. (3) We evaluated the performance of TransRP based on two different CNN-encoders: ResNet18 (He et al., 2016) and DenseNet121 and found that these two TransRP models both showed better prediction than CNN-based models. (4) We compared three approaches for effctive incorporating clinical data into TransRP.

2. Methods

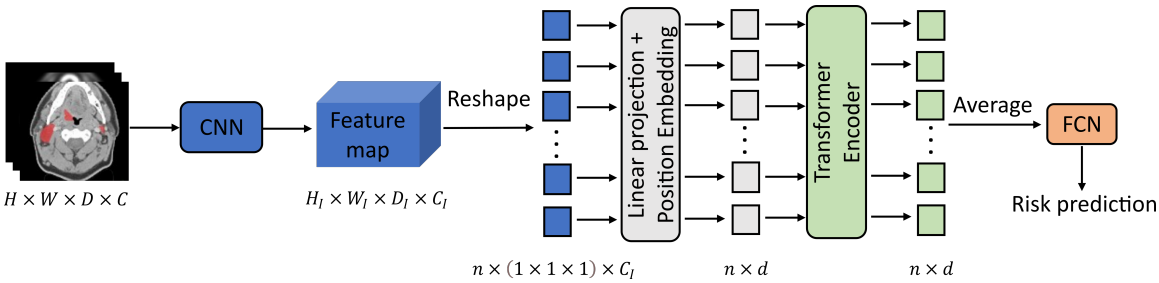

Figure 1: The architecture of the proposed TransMP based on imaging data.

The TransRP model in Figure 1, includes three main components: a CNN-based encoder, a Transformer-based encoder (ViT), and a fully connected network (FCN). The CNN-based encoder processes PET/CT/GTV volumes to extract image features, which are then passed to a linear projection and position embedding process for maintaininning spatial information. The embedded patches are fed into the transformer encoder, which uses self-attention to learn global contexts. The output of the Transformer encoder is then fed to FCN to for the time-to-event prediction of RFS. In the following sections, we will describe the TransRP architecture in further detail and different ways to incorporate clinical data into the TransRP model.

2.1. Model architecture

2.1.1. CNN-based encoder

The 3D ResNet18 (He et al., 2016; Hara et al., 2018) and 3D DenseNet121 (Huang et al., 2017) are two options for the CNN encoder in TransRP model. These models were implemented using the default settings in the MONAI 1.0.0 package with slight adjustments. The detailed architecture of the 3D ResNet18 is shown in Figure 3 of Appendix A, and the 3D DenseNet121 was constructed using the default settings in MONAI 1.0.0, with the exception of setting the kernel and stride to 1 in the average pooling of the third transition layer. Given an input image $x \in R^{H \times W \times D \times C}$ with size $H \times W \times D$ and C channels, the CNN encoder produces image feature maps $f_{image} \in R^{H_I \times W_I \times D_I \times C_I}$ with size $H_I \times W_I \times D_I$ and C_I channels.

2.1.2. Patch embedding

Initially, the image feature maps obtained from the CNN-based encoder $f_{image} \in R^{H_I \times W_I \times D_I \times C_I}$ are reshaped to $f_{image} \in R^{n \times (1 \times 1 \times 1) \times C_I}$, where $n = H_I \times W_I \times D_I$. This reshaped feature is then fed to a trainable linear projection, mapping each vectorized $1 \times 1 \times 1$ patch into a d-dimensional latent space to obtain patch embeddings $f_{emb} \in R^{n \times d}$. In order to encode the spatial information between the n patches, a learnable position embedding $f_{pos} \in R^{n \times d}$ is added to the patch embeddings to obtain the input for the Transformer encoder: $f_0 = f_{emb} + f_{pos}$.

2.1.3. Transformer encoder

The transformer encoder consists of L layers of multi-head attention (MSA) and multi-layer perceptron (MLP) blocks, as shown in Figure 3 of Appendix A. Therefore, the output of the ith layer of ViT is computed as Equation (1) and Equation (2):

$$f_i^{'} = MSA(LN(f_{i-1})) + f_{i-1} \tag{1}$$

$$f_i = MLP(LN(f_i^{'})) + f_i^{'} \tag{2}$$

where $LN()$ is the layer normalization and f_i denotes image representation. In this study, transformer encoder used the same setting with ViT-base (Dosovitskiy et al., 2020).

