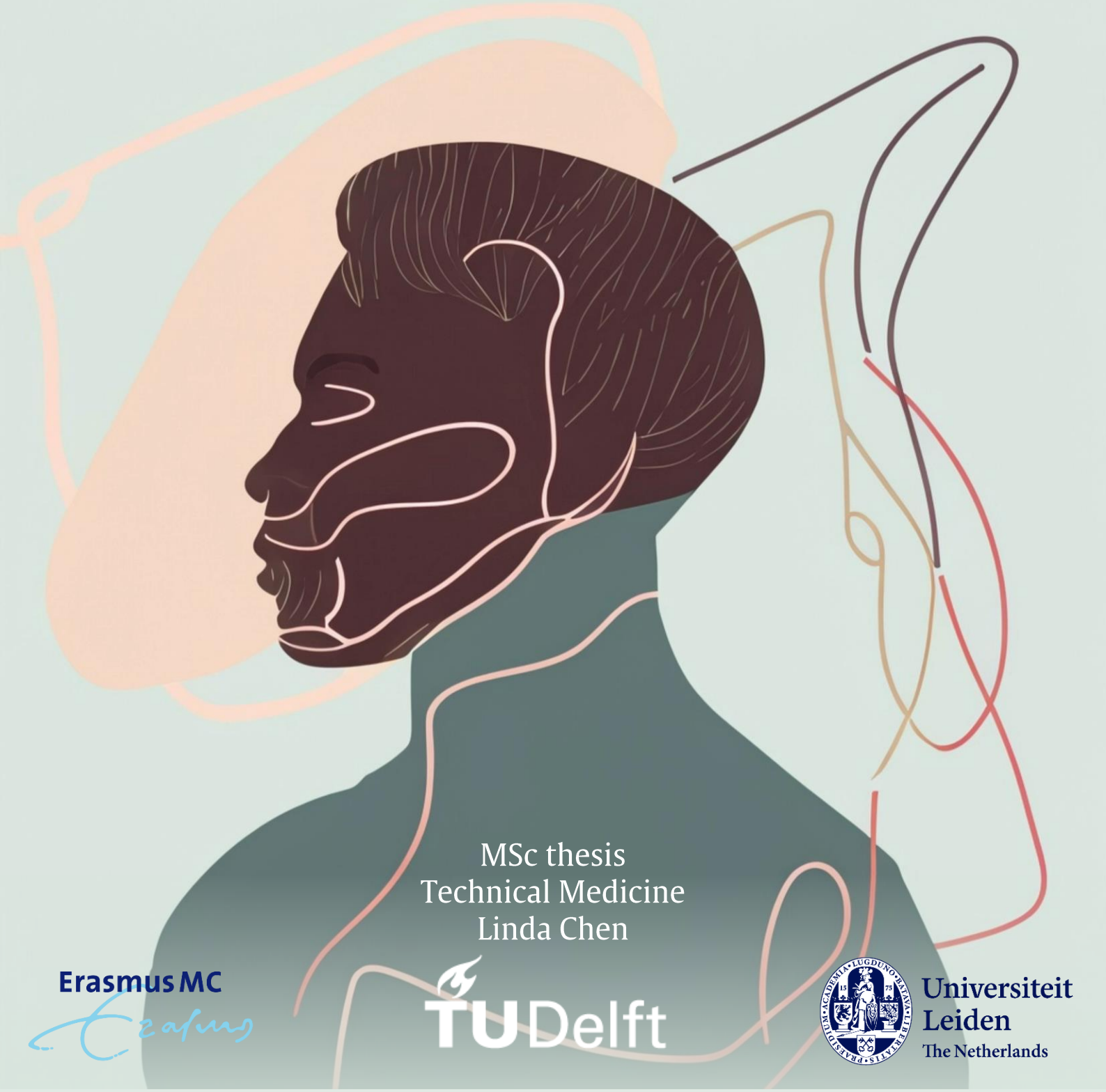
Predicting HPV status in oropharyngeal carcinoma using multiparametric MRI and clinical variables



MSc thesis
Technical Medicine
Linda Chen

Erasmus MC
TU Delft
Universiteit Leiden
The Netherlands

This page was intentionally left blank.

PREDICTING HPV STATUS IN OROPHARYNGEAL CARCINOMA USING MULTIPARAMETRIC MRI AND CLINICAL VARIABLES

Linda L. Chen

Student number: 4648242

23-08-2023

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University; Delft University of Technology; Erasmus University Rotterdam

Master thesis project (TM30004; 35 ECTS)

Dept. of Biomechanical Engineering, TUDELFT

12-12-2023 – 23-08-2023

Supervisors:

dr. ir. S.F. Petit	Technical supervisor
Dr. M. E. Capala	Medical supervisor
ir. I. Lauwers	Daily supervisor

Thesis committee members:

dr. ir. S.F. Petit	Erasmus MC (chair)
Dr. M.E. Capala	Erasmus MC
dr. ir. S.J.M. Habraken	LUMC and HollandPTC
ir. I. Lauwers	Erasmus MC
dr. J. F. Veenland	Erasmus MC and TU Delft

MSc thesis Technical Medicine (2023) 1–10

Original Article

Predicting HPV status in oropharyngeal carcinoma using multiparametric MRI and clinical variables

Linda L. Chen[a,b,c,*], Iris Lauwers[a], Marta Capala[a], Gerda Verduijn[a], Steven Petit[a]

[a] *Department of Radiotherapy, Erasmus MC Cancer Institute, Rotterdam, the Netherlands*
[b] *Faculty of Mechanical, Maritime and Materials Engineering, University of Technology Delft, Delft, the Netherlands*
[c] *Faculty of Medicine, Leiden University Medical Centre, Leiden, the Netherlands*

Abstract

Background: Histopathological examination in the diagnostic workflow of oropharyngeal squamous cell carcinoma (OPSCC) is essential. We aimed to develop a machine learning pipeline to predict human papillomavirus (HPV) status in OPSCC patients based on clinical variables and multiparametric magnetic resonance imaging (MRI).
Methods: In a dataset of OPSCC patients (n=59), we extracted features from three categories: clinical variables; histogram parameters from diffusion weighted imaging (DWI)-MRI model; radiomics based on T2-weighted and DWI-MRI. We performed ten-times repeated stratified five-fold cross-validation and divided each outer training set (80%) into an inner training set (80% of outer training set) and validation set (20% of outer training set) using five-fold stratified cross-validation. We performed three types of feature selection methods (LASSO, statistical analysis and manual selection), tuned and trained seven classifiers (logistic regression, k-nearest-neighbours, naive bayes, random forest, support vector machine, XGBoost and LightGBM) to find the optimal combination of features, hyperparameters and classifier on each inner training set. We ensembled the inner fold models to fit on the outer training set and tested on the outer test set (20%). We constructed additional models with subsets of the features.
Results: the combined model area under the curve was 0.793 ± 0.136. Models including clinical features outperformed models without clinical features ($p<0.001$). Features from all feature categories were selected for the combined model.
Conclusion: we were able to predict HPV status in OPSCC patients using multiparametric MRI and clinical variables with reasonable accuracy, though retraining and validating on larger, external datasets is needed before implementation in clinic.

Keywords:
machine learning, radiomics, magnetic resonance imaging, head and neck cancer, diffusion weighted imaging, human papillomavirus, oropharyngeal squamous cell carcinoma.

1. Introduction

Histopathological assessment for oropharyngeal squamous cell carcinoma (OPSCC) is an invasive, but essential step in the diagnostic workup of OPSCC, as a biopsy of the tumour confirms malignancy and determines the subsequent treatment choice [1]. Within OPSCC, human papillomavirus (HPV) status has been recognized as one of the most important risk factors [2, 3]. In general, patients with HPV-positive tumours are more responsive to treatment with irradiation, and thus have lower mortality rates [2, 4-6]. Currently, patients with locally advanced HPV-positive and HPV-negative tumours receive the same treatment, consisting of chemotherapy and radiation therapy (RT) [5, 7-12]. However, chemo-radiation frequently leads to side-effects such as xerostomia, dysphagia and dermatitis [7, 10-12]. To explore the possibility of treatment de-escalation for HPV-positive OPSCC, several studies are currently being conducted [5, 8, 13-15].

For possible treatment adaptation and expectation management, pre-treatment determination of HPV status is essential. Currently, histopathological analyses such as p16 immunohistochemistry (IHC) are used in clinical practice for determination of HPV status [16-18]. However, these methods are invasive, costly, and labour-intensive [16, 17, 19, 20]. Hence, we investigate the possibility of using data collected in routine

*Corresponding author
Email address: l.chen@erasmusmc.nl (Linda L. Chen)

clinical workflow to predict HPV status.

Throughout literature, clinical features such as alcohol abuse and smoking were found to be more prevalent in HPV-negative patients, as well as a higher T-stage in HPV-positive patients and higher N-stage in HPV-negative patients [21-24]. These clinical features could be predictive of HPV status, to which Magnetic Resonance Imaging (MRI) could be of added value, and several studies have shown differences between HPV-positive and HPV-negative tumours on MRI [25-31]. Moreover, MRI is non-invasive, represents the entire tumour, allows for retrospective analysis, and is integrated in the treatment planning workflow for OPSCC irradiation, as soft tissue can be best distinguished on MRI and is thus used for tumour delineation in clinical practice [30, 31].

