



HPV Positivity Prediction for Head and Neck Cancer using CT images

Oya Altınok^{1,2}, Matthew B. Schabath¹, Albert Guvenis²

¹Moffitt Cancer Center, Tampa, FL, USA

²Bogazici University, Istanbul, Türkiye
guvenis@boun.edu.tr

Precision medicine personalizes treatment using genetic, environmental, lifestyle, and test data, often analyzed by AI models trained on large patient populations. The goal is to maximize health benefits and minimize side effects. In this study, we used CT images to predict HPV positivity. This information is crucial in personalizing the treatment. While this information can be obtained from biopsy, imaging provides us with a non-invasive alternative method.

This particular retrospective study used publicly available TCIA data from 192 OPC patients (85% HPV+). CT scans were first resampled to $1 \times 1 \times 1 \text{ mm}^3$ using cubic interpolation, z-score normalized, and entropy maps generated for texture enhancement. Original and entropy-filtered CTs were fused. Missing values were imputed with the median, and features were z-score normalized. Data was stratified into 70% training and 30% test sets. SMOTE was applied to the training set to address class imbalance. Eighteen radiomic features were extracted from the entire tumor using both CT types. An optimizable Gentle-Boost ensemble model with decision tree base learners was built in MATLAB's Classification Learner (2024b) to predict HPV status. This model was chosen for its effectiveness with moderate feature sets and bias-variance reduction. Model performance was assessed using ROC AUC with 10-fold cross-validation on the training set; the 30% test set was held out.

The model yielded an AUC of 0.77 (training) and 0.75 (test). We therefore conclude that machine learning can predict H&N tumor characteristics from CT images, potentially impacting treatment choices. These results confirm prior findings. Work is underway to improve accuracy by using methods that take heterogeneity into account and better segmentation schemes. Further validation through multi-center prospective studies is necessary, along with harmonization of imaging data from different equipment and protocols. More balanced datasets should also be beneficial.

Keywords: HPV positivity, head and neck cancer, radiomics, machine learning, precision medicine