2.1.4. FCN and loss function

The output of transformer is averaged over n patches and input to a FCN for RFS prediction. The FCN consists of three layers with d, 64 and 1 nodes, respectively.

We adopted the Cox negative logarithm partial likelihood loss proposed in DeepSurv (Katzman et al., 2018), which is shown in Equation (3):

$$L = -\frac{1}{N_{E=1}} \sum_{i:E_i=1} (h_i - log \sum_{j \in R(T_i)} e^{h_j}) + \lambda \|\theta\|_2 \tag{3}$$

where E is the event indicator, $E = 1$ corresponds to a OPSCC patient with event and $E = 0$ indicates a censored patient, h is the predicted risk score by the model, $N_{E=1}$ is

the number of all patients with event, T is time-to-event or time-to-censored and $R(T_i)$ is the patients set whose T are not shorter than ith patient who had event. θ are network parameters and λ controls the contribution of L_2 regularization.

2.2. Incorporating clinical data

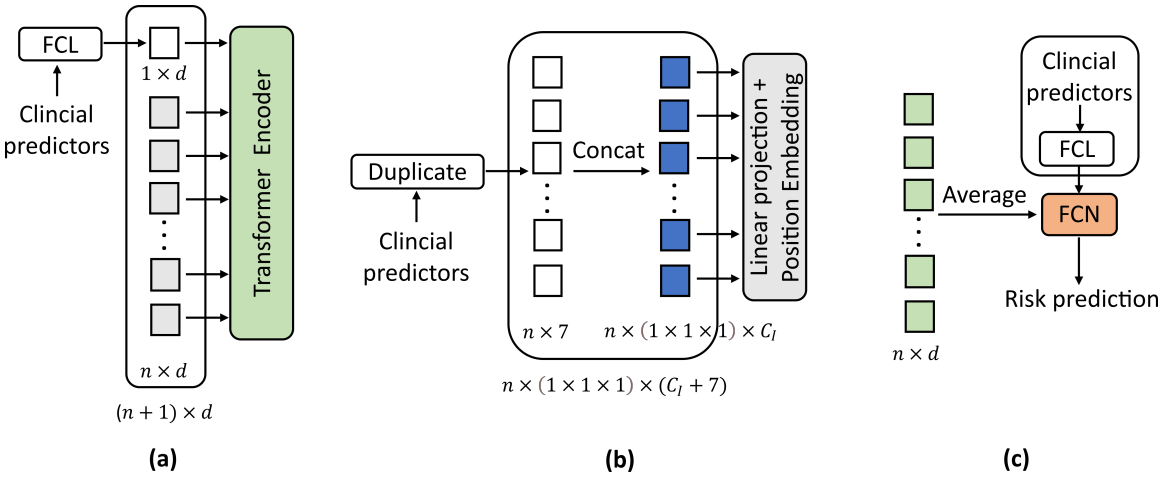

Figure 2: Three methods of incorporating clinical data into TransRP. FCL: fully connected layer.

In previous studies of Transformer-based models for outcome prediction in patients with OPSCC (Saeed et al., 2022) or nasopharyngeal cancer (Zheng et al., 2022), researchers utilized the Transformer encoder to learn and combine multi-modality information, including clinical and image data. These studies demonstrated that clinical data is essential for accurate outcome prediction. However, they did not compare the performance of their models with more simple methods for incorporating clinical data. Therefore, in this study, we compare three methods for incorporating clinical data into the TransRP model, as shown in Figure 2. The first two methods generally use the Transformer encoder to combine clinical and image features while the third utilizes a fully connected layer. We describe these methods in detail in next sections.

Method 1 (m1) was inspired by the study of Saeed et al. that used Transformer for PFS prediction in OPSCC patients (Saeed et al., 2022). As shown in Figure 2 (a), clinical predictors are first mapped to clinical features f_{clc} with size $1 \times d$. f_{clc} are then used as an additional embedded patch in the input to the Transformer encoder. The Transformer encoder is able to learn multi-modality information by learning the interaction between f_{clc} and the other image embedded patches.

