UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Centre de Formació Interdisciplinària Superior

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

UNIVERSITY OF TORONTO

Bachelor's degree thesis

Use of radiomic data to improve imputation of HPV (p16) status in oropharyngeal cancer

Aleix Lascorz Guiu

Supervised by:

Benjamin Haibe-Kains (UofT)

In partial fulfillment of the requirements for the

Bachelor's degree in Mathematics

Bachelor's degree in Informatics Engineering

October 2019

Contents

List of Figures

Abstract

The incidence of oropharyngeal cancer has been steadily increasing during the past decades. This increase is linked to human papillomavirus, one of the most common sexually transmitted diseases in Canada and worldwide. In this study, we take a novel approach to HPV status imputation by building machine learning models that utilize not only clinical data but also imaging features, aiming to show a significant improvement over classical models. The increase of performance between state of the art clinical models and our models will be assessed through the use of the RADCURE dataset from the Princess Margaret Cancer Centre, one of the biggest datasets in cancer research containing data of over one thousand oropharyngeal cancer patients.

by Aleix Lascorz Guiu

Keywords: machine learning, oropharyngeal cancer, imputation, radiomics, HPV, medicine
MSC2000 code: 92B20

Chapter 1

Introduction

Oropharyngeal cancer (OPC) is a type of head and neck cancer which affects mainly the oropharynx, but also possibly the base of the tongue, the throat, the soft palate and the side and back wall of the throat. Symptoms include the apparition of a painless swelling or lump in the neck, a sore throat or tongue and earache. Given that all of this may also be caused by other neck illnesses like the flu or the common cold, early diagnosis is uncommon. While its incidence is increasing, its cure rates are also improving. Treatment methods include chemotherapy and targeted therapy, radiotherapy and surgery to remove the cancer. Historically tobacco and alcohol were the major risk factors, but a recent shift in trends has resulted in human papillomavirus causing most of these tumours [4, 16].

Currently around 70% of OPC cancers are human papillomavirus (HPV) positive. HPV is one of the most common sexually transmitted infections in Canada and worldwide, with many strands which can cause symptoms ranging from genital warts to cancer. Specifically, HPV type 16 is the one most strongly linked with oropharyngeal cancer. While being the most prevalent, HPV-positive oropharyngeal cancer also has better survival rate than HPV-negative oropharyngeal cancer due to its greater response to all treatments [10]. Because of that, early tipification of HPV status in oropharyngeal cancer patients is extremely important, allowing for less intense therapy which significantly reduces treatment-related toxicity [9] [12].

Due to both the link between HPV and oropharyngeal cancer and the greater response of HPV-positive OPC to treatment, using the p16 test to determine HPV status of oropharyngeal cancer patients is now the regular practise in all hospitals in Canada. However, that was not the case a decade ago, when HPV testing was only performed on a minority of patients.

This motivates this study, which aims to find better imputation methods to retrospectively generate the HPV status of oropharyngeal cancer patients that were not tested. While many imputation models already exist, they use only clinical data to perform that imputation. However, recent developments in the field of radiomics, which focuses on extracting large amounts of features from radiographic medical images, have enabled a novel approach to imputation, using not only clinical features but also radiomic ones in hopes of further increasing the quality of the imputation.

By making use of a large dataset, which contains more than four thousand different cancer patients and their CT scans, this study will assess both the usefulness of said radiomic features and the improvement in the quality of imputation when using them.

Chapter 2

Background and previous work

There are almost 3000 papers in pubmed regarding HPV status and its relation to oropharyngeal cancer. Out of them, two papers specially illustrate the context of this study and are some of the most comprehensive pieces in the field. The first one is Habbous et al. [5], which uses a combination of clinical and epidemiological data, and the second one is Leijenaar et al. [7], which uses CT scans extracted from patients. The following sections will explore each one of them and their implications for this study.

2.1 Human papillomavirus in oropharyngeal cancer in Canada: analysis of 5 comprehensive cancer centres using multiple imputation.

In their work Habbous et al. [5] use social, clinical and demographic characteristics of 3643 patients diagnosed between 2000 and 2012 with oropharyngeal cancer cancer to estimate the HPV (p16) status of all patients with missing values. Through their study they find that their historical data fails to reflect the known increase of HPV positive oropharyngeal cancer during the last decades. This is caused by a significant selection bias in testing during the earlier years studied, when only patients that were already considered quite likely to be HPV positive were tested. Because of that, any modeling or study of HPV status using that dataset would yield biased results. However, after using their method of imputation, they identified a rise in HPV positive cancer incidence, with their best-probability cut point model reporting an increase from 47.3% in 2000 to 73.7% in 2012 (p<0.001). They trained one model for each cancer centre and then a model for all the data, randomly splitting the patients into a training and a testing dataset while preserving class proportions for the outcome. The global model obtained an area under the receiver operator curve[1] of 0.86 [95% CI (0.83 - 0.89)].

This paper serves to motivate the research for better methods of HPV status imputation. Bias in the testing process makes HPV not only very hard to model historically, but also diminishes its use as a prognostic tool for survival analysis. Datasets in medicine already have reasonably low amounts of patients, and discarding all those with unknown HPV status is not an option when training survival models. But given that HPV status is biased and with many missing values, the alternative is to eliminate HPV status from the survival analysis inputs, which results in a significant loss of relevant information. With solid HPV imputation models it would be possible to leverage all of the available patient data by calculating any missing HPV values.

2.2 Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: a multicenter study.

This study by Leijenaar et al. [7] focuses on the use of radiomic data to identify the HPV status of oropharyngeal cancer patients. To do that, 778 patients coming from four different cohorts were analysed to obtain 902 features from their CT images. Then different models were trained on those features together with clinical features to predict HPV status. Their models obtained areas under the receiver operator curves ranging from 0.70 to 0.80, which is similar to the results obtained by Habbous et al. [5]. The differentiating factor for the models was whether the training and validation sets contained patients with dental artifacts or not[2]. Their results show no significant difference between the performance on models trained on all data and only data without artifacts, and neither between using a validation set with all

[1] The area under the receiver operating curve represents the fraction of pairs in the data where the observation with a positive outcome has the higher probability of positive outcome predicted by the model. In the case of this study, the positive outcome is the patient being HPV positive, meaning that the model correctly ranks the probability of an HPV positive patient to have HPV higher than that of an HPV negative patient. Further description of this metric and the reasoning behind using it is provided in chapter 4.4.1

[2] Dental artifacts appear in CT scans when patients have metal teeth. They are zones of extreme bright values that degrade the quality of acquired images.

data and only data without artifacts.

While this paper uses radiomic features to predict HPV status, one key missing factor is a comparison between their model using both clinical and radiomic features and one built using only the clinical features as a baseline. Clinical features are more standardized and easier to obtain than radiomic features, an have been the ones used in most of the previous literature. Even though the aim of their paper is to provide proof of concept that molecular information can be derived from standard medical images, the lack of comparison with the trivial model results in an incomplete assessment of their usefulness. For this reason, this thesis will report a thorough comparison between purely radiomic, purely clinical and clinical plus radiomic models to ensure that there is significant difference between a hybrid model and a clinical one.

This previous papers provides a baseline expected performance for the models that will be built, taking their area under the receiver operator curve of 0.7636 [95% CI (0.6874–0.8399)] when training and validating on all data as a golden standard.

Chapter 3

Datasets

The full dataset used in this thesis is called RADCURE. It contains information of patients from the Toronto General Hospital diagnosed over the last two decades. This dataset was then split into OPC1, OPC2 and OPC3. OPC1 and OPC2 were already published at the beginning of the study, with OPC1 being the training dataset and OPC2 a smaller validation dataset. Afterwards OPC3 was collected, aiming to create a bigger validation dataset to investigate the long generalizability of the model.

The next sections will describe the patient distribution in the different datasets used.

3.1 RADCURE

RADCURE is a dataset that contains information of 1705 oropharyngeal cancer patients, with year of diagnosis ranging from 2004 to 2017. The amount of patients per diagnosis year is shown in figure 3.1.

This dataset was curated by eliminating all patients diagnosed during 2014 due to their low number, which made observations for that year too noisy. The final number of patients is 1696, diagnosed between 2005 and 2017. The HPV status was determined using p16 protein immunohistochemistry, a well known method to establish HPV status in oropharyngeal cancer. Some of the patients were untested, therefore having unknown HPV status.

The amount of HPV positive oropharyngeal squamous cell carcinoma cancers has been steadily increasing in the past years. However, the RADCURE dataset doesn't reflect this

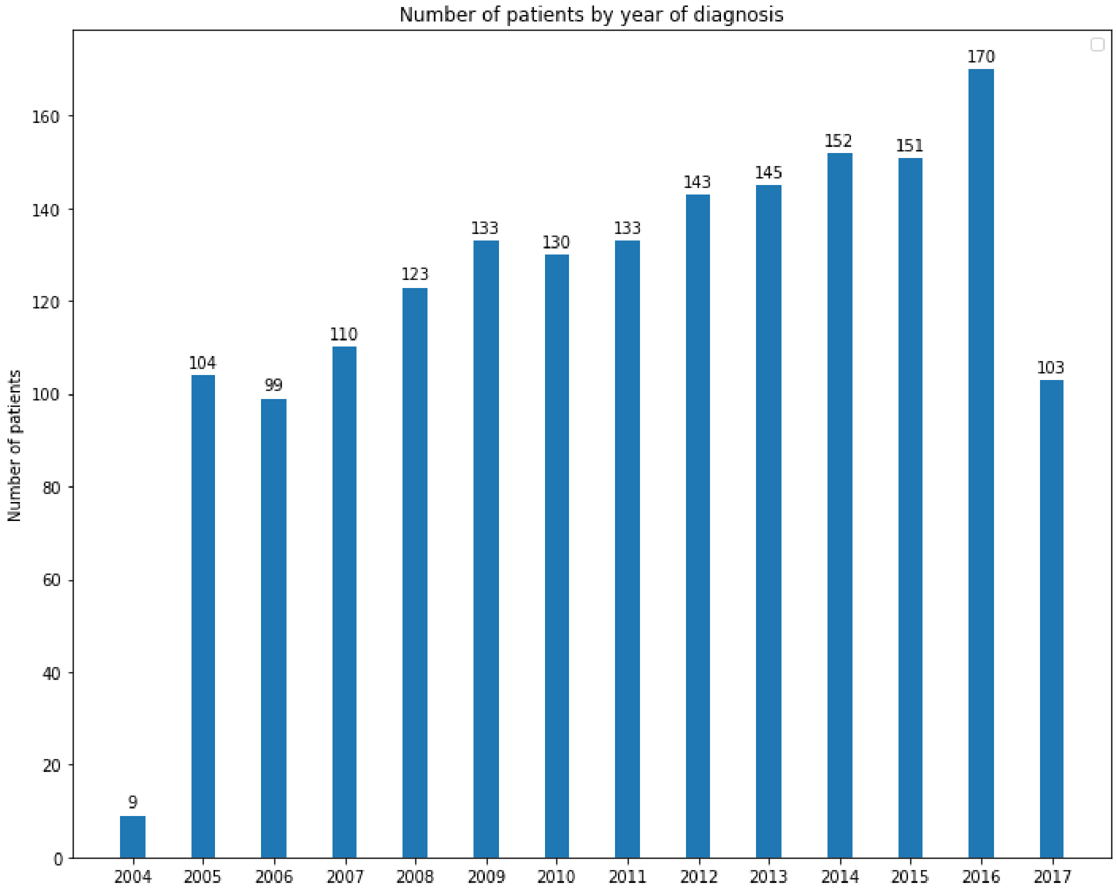

Figure 3.1: Number of patients diagnosed each year in the RADCURE dataset

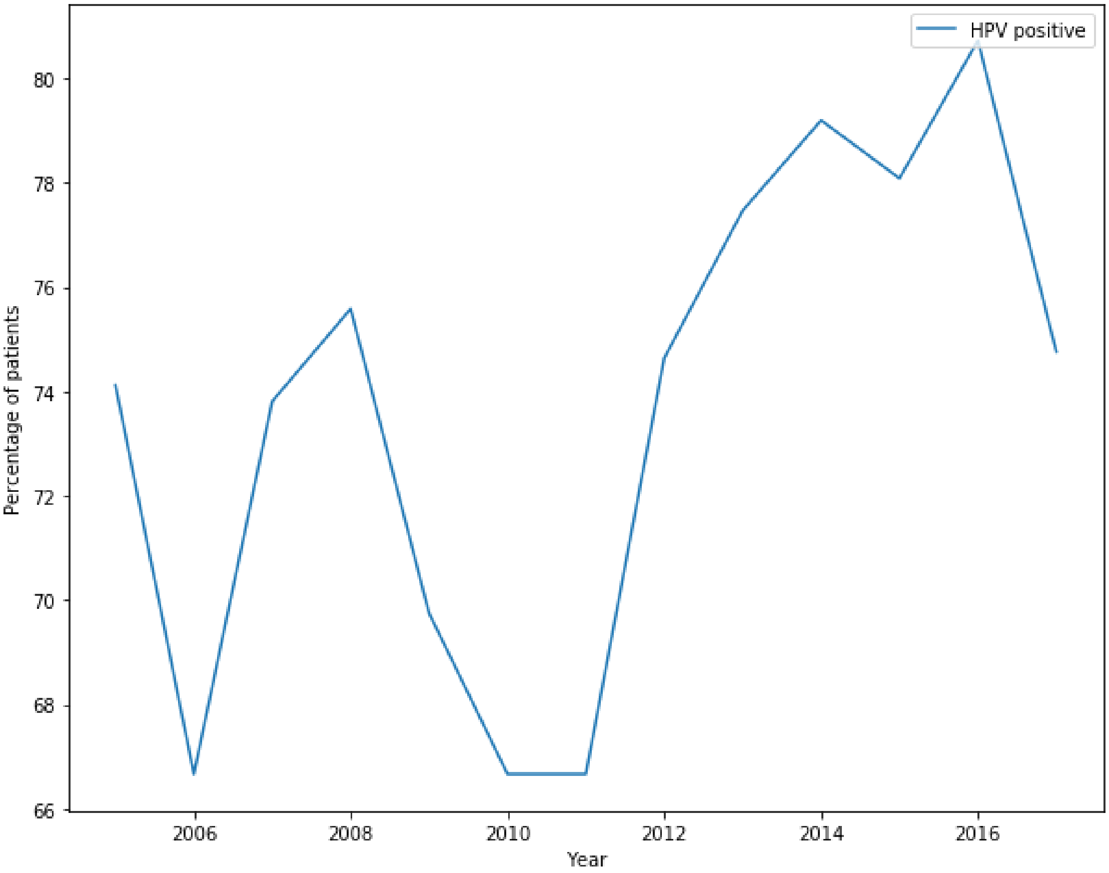

Figure 3.2: Percentage of HPV positive patients in RADCURE depending on the diagnosis year

trend.

