POLITECNICO DI MILANO
DEPARTMENT OF ELECTRONICS, INFORMATION AND BIOENGINEERING
DOCTORAL PROGRAMME IN BIOENGINEERING

MRI-BASED RADIOMIC ANALYSIS OF RARE TUMORS: OPTIMIZATION OF A WORKFLOW FOR RETROSPECTIVE AND MULTICENTRIC STUDIES

Doctoral Dissertation of:
Marco Bologna

Supervisor:
Prof. Luca Mainardi

Tutor:
Prof. Dario Gastaldi

The Chair of the Doctoral Program:
Prof. Andrea Aliverti

Academic year 2019-2020 – Cycle XXXII

Contents

Abstract

The main purpose of this thesis was the optimization of a workflow for the radiomic analysis of Magnetic Resonance Images (MRI) acquired with uncontrolled image acquisition protocols. The secondary aim was the application of the optimized workflow to build prognostic models for Overall Survival (OS) for Head and Neck Cancer (HNC) and Soft Tissue Sarcoma (STS), in order to show the feasibility of using radiomics in multicentric and/or multiprotocol datasets.

The first part of the work focused on a series of stability analyses performed using a virtual phantom (BrainWeb). The aim of these studies was two-fold: 1) to evaluate the effect of image preprocessing on the stability to imaging-related variability; 2) to select the features that are stable to such variations, in order to use them for the following analysis. Intensity standardization, image denoising, voxel size resampling and bias field correction were considered as potentially useful preprocessing steps. Intraclass Correlation Coefficient (ICC) was used to quantify features stability, and features with ICC>0.75 were considered stable. All the preprocessing steps (Gaussian filtering, N4ITK Bias field correction, B-spline spatial resampling and intensity standardization) had positive effects in increasing the stability of radiomic features. When including all the previous preprocessing step, 550 features, based on both T1-weighted (T1w) and T2-weighted (T2w) MRI were identified as stable, out of a total of 1072 (536 per image type).

Stability to uncertainties of the region of interest (ROI) was also investigated. Two sources of variability were considered: multiple segmentation and geometrical transformations of the ROI. Both tests were performed on real images of STS and HNC, considering T1w, T2w and apparent diffusion coefficient maps (ADC). In each test, features with ICC>0.75 were

considered stable. In total, 701 and 1057 features out of 1608 were stable for HNC and STS respectively. After properly combining these stable features sets with the results previously obtained on the BrainWeb dataset, the number of stable features was reduced to 410 and 617. These two sets of features were used for successive studies.

The postprocessing of the features was also optimized. In particular, features normalization and feature selection/dimensionality reduction were optimized in order to maximize the performance of a Cox proportional hazard regression model. Four different features normalization algorithms and 2 different features selection pipelines were tested. Harrell's C-index was used to quantify the models performance. It was found that the combination of Z-score normalization and a series of different features selection (pairwise correlation and cross-validated Multivariate-Cox) lead to the best performance in a retrospective multicentric HNC dataset (C-index 0.67).

After the optimization based on the results of the previous analyses, the radiomic workflow was used to identify signatures that were prognostic of OS in HNC and STS. In HNC, a five-features radiomic signature had a good prognostic value in both cross-validation (C-index 0.67) and independent validation (C-index 0.63) and in both cases the radiomic features improved the prognosis when added to the clinical ones (from 0.67 to 0.69 and from 0.69 to 0.72 for the cross-validation and independent validation respectively). Similar results were found after cross-validation of a radiomic model in STS (C-index 0.74, 0.74 and 0.78 for the radiomic, clinical and combined model).

The results show that with the right processing, radiomic analysis from non-standardized images is possible and provides a consistent improvement in the prognostic performance of survival model for OS.

List of abbreviations

ADC Apparent Diffusion Coefficient

AJCC American Joint Committee on Cancer

ANOVA ANalysis Of VAriance

AOP Azienda Ospedaliero-universitaria di Parma

CCC Concordance Correlation Coefficient

CE-T1w Contast Enhanced T1-weighted

CI Confidence Interval

CIS Carcinoma In Situ

CT Computed Tomography

DFS Disease-Free Survival

DWI Diffusion-Weighted Imaging

DWT Discrete Wavelet Transform

EPI Echo-Planar Imaging

ETL Echo-Train Length

FDG FlouroDeoxyGlucose

FDR False Discovery Rate

FFT Fast Fourier Transform

FID Free Induction Decay

FOS First Order Statistics

GLCM Grey Level Co-occurrence Matrix

GLDM Grey Level Dependence Matrix

GLRLM Grey Level Run Length Matrix

GLSZM Grey Level Size Zone Matrix

HNC Head and Neck Cancer

HPV Human PapillomaVirus

HR Hazard Ratio

HU Hounsfield Unit

IBSI Imaging Biomarker Standardization Initiative

ICC Intra-class Correlation Coefficient

INT Istituto Nazionale dei Tumori

INU Intensity Non-Uniformities

LASSO Least Absolute Shrinkage and Selection Operator

MAASTRO MAASTricht Radiation Oncology clinic

MRI Magnetic Resonance Imaging

NGTDM Neighboring Grey Tone Difference Matrix

NMR Nuclear Magnetic Resonance

OS Overall Survival

PCA Principal Component Analysis

PD Proton Density

PDw Proton Density-weighted

PET Positron Emission Tomography

RF Radio-Frequency

ROI Region Of Interest

SCB Spedali Civili di Brescia

SE Spin-Echo

SS Shape and Size

STS Soft Tissue Sarcoma

SUV Standardized Uptake Value

T1w T1-weighted

T2w T2-weighted

TE Time of Echo

TNM Tumor Node Metastasis

TR Time of Repetition

TSE Turbo Spin-Echo

UDUS Heinrich-Heine-Universität Düsseldorf

UICC Union for International Cancer Control

ULM university hospital of ULM

VUMC Vrije Universiteit Medical Center

WHO World Health Organization

CHAPTER 1

Introduction

In this chapter, an introduction to aims and structure of the thesis is presented, together with a brief outline on the clinical context in which this thesis develops.

1.1 Personalized medicine and radiomics

Personalized medicine is a new paradigm in medicine which consists in tailoring a therapy for a particular disease according to the characteristics of the single patient, taking individual variability into account [1]. The concept is not totally new: blood typing for transfusions is a first example of personalized medicine. However, the prospect of applying this concept broadly has been dramatically improved by the recent development of large-scale biologic databases [1] and by the development of the so-called "omics". "Omics" is a term referred to scientific disciplines in which high throughput extraction of features from different biological sources (i.e. genes, proteins, metabolites, etc...) is used to gather additional information and to provide biological signatures that may further cluster the patients, thus improving the efficacy of a treatment or the stratification of the prognosis [2].

Of particular interest for this thesis is the application of personalized medicine in cancer treatment. As a matter of fact, cancer is still the first cause of death worldwide [3] and, given the high heterogeneity of the disease, it is very difficult to treat it with standardized techniques. Application of personalized medicine in oncology is therefore of primary importance because it may lead to an improvement in cancer management and treatment, with increasing life expectancy for the patients. Among the "omics", genomics was the first to be used for stratifying patients in subgroups with different prognosis [4].

Although very promising, biology-related omics such as genomics have some limitations. The first limitation is that omics require biological samples as a starting point of the analysis and to get these sample biopsies are required, which are typically invasive procedures [5]. The second limitation is that the tumors are typically spatially heterogeneous and a localized biopsy may not be enough for a complete characterization. Therefore, multiple biopsies may be required, with increasing discomfort for the patient [5]. A further limitation is that omics are not currently part of the clinical practice [5] and therefore their used is associated with an increasing cost of cancer treatment.

The above mentioned issues led to the increasing interest in the field of radiomics [6, 7]. Radiomics is a new field of research in medical image analysis that involves the high-throughput extraction of quantitative imaging features with the intent of creating mineable databases from radiological images [6]. The underling hypothesis of radiomics is that the analysis of quantitative features extracted from a Region Of Interest (ROI) inside the image can provide more and better information than that of a physician, revealing predictive or prognostic associations between images and medical outcomes [6, 7].

Radiomics offers some advantages over the other omics. One advantage is the fact that radiomic features are derived from non-invasive imaging techniques, such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT) or Positron Emission Tomography (PET). This is in contrast with other omics, that require biopsies to get the tissue samples to be analyzed. Another advantage over traditional omics is that radiomic features, being extracted from the entire ROI of the tumor, account for tumor spatial heterogeneity [8]. Last, radiomics requires only data that are already part of the cancer management process and is therefore cost-free in terms of data acquisition, unlike the other omics that require additional tests. These are the reasons why radiomics has attracted so many attention in the last decade and the number of works in radiomics has been growing exponen-

tially [9].

Studies about the application of radiomics to clinical datasets have already been presented in literature for CT, MRI and PET [5, 10, 11]. However, the main limitation of these studies is the fact that the training of the radiomic model is performed on datasets coming from the same center and with strict image acquisition protocols [5, 10, 11]. The effect of image acquisition conditions on the measured radiomic features is still under investigation and understanding the effect of acquisition-related variability on the performance of radiomic signatures is still an open challenge for radiomics [6]. This is true in particular for MRI, in which the values of signal measured may strongly depend on the conditions in which the signal was acquired [6]. Radiomic features harmonization is of particular importance for all those rare tumors (such as head and neck cancer and soft tissue sarcoma, which are the focus of this thesis) for which the only way to collect a large number of patients necessary to train a radiomic model is to put together data coming from difference centers or from different retrospective studies, for which a strict standardization in the image acquisition parameters is typically not available.

1.2 Aims and objectives

The aim of this thesis was to define a workflow to create MRI-based radiomic signatures prognostic for Overall Survival (OS) for two categories of rare tumors: Head and Neck Cancer (HNC) and Soft Tissue Sarcoma (STS). The focus on MRI was due to the fact that, given the high intensity contrast in soft tissues, MRI is a particularly suited techniques to provide informative images of the districts were HNC and STS usually appear (head and neck for HNC and limbs for STS), but also to the fact that MRI is the technique that is affected the most by variations in image acquisition conditions, and in which the problem of features harmonization is more challenging. However, the arguments treated in this thesis may be easily transferred, with small variations, to other imaging techniques.

The creation of prognostic radiomic signatures is a multi-step process that is schematically illustrated in Figure 1.1. To effectively apply such workflow to real clinical datasets, the critical issues related to each step must be properly addressed to avoid bias that may reduce the generalizability of the results. The studies performed for this thesis dealt with the evaluation and design of proper solution to those critical issues. In particular, the objectives of the thesis were the following:

- The evaluation of the stability of radiomic features to variation in im-

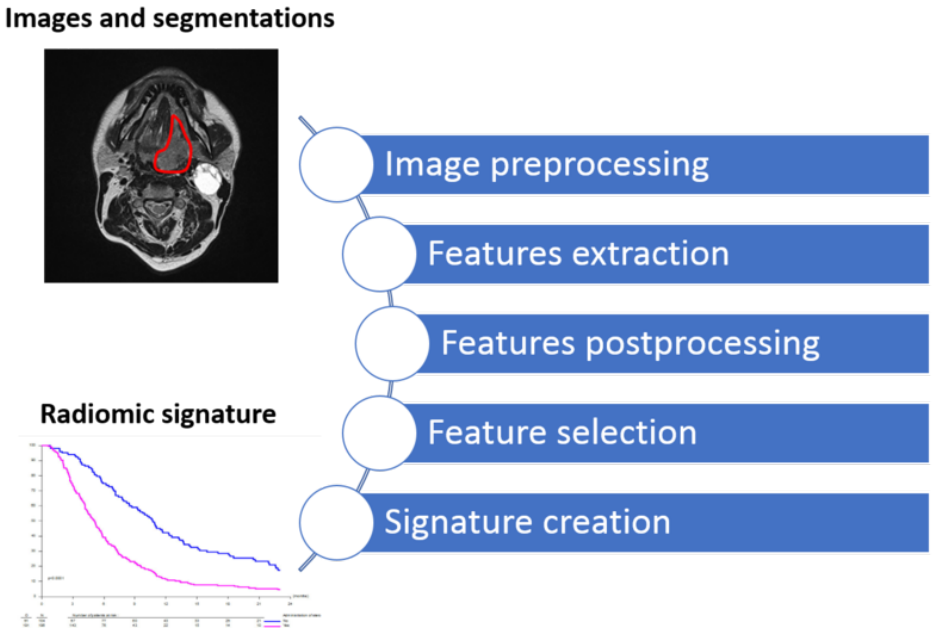

Figure 1.1: *Workflow for the creation of the radiomic signatures.*

age acquisition parameters.

- The evaluation of image preprocessing steps in reducing such variability.
- The evaluation of the stability of radiomic features to variations of the ROI.
- The identification of a set of stable radiomic features that are robust to the aforementioned sources of variability.
- The optimization of the postprocessing of the features, with focus on the selection of best features normalization and features selection algorithms.
- The validation of the results obtained with the optimized radiomic workflow.
- The evaluation of the added prognostic value of radiomics with respect to the models relying on established clinical variables.

1.3 Outline of the thesis

The different chapter of the thesis try to achieve the objectives outlined in the previous subsection. The main outline of the thesis is the following:

- Chapter 2 provides the minimal background about oncology, imaging, radiomics and machine learning that is necessary to completely understand the thesis. The reader which is already familiar with those topics may skip the chapter.
- Chapter 3 describes the experiments with a virtual MRI simulator (BrainWeb) performed to understand the effect of image preprocessing on features stability and to find a set of features that is stable to variations in image acquisition parameters.
- Chapter 4 describes the experiments performed on images of HNC ans STS patients to identify a set of radiomic features to uncertainties of the ROI.
- Chapter 5 covers the topic of features postprocessing. In particular, features normalization and features selection are described.
- Chapter 6 describes a clinical application of the radiomic workflow for the creation of a radiomic signature for OS in HNC.
- Chapter 7 describes a clinical application of the radiomic workflow for the creation of a radiomic signature for OS in STS.
- Chapter 8 gives a wrap-up of the results obtained in the thesis, also highlighting limitations and possible future developments and impact of the research.

CHAPTER 2

Background

This chapter gives a general introduction on the field of radiomics. Radiomics is a complex and multidisciplinary field involving oncology, medical imaging, machine learning and medical statistics. Although it is not possible to perfectly master all the pieces that make up a radiomic analysis, any person working on radiomics should have a minimal background in each, and the purpose of this chapter is to provide such background.

2.1 Radiomics: a general introduction

Radiomics is a new field of research in medical image analysis that involves the high-throughput extraction of quantitative imaging features with the intent of creating mineable databases from radiological images [6]. The underling hypothesis of radiomics is that the analysis of quantitative features extracted from a ROI inside the image can provide more and better information than that of a physician revealing predictive or prognostic associations between images and medical outcomes [6, 7]. If such hypothesis was proven true, radiomics would become a valuable tool to reach the goal of personalized medicine. That is the reason why the number of studies focusing on radiomics have been increasing in the last years [9].

The concept of radiomics gained particular importance after two studies performed in the last decade [12, 13]. In [12] a set of 14 qualitative imaging features from CT was able to predict 80% of the gene expression pattern in hepatocellular carcinoma. In [13], qualitative MRI features extracted from glioblastoma were able to predict immunohistochemically identified protein expression patterns. The modern concept of radiomics tries to extend those two studies by using high-number of quantitative image features accounting for tumor shape, signal intensity and texture [6].

As mentioned in Chapter 1, radiomics has several advantages over other omics. The first advantage is that does not require additional biopsies, since it is typically based on non-invasive image acquisition techniques such as CT, PET and MRI. The second advantage is that radiomics is low-cost, since the material used for radiomic analysis (i.e. the clinical images) is already acquired as part of the clinical routine. The last advantage is that radiomic can better characterize the spatial heterogeneity of the tumor, since the analysis is usually performed on the entire tumor mass, rather than on a small piece of it.

The process of radiomic analysis involves different steps, which are illustrated in Figure 2.1:

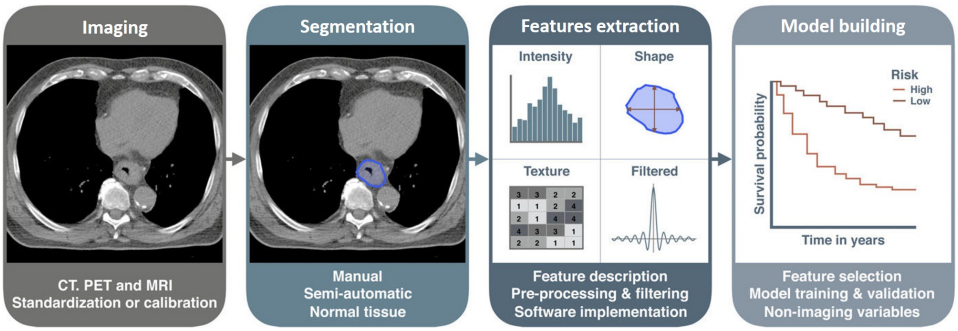

Figure 2.1: *Workflow of radiomic analysis and its four main steps: image acquisition; image segmentation; features extraction; model building. Adapted from [14].*

Image acquisition

The first step of radiomic analysis is to acquire the medical images of interest. Although radiomic analysis could potentially be applied to any imaging technique, most of research in radiomics focuses on non-invasive imaging techniques such as CT, MRI, PET, which are the most used in the clinical practice especially for oncology [14]. Other imaging techniques, like ultrasonography, may also be used [14].

Image preprocessing and segmentation

Radiomic features are typically not extracted from the entire image but from a ROI including the organ or tissue to be studied (e.g. a tumor or a metastasized lymph node). Segmentation could be semi-automatic, with partial input from the radiologist, or manual, when completely performed by the radiologist [14]. Although semi-automatic segmentation algorithms lead to a reduced inter- and intra-observer variability [15], they are not available for some districts and therefore manual segmentation is still widely used.

Segmentation is not the only operation that can be performed on the images before radiomic features extraction. Denoise filtering, discretization of the grey values, and resampling, for example, have been shown to improve the repeatability of the extracted radiomic features [16, 17].

Features extraction

Once the images have been segmented, the radiomic features can be extracted. The number features may vary from a few to several thousands depending on the study [18]. Typically radiomic features can be divided in three main categories. Shape and Size (SS) features, like volume and largest diameter, consider geometrical properties of the ROI. Intensity-based or First Order Statistics (FOS) features account for the statistical distribution of the grey levels inside the ROI. Textural features also take into account the spatial distribution of the grey level inside the ROI. Textural features are computed from different types of textural matrices, like Grey Level Co-occurrence Matrix (GLCM) [19], Grey Level Run Length Matrix (GLRLM) [20], Grey Level Size Zone Matrix (GLSZM) [21], Grey Level Dependence Matrix (GLDM) [22], Neighbouring Grey Tone Difference Matrix (NGTDM) [23]. FOS and textural features can be extracted from the original images but also from transformed images to get further details. Image transforms typically used mainly include wavelet decomposition [24] or Laplacian of Gaussian filtering [24], but other transformation such as logarithm, exponential and square root filtering may be used [24]. A more detailed explanation of the different categories of radiomic features is provided in Section 2.4.

Features processing and model development

Once the features have been extracted, they can be used to get further information for tumor characterization, e.g. by developing prognostic models. In order to do that, radiomic features have to undergo several processing

steps, such as normalization, features selection or dimensionality reduction.

Once the best features set has been defined, it can be used to develop models by traditional statistics or machine learning methods, using either supervised or unsupervised approaches.

2.2 Clinical background

2.2.1 Tumors and cancer

A tumor is a disease that is characterized by an abnormal growth of a body tissue that is caused when cells divide and grow excessively [25]. Tumor cells typically spread from the original site to other areas or organs through the lymphatic system, generating the so called metastasis (Figure 2.2), reducing the functionalities of the host organ and producing damage to the organism [26, 27]. Such tumors are called malignant (or cancerous), which is in contrast with the benignant tumors that remain confined in a specific region of the body [27].

According to the latest data provided by the World Health Organization (WHO), 18.1 millions new cancers were diagnosed in 2018, and the number of deaths due to cancer in the same year was 9.6 millions, making cancer the leading cause of death worldwide (1 death out of 6 is caused by cancer) [3,28], with men being the most affected (52% and 56% for occurrence and deaths respectively) [3]. In Italy, around 365.000 patients (54% men) are diagnosed with cancer every year and around 175.000 (54% men) people die because of the disease (around 30% of the deaths in Italy) [29, 30]. Tumors and cancers can appear in many sites, but the ones that are most frequently affected are lungs (11.6%), colon-rectum (10%), breast (11.6% in women) and prostate (7.1% in men) [3].

2.2.2 Head and neck cancer

The term head and neck cancer (HNC) is referred to the tumors that begin in the squamous cells that line the moist, mucosal surfaces inside the head and neck [31]. These squamous cell cancers are often referred to as squamous cell carcinomas of the head and neck (HNSCC). HNC can also begin in the salivary glands, but salivary gland cancers are relatively uncommon [31]. The sub-sites that are affected by HNC are oral cavity, pharynx, larynx, para-nasal sinuses, nasal cavity and salivary glands (Figure 2.3).

Worldwide, HNC accounts for more than 650,000 cases and 330,000 deaths annually, making it the sixth most common cancer [3, 32]. These

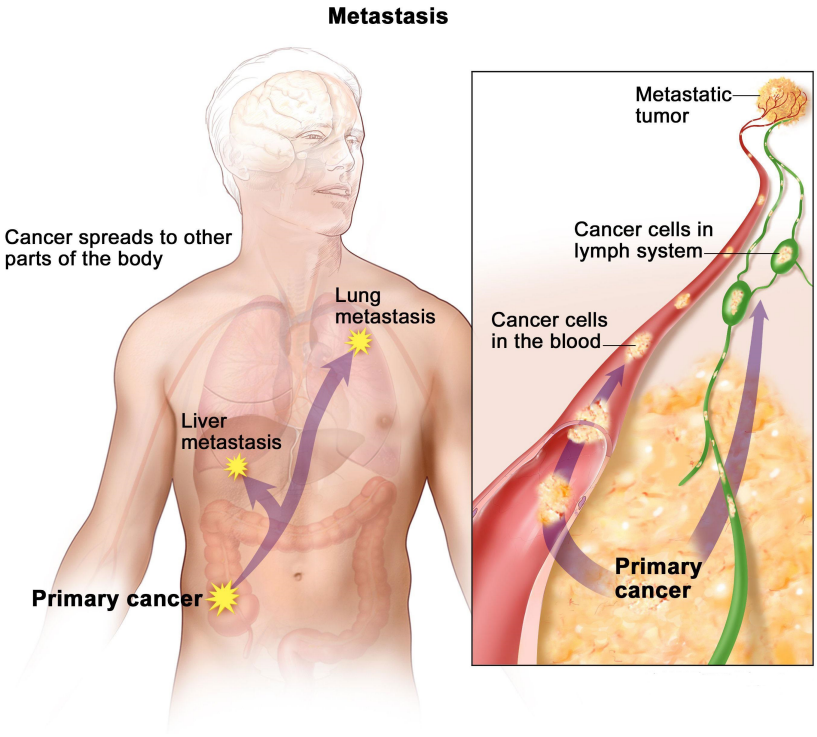

Figure 2.2: *Example of a cancer (colo-rectal) spreading through the lymphatic system generating a metastasis in a distant region of the body (liver) [26].*

cancers are more than twice as common among men as they are among women [32]. HNC are also diagnosed more often among people over age 50 than they are among younger people [31]. The most significant causes of all head and neck cancers are tobacco use and alcohol consumption, accounting for 80% of such cancers [33]. HNC, especially oropharyngeal cancer, may also be caused by infection with certain types of Human PapillomaVirus (HPV), especially HPV type 16 [34]. The number of cancers caused by HPV (HPV+) has increased in the last years [34]. HPV+ cancers have in general better prognosis than cancers caused by other factors (HPV-). Beside alcohol, smoke and HPV, there are other minor factors contributing to the insurgence of HNC, such as genetics, oral hygiene, and diet [31].

Although HNC is neither the most widespread or the most aggressive type of cancer [35], the social and psychological status after treatment remains a major concern [36]. As a matter of fact, the condition and its treatment can affect breathing, eating and communicating, and cause a change

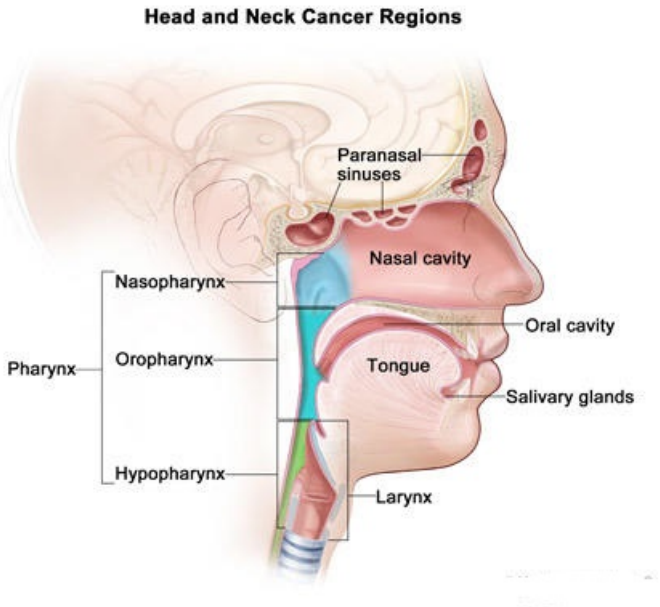

Figure 2.3: *Graphical representation of the districts that are part of the head and neck region [31].*

physical appearance, possibly leading to anxiety and depression [36]. The optimization and improvement of HNC treatment is therefore of primary importance.

2.2.3 Soft tissue sarcomas

Soft tissue sarcomas (STS) are a rare and heterogeneous group of tumors, arising in connective tissues embryologically derived from the mesenchyme [37]. There are dozens of subtypes arising from cartilage, muscle, blood vessels, nerves, and fat [37]. As Figure 2.4 shows, approximately 50% of sarcomas develop in the arms or legs with the remaining types originating in the thorax or abdomens (40%) or the head and neck area (10%) [38].

STS are rare diseases (<1% of all the tumors), with an estimated incidence of at most 4 people every 100,000, with a prevalence for men [37,39]. The median age of diagnosis for STS is 59 years. Risk factors for STS are age, genetic predisposition, presence of concurrent pathologies, exposures to chemicals and radiations, but the majority of the diagnosed STS is not directly caused by one of these risk factors [39].

The rarity of STS often reflects in a delay of the diagnosis, because STS

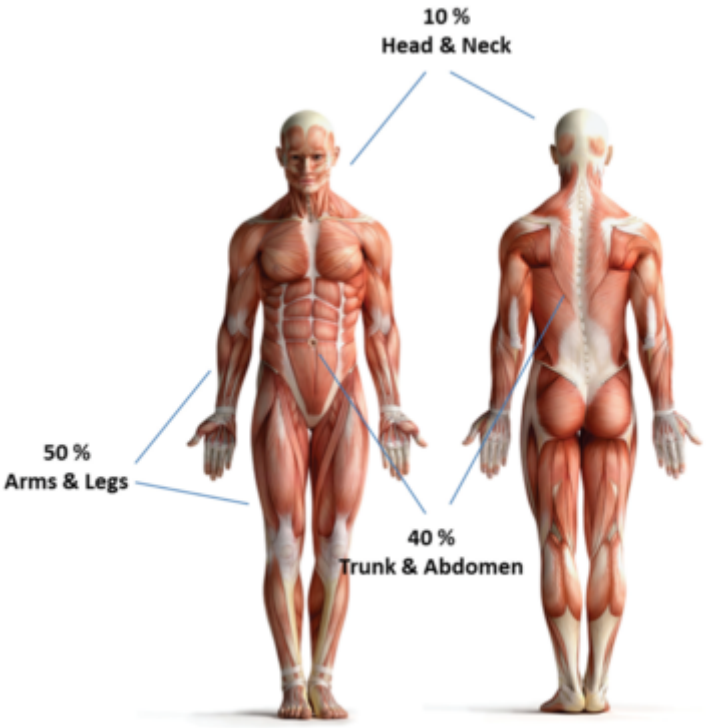

Figure 2.4: *Districts of the body affected by soft tissue sarcoma [38].*

may at first be mistaken as simple swelling due to injury. A late intervention due to a mistaken diagnosis may lead to worst prognosis after treatment. In case of limb STS it could even lead to amputation [40], with consequent reduction of the quality of life of the patients. To avoid this, early screening and proper clinical management of the pathology is required.

2.2.4 Cancer staging and grading

There are two main classification systems for tumors: tumor grading [41] and tumor staging [42].

Tumor grade is an index of the differentiation of the tumor cells and is determined through biopsy and subsequent analysis at the microscopy.Tumors with highly differentiated cells grow and spread at a slower rate compared to tumors with undifferentiated cells. The American Joint Committee on Cancer (AJCC) defined a 4 grades classification system [41]:

- **GX**: Grade cannot be assessed (undetermined grade);
- **G1**: Well differentiated (low grade);

- **G2**: Moderately differentiated (intermediate grade);
- **G3**: Poorly differentiated (high grade);
- **G4**: Undifferentiated (high grade).

Cancer stage refers to the size and/or extent of the original tumor and whether or not cancer cells have spread in the body [42]. Cancer stage is based on factors such as the location of the primary tumor, tumor size, regional lymph nodes involvement (the spread of cancer to nearby lymph nodes), and the number of tumors present. Stage can be evaluated by biopsy, lab tests for body fluid (e.g. blood, urine, etc...) and via imaging. The AJCC define the TNM classification system for cancer staging, that evaluates the three elements characterizing a cancer (primary tumor, involved lymph nodes, metastatic tumors) [41,42]. In the TNM system, the T refers to the size and extent of the main (or primary) tumor, the N refers to the the number of nearby lymph nodes that have cancer, and the M refers to whether the cancer has metastasized. The letters T, N and M are accompanied by numbers that give more details about the cancer. The following explains what the letters and numbers mean:

- **Primary tumor (T):**
  - **TX**: Main tumor cannot be measured.
  - **T0**: Main tumor cannot be found.
  - **T1, T2, T3, T4**: Refers to the size and/or extent of the main tumor. The higher the number after the T, the larger the tumor or the more it has grown into nearby tissues. T's may be further divided to provide more detail, such as T3a and T3b.
- **Regional lymph nodes (N):**
  - **NX**: Cancer in nearby lymph nodes cannot be measured.
  - **N0**: There is no cancer in nearby lymph nodes.
  - **N1, N2, N3**: Refers to the number and location of lymph nodes that contain cancer. The higher the number after the N, the more lymph nodes that contain cancer.
- **Distant metastasis (M):**
  - **MX**: Metastasis cannot be measured.
  - **M0**: Cancer has not spread to other parts of the body.

  - **M1**: Cancer has spread to other parts of the body.

Often the T, N and M combinations are grouped into five less-detailed stages, that are described as follows:

- **Stage 0**: abnormal cells are present but have not spread to nearby tissue. Also called Carcinoma In Situ (CIS); CIS is not cancer, but it may become cancer.
- **Stage 1-3**: cancer is present. The higher the number, the larger the cancer tumor and the more it has spread into nearby tissues.
- **Stage 4**:may refer to cancers that are locally very extensive (stage 4a), regionally very extensive (stage 4b) or that present distant metastases (stage 4c).

The TNM staging system has been refined over time and at present the 7^{th} and 8^{th} editions (TNM VII and TNM VIII respectively) are the ones currently used for cancer evaluation [43]. TNM VII edition was introduced by the AJCC in 2009 [43] and it has been the gold standard until 2017, when the Union for International Cancer Control (UICC) introduced the TNM VIII. In HNC, the main advantage of TNM VIII over TNM VII is that the former downstages tumors that are virus-related (like HPV+ tumors) and that usually have a better prognosis [44]. When making comparisons with tumor staging, TNM VIII is the one that was used as a reference, given its better prognostic performance in HNC, in particular in those sub-sites, like oropharynx, where HPV+ tumors are the majority [44].

2.2.5 Cancer treatment and personalized medicine

Cancer may be treated successfully if the stage is not advanced. There are several ways in which cancer could be treated and they are briefly described here [45]:

- **Surgery**: consisting of the physical removal of the tumor mass; surgery works best for solid tumors that are contained in one area, but fails when the tumor is spread throughout the body.
- **Radiotherapy**: consisting of the use of high dose of ionizing radiations to kill cancer or slow down its progression.
- **Chemotherapy**: consisting of the use of drugs to kill cancer cells; it is the only options when cancer is spread all over the body, but has many collateral effects, since it also affect the healthy cells and tissues.

- **Immunotherapy**: consisting of the use of drugs to help the immune system fighting cancer.
- **Targeted therapy**: consisting of the injection of specific molecules targeting specific components of the cancer cells, reducing their functionalities and causing their death; it is very effective but not very applicable due to higher cost.

Usually tumors are not treated with just one type of therapy but with a combination of them [46]. Moreover, while in the past the type of treatment for a particular type of cancer was standardized, nowadays the best combination of techniques for the treatment depends on the patient's prognostic group. Oncology is moving from a standardized approach to a more personalized approach with the final aim of reaching the goal of personalized medicine, i.e. a paradigm in which the treatment is tailored to the individual patients to maximize the efficacy [1]. Omics [2] may be a useful tool in this process, since they could allow to stratify patients according to their prognosis, providing additional information that the clinician can use to optimize patients treatment and follow-up.

2.3 An introduction to medical imaging

In the current section, the main imaging techniques used in radiomic analysis will be described, with particular focus on MRI, which was used in the thesis. The purpose of the current section is not to be completely exhaustive, but to provide the main characteristics of each imaging modality and to provide the minimum level of detail to understand the content of the next chapters. For more details on medical imaging the reader can refer to [47, 48].

2.3.1 Medical imaging in oncology

Medical imaging plays and important role in oncology. As a matter of fact, imaging, and non-invasive imaging techniques such as CT, PET and MRI, are used in different steps of the cancer management pipeline [49]. For example, medical imaging is used in tumor diagnosis and is the first way to assess the presence of a tumor in patients that present suspect symptoms. Also, during the surgery or radiotherapy planning, anatomical imaging (i.e. imaging that show the anatomy of the districts of the body), such as CT or MRI, is used to identify the region to operate/irradiate. Moreover, imaging is used to evaluate the effectiveness of a treatment like chemotherapy or

radiotherapy. Last, imaging is used in follow-up exams to promptly detect the recurrence of a tumor.

Given its predominant role in the whole cancer management process, it is easy to understand why so much attention has been paid to non-invasive medical imaging and radiomics lately [8].

2.3.2 Computed Tomography

Computed Tomography, or CT, is an imaging technique that uses X-ray radiation to draw tomographic images (or slices) of the body [47, 48]. The physical principle behind CT is the same of radiography, i.e. each tissue has different absorption to the X-ray radiation and the different tissues can be distinguished based on such differences [47, 48]. When a X-ray beam or fan (Figure 2.5) is emitted by an X-ray tube, the attenuated signal can be obtained by placing some X-ray detectors on the other side of the emitter. By combining the attenuation signals from multiple projection taken from different angles, a tomographic image of the original object can be obtained (Figure 2.5). The mathematical explanation of how CT images are reconstructed is beyond the scope of this thesis, but the reader may refer to [47, 48, 50].

The value of intensity in each CT image is proportional to a measure of attenuation, called Hounsfield Unit (HU), which is typically related to tissue density (the higher the density, the higher the HU and the attenuation). CT imaging allows to distinguish the tissues by their density and it is therefore an anatomical imaging technique.

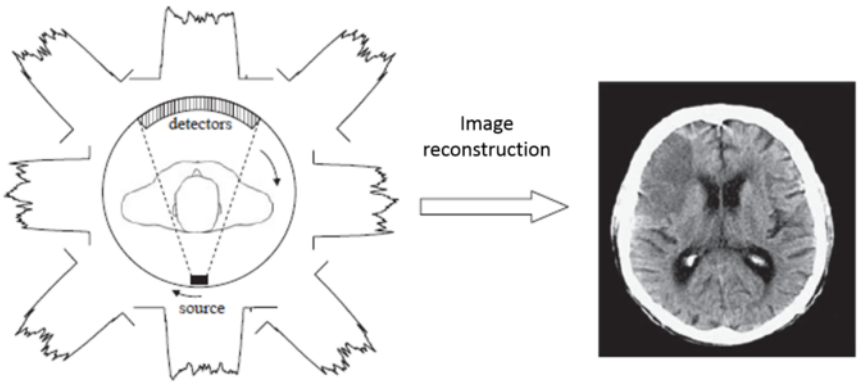

Figure 2.5: *The acquisition of multiple attenuation signals obtained by moving the X-ray source and detectors is used to reconstructed a Computed Tomography image. Adapted from [48]*

Beside non-invasiveness, CT has many advantages like the velocity of acquisition and the possibility of obtaining high resolution images. Moreover, the fact of being a quantitative imaging technique and the fact that CT is one of the most used imaging techniques in clinical practice has made it particularly suitable for radiomics.

2.3.3 Positron Emission Tomography

Positron Emission Tomography, or PET, is an imaging technique that uses γ-rays and radioactive tracers to draw tomographic images of the body showing regions with higher metabolic activity [51]. The physical phenomenon which PET is based on is the radioactive decay of some isotopes (like ^{18}F) that are inserted in molecules called radio-tracers. When the radioactive atom decays. Every time an atom decays, it emits a positively charged particle called positron, which, after annihilation with an electron in tissues, result in the formation of two γ-rays [48], which have trajectories 180° apart and strike solid-state detectors which are positioned in a series of complete rings around the patient (Figure 2.6A). Given any two detectors it is possible to track the initial position of the decaying atom (for the mathematical details see [48]). By analyzing the number of decays per second it is possible to obtain tomographic images (Figure 2.6B).

By adjusting the signal for patients weight and by quantity of radiotracer injected it is possible to compute a standardized metric called Standardized Uptake Value (SUV) [52] and use PET as a quantitative imaging technique.

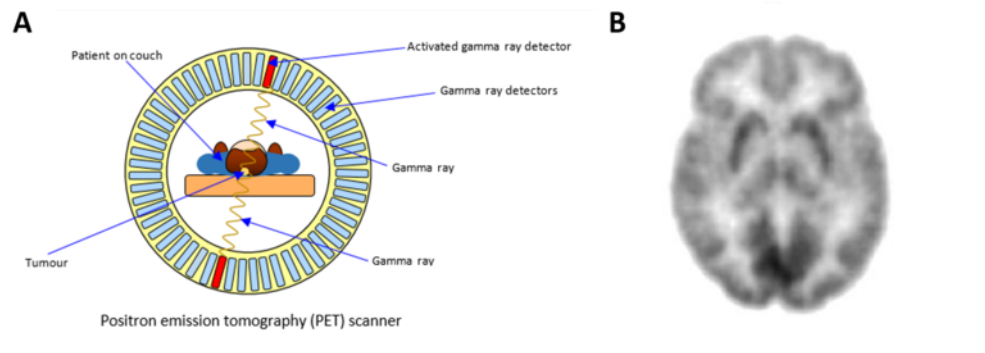

Figure 2.6: *PET image acquisition. A) Example of Positron Emission Tomography (PET) scanner showing how the PET signal is acquired. B) PET image showing the spatial distribution of radioactive decays. The darker regions show higher tracer concentration and are associated with a more intense metabolic activity. Adapted from [53].*

The radioactive tracers used in PET (e.g. ^{18}F-FlouroDeoxyGlucose or FDG, which is the most common) are typically a modified version of molecules that are part of the cellular metabolism and tend to flow more in the region of the body with high metabolic activity, like for example cancer cells [51]. Therefore, PET imaging allows to recognize an organ by its activity rather than by its physical properties. Therefore PET is defined as a functional imaging technique [51], since it does not show the anatomy of the organs, but shows which regions of the body are more active.

Since PET detects tumor metabolic activity rather than mass, it is useful to diagnose tumors in early-stage, when they are still not visible to anatomical imaging techniques such as CT or MRI. Disadvantages of PET include lower resolution and higher cost compared to CT and MRI and use of ionizing radiations (γ-rays).