Method 2 (m2) was inspired by the study of Zheng et al. that used a combination of a CNN and Transformenr encoder for survival prediction in nasopharyngeal cancer patients (Zheng et al., 2022). As illustrated in Figure 2 (b), seven clinical predictors are first duplicated to size $n \times 7$, where n is the original number of CNN-extracted feature patches. The

duplicated clinical predictors are then concatenated with the image feature patches with size $n \times (1 \times 1 \times 1) \times C_I$ to produce combined features with size $n \times (1 \times 1 \times 1) \times (C_I + 7)$. The Transformer encoder then uses these combined feature patches for RFS prediction.

Method 3 (m3) in Figure 2 (c) uses a fully connected layer with 64 nodes to extract clinical features and then concatenates them to the second last layer of FCN in TranRP, which is a very simple way.

3. Experiments and Discussion

3.1. Data and preprocessing

The HECKTOR 2022 dataset (https://hecktor.grand-challenge.org/Data/) consists of 489 OPSCC patients who were collected with pretreatment PET/CT scans, GTV of primary tumors and lymph nodes, clinical data and clinical endpoint from seven different centers. The potential clinical predictors we used include patient center (1: 7), age, gender (female: 0 vs. male: 1), weight, HPV status (positive vs. unknown vs. negative), chemotherapy (no: 0 vs. yes: 1) and surgery (no: 0 vs. yes: 1). The clinical endpoint is recurrence-free survival (RFS) with events defined as local, regional recurrence and distant metastasis and time-to-event is defined in days starting with the end of radiation therapy. Patients are censored at the time of death or the time of last follow up. Detailed description can be found in (Andrearczyk et al., 2023).

A bounding box containing GTV regions was generated with the size of $192 \times 192 \times 192mm^3$ for each patient, according to the brain segmentation position in PET (Detailed description in Appendix B). Then, the complete CT, PET and GTV image information within the bounding box was extracted and resampled to a $2 \times 2 \times 2mm^3$ resolution. The processed CT, PET and GTV were concatenated to obtain the input with size $96 \times 96 \times 96 \times 3$ for the model. Seven patients were excluded due to either abnormal PET values or no signal of brain in PET. The intensity valuse of CT and PET were truncated into ranges of [-200, 200] and [0, 25], respectively, an then were normalized to [0, 1]. Finally, 120 of 482 patients were randomly selected as the test set and 362 patients were stratified into five-fold in terms of HPV status and RFS events number.

3.2. Implementation details

The λ in Equation (3) was set to $5e^{-5}$. All models were trained using Pytorch 1.8.0 and MONAI 1.0.0 packages on a Tesla V100 GPU with 32G memory. A SGD optimizer with a momentum of 0.90 was initialized with a learning rate of 0.0002. The total training epoch was set to 200 and the learning rate decreased by a factor of 0.2 in 200th and 300th epoch, respectively. Early stopping with a patience number of 25 epochs was used to select and save the model with the highest validation performance. The risk predictions of patients in the test set were the average ensemble of predictions from five-fold cross validation models. Data augmentation included random flipping in three directions with a probability of 0.5 (p = 0.5), random affine transformation (p = 0.5) and random elastic distortion (p = 0.2). Oversampling was implemented to keep patients with events and censored patients balanced in the training set.

3.3. Comparison of methods

Because TransRP has two CNN-encoder options (ResNet18 and DenseNet121) and three optional methods (m1, m2 and m3) of incorporating clinical data. Therefore, we use TransRP(CNN-encoder name)-the method of using clinical data to denote different TransRP models. Additionally, the performance of all TransRP models were compared with a DeepSurv (Katzman et al., 2018) model which uses only clinical data, a ResNet18-based model and a DenseNet121-based method (Wahid et al., 2021) which is the winner of the task using CT/PET/GTVp for outcome prediction in HECKTOR 2021 (Andrearczyk et al., 2021). To make a fair comparison, all models were trained using the same implementation details as shown in 3.2. ResNet18 and DenseNet121 were implemented using the default setting in MONAI 1.0.0 expect setting the stride in the first convolution layer to 2 in ResNet18. Clinical data was combined to ResNet18 and DenseNet121 using method 3 by a fully connected layer. The comparison of C-index results obtained by different models in the test set is shown in Table 1.

Table 1: The C-index comparison of different models in the test set.