Figure 3.2 shows the percentage of HPV positive patients diagnosed with OPSCC cancer depending on the year that they were diagnosed. That percentage does not indicate a clear increase, with the statistical test failing to reject the null hypothesis of independence between year of diagnosis and percentage of HPV positive patients (p-value = 0.265), which motivates further analysis on the distribution of patients between three categories: tested positive, tested negative and not tested. That distribution in regards to year is presented in figure 3.3

Out of a total of 1696 patients in RADCURE from 2005 to 2017, 151 of them have no known HPV status, representing 8.90% of the whole dataset. This is accentuated in the earlier years, which are the most informative from a ML perspective to perform survival analysis because they have a longer follow-up time, with 24.31% of patients diagnosed from 2005 to

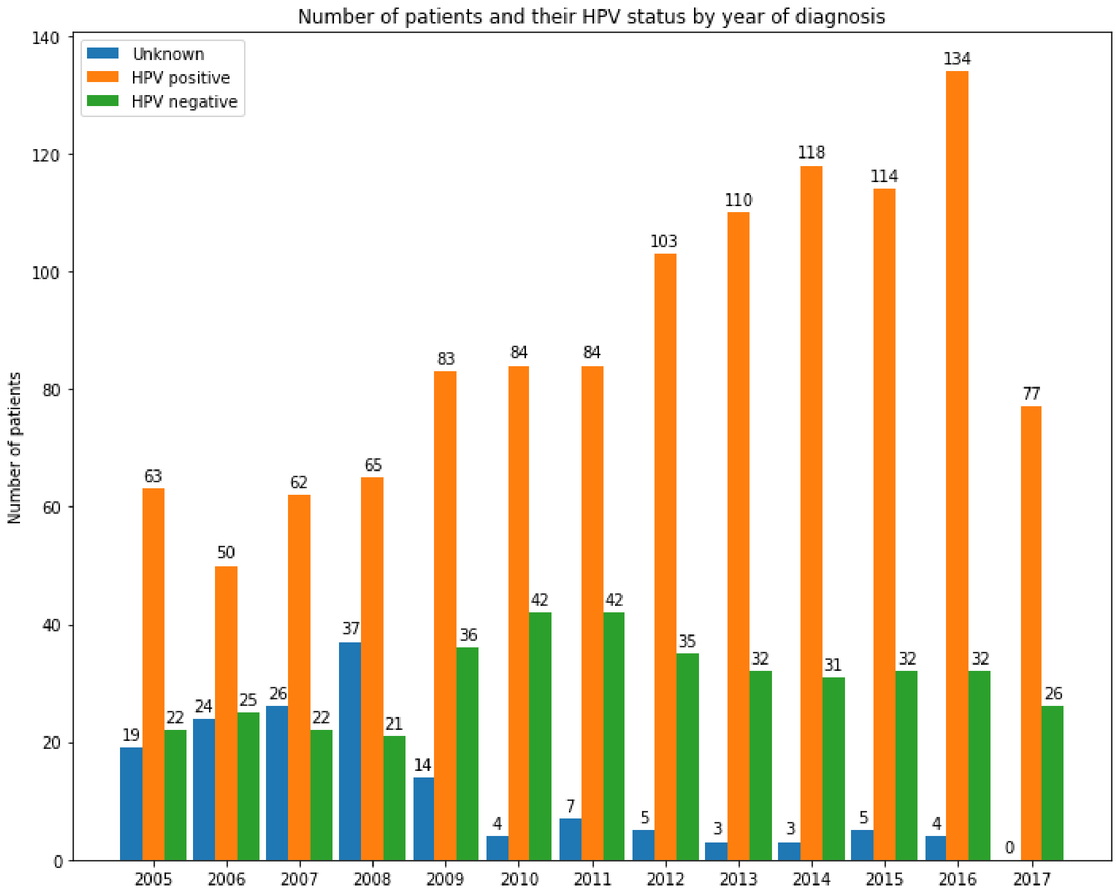

Figure 3.3: Comparison between the amount of patients with known and unknown HPV status depending on the year of diagnosis

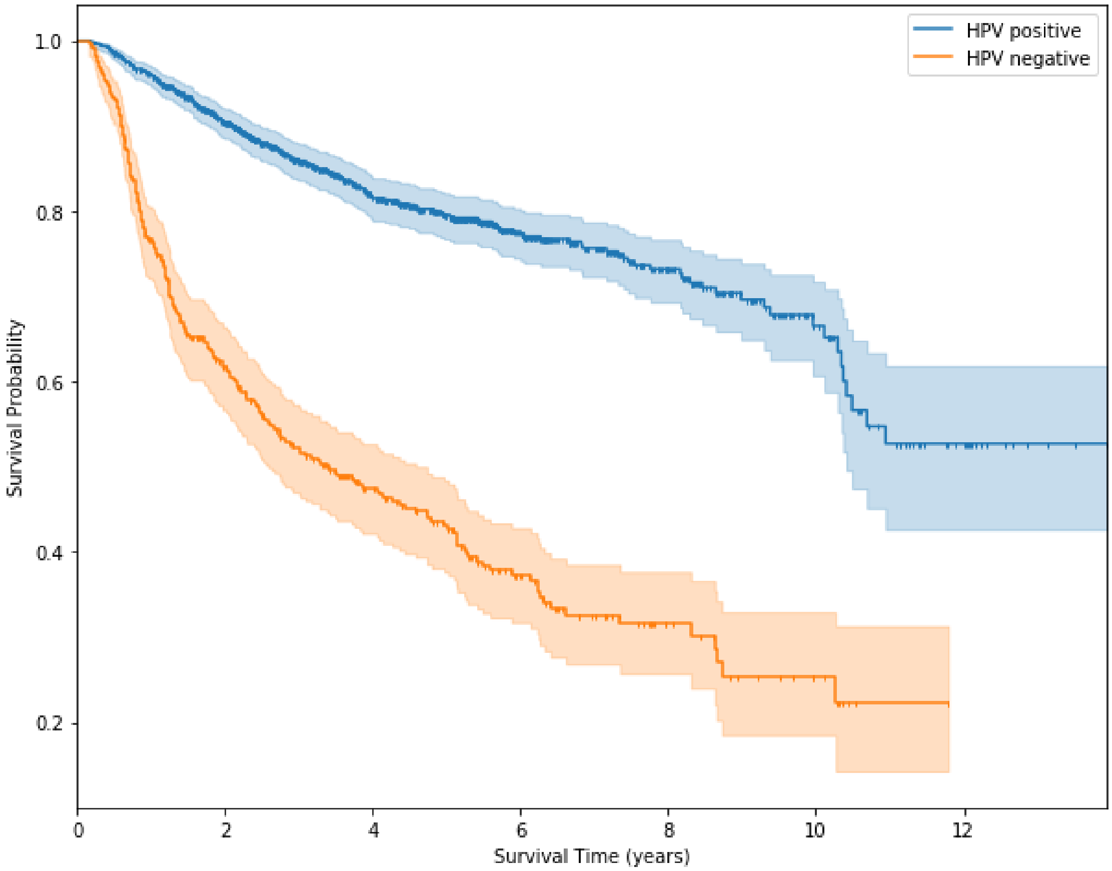

Figure 3.4: Kaplan-Meier estimator of survival depending on HPV status for all patients in RADCURE

2008 having untested HPV status. Since HPV status is strongly related with survival (as highlighted in figure 3.4, p-value < 0.005 on the logrank test for independence of survival regarding HPV status), this represents a very important loss of information and motivates the effort to find highly accurate methods to impute this missing values. Otherwise known trends such as the increase of HPV positive OPSCC cancer over the years can be hidden due to the missing data. This findings are consistent with the results described in *Habbous S, Chu KP, Lau H, et al. Human papillomavirus in oropharyngeal cancer in Canada: analysis of 5 comprehensive cancer centres using multiple imputation.*

While the 151 untested patients are not part of any of the datasets used in training nor validation, they were stored to test the imputation methods.

3.2 Separation method and possible bias

Figure 3.5 presents the year of diagnosis of the patients in each dataset. Patients in OPC1 were diagnosed between 2005 and 2010, patients in OPC2 were diagnosed between 2012 and 2015, and patients in OPC3 were diagnosed between 2005 and 2017. This difference is intentional, given that OPC1 and OPC2 were the first two datasets extracted from RADCURE. A subset of 606 patients was formed, and then that subset was split amongst those that were diagnosed before 2011 and after 2011. In the case of OPC3, patients were randomly selected from RADCURE (excluding those already in OPC1 and OPC2) and then were not split depending on the year of diagnostic. This creates some bias in the selection of patients between OPC2 and OPC3. However, given that the first subset of 606 patients was randomly selected from RADCURE and so was OPC3, this bias should not result in any inconsistencies further than those caused by the difference in year of diagnostic and treatment. Therefore, unless there is some concept drift in the RADCURE superset, OPC2 and OPC3 should behave almost identically as validation sets.

3.3 OPC1

OPC1 acted as the training set. It has a total of 421 patients with both CTV and GTV masks. The average age for patients in OPC1 is of 61 years, with a standard deviation of 10, a maximum of 89 and a minimum of 33. The distribution of the clinical features that are relevant for the model can be found in table 3.1. Subsite describes where the tumour is placed, ECOG Performance Status is a categorical variable describing the general well-being of cancer patients, stage summarizes the extent to which a cancer has developed, T status represents the extent of the primary tumour, N stage captures the number of nearby lymph nodes that the cancer has and M stage describes whether the tumour has metastasized. T, N and M are part of the TNM staging system, a standard for medical healthcare.

Out of the 421 patients, 294 are HPV positive and 127 are HPV negative, making the training set relatively unbalanced (70% to 30% class imbalance). Because of this the metric for performance will have to be chosen carefully, avoiding those that are most impacted by imbalance in classes.

Clinical variables and their occurrences in OPC1		
	Subcategory of the variable	**Number of ocurrences**
Sex	Male	338
	Female	83
ECOG PS	ECOG 0	265
	ECOG 1	108
	ECOG 2	38
	ECOG 3	8
	Unkown	2
Smoking status	Current smoker	135
	Ex-smoker	167
	Non-smoker	119
Drinking status	Heavy drinker	88
	Moderate drinker	51
	Light drinker	52
	Ex-drinker	37
	Non-drinker	188
	Unknown	5
T	T1	56
	T2	133
	T3	138
	T4a	65
	T4b	29
N	N0	72
	N1	39
	N2a	23
	N2b	151
	N2c	105
	N3	31
M	M0	421
Stage	I	5
	II	27
	III	58
	IVA	274
	IVB	57
Subsite	Base of the tongue	151
	Tonsil	114
	Tonsillar fossa	91
	Soft palate	27
	Tonsil pillar	11
	Lateral wall	10
	Vallecula	9
	Posterior wall	8
	Uvula	0

Table 3.1: Spread of the clinical features for the patients of OPC1

3.4 OPC2

OPC2 was the first dataset used for validation. It is smaller than OPC1 and has a total of 185 patients. The average age for patients is of 60 years, with a standard deviation of 9, a maximum of 85 and a minimum of 33. Table 3.2 compiles the values of the clinical variables.

The class imbalance is slightly higher than the one in OPC1, with 139 patients being HPV positive and 46 HPV negatives, causing a 75% to 25% class imbalance.

3.5 OPC3

Finally, OPC3 was the dataset obtained when the model had already been built and validated using OPC1 and OPC2. Having 499 patients, it is significantly bigger than OPC2, thus providing more confidence in its assessment of the model's performance. For that reason the model was re-validated on it. In this dataset the average age for patients is of 62 years, with a standard deviation of 9, a maximum of 87 and a minimum of 32. The spread of the clinical variables is shown in table 3.3.

The class imbalance is very close to the one in OPC2, albeit higher. 381 of the patients are HPV positive and only 118 are HPV negatives, resulting in a 76% to 24% class imbalance.

Clinical variables and their occurrences in OPC2		
	Subcategory of the variable	**Number of ocurrences**
Sex	Male	161
	Female	24
ECOG PS	ECOG 0	119
	ECOG 1	58
	ECOG 2	7
	ECOG 3	1
	Unkown	0
Smoking status	Current smoker	76
	Ex-smoker	62
	Non-smoker	47
Drinking status	Heavy drinker	38
	Moderate drinker	29
	Light drinker	51
	Ex-drinker	16
	Non-drinker	47
	Unknown	4
T	T1	26
	T2	71
	T3	46
	T4a	26
	T4b	16
N	N0	17
	N1	14
	N2a	6
	N2b	82
	N2c	54
	N3	12
M	M0	185
Stage	I	0
	II	9
	III	17
	IVA	135
	IVB	24
Subsite	Base of the tongue	41
	Tonsil	50
	Tonsillar fossa	65
	Soft palate	10
	Tonsil pillar	7
	Lateral wall	6
	Vallecula	0
	Posterior wall	4
	Uvula	2

Table 3.2: Spread of the clinical features for the patients of OPC2

Clinical variables and their occurrences in OPC3		
	Subcategory of the variable	**Number of ocurrences**
Sex	Male	401
	Female	98
ECOG PS	ECOG 0	298
	ECOG 1	175
	ECOG 2	23
	ECOG 3	3
	Unkown	0
Smoking status	Current smoker	155
	Ex-smoker	210
	Non-smoker	134
Drinking status	Heavy drinker	88
	Moderate drinker	62
	Light drinker	137
	Ex-drinker	48
	Non-drinker	147
	Unknown	15
T	T1	86
	T2	165
	T3	141
	T4a	83
	T4b	24
N	N0	55
	N1	47
	N2a	23
	N2b	201
	N2c	147
	N3	26
M	M0	499
Stage	I	8
	II	18
	III	54
	IVA	370
	IVB	49
Subsite	Base of the tongue	242
	Tonsil	112
	Tonsillar fossa	97
	Soft palate	16
	Tonsil pillar	16
	Lateral wall	4
	Vallecula	5
	Posterior wall	6
	Uvula	1

Table 3.3: Spread of the clinical features for the patients of OPC3

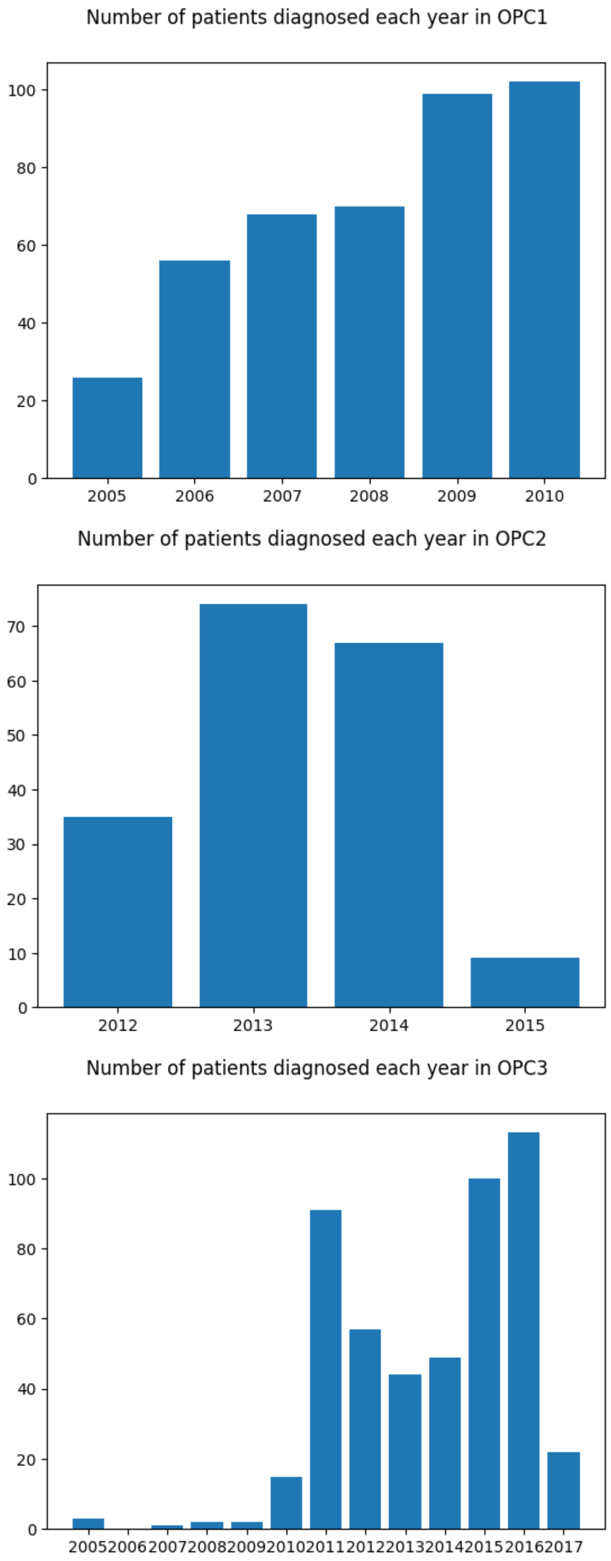

Figure 3.5: Spread of the year of diagnosis for the patients in each dataset

Chapter 4

Methodology

The objective is to use radiomic features from the patient's CT scans to improve HPV imputation. To improve imputation it is necessary to find the best predictive model for HPV status, so the following questions about how to generate and choose the features for the model have to be addressed:

1. Is CTV or GTV a better contour for image extraction?
2. How will the radiomic features be extracted?
3. Do radiomic features help make more accurate predictions?