Furthermore, MRI has the potential to provide information about the microstructure and tissue perfusion of the tumour using sequences such as Diffusion Weighted Imaging (DWI) [32]. DWI-MRI studies have demonstrated that the measure of water diffusion within tissue was significantly higher in HPV-negative tumours than HPV-positive tumours, and could be predictive of treatment response [21, 24, 32-43]. A number of studies reported the feasibility of using MRI to predict HPV status, though to our best knowledge, there have been no studies that make use of the non-gaussian intravoxel incoherent motion (NG-IVIM) model for this purpose, which gives insight in the microvascular perfusion, and inter- and intra-cellular diffusion in tumours [21-23, 25, 28, 29, 38, 44]. Moreover, radiomics is able to find discerning features in imaging through histogram or textural analysis, whilst ML makes it possible to analyze and utilize the vast amounts of data from all features [23, 28, 29, 45, 46].

Our goal was to develop a machine learning (ML) pipeline to predict HPV status in OPSCC patients based on clinical data, and radiological and radiomics features from T2-weighted (T2w) and DWI-MRI images, using different feature selection methods and classifiers.

2. Materials and methods

2.1. Patient population

We retrospectively reviewed all consecutive patients with histologically proven primary OPSCC who were treated at the Erasmus MC between April 2020 and April 2023 for whom written informed consent was obtained [47]. Inclusion criteria were patients who had: (i) histologically confirmed OPSCC; (ii) no distant metastases; (iii) pre-treatment DWI-MRI available on which the primary tumour was visible; (iv) received primary (chemo-)radiotherapy; and (v) HPV status determined through p16 IHC. Fifty-nine patients met our inclusion criteria and formed our study population. Patients were treated with radiotherapy with current clinical protocols of 70Gy intensity modulated RT or intensity modulated proton beam therapy in 35 fractions, with or without the addition of cisplatin or cetuximab [47]. IHC of p16 protein overexpression was performed on tissue samples as a surrogate marker for HPV status according to clinical practice, where strong and diffuse nuclear cytoplasmic immunostaining in >70% of the tumour cells was considered p16-positive [48, 49].

2.2. MRI data acquisition

All patients underwent MRI examination on a 1.5T GE MR450w (GE, Waukesha, WI, USA) using MR Radiation Oncology Suite Coils (GE, Waukesha, WI, USA) with the patients immobilized in head-first-supine RT treatment position. We applied the imaging protocol described by Verduijn et al., of which we used the T2-weighted (T2w) turbo spin echo and DWI images for this study (flip angle: 90 degrees; repetition time 6700ms; echo time 81.8ms; field of view 26×26cm; 4mm slice thickness; 0.2mm gap, 128×128 matrix; bandwidth: 1953.12 Hz/pixel) [47]. We acquired the DWI-MRI images using fifteen b-values (0, 10, 2×80, 130, 570, 2×770, 2×780, 790, and $4\times1500 s/mm^2$) in three orthogonal diffusion directions according to the pipeline by Sijtsema et al [50]. All gross tumour volumes (GTVs) were manually delineated on the T2w images by one expert observer (experienced head-and-neck radiation oncologist) or by one less-experienced observer, which were verified and corrected by the expert observer. DWI images were rigidly registered to the T2w images using rotation and translation with mattes mutual information and the tumour delineation was corrected to exclude air pockets on DWI images if necessary based on the b = 0 s/mm2 image [51].

2.3. Feature extraction

2.3.1. Clinical data

We collected HPV status, TNM classification (7th edition), alcohol status, and smoking status for each patient. Alcohol status was categorized as never (no alcohol abuse in history or present); occasional alcohol use (≤2 units/day for women, ≤3 units/day for men); active alcohol abuse (>2 units/day for women, >3 units/day for men); and alcohol abuse in history [52]. We categorized smoking status as never (≤100 cigarettes in life); past (>100 cigarettes in life, but stopped before treatment); current (active smoker).

2.3.2. Histogram parameters from NG-IVIM model from DWI-MRI

The NG-IVIM model *(Eq. 1)* is an extension of the apparent diffusion coefficient (ADC) model *(Eq. 2)* which in addition to intercellular diffusion also characterizes perfusion and restricted diffusion simultaneously, and thus gives more insight in the microvascular perfusion and inter- and intra-cellular diffusion compared to the ADC model, which exclusively measures intercellular diffusion [44].

Voxel-wise least square fitting of the NG-IVIM model *(Eq. 1)* and the ADC model *(Eq. 2)* was carried out in the GTVs as described in Sijtsema et al. in MATLAB 2021b [53, 54]:

$$S_b = S_0((1-f)(e^{-bD+\frac{1}{6}(bD)^2}) + fe^{-bD^*}) \quad \text{(Eq. 1)}$$

In Eq. 1, S_b is the signal at the b-value, S_0 is the signal at b=0s/mm^2, f is the perfusion fraction, D is the diffusion coefficient, K is the kurtosis and D^* is the pseudo-diffusion coefficient. The ADC was calculated according to

$$S_b = S_0 e^{-b[ADC]} \quad \text{(Eq. 2)}$$

where ADC is the apparent diffusion coefficient. We derived the D, D^*, K, f, and ADC for each GTV and computed the histogram mean, median, skewness, excess skewness, 10^{th} percentile, 90^{th} percentile, kurtosis, and excess kurtosis for each of the parameters.

2.3.3. Radiomics features from T2w and DWI-MRI

We normalized signal intensities for T2w scans in each GTV with zero mean and unit standard deviation (sd) prior to radiomics extraction to reduce intensity variations between MRI scans from different patients. For DWI-MRI, we extracted radiomics features from the b=0 and b=790 s/mm2 images, as these b-values were most prevalent in similar studies using DWI-MRI [21, 24, 32-43]. The open-source package Pyradiomics v3.0.1 was used to extract 108 radiomics features from each primary tumour, for each sequence (T2w, DWI-MRI$_{b=0}$ and DWI-MRI$_{b=790}$), in seven categories: shape, first order statistics, gray-level co-occurence matrix (GLCM) features, gray level dependence matrix (GLDM) features, gray-level run length matrix (GLRLM), gray level size zone matrix (GLSZM), Neighbouring Gray Tone Difference Matrix (NGTDM) [55].

2.4. Feature selection

We applied three feature selection methods to compare feature selection methods on pipeline performance and to choose the best performing feature set for each inner fold based on the inner training set.

2.4.1. Statistical testing for feature selection

We converted categorical variables with one-hot encoding for compatibility with all classifiers and performed the Chi-squared test to find differences between HPV-positive and HPV-negative groups. For the continuous variables, we performed the Shapiro-Wilk test for normality, and subsequently performed the independent two-sample t-test for normally distributed variables and the Mann-Whitney U test for not normally distributed Gaussian variables. For the dataset from the statistical feature selection, we only considered variables that were significantly different between the HPV-positive and HPV-negative groups. We considered p-values<0.05 as statically significant.

2.4.2. LASSO for feature selection

We applied Least Absolute Shrinkage and Selection Operator (LASSO) regression for feature selection, which is a linear model that aims to find a function best describing the data by minimizing the cost function, consisting of the sum of squared residuals and a L1-penalty term [56, 57]. LASSO aims to avoids overfitting by stimulating the reduction of the number of features (by setting the coefficients of superfluous features to zero) to lead to a smaller L1-penalty term. After selecting the weight of the penalization factor using GridSearchCV hyperparameter tuning based on the negative mean squared error, we trained and fitted LASSO on the training set and eliminated features with a coefficient of zero [58].

2.4.3. Manual feature selection

A manual feature selection was performed after discussion with a team including a medical physicist and radiation oncologists. The manual selection was based on features that were used in clinical practice or found in literature (Table 1) [21-24, 26, 32-43].

Table 1. Features from manual feature selection after discussion with a team including a medical physicist and radiation oncologists. ADC = apparent diffusion coefficient, D = diffusion coefficient, D* = pseudo-diffusion coefficient, f = perfusion fraction, K = kurtosis, T2w = T2-weighted.

Features in manual selection		
Clinical features	T-stage	
	N-stage	
	Smoking status	
	Alcohol status	
ADC parameter histogram	ADC	Mean
		Skewness
		Kurtosis
NG-IVIM parameter histogram	D	Mean
		Skewness
		Kurtosis
	f	Mean
	K	Mean
	D*	Mean
ADC radiomics	$b = 0\ s/mm^2$ intensity histogram	10^{th} percentile
		90^{th} percentile
T2 radiomics	Diagnostics	Number of voxels in tumour mask
	First order features	10^{th} percentile
		90^{th} percentile
		Kurtosis
		Skewness

2.5. Machine learning pipeline architecture

We divided the data *(n=59)* into the outer training set *(80%, n=48 or 49)* and test set *(20%, n=11 or 10)*, using ten-times stratified five-fold cross validation to reduce impact of each test set on the overall results (Fig. 1). The outer training set was then divided into an inner training set *(80%, n=39 or 40)* and validation set *(20%, n=9)* using stratified five-fold cross validation. A standardization scaler was fitted on the inner training set and applied to the inner training and validation sets. Afterwards, we applied the three methods of feature selection and subsequently trained and tuned the following seven classifiers: logistic regression, k-nearest neighbours, naive bayes, random forest, support vector machine, XGBoost, and LightGBM [59-67]. We performed hyperparameter tuning with GridSearch on the inner training set (Supplementary Data A).

