# HECKTOR2025 Challenge Report: Fully Automated Diagnoses of HPV Status using PET/CT Images and Clinical Information

Muzi Guo[1], Tianen Yu[2,3], Jian He[3], and Lei Xiang[1]

[1] Subtle Medical Inc., CA, USA
[2] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[3] Department of Nuclear Medicine, Nanjing Drum Tower Hospital, Affiliated Hospital of Medical School, Nanjing University, Nanjing, China

**Abstract.** Accurate prediction of human papillomavirus (HPV) status is essential for risk stratification and personalized treatment planning in head and neck cancer. In this work, we propose a multi-modal deep learning framework to classify HPV status using the HECKTOR25 Task 3 dataset, which provides 3D FDG-PET and CT scans with clinical data. Our approach leverages a 3D ResNet-18 architecture for imaging feature extraction, combined with a fully connected network to encode clinical variables, followed by multimodal fusion for final prediction. To address the significant class imbalance problem, we implemented a weighted cross-entropy loss. On internally held-out test splits, the model achieved a specificity of 0.9167 and a balanced accuracy of 0.9017, demonstrating robust intra-dataset performance. However, evaluation on the organizers external dataset—which contains cases from centers not included in the training data—yielded reduced performance (validation specificity 0.9048, balanced accuracy 0.6765), highlighting the challenges of cross-center generalization. These findings underscore the potential of multimodal deep learning for HPV status prediction and indicate that further strategies are required to enhance model robustness to inter-center variability.

**Keywords:** HECKTOR2025 · classification challenge · PET/CT · deep learning.

## 1 Introduction

### 1.1 Motivation

Head and neck cancer (HNC) represents a heterogeneous group of malignancies, with nearly 90% classified as squamous cell carcinomas (HNSCC) [1]. Well-established risk factors, such as tobacco and alcohol consumption, act synergistically to markedly increase disease incidence [2, 3]. Despite advances in therapeutic strategies, the overall 5-year survival rate remains poor, particularly among

patients with advanced disease. Current clinical workflows—primarily based on TNM staging, imaging, and histopathology—are often insufficient for early detection and accurate prognostication. Although biomarkers such as HPV status have improved risk stratification, their integration into routine clinical practice remains limited. Artificial intelligence (AI) offers promising opportunities by leveraging multimodal data—including medical imaging, radiotherapy dose distributions, and electronic health records—to enhance diagnostic accuracy and prognostic modeling. However, progress in AI-driven approaches for HNC has been constrained by the scarcity of large, diverse, and publicly available datasets, hindering clinical translation. The Head and Neck Tumor Segmentation and Outcome Prediction Challenge (HECKTOR 2025) [4] addresses this gap by providing a large-scale, multi-institutional, multimodal dataset collected from 10 international centers, creating new opportunities for the development and validation of AI methods in tumor detection, segmentation, outcome prediction, and treatment optimization.

### 1.2 Dataset

In this study, we focus on the classification task of predicting HPV status using FDG-PET/CT images in combination with available clinical information. Each case in both the training and test cohorts consists of a paired 3D FDG-PET volume and a corresponding 3D CT volume. The clinical variables include center, gender, age, tobacco and alcohol use, performance status, treatment (radiotherapy alone or chemoradiotherapy), M-stage (metastasis), and HPV status.

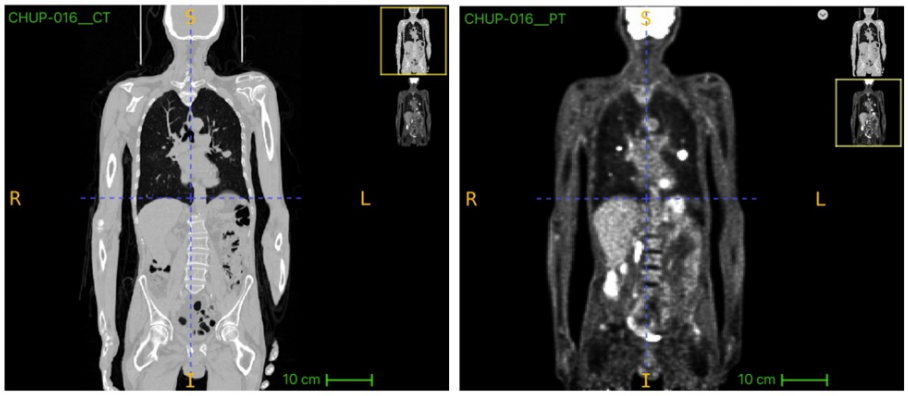


**Fig. 1.** An example coronal view of a PET image and its corresponding CT image with the tumor.

The training set comprises 588 cases collected from 6 centers, including 4 in Canada, 1 in France, and 1 in the United States. The test set contains 23

previously unseen cases from 3 centers and include approximately 80% HPV-positive and 20% HPV-negative cases. The CT and PET volumes were rigidly aligned to a common origin but remain heterogeneous in field of view(FOV) and resolution. Representative examples of CT and PET images, together with corresponding clinical information, are illustrated in Figures 1 and 2.