2.3.4 Magnetic resonance imaging

Nuclear Magnetic resonance

The physical principle behind MRI is the Nuclear Magnetic Resonance (NMR), i.e. the synchronized precession of nuclei of some isotopes when they are put inside a static magnetic field and are excited with energy of proper frequency [48]. The isotope used for MRI is ^{1}H due to its abundance in the molecules of the body. The nucleus of ^{1}H consist of a single proton, so the term proton and nucleus will be used interchangeably in this section.

At the atomic level, the ^{1}H nucleus is a charged particle which spins around an internal axis of rotation with a given value of angular momentum (P), it also has a magnetic moment (μ), and therefore can be thought of as a very small bar magnet with a north and south pole, as shown in Figure 2.7A. Each rotating nucleus is also called spin.

When considering a packet of spins in the absence of a magnetic field (Figure 2.7B) the net magnetic moment (or the net magnetization) is zero because the single spins are randomly oriented and their moments cancel out each other. When an external magnetic field is applied (Figure 2.7C), the spins tend to orient on the same direction of the magnetic field (called z-direction for convention) with either the same verse (low-energy state) or opposite verse (high-energy state). Spins in the low-energy state are slightly prevalent and this causes the packet of spins to have a non-null net magnetization M_0 in the same verse of the magnetic field (longitudinal magnetization).

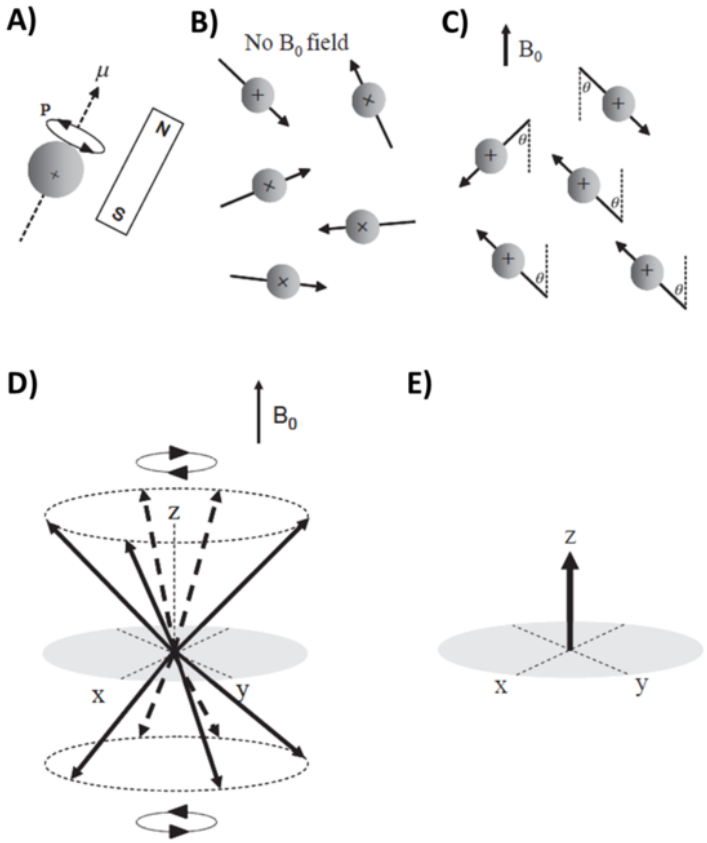

Figure 2.7: *Magnetization of ^{1}H nuclei in a static magnetic field. A) Single proton with its magnetic moment (spin). B) Spins in the absence of an external magnetic field. C) Spins in presence of external magnetic field. D) Vectors representing 3D magnetic moment of the single spins. E) Net magnetization due to the sum of all the magnetic moments: the transverse components cancel each other and only a net longitudinal magnetization remains. Adapted from [48].*

The spins also have a mote of precession around the direction of the magnetic field, with an angle of precession of θ (Figure 2.7C) and a frequency of precession f, called Larmor frequency, which can be computed as follows:

$$f = \gamma B_0 \tag{2.1}$$

where B_0 is the magnitude of the external magnetic field (in T) and γ is the gyromagnetic ratio of the spin (in MHz/T). For ^{1}H the gyromagnetic ratio is 42.575 MHz/T and the corresponding Larmor frequency with field strength of 1.5 T (the most used in clinical practice) is around 64 MHz,

in the band of the radio-frequencies. Due to this precession movement, the single spins have also a component of the magnetization vector that rotates in the plane perpendicular to the direction of the magnetic field, called transverse magnetization (Figure 2.7D). However, since the rotation of the spins are not in phase, macroscopically the packet of spin shows no transverse magnetization, but only the longitudinal one is present (Figure 2.7E).

In this status, the longitudinal magnetization of the packets of spins cannot be detected, since it is far smaller than the static magnetic field. However, when an external sinusoidal magnetic field B_1 varying with the Larmor frequency (a Radio-Frequency pulse, or RF pulse) is applied perpendicularly to the z-direction (in the x-direction for example), two things happen: 1) more spins switch from the low-level energy state to the high level-energy state; 2) the precession of the spins synchronize to the same phase. Macroscpically this results in the reduction of the longitudinal magnetization and in the generation of a non-null transverse magnetization. If the RF pulse is applied for long enough, the number of high-energy spins equals the ones of low-energy spins, reducing the longitudinal magnetization to 0 and maximizing the transverse magnetization (Figure 2.8). This is equivalent to a rotation of 90° around the x-axis and for this reason the RF pulse is called 90° RF pulse. If the RF pulse is applied for an even longer time all the protons will switch to a high-energy state causing the net magnetization to be opposite to the external magnetic fields. This is called 180° RF pulse.

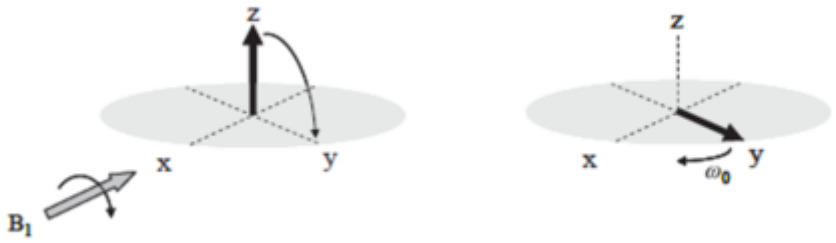

Figure 2.8: *Example of 90° pulse rotating the net magnetization vector [48].*

The transverse magnetization that is generated by a 90° RF pulse can be measured by a receiving coil that generates an electrical signal due to electromagnetic induction. Such signal is called Free Induction Decay (FID) and is the oscillating signal visible in (Figure 2.9). If the signal is mea-

sured with two perpendicular coils, it is possible to compute the module and the phase of the transverse magnetization and to express it as a complex number. Such transverse magnetization signal is the one that is used to reconstruct the image of the object/patient put in the scanner.

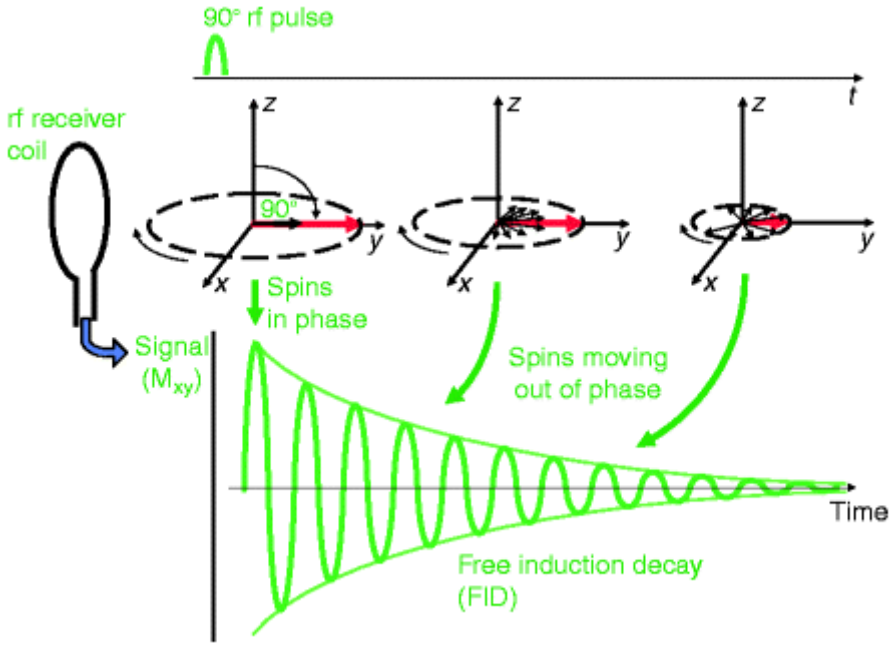

Figure 2.9: *The free induction decay (FID) measured by one coil after spins excitation with the 90° radio-frequency (RF) pulse. The oscillating signal is the one measured by the coil wile the envelope of the peaks of the FID is the modulus of the transverse magnetization [54].*

Magnetization relaxation

When the RF pulse stops, the spins tend to return to their resting position [47]. Some of the high-energy spins start to emit energy and return to the low-energy state. This causes a recovery of the initial longitudinal relaxation (Figure 2.10A-B). The level of longitudinal magnetization as a function of time t from the end of the RF pulse can be expressed by the following relation:

$$M(t) = M_0 \left(1 - e^{-\frac{t}{T1}}\right) \tag{2.2}$$

where M_0 is the original longitudinal magnetization and T1 is a time constant that is dependent on the type of tissue in which the excited spins

are placed in. This longitudinal recovery is also called spin-lattice relaxation or T1 relaxation [47].

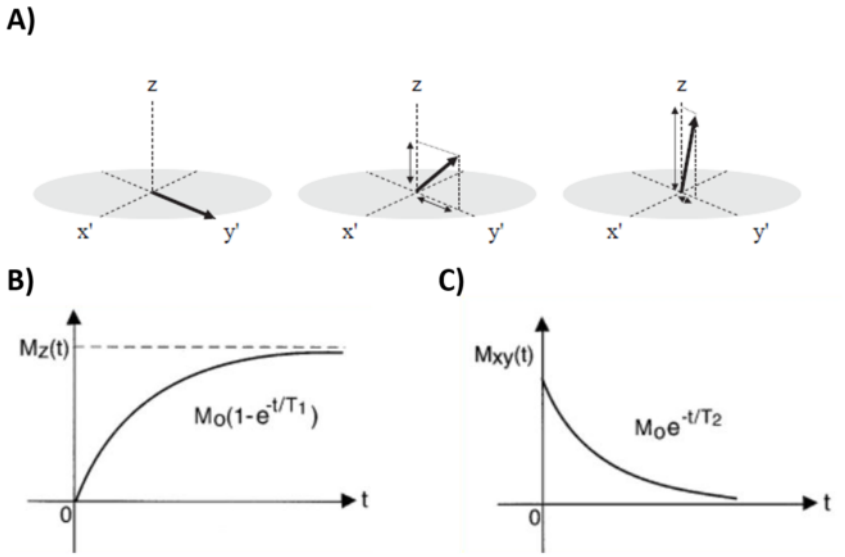

Figure 2.10: *Phenomenon of magnetization relaxation. A) 3D representation of the net magnetization vector. B) Recovery of the longitudinal magnetization over time. C) Decay of the transverse magnetization over time. Adapted from [48].*

Another effect that can be observed after the end of the RF pulse (Figure 2.10A-C) is the decay in the intensity of the transverse magnetization, which is due to the dephasing of the precession. Such decay in the signal can be expressed as a function of time:

$$M(t) = M_0 e^{-\frac{t}{T2}} \tag{2.3}$$

where M_0 is the maximal transverse magnetization and T2 is a time constant that is dependent on the type of tissue in which the excited spins are placed in. This loss of signal is called spin-spin relaxation or T2 relaxation [47].

When the same magnetic field strength is used, T1 and T2 (also called relaxation times) only depend on the properties of the tissues and can therefore be used to characterize and distinguish the different tissues inside an MRI image. Table 2.1 provides values of T1 and T2 for some reference tissues (in ms) for a static magnetic field of 1.5 T. It is important to see that T1 is always larger than T2 and this is due to the fact that T2-relaxation

only depends on the dephasing of the spins, while the recovery of the longitudinal magnetization also depend on the transition of the spin from high to low energy state, which is a much slower process [47].

TISSUES RELAXATION TIMES (1.5 T)		
Tissue	**T1 (ms)**	**T2 (ms)**
Water/Cerebrospinal fluid	4000	2000
Grey matter	900	90
Muscle	900	50
Liver	500	40
Fat	250	70
Tendon	400	5
Proteins	250	0.1-10

Table 2.1: *Values of T1 and T2 relaxation times (at 1.5 T) for different tissues in the body [55].*

One last thing to know about MRI relaxation is that, in the case of spin-spin relaxation, the real decay of the transverse magnetization is faster compared to the ideal one computed using Equation 2.3 (Figure 2.11). This happens because of inhomogeneities in the static magnetic field that accelerate the process of dephasing of the spins. This results in an exponential decay that is described by another time constant called T2*, shorter than T2 [48, 56]. This issue causes a reduction in the signal to noise ratio (SNR) of the images but it is typically solved by using alternative combinations of RF pulses (called pulse sequences).

Spin-echo pulse sequence

The Spin-Echo (SE) pulse sequence is the most basic among the pulse sequences that are used in the clinical practice to acquire the MRI signal [48, 57]. It was invented as a way to overcome the main flaws of the FID, i.e the reduction in SNR due to the T2*-decay and its sensitivity to local inhomogeneities [48].

In SE sequence an additional 180° RF pulse is added at time τ after the initial 90° pulse has been applied (Figure 2.12). Such 180° RF pulse causes a flip of the spins, but does not change the local magnetic field inhomogeneities that caused the T2*-decay. This causes the spins to see an

opposite local magnetic fields and the dephasing that has been accumulated in the interval [0; τ] is recovered in [τ; 2τ]. This causes a new increase in the transverse magnetization, called echo, which peaks at time 2τ, when the signal is equal to the one obtainable with the ideal T2-decay [47, 57].

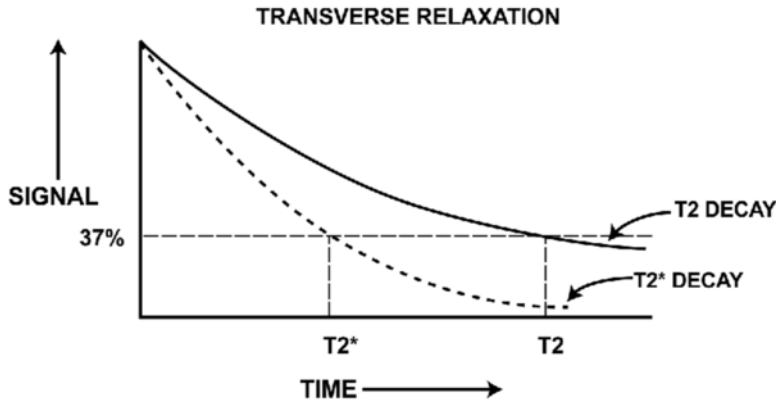

Figure 2.11: *Comparison of the decay of the transverse magnetization due to spin dephasing (T2-decay) and the ones due also to magnetic field inhomogenities (T2*-decay). Adapted from [58].*

The value 2τ is typically called Time of Echo (TE), which is one of the parameter used in the clinical practice to define an MRI acquisition [48,57]. The MRI signal is typically acquired only in the neighborhood of the TE [TE-Δt; TE+Δt].

When acquiring an MRI image, the process described above is repeated multiple times. Another important parameter that defines the SE pulse sequence is the Time of Repetition (TR), i.e. the time between one 90° RF pulse and the next one [48, 57].

MRI signal localization and image reconstruction

When creating an MRI image, the packets of spins are grouped in small portions of volume called voxels (pixels when referring to 2D images). In order to get the MRI signal from each voxel, it is necessary to repeat the RF pulse multiple times and each time, the location of the voxel of interest must be encoded. A brief description of the localization process will be given. For further details, refer to [48].

To localize the MRI signal coming from different voxels, MRI scanners used additional linear magnetic fields to adjust the precessional frequency

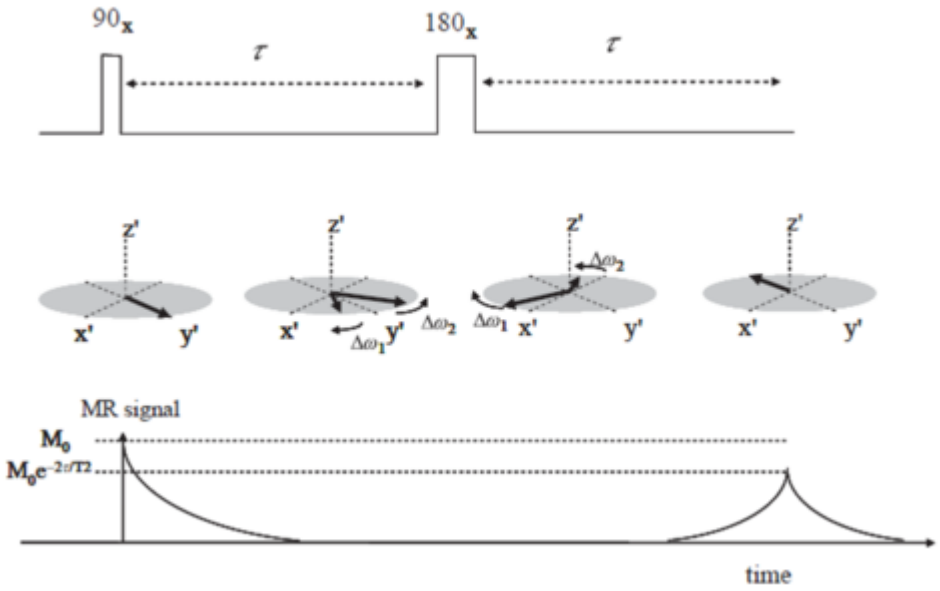

Figure 2.12: *Spin-echo pulse sequence. The addition of a 180° radio-frequency pulse at time τ after the initial 90° pulse causes a recovery in the transverse magnetization called echo [48].*

and the phase of the precession of the spins in the voxels of interest [57]. In particular, three types of gradients are applied, according to the axis of imaging (x-, y-,or z-axis). The slice encoding gradient selects the section to be imaged [57]. The phase-encoding gradient causes a phase shift in the spinning protons so that the MR imaging system computer can detect and encode the phase of the spin [57]. The frequency-encoding gradient also causes a shift that helps the MR system to detect the location of the spinning nuclei [57]. Because this shift of frequency usually occurs when the echo is read, it is also called the readout gradient. Once the MRI system processor has all of that information (i.e. the frequency and phase of each spin), it can compute the exact location and amplitude of the signal. That information is then stored in a row of a two dimensional matrix called k-space (Figure 2.13) as a matrix of complex numbers. The application of a 2D inverse Fourier transform of the k-space [48], provides the final image representing one slice of the MRI acquisition (Figure 2.13).

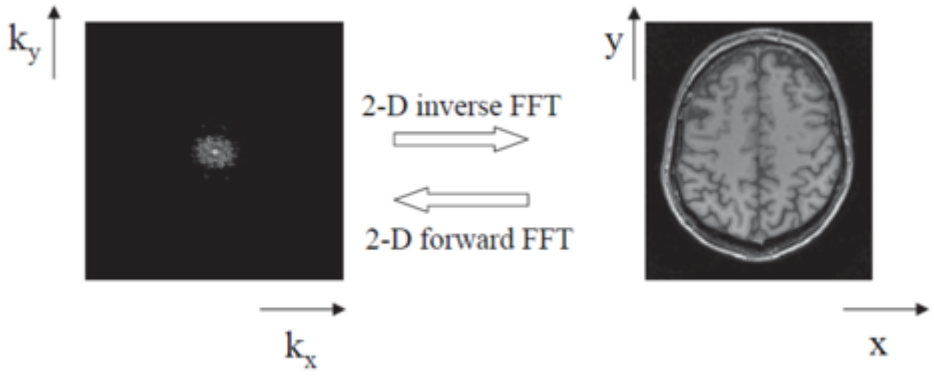

Figure 2.13: *The 2D inverse Fast Fourier Transform (FFT) can be used to compute the final magnetic resonance image. Adapted from [48].*

Spin-echo imaging

Multiple repetitions of the SE pulse sequence and a proper use of encoding gradients can be used to provide 3D MRI images. Unlike other imaging techniques such as CT or PET, the signal is not directly related to just one physical property of the tissue but rather depends on a combination of several biological properties, such as T1, T2 and the Proton Density (PD) inside each voxel, and image acquisition parameters, such as TR, TE and the strength of the local magnetic field [47, 48].

Given a 2D slice $I(x, y)$ the intensity measured at location (x, y) depends on different parameters as defined by the following equation [48]:

$$I(x,y) = K(x,y)\rho(x,y)\left(1 - e^{-\frac{TR}{T1(x,y)}}\right)e^{-\frac{TE}{T2(x,y)}} \tag{2.4}$$

where $\rho(x, y)$, $T1(x, y)$ and $T2(x, y)$ are respectively the mean PD, T1 and T2 of the pixel (x, y), TR and TE are the parameters used to define the SE pulse sequence, and K(x,y) is a multiplicative term that depends on several factors (static magnetic field, local inhomogeneities, and amplification system of the scanner).

By looking at Equation 2.5, it is possible to see how, with fixed biological properties, it is possible to control the signal and modify the contrast among tissues just by modifying the TR and TE [57, 59]. As a matter of fact, different types of MRI images can be obtained using the SE pulse sequence (Figure 2.14), each highlighting the differences due to a particular biological property.

When a short TR (typically less than 700 ms) and a short TE (typically

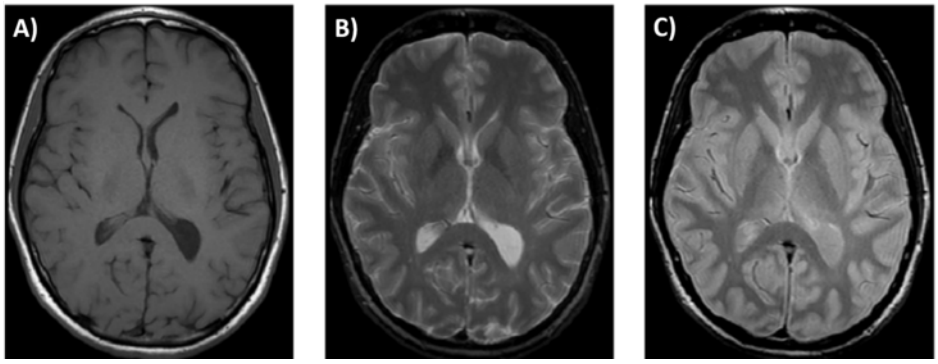

Figure 2.14: *Examples of images obtainable with the spin-echo sequence. A) T1-weighted image. B) T2-weighted image. C) Proton-density-weighted image.*

less than 25 ms) are used the contrast between tissues is mainly due to differences in T1. Images obtained in this way are called T1-weighted (T1w) images (Figure 2.14A). When both a long TR (typically higher than 2000 ms) and a long TE (typically higher than 60 ms) are used T2-weighted images (T2w) are obtained (Figure 2.14B). Last, images highlighting differences in PD, or PD-weighted (PDw), can be obtained by using long TR and short TE (Figure 2.14C). In this thesis, the focus will be more on the first two image types, since they are part of the datasets that will be analyzed.

Spin-echo images (and MRI images in general) can be further customized by using particular contrast agent (e.g. use of gadolinum to enhance contrast in T1-weighted images) or by using particular techniques such as fat-suppression that allow to remove the bright signal from fat, therefore enhancing the contrast between the remaining tissues [59].

Diffusion-weighted imaging

Diffusion-Weighted Imaging (DWI) is another type of MRI that provides images whose contrast gives information on the diffusion of hydrogen nuclei in the water molecules of the body [60]. It is a useful imaging technique in oncology as it can provide additional information compared to the SE images. As a matter of fact, tumors are often made by dense masses of tumor cells packed together, which have very different diffusion properties compared to the surrounding tissues [60].

Figure 2.15 present a modified version of SE sequence that could be used to obtain a DWI. The change consist in the application of a symmetric pair of diffusion-sensitizing gradients around the 180° refocusing pulse. In this situation, static molecules acquire phase information from the first dif-

fusion gradient, but information will be rephased by the second diffusion gradient without a significant change in the measured signal intensity [60]. By comparison, moving water molecules acquire different phase information from the first gradient, but because of their motion, their signal will not be completely rephased by the second gradient, thus leading to a signal loss [60]. The degree of water motion has been found to be proportional to the degree of signal attenuation.

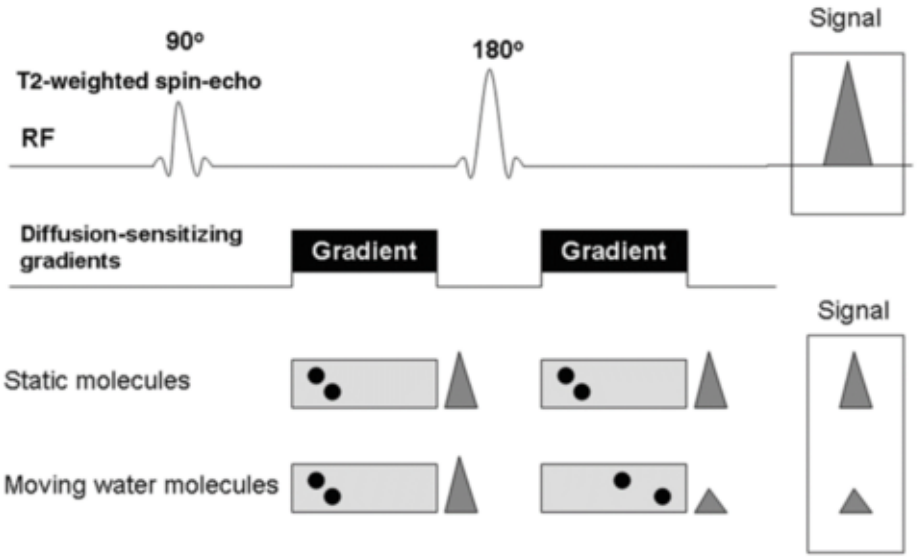

Figure 2.15: *Modified T2-weighted spin-echo that can be used to obtain a diffusion weighted image [60].*

The sensitivity of the DWI sequence to water motion can be varied by changing a parameter called b-value (measured in s/mm^2) which is defined as follows [61]:

$$b = \gamma^2 G^2 \delta^2 \left(\Delta - \frac{\delta}{3} \right) \tag{2.5}$$

where γ is the gyromagnetic ratio of the ^{1}H nuclei, G is the magnitude of the sensitizing gradients, δ is the duration of the two gradients and Δ is the time between the beginning of the first gradient and the beginning of the second one. The b-value is typically controlled by changing the values of G, leaving the other parameters constant. An image with b-value 0 s/mm^2 is a T2w image (typically performed with fat-suppression) and the higher the b-value, the higher the signal attenuation due to diffusion. Tissues with

high cellular concentration tend to maintain high signal in DWI even when large b-values are used (Figure 2.16A-B).

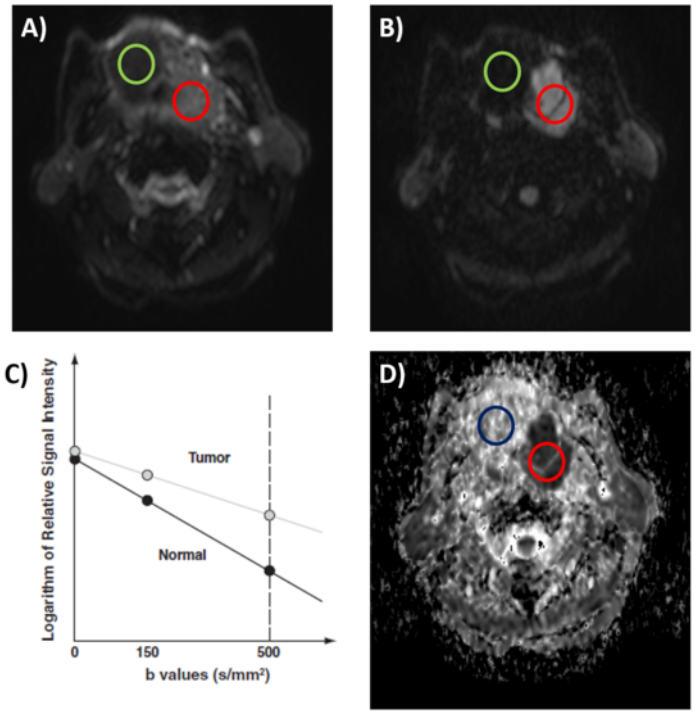

Figure 2.16: *Process of creation of an Apparent Diffusion Coefficient (ADC) map from diffusion-weighted images (DWI). A) DWI image with a b-value of 0 s/mm^2. B) DWI image with a b-value of 500 s/mm^2. C) Signal decay by b-value for tumor and normal tissue. D) Map of ADC.*

The relationship between signal decay and b-value can be modeled using a single exponential [60]:

$$S(x, y, b) = S(x, y, 0)e^{-bADC(x,y)} \tag{2.6}$$

where $S(x, y, 0)$ is the signal measured in pixel (x, y) using a b-value of 0 s/mm^2 and $ADC(x, y)$ is a property of the tissue called Apparent Diffusion Coefficient.

Equation 2.7 can be modified and transform into a linear equation (Figure 2.16C):

$$\ln S(x, y, b) = \ln S(x, y, 0) - bADC(x, y) \tag{2.7}$$

Given two or more DWI images computed with different b-values, the values of ADC can be computed pixel-wise to obtain maps of ADC (Figure 2.16D). In ADC maps, low intensity represents regions with lower diffusion coefficients (e.g. tumors). Therefore, ADC maps can be used as quantitative images, like CT or PET and unlike T1w or T2w MRI, since the measured intensities have a physical meaning. This also makes ADC maps more standardized across clinical centers and among scanners, as soon as the same magnetic field and range of b-values is used [62, 63]. However, ADC images typically have lower resolution compared to traditional MRI images and cannot be used to detect fine details in the image [64]. Also, longer acquisition times are required [64].

Fast MRI imaging

Although the pulse sequences that were previously presented could potentially be used to obtain the T1w, T2w and DWI images, they are not used in the clinical practice because they lead to long acquisition times (several hours for a single diagnostic exam). Therefore, faster pulse sequences are used.

Turbo Spin-Echo (TSE) [57], also called fast spin-echo, is a particular type of SE in which multiple echos are acquired during the same iteration (Figure 2.17). This allows to fill multiple lines of the k-spaces in one iteration. Besides TR and TE, the TSE pulse sequence is characterized by Echo-Train Length (ETL), i.e. the number of echos acquired in the same iteration. TSE is typically used in the clinical practice to acquire T1w and T2w image [57].

Echo-Planar Imaging (EPI) is a pulse sequence that is used to substantially reduce the acquisition time and that is used in particular for DWI acquisition [57]. In EPI, the phase-encoding gradient and the frequency-encoding (or readout) gradient are turned on and off very rapidly, a technique that allows the rapid filling of k-space in one iteration (Figure 2.18). All the DWI acquired in the thesis were acquired using EPI.

2.4 Radiomic features description

As described in Section 2.1, different categories of radiomic features may be extracted. The description of the features classes that is given in this subsection is not meant to be exhaustive, but will mainly refer to the ones that are used in this thesis and in the majority of the works involving radiomic analysis. The purpose of this subsection is to provide a general knowledge of the different mathematical tools that are used to compute the features of

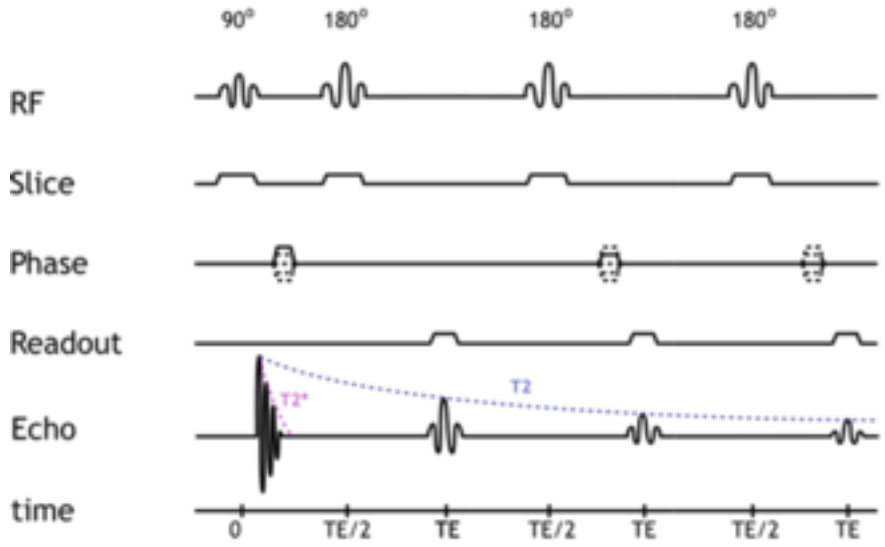

Figure 2.17: *Representation of turbo spin-echo pulse sequence. In the scheme, the temporal application of radio-frequency pulses (RF) is shown, as well as the one of encoding gradients (slice, phase and readout gradients) [65].*

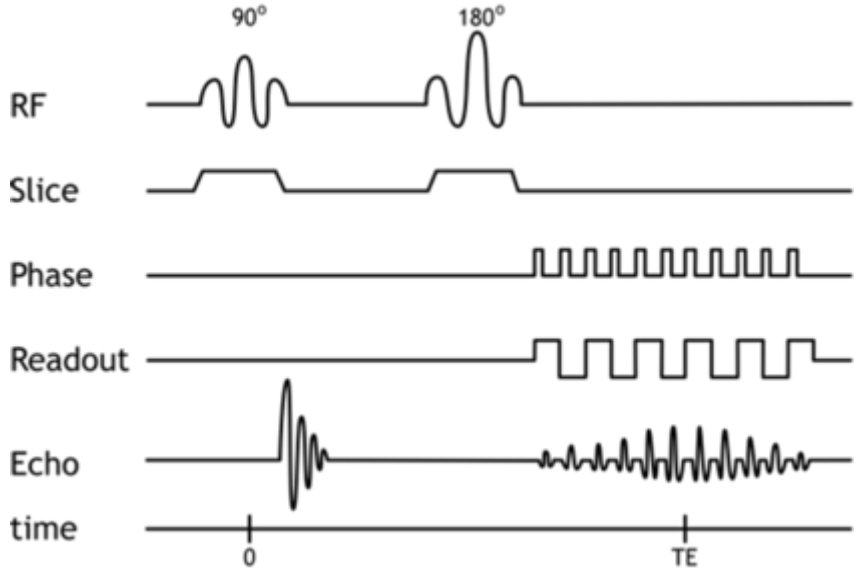

Figure 2.18: *Representation of echo-planar imaging pulse sequence. In the scheme, the temporal application of radio-frequency pulses (RF) is shown, as well as the one of encoding gradients (slice, phase and readout encoding gradients) [65].*

interest, but not to provide the complete list of all the radiomic features. For that, the reader may refer to the documentation of Pyradiomics [66], which is the software that was used to extract the radiomic features in this thesis, or to the manual of the Imaging Biomarker Standardization Initiative (IBSI, see [67]).

2.4.1 Shape and size features

Shape and Size features (SS) describe geometric aspects of a ROI, such as area and volume. In order to calculate these features, different representation of the ROI may be used [67]:

- A collection of voxels with each voxel taking up a certain volume.
- A voxel point set that consists of coordinates of the voxel centers.
- A surface mesh.

SS features in Pyradiomics are typically computed using a surface mesh. A surface mesh is a surface made of M adjacent triangular surfaces and N vertices. Each mesh can be described using a $Nx3$ matrix with the 3D coordinates of the points and a $Mx3$ matrix with the index of the points that are used to define each surface.

SS features may be computed both in 3D and 2D but for this thesis, only 3D features were used.

2.4.2 First order statistics

First Order Statistics (FOS) are features that describe the distribution of the different grey values inside the ROI. Features of this category may be further classified in intensity-based or histogram-based features [67]. The difference between the two is that to compute the latter, an histogram is required. The computation of the histogram requires the discretization of the distribution of the grey levels in bins. Two approaches can be used for the histogram discretization [67]: fixed bin size and fixed bin number.

Intensity-based statistical features are not meaningful if the intensity scale is arbitrary, like in MRI. So proper intensity standardization should be made before any radiomic analysis (see Chapter 3).

2.4.3 Textural features and textural matrices

Textural features provide spatial information about the distribution of the grey values inside the ROI [67]. Just as some FOS features are computed

from an histogram, textural features are computed from textural matrices. Like for histogram-based features, grey values discretization is usually performed prior to the computation of the textural matrices.

Pyradiomics uses 5 different textural matrices [66]: Grey Level Co-occurrence Matrix (GLCM); Grey Level Run Length Matrix (GLRLM); Grey Level Size Zone Matrix (GLSZM); Grey Level Dependence Matrix (GLDM); Neighbouring Grey Tone Difference Matrix (NGTDM).

The GLCM is a N_gxN_g matrix, N_g being the number of discrete grey values in an image, and describes the second order joint probability function of an image region constrained by a ROI. Fixed a given positive distance δ and a direction θ the (i, j) element of the GLCM describes how many times the element grey value j appears at a distance δ (e.g. one pixel) from the grey value i in the direction θ (e.g. horizontal). An example of GLCM computation is reported in Figure 2.19. The rules for GLCM computation are analogous in 3D, with the only difference in the number of possible θ directions.

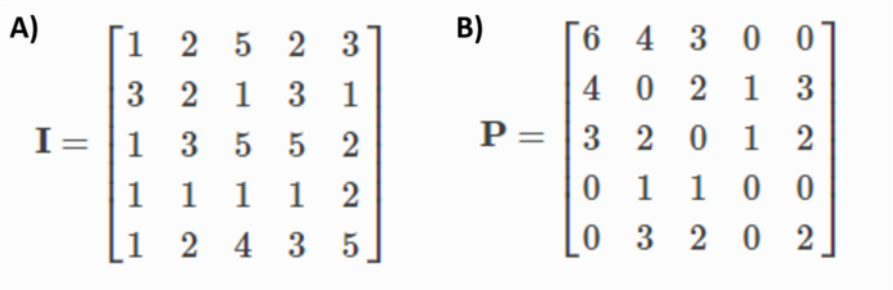

Figure 2.19: *Example of computation of a Grey Level Co-occurrence Matrix (GLCM). A) Image I with the original grey values. B) Corresponding GLCM. The matrix P in B) is computed using $\delta = 1$ (1 pixel distance) and $\theta = 0$ (horizontal direction, both left to right and right to left). Adapted from [66].*

The GLRLM is a matrix that quantifies grey level runs, which are defined as the length of consecutive pixels that have the same grey level value. Given a direction θ the (i, j) element of the GLRLM describes how many times the element grey value i appears consecutively for j times. Therefore, the dimension of the matrix is N_gxN_{max}, N_{max} being the maximum size in the image. An example of GLRLM computation is reported in Figure 2.20.

The GLSZM quantifies grey level zones in an image. A grey level zone is defined as a the number of connected voxels that share the same grey level intensity. The (i, j) element of the GLSZM equals the number of zones with grey level i and size j appear in image. An example of GLSZM

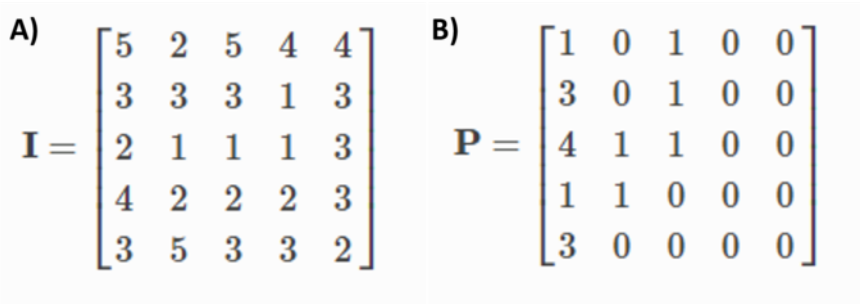

Figure 2.20: *Example of computation of a Grey Level Run Length Matrix (GLRLM). A) Image I with the original grey values. B) Corresponding GLRLM. The matrix P in B) is computed using $\theta = 0$ (horizontal direction). Adapted from [66].*

computation is reported in Figure 2.21. The GLSZM matrix displayed in Figure 2.21B is displayed as a 5x5 for simplicity, but in general a GLSMZ has size of N_gxN_p, where N_p is the total number of pixels in the image.