The next sections discuss how these questions will be tackled.

For this study, the data from patients with dental artifacts was not eliminated from the datasets. The decision to use all data instead of removing the one with dental artifacts is mostly based on the fact that it did not result in any significant difference in performance in Leijenaar et al. [7], and using data from all patients increases the size of the datasets and possibly the generalization power of models built with them.

Patients with unknown HPV status were removed from OPC1, OPC2 and OPC3 and stored to be imputed after an accurate enough model was found. To assess the quality of the imputation, the survival curves of patients with known HPV status will be compared with those of patients with imputed HPV status. If the imputation is accurate, a clear overlap between the imputed and the tested behaviours is expected, with the imputation showing the difference in life expectancy between HPV positive and HPV negative patients.

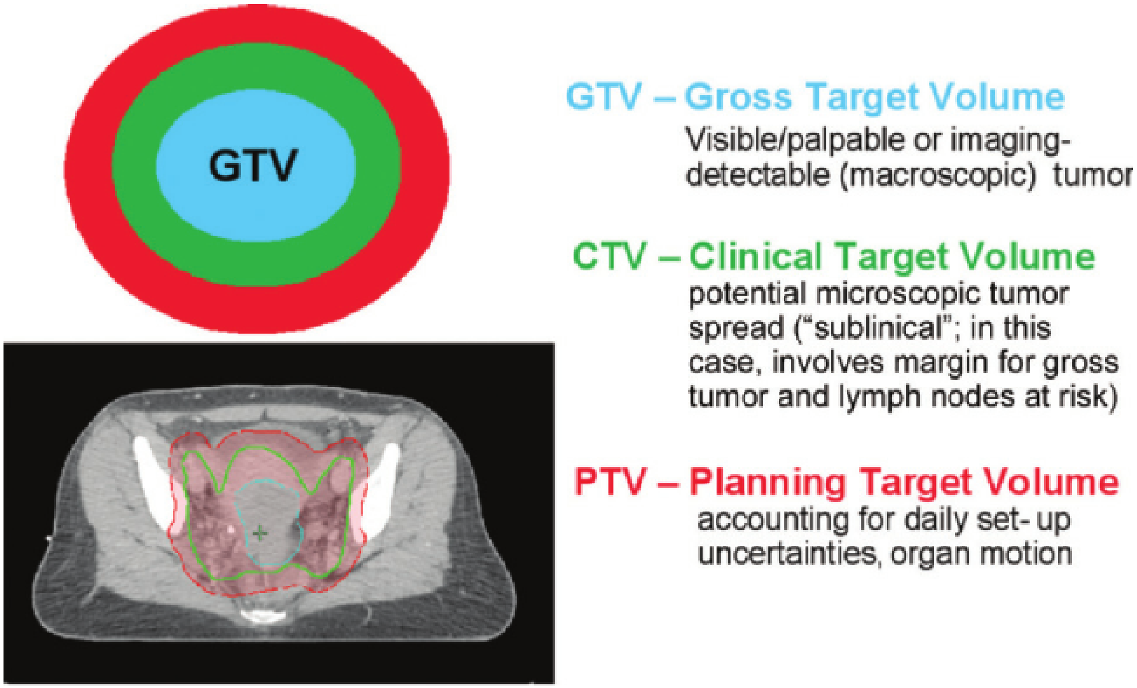
GTV
GTV – Gross Target Volume
Visible/palpable or imaging-detectable (macroscopic) tumor
CTV – Clinical Target Volume
potential microscopic tumor spread ("sublinical"; in this case, involves margin for gross tumor and lymph nodes at risk)
PTV – Planning Target Volume
accounting for daily set- up uncertainties, organ motion

Figure 4.1: Comparison of the different contours [17]

4.1 CTV vs GTV

When performing radiotherapy on a patient, volumes have to be defined for the treatment by a radiation oncologist. The main ones are GTV, CTV and PTV [1]. GTV stands for *gross tumour volume* and represents the position and extent of the main tumour. *Clinical target volume* or CTV is a combination of the GTV plus a a margin for sub-clinical disease spread. Finally, the PTV or *planning target volume* is a volume that contains the CTV and indicates the parts that will be treated. It is defined in a way such that it ensures the CTV will be dosed. Figure 4.1 In our datasets, patients usually have both a CTV and a GTV mask. This mask will be used in tandem with the CT scans of the patient to obtain the imaging features. Therefore, it is crucial to chose the one that performs the best for modelling. To asses their differences, all models and sets of input features will be trained and calculated using both masks. Furthermore, only patients that have both masks will be used in the training and validation sets to ensure that the results are comparable and that there's no confounding factors in the data.

4.2 Imaging features extraction

Each given patient will have his clinical information, his CT scans and his contour masks. Both the scans and the masks will be used to better extract the interesting features from the image. While the Python library known as PyRadiomics is the most common and explored method in the previous literature, it is also important to investigate other extraction strategies. Both PyRadiomics and its alternatives are better explained in the following subsections.

4.2.1 PyRadiomics

As aforementioned, PyRadiomics is a Python library that extracts features from a CT scan and its corresponding mask [3]. Those features are mathematically meaningful but not necessarily medically meaningful. They were chosen to cover a wide spectrum of the properties of the volumes so that they would be a good overall representation of the information contained in the scans. The version used in this study is PyRadiomics 2.1.2. When using it, the user can define a set of filters that will be applied to the scans from the allowed pool, which is compromised of:

1. *Original* (no filter applied)
2. *Wavelet filter* with 8 decompositions formed by combinations of High or Low pass filters in each dimension
3. *Laplacian of Gaussian filter* which enhances edges
4. *Square filter*
5. *Squareroot filter*
6. *Logarithm filter*
7. *Exponential filter*
8. *Gradient filter*
9. *LocalBinaryPattern2D* which performs the operation for each slice separately
10. *LocalBinaryPatter3D* which performs it on the whole volume

For each one of these filters, different sets of features can be extracted: b

1. *Firstorder* features describe the voxel intensity of the tumour
2. *Shape* features summarize the different geometrical properties of the tumour. For that reason, they are only computed in the original tumour volume
3. *Glcm* or Gray Level Co-occurrence Matrix is defined as the matrix such that the $(i, j)^th$ element represents the number of times the combination of levels i and j occur in two pixels of the image, separated by a distance of δ along angle θ. Features are calculated for an array of these matrices with different angles and distances, and then averaged.
4. *Glrlm* or Gray Level Run Length Matrix describes the length of same gray level consecutive pixel streaks. The $(i, j)^th$ element is equal to the number of streaks that have intensity i and length j along a given angle ω. Similarly to how glcm works, a set of configurable angles are used to calculate different matrices, and the features on each matrix then averaged.
5. *Glszm* or Gray Level Size Zone Matrix, where the $(i, j)^th$ element is the number of zones with voxel intensity i and volume j in the scans. This is uniquely defined, so features of this matrix are only computed once.
6. *Gldm* or Gray Level Dependence Matrix quantifies the number of voxels separated by a distance of δ or less that have a difference in value less than α of the central voxel. Averaging of features is performed like in the glcm and the glrlm.
7. *Ngtdm* or Neighbouring Gray Tone Difference Matrix indicates how different the intensity of a voxel is compared to its neighbours in distance δ or less, storing the sum of those absolute differences. Averaging is also performed.

Using all of the possible filters and extracting all of the possible sets of features, each volume results in almost 1700 features. Given the size of the datasets, it is infeasible to use all of them as inputs for the models. The R package mRMRe 2.0.9 (Parallelized Minimum Redundancy, Maximum Relevance Ensemble Feature Selection) was used [6]. This package performs mRMR on a set of continuous or categorical variables. In this study, features were selected using the training set such that they were the best representation for the HPV status. A total of 30 sets of 30 features were generated, each one using a different subset of the patients. The reason to generate different sets of input features is that, given the relatively small size of the training dataset, the choice might be quite noisy. To account for that, mRMRe selects 30 different subsets of the patients and performs mRMR on each one of them, making sure that the resulting sets of features are all different. During the training process, the first step was to study the differences between the 30 different subsets of 30 PyRadiomic features. To do that, a part of the training set was left out for tuning (20%) making sure to preserve the proportion of HPV positive patients. Logistic regression models were trained on a combination of clinical data and each one of the different PyRadiomics subsets and then evaluated. The results were almost identical across all subsets, and ensembling the models obtained with different subsets didn't boost the performance any further. Therefore, and for the sake of simplicity, a single subset of PyRadiomic features was chosen (the one with the greatest performance in tuning) for all subsequent models. Out of these 30 PyRadiomic features, 12 of them were eliminated due to their weights in the logistic regression not being statistically different from zero (95% CI) during training, leaving a subset of 18 radiomic features.

It is important to note that each volume of CT scans has a slightly different *slice thickness*, meaning a different distance between consecutive scan slices. That can be troublesome, since two different tumours both having a height of 10 slices might have very different actual heights. This also happens, albeit quite less commonly, in the width and length of the slices, where one pixel might represent different distances depending on the patient. To prevent that discrepancy from making the features across patients impossible to compare, PyRadiomics can be configured (and will be in this study) to automatically resample the scans so that the size of a voxel is the same in all of the volumes (1mm per dimension).

4.2.2 Pre-trained convolutional neural network architectures

While PyRadiomics is the de facto method to extract features from scans, other methods were explored for the sake of completion. CNNs trained on Imagenet can be used as feature extractors when removing some of the last layers (in the case of this study, only the last one), which is known as *transfer learning.* So the ResNet [19] and Inception [15] pretrained models were downloaded using the PyTorch zoo model, and their last layer removed to obtain two different deep feature extractors, both resulting in hundreds or thousands of features (depending on the version used).

Imagenet is a dataset that contains RGB images of size 224 by 224, which is quite different from the dimensions that CT scan volumes have. To bridge that discrepancy, each cancer volume was analysed using its mask to find the slice with the biggest tumour area. This slice slice was taken as the central one, and stacked with both its adjacent slices to create a set of 3 slices that were joined as a 3 channel image. Then, these slices were cropped around the center to have size 224 by 224 (or 229 and 229 in the case of the Inception v3 architecture). While this approach is not ideal, specially because it converts the RGB channels into 3 different slices, it has been used with some success in previous literature [8, 11, 13], and so was considered an interesting alternative to the usual method.

4.2.3 3D convolutional neural neural networks

One final approach was to use the volumes of the patients to train a 3D convolutional neural network. Each volume was analysed using its mask to find the center of the tumour, and then a cube of side 50 centered in the tumour was extracted. Volumes had all been resampled to have the same size of voxels, 1mm per side. However, this path presented many issues, mainly lack of enough training data and the long training times of the networks.

In an attempt to circumvent those issues, data was augmented using random translations, random rotations and random flips, with an augmentation factor of 128. These augmentation methods have been used in a variety of previous radiomic studies [18, 14]. A GPU computation cluster was used to reduce training time to 3 days, while also training different models in parallel.

Despite these efforts, 3D convolutional networks failed to perform significantly better than just random guessing, and were therefore disregarded for the rest of the study. While this outcome was to be expected, having only the 421 patients from OPC1 to train a 3D convolutional network, it was still worth pursuing just to corroborate that the dataset was in fact too small for it. Maybe in the upcoming years there willR be pretrained 3D convolutional network models that are easily accessible, much like what happens with 2D CNNs, and then this approach will be feasible with the help of transfer learning and fine tuning.

4.3 Machine learning models

The four different machine learning models used will be logistic regression, random forest (with the XGBoost algorithm for training), a neural network and an ensemble of the previous three models by averaging their predictions. Logistic regression and random forest were chosen to add some interpretability to the model, allowing analysis on what features are the most predictive and how exactly they affect predictions. Neural networks were added as a more complex system model, which is more black-box but might get better predictions through the exploration of higher level relationships between the input variables. Finally, an ensemble was used because it is a well known strategy in machine learning to ensemble different models to increase robustness and sometimes predictive power. Hyperparameter tuning will be performed using 5-fold cross-validation. Cross-validation will also be used to decide the amount of hidden layers and the neurons per hidden layer in the MLP model. The random forest model will be trained using the XGBOOST library for Python, which performs optimized distributed gradient boosting. Both logistic regression and the neural network model will be coded in Pytorch 1.2 (ref: Pytorch website, using L1 and L2 regularization for the logistic regression and L1 regularization for the neural network to reduce the likelihood of overfitting on the training data. Both models will also use the Adam optimizer. Finally, the loss function for all three models is BCE loss. A diagram of the experimental process is shown in figure 4.2.

4.4 Comparing different sets of inputs

The ultimate goal of these study is improving HPV status imputation in oropharyngeal cancer through the use of radiomic data. Therefore, it is crucial to perform a deep study on the effect of using radiomic data in addition to the usual clinical data. Previous works have failed to show a statistically significant improvement when using a combination of radiomic and clinical data over using only clinical data, which is easier to obtain and to interpret. In these study, each one of the four machine learning algorithms will be trained using only clinical data, only radiomic data, and then a combination of clinical and radiomic data. By doing that the performance of purely radiomic models can be assessed. While it is not expected that it will be the best performer, it is necessary to ensure that only radiomic models do significantly better than random guessing, and by how much. Furthermore, having an only clinical model as a baseline will provide a comparison model for the one using both sets of input features. Statistical tests will be performed to compare the different AUCs of the models, determining which ones are better and significantly different from others.