| PatientID | CenterID | Age | Gender | Tobacco Consumption | Alcohol Consumption | Performance Status | Treatment | M-stage | HPV Status |
|---|---|---|---|---|---|---|---|---|---|
| CHUM-012 | 1 | 61 | 1 | NaN | NaN | NaN | 1 | M0 | 1 |
| CHUM-014 | 1 | 59 | 1 | NaN | NaN | NaN | 1 | M0 | 1 |
| CHUM-016 | 1 | 58 | 1 | NaN | NaN | NaN | 1 | M0 | 1 |
| CHUM-021 | 1 | 63 | 1 | NaN | NaN | NaN | 1 | M0 | 1 |
| CHUM-022 | 1 | 56 | 1 | NaN | NaN | NaN | 1 | M0 | 1 |
| CHUM-026 | 1 | 67 | 0 | NaN | NaN | NaN | 1 | M0 | 1 |
| CHUM-029 | 1 | 50 | 1 | NaN | NaN | NaN | 1 | M0 | 1 |
| CHUM-033 | 1 | 68 | 1 | NaN | NaN | NaN | 1 | M0 | 1 |
| CHUM-036 | 1 | 51 | 1 | NaN | NaN | NaN | 1 | M0 | 1 |
| CHUM-037 | 1 | 70 | 0 | NaN | NaN | NaN | 1 | M0 | 1 |

**Fig. 2.** An example of the clinical information subset.

## 2 Method

### 2.1 Data Preprocessing

**Crop Head and Neck Region** Although most of the cases included only the head and neck region, a subset of cases from several centers consisted of whole-body scans. Because the analysis focused exclusively on the head and neck, the axial coverage was standardized by cropping the slice range. The slice range (in millimeters) was calculated as the product of the number of slices and the voxel size along the z-axis. A comprehensive review of the image sets indicated that a coverage of 516 mm was sufficient to cover the entire head and neck region. Consequently, for scans with an axial coverage exceeding 516 mm, the volume was cropped to this limit.

In addition, for certain cases—primarily from the CHUS center—where the CT slice range was smaller than the corresponding PET slice range, the PET volume was cropped along the z-axis to match the CT slice range, ensuring consistent spatial alignment between modalities.

**Align PET and CT images** To fully leverage CT information, spatial alignment was performed between PET and CT images. As described above, we first calculated the field of view (FOV) for PET and CT images, respectively. The absolute difference between the two FOVs was then computed. If the FOV of the PET images exceeded that of the CT images, the CT volume was symmetrically padded along the x and y axes to match the PET FOV. The number of padding pixels was determined by dividing the FOV difference by the CT pixel spacing. After padding, CT images were resampled to match the resolution of

PET images. In contrast, if the FOV of the PET images was smaller than that of the CT images, the CT volume was cropped to match the PET FOV before resampling.

**Data Normalization** Before intensity rescaling, CT images were adjusted using a mucosal window (window level = 40 HU, window width = 300 HU) to enhance soft tissue contrast and improve the conspicuity of small tumors. Subsequently, the intensities of the CT and PET images were independently normalized by dividing each by their respective mean intensities. After normalization, CT and PET images were concatenated to generate a two-channel input for the model.

**Clinical Information Processing** For the clinical data, z-score normalization was applied to age and gender variables. Tobacco consumption, alcohol consumption, performance status, and M-stage were encoded using one-hot encoding. The processed clinical features were then concatenated and fed into a fully connected network to extract high-level clinical representations.

### 2.2 Model Architecture

Our proposed network consists of two parallel branches designed to jointly leverage imaging and clinical information for classification (Fig. 3).

**Image Branch** The image branch uses a 3D ResNet-18 backbone implemented by MONAI [5] to extract deep volumetric features from the PET and CT inputs. After preprocessing step, the two modalities are concatenated along the channel dimension to form a two-channel input volume. The 3D ResNet-18 is an extension of the conventional 2D ResNet-18, with all convolutional and pooling layers replaced by their 3D counterparts, enabling the network to capture spatial context across three dimensions. The architecture consists of an initial $7\times7\times7$ convolutional layer (stride 2) followed by a $3\times3\times3$ max-pooling layer. Four residual stages are subsequently stacked, each containing two basic residual blocks with identity skip connections to facilitate gradient propagation. The number of filters doubles at each stage (64, 128, 256, 512), progressively enriching the learned feature representation. A global average pooling layer is applied to obtain a compact feature vector.