The NGTDM quantifies the difference between a grey value and the average grey value of its neighbours within distance δ. The NGTDM is a N_gx4 matrix (Figure 2.22). The $(i, 1)$ element of the matrix shows the grey value i, the elements $(i, 2)$ and $(i, 3)$ represent its absolute and relative frequency in the matrix and, given a size δ, the element $(i, 4)$ is the mean absolute difference in the grey values between each voxel with intensity i and the average grey value in its neighborhood.

The GLDM quantifies grey level dependencies in an image. A grey level dependency is defined as a the number of connected voxels within distance δ that are dependent on the center voxel. Given a parameter α, a voxel is said to be connected if $|i - j| \leqslant \alpha$. The (i, j) element of the GLDM describes the number of times a voxel with grey level i with j dependent voxels in its neighbourhood appears in image. An example of GLDM is presented in Figure 2.23.

2.4.4 Wavelet transform and wavelet features

In many of the radiomic studies of literature, radiomic features belonging to FOS and textural groups are extracted not just from the original images, but also from transformed versions, in order to provide additional insight on the object that is being imaged (e.g. the tumor). In particular, in this thesis, the wavelet decomposition was used.

Wavelet decomposition effectively decouples textural information by decomposing the original image, in a similar manner as Fourier analysis, in low âand high-frequencies [5]. A detailed mathematical detail of wavelet

A)

$$\mathbf{I} = \begin{bmatrix} 5 & 2 & 5 & 4 & 4 \\ 3 & 3 & 3 & 1 & 3 \\ 2 & 1 & 1 & 1 & 3 \\ 4 & 2 & 2 & 2 & 3 \\ 3 & 5 & 3 & 3 & 2 \end{bmatrix}$$

B)

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure 2.21: *Example of computation of a Grey Level Size Zone Matrix (GLSZM). A) Image I with the original grey values. B) Corresponding GLSZM. Adapted from [66].*

A)

$$\mathbf{I} = \begin{bmatrix} 1 & 2 & 5 & 2 \\ 3 & 5 & 1 & 3 \\ 1 & 3 & 5 & 5 \\ 3 & 1 & 1 & 1 \end{bmatrix}$$

B)

i	n_i	p_i	s_i
1	6	0.375	13.35
2	2	0.125	2.00
3	4	0.25	2.63
4	0	0.00	0.00
5	4	0.25	10.075

Figure 2.22: *Example of computation of a Neighbouring Grey Tone Difference Matrix (NGTDM). A) Image I with the original grey values. B) Corresponding NGTDM. Adapted from [66].*

A)

$$\mathbf{I} = \begin{bmatrix} 5 & 2 & 5 & 4 & 4 \\ 3 & 3 & 3 & 1 & 3 \\ 2 & 1 & 1 & 1 & 3 \\ 4 & 2 & 2 & 2 & 3 \\ 3 & 5 & 3 & 3 & 2 \end{bmatrix}$$

B)

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 2 & 1 \\ 1 & 2 & 3 & 0 \\ 1 & 4 & 4 & 0 \\ 1 & 2 & 0 & 0 \\ 3 & 0 & 0 & 0 \end{bmatrix}$$

Figure 2.23: *Example of computation of a Grey Level Dependence Matrix (GLDM). A) Image I with the original grey values. B) Corresponding GLDM obtained by setting $\alpha = 0$ and $\delta = 1$. Adapted from [66].*

theory is beyond the scope of this introductory material, and for that, the reader may refer to [24]. The following explanation provides the minimum knowledge to understand the applications of wavelet transform, in particular the 2D Discrete Wavelet Transform (DWT), in image processing and in radiomics is given.

Essentially, the DWT of an image up to level (or scale) J is performed through a cascade tree of low-pass and high-pass filters followed by down-sampling by a factor of 2. For a 2D image, performing one level of a 2D wavelet decomposition consists of filtering and down-sampling an image $I(x, y)$ both in the x and y directions, with both a 1D low-pass and high-pass filters. This results in four sub-bands (Figure 2.24): LL, LH, HL, HH. The LL band (upper-left in Figure 2.24) contains a coarse approximation of the original image, while the other bands contain information about high frequency changes in intensity in the horizontal, vertical and diagonal direction respectively (upper-right, lower-left and lower-right of Figure 2.24). Extracting the radiomic features from those sub-bands will result in new, potentially useful, information.

In case of 3D volumes, like in MRI exams, the combination of high and low-pass filters will results in 8 different possible combinations (Figure 2.25), and the FOS and textural features will be extracted from each of the sub-bands.

2.5 Machine learning and survival analysis

2.5.1 Machine learning: general introduction

Machine learning is a field of computes science that deals with the design and creation of algorithms that perform a task without being explicitly programmed, but rather by learning it directly from data through some sort of inference. The data that are given to the machine to make it learn (i.e. the training data) may take different form but can usually be reduced to an array of features describing the single instance or subject to be predicted [68]. Such features can be of different types: numerical (e.g. body weight), ordinal (e.g. tumor grade) or categorical (e.g. patient gender). Different types of machine learning exist and may be summarized as follows [68, 69]:

- **Supervised machine learning** in which each training datum (also called instance) includes both the features and an expected outcome, i.e a value (or an array of values) that should be obtained as a function of the input features. The goal for this type of machine learning is to automatically infer the best relationship between the features and the

Figure 2.24: *Example of first level 2D discrete wavelet decomposition of an image. The upper-left corner represents the LL sub-band (an approximation of the original image). The upper-right, lower-left and lower-right parts of the image represent the LH, HL and HH sub-bands respectively. The latter have been thresholded for better visualization.*

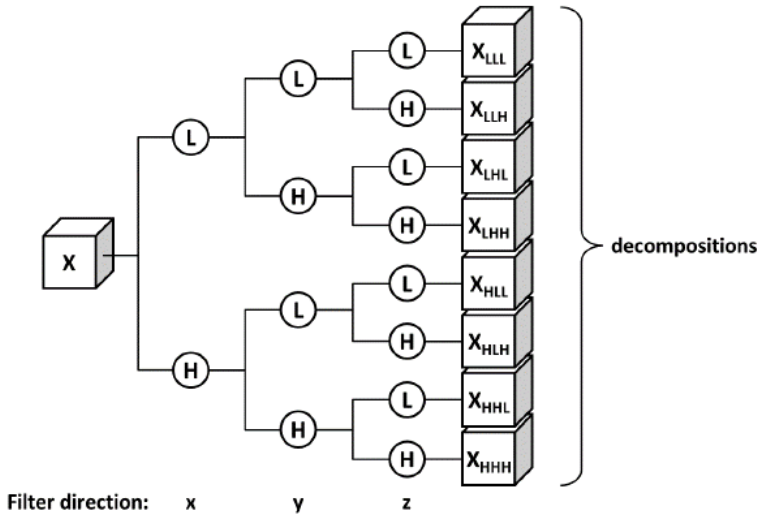

Figure 2.25: *Schematic representation of the 8 possible discrete wavelet decompositions of a 3D image [5].*

outcome.

- **Unsupervised machine learning** in which the training data includes only features without an outcome associated. The goal in this type of machine learning is to identify groups and patterns within the data in order to provide a better insight.
- **Semi-supervised machine learning** which is a type of supervised learning in which the expected outcomes are available for only a portion of the instances. The goals are the same of the supervised learning, but the partial lack of outcome variables requires the use of alternative methodologies.
- **Reinforcement learning** which is concerned with the problem of finding suitable actions to take in a given situation in order to maximize a reward. Here the learning algorithm is not given examples of optimal outputs, in contrast to supervised learning, but must instead discover them by a process of trial and error.

This thesis dealt with the first type of machine learning, and in particular with a subcategory of supervised learning called survival analysis. The concepts of supervised machine learning and survival analysis, as well as the relative workflows, will be described in details in the following subsections.

2.5.2 Supervised machine learning and survival analysis

As defined in Subsection 2.5.1, supervised machine learning is a type of machine learning problem in which both features and an expected outcome are available for each instance [68]. Different types of output variables may be used in supervised machine learning and, depending on the output variable, three main supervised learning problems may be identified.

In regression, the output variable is a continuous number [68]. In the context of radiomics, an example of regression problem is predicting the volume of a tumor after a particular treatment based on some quantitative features extracted from the baseline imaging.

In classification, the output variable is a categorical value [68]. In general, the models that are trained to solve a classification problem tend to assign a scoring system or a rule that associates the features to the probability of belonging to a specific class. Binary classification refers to the situation in which there are only two classes, while when more than two classes are present, the problem is called multi-class classification [68]. In the context

of radiomics, a problem of classification may be to automatically distinguish a benign from a malignant tumor using quantitative imaging features.

Survival analysis is a particular type of supervised analysis used in medical statistics in which features are used to predict a function $S(t)$ representing the probability of not experiencing an event at a certain point in time t [70].

$$S(t) = Pr\{T \geqslant t\} \tag{2.8}$$

where T is the time of the event.

In the context of oncology, survival problems of interest are, for example, the estimation of the OS probability (i.e. the probability of a patient to be alive at a certain point in time after being diagnosed with cancer) and Disease-Free Survival (DFS) probability (i.e the probability of a patient to be alive without tumors at a certain point in time after the treatment). The models used to predict $S(t)$ are called prognostic models.

In case of survival analysis the values of the output variable in the train data are not the value of $S(t)$, which are unknown, but the time-to-event for each patient, that can be used to estimate $S(t)$ for the population [70]. The estimation of $S(t)$ from the time-to-event data is performed using a non-parametric method called Kaplan-Meier method [70, 71]. The estimate of survival function according to the Kaplan-Meier method may be computed as follows [70, 71]:

$$S_{KM}(t) = \prod_{T_i \leqslant t} \frac{n_i - e_i}{n_i} \tag{2.9}$$

where T_i is the time to the i-th event, e_i is the number of events occurred between T_i and T_{i-1}, and n_i is the number of patients that are still in the study between T_i and T_{i-1}.

Since the Kaplan-Meier estimates depends on the number of patients at risk at time t and not on the initial number of patients, it can handle the presence of right-censored patients, i.e. patients whose follow-up is shorter than the duration of the study and for which the occurrence of the event is not known [70]. However, in order to make a proper estimate, it is necessary to specify which patients are censored [70]. Therefore, the outcome data for survival analysis problem actually consist of two variables, the time-to-event and a binary label indicating whether the patient is censored [70].

The Kaplan-Meier estimates of survival can also be visualized using the Kaplan-Meier curves (Figure 2.26).

Another important concept in survival analysis is the concept of hazard

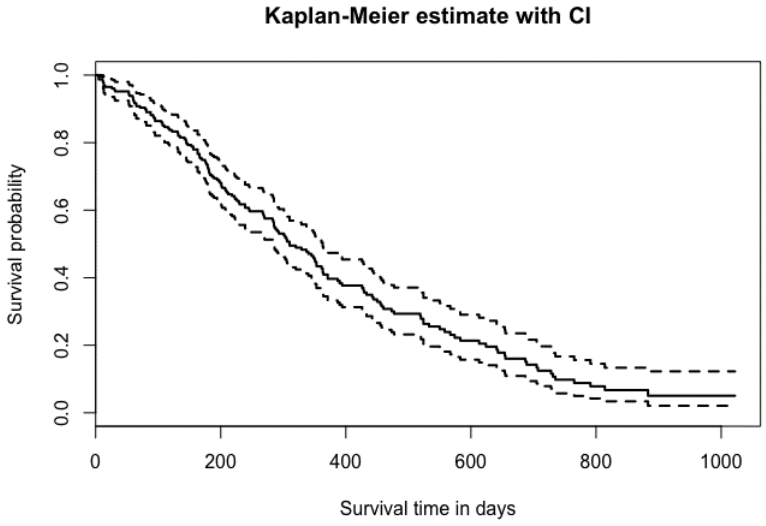

Figure 2.26: *Kaplan-Meier curve representing an estimate of survival probability, with the Confidence Interval (CI).*

function $h(t)$, which is the instantaneous probability of experiencing the event, mathematically defined as [72]:

$$h(t) = \lim_{dt \to 0} \frac{Pr\{t \leqslant T \leqslant t + dt | T \geqslant t\}}{dt} \tag{2.10}$$

The hazard function can be also expressed as the negative derivative of the natural logarithm of the survival function [72]:

$$h(t) = -\frac{d}{dt} \ln S(t) \tag{2.11}$$

In survival analysis the initial condition necessary to solve the differential equation is known a-priori (since $S(0) = 1$) and therefore by estimating the hazard rate is immediately possible to compute the survival function:

$$S(t) = e^{-\int_o^t h(x)dx} \tag{2.12}$$

The output of the algorithms used in survival analysis is typically the hazard function rather than the survival function, but one can easily derive the latter from the former.

2.5.3 Survival analysis: general workflow

Figure 2.27 depicts the general workflow used for the development of prognostic models. The workflow is quite general and is applicable to other types of supervised machine learning such as classification or regression.

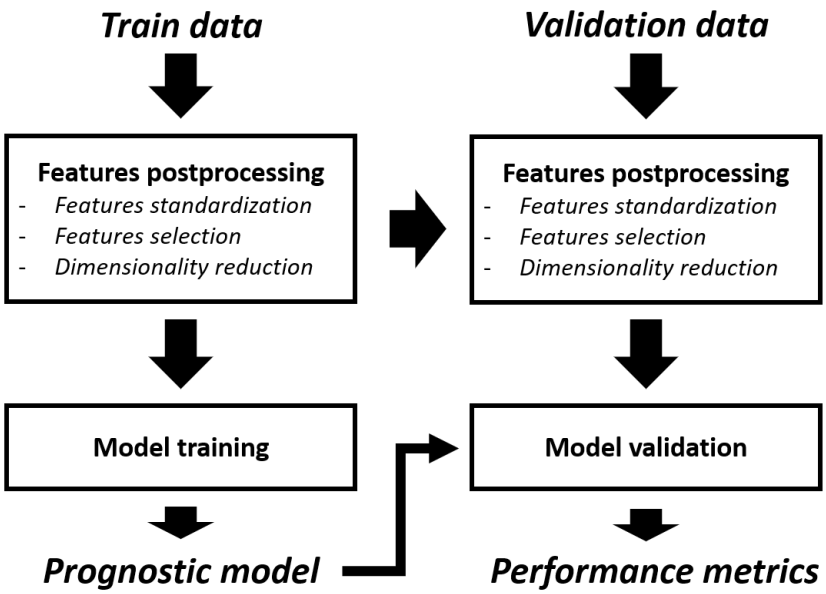

Figure 2.27: *Generic workflow for the development of prognostic models.*

Given a set of training data, it is possible to fit a prognostic model on those data such that a particular goodness-of-fit metric is maximized. In the case of survival analysis, the main prognostic models that is used in the literature is the Cox proportional Hazard regression model.

The features that are part of the training data may undergo a series of postprocessing steps in order to make them more usable in the prognostic models, for example by making the range of features uniform or by selecting a subset of features that is useful for the analysis.The postprocessing steps that were used in this thesis are reported in Figure 2.27, but the list is not exhaustive (see [68] for more). A detailed explanation of the postprocessing steps listed in Figure 2.27 is reported in Subsections 2.5.5 and 2.5.6.

Once the model is trained, some metrics can be used to quantify the performance of the model. However, the computations of such metrics on the same data that have been used to train the model provides a biased estimate of the quality of the model [68]. To avoid this, a validation on a separated

dataset must be performed. The validation data must undergo the same or equivalent postprocessing of the training data and be evaluated by the same model. The score evaluated on the validation data, which is unbiased, can then be used to evaluate the performance metrics of interest [68]. The ideal situation is the one in which an external, independent validation set is available, but this is not always the case. Therefore, some methodologies have been developed to overcome this issue and to perform model validation also when just one dataset is available (i.e. cross-validation, hold-out and bootstrap).

2.5.4 Cox proportional hazard regression model

Cox proportional hazard regression is used to predict the hazard function of a patient that is subjected to different risk factors [70, 72, 73]. According to this model, the hazard function $h(t)$ can be computed as follows:

$$h(t) = h_0(t)e^{\sum_{i=1}^{n} \beta_i x_i} \tag{2.13}$$

where x_i are the risk factors, β_i are the corresponding multiplicative coefficients and $h_0(t)$ is the baseline hazard function, i.e. the hazard function that is observed when all the risk factors are not present. The baseline hazard function is known a-priori or is non-parametrically estimated from the training data [70, 72]. The problem can be re-written as a linear regression problem:

$$\ln\left[\frac{h(t)}{h_0(t)}\right] = \sum_{i=1}^{n} \beta_i x_i \tag{2.14}$$

in which the output variable is the logarithm of the Hazard Ratio (HR), i.e. the ratio between the hazard rate of the group at risk and the one of the baseline group.

If the effect of the risk factors does not depend on time, which is an hypothesis of Cox models [70, 72], the hazard ratio in Equation 2.14 will only depend on the risk factors at $t = 0$ and so is his logarithm. The linear combination of features and coefficient computed in Equation 2.14 will therefore provide a patient-specific risk score, that in the context of radiomics is called radiomic signature or radiomic risk score.

The Cox proportional hazard regression model is a semi-parametric method, because part of Equation 2.13 ($h_0(t)$) is estimated non-parametrically or known a-priori, while the exponent of e depends on some parameters β_i that are fitted on the training data. The β_i coefficient are fitted in order to maximize a quality metric which is typically the partial log-likelihood l.

The partial log-likelihood is like a traditional likelihood but is computed only using the patients that have experienced an event. In the case of Cox proportional hazard regression, l can be computed as follows [73]:

$$l(\beta_i) = \sum_{i \in U} \left(\ln(HR_i) - \sum_{j \in \{T_j \geqslant T_i\}} \ln(HR_j) \right) \tag{2.15}$$

where T_i and T_j are the follow-up times for the i-th and j-th patient, U is the set of uncensored patients (i.e. patients who experience the event during the follow-up) and $\{T_j > T_i\}$ is the set of patients that have a follow up time longer than T_i. The inner summation is evaluated only on the patients with a time to event that is larger then T_i and this ensures that invalid pairs of time-to-event (i.e. pairs in which, due to censoring, the higher time to event cannot be determined) are not used, making it the optimal metric for right-censored data.

2.5.5 Feature normalization

Features normalization is the operation that ensures that all the features have the same (or similar) range of values [68]. This operation is typically performed because having features with similar ranges is a requirements of some postprocessing and models [68]. Even, even when not strictly required, features normalization may be advised because helps the convergence of the optimization algorithms used in model fitting [74]. Different methods of feature normalization exist [75], all with their pros and cons.

Z-score normalization [75] is one of the most frequently used method to normalize the distribution of a feature x by subtracting the mean μ and dividing by the standard deviation σ [75]:

$$x_{Norm} = \frac{x - \mu}{\sigma} \tag{2.16}$$

The median-mad normalization [75] is a generalization of the previous technique that is more suitable for non-normal distributions because it uses the median value and the median absolute deviation:

$$x_{Norm} = \frac{x - median(x)}{median(|x - median(x)|)} \tag{2.17}$$

The min-max normalization [75] is a normalization technique that normalizes the range of values of x between 0 and 1:

$$x_{Norm} = \frac{x - min(x)}{max(x) - min(x)} \tag{2.18}$$

It is the best in terms of providing a fixed range, because it ensures that the normalized values are always between 0 and 1, but it is very sensitive to the presence of outliers.

Hyperbolic tangent normalization [75] is a non-linear modification of the Z-score normalization in which the result of the normalization is put inside a non-linear function (the hyperbolic tangent). The non-linearity reduces the effect of the outliers. The formula describing the normalization method is the following:

$$x_{Norm} = 0.5 \left[tanh \left(0.01 \frac{x - \mu}{\sigma} \right) + 1 \right] \tag{2.19}$$

2.5.6 Feature selection and dimensionality reduction

Survival models, and machine learning algorithms in general, are typically designed to perform well in situations in which the number of instances is much larger than the number of features used [68]. Also in the case of Cox regression, it can be seen that the partial log-likelihood used to fit the models coefficient depend on the patients that experience an event. In order to uniquely fit an optimal model there must be a number of event larger or equal to the number of features used [68]. The minimal events/features ratio was defined to be around 10 [76]. If the events/features ratio is below that threshold, the risk of overfitting, i.e. the training of a model that perform very well on training data but poorly on the test data, increases [68].

One of the main issued of omics is that typically the number of analyzed features is much larger than the number of available patients [77], a situation which is called "curse of dimensionality".

One of the most important steps in the processing of omics features is features selection, i.e. the process of selection of a subset of meaningful features from a much larger set. Another analogous task is performed with the application of dimensionality reduction techniques, in which new features are created from the original, so that almost all the information of a dataset is maintained with a significant reduction in the features number [68].

There are countless different algorithms for features selection and dimensionality reduction and an exhaustive description is not possible. Therefore, in this subsection, only the methodologies that were considered in the thesis are explained.

Correlation based feature selection

The correlation-based feature selection is based on the assumption that a lot of the features in high-dimensional dataset are correlated [6]. Therefore, taking only one out of two correlated features may be a preliminary selection method.

Given any set of features and a correlation coefficient (either Pearson or Spearman), it is possible to compute a correlation matrix with representing the correlation between each pair of features. If the pairwise correlation coefficient between two features has a magnitude above a certain threshold, only one of the two is kept. The selection of the feature to keep may be done in different ways, but the typical choices are to keep the feature with the lower mean correlation coefficient with all the other features in the dataset, or to keep the one with the higher correlation with the outcome of interest. In this thesis, the former approach was used.

Significance-based features selection

Significance-based feature selection methods deal with the computation of a performance metric [5, 6, 77]. Features are sorted based on this performance metric and only the subset of the best features is selected. The cut-off could be either a threshold in the metric or a number of features.

One example of significance-based feature selection (that was also used in this thesis) consist in fitting a univariate Cox regression model for each feature and using a z-test to test the hypothesis that that coefficient is significantly different from 0. A p-value if provided for each test. Features with coefficients that are not significantly different from 0 ($p>0.05$) are then excluded.

If the number of univariate model considered is high, correction for multiple hypothesis testing must be performed [5, 6]. Several correction methods exist but the most used in case of large features set is the False Discovery Rate or FDR approach [78] that adjusts each p-value based on the distribution of p-values for all the features in order to remove the significant values that are most likely to be false positives.

Wrapper feature selection

A wrapper feature selection method is a selection method that involves the use of a machine learning algorithm and a performance metric. Such methods could be classified in forward selection, backward selection algorithms or hybrid [68]. In forward feature selection, features are progressively added to the machine learning algorithm in order to maximize the

performance metric of interest. In backward feature selection, the model starts with all the features and features are progressively removed until the performance metric is maximized. The hybrid features selection includes a combination of features inclusion and removal.

Independently on the type of selection scheme that is used, there must be a criteria defined a-priori to add or remove the features and to determine the optimal set of features. Many different criteria can be used for the purpose [68].

In the context of this thesis, a forward approach was applied to the data and features were added to the model according to their C-index (see Subsection 2.5.7). The optimal number of features was identified by maximizing the C-index obtained by an internal validation of the data (performed via bootstrap, see Subsection 2.5.8).

Principal component analysis

Principal Component Analysis (PCA) is a linear transformation that maps the original features space to a new space called the component space [68]. The features of this new space are called components and each of them is a linear combination of all the original features.

The properties of the components make the PCA an optimal choice for dimensionality reduction. First of all, the components are linearly independent vectors, so if the number of instances is N, the number of components is at most $N - 1$, and this allows a first dimensionality reduction. Also, the algorithm of PCA identifies the components in such a way that the first component explains the majority of the variance, the second explains the majority of the variance left and so on (Figure 2.28). Therefore, a large portion of the variance in the data is explained by using the first components (10-20 components usually explain more than 90% of the total variance in datasets with hundreds of components). Considering only the first components causes a large reduction in data dimensionality at the cost a relatively small loss in the variance of the data [68].

The mathematical detail of how the components are computed is out of the scope of the thesis, but the reader may refer to [68] for further details.

2.5.7 Performance metrics for survival models

There are different ways to evaluate the performance of a survival model, but the ones that are used in clinical studies are mainly two: the Harrel's C-index [79, 80] and the log-rank test between risk groups [81].

The Harrel's C-index [79, 80] of a prognostic score is defined as the

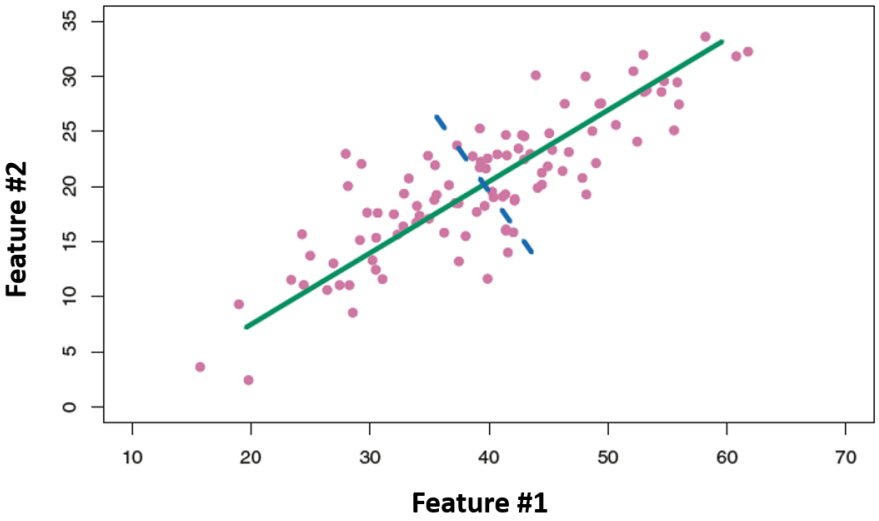

Figure 2.28: *Application of principal component analysis to a bi-dimensional dataset. The first component (green line) explains the majority of the variance in the data and the second component (blue dashed line) explains the remaining variance. It can be seen that the first component explains more variance than each of the original feature alone.*

probability that a the patient with the higher risk score s has a shorter time-to-event T:

$$C = Pr\{T_i < T_j | s_i > s_j\} \tag{2.20}$$

If all the patients in the dataset have experienced the event, Equation 2.20 can be used as it is. When some of the patients are censored, the computation of the probability must be done only on the pairs for which the concordance or discordance can be determined unambiguously [80].

Another way to evaluate the performance of a risk score is to define subgroups of patients based on thresholds of the score and to compare the Kaplan-Meier curves using the log-rank test [81]. An example is reported in Figure 2.29.

In the context of this thesis, two risk groups (high and low risk) were determined using the median value of the risk score in the training set as a threshold and the log-rank test was performed to detect significant differences in respective Kaplan-Meier curves.

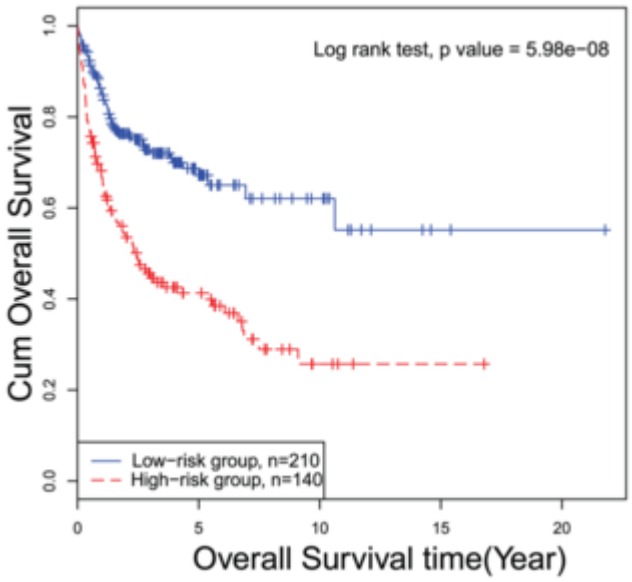

Figure 2.29: *High and low risk curves defined by thresholding of a risk score. According to the Kaplan-Meier test, the two curves are significantly different. In the figure, censored patients are marked by crosses.*

2.5.8 Model validation

As mentioned in Subsection 2.5.3, the evaluation of a prognostic model must not be done on the same data that were used to train the model, because that may give an overestimation of the quality of the results, especially if the model uses a lot of features. The ideal way of evaluating the performance of a prognostic model is to compute the risk score on a set of unseen patients called validation set and to use the performance metrics defined in Subsection 2.5.7 on that unbiased score.

The strongest validation of a model is provided when an independent external dataset is used, but that is not always available in practice. When just one dataset is available some techniques of internal validation (also called resampling methods) may still be used in order to provide an unbiased measurement of the performance of the model. Three main techniques exist for the purpose [68]: train-validation split, K-fold cross-validation and bootstrap.

Train-validation split

The train-validation split [68] approach consists in leaving a randomly selected portion of the original dataset out of the training process, creating

a separate train and validation sets. The percentage of left-out data may vary but it is typically a minority of the data (10-40%). The split may be stratified, meaning that the partition can be made in order to ensure that the proportion of events in the train and test is as similar as possible. This method is very simple but the results may depend on the particular split performed and so multiple iteration of train test should be performed to get the variance of the performance estimate [68]. Also, since machine learning methods tend to perform worse when trained on fewer observations, this suggests that the validation set performance metric may tend to underestimate the performance metric that would be computed on an independent dataset [68].

Bootstrap

Another way to perform internal validation is bootstrap [68]. Bootstrap consist in the creation of artificial training set by resampling with replacement of the original dataset (Figure 2.30). Each bootstrap training will have a different combination of the training data, some instances may be over-represented (i.e. appear more than once in the same training set) and some other may not be present at all. Each of the bootstrap model may be validated using the unselected patients obtaining a performance metric for each model. The array of performance metrics can be used to estimate a mean, standard deviation and confidence interval.

Figure 2.30: *Example of internal validation using bootstrap resampling. The training set are a combination with replacement of the original dataset. The validation sets consist of the left out instances for each iteration.*

In the context of this thesis bootstrap was used to provide a estimates of

the variance for the C-index for all the models that were tested.

K-fold cross-validation

In K-fold cross-validation [68], the initial dataset is divided in K partitions with approximately the same number of subjects each, called folds. In each of the K iteration of the algorithm, a different fold is used as the validation set while the other K-1 are used as the train set. The performance metric is evaluated K-times, every time on a different validation fold and the results are averaged to get the final performance estimate (Figure 2.31).

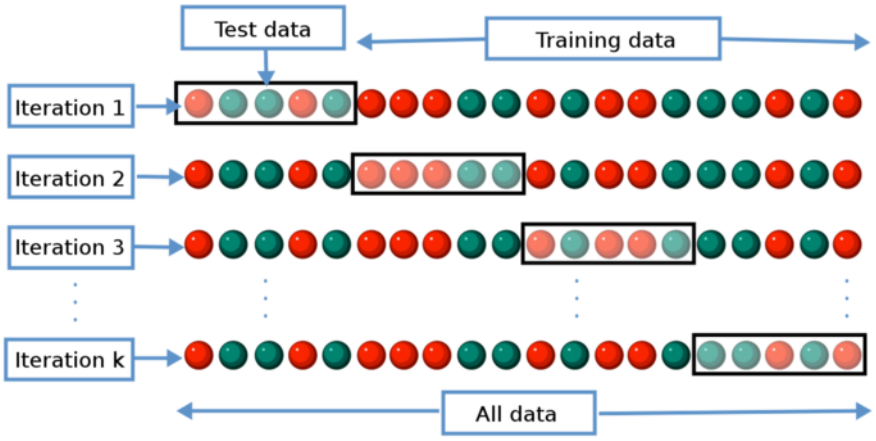

Figure 2.31: *Simple illustration of the K-fold cross-validation method.*

The advantage of K-fold is that it allows to obtain a lower variance in the estimate of the performance metric compare to train-test split [68]. Another advantage of K-fold cross-validation is that each subject in the initial datasets gets a unique unbiased estimate of the model score/prediction [68].

When the number of events per fold is low, computing separate metrics for each fold may lead to an overestimation of the variance of the results (and consequently, of the confidence interval). Since at the end of K-fold cross validation, an unbiased estimate of the model/prediction is available for each patient, such array of unbiased estimate can be used to compute the performance metric, and the confidence interval may be obtained through bootstrap.

CHAPTER 3

Stability analyses on a virtual phantom

This Chapter describes a series of experiments performed using simulated MRI to evaluate the stability of radiomic features to variations in image acquisition parameters and to sources of random noise. Effect of image preprocessing to features stability was also evaluated. Last, a set of radiomic features stable to imaging-related variability is identified.

3.1 Introduction

One of the main limitations of radiomics, that so far prevented its use in the clinical practice, is the dependence of radiomic features from image acquisition parameters/conditions [6, 8]. Sources of heterogeneity include, but are not limited to, systematic differences among scanners (e.g different reconstruction algorithms, different sensitivity of the instruments) or among centers (i.e. every center has its own image acquisition protocols), differences in the parameters that may be specific for the single acquisition (e.g. in order to increase signal-to-noise ratio or to reduce time of imaging when possible), or image artifacts (e.g. magnetic field inhomogeneity in MRI or metallic artifacts in CT). This heterogeneity may potentially affect the usefulness of radiomic features because the imaging-related variability

may mask the biology-related variability [6, 8]. This issue becomes particularly relevant for studies in which cohorts are multicentric or in which data are collected from multiple retrospective cohorts. Also, it becomes critical when MRI images are involved, since the measured signal is not tissue-specific and may strongly vary from acquisition to acquisition due to different acquisition parameters such as scanner, pulse sequence, TR/TE, pixel spacing and slice thickness [82].

Image preprocessing may be a way to reduce heterogeneity in the images and, consequently, to reduce heterogeneity in the radiomic features [16, 83–85]. Among these preprocessing techniques, the following may be cited [67]: voxel spatial resampling; intensity discretization; denoising; intensity standardization and intensity inhomogeneity correction (in MRI). The effect of these preprocessing techniques on features stability has been partially investigated in previous studies of literature [16, 83–85].

The assessment of feature stability is often performed through the use of a phantom, i.e. an object with known properties that is scanned by one or multiple machines in different conditions. Since the phantoms are typically described in the articles in which they are used (such in [86]), they are useful and reproducible tools to assess variability to imaging conditions, allowing to isolate the impact of the different acquisition parameters. To further increase the reproducibility of phantom results, different online resources have been provided, such as the open-source CT phantom described in [87], or the virtual MRI simulators BrainWeb [88].

Many phantom-based variability analyses have been reported in literature [16, 82, 84–86, 89–91]. Part of those studies investigated the effect of image preprocessing on features stability [16, 84, 85]. Most of the phantom studies focused on CT and PET [16, 84–86, 89, 90], with only a few studies related to MRI instead [82, 91]. Among those few, none investigated the effect of image preprocessing in the improvement of features stability.

The purpose of this chapter is to use virtual MRI simulations, obtained using Brainweb, to test the stability of radiomic features to four different sources of variability: a) intensity variations caused by modification in TR/TE; b) changes in voxels size due to modification in slice thickness and in-plane pixel spacing; c) image noise; d) intensity non-uniformity generated by inhomogeneity in the local magnetic field. The analysis focused on T1w and T2w MRI obtained with SE pulse sequence (see Subsection 2.3.4). Those type of images, and variants obtained with TSE are routinely performed in clinical practice for observation of both HNC and STS and can be obtained using BrainWeb software. The experiments described in the following sections were also part of publications [92, 93].

3.2 Materials and methods

3.2.1 BrainWeb simulated datasets

All the images dataset used in this chapter were obtained using the BrainWeb [88, 94], a virtual MRI simulator. Starting from 3 volumes map representing values of T1, T2 and PD (estimated using a 1.5 T Philips Gyroscan scanner), the simulator allows to create customized volumes using signal equations characterizing the different RF pulse sequences (Figure 3.1). By defining the pulse sequence and the proper image acquisition parameters, the user can define potentially infinite image acquisitions. Last, Brainweb allows to set other parameters not strictly related to the pulse sequence, such as the slice thickness, the level of noise and the level of intensity non-uniformity.

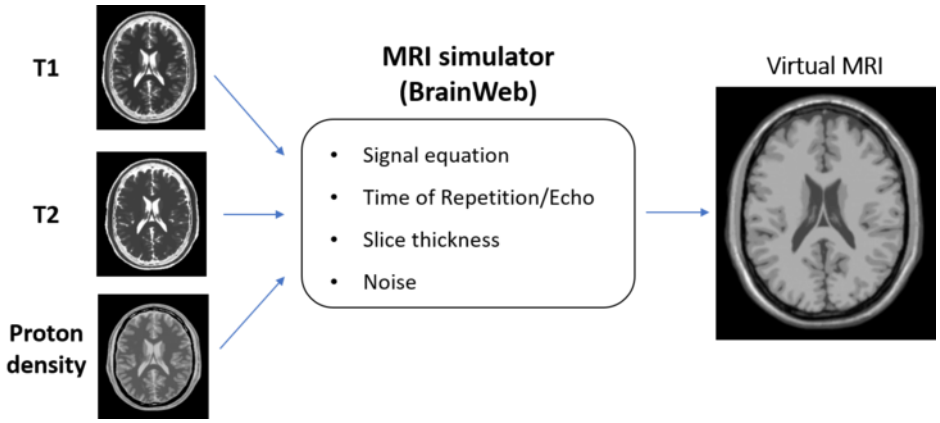

Figure 3.1: *Illustration of the process of creation of a virtual MRI using the BrainWeb simulator.*

In the context of this chapter, the pulse sequence used is the SE pulse sequence, whose signal equation is listed in Subsection 2.3.4, equation 2.5. The parameters that were varied for the creation of customized images were the following: a) time of repetition; b) time of echo; c) slice thickness; d) pixel spacing; e) noise level; f) intensity non-uniformities (INU).

The noise was modeled as a Gaussian with mean 0 and a standard deviation controlled by the user using the *noise level* parameter. Such parameter is the ratio (expressed in percentage) between the standard deviation of the noise and the average intensity of a reference tissue in the phantom (by default, the brightest tissue).

The type of non-uniformity is controlled by the *INU field* parameter,

which allows to select among 3 different non-uniformity fields. Also, the entity of the non-uniformity is controlled by the *INU level* parameter, which is the ratio (in percentage) between the range of non-uniformity and the intensity of a reference tissue (by default, the brightest tissue).

Brainweb was used to simulating 5 different datasets comprising both T1w and T2w MRI obtained using SE pulse sequence. Each dataset was used for a particular type of stability analyses as illustrated in Figure 3.2. DS1 was used to perform analyses of stability to variations in TR/TE and to select the best intensity standardization algorithm to increase stability. DS2 was used to evaluate the effect of voxel size resampling on features stability. Analyses on the effect of image denoising on features stability were performed on DS3. Effect of bias field correction was evaluated on DS4. Last DS5 was used to select a set of features that were stable to random variations of the aforementioned image acquisition parameters/conditions. All the datasets could be created directly from Brainweb except for DS2. As a matter of fact, Brainweb does not allow to control the pixel spacing, which is set to a fixed value of 1 mm.Therefore, the in-plane resampling was performed in MATLAB 2018a (Mathworks, Natick, MA, USA).

Tables 3.1-3.5 give details about the image acquisition parameters for the individual datasets. Variations in the image acquisition parameters were set in range similar to the ones observed in clinical practice. The maximum value for noise and INU levels were set to 9% and 40 % respectively, which were values higher or equal compared to the ones used in other studies of literature involving images obtained with BrainWeb [95, 96].