For the clinical inputs, the variables used will be age, sex, smoking status, drinking status,

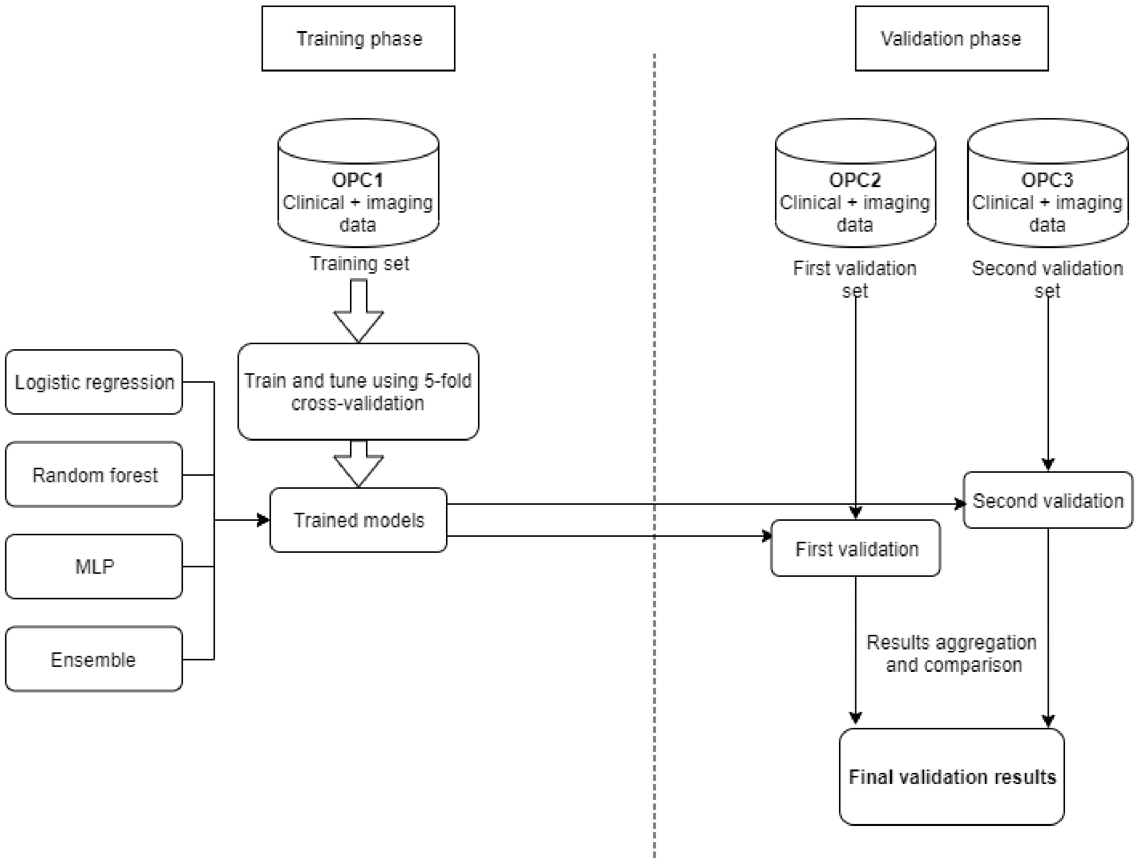

Figure 4.2: Diagram of the workflow for the training and validation of the models

subsite, ECOG Performance Status, stage, T status and N stage. However, the M stage was not used due to not having any patients with metastatic cancer in the training dataset. Most of the clinical variables had to be binned to make sure that each category had a significant amount of samples. The 30 radiomic features from PyRadiomics were, as mentioned previously, chosen using mRMRe from a set of 1700. Furthermore, early in tuning phase a total of 12 of them were eliminated due to not having odd ratios significantly different from zero in the logistic regression model. That resulted in a set of 18 radiomic input features that were relevant. Those features were original-shape-MinorAxis, lbp-2D-firstorder-InterquartileRange, wavelet-HHH-firstorder-Skewness, wavelet-LLH-firstorder-Mean, original-gldm-Dependence-NonUniformityNormalized, original-glcm-MaximumProbability, wavelet-HHH-glszm-SizeZone-NonUniformityNormalized, square-ngtdm-Busyness, exponential-ngtdm-Busyness, wavelet-HLL-glcm-Correlation, lbp-3D-k-ngtdm-Busyness, original-shape-Sphericity, lbp-3D-k-glszm-ZoneEntropy, exponential-firstorder-Kurtosis, exponential-ngtdm-Complexity, lbp-3D-m1 - firstorder-Skewness, lbp-2D-firstorder-90Percentile and wavelet-HLL-firstorder-Skewness. Most names are not very descriptive, but original shape features add some interpretability to what the model is looking at. Finally, subsets of 30 deep extracted radiomic features were also compiled using the ResNet and Inception architectures.

4.4.1 Area under the receiver operating characteristic curve

The metric used to describe the performance of the models will be the area under the receiver operating characteristic curve (AUC under the ROC curve or just AUC from now on), a well known metric used in classification problems. The ROC curve is plotted as the true positive rate (or TPR) versus the false positive rate (or FPR) of our model as we change the decision threshold. This curve monotonically increases from 0 to 1, given that a threshold of 0 will have a true positive rate of 1 (all samples are marked as positive, so there are no false negatives) but also a false positive rate of 1 (there are no true negatives since no data is being classified as negative). Similarly, when the threshold is 1 both the TPR and the FPR are 0. A perfect model will result in a ROC curve as shown in figure 4.3 since there is a certain threshold for which the TPR is 1 and the FPR is 0.

A more common example is shown in figure 4.4, in which case the ROC curve is above the straight line that connects the origin with point (1, 1) (which would be the expected performance for a model that is performing random guesses) while still not showing perfect

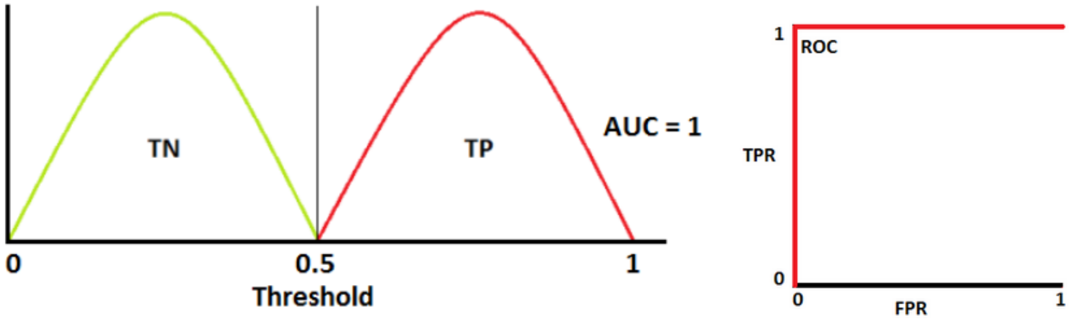

Figure 4.3: Model that perfectly distinguishes between two data classes and its ROC curve (Credit to: My Photoshopped Collection)

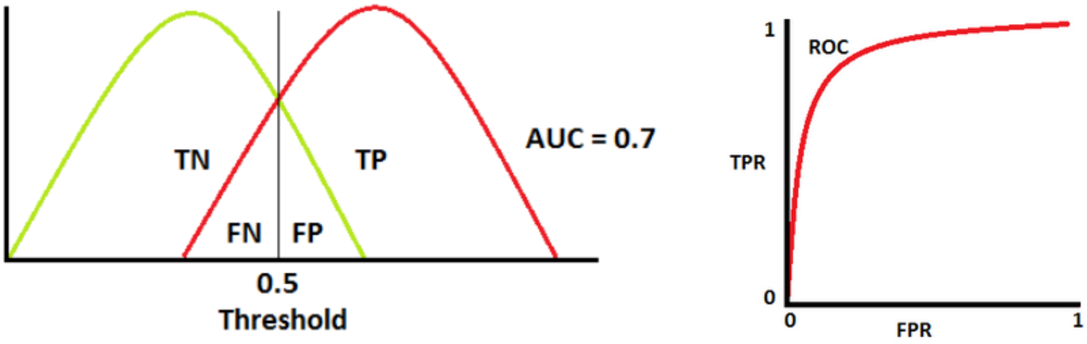

Figure 4.4: Model that mostly distinguishes between two data classes and its ROC curve (Credit to: My Photoshopped Collection)

predictive power.

In summary, the further up the ROC curve is, the better our model is performing. To capture this, the AUC is used. A random guessing model has an expected AUC of 0.5, while a perfect model has AUC of 1. Any results lower than 0.5 can be improved by simply flipping the predictions of the model. The AUC is identical to another performance metric, the concordance index (or CI), which is also quite studied. To calculate the CI, the model is given all possible pairs of one positive and one negative patient, and it has to make a prediction of which is which. The likelihood of the model correctly predicting which one of the two patients is the positive one is the concordance index. Another way to see it is the likelihood of properly ranking a positive patient higher than a negative patient in the prediction.

The two main reasons to use this metric over others are that it is the one canonically used in all previously cited works [4] [7] and that it helps circumvent the class imbalance present in both validation sets, while metrics like accuracy might be somewhat deceiving.

Chapter 5

Results

As mentioned in chapter 3, OPC1 was used for training, while both OPC2 and OPC3 were used for validation. Performing two different validations is quite an uncommon practise but if they were both sampled from the RADCURE superset, a significant difference in performance between them wouldn't be expected. However, as the next sections will show, there were notable mismatches in the performances for the validation datasets which will be further explored in the next chapter.

In each section, the first set of results presented will be those obtained while extracting imaging features with PyRadiomics to compare CTV and GTV. Then there will be the comparison of PyRadiomics with the deep feature extractors, followed by the results comparing only clinical models, only radiomic models and models with both inputs and a final subsection dedicated to summarizing the results. This breakup is performed for the sake of making the results more understandable and modular. During experimentation, a total of 64 different models (changing the input data, the contour type or the kind of model itself) were built and evaluated at the same time. Violin plots were built by creating 100 different and class-balanced subsets of 90% of OPC1 and training and then validating on them, while the overall performance is given for a single model trained using all of OPC1.

5.1 Results on OPC2

5.1.1 GTV vs CTV

The first question posed in chapter 4 is whether GTV or CTV contours have better prognostic value for HPV status in oropharyngeal cancer. All models were built using both GTV and CTV as masks, and the results proved that using imaging features obtained through the CTV masks resulted in better performance overall. Figure 5.1 shows a subset of these results, specifically when using both PyRadiomics and clinical data as the inputs for the models. It is clear that for each one of the four trained models, the AUC using CTV is better than the AUC using GTV. This disparity is very pronounced in the logistic regression, random forest and ensemble model but less significant in the MLP model, although still apparent.

Previous work in cancer treatment response prediction has shown the importance of imaging not only the main tumour but also some of the surrounding tissue [18]. This may be the main cause for the superior results of CTV versus GTV, given that CTV encompasses a bigger volume and therefore captures some of the surrounding tissues.

The next subsections will focus on results obtained using CTV given its better prognostic power.

5.1.2 PyRadiomics vs deep extracted features

Having selected CTV as the preferred mask, new imaging features were extracted using a different pretrained architectures. The performance using ResNet101 is shown in figure 5.2 compared to the PyRadiomics extraction. While only results with ResNet101 are being shown, not only Inception but also other versions of the ResNet architecture present in the PyTorch model zoo were tried. ResNet101 was the better performer and is therefore the one featured here. The purely ResNet101 models all get an AUC significantly above 0.5, showing that this method of extraction does result in some significant features. This is extremely interesting, given that only three slices of the patient are used, and that they are being sent in the RGB channels. Even when the method had been used before, it wasn't immediately clear just how well it would perform. However, purely PyRadiomics models still performed better every time, and combining PyRadiomic features with clinical ones also resulted in better results than combining the clinical data with ResNet101 features. Finally, mRMRe

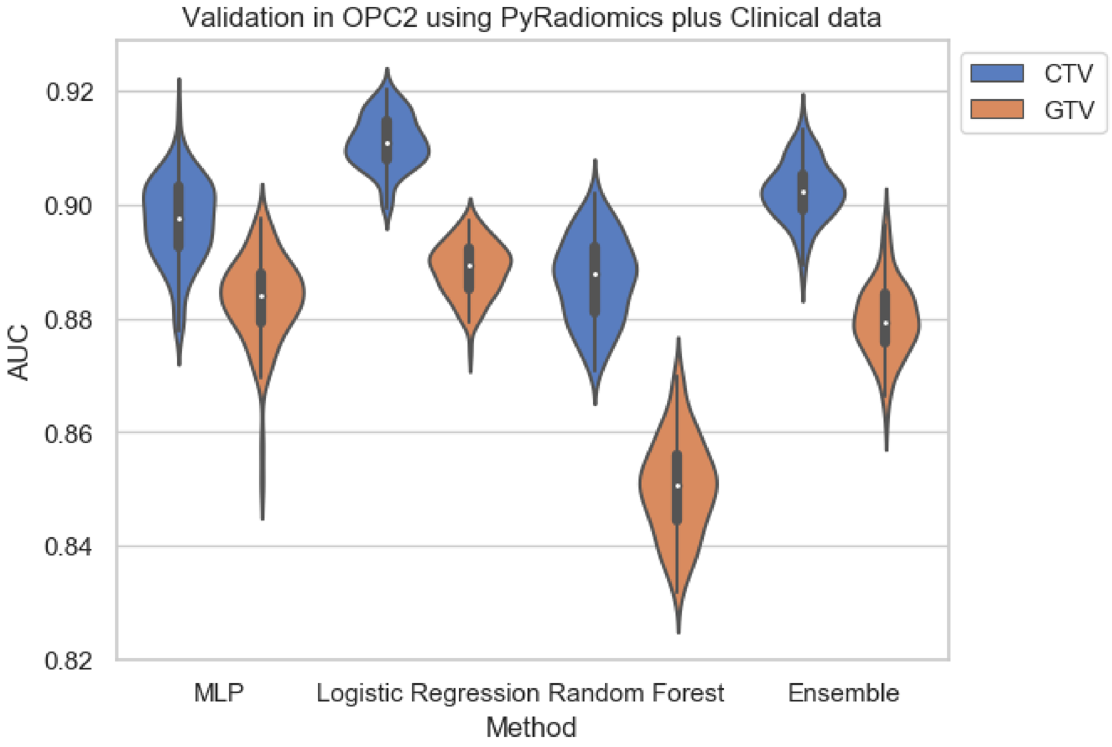

Figure 5.1: Comparison between using GTV and CTV in models that have clinical and PyRadiomics data as input

was performed to obtain a set of 30 features combining both PyRadiomics and Resnet101 features. While it performed better than purely ResNet101, it failed to improve upon the clinical plus pure PyRadiomics model.

Because of this results, PyRadiomics stayed the preferred method for feature extraction from the patient's CT scans, proving its greater descriptive power due to its use of not only three slices of the patient's scans but rather all of the volume.