Model	**C-index**
DeepSurv (Katzman et al., 2018)	0.601
ResNet18-m3	0.650
DenseNet121-m3 (Wahid et al., 2021)	0.676
TansMP(ResNet18)-m1	0.624
TansMP(ResNet18)-m2	0.644
TansMP(ResNet18)-m3	0.686
TansMP(DenseNet121)-m1	0.693
TansMP(DenseNet121)-m2	0.671
TansMP(DenseNet121)-m3	0.698

From Table 1, it is obvious that DenseNet121-m3 performed better than ResNet18-m3 and the clinical data-based DeepSurv model, which means that image features can improve RFS prediction and DenseNet121 is the better way of image feature extraction than ResNet18. When compared to ResNet18-m3, only TransMP(ResNet18)-m3 can obtain higher C-index of 0.686. TransMP(ResNet18)-m1 and TransMP(ResNet18)-m2, which uses transformer to combine clinical features and image features can not obtain higher C-index than ResNet18-m3. Compared with DenseNet121-m3, TransMP(DenseNet121)-m1 and TransMP(DenseNet121)-m3 can obtain higher C-indexes of 0.693 and 0.698, respectively. In conclusion, Method 3 (m3) which uses a fully connected layer to input clinical data into TransRP obtained highest C-indexes whether the CNN-encoder is ResNet18 or DenseNet121. This finding contradicts previous studies which thought that transformer were superior in combining multi-modality information, such as clinical and imaging data, for the outcomes prediction in OPSCC or nasopharyngeal cancer patients (Saeed et al., 2022; Zheng et al., 2022).

3.4. Ablation study

The ablation study was based on TransMP(DenseNet121)-m3 as it is the best predictive model, as shown in Table 1. The study compared various models that exclude CNN and Transformer encoders together, clinical data, CNN (DenseNet121) and Transformer (ViT), respectively. The results presented in Table 2 indicate that excluding any component led to a decrease in the C-index. Particularly, excluding CNN-encoder (ViT-m3) caused a large C-index decrease from 0.698 to 0.656. This may be due to the Transformer's inability to extract detailed image features when the training set is small. It is also worth noting that TransMP(DenseNet121) which does not use clinical data, achieved a better C-index than DenseNet121-m3 (0.686 vs. 0.676). This suggests that, by learning both local detailed and global image features, TransMP without clinical data can even outperform CNN-based models that incorporate clinical data.

Table 2: The ablation study of TansMP(DenseNet121)-m3.

Clinical	**CNN**	**Transformer**	**Model**	**C-index**
✓			DeepSurv	0.601
	✓	✓	TansMP(DenseNet121)	0.686
✓		✓	ViT-m3	0.656
✓	✓		DenseNet121-m3	0.676
✓	✓	✓	TansMP(DenseNet121)-m3	0.698

4. Conclusion

We propose a novel transformer-based model named TransRP, which shows good predictive performance for the recurrence-free survival in patients with oropharyngeal squamous cell carcinoma. The TransRP uses a convolution neural network (CNN) to extract abundant CT and PET image features and then a Transformer to learn global contexts. TransRP obtained higher prediction metric than several clinical and CNN-based models. Additionally, we found that clinical data incorporated into TransRP using a simple fully connected layer showed better prediction than using transformer to combine clinical and image features.

Acknowledgments

Acknowledgments withheld.

References

Parnian Afshar, Arash Mohammadi, Konstantinos N Plataniotis, Anastasia Oikonomou, and Habib Benali. From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities. *IEEE Signal Processing Magazine*, 36(4):132–160, 2019.

Mahul B Amin, Frederick L Greene, Stephen B Edge, Carolyn C Compton, Jeffrey E Gershenwald, Robert K Brookland, Laura Meyer, Donna M Gress, David R Byrd, and

David P Winchester. The eighth edition ajcc cancer staging manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA: a cancer journal for clinicians*, 67(2):93–99, 2017.

Vincent Andrearczyk, Valentin Oreiller, Sarah Boughdad, Catherine Cheze Le Rest, Hesham Elhalawani, Mario Jreige, John O Prior, Martin Vallières, Dimitris Visvikis, Mathieu Hatt, et al. Overview of the hecktor challenge at miccai 2021: automatic head and neck tumor segmentation and outcome prediction in pet/ct images. In *3D Head and Neck Tumor Segmentation in PET/CT Challenge*, pages 1–37. Springer, 2021.