3.2.2 Regions of interest

The stability analyses described in the following subsections were performed on either one of two ROI datasets referred to as *ROI-rect* and *ROI-bio*.

The *ROI-rect* set was composed by 3 rectangular ROIs (Figure 3.3A) that were segmented on one of the acquisitions (TR=500 ms, TE=9 ms, isotropic voxel size 1 mm) using the open source software 3D Slicer [97]. These regions were chosen in order to include areas of different levels of heterogeneity and size, similarly to what was done in [82]. The ROIs were segmented for 10 consecutive slices.

The *ROI-bio* set was obtained directly from the BrainWeb website [98]. In particular, 9 3D ROIs of different sizes representing different tissues (cerebrospinal fluid, grey matter, white matter, fat, muscle, skin, skull, glial matter and connective tissue) were considered. Since the ROIs were very

large, to reduce the computational complexity of the radiomic features extraction, only the 11 central slices of each ROI were used. An example of segmented ROI (white matter) is reported in Figure 3.3B.

Since all the images simulated with BrainWeb shared the same physical space, it was possible to use the same ROIs for all the MRI acquisitions, without the need of repositioning them in each dataset.

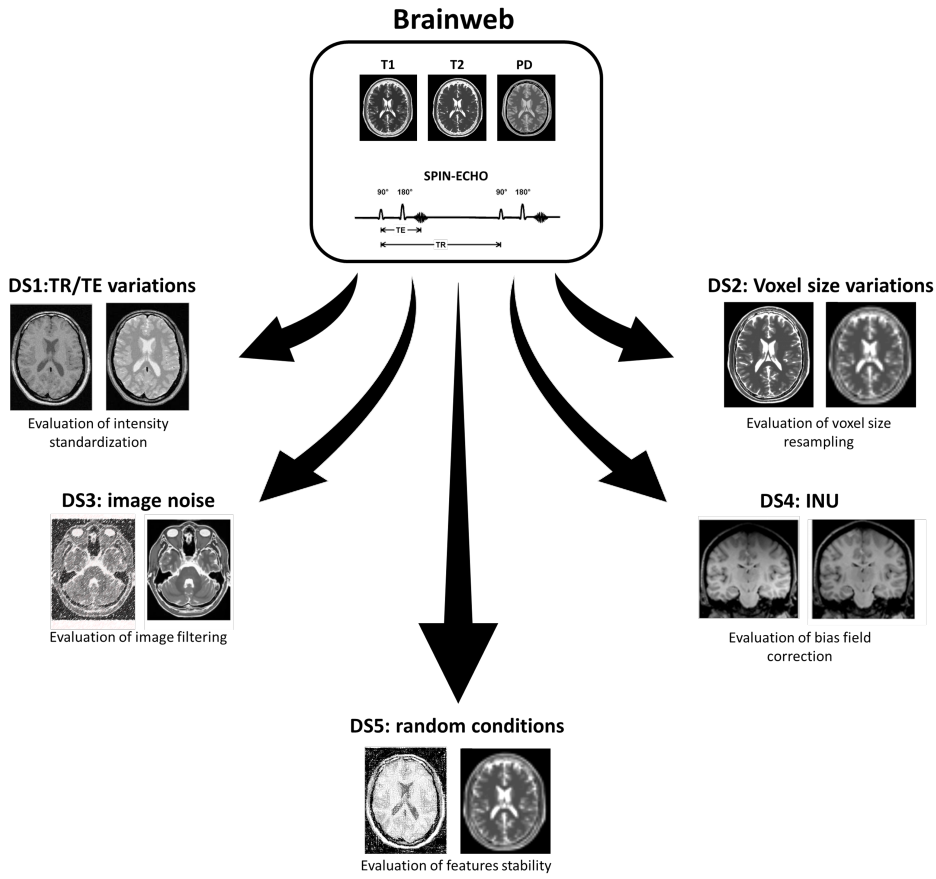

Figure 3.2: *Representation of the 5 dataset created with Brainweb and their purpose. TR: time of repetition. TE: time of echo. INU: intensity non-uniformities.*

3.2.3 Radiomic features extraction

In all the analyses performed with the virtual phantoms (and in general for all the analyses described in the dissertation), the radiomic features extraction was performed using a MATLAB wrapper of the open-source software Pyradiomics (version 2.1.0) [99].

ACQUISITION PARAMETERS (DS1)		
Image sequence	**T1w**	**T2w**
Number of images	42	48
Pulse sequence	Spin-echo	Spin-echo
Magnetic field	1.5 T	1.5 T
Time of repetition	- Range: 350-650 ms - Step: 50 ms	- Range: 2000-9000 ms - Step: 1000 ms
Time of echo	- Range: 5-15 ms - Step: 2 ms	- Range: 80-130 ms - Step: 10 ms
Slice thickness	1 mm	1 mm
Pixel spacing	1 mm	1 mm
Noise level	0 %	0 %
Intensity non-uniformity	None	None

Table 3.1: *Image acquisition parameters for the T1-weighted (T1w) and T2-weighted (T2w) images of the DS1 dataset. Varying parameters are highlighted in red and expressed by their range of values and step.*

The categories of features that were considered for analyses were shape and size (14 features), FOS (18 features) and textural features (75 features). Textural features were calculated from the following matrices: GLCM (24 features); GLRLM (16 features); GLSZM (16 features); NGTDM (5 features); GLDM (14 features). Also, for FOS and textural features, the features could also be computed for the 8 volumes resulting from the first level wavelet decomposition. For a better description of the features refer to Section 2.4 or [5, 24, 66].

Prior to performing the features extraction, the discretization of the histogram of the grey values is performed. Instead of the fixed bin size discretization (the default of Pyradiomics), which may not be the best choice in case of images with arbitrary intensity units such as MRI [67], a fixed bin number intensity discretization was used. In particular, a 32 bins histogram discretization was used, as done in a previous study of MRI [100].

3.2.4 Metric for stability quantification

Before explaining the analyses performed on the different dataset it is important to clearly define the metric that was used to quantify the stability

ACQUISITION PARAMETERS (DS2)		
Image sequence	**T1w**	**T2w**
Number of images	28	28
Pulse sequence	Spin-echo	Spin-echo
Magnetic field	1.5 T	1.5 T
Time of repetition	500 ms	6000 ms
Time of echo	9 ms	100 ms
Slice thickness	- Range: 1-7 mm - Step: 1 mm	- Range: 1-7 mm - Step: 1 mm
Pixel spacing	- Range: 1-4 mm - Step: 1 mm	- Range: 1-4 mm - Step: 1 mm
Noise level	0 %	0 %
Intensity non-uniformity	None	None

Table 3.2: *Image acquisition parameters for the T1-weighted (T1w) and T2-weighted (T2w) images of the DS2 dataset. Varying parameters are highlighted in red and expressed by their range of values and step.*

ACQUISITION PARAMETERS (DS3)		
Image sequence	**T1w**	**T2w**
Number of images	10	10
Pulse sequence	Spin-echo	Spin-echo
Magnetic field	1.5 T	1.5 T
Time of repetition	500 ms	6000 ms
Time of echo	9 ms	100 ms
Slice thickness	1 mm	1 mm
Pixel spacing	1 mm	1 mm
Noise level	9 %	9 %
Intensity non-uniformity	None	None

Table 3.3: *Image acquisition parameters for the T1-weighted (T1w) and T2-weighted (T2w) images of the DS3 dataset. Varying parameters are highlighted in red.*

ACQUISITION PARAMETERS (DS4)		
Image sequence	**T1w**	**T2w**
Number of images	4	4
Pulse sequence	Spin-echo	Spin-echo
Magnetic field	1.5 T	1.5 T
Time of repetition	500 ms	6000 ms
Time of echo	9 ms	100 ms
Slice thickness	1 mm	1 mm
Pixel spacing	1 mm	1 mm
Noise level	0 %	0 %
Intensity non-uniformity	- 3 inhomogeneity fields - 1 reference - INU level: 40 %	- 3 inhomogeneity fields - 1 reference - INU level: 40 %

Table 3.4: *Image acquisition parameters for the T1-weighted (T1w) and T2-weighted (T2w) images of the DS4 dataset. Varying parameters are highlighted in red. INU: intensity non-unifomity.*

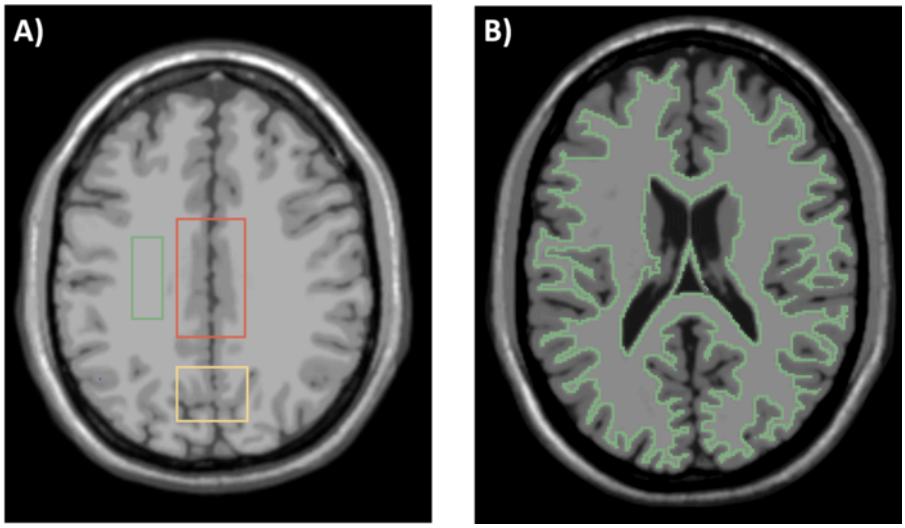

Figure 3.3: *Examples of the two classes of region of interest (ROI) used for the analyses, superimposed on a reference T1-weighted image. A) ROI-rect set. B) Segmentation of the white matter belonging to the ROI-bio set.*

ACQUISITION PARAMETERS (DS5)		
Image sequence	**T1w**	**T2w**
Number of images	50	50
Pulse sequence	Spin-echo	Spin-echo
Magnetic field	1.5 T	1.5 T
Time of repetition	350-650 ms	2000-9000 ms
Time of echo	5-15 ms	80-130 ms
Slice thickness	1-7 mm	1-7 mm
Pixel spacing	1-4 mm	1-4 mm
Noise level	0-9 %	0-9 %
Intensity non-uniformity	0-40 % (3 INU fields)	0-40 % (3 INU fields)

Table 3.5: *Image acquisition parameters for the T1-weighted (T1w) and T2-weighted (T2w) images of the DS5 dataset. Each of the 50 images is a random combination of the parameter in the table within the listed ranges. INU: intensity non-uniformity.*

in all the experiments: the Intra-class Correlation Coefficient (ICC) [101]. Given a n-by-k matrix of values representing one feature measured from n different instances (e.g. the different ROIs) on k different conditions (e.g. the different imaging conditions), the ICC quantifies the agreement between corresponding measurements in the different conditions. If a feature has an ICC of 1, it means that the changes in the factor of interest caused no changes in the features, otherwise the lower the value of ICC the lower the stability of the feature.

There are different types of ICC [101]. The one used for all the analyses of this dissertation was the one measuring the agreement in a two-way random effects model, equivalent to the (A,1) model described in [101], and it is computed as follows:

$$ICC(A,1) = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)} \tag{3.1}$$

Where MS_R, MS_C and MS_E are the between-ROI (row), between-acquisitions (column) and residual (else) mean squares respectively.

Although there is no clearly defined value of ICC to distinguish between stable and unstable features, in [102] the threshold of 0.75 is used to define

good stability, and thus this value was adopted for all the stability analyses in this thesis. The computation of the ICCs, as well as other analyses presented in the thesis, were performed in MATLAB.

3.2.5 Identification of the best intensity standardization algorithm

The purpose of this first analysis was to identify the best intensity standardization to increase the stability of radiomic features to changes in TR/TE [92]. The analysis was performed on the set of T1w images of the DS1 dataset using the ROIs of the *ROI-rect* set.

The following analysis was limited to FOS and textural features based on GLCM and GLRLM, which are the most common categories of features used for the radiomic analyses. SS features are not affected by TR and TE, but only depend on the ROI, and were therefore excluded from the analysis. In total, 58 features were considered.

Three different intensity standardization algorithms were tested: histogram stretching, Z-score normalization, and histogram matching. These techniques were chosen because are widely known and have already been used as a preprocessing step in MRI studies [103–105].

Histogram stretching consists in a linear mapping of an intensity range $[I_{Min}; I_{Max}]$, which is image-specific, to a new intensity range $[I_{NewMin}; I_{NewMax}]$, which is defined by the user and that is independent on the particular image. The mapping is performed according to the following equation [103]:

$$I_{Norm} = (I - I_{Min})\frac{I_{NewMax} - I_{NewMin}}{I_{Max} - I_{Min}} + I_{NewMin} \tag{3.2}$$

where I is the original intensity value of a specific voxel and I_{Norm} is the same intensity after the normalization. The value I_{Min} and I_{Max} do not need to be the maximum and minimum, but just a low and high intensity reference. For the experiment, I_{Min} and I_{Max} were actually defined as the quantiles 0.02 and 0.98 of the distribution of intensity. These quantiles were used because they are less sensitive to image noise compared to maximum and minimum. Values of I_{NewMin} and I_{NewMax} were set to 0 and 5000 respectively. This range was chosen so that it was larger than any other range of intensities observed for the MRI images, in order not to cause any loss of information due to quantization.

Z-score normalization (the deafault method in Pyradiomics) is another linear intensity normalization technique which consist in standardizing the distribution of grey values in the MRI image by subtracting the mean μ and dividing by the standard deviation σ [105]:

$$I_{Norm} = \frac{I - \mu}{\sigma} \tag{3.3}$$

Histogram matching is a standardization techniques that consist in non-linearly changing the grey values so that the histogram of the MRI image of interest is made as close as possible to the one of a reference MRI image. The detailed algorithm for histogram matching is described in [24]. In this experiment, the histogram of the MRI displayed in Figure 3.3 was used as the reference histogram.

Radiomic features were extracted for the 3 rectangular ROIs and for all the same acquisitions of the phantom. This latter operation was performed 4 times, one for the original DS1 dataset and the other 3 for intensity-standardized version of DS1 (one for each algorithm). By performing stability analyses on all the datasets, it was possible to obtain 3 different n-by-4 matrix of ICCs, one for each category of feature (FOS, GLCM and GLRLM). In each of the matrix, the columns represented the arrays of ICCs obtained with the 4 different standardization techniques (no standardization, histogram stretching, histogram matching and Z-score).

In order to evaluate the presence of significant differences on features stability due to the method of intensity standardization, a Friedman test was applied to each ICC matrix. In order to evaluate which groups presented significant differences between each other, post-hoc comparisons with two-sided Wilcoxon signe-rank tests and Tukey-Kramer correction for multiple testing were performed.

3.2.6 Effect of intensity standardization on features stability

The purpose of this analysis was to identify whether intensity standardization could increase the stability of radiomic features to changes in TR/TE [93]. The analysis was performed on both T1w and T2w MRI of the DS1 dataset using the ROIs of the *ROI-bio* set. FOS and all the textural features were used for this analysis, for a total of 93 features.

In order to evaluate the effect of intensity standardization of features stability to TR/TE variations, radiomic features were extracted for the 9 ROIs of the *ROI-bio* set, using both a standardized and a non-standardized version of the DS1 dataset. The method that was chosen as the best based on the previous analysis (Subsection 3.2.5) was used for intensity standardization. At the end of the analyses, 4 different n-by-2 ICC matrices were obtained, one for each combination of features category (FOS, and textural) and image sequence (T1w and T2w).

In order to statistically evaluate the improvement on features stability

due to the intensity standardization, a two-sided Wilcoxon signed rank test was applied to each ICC matrix.

3.2.7 Effect of voxel size resampling on features stability

The purpose of the analysis performed on DS2 was to understand whether voxel size resampling to a common isotropic resolution could be used to increase the stability to differences in pixel spacing and slice thickness [93]. Figure 3.4 shows and example of T2w images from the DS2 dataset.

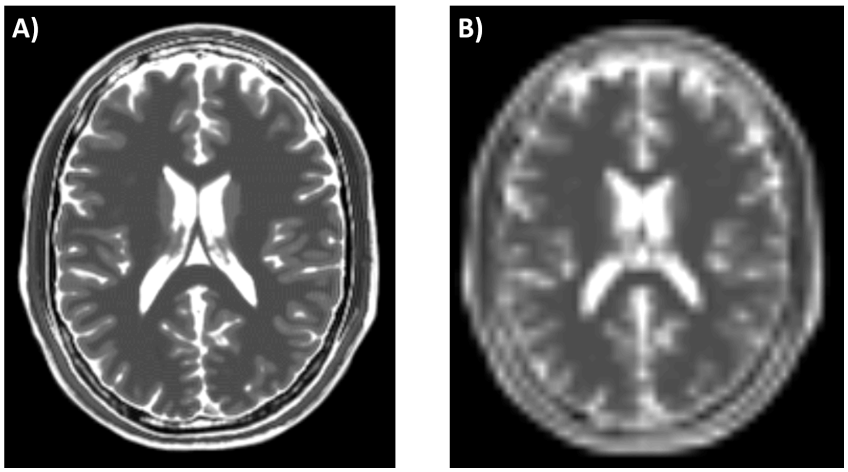

Figure 3.4: *Axial slice of a T2-weighted image of DS2. A) high resolution image. B) Low resolution image.*

In total, 107 features belonging to FOS, textural and shape and size categories were extracted from each of the 9 ROIs of the *ROI-bio* set, and for each image type (T1w or T2w). Features were extracted before and after resampling the voxel size to an isotropic resolution of 1 mm. B-spline interpolation was used to resample the images.

At the end of the analysis, 6 different n-by-2 ICC matrices were obtained, one for each combination of features category (FOS, textural, shape and size) and image sequence (T1w and T2w), and a two-sided Wilcoxon signed-rank test was applied to each matrix, to highlight statistically significant differences between ICC values before and after the resampling.

3.2.8 Effect of image denoising on features stability

Since the presence of random image noise may reduce the stability of radiomic features, dataset DS3 was used to investigate whether image fil-

tering could be used to increase the stability of radiomic features to such noise [93]. Examples of noisy images in DS3 are illustrated in Figure 3.5.

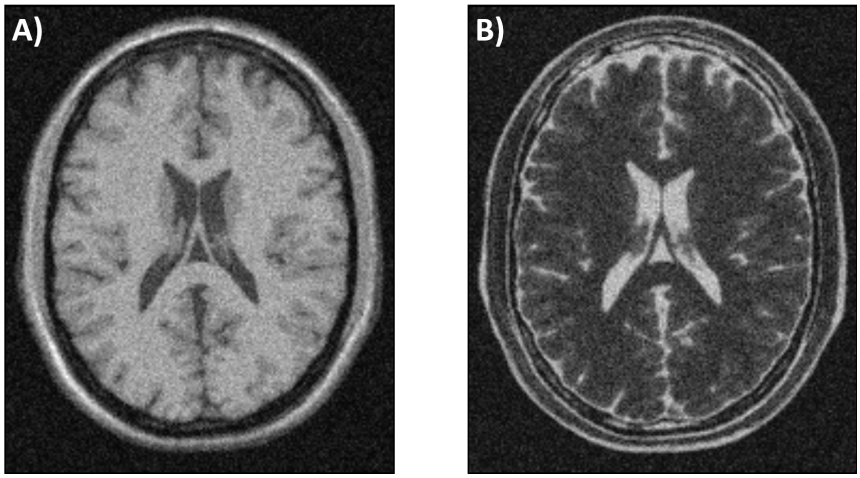

Figure 3.5: *Examples of nosiy MRI images from DS3 (noise level 9%). A) T1-weighted image. B) T2-weighted image.*

For this analysis, the same features set described in Subsection 3.2.6 was used, since image noise is not supposed to influence the shape features. The *ROI-bio* set was used for the extraction. Features extraction and ICC computation were performed twice (before and after the filtering). The denoising of the images was performed by a 3D Gaussian filter 3x3x3 voxel kernel and $\sigma = 0.5$ obtained by the *imgaussfilt3* MATLAB function.

At the end of the analysis, 4 different n-by-2 ICC matrices were obtained, one for each combination of features category (FOS, textural) and image sequence (T1w and T2w), and a two-sided Wilcoxon signed-rank test was applied to each matrix, to highlight statistically significant differences in the ICC due to Gaussian filtering.

3.2.9 Effect of bias field correction on features stability

The analysis performed on DS4 was used to investigate whether bias field correction could be used to increase the stability of radiomic features to INU caused by magnetic field inhomogeneities [93]. Example of images with inhomogeneity fields are reported in Figure 3.6.

For this analysis, the features set described in Subsection 3.2.6 was used, and the features were extracted from the ROI of the *ROI-bio*. Features extraction and ICC computation were performed twice (before and after the correction). The bias field correction was performed using N4ITK [106].

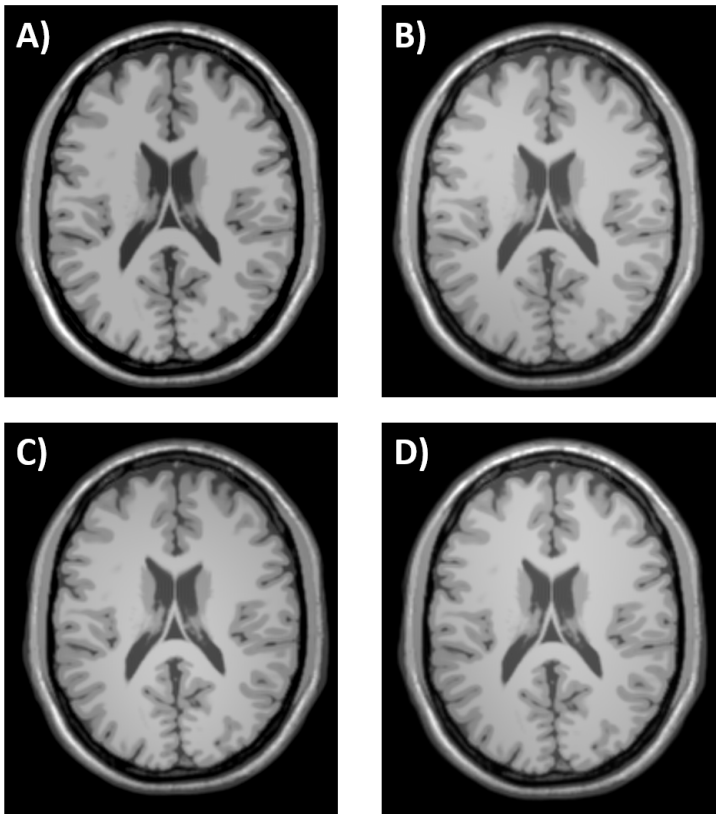

Figure 3.6: *Example of T1-weighted MRI images with the addition of intensity non-uniformities (INU). The INU level was set to 40% of the brightest tissue. A) Reference image without non-uniformities. B-C-D) Same T1-weighted MRI with different non-uniformity fields.*

N4ITK is an improvement of the popular N3 algorithm for inhomogeneity correction [107] and it has been implemented using ITK framework [108].

At the end of the analysis, 4 different n-by-2 ICC matrices were obtained and a two-sided Wilcoxon signed-rank test was applied to each matrix, to statistically evaluate the effect of bias field correction in each category of features and for each image sequence.

3.2.10 Stable features identification

The purpose of the analysis performed on DS5 was to investigate which features are stable enough to be used even in situation in which the image acquisition parameters may vary.

The set of 1072 radiomic features (536 per image type) used for this

analyses included the shape and size features, the FOS and some of the textural features (GLCM and GLRLM based), computed for both the original image and the wavelet transforms. These categories of features were also the ones considered for the following chapters. The choice of this features set is due to the fact that they are included in almost all the software used for radiomic features extraction and so signature obtained with such features could be easily reproduced.

Images of DS5 were acquired with random combination of image acquisition parameters withing the range that can be encountered in the clinical practice, and to reduce the imaging related variability the images were preprocessed with the best combination of preprocessing steps, defined according to the results of subsections 3.2.5-3.2.9. Radiomic features were extracted from such preprocessed images. The ROIs of the *ROI-bio* set were used.

After the stability analyses it was possible to obtain a value of ICC for each feature and to classify it in stable (ICC>0.75) or unstable.

3.3 Results

3.3.1 Identification of the best intensity standardization algorithm

The boxplots in Figure 3.7 present the values of ICC for the different features classes, grouped by type of intensity standardization algorithm used. Asterisk and triangles represent significantly increases or decreases compared to the baseline situation (no intensity standardization).From Figure 3.7A it can be seen that all the intensity standardization algorithm significantly improved the stability of the FOS features ($p=8.12*10^{-6}$ for Friedman Test, $p<0.02$ in corrected post-hoc comparisons). Histogram matching showed a higher median values but the difference with the other standardization algorithm was non significant ($p>0.16$ in post-hoc comparisons). From Figure 3.7B-C it can be seen that the intensity standardization had a lower effect on the stability of textural features. However, histogram matching, caused a small but sistematic reduction ($p=8.36*10^{-9}$ in corrected post-hoc comparisons) in the values of ICCs for GLCM features.

From these results it is clear how intensity standardization plays an important role in increasing the stability of FOS features. Also, histogram matching reduced the stability of GLCM features and therefore should be excluded from future analyses. The other two methods were equivalent. In the following analyses, Z-score was used since it is the default method for intensity standardization in Pyradiomics. This also helps increasing the re-

producibility of the stability analyses shown in this and in the next chapters.

3.3.2 Effect of intensity standardization on features stability

Figure 3.8 shows the boxplots (grouped by features category and image type) with the values of ICC for the features extracted from the original and standardized images (light and dark blue boxplots respectively). FOS features extracted from the ROI-bio set were more stable compared to the features extracted from the *ROI-rect* set used in the previous experiment. As a matter of fact, almost all the values of ICC computed were above the threshold of 0.75 (dashed black line in Figure 3.8A). The effect of intensity standardization however was the same: Z-score normalization caused a significant increase in FOS features stability (T1w: median increase 0.09 [0.06-0.11], p = $7.37*10^{-4}$; T2w: median increase 0.11 [0.05-0.14], p=$8.44*10^{-3}$). Intensity standardization also caused a significant decrease in stability to variations of TR/TE for the textural features (T1w: median decrease $5.00*10^{-4}$ [$2.30*10^{-5}$-$1.50*10^{-3}$], p=$3.67*10^{-7}$; T2w: median decrease $1.40*10^{-4}$ [3.21 *10^{-5}-$1.66*10^{-3}$], p=$1.93*10^{-11}$). However, the reduction in stability was very small and there was no change in the number of unstable features.

3.3.3 Effect of voxel size resampling on features stability

Figure 3.9 shows boxplots representing the values of ICCs of the features obtained when the same phantom was acquired with different voxel sizes. Values are shown for both T1w and T2w MRI, and for both original and isotropically resampled images. It can be seen that heterogeneity in voxel size reduced the values of ICCs of textural features in particular, with almost all being below 0.75. SS and FOS features were also affected but most of the values of ICCs were above 0.75.

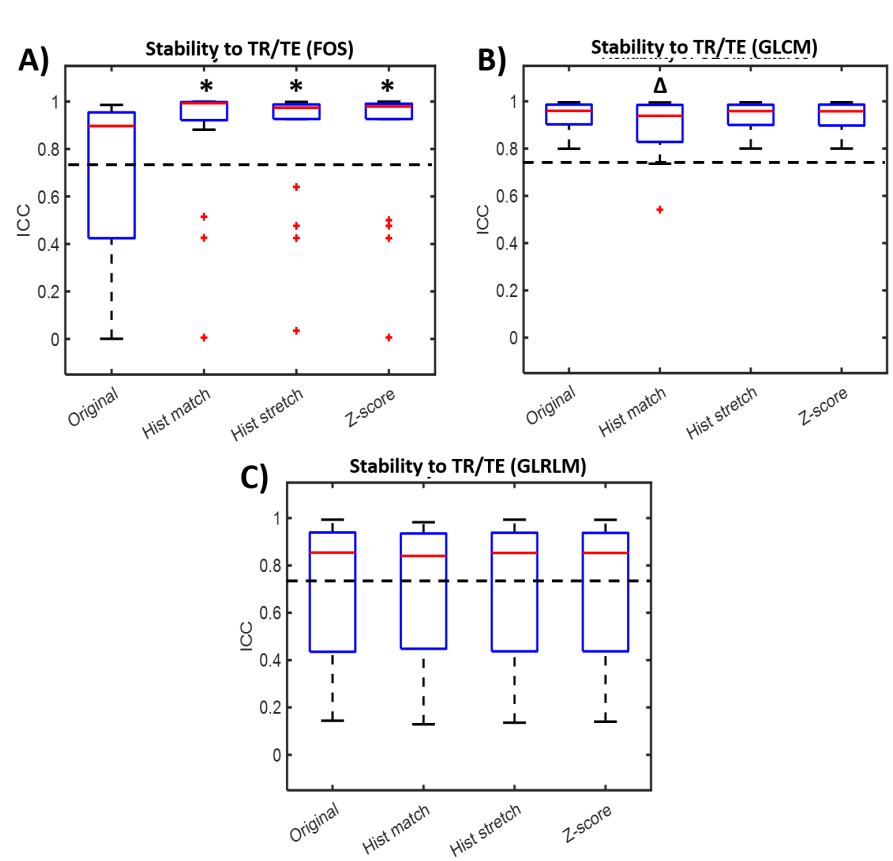

Figure 3.7: *Boxplots showing the values of Intra-class Correlation Coefficient (ICC) quantifying stability to variations in Time of Repetition (TR) and Time of Echo (TE) for the different features categories and intensity standardization algorithms. Significant improvements in stability compared to the original images are marked with asterisks, while significant reductions are marked in triangles. A) First Order Statistics (FOS) features. B) Textural features based on Grey Level Co-occurrence Matrix (GLCM). C) Textural features based on Grey Level Run Length Matrix (GLRLM). The dashed line represent the threshold of ICC=0.75*

From Figure 3.9A it can be seen that harmonizing the voxel size significantly improves the stability of the extracted features (T1w/T2w: median increase 0.02 [$2.20*10^{-3}$-0.09], p = $1.22*10^{-4}$), for which the values of ICC went back to 1. Resampling to uniform resolution also increased the stability of and textural features in both T1w and T2w MRI (Figure 3.9C). The effect was statistically significant only for T1w MRI (median increase 0.14 [0.01-0.28], p = $1.87*10^{-7}$) but not in T2w MRI (median increase 0.01

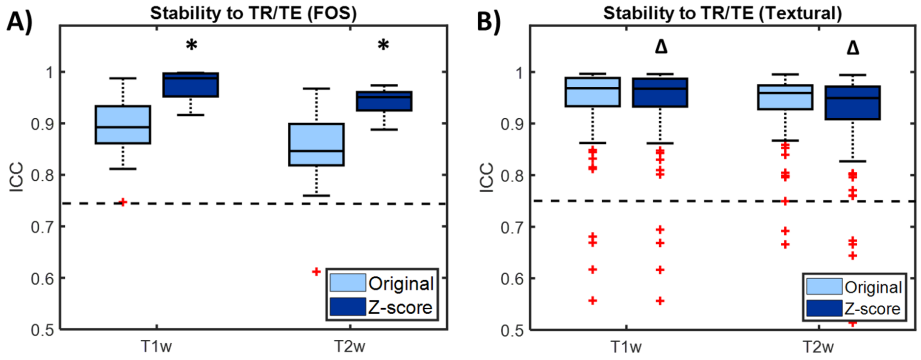

Figure 3.8: *Boxplots showing the values of Intra-class Correlation Coefficient (ICC) quantifying stability to variations in Time of Repetition (TR) and Time of Echo (TE) for the different features categories and image types. Significant improvements due to Z-score normalization are marked with asterisks, while significant reduction are marked in triangles. A) First Order Statistics (FOS) features. B) Textural features. The dashed line represent the threshold of ICC=0.75.*

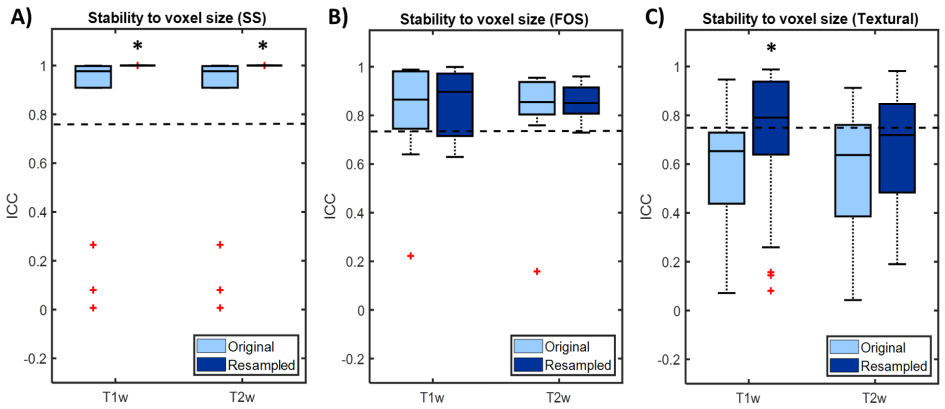

Figure 3.9: *Boxplots showing the values of Intra-class Correlation Coefficient (ICC) quantifying stability to variations in voxel size for the different features categories and image types. Significant improvements in stability due to resampling to uniform voxel size are marked with asterisks. A) Shape and Size (SS) features. B) First Order Statistics (FOS). C) Textural features. The dashed line represent the threshold of ICC=0.75.*

[-0.11 to 0.18], p = 0.17). As seen in Figure 3.9B, stability of FOS features is not improved by resampling (T1w median increase -0.01 [-0.03 to 0.01], p = 0.21; T2w median increase -0.03 [-0.04 to 0.01], p = 0.29)

3.3.4 Effect of image denoising on features stability

Figure 3.10 shows boxplots representing the values of ICC of the features when the same phantom is acquired with different random noise. Values are shown for both T1w and T2w images, and for both original and Gaussian filtered images.

From Figure 3.10 it can be seen that noise did not have much effect on the stability of radiomic features. In fact, most of the features were above the threshold of stability even before denoising. Gaussian filtering caused a small but significant increase in ICC for both FOS (T1w: median increase $7.76*10^{-4}$ [$4.32*10^{-6}$-$4.50*10^{-3}$], p = $1.96*10^{-4}$; T2w: median increase $1.84*10^{-4}$ [$5.80*10^{-6}$-$7.90*10^{-3}$], p = $3.26*10^{-4}$) and textural features (T1w: median increase $7.70*10^{-3}$ [$5.17*10^{-5}$-0.03], p = $8.53*10^{-6}$; T2w: median increase $2.10*10^{-3}$ [$-1.00*10^{-3}$ to $9.30*10^{-3}$], p = 0.03).

3.3.5 Effect of bias field correction on features stability

Figure 3.11 shows boxplots representing the values of ICC of the features when the same phantom was acquired with different intensity non-uniformity fields. Values are shown for both T1w and T2w images, and before and after the application of bias-field correction.

From Figure 3.11 it can be seen that the majority of features were above the threshold of stability. In both T1w and T2w MRI, the non-uniformity correction significantly increased ICC values of both FOS (T1w: median increase 0.09 [$6.40*10^{-3}$-0.18], p = $5.35*10^{-4}$; T2w: median increase $7.40*10^{-3}$ [$5.50*10^{-3}$-0.02], p = $1.96*10^{-4}$) textural features (T1w: median increase 0.05 [0.02-0.15], p = $3.38*10^{-11}$; T2w: median increase 0.01 [$4.90*10^{-3}$-0.04], p = $2.48*10^{-11}$).

3.3.6 Stable features identification

After the images of the DS5 underwent the best preprocessing pipeline (Gaussian filtering, bias-field correction, resampling and intensity standardization, in this order), an ICC was computed for each features and was used to classify the images in stable or unstable. Figure 3.12 show some pie charts displaying the proportions of stable and unstable features (in blue and orange respectively). It is possible to see that, for both image types,

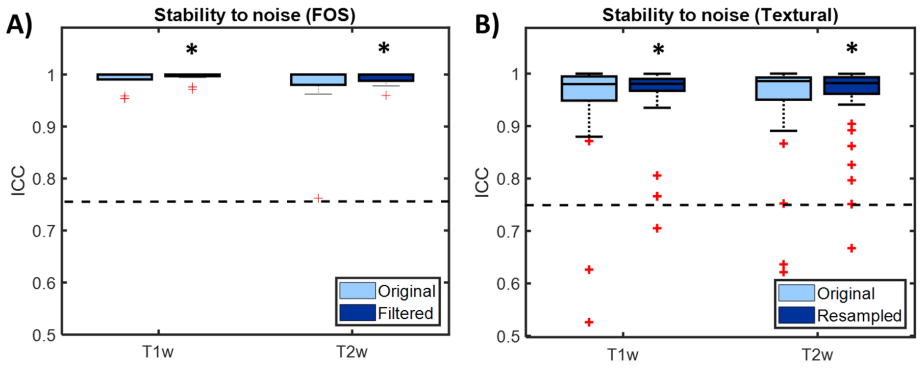

Figure 3.10: *Boxplots showing the values of Intra-class Correlation Coefficient (ICC) quantifying stability to image noise for the different features categories and image types. Significant improvements in stability due to Gaussian filtering are marked with asterisks. A) First Order Statistics (FOS). B) Textural features. The dashed line represent the threshold of ICC=0.75.*

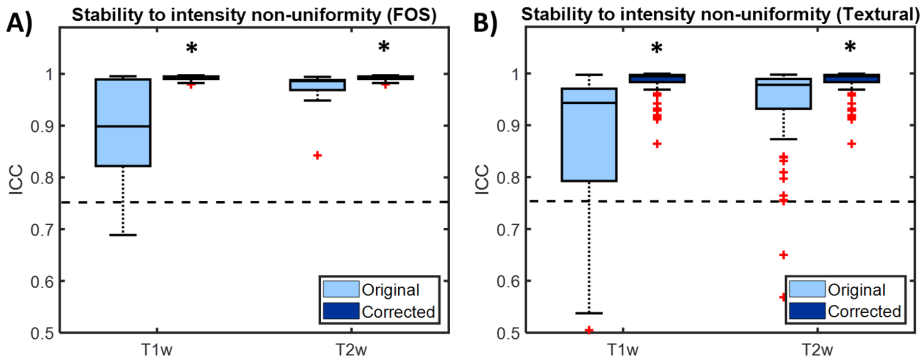

Figure 3.11: *Boxplots showing the values of Intra-class Correlation Coefficient (ICC) quantifying stability to intensity non-uniformities for the different features categories and image types. Significant improvements in stability due to N4ITK algorithm are marked with asterisks. A) First Order Statistics (FOS). B) Textural features. The dashed line represent the threshold of ICC=0.75.*

only around half of the 536 radiomic features was considered stable (49.63 % and 52.99 % for T1w and T2w respectively).