5.1.3 Different kinds of inputs

Figure 5.5 shows the AUC obtained using either only clinical data, only PyRadiomics features or both. Across all four models, combining clinical data and PyRadiomics features gives the best overall performance. Furthermore, the performance of the models that use only clinical data is significantly lower, showing that imaging features are key to improve the predictions. This is extremely important, since imaging features are sometimes too hard to analyze, preventing models that use them from achieving a significant improvement over purely clinical models. The models using only radiomic features perform the worst of all three, but still significantly better than random guessing and similar to only clinical, further showing the relevance of CT scans as indicators of HPV status for tumours. Finally, the increase in performance when combining both clinical and radiomic features means that while they have similar performances separately, the information obtained through these two sources of data is significantly different, hence complementing each other when combined. A comprehensive list of all the results and their confidence intervals can be found in table 5.1.

The best result is obtained with logistic regression and clinical plus PyRadiomics inputs, with an AUC of 0.91 [95% CI (0.90 - 0.92)]. In comparison, the AUC of logistic regression with only clinical inputs is 0.89 [95% CI (0.88 - 0.89)], and for only PyRadiomics features it is 0.82 [95% CI (0.80 - 0.84)]. A comparison of the ROC curves of these three logistic regression models is shown in figure 5.3, and their respective precision-recall curves are shown in figure 5.4. In both figures the hybrid model is the best performing one, and all three are above the performance of the trivial models. Although there is no statistically significant difference between the logistic regression, MLP and ensemble models when trained on clinical data plus PyRadiomic features, given the fact that logistic regression is both the fastest model to train and the easiest to interpret, it is favored over the other two, and will therefore be the one used moving forward in this study.

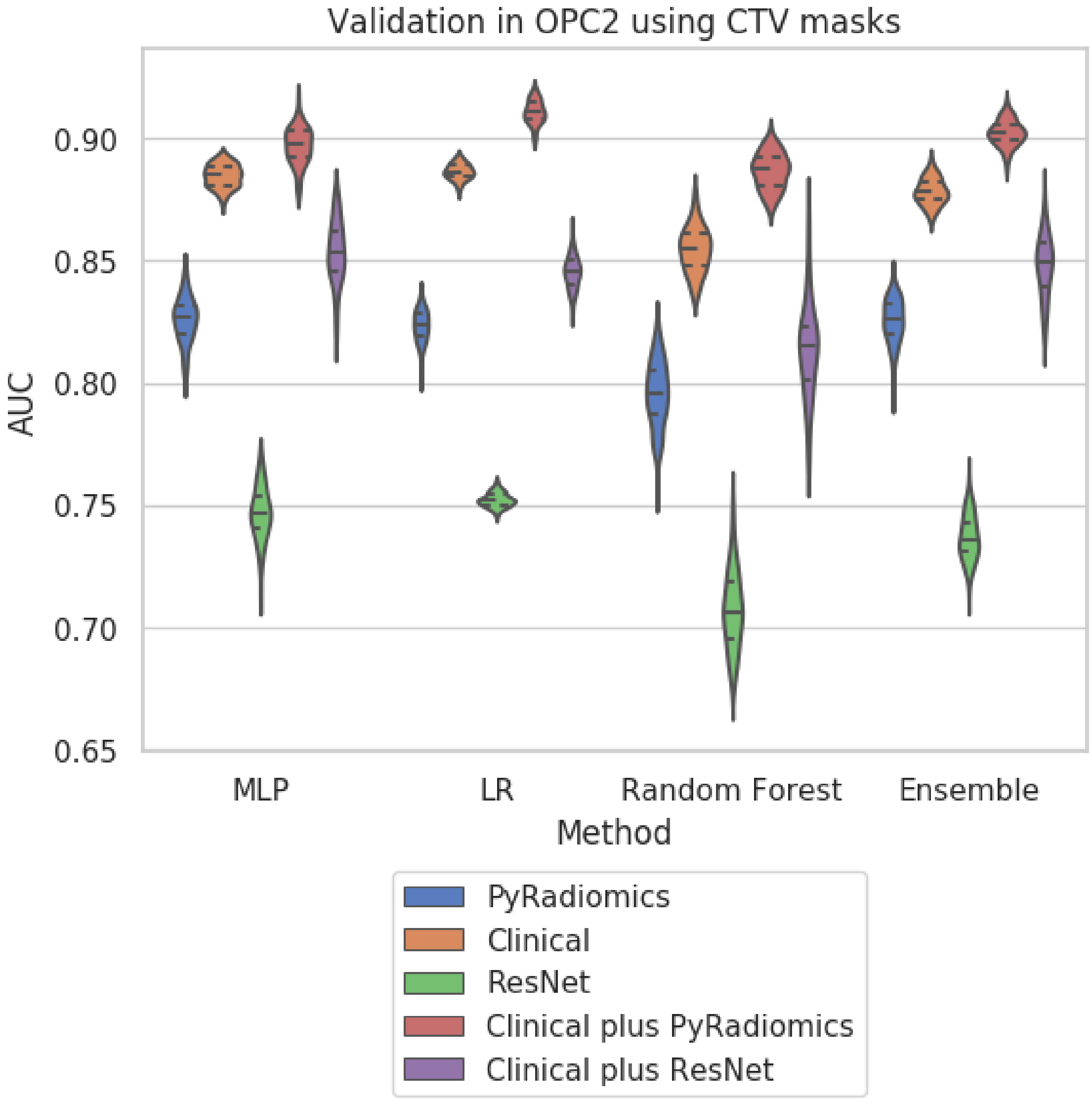

Figure 5.2: Comparison between using ResNet101 and PyRadiomics features in models that have CTV masks

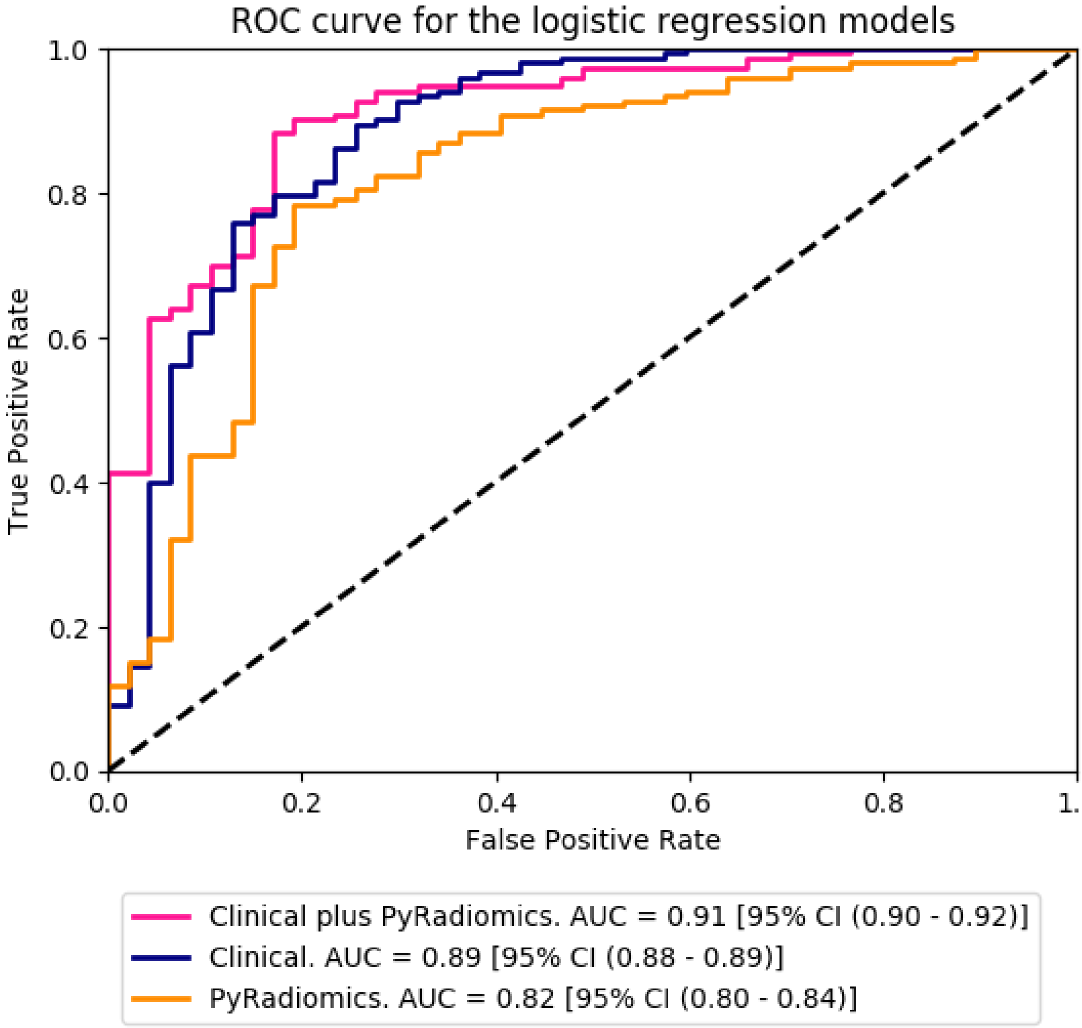

Figure 5.3: Receiver operating characteristic curves of the logistic regression models trained with only clinical, only radiomic and clinical plus radiomic features.

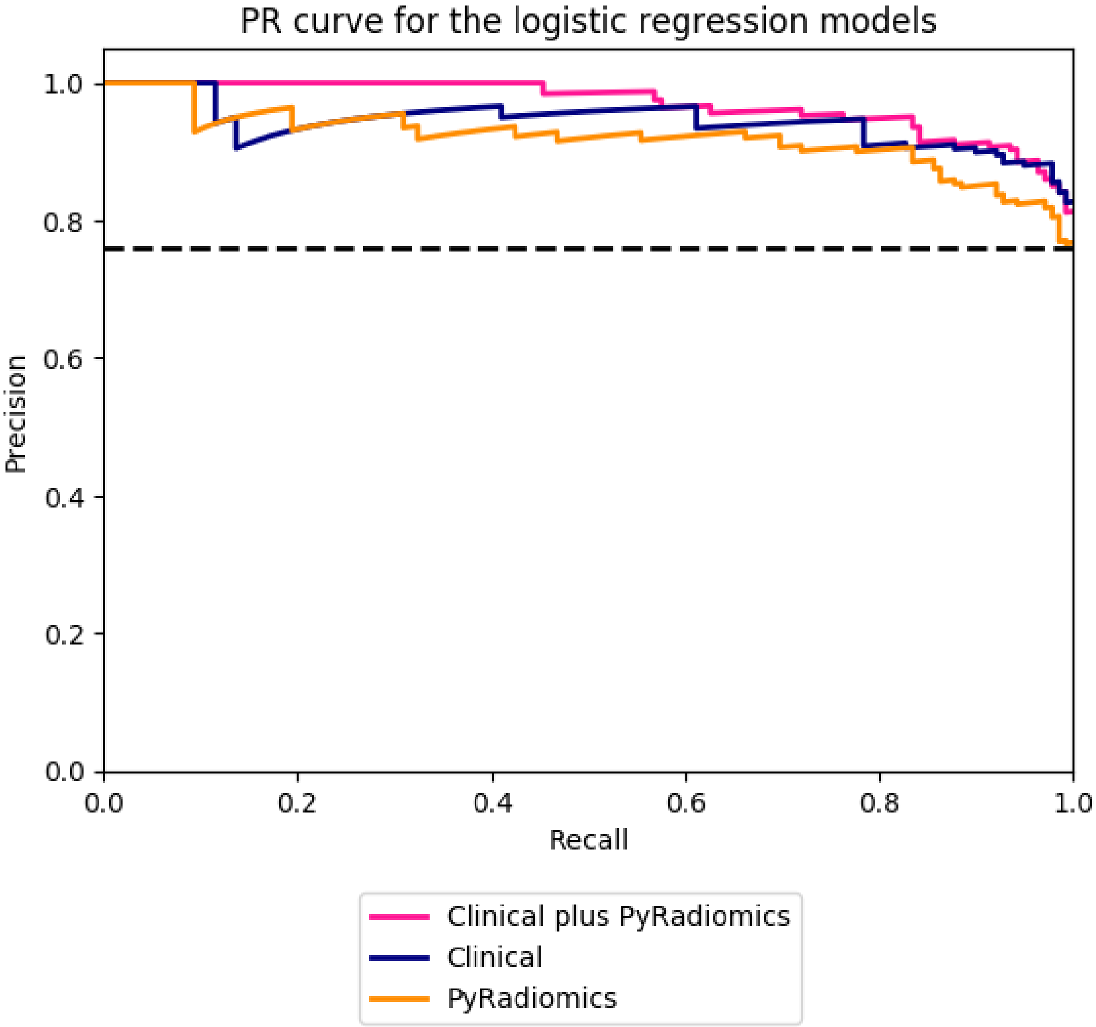

Figure 5.4: Precision-recall curves of the logistic regression models trained with only clinical, only radiomic and clinical plus radiomic features.

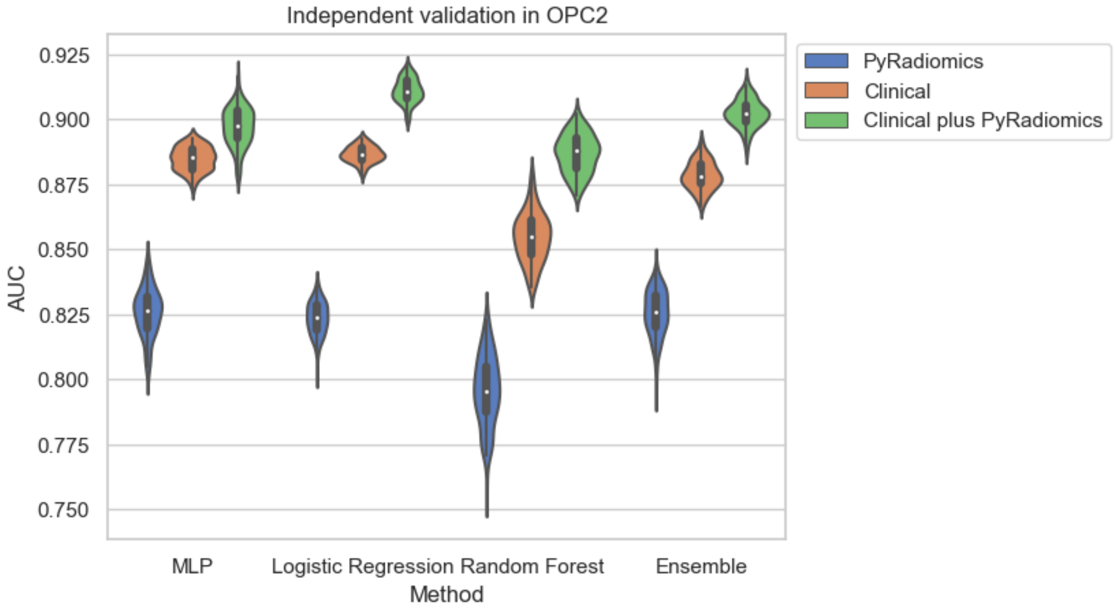

Figure 5.5: AUC of the different models built depending on the input

Concerns were expressed by medical practitioners during this study as to whether the model was indeed using a combination of all input features or mainly smoking status, age and tumour volume. The first two are well known indicators of HPV status, while the later tends to be highly predictive in the radiomics field [18]. To further investigate that possibility, a minimal model was built using these three features. The results obtained are presented in table 5.2, with AUCs significantly below those of the chosen model. These results, combined with the already observed improvement of performance when using the imaging features serve solidly prove that CT scans contain crucial information to maximize the predictive power of HPV status imputation models.