Vincent Andrearczyk, Valentin Oreiller, Sarah Boughdad, Catherine Cheze Le Rest, Hesham Elhalawani, Mario Jreige, John O Prior, Martin Vallières, Dimitris Visvikis, Mathieu Hatt, et al. Overview of the hecktor challenge at miccai 2022: Automatic head and neck tumor segmentation and outcome prediction in pet/ct. In *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer, 2023.

K Kian Ang, Jonathan Harris, Richard Wheeler, Randal Weber, David I Rosenthal, Phuc Felix Nguyen-Tân, William H Westra, Christine H Chung, Richard C Jordan, Charles Lu, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. *New England Journal of Medicine*, 363(1):24–35, 2010.

Paula Bos, Michiel WM van den Brekel, Zeno AR Gouw, Abrahim Al-Mamgani, Marjaneh Taghavi, Selam Waktola, Hugo JWL Aerts, Jonas A Castelijns, Regina GH Beets-Tan, and Bas Jasperse. Improved outcome prediction of oropharyngeal cancer by combining clinical and mri features in machine learning models. *European Journal of Radiology*, 139: 109701, 2021.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Nai-Ming Cheng, Jiawen Yao, Jinzheng Cai, Xianghua Ye, Shilin Zhao, Kui Zhao, Wenlan Zhou, Isabella Nogues, Yuankai Huo, Chun-Ta Liao, et al. Deep learning for fully automated prediction of overall survival in patients with oropharyngeal cancer using fdg-pet imaging. *Clinical Cancer Research*, 27(14):3948–3959, 2021.

André Diamant, Avishek Chatterjee, Martin Vallières, George Shenouda, and Jan Seuntjens. Deep learning in head & neck cancer outcome prediction. *Scientific reports*, 9(1):1–10, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Stefan P Haider, Tal Zeevi, Philipp Baumeister, Christoph Reichel, Kariem Sharaf, Reza Forghani, Benjamin H Kann, Benjamin L Judson, Manju L Prasad, Barbara Burtness, et al. Potential added value of pet/ct radiomics for survival prognostication beyond ajcc 8th edition staging in oropharyngeal squamous cell carcinoma. *Cancers*, 12(7):1778, 2020.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1): 1–12, 2018.

Dong Wook Kim, Sanghoon Lee, Sunmo Kwon, Woong Nam, In-Ho Cha, and Hyung Jun Kim. Deep learning-based survival prediction of oral cancer patients. *Scientific reports*, 9(1):1–10, 2019.

Mingyuan Meng, Lei Bi, Dagan Feng, and Jinman Kim. Radiomics-enhanced deep multi-task learning for outcome prediction in head and neck cancer. *arXiv preprint arXiv:2211.05409*, 2022.

Brian O'Sullivan, Shao Hui Huang, Jie Su, Adam S Garden, Erich M Sturgis, Kristina Dahlstrom, Nancy Lee, Nadeem Riaz, Xin Pei, Shlomo A Koyfman, et al. Development and validation of a staging system for hpv-related oropharyngeal cancer by the international collaboration on oropharyngeal cancer network for staging (icon-s): a multicentre cohort study. *The Lancet Oncology*, 17(4):440–451, 2016.

Shuchao Pang, Matthew Field, Jason Dowling, Shalini Vinod, Lois Holloway, and Arcot Sowmya. Training radiomics-based cnns for clinical outcome prediction: Challenges, strategies and findings. *Artificial Intelligence in Medicine*, 123:102230, 2022.

Louis Rebaud, Thibault Escobar, Fahad Khalid, Kibrom Girum, and Irene Buvat. Simplicity is all you need: Out-of-the-box nnunet followed by binary-weighted radiomic model for segmentation and outcome prediction in head and neck pet/ct.

Numan Saeed, Ikboljon Sobirov, Roba Al Majzoub, and Mohammad Yaqub. Tmss: An end-to-end transformer-based multimodal network for segmentation and survival prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 319–329. Springer, 2022.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Kareem A Wahid, Renjie He, Cem Dede, Abdallah SR Mohamed, Moamen Abobakr Abdelaal, Lisanne V van Dijk, Clifton D Fuller, and Mohamed A Naser. Combining tumor segmentation masks with pet/ct images and clinical data in a deep learning framework for improved prognostic prediction in head and neck squamous cell carcinoma. In *3D Head and Neck Tumor Segmentation in PET/CT Challenge*, pages 300–307. Springer, 2021.