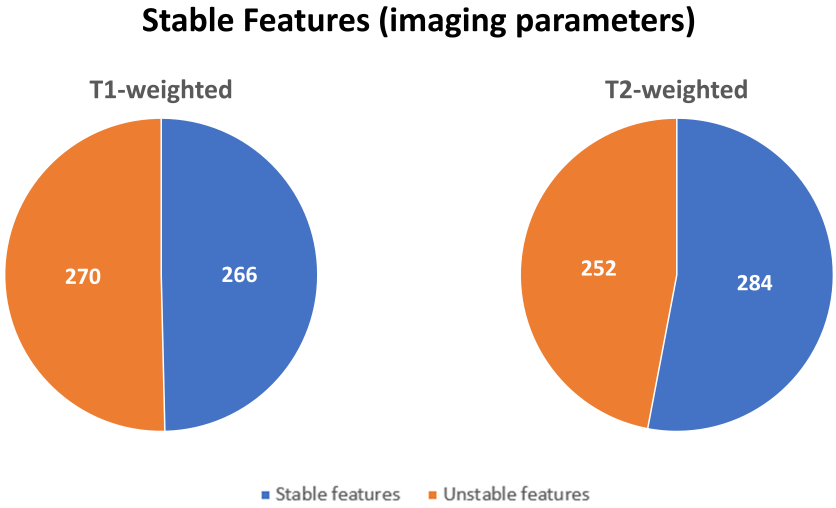

Figure 3.12: *Pie charts showing the proportion of stable radiomic features for T1-weighted and T2-weighted images.*

Figure 3.13 represents the same stable features, grouped by features class. From the figure it is possible to notice how SS features were the one with the best stability since all of them were selected. FOS features also showed good stability properties, with 15 out of 18 features (83.33 %) being classified as stable for both T1w and T2w images. The proportion of stable textural features, although lower compared to FOS or SS, is still high, with better results for T1w images compared to T2w images (31 and 26 features respectively). Wavelet are the features with the worst properties, with less than half of them being classified as stable (38.43 % and 42.72 % for T1w and T2w respectively).

3.4 Discussion

The goal of the analyses performed on the virtual phantom was to understand whether and how much image preprocessing could help in increasing the stability of radiomic features (quantified by ICC) to variations in quantitative image acquisition parameters (TR, TE, voxel size) or to other random sources of variability (image noise or magnetic field non-uniformities). Looking at the results presented in Subsections 3.3.2-3.3.5, it is possible

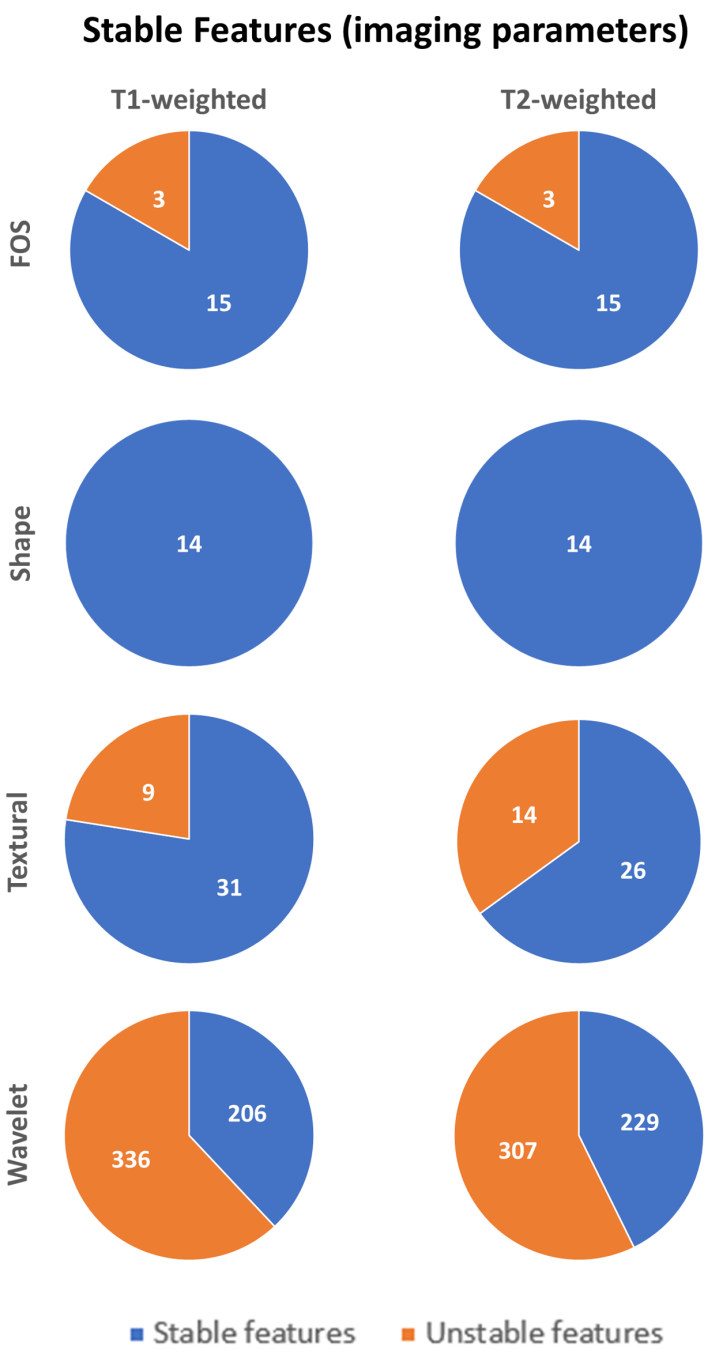

Figure 3.13: *Pie charts showing the proportion of stable radiomic features for T1-weighted and T2-weighted images, divided by class of features.*

to see how image preprocessing improved the stability of at least one category of radiomic features.

Changes in TR/TE were observed to produce differences in radiomic features especially in case of FOS features. It was found that intensity standardization improves the stability of FOS radiomic features to changes in signal intensity due to variations in TR/TE. In Subsection 3.3.1 the improvement was observed when using any of the 3 main intensity standardization algorithm (histogram matching, histogram stretching and Z-score). That result seems to suggest that that even studies involving different intensity standardization algorithm (such as [103–105]) can be comparable. However it is worth noting that histogram matching cause a low but significant reduction in the stability of GLCM-based features and should maybe be avoided in future studies.

Differences in voxel size may affect the stability of radiomic features, especially for textural features, as shown in Subsection 3.3.3. Our results confirmed what was previously found in literature on different phantom and non-phantom studies in CT and PET [16, 84, 85, 109, 110]. We found that resampling to a common resolution improves the stability to variations in voxel size. This effect was significant for shape and size features, but there an effect on textural features, although not always significant, was observed as well. The positive effect on stability provided by voxel size resampling that was found in this study is in agreement with other studies of literature [16, 84, 85].

From the results observed in Subsection 3.3.4, it seems that Gaussian noise has little effect on the stability of radiomic features. This observations were in apparent contrast with what is observed in other studies of literature [111]. This may depend on the fact that larger ROIs were used in this study compared to [111], which may contribute to reduce the effect of noise. However, the increase in stability of the features due to smoothing is in agreement with what has been found in other studies of literature [112].

To the knowledge of the authors, stability to INU has not been evaluated in literature. In this study it was found that INU may reduce radiomic features stability, especially for T1w MRI, and that bias field correction may increase the ICC to values that are close to 1. This is particularly important because INU are often present in MRI images and having a way to successfully deal with this type of artifacts is a necessary step prior to any quantitative image analysis.

Based on the results of Subsections 3.3.2-3.3.5 it is clear how the optimal preprocessing method should include all the four analyzed step. In Subsection 3.3.6, stability of radiomic features was evaluated for images

acquired with random acquisition conditions after they underwent optimal preprocessing, and the proportion of stable radiomic features was calculated. The knowledge of stable and unstable features is important because it is a first criteria that could be used to guide features selection.

Although some other studies investigated stability of radiomic features to changes in MRI acquisition parameters, to the knowledge of the authors, the collection of studies presented in this chapter was the most exhaustive and it was also the first one to quantitatively observe the effect of image preprocessing on radiomic features stability. Another advantage of the presented studies is that the experiments were performed using a virtual simulator (BrainWeb) that is freely available online. Therefore, the presented results are completely reproducible.

The presented studies are not exempt from limitations. One limitation is the fact that the custom MRI simulation was performed only on one phantom, in one district (the brain) and without considering pathological tissue. These limitations can be partially addressed. Since many different tissues were considered for the stability analyses that were performed, it is reasonable that the results on features stability can be translated also to other district of the body and to pathological tissues, even though this assumption has to be verified in future studies.

Another limitation is related to the fact that BrainWeb does not allow to study the behaviour of other type of MRI images, such as contrast enhanced or ADC images. However, we assume that image preprocessing, when applicable, will lead to the same positive results.

Another limitation is the fact that, by using BrainWeb, it was not possible to evaluate the effect of preprocessing on harmonizing the differences due to systematic source of variability, such as MRI scanner, which are known to have an effect on the radiomic features [84, 113, 114].

The presented method provides a first example of using stability to perform an preliminary features selection. However, the knowledge of how image acquisition parameters affect the radiomic features may be used alternatively. A recent work [115] suggested the possibility to use the information of condition-dependency of the features for data augmentation instead of feature selection, avoiding possible information loss due to the removal of some features. Side studies are used to estimate the parameters of the noise distributions that allow the generation of the augmented data and BrainWeb could be an optimal framework to perform such side studies, at least for as far as T1w and T2w MRI are concerned. The combination of BrainWeb-based studies and data augmentation could definitely be an interesting future development for studies of MRI radiomics.

In conclusion, with the experiments presented in this chapter it was possible to show that applying a correct preprocessing to MRI images helps increasing the radiomic features stability even when the images are acquired in different conditions. The preprocessing techniques analyzed in the chapter will therefore be used as a part of the workflow to develop the radiomic-based prognostic signature for survival analysis.

CHAPTER 4

Stability analyses for segmentation uncertainties

This chapter describes the experiments performed to evaluate the stability of radiomic features to variations in the ROI. The analysis was performed separately on both STS and HNC. In particular, stability to multiple segmentations and geometrical transformations of the ROI are evaluated as possible sources of variability. The possibility to infer the results of multiple segmentations by ROI geometrical manipulation was also evaluated. For each type of cancer, a set of features stable to ROI variability was identified. Last, the two sets of stable features were intersected with the one identified in Chapter 3 to define the final lists of features to be used for the future analyses.

4.1 Introduction

In the previous chapter the stability of radiomic features to variation in image acquisition conditions have been investigated through experiments on virtual MRIs. However, another source of variability may be encountered after the image acquisition and it is related to the variability in the ROI seg-

mentation that is used to define the mask to extract the features. As a matter of fact, tumor segmentation is a process that is typically performed manually or semi-automatically and this leads to uncertainty in the segmentation that eventually leads to uncertainties in the radiomic features [14].

The problem of ROI uncertainties is well known. Previous studies of literature investigated this issue by using ICC to evaluate the stability of radiomic features to multiple segmentations, performed by either one or multiple radiologist [5, 116, 117] also using the information to perform a preliminary feature selection. This is the most intuitive solution, but often unrealistic as the production of multiple segmentations by more than one radiologist is very time consuming procedure which subtracts time to the clinical routine and is typically not performed. Semi-automatic or automatic segmentation methods may reduce this variability [15], but such methods are designed for specific body district and they are not available for all the cases.

Another method to evaluate the stability of radiomic features to ROI uncertainties is to use geometrical perturbation of the ROIs [118–121], which is an attempt to mimic the effect of multiple segmentations without spending time to perform them. Such type of perturbation can also be used to mimic the misalignment of the ROI that may happen due to bad image registration in multi-modality imaging studies.

In this chapter, a stability analysis to both multiple segmentation and ROI geometrical transformation was performed. The purpose of the analysis is to identify a set of stable features to ROI uncertainties. Such set is then intersected with the set of stable features identified with the experiments of virtual phantom in Chapter 3 to define the final set of features to be used for the following analyses. Also, the possibility of using ROI transformation (which is fast and semi-automatic) as a surrogate of multiple segmentations (which is time-consuming) in stability analyses will be investigated

Also, a comparison of multiple segmentations and ROI perturbations, was performed to try to compare the two types of test, to understand if the latter method (that can be automatized and need only one set of ROI), can be used to infer the results of the former (which is more time consuming).

4.2 Materials and methods

4.2.1 Image dataset

Two different image datasets were used for this experiment, one for each of the types of cancer that were investigated for this thesis (HNC and STS).

The first dataset (called HNC dataset from now on) was made up of MRI images from 15 different patients affected by HNC, scanned at Istituto Nazionale dei Tumori in Milan, Italy (INT). MRI images include T1w and T2w images, as well as ADC maps (Figure 4.1). ADC images were obtained as described in Subsection 2.3.4 by fitting an exponential decay on DWI images acquired with different b-values in the range of 0-1000 s/mm^2. Details on the image acquisition parameters for the HNC dataset are reported in Table 4.1.

The second dataset (called STS dataset from now on) included MRI images from 15 different patients affected by STS, scanned at INT. The same MRI sequences used for the HNC dataset were considered (see examples in Figure 4.2). ADC images were obtained as described in Subsection 2.3.4 by fitting an exponential decay on DWI images acquired with different b-values in the range of 0-1000 s/mm^2. Details on the image acquisition parameters for the STS dataset are reported in Table 4.2.

4.2.2 Regions of interest

For the images of the HNC and STS datasets, the main tumor was manually segmented by two radiologist from INT, with at least 10 years of experience each. Each radiologist performed his/her own segmentation on the T2w images. T2w images were used because the tumor can be easily distinguished from the surrounding tissue. All the segmentations were performed using the open source software 3D slicer [97]. Figure 4.3A displays an example of double segmentation of the same tumor performed by the two radiologists.

Another set of ROIs was obtained by applying geometrical transformations to the segmentation performed by the most expert radiologist of the two. ROIs were translated positively and negatively in both x and y directions (in-plane directions) as done in previous experiments on ADC images from our research group [119, 120]. In particular, translations of 10 % of the length of the bounding box of the tumor were considered, which is a value similar or higher compared to the one obtainable with multiple segmentation (compare Figures 4.3A and 4.3B). Only translations were considered, because they were found out to be the transformations that cause larger changes in the radiomic features [67, 120]. For each patients, 5 different ROIs were available (the original segmentation and the 4 translated versions).

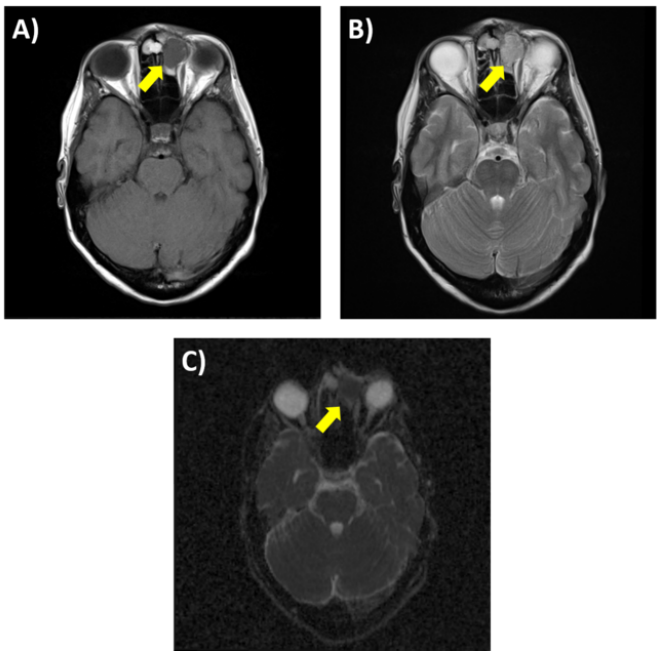

Figure 4.1: *Images from a sample patient of the head and neck cancer (HNC) dataset. In every image, the tumor mass is pointed by a yellow arrow. A) T1-weighted image. B) T2-weighted image. C) Apparent diffusion coefficient map.*

HNC DATASET ACQUISITION PARAMETERS			
Image sequence	**T1w**	**T2w**	**ADC**
Scanner	Siemens Avanto	Siemens Avanto	Siemens Avanto
Number of images	15	15	15
Pulse sequence	Spin-echo	Spin-echo	Echo-planar
Magnetic field	1.5 T	1.5 T	1.5 T
Time of repetition	359-650 ms	2950-7400 ms	3271-10127 ms
Time of echo	9-15 ms	75-124 ms	63-93 ms
Slice thickness	3-5 mm	3-5 mm	3-5 mm
Slice spacing	3.3-6.0 mm	3.3-6.0 mm	3.3-6.0 mm
Pixel spacing	0.31-0.90 mm	0.26-0.82 mm	0.94-2.18 mm

Table 4.1: *Synthetic description of the head and neck cancer (HNC) dataset. Parameters are shown by image sequence: T1-weighted (T1w), T2-weighted (T2w) and apparent diffusion coefficient maps (ADC).*

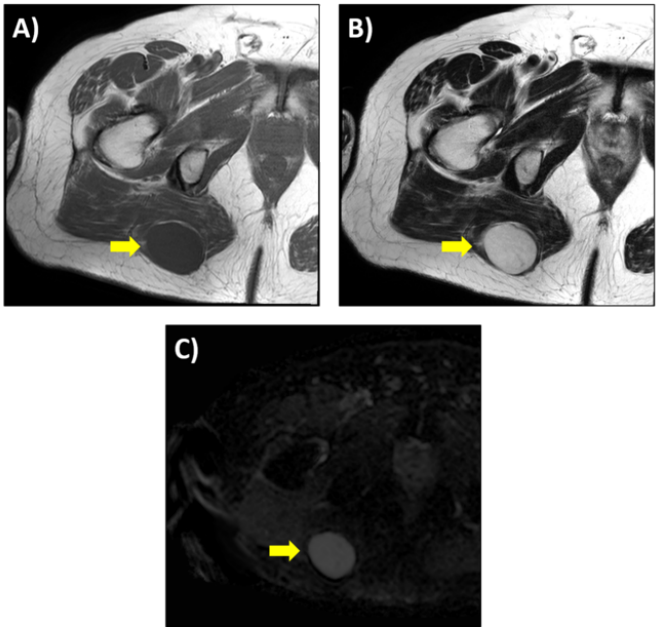

Figure 4.2: *Images from a sample patient of the soft tissue sarcoma (STS) dataset. In every image, the tumor mass is pointed by a yellow arrow. A) T1-weighted image. B) T2-weighted image. C) Apparent diffusion coefficient map.*

STS DATASET ACQUISITION PARAMETERS			
Image sequence	**T1w**	**T2w**	**ADC**
Scanner	Philips Achieva	Philips Achieva	Philips Achieva
Number of images	15	15	15
Pulse sequence	Spin-echo	Spin-echo	Echo-planar
Magnetic field	1.5 T	1.5 T	1.5 T
Time of repetition	497-746 ms	3000-5065 ms	5400-8011 ms
Time of echo	7-10 ms	80-132 ms	64-85 ms
Slice thickness	4-5 mm	4-5 mm	4-5 mm
Slice spacing	4.4-6.0 mm	4.4-6.0 mm	4.4-6.0 mm
Pixel spacing	0.38-1.02 mm	0.35-1.22 mm	1.34-2.08 mm

Table 4.2: *Synthetic description of the soft tissue sarcoma (STS) dataset. Parameters are shown by image sequence: T1-weighted (T1w), T2-weighted (T2w) and apparent diffusion coefficient maps (ADC).*

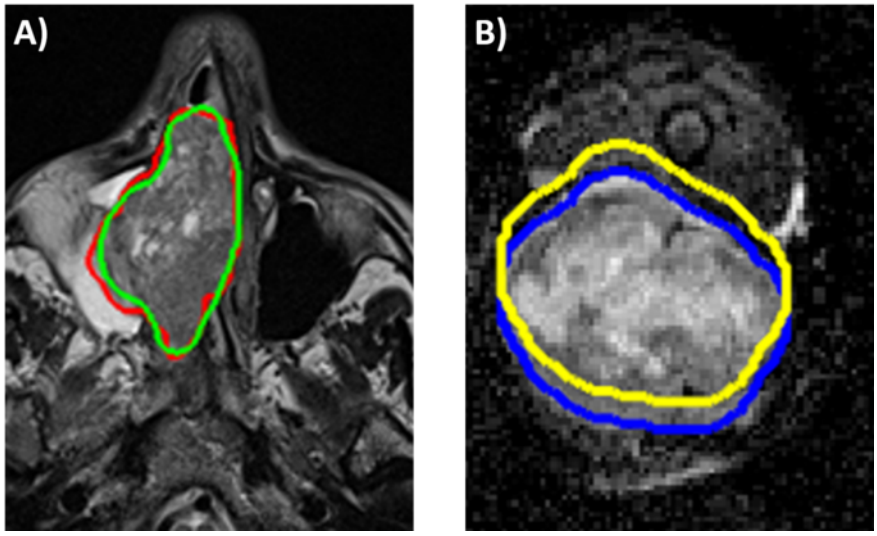

Figure 4.3: *Examples of modified regions of interests used for the radiomic stability analysis. A) T1-weighted image in which the tumor has been segmented by two radiologist. B) Apparent diffusion coefficient map in which the segmentation has been translated 10% of the bounding box.*

4.2.3 Image preprocessing

For T1w and T2w images, the optimized preprocessing defined in Chapter 3 was applied. First, a 3D Gaussian filter with a 3x3x3 voxel kernel and $\sigma = 0.5$ was used to denoise the images. Then, the N4ITK algorithm [106] was used for the correction of intensity-non uniformity. After that, intensity standardization (Z-score) and voxel size resampling (with B-spline interpolation) were performed directly from Pyradiomics. The voxel size was resampled to isotropic resolution of 2 mm. The value of 2 mm was used instead of 1 mm (as in Chapter 3) for reasons of computational complexity during the calculation of radiomic features. In fact, some images of the STS dataset are very large in terms of field of view and the use of 1 mm resolution led to large 3D matrices and excessive memory requirements. The resolution of 2 mm was used in other studies of literature with good results [122], so resolution was assumed to be sufficient to avoid the loss of important information. This value of resolution was also used for all the analyses of the next chapters.

ADC had a different preprocessing compared to T1w and T2w images. Intensity standardization and inhomogeneity correction were not performed on ADC maps, because the intensities in ADC maps represent a physical

quantity and have been shown to be consistent among acquisitions, provided some conditions, like long enough TR, same range of b-values and same magnetic field strength [62, 63, 123]. Prior to features extraction, intensity values were windowed between 0 and 4000*10^{-6} mm^2/s, in order to remove non-physiological values due to image noise in the DWI used to fit the ADC maps. Inhomogeneity correction was not performed because, even if inhomogeneities may be present in the original DWIs, ADC images depend on the difference in the DWI signal intensities and systematic differences do not affect them.

4.2.4 Radiomic features extraction

A set of 536 features were extracted for each image type (T1w, T2w and ADC) for a total of 1608 features. Detail of features categories and features extraction parameters were described in Subsections 3.2.3 and 3.2.10.

4.2.5 Stability analysis for ROI uncertainties

The general workflow of the stability analysis for uncertainties in the ROI is presented in Figure 4.4. Two separate tests were performed in parallel: multiple segmentations and ROI geometrical transformation.

In both tests, radiomic features were extracted from each tumor, using all the ROIs available (2 in the multiple segmentation test and 5 in the ROI translation test). At the end of each test, a n-by-m matrix was available for each feature, where n is the number of patients and m is the number of ROI used for the features extraction of each patient. From such matrices it was possible to compute two arrays of ICCs (one for each test) and identify the stable features (ICC>0.75). Features that passed both tests were considered stable to uncertainties of the ROI.

Two separate stability analyses like the one previously described were performed for the images of HNC and STS dataset, because the stability of radiomic features to ROI uncertainties may be district dependent [120,124].

4.2.6 Comparison of multiple segmentation and ROI transformations

The stable features sets obtained by the two tests (multiple segmentations and ROI geometrical transformations) were compared. Pearson correlation coefficients and scatter plots were used to evaluate correlation between ICCs of corresponding features in the two tests. Also, a confusion matrix was computed for both the STS and HNC datasets, in order to understand if one test was more restrictive than the other.

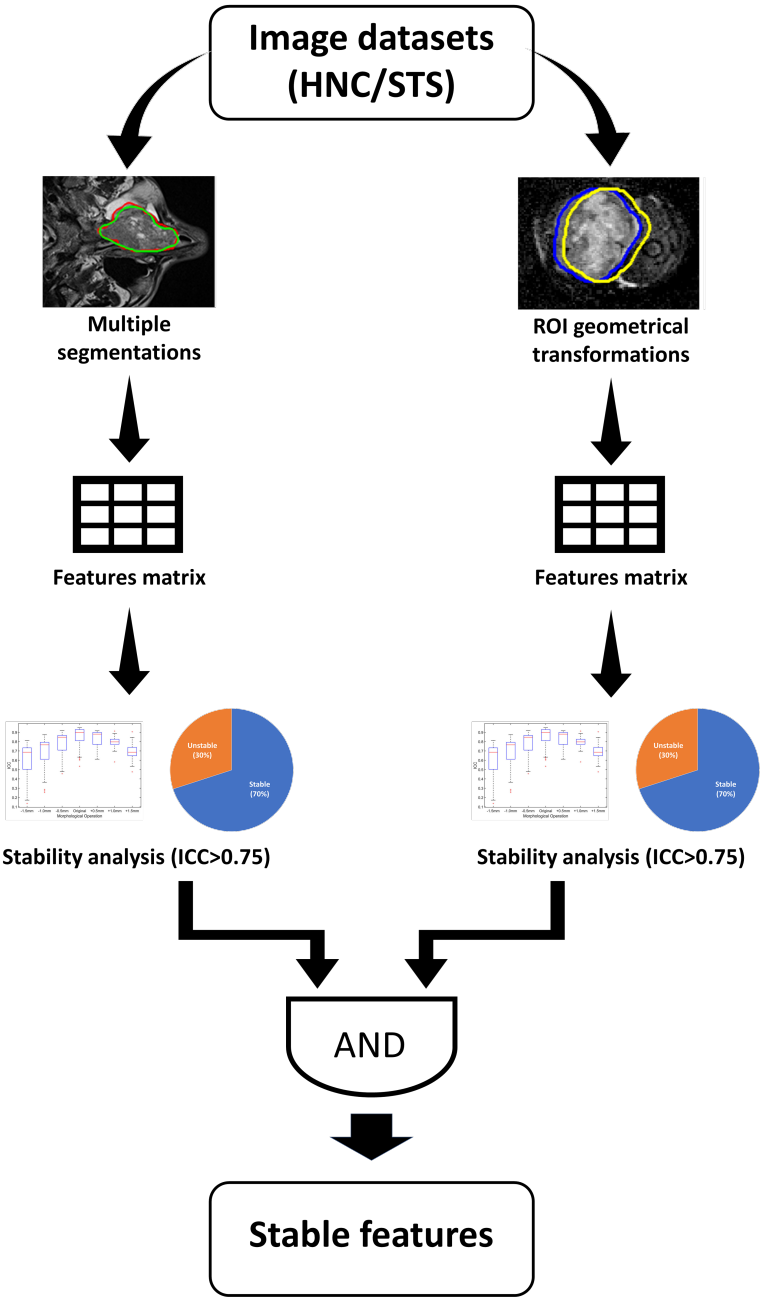

Figure 4.4: *Workflow of the analysis of stability of radiomic features to segmentation uncertainties. Features with Intra-class Correlation Coefficient (ICC) higher than 0.75 in both tests were considered stable. The analysis was performed separately for Head and Neck Cancer (HNC) and Soft Tissue Sarcomas (STS) dataset.*

4.2.7 Definition of the final stable features set

The final set of radiomic features to be used for an optimal radiomic analyses should be stable to both uncertainties in the ROI and variations in the image acquisition conditions. Therefore, for each image type and body district, the intersection with the stable features set obtained from Chapter 3 was performed.

BrainWeb does not allow to simulated DWI and ADC images and so those type of images could not be used for the stability analyses in Chapter 3. In general ADC images should be more stable compared to T1w and T2w images, because they are less effected by changes in TR/TE and inhomogeneities in the magnetic fields [62,63,123], so it is reasonable to expect a higher number of stable features. However, to go towards safety, only the features that were stable for both T1w and T2w images (n=229) were used as a surrogate of a stable features set for ADC.

4.3 Results

4.3.1 Stability analysis for ROI uncertainties

Figures 4.5-4.6 display the proportion of radiomic features that was found stable to variations in the ROI due to both multiple segmentation and geometrical transformations, for the HNC and STS dataset respectively. Figures 4.7-4.8 show analogous results divided by class of features.

The pie charts in Figures 4.5-4.6 show that the T1w-based features tend to be less stable to variation in the ROI compared to T2w and ADC images and this was true for both the STS and HNC dataset, even if the difference is less evident in STS. It can be also observed that the number of stable features was higher for the STS dataset compared to the HNC dataset, independently on the type of MRI considered.

By looking at the different classes of features (Figures 4.7-4.8) it is possible to see that in general, SS features are the most stable features, while for the other features classes there is no clear ranking based on stability.

4.3.2 Comparison of multiple segmentation and ROI transformations

Figure 4.9 shows the scatter plots of the ICC of the two tests for both HNC and STS datasets. It is possible to see that there is a significant ($p<10^{-263}$) correlation between the two metrics (Pearson correlation coefficient 0.73 and 0.81 for the HNC and STS respectively). The concordance between the two types of tests can also be appreciated by looking at the confusion matrices in Figure 4.10.

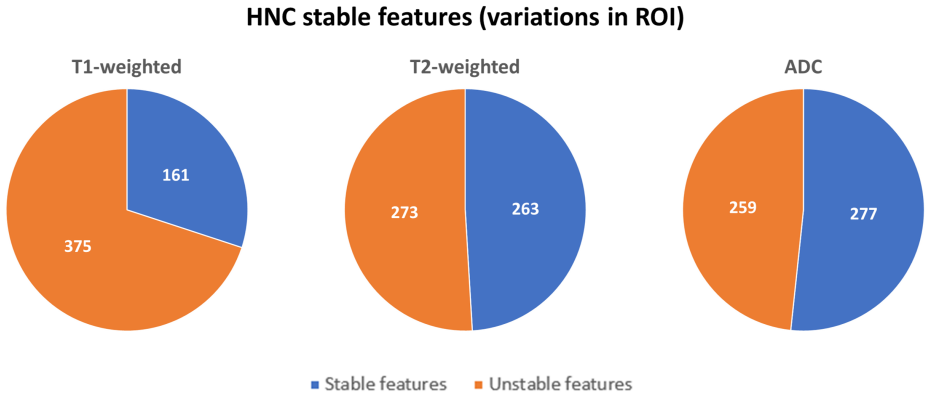

Figure 4.5: *Pie charts displaying the proportions of features stable to uncertainties in the Region Of Interest (ROI) in the Head and Neck Cancer (HNC) dataset. Results are grouped by type of MRI images: T1-weighted, T2-weighted and Apparent Diffusion coefficient (ADC) maps.*

4.3.3 Definition of the final stable features set

The total number of features stable to image acquisition parameters was 779 (266 T1w, 284 T2w and 229 surrogates features for ADC). When intersecting these 779 features with the features that were considered stable for variations in the ROIs (701 and 1057 for HNC and STS datasets respectively), the final features sets were obtained. The numbers of features that were stable to both imaging-related variability and ROI-related variability were 410 and 617 for the HNC and STS datasets respectively. This is depicted in Figures 4.11-4.12, which also shows the distributions by image type. The full list of features can be found in appendix A.

4.4 Discussion

The experiment presented in this chapter gave some insight on the stability of radiomic features to different types of uncertainties in the ROI. In particular, multiple segmentations and ROI translations were considered.

By looking at the results of Subsection 4.3.1 it may be inferred that stability is district dependent, as it was found in previous studies of literature [120, 124]. In the context of multiple segmentations and ROI manipulation, the observed difference in the values of ICC between tumor sites may be explained in two ways: HNC are more difficult to see and segment compared to STS and therefore higher inter-reader variability is expected; STS are larger than HNC and so when similar perturbation are applied to

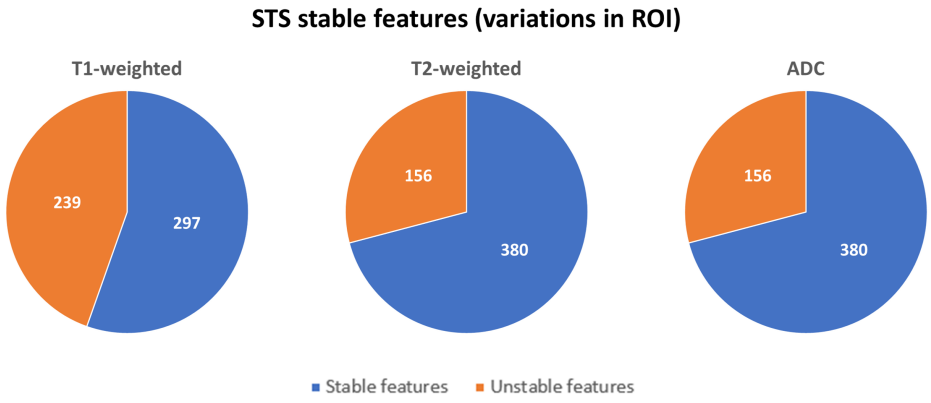

Figure 4.6: *Pie charts displaying the proportions of features stable to uncertainties in the Region Of Interest (ROI) in the Soft Tissue Sarcoma (STS) dataset. Results are grouped by type of MRI images: T1-weighted, T2-weighted and Apparent Diffusion coefficient (ADC) maps.*

the ROI, the percentage of the ROI that is affected by the variation is lower.

Another finding of the experiment was the dependence of stability to the particular imaging sequence considered. In particular, T1w features presented a lower stability to ROI uncertainties compared to T2w and ADC images. Such difference in stability among the different imaging sequences was also reported in other studies of literature [125]. The phenomenon may be explained by reasoning on the ICC metric. The ICC involves the ratio between a a patient-related variability (MS_R in Equation 3.1) in and a condition-related variability (MS_C in Equation 3.1). In T2w and ADC images the heterogeneity of the tissues can be better appreciated compared to T1w images. Therefore, in T2w and ADC the patient-related variability is higher and consequently the values of ICC are higher when the same entity of ROI-related variability is applied.

By looking at the scatter plots in Figure 4.9 it is possible to see how the ICC for the two tests were correlated. This is reasonable since both tests deal with small modifications of the ROIs. The correlation coefficients were 0.73 and 0.81 which means that one test explains around 50-65 % of the variability observed in the other test. Also, by looking at Figure 4.10 it can be seen that when it comes to select the stable features, only a minority of features (around 4-5 %) that resulted stable to the ROI transformation test was unstable to the multiple segmentation test. Therefore, this could be the proof that, when it comes to select a set of stable features, ROI geometrical transformation may be used as a good surrogate of multiple segmentation,

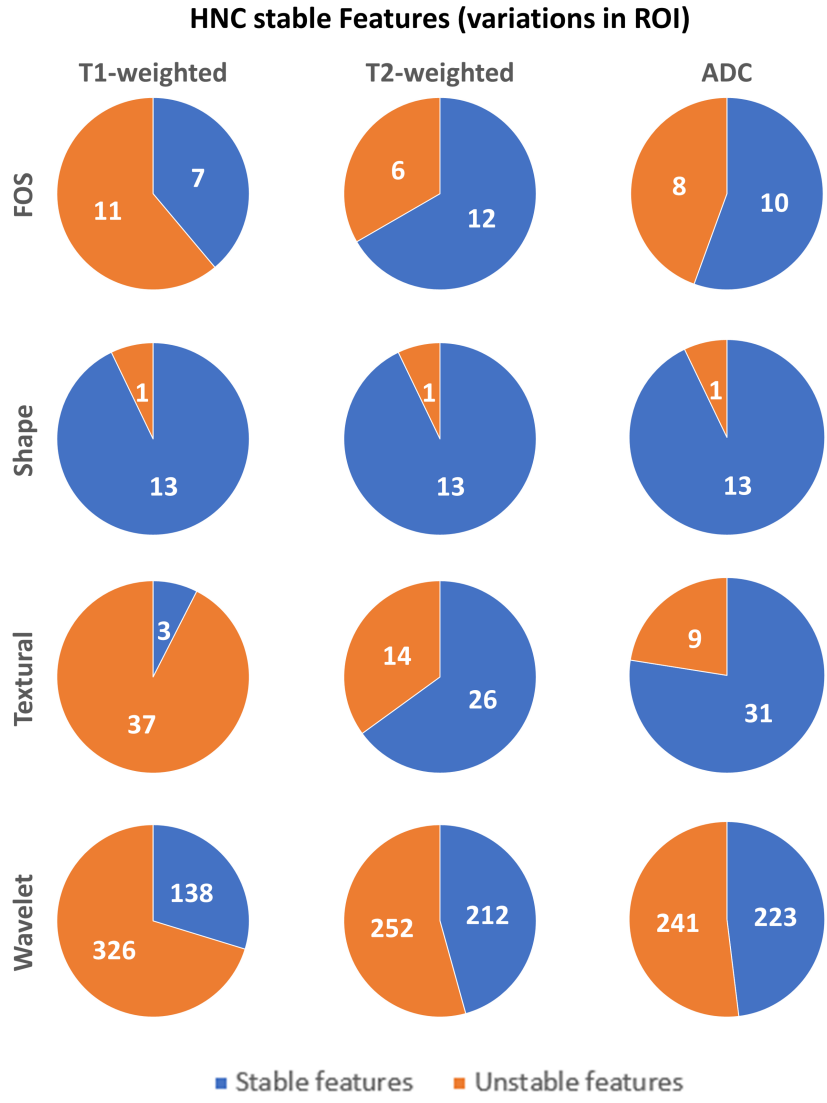

Figure 4.7: *Pie charts displaying the proportions of features stable to uncertainties in the Region Of Interest (ROI) in the Head and Neck Cancer (HNC) dataset, grouped by feature class and type of MRI image.*

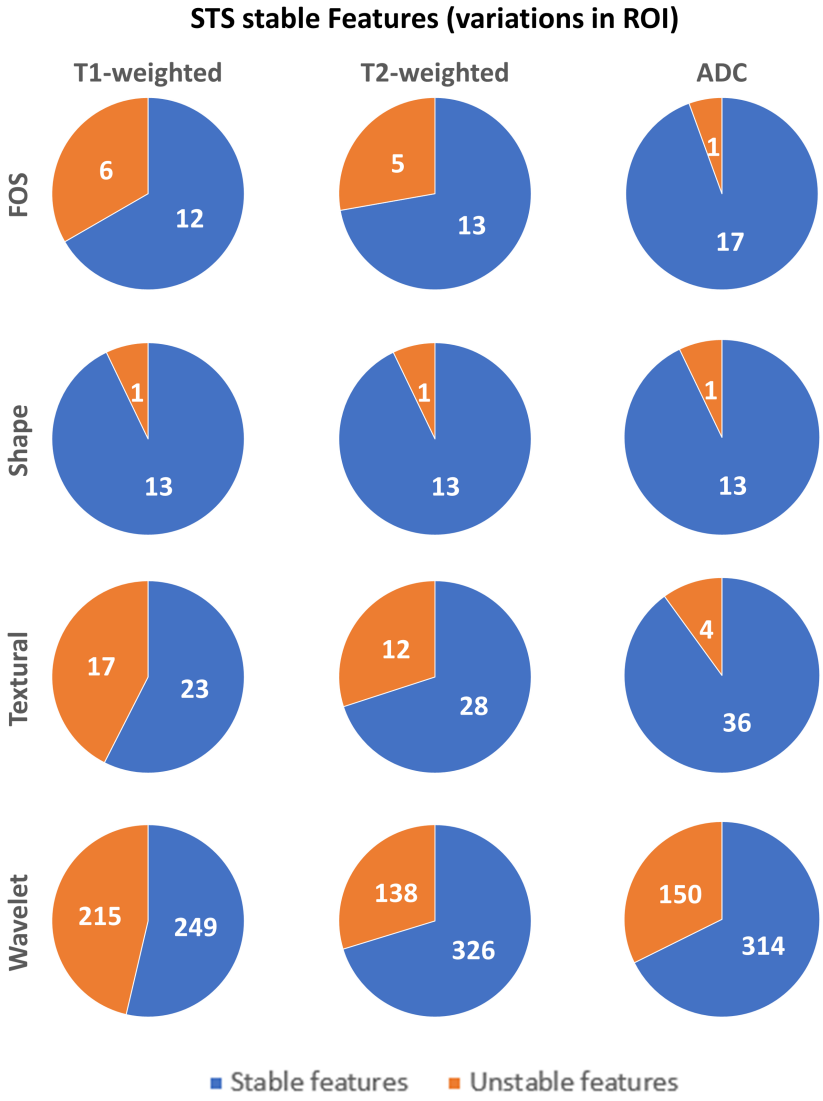

Figure 4.8: *Pie charts displaying the proportions of features stable to uncertainties in the Region Of Interest (ROI) in the Head and Neck Cancer (HNC) dataset, grouped by feature class and type of MRI image.*

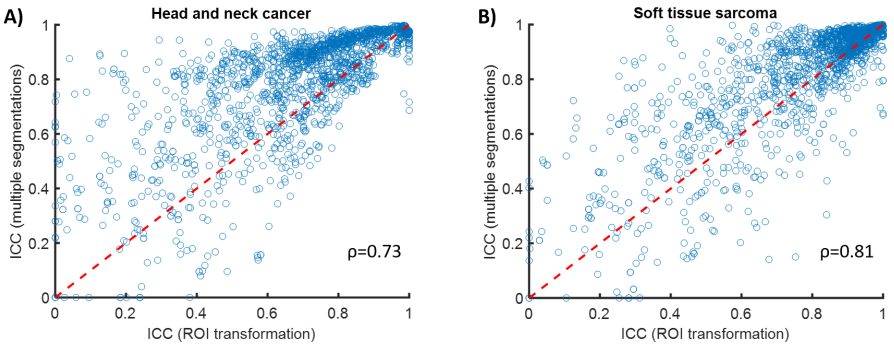

Figure 4.9: *Scatter plot representing the correlation between the Intra-class Correlation Coefficients (ICC) for the multiple segmentations and the ROI geometrical transformation tests. A) Head and neck cancers dataset. B) Soft tissue sarcomas dataset. Red dashed line is the bisector of the first and third quadrant.*

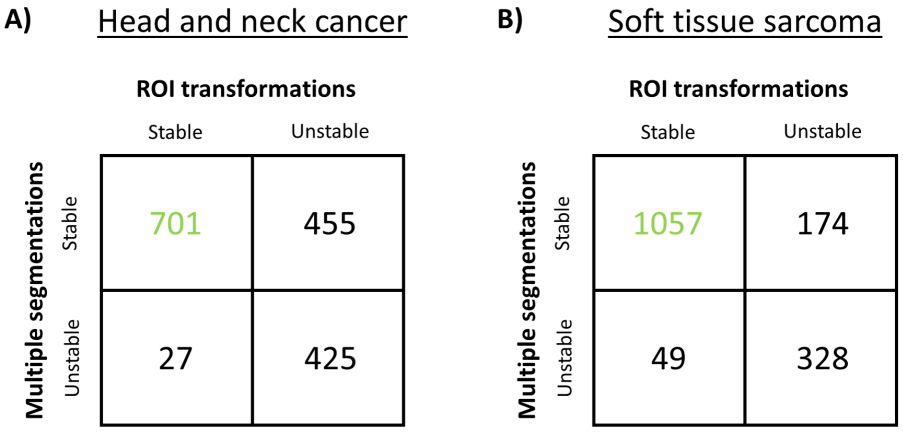

Figure 4.10: *Confusion matrices with the four different combinations of features stability for the two tests (multiple segmentations and ROI geometrical transformations). The number in greens represents the features stable to variations of the ROI. A) Head and neck cancer dataset. B) Soft tissue sarcoma dataset.*

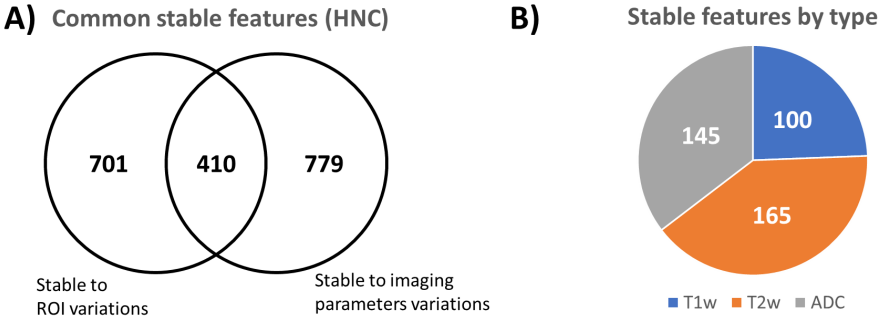

Figure 4.11: *A) Intersection of the imaging-stable and ROI-stable features sets in the Head and Neck Cancer datasets (HNC). B) Distribution of the common stable features as a function of the image type.*

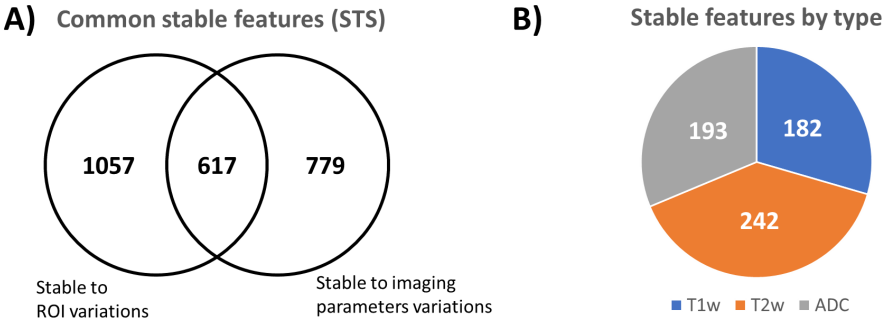

Figure 4.12: *A) Intersection of the imaging-stable and ROI-stable features sets in the Soft and Tissue Sarcoma datasets (STS). B) Distribution of the common stable features as a function of the image type.*

with an overall increase in the efficiency of the radiomic analysis (i.e. less time spent for segmentation).