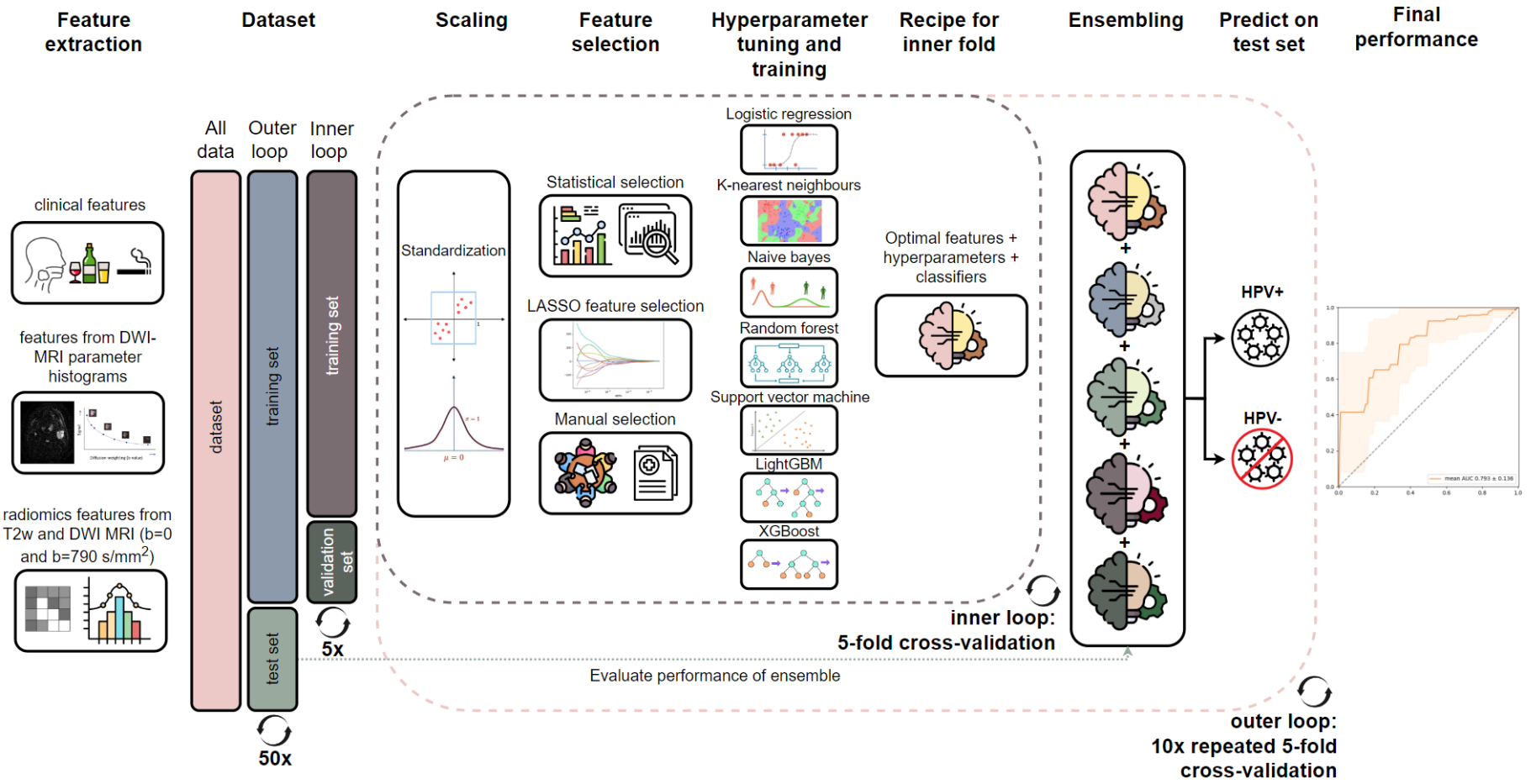


Figure 1. A schematic overview of the machine learning pipeline to predict human papillomavirus (HPV) status in oropharyngeal squamous carcinoma. Feature extraction is from three domains: clinical features, features from diffusion-weighted imaging magnetic resonance imaging (DWI-MRI) parameter histograms, and radiomics features. The dataset is divided in a test and training set (outer loop), after which the training set is split in a training and validation set (inner loop). Afterwards, scaling by standardization is performed based on the inner training set. Parallel feature selection is performed using three methods: statistical feature selection, Least Absolute Shrinkage and Selection Operator (LASSO) feature selection, and manual selection. Seven classifiers are tuned and fitted using the inner training set. The hyperparameters with the highest AUC on the training set is selected per classifier and set of features. Subsequently, these 21 classifier-feature-set combinations are applied to the validation set and the classifier with the highest AUC on the validation set is selected and saved for the fold. The five models from the inner loop are ensembled, which is fitted on the outer training set and tested on the test set. This is repeated for the ten times five-fold cross-validation on the total data set. The predictions are saved, and after all folds, the final performance is plotted for all fifty ensemble models.

We fitted each classifier on the inner training set and validated with the corresponding validation set. Subsequently, we compared the twenty-one resulting classifiers (three feature sets, seven classifiers) and selected the classifier with the highest area under the curve (AUC) on the validation set. After the five folds of inner cross validation, we ensembled the classifiers with soft voting, where the predicted probabilities of the five classifiers are averaged and converted to a class, which we trained and validated on the outer training set and test set, respectively [68]. For the fifty folds (ten times repeated five-fold cross validation) of the outer loop, we reported the mean and sd of the accuracy, AUC, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Python 3.8 was used for all ML-related programming [69].

2.6. Additional explorative models

Apart from the combined model, which integrated all three feature categories (clinical features, features from DWI-MRI parameters, and radiomics features), we constructed seven additional models. Without changing the pipeline architecture, three models used one of the three feature categories exclusively as input, and three used each combination of two feature categories as input. These six models offered insight in the added value of each of the feature categories. Lastly, we constructed one model using manually selected features exclusively to omit the effect of the feature selection. We performed the Kruskal-Wallis test followed by the Dunn's test as post-hoc test on AUC values to assess differences in performance over the eight models and considered $p<0.05$ as statistically significant [70, 71].

3. Results

3.1. Patient characteristics

Fifty-nine patients were included in our study, of whom 48 were male (81.36%) (Table 2). Among the 59 patients, 28 were HPV-positive and 31 were HPV-negative, and the overall mean age was 61.2 years. The most frequent alcohol status were occasional alcohol use and past alcohol abuse for HPV-positive and HPV-negative patients, respectively ($p<0.001$). There were no other differences between the two groups in baseline characteristics.

3.2. Recipe selection

Table 3 shows the ten most frequently selected features for the statistical feature selection method and LASSO feature selection method, which contains features from all feature categories. LASSO feature selection yielded between zero and twelve features per fold, statistical feature selection between four and eleven features per fold. Six out of ten most-selected features were in both lists: alcohol abuse in history, occasional alcohol use, 25% quartile of ADC histogram, first order skewness and GLCM cluster shade of the b=0 s/mm^2 DWI image, and the 25% quartile of the D histogram. The full list of features with selected frequency is in Supplementary Data B.

Out of the 250 selected models (= 5 inner folds × 5 outer folds × 10 repeats), the manually selected features were preferred 97 times (38.8%), the features from LASSO feature selection 89 times (35.6%), and the features from statistical feature selection 64 times (25.6%). Out of the 250 classifiers, naive bayes was selected 60 times (24.0%), followed by logistic regression (56 times, 22.4%), k-nearest neighbours (34 times, 13.6%), random forest (33 times, 13.2%), LightGBM (26 times, 10.4%), support vector machine (25 times, 10.0%), and XGBoost (16 times, 6.4%).

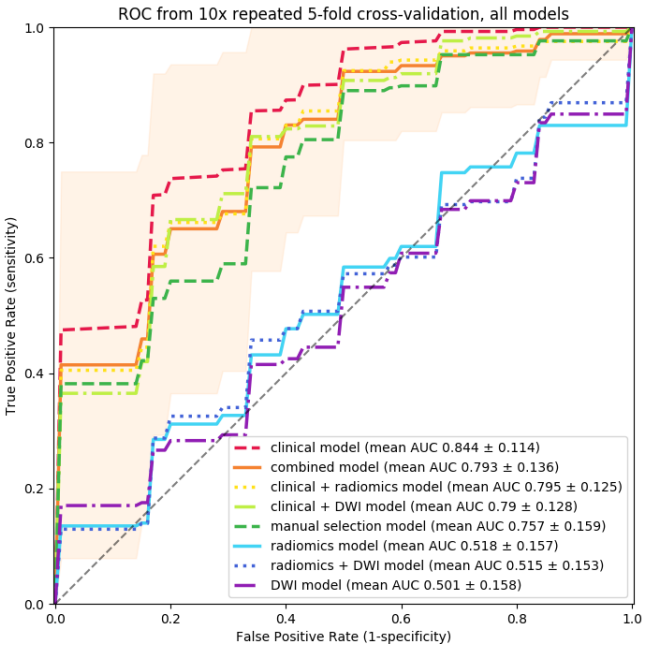


Figure 2. Mean receiver-operating characteristic (ROC) curve for predicting human papillomavirus (HPV) status of oropharyngeal squamous cell carcinoma for all eight models. The combined model is shown in orange with the mean area under the curve (AUC) and standard deviation of 0.793±0.136. The clinical model shows the highest area under the curve (AUC) score and the three models not containing any clinical features (radiomics model, radiomics + diffusion weighted imaging (DWI) model and DWI model) perform similarly just above 0.5. Models combining clinical features with MRI features perform similarly with mean AUCs between 0.75 and 0.795.