Kai Wang, Yunxiang Li, Michael Dohopolski, Tao Peng, Weiguo Lu, You Zhang, and Jing Wang. Recurrence-free survival prediction under the guidance of automatic gross tumor volume segmentation for head and neck cancers. *arXiv preprint arXiv:2209.11268*, 2022a.

Yiling Wang, Elia Lombardo, Michele Avanzo, Sebastian Zschaek, Julian Weingärtner, Adrien Holzgreve, Nathalie L Albert, Sebastian Marschner, Giuseppe Fanetti, Giovanni Franchin, et al. Deep learning based time-to-event analysis with pet, ct and joint pet/ct for head and neck cancer prognosis. *Computer Methods and Programs in Biomedicine*, page 106948, 2022b.

Nora Würdemann, Steffen Wagner, Shachi Jenny Sharma, Elena-Sophie Prigge, Miriam Reuschenbach, Stefan Gattenlöhner, Jens Peter Klussmann, and Claus Wittekindt. Prognostic impact of ajcc/uicc 8th edition new staging rules in oropharyngeal squamous cell carcinoma. *Frontiers in oncology*, 7:129, 2017.

Hanci Zheng, Zongying Lin, Qizheng Zhou, Xingchen Peng, Jianghong Xiao, Chen Zu, Zhengyang Jiao, and Yan Wang. Multi-transsp: Multimodal transformer for survival prediction of nasopharyngeal carcinoma patients. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–243. Springer, 2022.

Appendix A. The detailed architectures of ResNet18 and Transformer encoder

The detailed architectures of ResNet18 and Transformer encoder are shown in Figure 3.

Appendix B. Flow of bounding box determination

The bounding box extraction method was inspired by the official code provide by HECKTOR challenge (https://github.com/voreille/hecktor). Firstly, the brain of each patient was segmented using PET threshing with the threshold 3-5. After getting brain segmentation, a point p with position (x, y, z) was determined as the point $30mm$ lower from the lowest voxel of the brain. Then, bounding box was obtained according to the position $(x - 96 : x + 96, y - 84 : y + 108, z - 72 : z + 120)$ where x, y and z denote coronal, sagittal and axial plans, respectively.

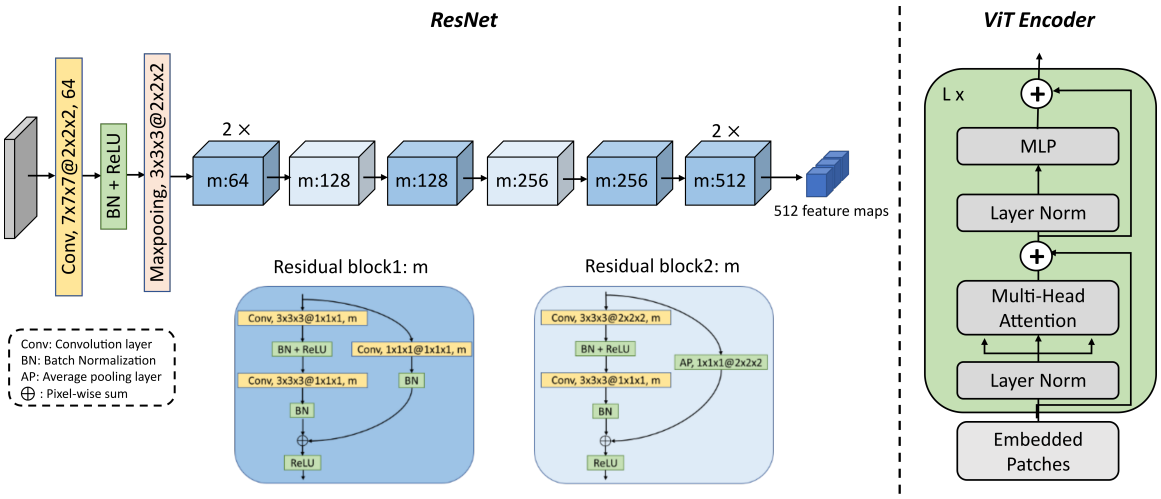

Figure 3: The detailed architecture of ResNet and ViT encoder. The notation "a×b×c@i×j×k" denotes a kernel size of a×b×c and a stride of i×j×k. m is the number of convolution kernel in residual blocks. MLP: Multi-layer peceptron.