The two stability analyses presented in Chapter 3-4 were used to perform a first preliminary features selection, like previously done in literature [5,117]. When making an intersection between imaging-stable features and ROI-stable features, 410 and 617 were found stable for HNC and STS respectively. These features sets were used for the analyses in the following chapters.

The stability based feature selection applied in the Chapters 3-4 are not exempt from limitations. The first limitation is that no test-retest was considered, so there is no information available about the repeatability of radiomic features. This was due to the fact that test-retest can be evaluated only using a real phantom, which was not available. However, there seems to be a correlation between the ICC for test-retest and for multiple segmentations [116], and the most stable features for multiple segmentations are also the ones with the highest test-retest repeatability. Also, test-retest variability affects has less impact on radiomic features compared to changes in the image acquisition parameters [113]. Therefore, it is safe to assume that the selected features set, which is stable to imaging-related variability, will also be stable to test-retest. Another limitation is that bias in the radiomic features due to systematic sources of variability, like in case of acquisitions with different scanners, were not considered, even if they are known to have a strong negative effect on stability [86, 113]. This is a limitation due to the fact that batch-related variability (e.g. scanner related variability) may be evaluated only with one or more real phantoms, which were not available. One last limitation is that the threshold for ICC used to define the stable features set is somehow arbitrary. When defining the threshold of 0.75, [102] was used as a reference, but other values were used in literature, either higher or lower [118, 121]. A low threshold removes a lower number of features but also leads to results that are less reproducible. A higher threshold increases the reproducibility of the results but may lead to a larger loss of the total information contained in the original dataset and to worse results for the prognostic models. Future studies may be required to better understand the optimal trade-off between stability of the results and performance of the final model.

In conclusion, the result of this chapter provided a a list of features stable to ROI uncertainties was found, and demonstrated that most of this list can be obtained just by using geometrical transformation of the ROI. The results of this and Chapter 3 may help removing unstable features that will negatively affect any radiomics-based model. The methodology is highly

recommended for all those situations, like multicentric studies or collection of retrospective patients, in which the analyzed sources of variability are present.

CHAPTER 5

Postprocessing optimization for radiomic analysis

This chapter describes the analysis performed to optimize the postprocessing for the radiomic features. The experiments were performed on a multicentric dataset of patients affected by HNC The combination of features normalization algorithm (4 methods) and feature selection (2 pipelines) was set in order to maximize the prognostic performance of a Cox proportional hazard regression model for OS.

5.1 Introduction

In the context of radiomic analysis, features postprocessing refers to all the steps that are performed on the extracted features before their used to build the predictive or prognostic model. The focus of the experiment described in this chapter was the optimization of a postprocessing pipeline, in particular for features normalization and feature selection/dimensionality reduction.

In radiomics, different approaches for features selection and dimensionality reduction are used, which may be divided in supervised or purely

unsupervised. Purely unsupervised algorithms reduce the number of features without using any information about the endpoint that has to be predicted [126]. Those may include a-priori information based on features stability (see Chapter 4 or [127]), reduction based on PCA (see Subsection 2.5.6 or [126]), features pairwise-correlation [122] or features clustering [128]. Supervised feature selection methods use the information of the outcome to evaluate the predictive or prognostic performance of different combinations of features and select the subset that maximize such performance. Examples of supervised feature selection algorithms applied to radiomics include the Least Absolute Shrinkage and Selection Operator (LASSO) [11], significance evaluation in univariate [5] or multivariate Cox analysis , minimum redundancy maximum relevance [77], supervised-PCA [129], and others (see [77,127]). Supervised feature selection methods tend to perform better, but unsupervised method have the advantage of not needing any label and to be less prone to overfitting when the number of samples is low [126].

Features normalization is the operation that ensures that all the features have the same (or similar) range of values [68]. This operation is typically performed because having features with similar ranges is a requirements of some non-scale invariant methodologies as k-nearest neighbors classifier and PCA [68]. When those kind of models/operations are not used, features normalization may be avoided and as a matter of fact, there are studies in which it was not performed [5, 11]. However, even when not strictly required, features normalization may be advised because it helps the convergence of the optimization algorithms used in model fitting [74]. No study investigating the advantages or disadvantage of normalization for radiomics has been proposed so far. Among the studies on radiomics that performed features normalization, Z-score normalization [77, 103] and its non-parametric equivalent [117] were the most used , but other methods exist [75]. A comparison of different normalization algorithms for radiomic analysis has not been performed yet. Moreover, feature normalization may affect the performance of the feature selection algorithms, and an analysis trying to find the best combination of normalization and feature selection has not been performed.

In the study presented in this chapter, the above mentioned steps of features postprocessing were investigated. The impact of different combinations of features normalization methods and feature selection/dimensionality reduction algorithms on the performance of a Cox model prognostic for OS were evaluated. The best combination that resulted from this analysis was chosen as part of the designed pipeline for radiomic analysis.

5.2 Materials and methods

5.2.1 Image dataset

The dataset used for this study (from now on called BD dataset) was part of a larger database of patients with HNC collected for the European project *BD2Decide: Big Data and models for personalized Head and Neck Cancer decision support* funded through the H2020 research program [130], containing clinical, genomic and imaging data of 1541 patients with advanced HNC (stage III-IV accordin to TNM VII) from 4 different sub-sites (hypopharynx, larynx, oral cavity and oropharynx). Patients were acquired from 7 different clinical centers across 3 countries (Germany, Italy and the Netherlands): the Azienda Ospedaliero-universitaria di Parma (AOP), in Parma, Italy; the Istituto Nazionale dei Tumori (INT), in Milan, Italy; the Spedali Civili di Brescia (SCB), in Brescia, Italy; the Heinrich-Heine-Universität Düsseldorf (UDUS), in Düsseldorf, Germany; the university hospital of Ulm (ULM), in Ulm, Germany; the Maastricht Radiation Oncology clinic (MAASTRO), in Maastricht, the Netherlands; the Vrije Universiteit Medical Center (VUMC), in Amsterdam, the Netherlands. Follow-up data (including death and cancer recurrences) were also acquired for each patient of the dataset.

The dataset actually comprised two main subsets: one that was retrospectively collected (called BD-Retro from now on), containing 1086 patients, and one that was prospectively collected for the BD2Decide project (called BD-Prosp from now on), containing 455 patients.

The dataset used in this study (called BD1 for short) included the group of retrospective patients for the BD2Decide project, whose images satisfied the following inclusion criteria: availability of MRI baseline examination and availability of both T1w and T2w MRI acquired with SE or TSE pulse sequences. In total, 262 patients from 4 different clinical centers (AOP, INT, SCB and VUMC) were chosen. Clinical information about the patients of interest are reported in Table 5.1. Information about the image acquired from each center are reported in Tables 5.2-5.5, where it is possible to see that most of the image acquisition parameters are in the range defined during the experiments with Brainweb (compare with Table 3.5).

CLINICAL DATA (BD1 DATASET)	
Number of patients	262
Age (median and IQR)	60 years [54-67]
Sex	Female: 78 (30%) Male: 184 (70%)
Stage TNM VII	Stage III: 54 (21%) Stage IV: 212 (79%)
Stage TNM VIII	Stage I: 26 (10%) Stage II: 18 (7%) Stage III: 79 (30%) Stage IV: 139 (53%)
Subgroups	Hypopharynx: 14 (5%) Larynx: 20 (8%) Oral cavity: 126 (48%) Oropharynx (HPV+): 71 (27%) Oropharynx (HPV-): 31 (12%)
Treatment	Surge: 29 (11%) Surge+Rad: 56 (21%) Surge+Rad+Chem: 55 (21%) Rad+Chem: 114 (44%) Other: 7 (3%) Unknown: 1 (<1%)
Follow-up time (median and IQR)	65 months [47-75]
Number of deaths	105 (40%)
Number of recurrences	155 (59%)

Table 5.1: *Clinical and demographic characteristics of the 262 patients of the BD1 dataset. Age and follow-up time are displayed as median and inter-quartile range (IQR).*

5.2.2 Image segmentation

For each patient, the main tumor was manually segmented. Each clinical center had his own radiologist performing the segmentation. The segmentation was performed using the T2w MRI as the reference and the same ROI was used also for T1w images (Figure 5.1). As a matter of fact, the reference system was the same for both T1w and T2w and only small misalignment between T1w and T2w are present, compatible with the ones analyzed in Chapter 4 (compare Figure 4.3 and Figure 5.1).

BD1 IMAGING DETAILS (AOP)		
Image sequence	**T1w**	**T2w**
Number of images	54	54
Scanner	Philips Achieva	Philips Achieva
Magnetic field	1.5 T	1.5 T
Pulse sequence	Spin-echo	Spin-echo
Time of repetition	450-984 ms	2220-9273 ms
Time of echo	7-12 ms	82-110 ms
Slice thickness	3-5 mm	3-5 mm
Slice spacing	3.4-7 mm	3.4-7 mm
Pixel spacing	0.43-0.65 mm	0.39-0.59 mm

Table 5.2: *Synthetic description of the imaging acquisition parameters for patients of BD1 dataset acquired at the Azienda Ospedaliero-universitaria di Parma (AOP). Parameters are shown by image sequence: T1-weighted (T1w) and T2-weighted (T2w).*

BD1 IMAGING DETAILS (INT)		
Image sequence	**T1w**	**T2w**
Number of images	182	182
Scanner	Siemens Avanto	Siemens Avanto
Magnetic field	1.5 T	1.5 T
Pulse sequence	Spin-echo	Spin-echo
Time of repetition	324-961 ms	2110-9141 ms
Time of echo	7-26 ms	80-134 ms
Slice thickness	2.7-7 mm	2.7-7 mm
Slice spacing	3.15-7.8 mm	3.3-8.05 mm
Pixel spacing	0.36-0.9 mm	0.29-0.98 mm

Table 5.3: *Synthetic description of the imaging acquisition parameters for patients of BD1 dataset acquired at the Istituto Nazionale dei Tumori (INT). Parameters are shown by image sequence: T1-weighted (T1w) and T2-weighted (T2w).*

BD1 IMAGING DETAILS (UDUS)		
Image sequence	**T1w**	**T2w**
Number of images	2	2
Scanner	Siemens Avanto	Siemens Avanto
Magnetic field	1.5 T	1.5 T
Pulse sequence	Spin-echo	Spin-echo
Time of repetition	573-645 ms	3400-3423 ms
Time of echo	11 ms	82 ms
Slice thickness	4 mm	4 mm
Slice spacing	4.4-4.8 mm	4.4-4.8 mm
Pixel spacing	0.38-0.63 mm	0.38-0.63 mm

Table 5.4: *Synthetic description of the imaging acquisition parameters for patients of BD1 dataset acquired at the Heinrich-Heine-Universität Düsseldorf (UDUS). Parameters are shown by image sequence: T1-weighted (T1w) and T2-weighted (T2w).*

BD1 IMAGING DETAILS (VUMC)		
Image sequence	**T1w**	**T2w**
Number of images	24	24
Scanner	GE Signa HDxt	GE Signa HDxt
Magnetic field	1.5 T	1.5 T
Pulse sequence	Spin-echo	Spin-echo
Time of repetition	440-780 ms	3440-8560 ms
Time of echo	9-17 ms	111-118 ms
Slice thickness	3-4 mm	4 mm
Slice spacing	3.3-4.4 mm	4.4-4.8 mm
Pixel spacing	0.45-0.59 mm	0.38-0.63 mm

Table 5.5: *Synthetic description of the imaging acquisition parameters for patients of BD1 dataset acquired at the Vrije Universiteit Medical Center (VUMC). Parameters are shown by image sequence: T1-weighted (T1w) and T2-weighted (T2w).*

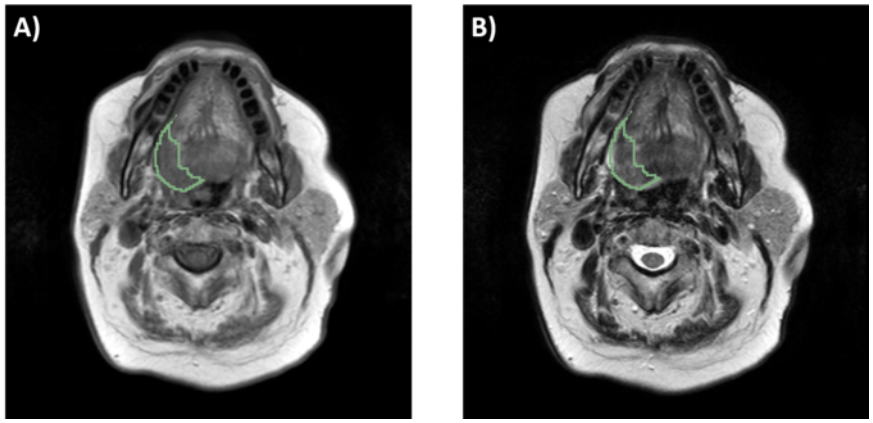

Figure 5.1: *Example of T1-weighted image (A) and T2-weighted image (B) from a patient of BD1 dataset. Segmentation of the main tumor is performed on T2-weighted image and used for both image types.*

5.2.3 Image preprocessing

All the preprocessing techniques described in Chapters 3-4 for T1w and T2w images were applied to the T1w and T2w MRI prior to the radiomic features extraction. First, a 3D Gaussian filter with a 3x3x3 voxel kernel and $\sigma = 0.5$ was used to denoise the images. Then, the N4ITK algorithm [106] was used for the correction of intensity-non uniformities. Intensity standardization was performed using Z-score. Voxel size resampling to an isotropic resolution of 2 mm was performed using B-spline interpolation.

5.2.4 Radiomic features extraction

The set of 265 stable features for T1w and T2w (as described in Chapter 4 or Appendix A) was used. The extracted features are grouped as follows: 100 features for T1w MRI (13 shape, 7 FOS, 1 GLCM, 2 GLRLM and 77 wavelet); 165 features for T2w MRI (13 shape, 12 FOS, 9 GLCM, 12 GLRLM and 119 wavelet). A fixed bin number intensity discretization (32 bins) was used prior to the features extraction.

5.2.5 Methods for features normalization

Four different normalization algorithms were used to normalize the ranges of the features, each with its own pros and cons [75]: Z-score normalization, median-mad normalization, min-max normalization and hyperbolic tangent normalization. For a better descriptions of the methods, refer to Subsection 2.5.5.

5.2.6 Features selection pipelines

Two different pipelines for features selection were considered, comprising both supervised and unsupervised selection (Figure 5.2).

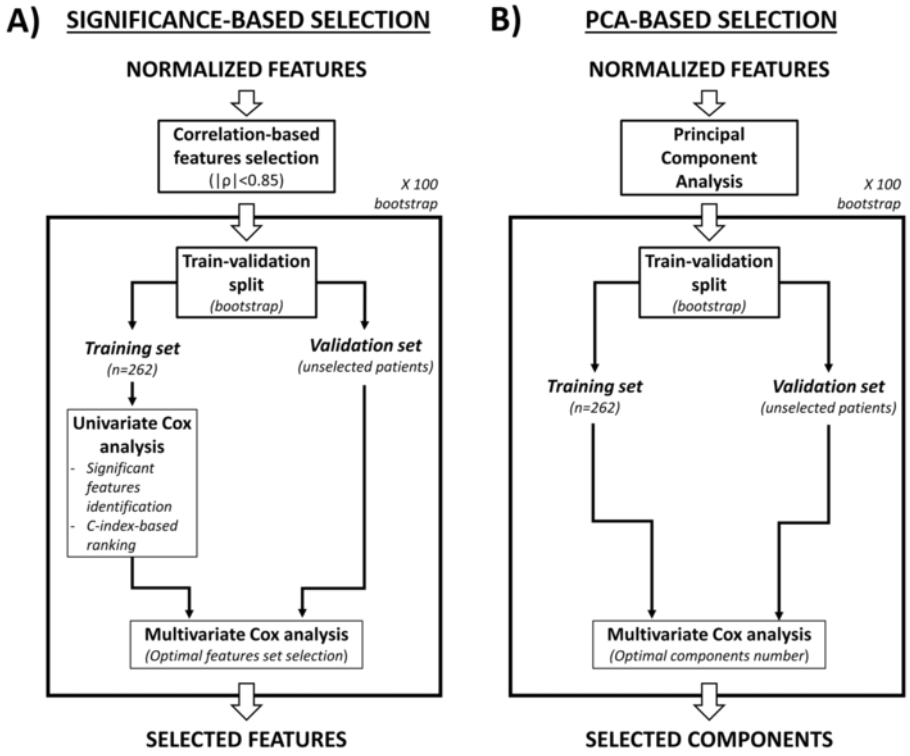

Figure 5.2: *Schematic description of the two feature selection pipelines used. A) Significance-based selection. B) PCA-based selection.*

Significance-based selection for simplicity, involved the use of the original radiomic features (Figure 5.2A) and the algorithm works as follows:

1. Spearman correlation coefficients are calculated for each pair of features. Whenever the absolute value of Spearman correlation coefficient of a is above 0.85, one of the two features is removed. In particular, the feature with the lowest mean correlation coefficient with all the other n features (264 in this case) is kept while the other is removed.
2. A training and a validation set are generated. The training set included 262 patients generated using bootstrap. The unselected patients from the original dataset constitute the internal validation set.

3. A univariate Cox proportional hazard regression model (see Subsection 2.5.4 for further details) for prediction of OS is fitted on the training set using each of the features selected at step 1. Features that are not significantly associated with OS (p>0.05 for log-rank test) are excluded. Correction for false discovery rate is performed using FDR correction as described in [78] (see also Subsection 2.5.6). In case all features presented a p>0.05, only one feature (the one with the lowest p-value) is selected.

4. The features selected from step 3 are sorted by their Harrell C-index (see Subsection 2.5.7) on the training set and they are progressively added to a multivariate Cox proportional hazard regression model fitted on the training data. The model is evaluated on the internal validation set and the validation C-index is computed. The combination of features that maximizes the validation C-index of features is selected.

5. Steps 2-4 are repeated 100 times and for each iteration the optimal radiomic feature set is stored. In total 100 different features sets are available.

6. The N features that are selected more often throughout the 100 sets are picked, N being the rounded average length of the 100 optimal features sets.

PCA-based selection (Figure 5.2B), involves the use of PCA and therefore the resulting features are linear combinations of the original features. The pipeline works as follows:

1. PCA is applied to the features.

2. A training and a validation set are generated, as explained for the significance-based pipeline.

3. The components are ranked by explained variance and progressively added to a Cox multivariate regression model.The model is evaluated on the internal validation set and the validation C-index is computed. The number of components that maximizes the validation C-index is selected.

4. Steps 2-3 are repeated 100 times and for each iteration the best radiomic features set is stored The first N components are picked for the final model, N being the average number of components selected throughout the 100 iterations.

5.2.7 Comparison of features processing pipelines

In total, 8 different postprocessing pipeline, given by the combination of the 2 features selection pipelines (Subsection 5.2.6) and the 4 features normalization methods (Subsection 5.2.5), were evaluated. The features/components obtained with the 8 pipelines were used to fit Cox proportional hazard regression models for OS [73]. To evaluate the best features processing pipeline, 10-fold cross-validation was used (as described in Subsection 2.5.8 and illustrated in figure 5.3). In each iteration, the parameters of the features processing pipeline (e.g μ and σ for Z-score normalization, the list of the selected features, etc...) and the coefficients of the Cox model were learned from the training set and then used to predict the signature (i.e. the linear combination of the selected features as defined by the Cox model) in the patients of the validation set. In this way, a unique unbiased estimate of the radiomic signature could be computed for each patient. The metric used to evaluate the quality of the pipeline was the Harrell's C-index [80] between the OS and the cross-validated signature. Confidence intervals for the C-indexes of the models were obtained through bootstrap. Statistical comparison was performed using 2-way ANOVA for repeated measures, with the factors being the normalization algorithm and the features selection pipeline. Repeated measure 2-way ANOVA also allowed to understand if there is a significant interaction effects between the normalization algorithm and the features selection pipeline. Two-sided unpaired t-tests with post-hoc comparisons were used to identify significantly different pairs. Tukey-Kramer method was used to correct for multiple hypothesis testing.

5.3 Results

Figure 5.4 shows the distributions of C-indexes obtained for each of the 8 different combinations of features normalization and feature selection pipeline. By looking at the figure, it seems that an interaction effect between the two factors is present. As a matter of fact, the choice of the best features selection pipeline depends on the normalization algorithm used. When Z-score normalization was used, the significance based selection is the one that gives the best performance, while the opposite happens if the hyperbolic tangent normalization is used. The 2-way ANOVA for repeated measures identifies an interaction-term that is significantly different from 0 and confirms what has already been observed qualitatively ($p=8*10^{-14}$). Therefore, each combination of normalization and selection has to be treated independently.

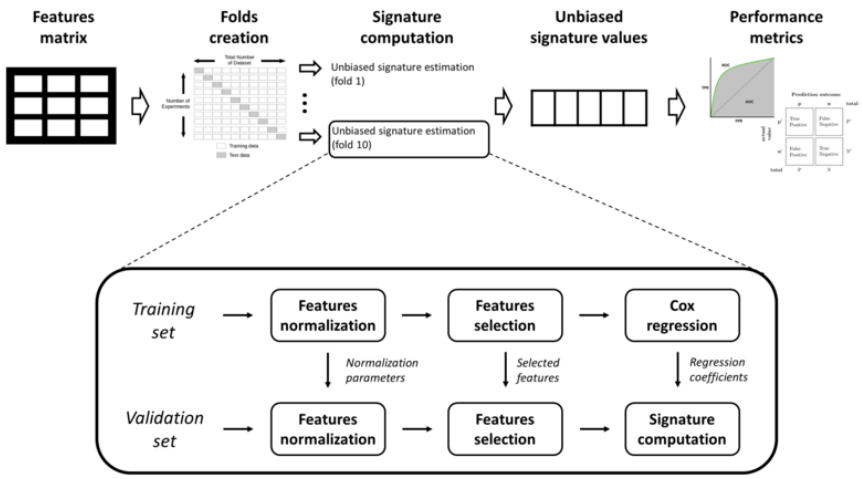

Figure 5.3: *Schematics of the 10-fold cross-validation used to validate the performance of a prognostic model for overall survival.*

Among the 8 different combinations, the one that led to the best performance was the combination of Z-score normalization and the significant based pipeline (mean C-index: 0.67, 95% CI: [0.61-0.73]).

5.4 Discussion

In this chapter, an investigation on features normalization and features selection was performed, in order to optimize the postprocessing for the radiomic analysis.

One of the findings of this experiment was that there is an interaction effect between the features normalization algorithms and the feature selection algorithm, that has been statistically proven by a 2-way ANOVA for repeated measures. This means that, whenever dealing with a radiomic-based survival analysis, features selection and features normalization methods cannot be independently optimized, but all the possible combinations must be tested and the best one must be chosen. A similar approach was used in [77] to optimize the combination of feature selection and classification algorithm.

By looking at Figure 5.4 it can be seen that significance-based feature selection works better than PCA-based feature selection for the majority of the normalization algorithms. This result may indicate that the use of the original features may be preferable to the remapped features obtained

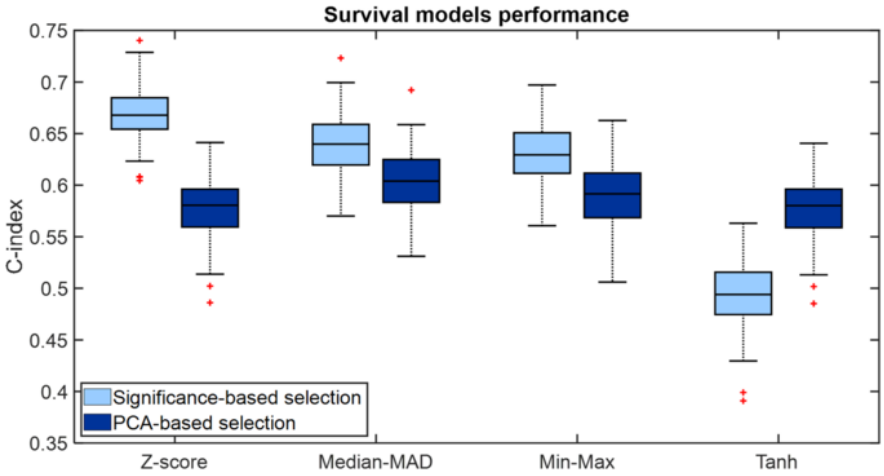

Figure 5.4: *Distributions of C-indexes for the different combinations of features normalization and feature selection.*

after PCA. This may be due to the fact that a lot of the features are just noise and are non informative, thus reducing the performance of the single components of the PCA. The result was in line with what was reported in literature for classification problems [126]. The best feature processing pipeline consisted in the combination of Z-score normalization and the significance-based feature selection (mean C-index: 0.67, 95% CI: [0.61-0.73]).

Looking at Figure 5.4, it is possible to see that the significance-based features selection is much more sensitive on the features normalization method compared to the PCA-based feature selection (median C-index ranges 0.49-0.67 vs 0.58-0.60). The results may be explained by the fact that, even though the original features are strongly affected by the normalization methodology, the components obtained after PCA are more stable.

Based on the results of the experiment reported in this chapter and the ones shown in Chapters 3-4, the final workflow for radiomic-based survival analysis in multicentric studies can be completely defined. Such workflow is schematically illustrated in Figure 5.5.

Although a first optimization of the postprocessing pipeline was performed, the experiment is not exempt from limitations. One limitation is that the analysis is not exhaustive since many more combinations of normalization algorithms and feature selection pipeline could have been tested. However, the analysis performed in this chapter provides some important

information (i.e. interaction between normalization and selection, best performance using the original features). Future analysis may try to integrate this information using new normalization/selection methods.

In this analysis (an in the final workflow) only Cox regression was used to create the signatures, but some other survival models exist. The application of just one survival model could be considered another limitation of the study. However, almost all the studies of radiomics for survival analysis used the Cox model and therefore the use of Cox models was a choice to make the comparison with previous studies of literature easier.

One last limitation of the workflow described in Figure 5.5 is that some postprocessing steps, like batch effect correction or missing data imputation, that may improve the performance of the final model, were not included. Batch effect correction refers to the statistical correction of systematic differences that are observed in features coming from different batches (e.g. different instrumentation, different hospitals, etc...). Among the batch effect correction methods, ComBat is the most used for radiomic analysis [131–133], with positive results. However, ComBat correction requires to specify some clinical covariates to work properly and does not guarantee optimal results when such covariates are confounded with the batches, as could happen with the datasets presented in this thesis [134], an issue that to date has not been resolved yet. Therefore, it was decided not to use ComBat within the design of the workflow. Similarly, missing data imputation was not included in the radiomic workflow, since missing data were not an issue for the datasets analysed in the thesis, which was composed of images sequences (T1w and T2w MRI) that, being part of the clinical routine, were performed for all the patients with MRI imaging available.

In conclusion, in this chapter the optimization of postprocessing of radiomic features (including features normalization and selection) was performed. Although the results are non-exhaustive, they were useful to guide the creation of the workflow for the development of radiomic-based prognostic models for survival for HNC and STS, that are described more in details in Chapters 6 and 7.

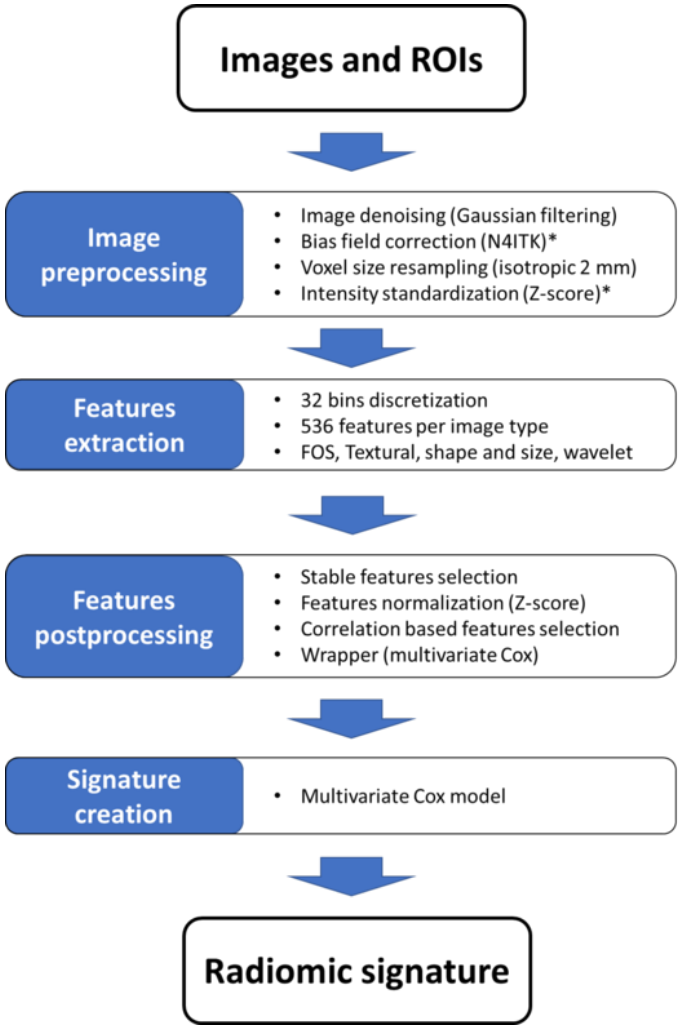

Figure 5.5: *Block diagram describing the steps of the workflow for the creation of the radiomic signature for survival analyses. Steps marked with asterisks are performed only for T1-weighted and T2-weighted images.*

CHAPTER 6

Radiomics-based survival models for head and neck cancer

This chapter describes the application of an optimized radiomic workflow for the development of an MRI-based prognostic signature for OS in patients affected by HNC. The signature was trained its prognostic power was evaluated using both cross-validation and external validation. Evaluation of the added prognostic power of radiomics was also performed.

6.1 Introduction

HNC is considered a rare pathology, as it accounts for only 3% of the total cancers worldwide [135]. Despite this fact HNC is one of the most studied types of cancer in terms of radiomics [9, 136]. Among the possible applications of radiomics to HNC, the development of prognostic models (for OS, DFS, etc...) is the one that has been explored the most.

Most of the studies related to the development of radiomic-based survival models for head and neck cancer was based on either CT or PET [5, 10, 137–139]. Among those studies, the most important is the one described in [5], where 4-features prognostic signature for OS trained on CT

images was proven to be prognostic in two different HNC datasets with 136 and 95 patients respectively (Harrel C-index 0.69 in both). Further studies confirmed the prognostic value of the same signature for CT on three additional datasets of oropharyngeal squamous cells carcinomas (OPSCC) with more than 200 patients each [137], obtaining value of C-index of 0.63-0.65.

When it comes to HNC, studies on MRI-radiomics are less common than studies on CT-radiomics, because of issue related to MRI signal, that make it more difficult to use it to perform a quantitative analysis [9, 136]. However, MRI is more versatile than CT or PET and has excellent soft tissue contrast, so it still holds a lot of potential for tumor characterization. Therefore, more recent studies tried to create prognostic model for HNC by MRI-radiomics. In [140] the value of ADC images for prognosis of DFS was evaluated on a dataset of 175 patients and it was shown that ADC_{high} (ADC computed using images with b-values>500 s/mm^2) was an independent prognostic factor and led to a C-index of 0.62. However, DWI and ADC images are not part of the clinical routine for HNC in the majority of the hospitals and therefore it would be difficult to use such model in the clinical practice. MRI-based signatures based on contrast-enhanced T1w images (CE-T1w) and T2w images (with or without fat suppression) were developed for survival prognosis in nasopharyngeal carcinoma (NPC) from large datasets (>100 patients each) obtained from hospitals of China, where NPC is endemic [11, 141–143]. Limitations of the aforementioned studies are the fact that they all used images acquired with the same protocols and the same good performance is not guaranteed when heterogeneity in image acquisition parameters is present. Moreover, signatures that perform well for NPC do not necessarily perform well for HNC in general.

The purpose of the study presented in this chapter was to use the radiomic workflow presented in Figure 5.5 to train a center- and parameter-robust prognostic signature for OS in HNC. External validation of the signature was also performed and the addictive value of radiomics to the features used in the clinical practice was evaluated.

6.2 Materials and methods

6.2.1 Image datasets

Two different datasets of HNC patients were used for this study, both coming from the BD dataset introduced in Subsection 5.2.

The first dataset was the BD1 dataset introduced in Subsection 5.2, containing 262 patients with advanced HNC from different sub-sites coming

from 4 different clinical centers, for which both T1w and T2w MRI were available. This dataset was used to train the radiomic signature and to perform a preliminary cross-validation. For further details on the dataset refer to Section 5.2, in particular to Table 5.1 for clinical data and Tables 5.2-5.5.

The second dataset (called BD2 dataset from now on) was a subset of the BD-Prosp dataset (see Section 5.2) which consisted of 232 patients with HNC from various sub-sites with baseline T1w and T2w MRI available. BD2 dataset included patients coming from 4 different clinical centers (AOP, INT, SCB and ULM).

Clinical information about the patients of BD2 are reported in Table 6.1, where a comparison with the data of BD1 was also performed. χ^2 tests and Mann-Whitney test were used for perform statistical comparisons among the clinical variables in the two sets, for the categorical and clinical variables respectively. Patients of the BD2 dataset had significantly shorter follow-up and reduced number of events. Also, the percentages of the different types of treatments was not the same in the two sets.

Most of the image acquisition parameters used to acquire the patients of the two datasets were in the range of values used to acquired the virtual MRI used for the stability analyses of Chapter 3 (compare Tables 5.2-5.5 and Tables 6.2-6.5 with Table 3.5).

6.2.2 Image segmentation

For each patient, the main tumor was manually segmented. Each clinical center had his own radiologist performing the segmentation. The segmentation was performed using the T2w MRI as the reference and the same ROI was used also for T1w images, since only small misalignment between T1w and T2w are present, to which the majority of radiomic features is stable (as verified in Chapter 4).

6.2.3 Image preprocessing

The optimal preprocessing pipeline defined in Chapters 3-4 and listed in Figure 5.5 was applied to the T1w and T2w MRI prior to the radiomic features extraction. First, a 3D Gaussian filter with a 3x3x3 voxel kernel and $\sigma = 0.5$ was used to denoise the images. Then, the N4ITK algorithm [106] was used for the correction of intensity non-uniformities. Intensity standardization was performed using Z-score. Voxel size resampling to an isotropic resolution of 2 mm was performed with B-spline interpolation.