3.3. Pipeline performance

We report an AUC of 0.793±0.136 (Fig. 2) for the combined model over the fifty folds (ten times repeated five-fold cross validation), with a sensitivity of 0.793±0.195, a specificity of 0.631±0.169, PPV of 0.643±0.125, and NPV of 0.816 ±0.159. For the models only making use of clinical features, radiomics features, and features from DWI-MRI parameter histograms, we report AUCs of 0.844±0.114, 0.518±0.157, and 0.501±0.158, respectively. For the models combining sets of two feature categories, we report AUCs of 0.795±0.125 for the clinical + radiomics model, 0.790±0.128 for the clinical + DWI model, and 0.515±0.153 for the radiomics + DWI parameters model. For the model with manually selected features, we report a mean AUC of 0.757±0.159. Performance metrics for all models can be found in Table 3. Kruskall-Wallis test over all eight models showed a significant difference ($p<0.001$), where

Table 2. Baseline patient and tumour characteristics, comparison between human papillomavirus (HPV)-positive and HPV-negative patients. ADC = apparent diffusion coefficient, D = diffusion coefficient, D* = pseudo-diffusion coefficient, f = perfusion fraction, HPV = human papillomavirus, K = kurtosis, sd = standard deviation.

Patient and tumour characteristics	HPV-positive (n=28)	HPV-negative (n=31)	p-value
Age (years, mean ± sd)	61.679 ± 6.872	60.839 ± 9.785	0.849
Sex (male, %)	24 (86%)	24 (77%)	0.630
T-stage			0.091
- T1-T2 (n, %)	20 (71%)	17 (55%)	
- T3-T4 (n, %)	8 (29%)	14 (45%)	
N-stage			0.186
- N0 (n, %)	6 (21%)	10 (32%)	
- N+ (n, %)	22 (79%)	21 (68%)	
Alcohol			**<0.001**
- Never alcohol use (n, %)	5 (18%)	3 (10%)	
- Past alcohol abuse (n, %)	0 (0%)	13 (42%)	
- Occasionally (n, %)	21 (75%)	10 (32%)	
- Active alcohol abuse (n, %)	2 (7%)	5 (16%)	
Smoking			0.092
- Never (n, %)	6 (21%)	1 (3%)	
- Past (n, %)	12 (43%)	15 (48%)	
- Current (n, %)	10 (36%)	15 (48%)	
ADC (mm^2/s, mean ± sd)	$1.17 \cdot 10^{-3} \pm 2.29 \cdot 10^{-4}$	$1.29 \cdot 10^{-3} \pm 4.08 \cdot 10^{-4}$	0.291
D (mm^2/s, mean ± sd)	$1.24 \cdot 10^{-3} \pm 1.67 \cdot 10^{-4}$	$1.31 \cdot 10^{-3} \pm 2.59 \cdot 10^{-4}$	0.189
D* (mm^2/s, mean ± sd)	$2.57 \cdot 10^{-2} \pm 5.16 \cdot 10^{-3}$	$2.48 \cdot 10^{-2} \pm 1.03 \cdot 10^{-2}$	0.072
K (mean ± sd)	$8.99 \cdot 10^{-1} \pm 2.58 \cdot 10^{-1}$	$8.10 \cdot 10^{-1} \pm 2.47 \cdot 10^{-1}$	0.109
f (mean ± sd)	$2.00 \cdot 10^{-1} \pm 5.78 \cdot 10^{-2}$	$2.19 \cdot 10^{-1} \pm 6.79 \cdot 10^{-2}$	0.265

Table 3. Performance metrics of all prediction models for human papillomavirus (HPV) status. AUC = area under the curve, DWI = diffusion weighted imaging, NPV = negative predictive value, PPV = positive predictive value, sd = standard deviation.

Performance metrics (mean ± sd)	Combined model	Clinical model	Clinical + radiomics model	Clinical + DWI model	Manual selection model	Radiomics model	Radiomics + DWI model	DWI model
Accuracy	0.701 ± 0.109	0.745 ± 0.128	0.722 ± 0.107	0.706 ± 0.127	0.664 ± 0.143	0.517 ± 0.145	0.521 ± 0.142	0.496 ± 0.129
AUC	0.793 ± 0.136	0.844 ± 0.114	0.795 ± 0.125	0.790 ± 0.128	0.757 ± 0.159	0.518 ± 0.157	0.515 ± 0.153	0.501 ± 0.158
Sensitivity	0.793 ± 0.195	0.819 ± 0.212	0.813 ± 0.186	0.856 ± 0.192	0.796 ± 0.237	0.506 ± 0.182	0.515 ± 0.246	0.536 ± 0.245
Specificity	0.631 ± 0.169	0.690 ± 0.204	0.648 ± 0.153	0.590 ± 0.217	0.558 ± 0.209	0.532 ± 0.208	0.529 ± 0.219	0.462 ± 0.238
PPV	0.643 ± 0.125	0.699 ± 0.160	0.656 ± 0.130	0.642 ± 0.136	0.595 ± 0.169	0.488 ± 0.196	0.461 ± 0.194	0.445 ± 0.161
NPV	0.816 ± 0.159	0.857 ± 0.158	0.844 ± 0.144	0.855 ± 0.200	0.814 ± 0.194	0.562 ± 0.144	0.580 ± 0.197	0.552 ± 0.223

the models including clinical features outperformed the models excluding clinical features significantly ($p<0.001$). There were no significant differences within the models including clinical features and within the models excluding clinical features (Supplementary data C).

4. Discussion

In this study, we aimed to develop an ML pipeline using clinical, radiological and radiomics features from T2w and DWI-MRI to predict HPV status in OPSCC patients. Our findings show that it is possible to predict HPV status with reasonable accuracy using a pipeline with features from multiparametric MRI and clinical practice, with multiple feature selection methods and classifiers. To our knowledge, there has been no other study investigating the use of such an extensive set of feature selection methods and classifiers for this purpose.

Overall, the features that remained after feature selection were in line with findings from literature. Drinking habits were

predictive features in our dataset, which has been reported by several authors [21-24]. However, we did not find significant differences in smoking habits and T- and N-stage as is described in other studies [21-24].

MRI provides superior soft tissue contrast when compared to other modalities and recent studies found MRI-based radiomics features that were predictive for HPV status [23, 26, 72]. In our analysis, we found two T2w-radiomics in the most selected list of features, which were the Dependence Variance from GLDM and the energy. These describe the heterogeneity of a tumour and the sum squared values of all voxel intensities, respectively. This is in line with histopathological findings that HPV-negative tumours are more heterogeneous, which is caused by a higher prevalence of keratin pearls, intratumoural necrosis, haemorrhage and other factors, which in turn leads to different intensity histograms in HPV-positive and HPV-negative tumours [35]. However, few of the features derived from T2w-MRI were in the most selected features, indicating this variable category was of limited added value.

Regarding DWI-MRI, several authors report a lower ADC in HPV-positive tumours due to higher cellularity and thus, less water diffusion [21, 24, 32-43]. To our knowledge, there have been no other studies applying the NG-IVIM model to predict HPV status. Two authors investigated the IVIM model, of which the NG-IVIM model is an extension, and found that the diffusion coefficient D was significantly higher in HPV-negative compared to HPV-positive tumours [21, 38]. We did not find a correlation between ADC and HPV status, which may have been due to a high prevalence of smokers in the HPV-positive group in our dataset relative to other datasets, affecting the ADC [34, 39]. Though we did not find differences in ADC and D, we did find several of the features from the NG-IVIM parameter histograms and radiomics in the ten most selected features from the LASSO and statistical feature selection methods, indicating their added value.

As for the feature selection methods, we found the manual feature selection method was the most frequently selected option. The features that were selected for the manual dataset were chosen after discussion with experts and extracting findings from other studies, but without knowledge about our dataset beforehand. The manual selection was composed of thirty features whilst the LASSO and statistical feature selection always resulted in less than thirteen features. As the classifier with manual selection was trained on more features, this also increased the chance on overfitting. The statistical dataset was least selected of the three methods, possibly due to small inner training sets, leading to few features being significantly different in the two groups.

In our case, naive bayes and logistic regression were the most frequently selected classifiers. In a similar study by Marzi et al., naive bayes was also selected after comparison of several classifiers, possibly due to its properties of robustness to noisy data and the ability to work well on small sample sizes [22, 73]. Logistic regression is less inclined to overfitting with regularization techniques and was applied in multiple other studies with a similar aim [23, 25, 28, 29, 46, 74]. Therefore, both classifiers are suitable for our small, heterogeneous dataset.

Earlier studies found that models combining clinical and MRI features outperformed models using exclusively MRI or clinical features to predict HPV status in OPSCC [22, 23]. With our additional models, we were able to investigate the added value of each feature category for the model performance. We found that the models without clinical features performed around the randomness level, whilst adding other feature categories to the clinical model did not improve performance. Meanwhile, we did find that clinical as well as MRI parameters were chosen in our feature selection processes, which is in line with the findings from the other studies [22, 23].

The results of our study should be considered alongside several limitations. The most important limitation is that our study was performed on a small dataset. It is likely that due to the small training set size, each classifier overtrained, which made the performance on the test set lower. We tried to overcome the overtraining by using ensembling and to make our results more generalizable and transparent by repeating the performing ten-times repeated five-fold cross validation in the outer loop and reporting the sd. Other studies with a similar study sizes also report similar AUC values of between 0.744 to 0.77 [29, 46, 75]. Ravanelli et al. report an AUC of 0.944 with a study population of 59 patients, but do not describe the training process [39]. Studies with similar aims that use study populations of a hundred participants or more report AUC values between 0.80 and 0.871 [22, 23, 28, 38]. Because of the size of our dataset, outliers were of more influence than in larger datasets. Hence, applying our pipeline to a larger, multi-centre dataset, is likely to deliver more robust and generalizable results, whilst potentially improving the performance.