It is worth noting that the model built using both clinical and radiomic features does not include the volume of the tumour. The feature was eliminated during the mRMRe feature selection, and adding it manually didn't result in an improve in performance during tuning. This may be surprising, specially because this feature is highly predictive in other regards, for example survival. However, while the literature shows relation between nodal volume and HPV status [2], a relation between gross tumour volume and HPV status is yet to be found.

Mean AUC of the models in OPC2 and their 95% CI			
	PyRadiomics	**Clinical**	**Clinical plus PyRadiomics**
MLP	0.83 (0.81 - 0.84)	0.88 (0.88 - 0.89)	0.90 (0.88 - 0.91)
Logistic regression	0.82 (0.80 - 0.84)	0.89 (0.88 - 0.89)	0.91 (0.90 - 0.92)
Random forest	0.80 (0.77 - 0.82)	0.85 (0.84 - 0.88)	0.89 (0.87 - 0.90)
Ensemble	0.83 (0.80 - 0.84)	0.88 (0.87 - 0.89)	0.90 (0.89 - 0.91)

Table 5.1: AUC values for the different combinations of model and kind of input.

Mean AUC of the models in OPC2 and their 95% CI	
MLP	0.75 (0.73 - 0.76)
Logistic regression	0.75 (0.75 - 0.76)
Random forest	0.73 (0.71 - 0.75)
Ensemble	0.75 (0.73 - 0.76)

Table 5.2: AUC values for the models built using only the smoking status, age and tumour volume.

5.1.4 Analysis and summary

As mentioned in 5.1.3, logistic regression is favored because of its interpretability as well as its training speed. Table 5.3 summarizes the coefficients of the different clinical inputs of the model. Analyzing both the magnitude of each one of them, as well as their sign, provides insight on different trends observed from a clinical perspective in HPV positive oropharyngeal squamous cell carcinoma (OPSCC) cancer patients. Note that the weights depend only on the training set OPC1 but not on the validation set. Therefore, the analysis of weights will be performed only once.

Age has one of the highest absolute values and a negative sign, indicating that it is very informative and that the risk of being HPV positive decreases for older patients. That statement has already been proven in previous work, and derives from HPV being a sexually transmitted disease, which makes the likelihood of contagion lower when in a monogamous relationship. Similarly, smoking decreases the likelihood of an OPSCC cancer being HPV positive because smokers tend to have OPSCC cancers caused by their smoking habits. On this particular note, while being an ex-smoker has no statistical significance in the model, it is worth mentioning that it's only one of the three options, setting the overall smoking condition as relevant and with high absolute values. Similarly to smoking, drinking is also

Coefficients of the clinical features and their 95% CI	
Age	-0.0276 (-0.0317, -0.0226)
Sex	0.0140 (0.0088, 0.0192)
ECOG PS	-0.0193 (-0.0243, -0.0140)
T	-0.0154 (-0.0205, -0.0108)
N	0.0255 (0.0206, 0.0311)
Stage	0.0172 (0.0113, 0.0217)
Base of tongue subsite	0.0118 (0.0077, 0.0164)
Tonsil subsite	0.0158 (0.0106, 0.0204)
Other subsite	-0.0416 (-0.0465, -0.0368)
Current smoker	-0.0289 (-0.0344, -0.0238)
Ex-smoker	0.0009 (-0.0035, 0.0065)
Non-smoker	0.0284 (0.0243, 0.0335)
Ex-drinker	-0.0110 (-0.0179, -0.0052)
Drinker	-0.0162 (-0.0216, -0.0107)
Non-drinker	0.0221 (0.0173, 0.0274)

Table 5.3: Weights of the clinical features used in conjunction with radiomic features in the logistic regression model. Red indicates a weight isn't statistically significant.

very predictive of HPV status.

Regarding PyRadiomic features, the one with the greatest magnitude is the sphericity of the tumour, with a value of 0.0256 [95% CI (0.0200, 0.0311)], which indicates that tumours with a more spherical shape are more likely to be HPV positive. All chosen PyRadiomic features and their weights are summarized in table 5.4. This values were presented to a group of medical staff at Princess Margaret Cancer Centre. While the exact values of the features are impossible to estimate even by practitioners, they agreed with the sign of all of the clinical features, as well as with sphericity being somewhat linked to HPV status in OPSCC cancer.

To summarize all the findings using OPC2 as a validation set, the answers to the questions posed are:

1. CTV has the best performance to predict HPV status
2. The optimal method for feature extraction is PyRadiomics, even over deep feature extractors
3. Radiomic data significantly improves HPV status prediction over purely clinical models

Coefficients of the PyRadiomics features and their 95% CI	
original-shape-MinorAxis	-0.0156 (-0.0211, -0.0114)
lbp-2D-firstorder-InterquartileRange	0.0085 (0.0027, 0.0134)
wavelet-HHH-firstorder-Skewness	0.0072 (-0.0009, 0.0118)
wavelet-LLH-firstorder-Mean	0.0137 (0.0074, 0.0193)
original-gldm-DependenceNon-Uniformity-Normalized	-0.0175 (-0.0240, -0.0128)
original-glcm-MaximumProbability	0.0113 (0.0057, 0.0173)
wavelet-HHH-glszm-SizeZoneNon-Uniformity-Normalized	0.0047 (-0.0002, 0.0100)
square-ngtdm-Busyness	-0.0165 (-0.0220, -0.0100)
exponential-ngtdm-Busyness	-0.0145 (-0.0197, -0.0090)
wavelet-HLL-glcm-Correlation	-0.0028 (-0.0083, 0.0034)
lbp-3D-k-ngtdm-Busyness	-0.0235 (-0.0280, -0.0185)
original-shape-Sphericity	0.0217 (0.0174, 0.0269)
lbp-3D-k-glszm-ZoneEntropy	-0.0099 (-0.0151, -0.0048)
exponential-firstorder-Kurtosis	-0.0010 (-0.0059, 0.0051)
exponential-ngtdm-Complexity	0.0059 (0.0004, 0.0109)
lbp-3D-m1-firstorder-Skewness	-0.0055 (-0.0103, 0.0002)
lbp-2D-firstorder-90Percentile	0.0120 (0.0077, 0.0172)
wavelet-HLL-firstorder-Skewness	0.0079 (0.0011, 0.0130)

Table 5.4: Weights of the PyRadiomics features used in conjunction with clinical features in the logistic regression model. Red indicates a weight isn't statistically significant.

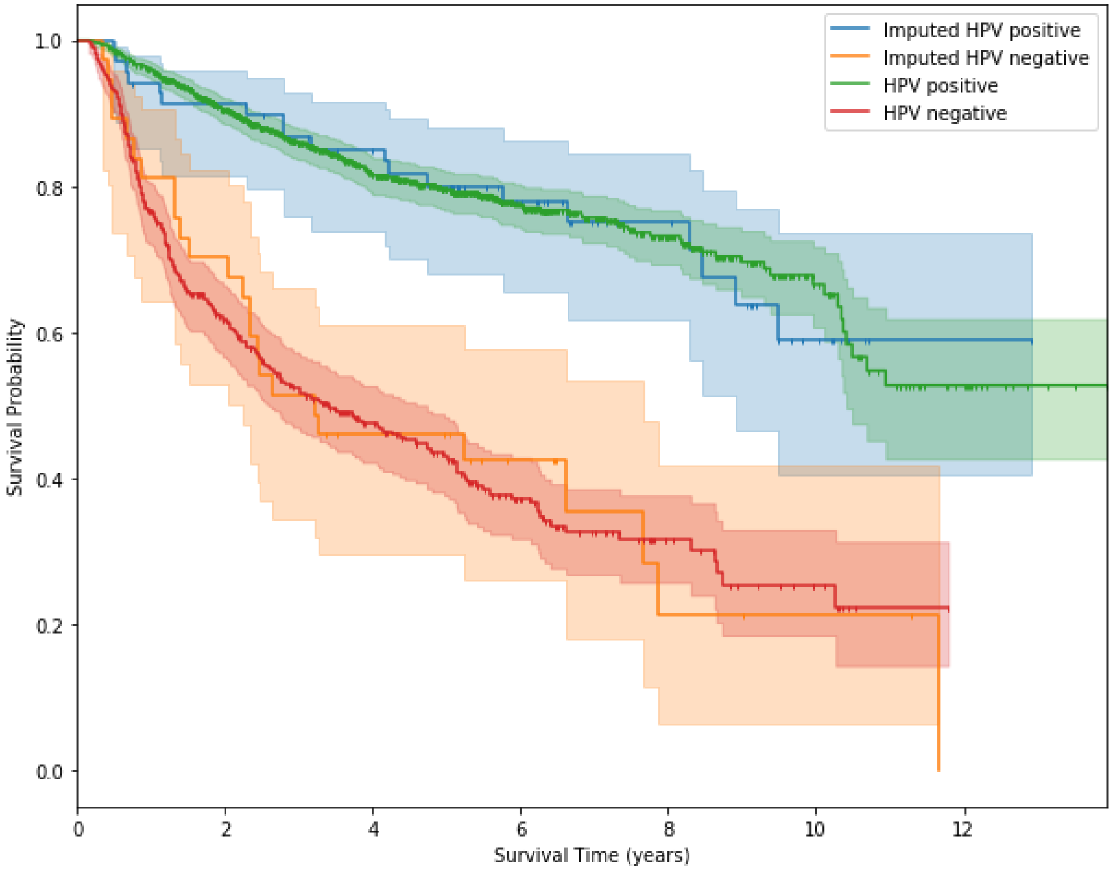

Figure 5.6: Comparison between the KM estimator curves for patients with known HPV status and those with imputed HPV status

Therefore, it is beneficial to use patient's CT scans whenever there is the chance to improve HPV status imputation.

As a proof of concept, the KM estimator curves for the RADCURE patients with known HPV status were compared with those obtained by imputing the HPV status in patients from OPC1 and OPC2 that were untested (and therefore not part of either the training or the validation dataset as explained in chapter 4). Figure 5.6 presents an overlay of the real KM estimator curves depending on HPV status and the KM estimator curves for the patients with imputed HPV status. While the confidence intervals are reasonably wide due to the low amount of untested patients with available radiomic data, the behaviours are evidently similar between the imputed and the tested group, further showcasing the model's prognostic and differentiation capabilities and demonstrating the usefulness of radiomic data in HPV status imputation.

5.2 Results on OPC3

Having validated in OPC2 and analyzed both the weights of the model and the imputation quality, the model was validated again in OPC3. The second validation was expected to give approximately the same performance, given that both OPC2 and OPC3 were obtained from the RADCURE superset, but with a tighter confidence interval due to the bigger amount of patients in OPC3. However, that was not the case, as will be shown in the next subsections.

5.2.1 GTV vs CTV

It is common practise to define more than one CTV contour, depending on how much irradiation is going to be used in each zone. Patients of OPC2 only had their CTV70 contours, which are those that cover the biggest volume off all the CTVs. In the OPC3 dataset most patients had more than one CTV contour stored and their respective PyRadiomic features. Initial studies showed no significant difference in performance between different variations of CTVs (CTV70, CTV64, CTV50, etc.), so CTV70 was chosen for the sake of consistency and to enable comparison with the results from OPC2. Figure 5.7 shows the difference in performance between using CTV70 and GTV features in the models that have both PyRadiomics and clinical data as inputs. While some of the models still perform better using CTV than GTV, the difference is very minor in the neural network and the logistic regression models. Only the random forest model shows a significant improvement, and as a consequence the ensemble model also favors CTV contours. Overall, the results are moderately similar to those in OPC2, but quite less dramatic.

5.2.2 PyRadiomics vs deep extracted features

Similarly to what happened when validating in OPC2, the deep feature extractor with the best results was the ResNet101 pretrained model. Its results and how they compare to using PyRadiomics are shown in figure 5.8. While the general behaviour observed previously still holds, it is worth mentioning that the difference between models using PyRadiomics and those using ResNet101 has shrunk significantly when combining those features with clinical ones. However, given the performance difference between the purely ResNet101 model and the purely PyRadiomics model, the shrinkage is better explained by the reduced difference

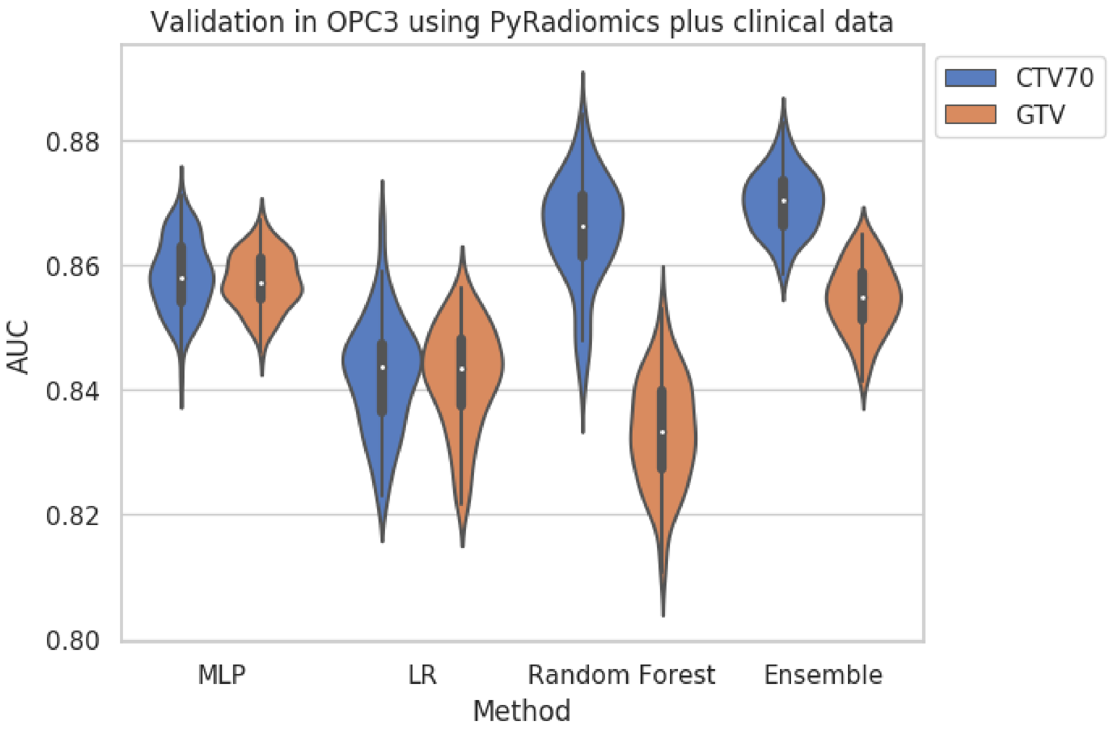

Figure 5.7: Comparison between using GTV and CTV in models that have clinical and PyRadiomics data as input

between purely clinical models and those that combine clinical features with imaging features. Since imaging features do not increase the performance by much (even reducing it in some cases), they are comparatively closer between them. In other words, the fact that the clinical plus PyRadiomics model is closer in performance to the clinical plus Resnet101 model than when validating in OPC2 is not due to an improvement in the performance of ResNet features, but is rather caused by the imaging features (both PyRadiomics and ResNet101) failing to significantly improve performance when combined with clinical ones. The issue of imaging features not boosting the performance of the clinical model will be further discussed in the next subsection.