CLINICAL DATA BD2DECIDE			
Feature	**BD1**	**BD2**	**p-value**
Number of patients	262	232	-
Age (median and IQR)	60 years [54-67]	61 years [54-69]	0.54
Sex	Female: 78 (30%) Male: 184 (70%)	Female: 71 (31%) Male: 161 (69%)	0.84
Stage TNM VII	Stage III: 54 (21%) Stage IV: 212 (79%)	Stage III: 35 (15%) Stage IV: 197 (85%)	0.11
Stage TNM VIII	Stage I: 26 (10%) Stage II: 18 (7%) Stage III: 79 (30%) Stage IV: 139 (53%)	Stage I: 31 (13%) Stage II: 15 (7%) Stage III: 54 (23%) Stage IV: 132 (57%)	0.29
Subgroups	Hypopharynx: 7 (4%) Larynx: 24 (10%) Oral cavity: 109 (47%) Oropharynx (HPV+): 67 (29%) Oropharynx (HPV-): 24 (10%)	Hypopharynx: 7 (4%) Larynx: 24 (10%) Oral cavity: 109 (47%) Oropharynx (HPV+): 67 (29%) Oropharynx (HPV-): 24 (10%)	0.45
Treatment	Surge: 29 (11%) Surge+Rad: 56 (21%) Surge+Rad+Chem: 55 (21%) Rad+Chem: 114 (44%) Other: 7 (3%) Unknown: 1 (<1%)	Surge: 37 (16%) Surge+Rad: 39 (17%) Surge+Rad+Chem: 45 (19%) Rad+Chem: 75 (32%) Other: 14 (6%) Unknown: 22 (10%)	0.04
Follow-up time (median and IQR)	65 months [47-75]	28 months [24-33]	$7.33*10^{-17}$
Number of deaths	105 (40%)	45 (19%)	$6.08*10^{-7}$
Number of recurrences	107 (41%)	71 (31%)	0.02

Table 6.1: *Clinical and demographic characteristics of the Patients of both BD1 and BD2 dataset. Age and follow-up time are displayed as median and inter-quartile range (IQR). p-values are obtained via χ^2 tests (for categorical variable) and Mann-Whitney tests (for continuous variables). Variables with different distributions between BD1 and BD2 are highlighted in red*

BD2 IMAGING DETAILS (AOP)		
Image sequence	**T1w**	**T2w**
Number of images	45	45
Scanner	Philipps Achieva	Philipps Achieva
Magnetic field	1.5 T	1.5 T
Pulse sequence	Spin-echo	Spin-echo
Time of repetition	367-960 ms	1500-7299 ms
Time of echo	8-20 ms	60-120 ms
Slice thickness	3-4 mm	2-4 mm
Slice spacing	3.3-4.8 mm	3.3-6.4 mm
Pixel spacing	0.35-0.90 mm	0.27-0.78 mm

Table 6.2: *Synthetic description of the imaging acquisition parameters for patients of BD2 dataset acquired at the Azienda Ospedaliero-universitaria di Parma (AOP). Parameters are shown by image sequence: T1-weighted (T1w) and T2-weighted (T2w).*

BD2 IMAGING DETAILS (INT)		
Image sequence	**T1w**	**T2w**
Number of images	140	140
Scanner	Siemens Avanto	Siemens Avanto
Magnetic field	1.5 T	1.5 T
Pulse sequence	Spin-echo	Spin-echo
Time of repetition	371-896 ms	1400-13518 ms
Time of echo	8-26 ms	80-134 ms
Slice thickness	3-6 mm	2.5-6 mm
Slice spacing	3.3-6.6 mm	3.3-6.6 mm
Pixel spacing	0.34-1.00 mm	0.29-1.00 mm

Table 6.3: *Synthetic description of the imaging acquisition parameters for patients of BD2 dataset acquired at the Istituto Nazionale dei Tumori (INT). Parameters are shown by image sequence: T1-weighted (T1w) and T2-weighted (T2w).*

BD2 IMAGING DETAILS (SCB)		
Image sequence	**T1w**	**T2w**
Number of images	43	43
Scanner	Siemens Aera	Siemens Aera
Magnetic field	1.5 T	1.5 T
Pulse sequence	Spin-echo	Spin-echo
Time of repetition	380-720 ms	2550-7320 ms
Time of echo	9-30 ms	99-128 ms
Slice thickness	2-3 mm	3-4 mm
Slice spacing	2.4-5.4 mm	3.3-6 mm
Pixel spacing	0.40-0.78 mm	0.35-0.78 mm

Table 6.4: *Synthetic description of the imaging acquisition parameters for patients of BD2 dataset acquired at the Spedali Civili di Brescia (SCB). Parameters are shown by image sequence: T1-weighted (T1w) and T2-weighted (T2w).*

BD2 IMAGING DETAILS (ULM)		
Image sequence	**T1w**	**T2w**
Number of images	4	4
Scanner	Siemens Skyra	Siemens Skyra
Magnetic field	3 T	3 T
Pulse sequence	Spin-echo	Spin-echo
Time of repetition	592-821 ms	4800-6720 ms
Time of echo	9-20 ms	104 ms
Slice thickness	5 mm	5 mm
Slice spacing	5.5 mm	5.5 mm
Pixel spacing	0.60-0.65 mm	0.49 mm

Table 6.5: *Synthetic description of the imaging acquisition parameters for patients of BD2 dataset acquired at the university hospital of Ulm (ULM). Parameters are shown by image sequence: T1-weighted (T1w) and T2-weighted (T2w).*

6.2.4 Radiomic features extraction

A total of 265 stable features for T1w and T2w (Appendix A) was used. The extracted features are grouped as follows: 100 features for T1w MRI (13 shape, 7 FOS, 1 GLCM, 2 GLRLM and 77 wavelet); 165 features for T2w MRI (13 shape, 12 FOS, 9 GLCM, 12 GLRLM and 119 wavelet). A fixed bin number intensity discretization (32 bins) was used prior to the features extraction.

6.2.5 Prognostic models training

The data of BD1 dataset were used to train 3 different prognostic models for OS. An illustration of the training process for the 3 models is illustrated in Figure 6.1. Each source of data (radiomics or clinics) underwent its own postprocessing pipeline, and at the end of the pipeline an optimal set of features was selected. The features sets were used to train a radiomic and clinical signature using multivariate Cox proportional hazard regression [73]. Last, a combined signature was obtained by training a Cox regression model on the combination of the radiomic and clinical features sets.

For radiomic features, the postprocessing pipeline is the one described in Figure 5.5. Z-score normalization for the standardization of the ranges of features. Then, the significance-based selection pipeline (described in Subsection 5.2 and Figure 5.2A) was used to choose the optimal features set.

The clinical variables of interest were the following (see also Tables 5.1 and 6.1): age at diagnosis, sex, stage TNM (version VIII), HPV status, tumor sub-site. Categorical variables (such as tumor sub-site) were represented as dummy variables [68]. The features selection was performed by selecting only the features that were significantly associated with OS in univariate Cox regression.

6.2.6 Validation of the radiomic signature

To evaluate the prognostic performance of the radiomic signature both internal cross-validation on the BD1 dataset and external validation on the BD2 datasets were performed. Internal cross-validation was performed using 10-fold cross validation, in order to obtain an unbiased estimate of each signature for each patient (as previously illustrated in Figure 5.3). Such estimates were used to compute the Harrell's C-index for each model and confidence intervals were obtained using bootstrap (100 iterations). Moreover, in both BD1 and BD2 datasets patients were split in high and low risk

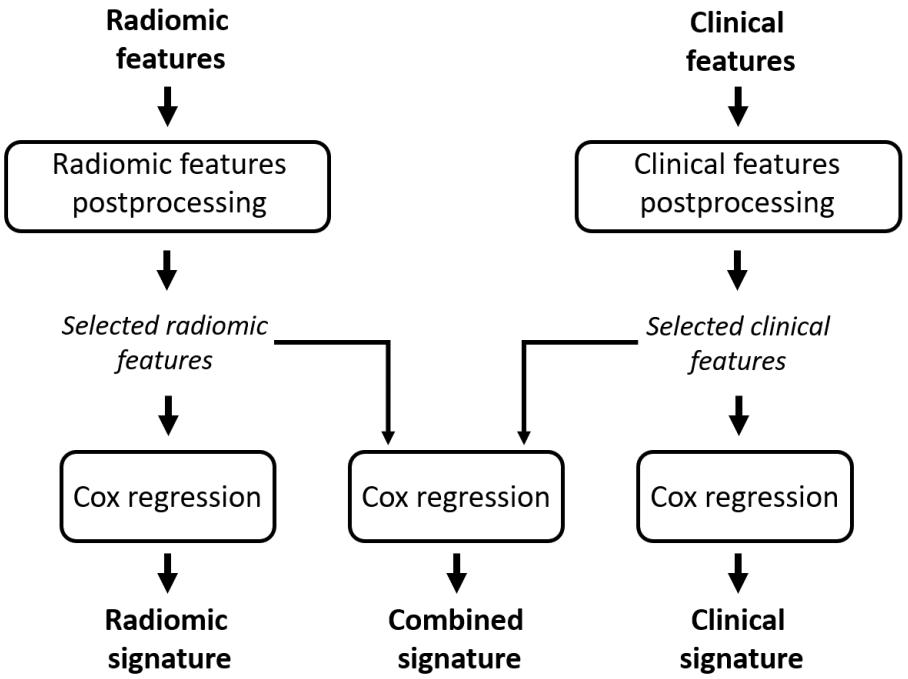

Figure 6.1: *Schematic description of the process of the creation of the prognostic signatures for overall survival from radiomic and clinical data.*

groups and the Kaplan-Meier curves were drawn for each group. Log-rank tests were used to compare the Kaplan-Meier curves for high and low risk patients. The median value of the signature on the training dataset (BD1) was used as a threshold to split the high and low risk groups.

6.2.7 Correlation between radiomic signature and clinical variables

The radiomic signature could potentially be highly correlated with other clinical variables. To ensure this was not the case, the correlation of the radiomic signature with the selected clinical features was evaluated. Statistically significant associations were identified using Spearman correlation coefficient and/or Kruskal-Wallis test. The analysis was performed on the merged BD1 and BD2 dataset to maximize the sample size, and since it is reasonable to think that the correlation with clinical variables is not dataset-specific.

6.2.8 Radiomic signature dependency on vendor and center

The radiomic signature could also potentially be affected by factors non related to tumor biology, like the MRI-scanner used or the center where the acquisition have been performed. To ensure the radiomic signature was not dependent from these factors, a 2-way ANOVA was performed. The analysis was performed separately on dead and alive patients at the end of the follow-up because it is reasonable to think that patients with the worst outcome will have a significantly higher signature. Moreover, for this analysis, the data of BD1 and BD2 datasets were merged, in order to maximize the number of samples in each subgroup, and since it is reasonable to think that the batch effect due to MRI-scanner and center is not dataset-specific.

6.2.9 Evaluation of added prognostic value of radiomics

To evaluate the added prognostic value of the radiomic, three different analyses were performed.

In the first analysis the prognostic power of the radiomic signature was evaluated by looking at its HR and p-value in a multivariate Cox regression model with the other selected clinical variables.

For the second analyses, the different subgroups defined by stage, HPV and tumor sub-site were considered. The Kaplan-Meier analysis described in Subsection 6.2.6 was repeated for each subgroup. This was done to assess whether the discriminatory power of the radiomic classification was high in each of the groups or was subgroup dependent.

For the third analyses, the C-indexes of the clinical, radiomic and combined signature (trained as described in 6.2.6) was computed to assess whether the prognostic performance of the radiomic or combined signature was better than the one of the clinical signature alone.

6.3 Results

6.3.1 Prognostic models training

After the signature training pipeline, 5 and 4 features were selected and used for the radiomic and clinical model respectively. Among the radiomic features, one of them was tumor volume and the others were wavelet features (1 related to texture and 3 related to FOS features). Table 6.6 lists the mean and standard deviation of the radiomic features, which were used for the Z-score normalization. Among the clinical features, one was TNM, one was HPV and the others referred to tumor sub-site. This results is in line with the fact that HPV, stage TNM and tumor location are prognostic

factors for survival. Details of the signatures used for the 3 Cox regression models (clinical, radiomics and combined) are reported in Table 6.7.

SELECTED RADIOMIC FEATURES DETAILS		
Features names	**Mean**	**Standard deviation**
T1w-waveletLHL-firstorder-90Percentile	0.45	0.28
T2w-original shape-VoxelVolume	16.58 cm^3	17.54 cm^3
T2w-waveletHHL-glrlm-GreyLevelNonUniformityNormalized	0.07	0.02
T2w-vaweletLLL-firstorder-InterquartileRange	1.33	0.56
T2w-vaweletLLL-firstorder- Range	6.43	3.37

Table 6.6: *Mean and standard deviation of the selected radiomic features. This values were used to compute the Z-scored versions of the features. Numeric values are reported up to the second decimal digit.*

SELECTED FEATURES COEFFICIENTS			
Features names	**Coefficient (clinical)**	**Coefficient (radiomic)**	**Coefficient (combined)**
T1w-waveletLHL-firstorder-90Percentile	-	-0.27	-0.15
T2w-original shape-VoxelVolume	-	0.13	0.03
T2w-waveletHHL-glrlm-GreyLevelNonUniformityNormalized	-	0.08	0.11
T2w-vaweletLLL-firstorder-InterquartileRange	-	0.13	0.09
T2w-vaweletLLL-firstorder- Range	-	0.20	0.15
TNM VIII	-0.80	-	0.74
HPV status	0.09	-	0.08
Oropharynx	-0.10	-	-0.18
Oral cavity	0.38	-	0.05

Table 6.7: *Coefficients of radiomic and clinical features for the three models tested: clinical, radiomic and combined. Coefficients values are displayed up to the second decimal digit.*

6.3.2 Validation of the radiomic signature

The C-index of the radiomic signature after the 10-fold cross-validation in BD1 was 0.67 (95 % CI [0.61-0.73]) and the C-index computed in the external validation on BD2 was 0.63 (95 % CI [0.53-0.73]).

The Kaplan-Meier curves for the high and low risk groups according to radiomics are displayed in Figure 6.2. The curves are displayed for the first 60 months, in order to make the results of BD1 and BD2 comparable. By looking at the p-values of the log-rank tests it is possible to observe that in both BD1 and BD2 datasets the curves for high risk and low risk groups are significantly different (p=0.001 and p=0.016 respectively). In particular, high risk patients had the worst outcome.

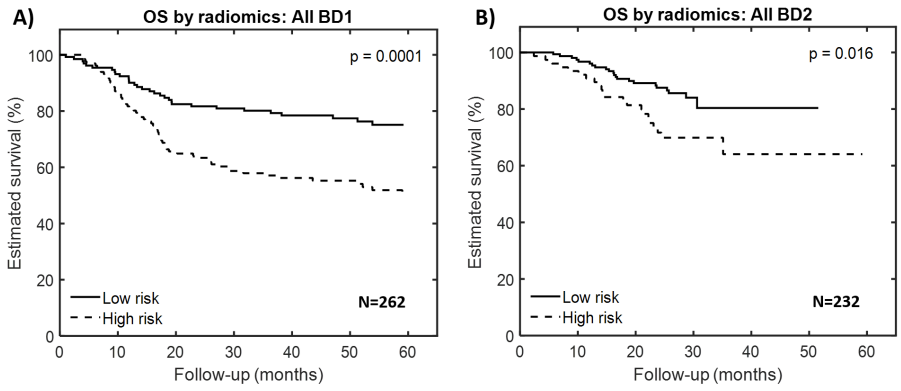

Figure 6.2: *Kaplan-Meier curves for overall survival (OS) in high and low risk groups according to radiomics. Sample sizes and p-values of log-rank tests are also indicated. A) Cross-validation on the training set (BD1). B) External validation set (BD2).*

6.3.3 Correlation between radiomic signature and clinical variables

Figures 6.3-6.5 illustrate the dependency of the radiomic signature on the prognostic clinical features: Stage TNM VIII, HPV status and tumor sub-site. Results are displayed for merged BD1 and BD2 datasets.

Figure 6.3 shows the distribution of the radiomic signature by Stage TNM VIII. It can be seen that the values of signature were significantly different across stage ($p=1.31*10^{-11}$ for Kruskal-Wallis test). In particular higher stages corresponded to higher signature values, with stage I-II presenting significantly lower value compared to stage III-IV ($p<0.01$ in post-hoc comparisons).

Figure 6.4 shows the distribution of the radiomic signature by HPV status. The radiomic signature was significantly lower for HPV+ patients ($p=7.00*10^{-8}$ for Mann-Whitney test).

Figure 6.5 shows the distribution of the radiomic signature by tumor sub-site. There were significant differences among the distributions of the radiomic signature across the different sub-sites ($p=1.15*10^{-10}$ for Kruskal-Wallis test), with significantly higher value for oral cavity compared to oropharynx ($p=3.98*10^{-9}$) and larynx ($p=9.98*10^{-4}$). the signature for oral cavity was the highest but the difference with hypopharynx was not significant (p=0.99).

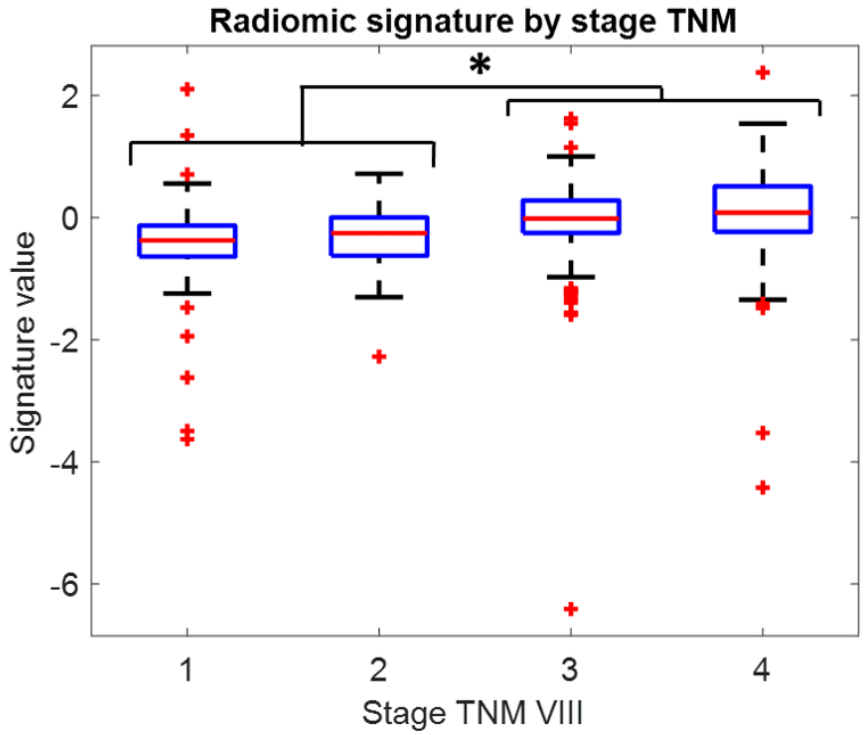

Figure 6.3: *Boxplots representing the distribution of radiomic signature across stages. Signature is significantly higher for stage III-IV than for stage I-II, as highlighted by the asterisk.*

6.3.4 Radiomic signature dependency on vendor and center

The results of the 2-way ANOVA performed to evaluate the effect of scanner and center are displayed in Table 6.8. According to the results, the scanner vendor had no effect on the value of the radiomic signature. The clinical center had an effect on the radiomic signature for both alive and dead patients (p=0.0053 and p=0.0001 respectively). However, this effect may depend on the fact that one of the clinical prognostic features that are correlated with the radiomic signature (see Subsection 6.3.3) is confounded with the clinical center. To account for this, a 3-way ANOVA accounting for scanner, center and tumor sub-site was performed. In this new analysis, whose results are shown in Table 6.9, no influence of the recording center was observed.

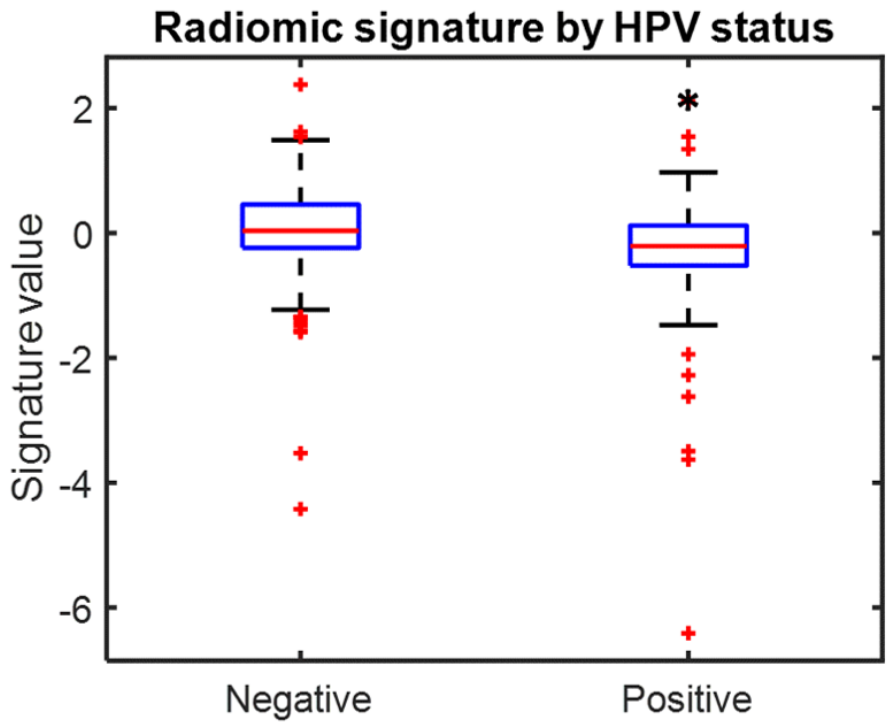

Figure 6.4: *Boxplots representing the distribution of radiomic signature for HPV positive and negative patients. The black asterisk highlights the significantly lower value of the signature for HPV positive patients.*

2-WAY ANOVA RESULTS		
Factor	**p-value (alive patients)**	**p-value (dead patients)**
Vendor	0.4912	0.9063
Center	0.0053	0.0001

Table 6.8: *Significance of the effect of center and scanner vendor on the radiomic signature, as defined by a 2-way ANOVA. Results are displayed for both alive and dead patients. Significant effect are highlighted in red.*

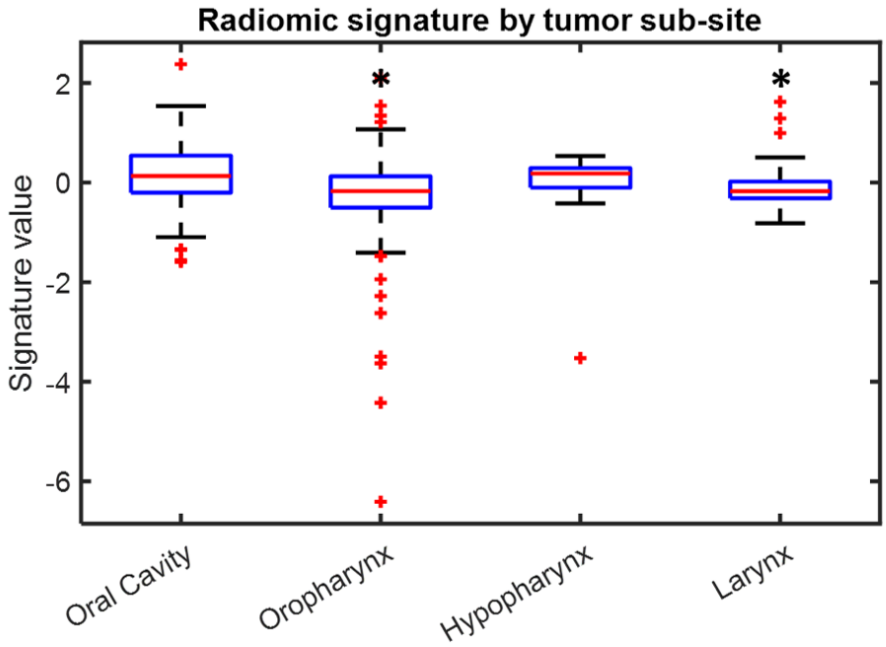

Figure 6.5: *Boxplots representing the distribution of radiomic signature across sub-sites. Asterisk highlights significantly lower values compared to the oral cavity group.*

3-WAY ANOVA RESULTS		
Factor	**p-value (alive patients)**	**p-value (dead patients)**
Vendor	0.6290	0.9260
Center	0.2920	0.5490
Tumor sub-site	<0.0001	0.0130

Table 6.9: *Significance of the effect of center and scanner vendor on the radiomic signature, as defined by a 2-way ANOVA. Results are displayed for both alive and dead patients. Significant effect are highlighted in red.*

6.3.5 Evaluation of added prognostic value of radiomics

Multivariate Cox analyses

The results of the multivariate Cox analyses are displayed in Tables 6.10 and 6.11 for BD1 and BD2 datasets respectively. In both tables the following prognostic variables are evaluated: radiomic signature; stage TNM

VIII; HPV status; tumor sub-site. The tables display the values of HR for the different variables and the associated p-values (p<0.05 if the HR is significantly different from 1).

MULTIVARIATE COX ANALYSIS (BD1)		
Feature	**Hazard ratio**	**p-value**
Radiomic signature	1.56	0.0050
Stage TNM VIII	1.84	0.0075
HPV status (positive vs negative)	1.02	0.9653
Oral cavity vs Hypopharynx	0.98	0.9497
Oropharynx vs Hypopharynx	0.65	0.3323
Larynx vs Hypopharynx	0.43	0.1654

Table 6.10: *Results on multivariate analysis on BD1 dataset.Results are displayed in terms of hazard ration and corresponding p-value. Significantly prognostic features are highlighted in red.*

MULTIVARIATE COX ANALYSIS (BD2)		
Feature	**Hazard ratio**	**p-value**
Radiomic signature	2.38	0.0118
Stage TNM VIII	1.24	0.5481
HPV status (positive vs negative)	0.29	0.1548
Oral cavity vs Hypopharynx	0.13	0.0025
Oropharynx vs Hypopharynx	0.25	0.0067
Larynx vs Hypopharynx	0.11	0.0038

Table 6.11: *Results on multivariate analysis on BD2 dataset.Results are displayed in terms of hazard ration and corresponding p-value. Significantly prognostic features are highlighted in red.*

When put in a multivariate Cox model, the value of the signature maintained its significance in both BD1 and BD2 datasets (BD1: HR=1.56, p=0.0050; BD2: HR=2.38, p=0.0118). The significance of the clinical variables depended on the particular dataset: in the BD1 dataset, stage was significantly prognostic; in the BD2 dataset, tumor sub-site was significantly prognostic. This may happen because in BD2 there is a non-uniform distribution of stages in some tumor sub-sites (e.g. all the hypopharynx are stage IV), so that the stage variable becomes confounded with the tumor sub-site variables.

Stratified Kaplan-Meier analysis

Figures 6.6-6.14 show the results of the stratified Kaplan-Meier analysis. Each figure displays the Kaplan-Meier curves for the high-risk and low risk groups defined according to radiomics as explained in Subsection 6.2.6, but in smaller, subgroups defined by the clinical variables of interest (stage, HPV status and tumor sub-site).

The results of the Kaplan-Meier analyses for stage I-III and stage IV are displayed in Figure 6.6 and Figure 6.7 respectively. The survival curves of high and low risk groups were significantly different for stage IV patients (log-rank test p-value 0.0001 and p=0.016 for the BD1 and BD2 patients respectively), but for stage I-III patients, the split between the curves was significant only in the training set (p=0.02).

The results of the Kaplan-Meier analysis stratified by tumor sub-site are displayed in Figures 6.8-6.13. The survival curves of high and low risk groups were significantly different for stage IV patients (log-rank test p-value 0.0001 and p=0.016 for the BD1 and BD2 patients respectively), but for stage I-III patients, the split between the curves was significant only in the training set (p=0.02).

For the patients with oral cavity cancer (Figure 6.8), radiomic caused a significant split in the survival in the BD1 dataset (log-rank p=0.026), while for the BD2 dataset the difference was not significant, even though the p-value of the log-rank test was close to 0.05 (p=0.074).

Figure 6.9 showed the results for patients affected by the oropharyngeal cancer. The high and low risk patients presented significantly different survival curves in the BD1 dataset (p=0.0089), but not in the BD2 dataset (p=0.15). After a further stratification in HPV+ and HPV- patients (Figures 6.10-6.11), no significant difference was found (p>0.065).

For the subgroups of patients with laryngeal (Figure 6.12) or hypopharyngeal cancer (Figure 6.13), no significant split between the Kaplan-Meier curves for high and low risk patients was observed (p>0.45).

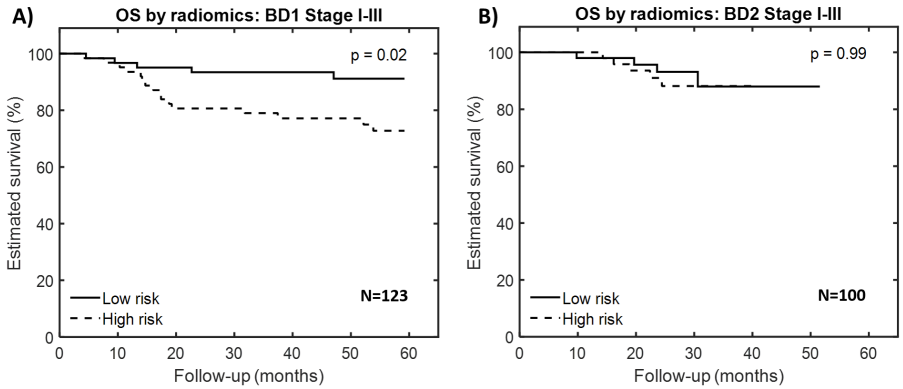

Figure 6.6: *Kaplan-Meier curves for overall survival (OS) in high and low risk groups according to radiomics. Results are displayed for patients with stage I-III tumors. Sample sizes and p-values of log-rank tests are also indicated. A) Cross-validation on the training set (BD1). B) External validation set (BD2).*

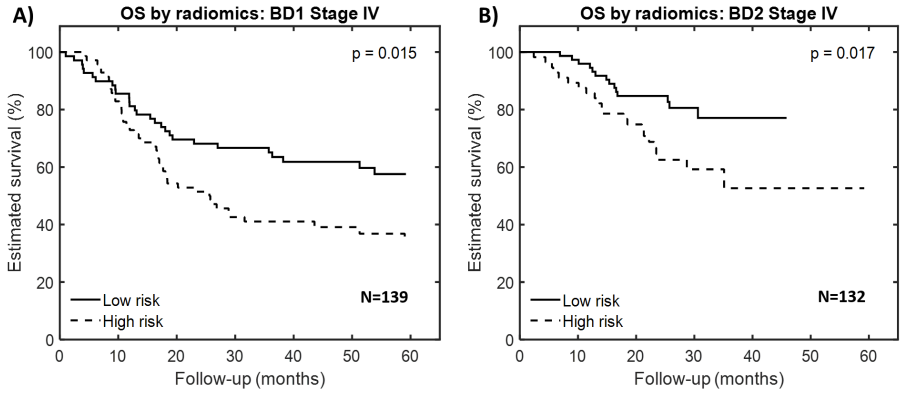

Figure 6.7: *Kaplan-Meier curves for overall survival (OS) in high and low risk groups according to radiomics. Results are displayed for patients with stage IV tumors. Sample sizes and p-values of log-rank tests are also indicated. A) Cross-validation on the training set (BD1). B) External validation set (BD2).*

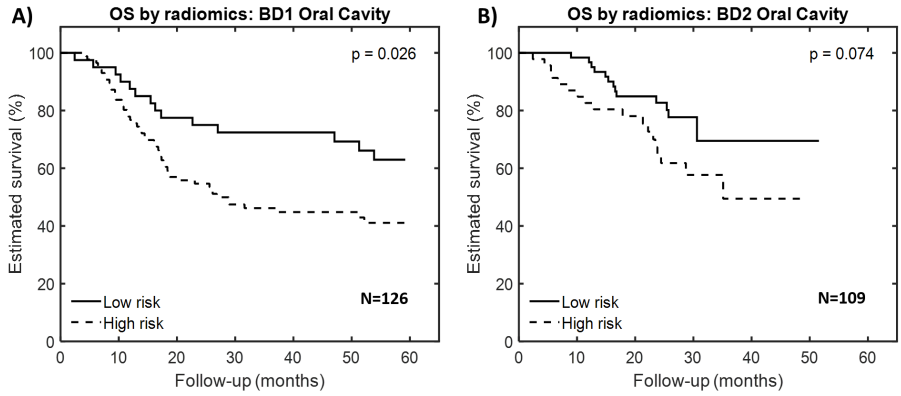

Figure 6.8: *Kaplan-Meier curves for overall survival (OS) in high and low risk groups according to radiomics. Results are displayed for patients with tumors of the oral cavity. Sample sizes and p-values of log-rank tests are also indicated. A) Cross-validation on the training set (BD1). B) External validation set (BD2).*

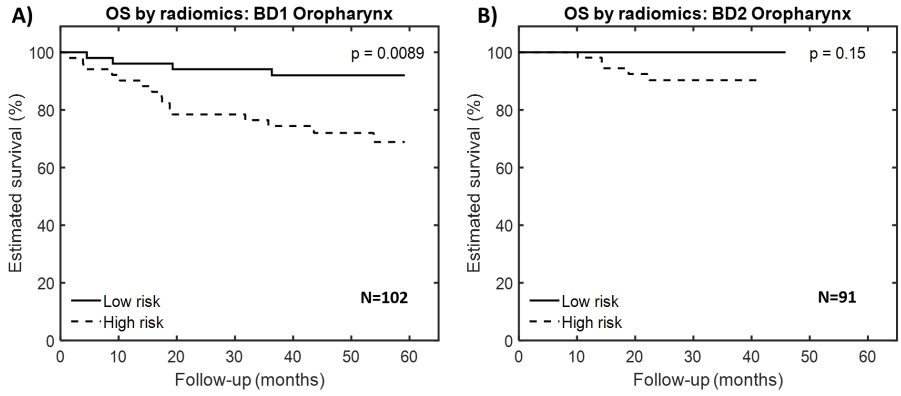

Figure 6.9: *Kaplan-Meier curves for overall survival (OS) in high and low risk groups according to radiomics. Results are displayed for patients with tumors of the oropharynx. Sample sizes and p-values of log-rank tests are also indicated. A) Cross-validation on the training set (BD1). B) External validation set (BD2).*

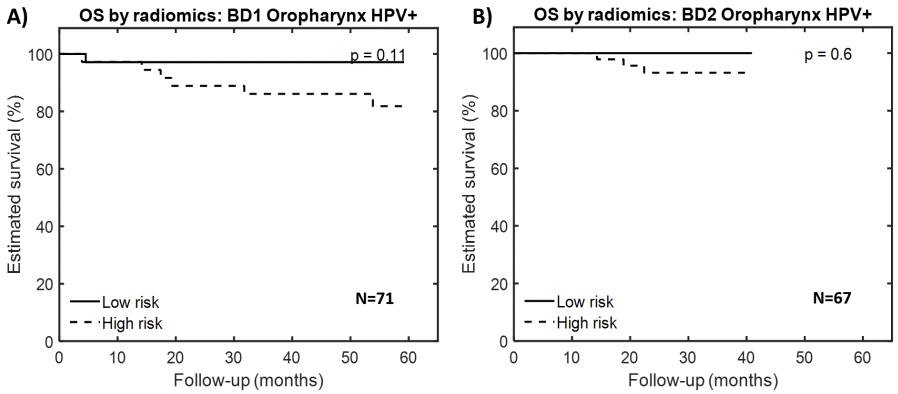

Figure 6.10: *Kaplan-Meier curves for overall survival (OS) in high and low risk groups according to radiomics. Results are displayed for patients with HPV+ oropharyngeal cancer. Sample sizes and p-values of log-rank tests are also indicated. A) Cross-validation on the training set (BD1). B) External validation set (BD2).*

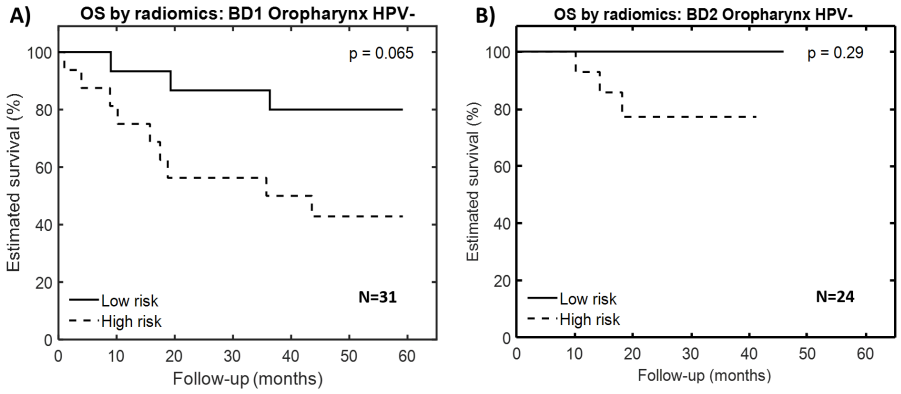

Figure 6.11: *Kaplan-Meier curves for overall survival (OS) in high and low risk groups according to radiomics. Results are displayed for patients with HPV- oropharyngeal cancer. Sample sizes and p-values of log-rank tests are also indicated. A) Cross-validation on the training set (BD1). B) External validation set (BD2).*

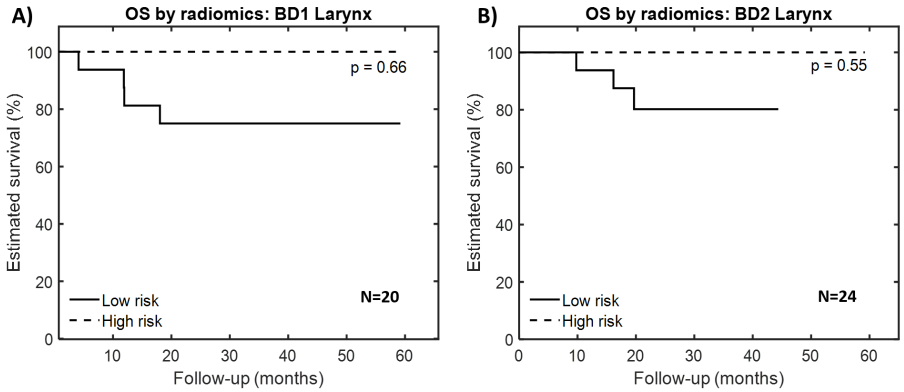

Figure 6.12: *Kaplan-Meier curves for overall survival (OS) in high and low risk groups according to radiomics. Results are displayed for patients with laryngeal cancer. Sample sizes and p-values of log-rank tests are also indicated. A) Cross-validation on the training set (BD1). B) External validation set (BD2).*

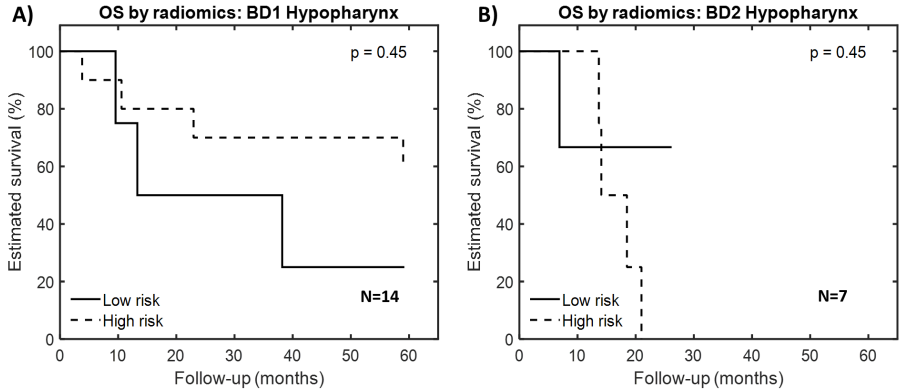

Figure 6.13: *Kaplan-Meier curves for overall survival (OS) in high and low risk groups according to radiomics. Results are displayed for patients with Hypopharyngeal cancer. Sample sizes and p-values of log-rank tests are also indicated. A) Cross-validation on the training set (BD1). B) External validation set (BD2).*

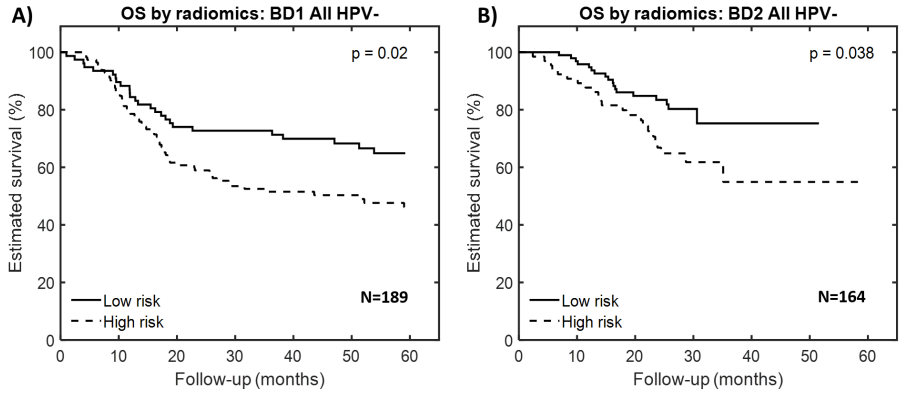

Figure 6.14: *Kaplan-Meier curves for overall survival (OS) in high and low risk groups according to radiomics. Results are displayed for patients with HPV- cancer. Sample sizes and p-values of log-rank tests are also indicated. A) Cross-validation on the training set (BD1). B) External validation set (BD2).*

The last analysis was performed by stratifying the patients by HPV status. For this analysis, the HPV- group included HPV- oropharyngeal tumors, but also all the patients with tumors in other sub-sites. The approximation of treating all non-oropharyngeal tumors as HPV- was done because HPV+ status does not positively affect prognosis, and so, from a prognostic point of view, they are closer to HPV- oropharyngeal cancer. For the patients with HPV- tumors (Figure 6.14), the radiomics-based risk classification provided two groups with significantly different survival curves in both BD1 and BD2 datasets (p-values 0.02 and 0.038 respectively). Differences in survival curves of patients with HPV+ tumors (all oropharyngeal cancers) were not significant, as displayed in Figure 6.10.