Moreover, the ground truth for our dataset consisted of the results of p16 IHC. Considering the limited sensitivity and specificity of 0.56-1.00 and 0.79-0.93 of p16 IHC when compared to HPV PCR, this is an inherent limitation of our dataset [16-18, 76]. In future studies, replicating other studies with an approach of confirming a positive p16 result with a HPV PCR test could lead to a more reliable ground truth [22, 24, 26, 35, 38, 40, 75, 77, 78].

Our model cannot outperform invasive tests for HPV status, such as p16 IHC that have an AUC, sensitivity and specificity of 0.96, 0.56-1.00, and 0.79-0.93, respectively [18, 76]. That being said, models such as ours could be complementary to the workflow. Once retrained on a larger dataset and fully integrated, the model has the potential to run simultaneously with the laboratory tests, and provide an independent HPV status prediction as an additional safeguard. For this purpose, it is an advantage that the pipeline makes use of data that will be gathered in the clinical workflow in any case, which increases the ease of implementation.

5. Conclusion

This study shows the potential of using multiparametric MRI combined with clinical variables for the prediction of HPV status in OPSCC patients. Although the pipeline should be retrained on a larger, multi-centre dataset, the input for the pipeline consists of data that will be generated in the clinical workflow in

any case. This contributes to the ease of implementation once a satisfactory model has been found. Future research in this field may contribute to working towards replacement of laboratory testing with more accessible, non-invasive solutions with data gathered in the clinical workflow.

6. Conflict of interest statement

The authors declare to have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

7. Declaration of generative AI in scientific writing

Adobe Firefly Text to Image was used for the generation of the basis of the front page.

8. References

1. Rivera C, Venegas B. Histological and molecular aspects of oral squamous cell carcinoma (Review). Oncol Lett. 2014;8(1):7-11.
2. Gillison ML, Koch WM, Capone RB, Spafford M, Westra WH, Wu L, et al. Evidence for a causal association between human papillomavirus and a subset of head and neck cancers. J Natl Cancer Inst. 2000;92(9):709-20.
3. de Perrot T, Lenoir V, Domingo Ayllón M, Dulguerov N, Pusztaszeri M, Becker M. Apparent Diffusion Coefficient Histograms of Human Papillomavirus-Positive and Human Papillomavirus-Negative Head and Neck Squamous Cell Carcinoma: Assessment of Tumor Heterogeneity and Comparison with Histopathology. AJNR Am J Neuroradiol. 2017;38(11): 2153-60.
4. Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tân PF, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. N Engl J Med. 2010;363(1): 24-35.
5. Marur S, Burtness B. Oropharyngeal squamous cell carcinoma treatment: current standards and future directions. Curr Opin Oncol. 2014;26(3):252-8.
6. Brierley JD, Gospodarowicz MK, Wittekind C. The TNM classification of malignant tumours. 8. Oxford: Wiley Blackwell. 2017.
7. Adelstein DJ, Li Y, Adams GL, Wagner H, Jr., Kish JA, Ensley JF, et al. An intergroup phase III comparison of standard radiation therapy and two schedules of concurrent chemoradiotherapy in patients with unresectable squamous cell head and neck cancer. J Clin Oncol. 2003;21(1):92-8.
8. Golusinski P, Corry J, Poorten VV, Simo R, Sjögren E, Mäkitie A, et al. De-escalation studies in HPV-positive oropharyngeal cancer: How should we proceed? Oral Oncol. 2021;123: 105620.
9. Windon MJ, D'Souza G, Rettig EM, Westra WH, van Zante A, Wang SJ, et al. Increasing prevalence of human papillomavirus-positive oropharyngeal cancers among older adults. Cancer. 2018;124(14):2993-9.
10. Bozec A, Culié D, Poissonnet G, Demard F, Dassonville O. Current Therapeutic Strategies in Patients with Oropharyngeal Squamous Cell Carcinoma: Impact of the Tumor HPV Status. Cancers (Basel). 2021;13(21).
11. Svajdova M, Dubinsky P, Kazda T, Jeremic B. Human Papillomavirus-Related Non-Metastatic Oropharyngeal Carcinoma: Current Local Treatment Options and Future Perspectives. Cancers (Basel). 2022;14(21).
12. Tolentino Ede S, Centurion BS, Ferreira LH, Souza AP, Damante JH, Rubira-Bullen IR. Oral adverse effects of head and neck radiotherapy: literature review and suggestion of a clinical oral care guideline for irradiated patients. J Appl Oral Sci. 2011;19(5):448-54.
13. Wittekindt C, Wagner S, Bushnak A, Prigge ES, von Knebel Doeberitz M, Würdemann N, et al. Increasing Incidence rates of Oropharyngeal Squamous Cell Carcinoma in Germany and Significance of Disease Burden Attributed to Human Papillomavirus. Cancer Prev Res (Phila). 2019;12(6):375-82.
14. Economopoulou P, Kotsantis I, Psyrri A. De-Escalating Strategies in HPV-Associated Head and Neck Squamous Cell Carcinoma. Viruses. 2021;13(9).
15. Mensour EA, Alam S, Mawani S, Bahig H, Lang P, Nichols A, et al. What is the future of treatment de-escalation for HPV-positive oropharyngeal cancer? A review of ongoing clinical trials. Front Oncol. 2022;12:1067321.
16. Walline HM, Komarck C, McHugh JB, Byrd SA, Spector ME, Hauff SJ, et al. High-risk human papillomavirus detection in oropharyngeal, nasopharyngeal, and oral cavity cancers: comparison of multiple methods. JAMA Otolaryngol Head Neck Surg. 2013;139(12):1320-7.
17. Venuti A, Paolini F. HPV detection methods in head and neck cancer. Head Neck Pathol. 2012;6 Suppl 1(Suppl 1):S63-74.
18. Wang H, Zhang Y, Bai W, Wang B, Wei J, Ji R, et al. Feasibility of Immunohistochemical p16 Staining in the Diagnosis of Human Papillomavirus Infection in Patients With Squamous Cell Carcinoma of the Head and Neck: A Systematic Review and Meta-Analysis. Front Oncol. 2020;10:524928.
19. Kim KY, Lewis JS, Jr., Chen Z. Current status of clinical testing for human papillomavirus in oropharyngeal squamous cell carcinoma. J Pathol Clin Res. 2018;4(4):213-26.
20. Duncan LD, Winkler M, Carlson ER, Heidel RE, Kang E, Webb D. p16 immunohistochemistry can be used to detect human papillomavirus in oral cavity squamous cell carcinoma. J Oral Maxillofac Surg. 2013;71(8):1367-75.
21. Vidiri A, Marzi S, Gangemi E, Benevolo M, Rollo F, Farneti A, et al. Intravoxel incoherent motion diffusion-weighted imaging for oropharyngeal squamous cell carcinoma: Correlation with human papillomavirus Status. Eur J Radiol. 2019;119: 108640.
22. Marzi S, Piludu F, Avanzolini I, Muneroni V, Sanguineti G, Farneti A, et al. Multifactorial Model Based on DWI-Radiomics to Determine HPV Status in Oropharyngeal Squamous Cell Carcinoma. Applied Sciences-Basel. 2022;12(14).
23. Bos P, van den Brekel MWM, Gouw ZAR, Al-Mamgani A, Waktola S, Aerts H, et al. Clinical variables and magnetic resonance imaging-based radiomics predict human papillomavirus status of oropharyngeal cancer. Head Neck. 2021;43(2):485-95.
24. Schouten CS, de Graaf P, Bloemena E, Witte BI, Braakhuis BJ, Brakenhoff RH, et al. Quantitative diffusion-weighted MRI parameters and human papillomavirus status in oropharyngeal squamous cell carcinoma. AJNR Am J Neuroradiol. 2015; 36(4):763-7.
25. Chan MW, Yu E, Bartlett E, O'Sullivan B, Su J, Waldron J, et al. Morphologic and topographic radiologic features of human papillomavirus-related and -unrelated oropharyngeal carcinoma. Head Neck.