**Clinical Branch** The clinical branch processes tabular clinical data. Age and gender are normalized using z-score normalization, while tobacco consumption, alcohol consumption, performance status, and M-stage are one-hot encoded. These preprocessed features are passed through a two-layer fully connected network with ReLU activation to extract high-level clinical feature representations.

**Feature Fusion and Classification** The feature vectors from the image and clinical branches are concatenated and fed into a final fully connected layer to output class probabilities. This design allows the model to jointly learn complementary information from imaging and clinical data, improving the classification performance.

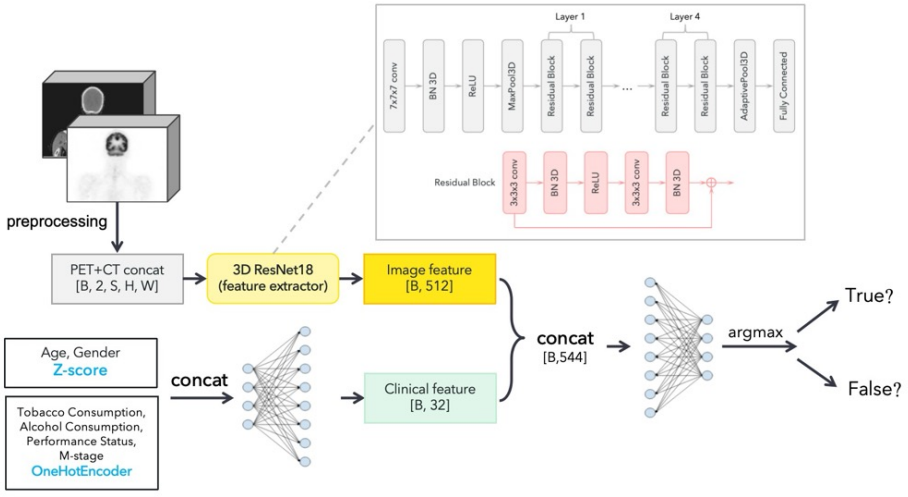


**Fig. 3.** An illustration of the training pipeline for the classification network.

### 2.3 Loss Function

Because the classification dataset contained 530 HPV-positive and 58 HPV-negative cases, we adopted a weighted cross-entropy loss to address the severe class imbalance. This loss extends the standard cross-entropy by incorporating class-specific weighting factors, thereby increasing the contribution of minority classes during training.

$$\mathcal{L}_{WCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c \cdot y_{i,c} \log\left(p_{i,c}\right) \tag{1}$$

where $w_c$ represents the weight assigned to class $c$, $y_{i,c}$ is the ground truth indicator, $p_{i,c}$ is the predicted probability, $N$ is the number of samples, and $C$ is the number of classes. For each sample, the standard cross-entropy was multiplied by a class-specific weight, typically set as the inverse of the class frequency:

$$w_c = \frac{N}{C \cdot N_c} \tag{2}$$

where $N_c$ denotes the number of samples in class $c$.

## 3 Experiment Setup

We utilized the HECKTOR25 Task 3 dataset and randomly divided it into five folds after holding out 20% test set, training a separate model for each fold. No additional data or pre-trained models were used. All input volumes were resized to $192 \times 192 \times 192$ to accommodate GPU memory constraints. Model training was performed using the AdamW optimizer with an initial learning rate of $2 \times 10^{-4}$, which was gradually reduced to zero by the end of training via a cosine annealing scheduler. Each model was trained for 100 epochs with a batch size of 1 on a single NVIDIA GeForce RTX 3090 GPU (24 GB).