5.2.3 Different kinds of inputs

In figure 5.9 the performance of the different models and input features is compared. The results are inconsistent in many regards to the ones obtained when validating on OPC2. First of all, the AUCs are overall lower. While the peak performance in OPC2 was of 0.91 AUC, in OPC3 the best one is 0.87. Secondly, the best performing model is the ensemble followed by random forest, while logistic regression (the best performing model in OPC2) is the one obtaining the worst results in OPC3. Conversely, random forest was the model with the lowest AUC in OPC2. Finally, the gap between the performance of the purely clinical models and the models using both clinical and radiomic features is not statistically significant. While the AUC of the ensemble using both sets of inputs is the highest at 0.87 [95% CI (0.86 - 0.88)], the ensemble model, logistic regression and neural net using only clinical features all get AUC scores that are not statistically different from it. A compilation of all the AUCs obtained using CTV is presented in table 5.5.

Even though purely PyRadiomics models still perform significantly better than random guessing, proving that radiomic features are predictive of HPV status, the observed improvement in OPC2 when combining both clinical and radiomic features is not apparent when validating on the OPC3 dataset.

This results was very unexpected, given that if OPC2 and OPC3 were sampled from the same superset and their features equally extracted there should be no significant difference in the performances obtained when validating on either one of them. The same analysis described in this subsection was performed using only those patients from OPC3 that were diagnosed after 2011, to account for the possible concept drift caused by year of diagnostic. However,

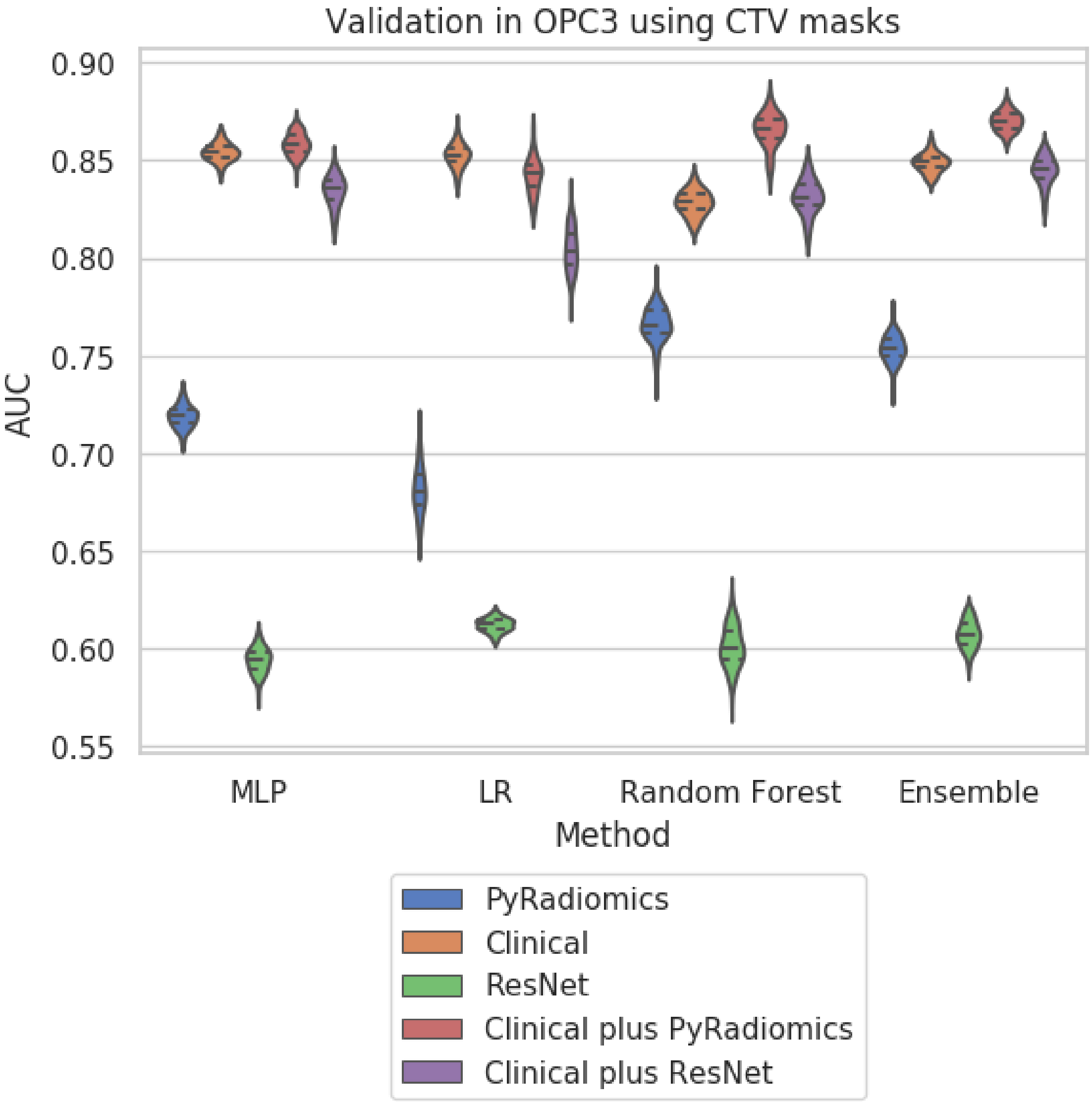

Figure 5.8: Comparison on between using ResNet101 and PyRadiomics features in models that have CTV masks

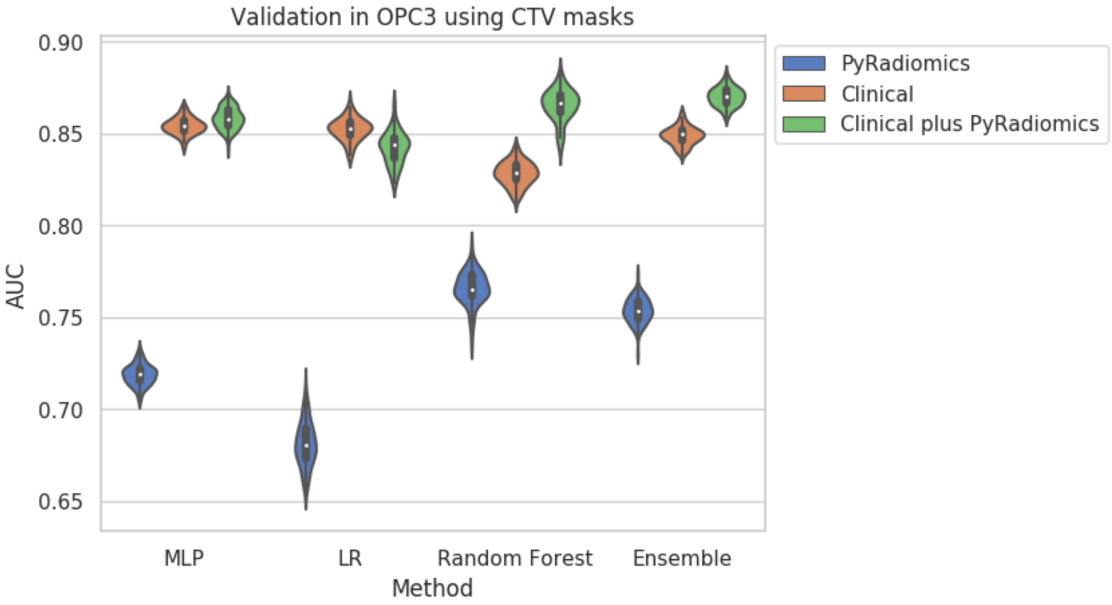

Figure 5.9: AUC of the different models built depending on the input

Mean AUC of the models in OPC3 and their 95% CI			
	PyRadiomics	**Clinical**	**Clinical plus PyRadiomics**
MLP	0.72 (0.71 - 0.73)	0.85 (0.84 - 0.86)	0.86 (0.85 - 0.87)
Logistic regression	0.68 (0.66 - 0.70)	0.85 (0.84 - 0.87)	0.84 (0.82 - 0.86)
Random forest	0.77 (0.75 - 0.78)	0.83 (0.82 - 0.84)	0.87 (0.85 - 0.88)
Ensemble	0.75 (0.74 - 0.77)	0.85 (0.84 - 0.86)	0.87 (0.86 - 0.88)

Table 5.5: AUC values for the different combinations of model and kind of input.

performance of the model showed no significant differences between using all of OPC3 or this subset of patients.

The next chapter will be dedicated to analysing the possible reasons that may have caused an inconsistency in the datasets.

Chapter 6

Posterior analysis

As mentioned in the previous chapter, results obtained in OPC2 and OPC3 are extremely inconsistent. To summarize, while the clinical models obtain relatively similar performances (0.89 best AUC in OPC2 and 0.85 best AUC in OPC3), the purely imaging models performed significantly worse in OPC3 than in OPC2. That resulted in a decrease in the gap between clinical models and hybrid models, making their performances not statistically different. To understand what caused this effect, different studies were performed on the data from OPC2 and OPC3 to detect if there was some signal of selection bias. While there is the known variation in year of diagnostic, the fact that performance of the model when validating in all of OPC3 and in only the patients of OPC3 that were diagnosed after 2011 does not change indicates that the bias is not due to this particular factor. Therefore, all the following studies were performed using all of OPC3.

The first approach was to train a model to distinguish OPC2 and OPC3. While it should be infeasible if the patients for each set had consistent feature distributions, the results in the validation phase strongly hint at some inherent discrepancies. The models used were logistic regression, neural network, random forest and an ensemble of these three models, the same as in the previous study. Out of the 684 patients (185 from OPC2 and 499 from OPC3) a total of 137, representing roughly 20% of the dataset, were randomly selected to serve as a validation set, making sure to preserve class proportions. The models were trained using only PyRadiomics features, only clinical features and a combination of both. The AUCs obtained are shown in table 6.1. Even though comparing different machine learning models is not they aim of this table, it is worth mentioning that ensemble has the best mean AUC for all input types, although that difference is not significant. The confidence intervals are somewhat wide, but two important facts can be observed. First of all, the model that uses only PyRadiomic obtains a higher AUC than the one using only clinical data, hinting at a greater difference in imaging features than in clinical features. However, due to the performances not being statistically different in the 95% CI, no definitive conclusion comparing the stability of PyRadiomics and clinical features can be obtained. Secondly, and most importantly, all of the models presented obtain better results than random guessing. This clearly proves that there's a difference in patients in OPC2 and OPC3, given that machine learning models are capable of distinguishing them. That is a crucial finding, and confirms the hypothesis formulated after comparing the validation results in both sets.

While the difference between OPC2 and OPC3 is enough to disprove the assumption that validation on them should yield similar results, it also motivates further research on the nature of that difference. To obtain a first indicator, the weights of the logistic regression models built to separate them were examined following the method used in subsection 5.1.4. A total of 7 PyRadiomics features had relevant weights. Those features were *wavelet-LLH-firstorder-Mean*, *original-glcm-MaximumProbability*, *wavelet-HLL-glcm-Correlation*, *lbp-3D-k-ngtdm-Busyness*, *original-shape-Sphericity*, *square-ngtdm-Busyness* and *exponential-ngtdm-Busyness*. Barplots comparing the values of each one of them in OPC1, OPC2 and OPC3 are shown in figure 6.1. Some of the plots show a significant difference in the distribution of OPC1, OPC2 and OPC3.

To assess the differences more accurately, the Kolmogorov-Smirnov 2-sided test was performed for each PyRadiomics feature between OPC1 and OPC2, OPC2 and OPC3 and OPC1 and OPC3. This is a two-sided test for the null hypothesis that 2 independent samples are drawn from the same continuous distribution. In this case, this test serves to determine in which cases there's a significant difference in the distribution of the features. *Square-ngtdm-Busyness*, *original-shape-Sphericity* and *wavelet-HLL-glcm-Correlation* have $pvalue < 0.05$ for all three pairs, meaning that the distributions in OPC1, OPC2 and OPC3 for these variables are all different. In the case of *wavelet-LLH-firstorder-Mean*, the only statistically significant difference is between OPC1 and OPC3. Conversely, the only statistically significant difference for *lbp-3D-k-ngtdm-Busyness* is between OPC1 and OPC2. *Original-glcm-MaximumProbability* shows significant difference between OPC1 and OPC2 and between OPC2 and OPC3, and finally *exponential-ngtdm-Busyness* shows significant difference between OPC1 and OPC2 and between OPC1 and OPC3. While this whole analysis is quite cumbersome, it illustrates the disparities amongst the three datasets. Not only are there many features that have different distributions in OPC2 and OPC3, but OPC1 is also largely inconsistent with the previous two.

The same study is performed for clinical variables. The features used for prediction by the models are age, sex and subsite. The age distributions for each dataset are shown in 6.2. While they seem almost identical, the K-S 2 sample test has a significant pvalue for the OPC2 and OPC3 pair. In the case of sex and subsite, they are categorical variables, so the K-S test can not be performed. Instead a chi-square test of independence of variables in a contingency table is used, which tests for independence between the observed frequencies in a contingency table. In this case, the rows of the contingency tables will be each one of the three datasets, and the columns will be each possible category. For sex, OPC1 has a 20% of female patients, in OPC2 they represent 13% of the dataset and in OPC3 female patients are 20% of the total. While it appears that the numbers in OPC2 are reasonably lower, the dataset is also smaller, making variability greater. Overall, the chi-squared test fails to reject the null-hypothesis of independence with a pvalue of 0.10, meaning that there is no significant difference in the distributions. In the case of subsite, the chi-square test has a $pvalue < 0.05$, rejecting the null hypothesis that distribution is independent of the dataset. Therefore there is also a significant difference in the subsites depending on the dataset, and in fact between all pairs.

To summarize, while their cause is still unclear, differences are identified between OPC1, OPC2 and OPC3 in terms of data distribution both in clinical and in imaging features. Furthermore, the differences in imaging features are substantially more abundant and severe. A total of seven out of eighteen imaging features were found to be inconsistent between datasets, with some of them not being consistent amongst any of the three pairs. Comparatively, only two of the clinical features were inconsistent, with only one case of inconsistency across all pairs. This factor may partially explain why the clinical performance had less variance when validating in OPC2 or OPC3 compared to the PyRadiomics models.