2017; 39(8):1524-34.
26. Giannitto C, Marvaso G, Botta F, Raimondi S, Alterio D, Ciardo D, et al. Association of quantitative MRI-based radiomic features with prognostic factors and recurrence rate in oropharyngeal squamous cell carcinoma. Neoplasma. 2020; 67(6):1437-46.
27. Huang YH, Yeh CH, Cheng NM, Lin CY, Wang HM, Ko SF, et al. Cystic nodal metastasis in patients with oropharyngeal squamous cell carcinoma receiving chemoradiotherapy: Relationship with human papillomavirus status and failure patterns. Plos One. 2017;12(7): e0180779.
28. Park YM, Lim JY, Koh YW, Kim SH, Choi EC. Machine learning and magnetic resonance imaging radiomics for predicting human papilloma virus status and prognostic factors in oropharyngeal squamous cell carcinoma. Head Neck. 2022; 44(4):897-903.
29. Sohn B, Choi YS, Ahn SS, Kim H, Han K, Lee SK, et al. Machine Learning Based Radiomic HPV Phenotyping of Oropharyngeal SCC: A Feasibility Study Using MRI. Laryngoscope. 2021;131(3):E851-E6.
30. Widmann G, Henninger B, Kremser C, Jaschke W. MRI Sequences in Head & Neck Radiology - State of the Art MRI-Sequenzen in der Kopf-Hals-Radiologie - State of the Art. Rofo. 2017;189(5):413-22.
31. Paczona VR, Capala ME, Deák-Karancsi B, Borzási E, Együd Z, Végváry Z, et al. Magnetic Resonance Imaging 2013; Based Delineation of Organs at Risk in the Head and Neck Region. Advances in Radiation Oncology. 2023;8(2).
32. Driessen JP, van Bemmel AJM, van Kempen PMW, Janssen LM, Terhaard CHJ, Pameijer FA, et al. Correlation of human papillomavirus status with apparent diffusion coefficient of diffusion-weighted MRI in head and neck squamous cell carcinomas. Head and Neck-Journal for the Sciences and Specialties of the Head and Neck. 2016; 38:E613-E8.
33. Cao Y, Aryal M, Li P, Lee C, Schipper M, Hawkins PG, et al. Predictive Values of MRI and PET Derived Quantitative Parameters for Patterns of Failure in Both p16+and p16-High Risk Head and Neck Cancer. Front Oncol. 2019;9.
34. Chan MW, Higgins K, Enepekides D, Poon I, Symons SP, Moineddin R, et al. Radiologic Differences between Human Papillomavirus-Related and Human Papillomavirus-Unrelated Oropharyngeal Carcinoma on Diffusion-Weighted Imaging. ORL J Otorhinolaryngol Relat Spec. 2016;78(6):344-52.
35. De Perrot T, Lenoir V, Ayllón MD, Dulguerov N, Pusztaszeri M, Becker M. Apparent diffusion coefficient histograms of human papillomavirus-positive and human papillomavirus-negative head and neck squamous cell carcinoma: Assessment of tumor heterogeneity and comparison with histopathology. Am J Neuroradiol. 2017;38(11):2153-60.
36. Freihat O, Toth Z, Pinter T, Kedves A, Sipos D, Cselik Z, et al. Pre-treatment PET/MRI based FDG and DWI imaging parameters for predicting HPV status and tumor response to chemoradiotherapy in primary oropharyngeal squamous cell carcinoma (OPSCC). Oral Oncol. 2021;116:105239.
37. Nakahira M, Saito N, Yamaguchi H, Kuba K, Sugasawa M. Use of quantitative diffusion-weighted magnetic resonance imaging to predict human papilloma virus status in patients with oropharyngeal squamous cell carcinoma. Eur Arch Otorhinolaryngol. 2014;271(5):1219-25.
38. Piludu F, Marzi S, Gangemi E, Farneti A, Marucci L, Venuti A, et al. Multiparametric MRI Evaluation of Oropharyngeal Squamous Cell Carcinoma. A Mono-Institutional Study. 2021.
39. Ravanelli M, Grammatica A, Tononcelli E, Morello R, Leali M, Battocchio S, et al. Correlation between Human Papillomavirus Status and Quantitative MR Imaging Parameters including Diffusion-Weighted Imaging and Texture Features in Oropharyngeal Carcinoma. AJNR Am J Neuroradiol. 2018; 39(10):1878-83.
40. Lenoir V, Delattre BMA, M'Ra DY, De Vito C, de Perrot T, Becker M. Diffusion-Weighted Imaging to Assess HPV-Positive versus HPV-Negative Oropharyngeal Squamous Cell Carcinoma: The Importance of b-Values. AJNR Am J Neuroradiol. 2022;43(6):905-12.
41. Connor S, Anjari M, Burd C, Guha A, Lei M, Guerrero-Urbano T, et al. The impact of Human Papilloma Virus status on the prediction of head and neck cancer chemoradiotherapy outcomes using the pre-treatment apparent diffusion coefficient. Br J Radiol. 2022;95(1130): 20210333.
42. Peltenburg B, Driessen JP, Vasmel JE, Pameijer FA, Janssen LM, Terhaard CHJ, et al. Pretreatment ADC is not a prognostic factor for local recurrences in head and neck squamous cell carcinoma when clinical T-stage is known. Eur Radiol. 2020;30(2):1228-31.
43. Han M, Lee SJ, Lee D, Kim SY, Choi JW. Correlation of human papilloma virus status with quantitative perfusion/diffusion/ metabolic imaging parameters in the oral cavity and oropharyngeal squamous cell carcinoma: comparison of primary tumour sites and metastatic lymph nodes. Clin Radiol. 2018;73(8): 757.e21-.e27.
44. Lu Y, Jansen JF, Mazaheri Y, Stambuk HE, Koutcher JA, Shukla-Dave A. Extension of the intravoxel incoherent motion model to non-gaussian diffusion in head and neck cancer. J Magn Reson Imaging. 2012;36(5):1088-96.
45. van Dijk LV, Fuller CD. Artificial Intelligence and Radiomics in Head and Neck Cancer Care: Opportunities, Mechanics, and Challenges. Am Soc Clin Oncol Educ Book. 2021; 41:1-11.
46. Suh CH, Lee KH, Choi YJ, Chung SR, Baek JH, Lee JH, et al. Oropharyngeal squamous cell carcinoma: radiomic machine-learning classifiers from multiparametric MR images for determination of HPV infection status. Sci rep. 2020;10(1): 17525.
47. Verduijn GM, Capala ME, Sijtsema ND, Lauwers I, Hernandez Tamames JA, Heemsbergen WD, et al. The COMPLETE trial: HolistiC early respOnse assessMent for oroPharyngeaL cancEr paTiEnts; Protocol for an observational study. BMJ Open. 2022;12(5).
48. Lassen P, Overgaard J. Scoring and classification of oropharyngeal carcinoma based on HPV-related p16-expression. Radiother Oncol. 2012;105(2):269-70.
49. Hashmi AA, Younus N, Naz S, Irfan M, Hussain Z, Shaikh ST, et al. p16 Immunohistochemical Expression in Head and Neck Squamous Cell Carcinoma: Association With Prognostic Parameters. Cureus. 2020;12(6):e8601.
50. Sijtsema ND, Petit SF, Poot DHJ, Verduijn GM, van der Lugt A, Hoogeman MS, et al. An optimal acquisition and post-processing pipeline for hybrid IVIM-DKI in head and neck. Magn Reson Med. 2021;85(2):777-89.
51. Mattes D, Haynor DR, Vesselle H, Lewellyn TK, Eubank W, editors. Nonrigid multimodality image registration. Proc SPIE; 2001.
52. CBS. Alcoholgebruik: CBS; 2022 [Available from: https://www.cbs.nl/nl-nl/nieuws/2022/10/overgewicht-roken-en-alcoholgebruik-nauwelijks-gedaald-sinds-2018/alcoholgebruik.
53. Sijtsema ND, Lauwers I, Verduijn GM, Hoogeman MS, Poot DHJ, Hermandez Tamames JA, et al. Non Gaussian Intra Voxel Incoher-

ent Motion Diffusion Weighted Imaging, HPV status and reponse in oropharyngeal carcinoma. 2023.
54. Inc. TM. MATLAB version: 9.11.0 (R2021b). Natick, Massachusetts, United States: The MathWorks Inc.
55. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017;77(21): e104-e7.
56. Ranstam J, Cook JA. LASSO regression. British Journal of Surgery. 2018;105(10):1348-.
57. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996;58(1):267-88.
58. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2625-830.
59. 1.3.0 s-l. sklearn.linear_model.LogisticRegression [Available from: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.]
60. 1.3.0 s-l. sklearn.ensemble.RandomForestClassifier [Available from: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.]
61. 1.3.0 s-l. sklearn.neighbors.KNeighborsClassifier [Available from: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html.]
62. XGBoost parameters: dmlc XGBoost; 2023 [Available from: https://xgboost.readthedocs.io/en/stable/parameter.html.
63. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; New York, NY, USA2016.
64. 1.3.0 s-l. 1.9 Naive Bayes [Available from: https://scikit-learn.org/stable/modules/naive_bayes.html.
65. 1.3.0 s-l. 1.4 Support Vector Machines [Available from: https://scikit-learn.org/stable/modules/svm.html.
66. Parameters: LightGBM; 2023 [Available from: https://lightgbm.readthedocs.io/en/latest/Parameters.html.
67. Ke G, Meng Q, Finely T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 30 (NIP 2017)2017.
68. Wang H, Yang Y, Wang H, Chen D, editors. Soft-Voting Clustering Ensemble. Multiple Classifier Systems; 2013; Berlin, Heidelberg: Springer Berlin Heidelberg.
69. van Rossum G. Python reference manual. CWI; 1995.
70. McKight PE, Najab J. Kruskal-Wallis Test. The Corsini Encyclopedia of Psychology2010. p. 1-.
71. Dinno A. Nonparametric Pairwise Multiple Comparisons in Independent Groups using Dunn's Test. The Stata Journal. 2015;15(1):292-300.
72. Vishwanath V, Jafarieh S, Rembielak A. The role of imaging in head and neck cancer: An overview of different imaging modalities in primary diagnosis and staging of the disease. J Contemp Brachytherapy. 2020;12(5):512-8.
73. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. Radiographics. 2017;37(2):505-15.
74. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of Clinical Epidemiology. 1996;49(11):1225-31.
75. Ahn Y, Choi YJ, Sung YS, Pfeuffer J, Suh CH, Chung SR, et al. Histogram analysis of arterial spin labeling perfusion data to determine the human papillomavirus status of oropharyngeal squamous cell carcinomas. Neuroradiology. 2021;63(8):1345-52.
76. Robinson M, Schache A, Sloan P, Thavaraj S. HPV specific testing: a requirement for oropharyngeal squamous cell carcinoma patients. Head Neck Pathol. 2012;6 Suppl 1(Suppl 1):S83-90.
77. Vidiri A, Gangemi E, Ruberto E, Pasqualoni R, Sciuto R, Sanguineti G, et al. Correlation between histogram-based DCE-MRI parameters and F-18-FDG PET values in oropharyngeal squamous cell carcinoma: Evaluation in primary tumors and metastatic nodes. Plos One. 2020;15(3).
78. Driessen JP, van Kempen PM, van der Heijden GJ, Philippens ME, Pameijer FA, Stegeman I, et al. Diffusion-weighted imaging in head and neck squamous cell carcinomas: a systematic review. Head Neck. 2015;37(3):440-8.