## 4 Results

**Table 1.** Table captions should be placed above the tables.

| Fold | Best Validation Accuracy | Test Accuracy | Balanced Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|---|
| 1 | 0.8941 | 0.8559 | 0.8090 | 0.8679 | 0.7500 |
| 2 | 0.9529 | 0.8644 | 0.8876 | 0.8585 | 0.9167 |
| 3 | 0.8824 | 0.8305 | 0.8687 | 0.8208 | 0.9167 |
| 4 | 0.9059 | 0.8898 | 0.9017 | 0.8868 | 0.9167 |
| 5 | 0.9286 | 0.8559 | 0.8459 | 0.8585 | 0.8333 |

Based on our data splits, the results of a single-run 5-fold cross-validation are summarized in Table 1. The evaluation metrics include the best validation accuracy, test accuracy, balanced accuracy, sensitivity, and specificity. Validation and test accuracies were computed as the number of correctly classified cases divided by the total number of cases in the corresponding dataset. Specificity, sensitivity, and balanced accuracy were calculated as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

$$balanced\ accuracy = \frac{specificity + sensitivity}{2} \tag{5}$$

where $TN$, $FP$, $TP$, and $FN$ denote true negatives, false positives, true positives, and false negatives, respectively. As indicated by the challenge organizers, the official ranking is the balanced accuracy. Therefore, we selected the fold-4 model for final validation and testing submissions. The test results remain undisclosed, while on the validation set we achieved a specificity of 0.9048

and a balanced accuracy of 0.6765, suggesting that the model performs well at correctly identifying HPV-negative cases but struggles to detect HPV-positive cases. This imbalance in performance may reflect the skewed class distribution in the training data and indicates that additional strategies—such as more aggressive class rebalancing, data augmentation, or cost-sensitive learning—may be required to improve sensitivity without sacrificing specificity.

## 5 Conclusion and Discussion

In this work, we developed a multimodal classification framework for HPV status prediction using the HECKTOR25 Task 3 dataset, which contains 3D PET/CT images and available clinical information. The proposed model, based on a 3D ResNet-18 backbone with a parallel clinical feature branch, was trained using a weighted cross-entropy loss to mitigate class imbalance. Our five-fold cross-validation results demonstrated high specificity (0.9048) but relatively low sensitivity, resulting in a balanced accuracy of 0.6765 on the validation set. These findings suggest that the model performs well in identifying HPV-negative cases but requires further improvement for detecting HPV-positive cases.

We investigated the potential reasons why the model performed well on the internal training dataset but showed reduced performance on the external dataset. One contributing factor may be that some centers in the external dataset were not included in the training data.

We further analyzed the distribution of HPV-negative cases and misclassified instances across different centers.

**Table 2.** Distribution of HPV negative and wrong prediction cases in Task 3 dataset across different centers.

| Center | Total Cases | HPV Negative | %HPV Negative | Wrong Prediction | %Wrong Prediction |
|---|---|---|---|---|---|
| CHUM | 22 | 1 | 4.55% | 5 | 23% |
| CHUP | 58 | 37 | 63.97% | 4 | 7% |
| CHUS | 33 | 8 | 24.24% | 6 | 18% |
| HGJ | 38 | 10 | 26.32% | 7 | 18% |
| HMR | 2 | 0 | 0.00% | 0 | 0% |
| MDA | 435 | 2 | 0.46% | 1 | 0% |

Centers such as HMR and MDA, with very few negative cases, and CHUP, with a relatively balanced class distribution, showed higher model performance. In contrast, CHUM, CHUS, and HGJ centers, which have both limited total cases and low numbers of negative cases, showed reduced performance. These findings underscore the importance of accounting for class distribution not only at the dataset level but also within each center. Future work should incorporate strategies such as data resampling, reweighting, or selective sampling to mitigate center-specific imbalances, thereby enhancing the model generalizability across heterogeneous clinical settings.

## References

1. Bhat, G.R., Hyole, R.G. and Li, J., 2021. Head and neck cancer: Current challenges and future perspectives. In Advances in cancer research (Vol. 152, pp. 67-102). Academic Press.
2. Dal Maso, L., Torelli, N., Biancotto, E., Di Maso, M., Gini, A., Franchin, G., Levi, F., La Vecchia, C., Serraino, D. and Polesel, J., 2016. Combined effect of tobacco smoking and alcohol drinking in the risk of head and neck cancers: A re-analysis of case–control studies using bi-dimensional spline models. European journal of epidemiology, 31(4), pp.385-393.
3. Dhull, A.K., Atri, R., Dhankhar, R., Chauhan, A.K. and Kaushal, V., 2018. Major risk factors in head and neck cancer: a retrospective analysis of 12-year experiences. World journal of oncology, 9(3), p.80.
4. Saeed, N., Hassan, S., Hardan, S., Aly, A., Taratynova, D., Nawaz, U., Khan, U., Ridzuan, M., Eugene, T., Metz, R. and Delpon, G., 2025. A Multimodal Head and Neck Cancer Dataset for AI-Driven Precision Oncology. arXiv preprint arXiv:2509.00367.
5. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D. and Nath, V., 2022. Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701.