AUC of the model to distinguish in OPC2 and OPC3 and its 95% CI			
	PyRadiomics	Clinical	Clinical plus PyRadiomics
MLP	0.71 (0.65 - 0.79)	0.65 (0.55 - 0.75)	0.74 (0.66 - 0.82)
Logistic regression	0.71 (0.63 - 0.80)	0.65 (0.57 - 0.73)	0.74 (0.66 - 0.83)
Random forest	0.72 (0.63 - 0.81)	0.65 (0.56 - 0.73)	0.75 (0.67 - 0.83)
Ensemble	0.73 (0.66 - 0.81)	0.66 (0.57 - 0.74)	0.76 (0.68 - 0.83)

Table 6.1: AUC values obtained depending on the model and the input type.

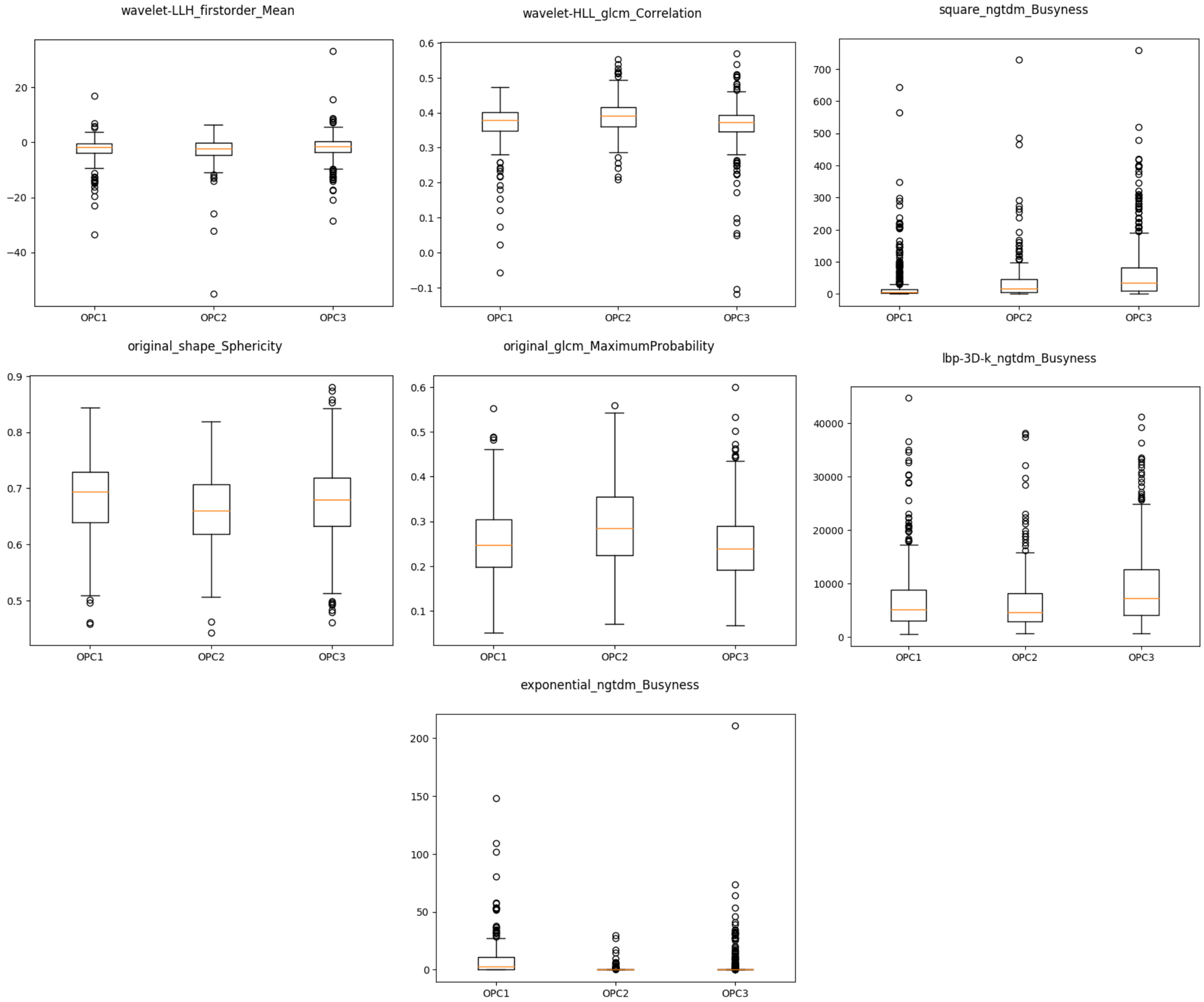

Figure 6.1: Boxplots of the seven PyRadiomic features that were used by the models to distinguish OPC2 and OPC3

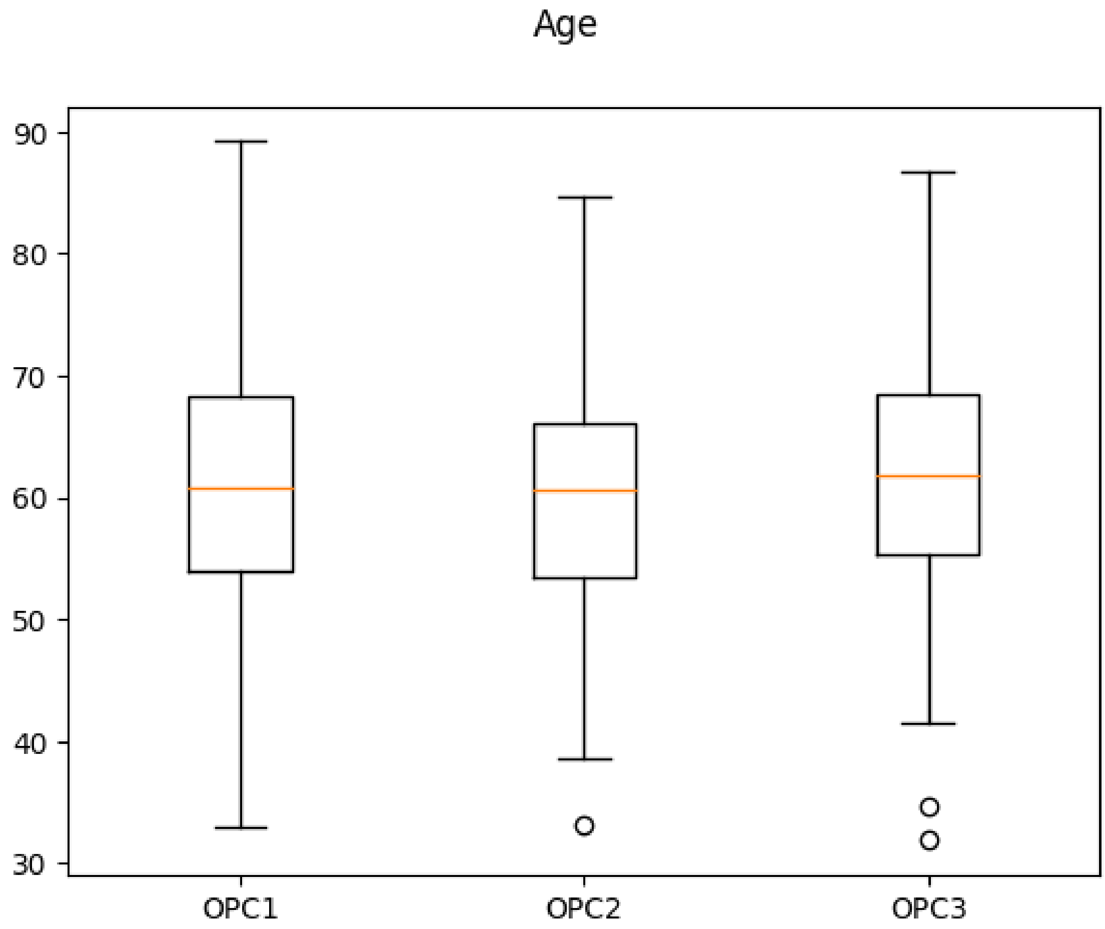

Figure 6.2: Boxplots of the age distribution in each dataset

Chapter 7

Conclusions and future work

This study was initially aimed at improving imputation of HPV (p16) status of oropharyngeal cancer patients. To do it, three different datasets were used. The two main objectives were building a model that was superior to purely clinical ones by using imaging features, and exploring its imputation capability using prognosis as an indirect validation. It was also important to answer questions about what imaging features to use, as well as what clinical masks. While results were very promising in the first validation phase, the second validation phase resulted in improvements in performance when using hybrid data that were not statistically significant. Because of that, the two validation datasets were examined to assess any inconsistencies that may have caused the difference in performance values. Different models consistently predicted the dataset to which patients belonged, showing that feature distributions were not indistinguishable between OPC2 and OPC3. This posterior analysis did not find the exact reason for that drift, although many factors could have caused it, such as the use of different settings when extracting features for OPC2 and OPC3, different versions of the imaging program or bias in the selection of the patients for each dataset, to name a few examples. While a number of reasons may be the ones that result in different features, one important finding was that clinical features were more consistent than imaging ones, both because they did not vary as much between datasets and because even when their distribution varied the effect they had on the prediction of HPV status was almost unchanged. That consistency is one of the greatest strengths of clinical variables, making them more reliable for long term predictions and less prone to be erroneous as years pass of imaging protocols change, an issue that radiomic features currently face.

Moving forward, there are a number of analysis to be performed that would help better understand the phenomenon that caused the inconsistencies between OPC1, OPC2 and OPC3. The main one would be identifying the exact cause of this effect, by contacting the different research staff that was worked in the imaging features extraction and the patient selection. Once found, appropriate measures could be implemented to either avoid that happening again or make sure that the consistency of datasets was known a priori by researchers. Secondly, ranking patients on how consistent their data was with that of patients in other datasets might reveal different profiles of patients that are more and less prone to be affected by changes in imaging features. This would add a confidence value to predictions, indicating when the model is most reliable and when it is under the risk of predicting incorrectly due to a drift in the data. Finally, once all the data was standardized and comparable, a new attempt at improving HPV (p16) imputation in oropharyngeal cancer by using imaging features could be made. Furthermore, the performance of models built for oropharyngeal cancer could be assessed for other kinds of head and neck cancer to find how well the models generalized.

Although strictly speaking this study did not achieve any of its initial goals, valuable insight on the current issues of radiomic features was gained, and it provides both a cautionary tale for future studies and different paths to eventually integrate imaging data in not only HPV prediction models, but in most medical models that use radiomic features.

Bibliography

[1] N. G. Burnet, S. J. Thomas, K. E. Burton, and S. J. Jefferies. Defining the tumour and target volumes for radiotherapy, Oct 2004. URL here.

[2] K. S. Davis, C. M. Lim, D. A. Clump, D. E. Heron, J. P. Ohr, S. Kim, U. Duvvuri, J. T. Johnson, and R. L. Ferris. Tumor volume as a predictor of survival in human papillomavirus-positive oropharyngeal cancer, Apr 2016. URL here.

[3] J. J. v. Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, H. J. Aerts, and et al. Computational radiomics system to decode the radiographic phenotype, Nov 2017. URL here.

[4] S. Habbous, K. P. Chu, X. Qiu, A. La Delfa, L. T. G. Harland, E. Fadhel, A. Hui, B. Perez-Ordonez, I. Weinreb, F.-F. Liu, and et al. The changing incidence of human papillomavirus-associated oropharyngeal cancer using multiple imputation from 2000 to 2010 at a comprehensive cancer centre, Dec 2013. URL here.

[5] S. Habbous, K. P. Chu, H. Lau, M. Schorr, M. Belayneh, M. N. Ha, S. Murray, B. O'Sullivan, S. H. Huang, S. Snow, and et al. Human papillomavirus in oropharyngeal cancer in canada: analysis of 5 comprehensive cancer centres using multiple imputation, Aug 2017. URL here.

[6] D. Jay, Simon, Olsen, Nehme, Bontempi, and Benjamin. mrmre: an r package for parallelized mrmr ensemble feature selection, Jul 2013. URL here.

[7] R. T. Leijenaar, M. Bogowicz, A. Jochems, F. J. Hoebers, F. W. Wesseling, S. H. Huang, B. Chan, J. N. Waldron, B. O'Sullivan, D. Rietveld, and et al. Development and validation of a radiomic signature to predict hpv (p16) status from standard ct imaging: a multicenter study, Jun 2018. URL here.

[8] J. J. Nappi, T. Hironaka, D. Regge, and H. Yoshida. Deep transfer learning of virtual

endoluminal views for the detection of polyps in ct colonography. In *Medical Imaging 2016: Computer-Aided Diagnosis*, volume 9785, page 97852B. International Society for Optics and Photonics, 2016.

[9] W. Owadally, C. Hurt, H. Timmins, E. Parsons, S. Townsend, J. Patterson, K. Hutcheson, N. Powell, M. Beasley, N. Palaniappan, and et al. Pathos: a phase ii/iii trial of risk-stratified, reduced intensity adjuvant treatment in patients undergoing transoral surgery for human papillomavirus (hpv) positive oropharyngeal cancer, Aug 2015. URL here.

[10] F. Petrelli, E. Sarti, and S. Barni. Predictive value of human papillomavirus in oropharyngeal carcinoma treated with radiotherapy: An updated systematic review and meta-analysis of 30 trials, May 2014. URL here.

[11] H. Ravishankar, P. Sudhakar, R. Venkataramani, S. Thiruvenkadam, P. Annangi, N. Babu, and V. Vaidya. Understanding the mechanisms of deep transfer learning for medical images. In *Deep Learning and Data Labeling for Medical Applications*, pages 188–196. Springer, 2016.

[12] S. E. Samuels, Y. Tao, T. Lyden, M. Haxer, M. Spector, K. M. Malloy, M. E. Prince, C. R. Bradford, F. P. Worden, M. Schipper, and et al. Comparisons of dysphagia and quality of life (qol) in comparable patients with hpv-positive oropharyngeal cancer receiving chemo-irradiation or cetuximab-irradiation, Mar 2016. URL here.

[13] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

[14] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, H. J. W. L. Aerts, and et al. Computational radiomics system to decode the radiographic phenotype, Nov 2017. URL here.

[15] Vanhoucke, Vincent, Sergey, Jonathon, and Zbigniew. Rethinking the inception architecture for computer vision, Dec 2015. URL here.

[16] D. J. Weatherspoon, A. Chattopadhyay, S. Boroumand, and I. Garcia. Oral cavity and

oropharyngeal cancer incidence trends and disparities in the united states: 2000-2010, Aug 2015. URL here.

[17] M. Welliver, W. Yuh, J. Fielding, K. Macura, Z. Huang, A. Ayan, F. Backes, G. Jia, M. Moshiri, J. Zhang, and N. Mayr. Imaging across the life span: Innovations in imaging and therapy for gynecologic cancer. *Radiographics : a review publication of the Radiological Society of North America, Inc*, 34:1062–1081, 07 2014. doi: 10.1148/rg.344130099.

[18] Y. Xu, A. Hosny, R. Zeleznik, C. Parmar, T. Coroller, I. Franco, R. H. Mak, and H. J. Aerts. Deep learning predicts lung cancer treatment response from serial medical imaging. *Clinical Cancer Research*, 2019. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-18-2495. URL here.

[19] Zhang, Ren, Sun, and Jian. Deep residual learning for image recognition, Dec 2015. URL here.