Supplementary data

Supplementary data A: tuned hyperparameters for each classifier

Classifier	Tuned hyperparameters	
Logistic regression [1, 2]	Solver:	lbfgs, saga, liblinear
	Penalty:	None, L1, L2, ElasticNet
	C (inverse of regularization strength):	100, 10, 1
Random forest [2, 3]	Max_features:	sqrt, log2, 1, 5, 10, 20
	N_estimators:	10, 100, 1000
Naïve bayes [2, 4]	None	
Support vector machine [2, 5]	Kernel:	linear, poly, rbf, sigmoid
	C (inverse of regularization strength):	100, 10, 1, 0.1, 0.001
K-nearest neighbours [2, 6]	Weights:	uniform, distance
	N_neighbours:	3, 5, 8
LightGBM [7, 8]	Boosting_type:	gbdt, rf, dart
	Learning_rate:	0.1, 0.01, 0.001
	Max_depth:	3, 6, 8
	Min_child_samples:	1, 2

	Bagging_freq:	1, 5
	Bagging_fraction:	0.2, 0.6
XGBoost [9, 10]	Eta:	0.05, 0.1, 0.2, 0.3
	Min_child_weight:	1, 2, 3, 4
	Max_depth:	3, 6, 8
	Alpha:	0, 0.2, 0.5, 0.7, 1
	Lambda:	1, 5, 10

References:

1. 1.3.0 s-l. sklearn.linear_model.LogisticRegression [Available from: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.
2. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2625-830.
3. 1.3.0 s-l. sklearn.ensemble.RandomForestClassifier [Available from: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.
4. 1.3.0 s-l. 1.9 Naive Bayes [Available from: https://scikit-learn.org/stable/modules/naive_bayes.html.
5. 1.3.0 s-l. 1.4 Support Vector Machines [Available from: https://scikit-learn.org/stable/modules/svm.html.
6. 1.3.0 s-l. sklearn.neighbors.KNeighborsClassifier [Available from: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html.
7. Ke G, Meng Q, Finely T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 30 (NIP 2017)2017.
8. Parameters: LightGBM; 2023 [Available from: https://lightgbm.readthedocs.io/en/latest/Parameters.html.
9. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; New York, NY, USA2016.
10. XGBoost parameters: dmlc XGBoost; 2023 [Available from: https://xgboost.readthedocs.io/en/stable/parameter.html.

Supplementary data B: full list of features and number of times selected in feature selection

Features from statistical feature selection

Feature	Count (out of 250)
clinical_alcohol_active	168
ADC_25% quartile	160
clinical_alcohol_in history	156
clinical_alcohol_never	144
clinical_alcohol_occasionally	131
Dstar_sd	109
DWI_b0_original_firstorder_Skewness	105
ADC_median	86
DWI_b0_original_glcm_ClusterShade	85
D_25% quartile	85
T2_original_glszm_SizeZoneNonUniformity	67
Dstar_25% quartile	62
DWI_b0_original_shape_Flatness	52
K_75% quartile	45
DWI_b790_original_shape_Flatness	42
K_sd	41
Dstar_mean	37
Dstar_min	37
T2_original_gldm_DependenceVariance	31
D_median	30
Dstar_75% quartile	29
Dstar_excess kurtosis	24
K_mean	24
Dstar_kurtosis	24
clinical_smoking_current	23
clinical_T-stage	21
f_kurtosis	21
f_excess kurtosis	21
clinical_smoking_never	21
DWI_b790_original_glszm_SmallAreaLowGrayLevelEmphasis	20
clinical_smoking_past	19
T2_original_shape_Flatness	18
ADC_75% quartile	16
D_mean	16
T2_original_shape_Elongation	14
T2_original_glszm_SizeZoneNonUniformityNormalized	14
DWI_b790_original_glcm_Imc1	14
T2_original_glszm_SmallAreaLowGrayLevelEmphasis	13

K_median	13
f_25% quartile	12
K_kurtosis	11
K_excess kurtosis	11
f_median	9
D_excess kurtosis	9
D_kurtosis	9
DWI_b790_original_glcm_MCC	8
DWI_b790_original_glcm_Imc2	7
DWI_b790_original_shape_Elongation	7
DWI_b790_original_firstorder_Minimum	7
Dstar_adj skewness	7
f_skewness	7
Dstar_skewness	7
DWI_b0_original_gldm_LargeDependenceLowGrayLevelEmphasis	6
f_mean	6
K_25% quartile	6
DWI_b790_original_ngtdm_Contrast	6
f_adj skewness	6
DWI_b0_original_shape_Elongation	6
D_75% quartile	6
DWI_b790_original_gldm_DependenceEntropy	5
T2_original_gldm_LargeDependenceHighGrayLevelEmphasis	5
T2_original_glszm_SmallAreaEmphasis	5
DWI_b0_original_glszm_LargeAreaLowGrayLevelEmphasis	5
DWI_b790_original_glcm_Correlation	5
K_adj skewness	4
D_max	4
DWI_b790_original_shape_Maximum2DDiameterRow	4
DWI_b0_original_shape_Maximum2DDiameterRow	4
K_max	4
D_adj skewness	4
ADC_mean	4
K_skewness	4
D_skewness	4
DWI_b790_original_glcm_ClusterShade	3
DWI_b790_original_gldm_DependenceVariance	3
DWI_b790_original_firstorder_90Percentile	3
DWI_b0_original_glrlm_GrayLevelNonUniformityNormalized	3
ADC_sd	3
DWI_b0_original_ngtdm_Strength	3
T2_original_glcm_ClusterShade	3
DWI_b790_original_firstorder_Variance	3
DWI_b0_original_firstorder_Uniformity	3
T2_original_glszm_LowGrayLevelZoneEmphasis	3

ADC_max	3
DWI_b0_original_firstorder_RobustMeanAbsoluteDeviation	3
T2_original_glrlm_RunEntropy	3
clinical_N-stage_0	3
DWI_b790_original_firstorder_MeanAbsoluteDeviation	3
DWI_b0_original_firstorder_InterquartileRange	3
DWI_b0_original_shape_Maximum3DDiameter	2
DWI_b790_original_glszm_SmallAreaEmphasis	2
clinical_N-stage_1	2
T2_original_firstorder_Skewness	2
f_75% quartile	2
f_min	2
DWI_b790_original_gldm_SmallDependenceLowGrayLevelEmphasis	2
DWI_b0_original_firstorder_Median	2
DWI_b0_original_shape_Sphericity	2
DWI_b790_original_firstorder_RootMeanSquared	2
DWI_b0_original_firstorder_Entropy	2
clinical_N-stage_2a	2
DWI_b0_original_glcm_Autocorrelation	2
DWI_b0_original_glcm_Imc2	2
DWI_b790_original_shape_Maximum3DDiameter	2
DWI_b0_original_firstorder_Kurtosis	2
DWI_b790_original_glszm_LowGrayLevelZoneEmphasis	2
ADC_adj skewness	2
DWI_b790_original_firstorder_Kurtosis	2
clinical_N-stage_2b	2
DWI_b790_original_glcm_Autocorrelation	2
DWI_b0_original_glcm_SumEntropy	1
DWI_b790_original_glcm_SumSquares	1
DWI_b790_original_ngtdm_Strength	1
DWI_b0_original_glcm_Imc1	1
DWI_b790_original_glcm_InverseVariance	1
DWI_b0_original_glcm_ClusterTendency	1
DWI_b0_original_gldm_DependenceVariance	1
DWI_b0_original_glszm_SizeZoneNonUniformityNormalized	1
DWI_b790_original_firstorder_RobustMeanAbsoluteDeviation	1
DWI_b0_original_glrlm_HighGrayLevelRunEmphasis	1
T2_original_shape_Maximum2DDiameterRow	1
DWI_b790_original_glszm_ZonePercentage	1
DWI_b790_original_glcm_ClusterTendency	1
DWI_b0_original_ngtdm_Complexity	1
DWI_b790_original_glrlm_GrayLevelVariance	1
DWI_b0_original_glcm_JointEnergy	1
DWI_b0_original_glcm_SumSquares	1
DWI_b0_original_firstorder_MeanAbsoluteDeviation	1

DWI_b0_original_firstorder_90Percentile	1
clinical_N-stage_2c	1
DWI_b790_original_gldm_LargeDependenceLowGrayLevelEmphasis	1
DWI_b790_original_firstorder_Uniformity	1
DWI_b0_original_glcm_JointEntropy	1
DWI_b0_original_glcm_InverseVariance	1
DWI_b790_original_gldm_LowGrayLevelEmphasis	1
DWI_b790_original_gldm_SmallDependenceEmphasis	1
DWI_b790_original_gldm_GrayLevelVariance	1
DWI_b0_original_gldm_LargeDependenceEmphasis	1
DWI_b0_original_glrlm_RunEntropy	1
ADC_skewness	1
DWI_b790_original_glrlm_HighGrayLevelRunEmphasis	1
T2_original_gldm_DependenceNonUniformityNormalized	1
T2_original_ngtdm_Coarseness	1
DWI_b0_original_glszm_LargeAreaEmphasis	1
T2_original_ngtdm_Strength	1
DWI_b790_original_firstorder_InterquartileRange	1
K_min	1
T2_original_firstorder_Kurtosis	1
DWI_b790_original_firstorder_Maximum	1
DWI_b790_original_glrlm_LongRunHighGrayLevelEmphasis	1
DWI_b0_original_firstorder_Minimum	1
DWI_b790_original_shape_LeastAxisLength	1
DWI_b0_original_glrlm_RunVariance	1
T2_original_firstorder_Minimum	1
DWI_b0_original_shape_MajorAxisLength	1
DWI_b0_original_ngtdm_Contrast	1
DWI_b0_original_glcm_Idn	1
DWI_b790_original_gldm_LargeDependenceHighGrayLevelEmphasis	1
DWI_b0_original_glszm_SmallAreaHighGrayLevelEmphasis	1
DWI_b0_original_glcm_Idmn	1
T2_original_firstorder_Energy	1
DWI_b790_original_shape_MajorAxisLength	1
DWI_b0_original_gldm_HighGrayLevelEmphasis	1
DWI_b0_original_shape_SurfaceVolumeRatio	1
T2_original_shape_SurfaceVolumeRatio	1
DWI_b0_original_gldm_SmallDependenceHighGrayLevelEmphasis	1
T2_original_shape_MajorAxisLength	1
DWI_b0_original_glrlm_ShortRunEmphasis	1
DWI_b790_original_firstorder_Median	1
DWI_b0_original_glcm_ClusterProminence	1
DWI_b0_original_glcm_Idm	1
DWI_b0_original_glszm_LargeAreaHighGrayLevelEmphasis	1
DWI_b790_original_glcm_ClusterProminence	1

DWI_b790_original_glrlm_RunEntropy	1
DWI_b0_original_glcm_Correlation	1
T2_original_firstorder_TotalEnergy	1
DWI_b0_original_glrlm_RunLengthNonUniformityNormalized	1
DWI_b790_original_firstorder_Mean	1
DWI_b0_original_shape_LeastAxisLength	1
DWI_b0_original_glcm_MCC	1
DWI_b0_original_glszm_ZoneVariance	1
DWI_b0_original_firstorder_Variance	1

Features from LASSO feature selection

Feature	Count (out of 250)
clinical_alcohol_in history	228
clinical_alcohol_occasionally	156
DWI_b0_original_firstorder_Skewness	111
T2_original_gldm_DependenceVariance	103
clinical_N-stage_3	103
DWI_b0_original_glcm_ClusterShade	100
T2_original_firstorder_Energy	97
DWI_b790_original_glszm_SmallAreaLowGrayLevelEmphasis	97
ADC_25% quartile	96
Dstar_25% quartile	92
clinical_smoking_never	90
T2_original_firstorder_TotalEnergy	74
clinical_N-stage_1	68
clinical_T-stage	68
DWI_b0_original_shape_Flatness	41
f_25% quartile	33
T2_original_glszm_SizeZoneNonUniformity	25
DWI_b790_original_ngtdm_Busyness	24
T2_original_shape_Elongation	24
DWI_b790_original_glcm_MCC	21
Dstar_sd	20
DWI_b790_original_shape_Elongation	19
Dstar_max	18
DWI_b790_original_shape_Flatness	18
T2_original_ngtdm_Strength	16
K_25% quartile	15
D_25% quartile	15
DWI_b790_original_ngtdm_Contrast	14
clinical_N-stage_0	12
f_kurtosis	11
T2_original_ngtdm_Complexity	10
DWI_b790_original_glcm_Imc1	10
DWI_b790_original_gldm_DependenceEntropy	10
DWI_b790_original_glcm_Correlation	10
D_kurtosis	10
Dstar_75% quartile	10
DWI_b0_original_firstorder_Kurtosis	9
Dstar_min	9
f_excess kurtosis	9
clinical_smoking_past	9
f_skewness	9

Dstar_kurtosis	9
D_excess kurtosis	8
ADC_median	8
DWI_b790_original_glcm_Imc2	8
ADC_kurtosis	8
Dstar_excess kurtosis	7
clinical_alcohol_active	7
ADC_excess kurtosis	7
T2_original_glszm_SmallAreaLowGrayLevelEmphasis	6
DWI_b790_original_firstorder_Minimum	6
clinical_smoking_current	5
f_adj skewness	5
T2_original_glcm_Idmn	5
DWI_b790_original_glszm_LowGrayLevelZoneEmphasis	5
T2_original_glcm_ClusterProminence	4
clinical_alcohol_never	4
clinical_N-stage_2c	4
K_kurtosis	4
K_75% quartile	4
DWI_b0_original_glszm_LargeAreaHighGrayLevelEmphasis	4
T2_original_glcm_Correlation	4
K_adj skewness	4
T2_original_gldm_LargeDependenceHighGrayLevelEmphasis	4
DWI_b790_original_gldm_SmallDependenceLowGrayLevelEmphasis	4
clinical_N-stage_2a	4
T2_original_firstorder_Minimum	3
DWI_b790_original_glszm_LargeAreaLowGrayLevelEmphasis	3
DWI_b0_original_glcm_Correlation	3
ADC_sd	3
D_median	3
DWI_b790_original_glcm_ClusterShade	3
D_mean	3
DWI_b0_original_glszm_LowGrayLevelZoneEmphasis	2
T2_original_firstorder_Maximum	2
DWI_b0_original_glcm_Idmn	2
DWI_b790_original_glszm_ZoneVariance	2
DWI_b790_original_gldm_DependenceVariance	2
DWI_b790_original_glszm_LargeAreaHighGrayLevelEmphasis	2
DWI_b0_original_glcm_MCC	2
DWI_b790_original_glszm_SmallAreaEmphasis	2
T2_original_glszm_LowGrayLevelZoneEmphasis	2
ADC_mean	2
T2_original_glrlm_RunEntropy	2
f_mean	2
DWI_b790_original_firstorder_Kurtosis	2

D_min	2
DWI_b790_original_glrlm_GrayLevelNonUniformityNormalized	2
DWI_b790_original_glszm_GrayLevelNonUniformityNormalized	2
D_skewness	2
K_sd	2
K_excess kurtosis	2
K_median	2
f_median	2
T2_original_gldm_DependenceNonUniformityNormalized	2
DWI_b0_original_gldm_DependenceVariance	2
ADC_max	2
T2_original_ngtdm_Contrast	1
T2_original_shape_SurfaceVolumeRatio	1
DWI_b790_original_gldm_LargeDependenceLowGrayLevelEmphasis	1
T2_original_glszm_LargeAreaLowGrayLevelEmphasis	1
T2_original_glszm_ZoneEntropy	1
T2_original_glrlm_LongRunLowGrayLevelEmphasis	1
DWI_b0_original_ngtdm_Busyness	1
T2_original_gldm_SmallDependenceLowGrayLevelEmphasis	1
T2_original_glszm_SizeZoneNonUniformityNormalized	1
T2_original_firstorder_Variance	1
T2_original_glszm_SmallAreaEmphasis	1
f_75% quartile	1
DWI_b0_original_ngtdm_Contrast	1
T2_original_glcm_Imc1	1
T2_original_firstorder_Range	1
T2_original_ngtdm_Coarseness	1
K_max	1
clinical_N-stage_2b	1
DWI_b0_original_glszm_SmallAreaLowGrayLevelEmphasis	1
D_max	1
f_min	1
D_adj skewness	1
K_mean	1
ADC_skewness	1
DWI_b0_original_glcm_Idn	1
ADC_adj skewness	1
f_max	1
DWI_b790_original_firstorder_Uniformity	1
DWI_b790_original_ngtdm_Strength	1
DWI_b790_original_glcm_DifferenceAverage	1
DWI_b0_original_glszm_SizeZoneNonUniformityNormalized	1
T2_original_glrlm_LongRunHighGrayLevelEmphasis	1
T2_original_firstorder_Kurtosis	1

Supplementary data C: results from Dunn's test, p-values between each pair of models

	clinical model	*combined model*	*clinical + radiomics model*	*clinical + DWI model*	*manual selection model*	*radiomics model*	*radiomics + DWI model*	*DWI model*
clinical model	1	1	1	0.917	0.129	<0.001	<0.001	<0.001
combined model	1	1	1	1	1	<0.001	<0.001	<0.001
clinical + radiomics model	1	1	1	1	1	<0.001	<0.001	<0.001
clinical + DWI model	0.917	1	1	1	1	<0.001	<0.001	<0.001
manual selection model	0.129	1	1	1	1	<0.001	<0.001	<0.001
radiomics model	<0.001	<0.001	<0.001	<0.001	<0.001	1	1	1
radiomics + DWI model	<0.001	<0.001	<0.001	<0.001	<0.001	1	1	1
DWI model	<0.001	<0.001	<0.001	<0.001	<0.001	1	1	1