Comparison of clinical, radiomic and combined signatures

The distributions of C-index obtained by the clinical, radiomic and combined signature described in Subsections 6.2.5 and 6.3.1 are displayed in Figures 6.15-6.16, for both BD1 and BD2 datasets. The features used in the clinical, radiomic and combined model were the ones listed in Table 6.7.

In BD1 dataset, the radiomic and clinical signature performed similarly (median C-index 0.67, 95% CI [0.61-0.73] for both the models), but the best signature was the combined one (median C-index 0.69, 95% CI [0.63-0.75]). In the BD2 dataset the performance of the radiomic signature was lower compared to the clinical signature (median C-index 0.63 vs 0.69,

95%CI [0.53-0.73] vs [0.61-0.77]) but the combined signature was the best in this case as well (median C-index 0.72, 95% CI [0.64-0.80]).

6.4 Discussion

The experiment performed in this chapter gave some insight about the usefulness of the developed pipeline in the creation of a radiomic signature for OS in HNC.

The radiomic signature was composed of 5 features. Among those, one was voxel volume, which confirms the role of volume in the prediction of OS. The feature *T1w-waveletLHL-firstorder-90Percentile* is related to the intensity of the high-pass filtered signal in T1w, while the features on the T2w account for the differences in texture (the grey level non-uniformity normalized) or for ranges of intensities (range and inter-quartile range). This is in line with the fact that T2w images provide better contrast than T1w images (which is the reason why T1w images are often contrast-enhanced) and the tumor heterogeneity can be better appreciated in that type of images.

The radiomics signature had a C-index of 0.67 and 0.63 in the training and validation set respectively. This values are in line with other prognostic signature for OS presented in other studies [5, 137, 138]. Unlike those studies though, the signature presented in this chapter has been obtained by training on a multicentric cohort. This is a further proof of the fact that an adequate pre-processing of the images and a proper choice of the features to use can lead to the application of radiomics in a multicentric context.

The multivariate Cox analysis showed how, in both BD1 and BD2, the radiomic signature maintained a significant prognostic value, meaning that it adds independent prognostic information. The added value of radiomic features is also proved by the fact that the combined signature, obtained as a linear combination of radiomics and clinical features, performed better than the signatures based on clinical and radiomic features alone.

The radiomic-based classification of the patients in high and low risk led to groups with significantly different survival curves in both BD1 and BD2 datasets. However, when considering the different subgroups (by stage, HPV or sub-site) this significance in the survival curves of high and low risk patients was not always maintained. Radiomic signature showed good stratification of survival for stage IV, oral cavity cancers and HPV- patients, but lower stratification power in the other subgroups. For some groups (Larynx and oropharynx) the lack of significance may simply be related to the low number of cases or events. A future development for this study

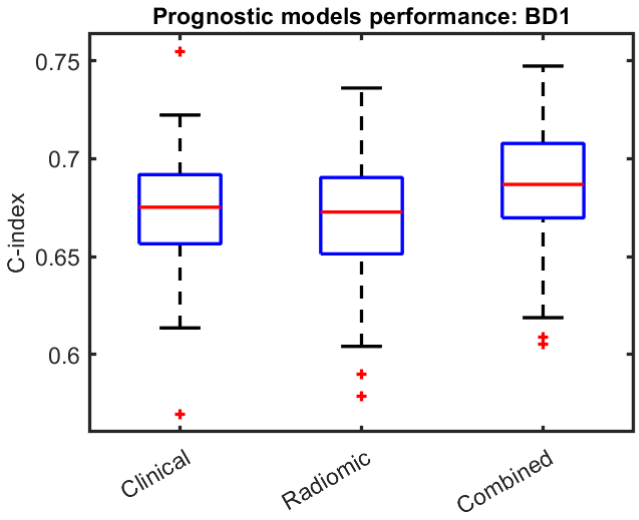

Figure 6.15: *Distributions of C-index for the cross-validated signature values in the BD1 dataset. Distributions were obtained by bootstrap of the original values (100 iterations). The combination of radiomic and clinical features resulted in the model with the best performance.*

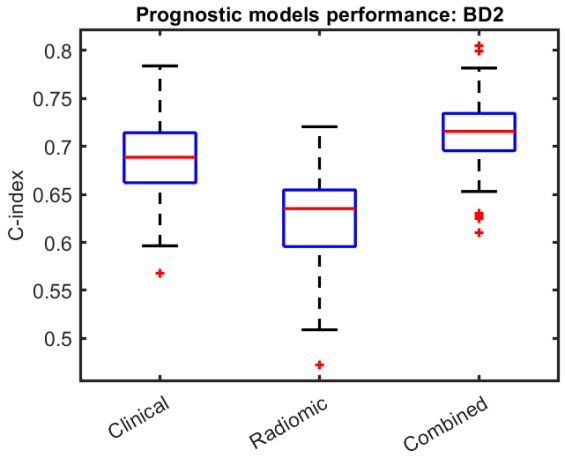

Figure 6.16: *Distributions of C-index for the signatures in BD2 dataset. Distribution were obtained by bootstrap (100 iterations). The combination of radiomic and clinical features results in the model with the best performance.*

would be to increment the number of patients per dataset in order to start using the optimized workflow to train a better model that allows a good patient stratification for those sub-sites as well.

The positive synergy between radiomics and clinical was a recurrent results in many studies of radiomics [11, 138] and underlines the fact that radiomics is a source of information that is independent from the one obtainable with the routine clinical exams and it is therefore a useful tool to support the decision of the clinicians in the treatment and monitoring of a patient with HNC.

It must be noted though that the performance of the radiomics signature reduced when moving from the training set (BD1) to the test set (BD2) and that was not the case for the clinical and combined model. This may be due to the fact that, although many precautions were taken to harmonize the features as much as possible, a perfect standardization of the features has not been reached yet. Of course in the future new methods to fatherly increase the harmonization of the features coming from different centers/scanners could be developed. Also, the addition of ADC maps, which are known to have a low inter-scanner variability [62, 63], could further improve the performance of a signature in a more generalized context. However, the results shown demonstrated that the developed workflow could reduce a sufficient amount of variability to allow the creation of a prognostic signature that can be used even in a more general context and that could be potentially applicable in the clinical practice.

In conclusion, the optimized workflow for radiomic-based survival analysis displayed in Figure 5.5 was used on a real HNC dataset to train a prognostic radiomic signature, which was successfully tested using both cross-validation and external validation. The signature may provide a useful tool to aid prognosis for advanced stage HNC patients.

CHAPTER 7

Radiomics-based survival models for soft-tissue sarcoma

This chapter describes the application of an optimized radiomic workflow for the development of an MRI-based prognostic signature for OS in patients affected by STS. The signature was trained its prognostic power was evaluated using cross-validation. Evaluation of the added prognostic power of radiomics was also performed.

7.1 Introduction

STS is a rare type of cancer accounting for less than 1% of the total cancers worldwide, with an incidence of one third of that of HNC [144]. Given the rarity of the tumor, it does not surprise that the number of studies involving STS is small compared to tumor of other districts. However, the use of radiomics in STS has been explored for different applications including tumor non-invasive characterization, metastasis prediction, treatment response and survival model [100, 117, 145–148].

One of the most frequent application of radiomics in STS is non-invasive identification of tumor grading which has been performed using both CT

[117] and MRI [100, 146]. In the case of MRI, radiomics-based tumor grading has been performed using mono-modality MRI, and previous studies have focused on T2w MRI [146] and ADC maps [100]. All the MRI based models achieved good results (AUC>0.8 and accuracy>0.8), but have the limitation of the small sample size. Studies involving CT-radiomics on larger datasets [117] did not perform as well (AUC=0.64).

Another application of radiomics in STS is prediction of distant metastasis. In [145] a fused PET/MRI model was developed to predict lung metastasis by a radiomic analysis of the primary tumor (the STS). The results of bootstrap cross-validation were excellent in terms of AUC, sensitivity and specificity (0.984, 0.955 and 0.926 respectively). The limitations of the model are that it requires both PET and MRI, which are not always performed jointly, and that it was obtained using a well defined image acquisition protocol, which reduces the possibility to extend the model to a more generalized context.

In [148] MRI-radiomics of STS was used for prediction of response to induction chemotherapy. In particular, a random-forest classifier on delta-radiomics, i.e. difference between radiomic features at two different time points, from T1w and T2w MRI was trained on 65 patients and validated trough train-test split, obtaining high values of AUC and sensitivity (0.86 and 94% respectively) but low level of specificity (66%).

In terms of survival analysis, models based on both CT- and MRI-radiomics were developed for STS, for both DFS and OS [117, 147]. In [117] CT-radiomics was used to develop signatures survival models that showed high C-index for OS, distant metastasis free survival and loco-regional recurrence free survival (0.73, 0.68 and 0.77) respectively. In [147], T1w-MRI was used to create a prognostic signature for OS that significantly improved the performance of a clinical model (up to a C-index of 0.74). The limitation of this study is that the follow-up was limited to 3-years, so the performance of the model on longer times could not be determined. Also, only one imaging modality was explored and still leaves room for investigating the additional power given by multi-modality MRI.

The experiment presented in this chapter deal with the creation of survival models for STS. In particular, ADC maps and different MRI sequences (T1w pre- and post-contrast, as well as T2w) were used to develop prognostic models of OS in STS.

7.2 Material and methods

7.2.1 Image dataset

The experiment presented in this section was based on a monocentric retrospective dataset collected for the ongoing study *Integration of radiOMics, genomICS and immunoprofiling into predictive and prognostic models in soft tissue SARComa patients* (from now on called SARCOMICS for short), funded by a research grant from the Italian Ministry of Health. The dataset contained MRI images of patients with STS of the limbs that were acquired between 2011 and 2015 at INT, using non-standardized image acquisition protocols. Two different scanners were used to acquire the images and parameters such as TR, TE and image resolution were not controlled. The dataset used for this experiment (called SAR1 from now on) was composed of the patients of the SARCOMICS dataset that fulfilled the following inclusion criteria: availability of T1w MRI, both pre- and post-contrast, acquired with TSE pulse sequence; availability of T2w MRI acquired with TSE pulse sequence; availability of DWI images acquired using at least two b-values in the range 0-1000 s/mm^2, acquired using EPI pulse sequence. In total, 91 patients were selected for the experiment. Main clinical and follow-up data for the selected patients are reported in Table 7.1. Details of the image acquisition parameters are listed in Table 7.2. By comparing Table 7.2 and Table 3.5, it is possible to see that most of the acquisition parameters were in the range used for the stability analyses of Chapter 3. For the selected patients, ADC maps were obtained as described in Subsection 2.3.4 by fitting an exponential decay on DWI images acquired with the different b-values.

7.2.2 Image segmentation

For each patient, the main tumor was manually segmented by a radiologist with more than 10 years of experience. The segmentation was performed using the T2w MRI as the reference and the same ROI was used also for T1w images and ADC maps, since only small misalignment are present (see Figure 7.1), and in Chapter 4 it has been shown that the majority of the features to which the majority of radiomic features is stable to shifts of this entity.

7.2.3 Image preprocessing

The optimal preprocessing pipeline described in Chapters 3-4 and Figure 5.5 was applied to the T1w and T2w MRI prior to the radiomic features ex-

CLINICAL DATA (SAR1 DATASET)	
Number of patients	91
Age (median and IQR)	59 years [46-71]
Sex	Female: 40 (44%) Male: 51 (56%)
Histological grade	Grade 1: 22 (24%) Grade2: 21 (23%) Grade 3: 48 (53%)
Treatment	Surge: 51 (56%) Surge+Rad: 8 (9%) Surge+Chem: 5 (5%) Surge+Rad+Chem: 27 (30%)
Follow-up time (median and IQR)	54 months [42-72]
Number of deaths	16 (18%)
Number of recurrences	27 (30%)

Table 7.1: *Clinical and demographic characteristics of the 91 patients of the SAR1 dataset. Age and follow-up time are displayed as median and inter-quartile range (IQR).*

traction. First, a 3D Gaussian filter with a 3x3x3 voxel kernel and $\sigma = 0.5$ was used to denoise the images. Then, the N4ITK algorithm [106] was used for the correction of intensity-non uniformities. Intensity standardization was performed using Z-score. Voxel size resampling to an isotropic resolution of 2 mm was performed using B-spline interpolation.

ADC had a different preprocessing compared to T1w and T2w images. For reasons that were explained in Subsections 2.3.4 and 4.2.3, intensity standardization and inhomogeneity correction were not performed. Prior to features extraction, intensity values were windowed between 0 and $4000*10^{-6}$ mm^2/s, in order to remove non-physiological values due to image noise in the DWI used to fit the ADC maps.

7.2.4 Radiomic features extraction

A set of 799 stable features was used for the analysis (see Appendix A). The extracted features were grouped as follows: 182 features for pre-contrast T1w MRI (13 shape, 12 FOS, 10 GLCM, 8 GLRLM and 139 wavelet); 182 features for post-contrast T1w MRI (the same as pre-contrast MRI); 242

SAR1 DATASET ACQUISITION PARAMETERS			
Image sequence	**T1w/T1wCont**	**T2w**	**ADC**
Scanner	Philips Achieva: 78 (87%) Siemens Avanto: 13 (13%)	Philips Achieva: 78 (87%) Siemens Avanto: 13 (13%)	Philips Achieva: 78 (85%) Siemens Avanto: 13 (15%)
Number of images	91	91	91
Pulse sequence	Spin-echo	Spin-echo	Echo-planar
Magnetic field	1.5 T	1.5 T	1.5 T
Time of repetition	411-745 ms	3000-6500 ms	3972-11145 ms
Time of echo	7-14 ms	80-153 ms	64-88 ms
Slice thickness	3-5 mm	3-5 mm	4-5 mm
Slice spacing	3.9-6.5 mm	3.9-6.5 mm	4-6.5 mm
Pixel spacing	0.3-1.13 mm	0.34-1.22 mm	1.28-2.34 mm

Table 7.2: *Description of the image acquisition details for the 91 patients of the SAR1 dataset.*

features for T2w MRI (13 shape, 13 FOS, 10 GLCM, 11 GLRLM and 195 wavelet); 193 features for ADC (13 shape, 14 FOS, 9 GLCM, 13 GLRLM and 144 wavelet). A fixed bin number intensity discretization (32 bins) was used prior to the features extraction.

7.2.5 Prognostic models training

The data of SAR1 dataset were used to train 3 different prognostic signatures for OS. An illustration of the training process for the 3 models was previously illustrated in Figure 6.1. Each source of data (radiomics or clinics) underwent its own postprocessing pipeline, and at the end of the pipeline an optimal set of features was selected. The features sets were used to train a radiomic and clinical signature using multivariate Cox proportional hazard regression [73]. Last, a combined signature was obtained

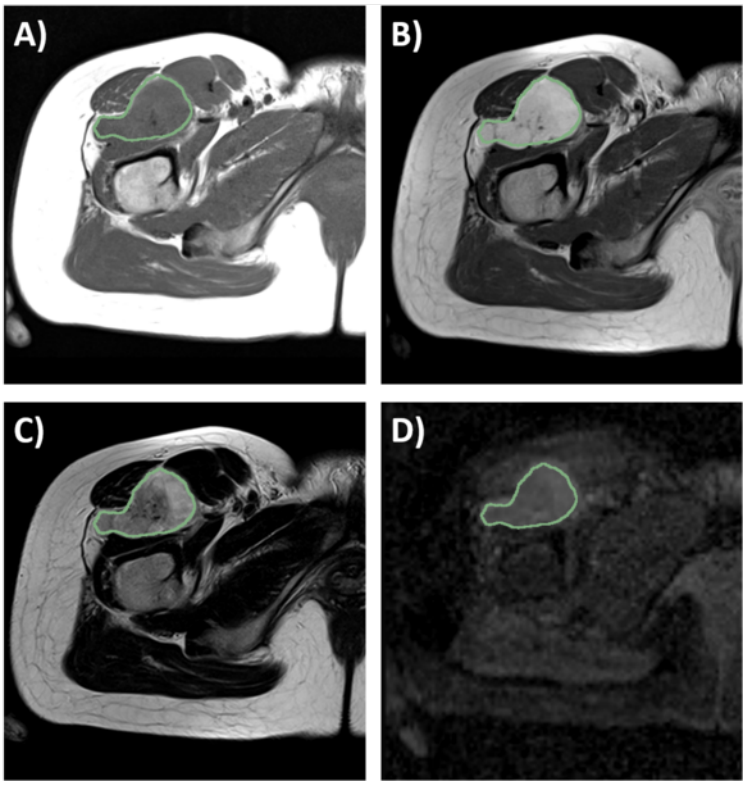

Figure 7.1: *Examples of images available for the patients of the SAR1 dataset. A) Pre-contras T1-weighted image. B) Post-contrast T1-weighted image. C) T2-weighted image. D) Apparent diffusion coefficient maps. The segmentation of the tumor is also shown.*

by training a Cox regression model on the combination of the radiomic and clinical features sets.

For radiomic features, the postprocessing pipeline is the one described in Figure 5.5. Z-score normalization for the standardization of the ranges of features. Then, the significance-based selection pipeline (described in Subsection 5.2 and Figure 5.2A) was used to choose the optimal features set.

The clinical variables of interest were the following (see also Table 7.1): age at diagnosis, sex, histological grade. Sex was represented as a dummy variable (1 for male, 0 for female) [68]. The features selection was performed by selecting only the features that were significantly associated with OS in univariate Cox regression.

7.2.6 Validation of the radiomic signature

To evaluate the prognostic performance of the radiomic signature an internal cross-validation was performed using 10-fold cross validation, in order to obtain an unbiased estimate of each signature for each patient (as previously illustrated in Figure 5.3). Such estimates were used to compute the Harrell's C-index for each model and confidence intervals were obtained using bootstrap (100 iterations). Moreover, patients were split in high and low risk groups and the Kaplan-Meier curves were drawn for each group. Log-rank tests were used to compare the Kaplan-Meier curves for high and low risk patients. The median value of the unbiased radiomic signature in the SAR1 was used as a threshold to split the high and low risk groups.

7.2.7 Correlation between radiomic signature and clinical variables

The radiomic signature could potentially be highly correlated with other clinical variables. To ensure this was not the case, the correlation of the radiomic signature with the selected clinical features was evaluated. Statistically significant associations were identified using Spearman correlation coefficient and/or Kruskal-Wallis test.

7.2.8 Radiomic signature dependency on scanner

The radiomic signature could also potentially be affected by factors not related to tumor biology. In the case of the SAR1 dataset the only possible batch was the MRI-scanner (Philips or Siemens). To evaluate signature differences in radiomic signature due to scanner a Mann-Whitney test was used. The analysis was performed separately on dead and alive patients at the end of the follow-up because it is reasonable to think that patients with the worst outcome will have a significantly higher signature.

7.2.9 Evaluation of added prognostic value of radiomics

To evaluate the added prognostic value of the radiomic, three different analyses were performed.

In the first analysis the prognostic power of the radiomic signature was evaluated by looking at its HR and p-value in a multivariate Cox regression model with the other selected clinical variables.

For the second analyses, the different subgroups defined by grade (I-II vs III) were considered. The Kaplan-Meier analysis described in Subsection 7.2.6 was repeated for each subgroup. This was done to assess whether the discriminatory power of the radiomic classification was high in each of the

groups or was subgroup dependent.

For the third analyses, the C-indexes of the clinical, radiomic and combined signature (trained as described in 7.2.6) was computed to assess whether the prognostic performance of the radiomic or combined signature was better than the one of the clinical signature alone.

7.3 Results

7.3.1 Prognostic models training

Only one feature was selected for the clinical and radiomic models. In particular, the clinical feature was tumor grade, while the radiomic feature was *T1-waveletHLL-firstorder-Median*. Table 7.3 lists the mean and standard deviation of the radiomic feature, which were used for the Z-score normalization. The combined model was obtained by using both the features. Table 7.4 show the coefficients of the features in the different models.

SELECTED RADIOMIC FEATURES DETAILS		
Features names	**Mean**	**Standard deviation**
T1w-waveletHLL-firstorder- Median	0.005	0.008

Table 7.3: *Mean and standard deviation of the selected radiomic feature. These values were used to compute the Z-scored versions of the feature. Numeric values are reported up to the second decimal digit.*

SELECTED FEATURES COEFFICIENTS			
Features names	**Coefficient (clinical)**	**Coefficient (radiomic)**	**Coefficient (combined)**
T1w-waveletHLL-firstorder- Median	-	-0.83	-0.52
Grade	2.21	-	1.91

Table 7.4: *Coefficients of radiomic and clinical features for the three models tested: clinical, radiomic and combined. Coefficients values are displayed up to the second decimal digit.*

7.3.2 Validation of the radiomic signature

The C-index of the radiomic signature after the 10-fold cross-validation in SAR1 dataset was 0.74 (95 % CI [0.64-0.84]).

The Kaplan-Meier curves for the high and low risk groups according to radiomics are displayed in Figure 7.2. The p-value of the log-rank test was p=3.94*10^{-4}, highlighting a significant survival between the two groups. In particular, high risk patients had the worst outcome, as expected.

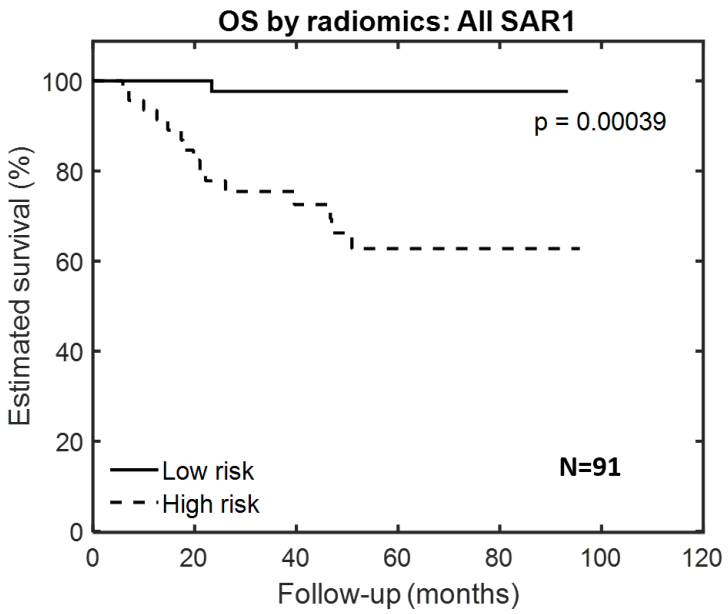

Figure 7.2: *Kaplan-Meier curves for overall survival (OS) in high and low risk groups according to radiomics. Sample sizes and p-values of log-rank tests are also indicated. Results are displayed for the cross-validation on SAR1 dataset*

7.3.3 Correlation between radiomic signature and clinical variables

Figure 7.3 illustrates the dependency of the radiomic signature on histological grade. There were significant differences among the grade (p=0.0021 for Kruskal-Wallis test). In particular, the median signature for grade 3 compared to of each grade is significantly higher compared to the ones for Grade 1 and 2 (p=0.0107 and p=0.0144 in post-hoc comparisons respectively).

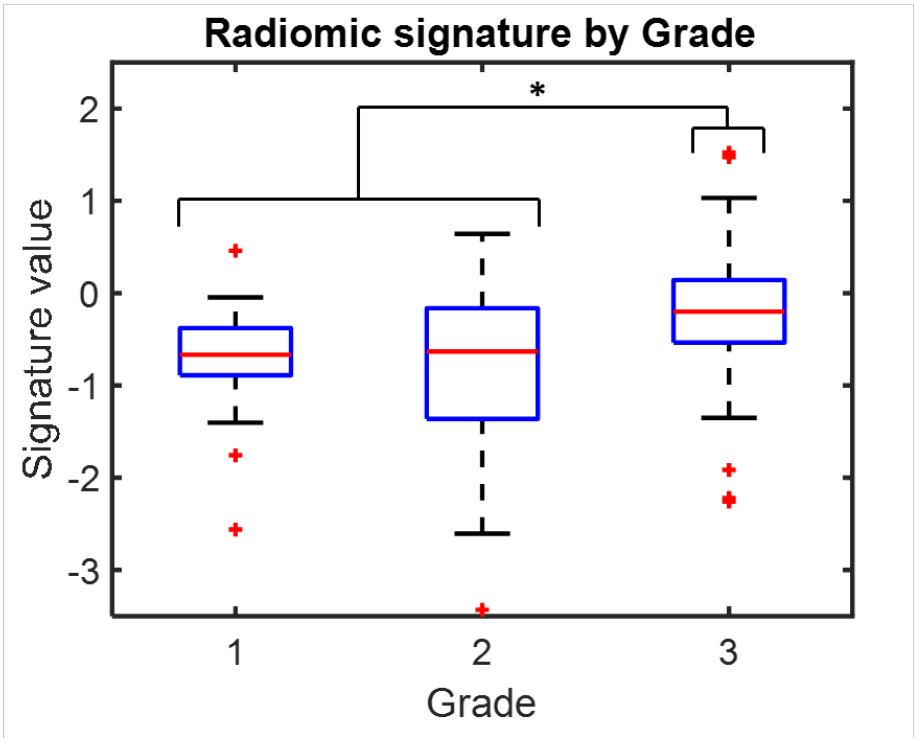

Figure 7.3: *Boxplots representing the distribution of radiomic signature across grades. Signature is significantly higher for grade 3 that for grade 1-2, as highlighted by the asterisk.*

7.3.4 Radiomic signature dependency on scanner

Figures 7.4A-B show the boxplots with the distribution of the radiomic signature across scanners for the alive and dead patients respectively. No significant difference was found between signature in the two scanners. The p-value of the Mann-Whitney test were 0.4 and 0.93 for the alive and dead patients respectively.

7.3.5 Evaluation of added prognostic value of radiomics

Multivariate Cox analyses

The results of the multivariate Cox analyses of the unbiased signature value in SAR1 dataset are displayed in Table 7.5. In both tables the following prognostic variables are evaluated: radiomic signature; tumor grade. The tables display the values of HR for the different variables and the associated p-values ($p<0.05$ if the HR is significantly different from 1).

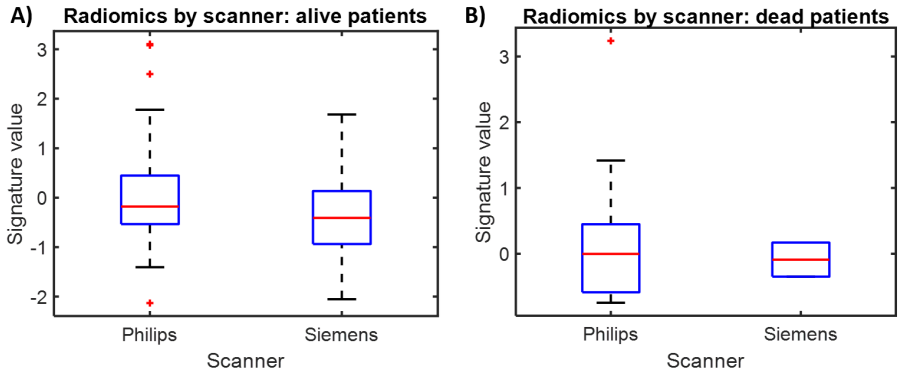

Figure 7.4: *Boxplots representing the distribution of radiomic signature across scanners. A) Alive patients. B) Dead patients.*

MULTIVARIATE COX ANALYSIS (SAR1)		
Feature	**Hazard ratio**	**p-value**
Radiomic signature	1.86	0.0479
Tumor grade	6.75	0.0376

Table 7.5: *Results on multivariate analysis on SAR1 dataset.Results are displayed in terms of hazard ration and corresponding p-value. Significantly prognostic features are highlighted in red.*

When put in a multivariate Cox model, the value of the signature maintained its significance, although the p-values was close to the limit of significance (HR=1.86, p=0.0479).

Stratified Kaplan-Meier analysis

Figure 7.5 of the Kaplan-Meier analysis, stratified by tumor grade (1-2 vs 3). For grade 3 patients, the Kaplan-Meier curves for the high and low risk groups defined by radiomics were significantly different (p=0.043 for log-rank test), with high risk patients showing the worst outcome. No significant difference was found for Grade 1-2 patients (p=0.98).

Comparison of clinical, radiomic and combined signatures

Figure 7.6 shows the distributions of C-indexes for the three models obtained after 10-fold cross-validation on dataset SAR1. The radiomic and clinical model performed similarly (radiomic: median C-index 0.74, 95%

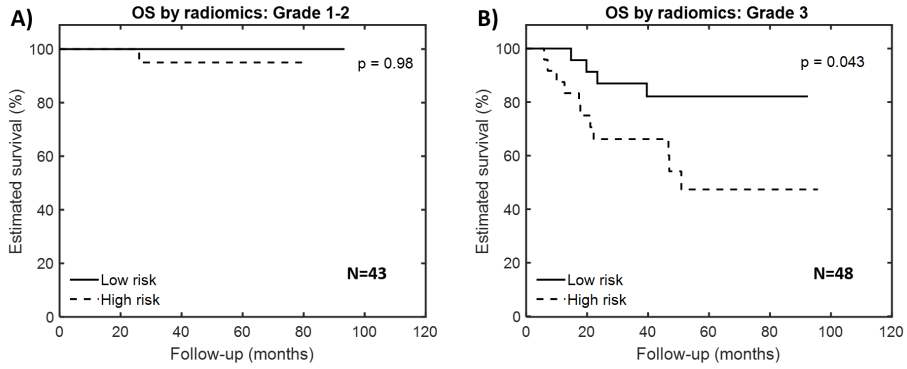

Figure 7.5: *Kaplan-Meier curves for overall survival (OS) in high and low risk groups according to radiomics. Sample sizes and p-values of log-rank test are also displayed. A) Grade 1-2 tumors. B) Grade 3 tumors.*

CI [0.64-0.84]; clinical: median C-index 0.74, 95% CI [0.68-0.80]), while the combined model performed better (median C-index 0.78, 95% CI [0.70-0.86]).

7.4 Discussion

In this chapter, the usefulness of MRI-radiomics for the development of prognostic models for OS in STS has been investigated. As observed in Chapter 6 for HNC, the addition of radiomic features provides independent prognostic information.

In the case of STS, the only radiomic features that was used in the radiomic model was *T1w-waveletHLL-firstorder-Median*, which was associated with a negative Cox coefficient (-0.52 and -0.83 for the radiomic and combined model respectively, see Table 7.4). This means that tumors that are iper-intense wavelet transform of the T1w images are associated to a lower risk of death. This behaviour was in agreement with the one observed for HNC, in which the feature *T1w-waveletLHL-firstorder-90Percentile*, which has a behaviour that is similar to *T1w-waveletHLL-firstorder-Median*, was associated to a negative Cox coefficient (-0.27 and -0.15 for the radiomic and combined model, see Table 6.7).

When comparing the cross-validation C-index obtained from SAR1 with the results presented in Subsection 6.3.5 for HNC, it can be seen that the values obtained for STS are higher. This may suggest that, although MRI could be used to identify phenotypic differences in both districts, it is particularly useful for STS. However, the better result could be also due to the

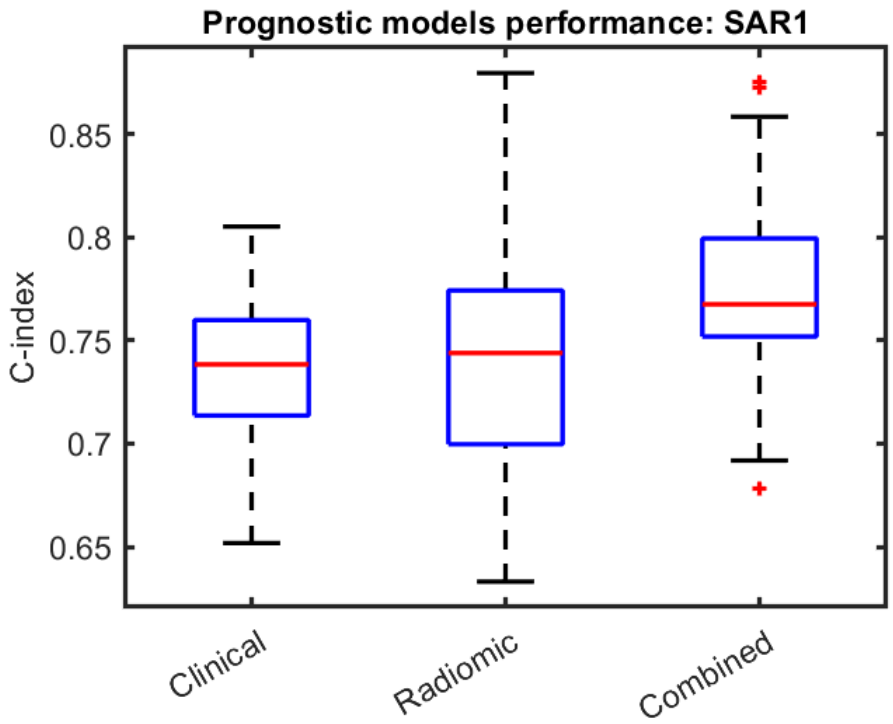

Figure 7.6: *Distributions of C-index for the cross-validated signature values in the SAR1 dataset. Distributions were obtained by bootstrap of the original values (100 iterations). The combination of radiomic and clinical features resulted in the model with the best performance.*

fact that while the BD2Decide datasets only include patients with advanced disease, the same cannot be said for the SARCOMICS dataset.

The results showed that the radiomic signature was correlated with histological grade and was significantly higher for grade 3 tumors. However, the multivariate Cox analysis showed that the radiomic signature added independent prognostic information. As a matter of fact, radiomic could further stratify survival in the group of grade 3 patients. Also, the combined signature showed a better cross-validation c-index compared to the clinical and radiomic ones (0.79 vs 0.74). These results show the potential of radiomic in the development of prognostic models for OS in higher grade STS, but also in distinguishing lower grade from higher grade tumors (as previously found by [100]).

Only two other studies investigated the prognostic performance of radiomic features on OS. In [117], a gradient boosting based on radiomic

features was used to develop a CT-based radiomic signature for OS. The signature was trained on a monocentric dataset of 83 patients and validated on two different validation datasets of 87 and 51 patients respectively, obtaining C-index of 0.72, 0.73 and 0.59 respectively. The addition of clinical variables augmented the C-index up to 0.76. In [147], a LASSO-Cox regression was trained on a dataset of 165 patients using different combinations of radiomic features (from T1w MRI) and clinical features (age and grade) and validated on an external dataset of 61 patients. The developed models performed well on both datasets, with C-index up to 0.78. The mean C-index obtained by cross-validation in the SAR1 dataset was 0.74 for the radiomic alone and 0.78 for the combined model, showing values that are in the same range as the one of the above mention studies of literature. The result of the SAR1 dataset is particularly important because it was obtained by training from images acquired with non-standardized parameters. Also, while in [147], the C-index was calculated for a follow-up of maximum 3 years, in the experiment presented in this chapter it was shown that the same performance could be reached also for longer follow-ups (up to 8 years).

The study still has some limitations and the main one is the lack of an independent validation dataset. The acquisition of such dataset is part of the SARCOMICS project but the enrollment of the prospective patients is still ongoing and so no independent validation is available at the moment. The only validation that was performed was provided by cross-validation which is a first fundamental step when developing any statistical model but is not enough to provide definitive results. The collection of the prospective data is therefore the logical next step for this analysis.

In conclusion, in this chapter the radiomic workflow developed for this thesis was applied to an STS cohort including MRI images acquired with uncontrolled image acquisition parameters. Although an external validation is missing the results obtained through cross-validation are promising and comparable to the ones of previous studies of literature. This is a further proof of the fact that a proper preprocessing of the data can make radiomics applicable also for images acquired with different protocols.

CHAPTER 8

Conclusions

8.1 Summary of the main results

In this thesis, a workflow for MRI radiomic analysis that allowed to train prognostic models using MRI images collected with non-standardized acquisition protocols was developed. The workflow could be used to successfully train a model starting from multicentric MRI cohorts, which is a necessity when dealing with rare pathologies like HNC or STS, for which large monocentric cohorts are often not available.

This section reports a summary of the main results of the thesis that are reported with respect to the initial objectives as defined in Section 1.2.

Evaluation of features stability to imaging-related variability

In order to asses the stability of radiomic features to variation, a series of experiments were performed using simulated MRI images (T1w and T2w) obtained using a virtual phantom (BrainWeb) as described in Chapter 3. Four main different sources of variability were considered: 1) variations in TR/TE; 2) variations in voxels size; 3) image noise; 4) intensity non-uniformities;

The results showed that the stability of the radiomic features to imaging

related-variability was not dependent on the type of MRI used (T1w and T2w), but was mainly related to the features category.

Variations in TR/TE affected the FOS features more than the textural features. This is probably due to the fact that textural features depend on the number of bins used for grey-level discretization (which was set at 32 bins in all the analysis of the thesis) and are not affected by linear transformations of the histogram (scaling or translation), unlike many FOS features (like mean or standard deviation) that depend on the exact value of MRI signal intensity.

Voxel size was the factor causing the largest variability in radiomic features. In particular, textural features were the most effected. FOS and SS features are also affected, but the majority of the values of ICC (used to quantify stability) were higher compared to the ones of the textural features. The low values of ICC for textural features underline how important it is to either control the parameters that define voxel size (e.g. pixel spacing, spacing between slices) or to correct for lack of standardization of the parameters.

Random Gaussian noise had little effect on features values. The effect of noise was more evident in textural features, presenting lower ICC values compared to FOS. This may be due to the fact that FOS features, unlike textural features, do not depend on the spatial distribution of the gray values but only on the histogram of the grey values. Since the Gaussian noise used was a white noise (null mean), only minor changes in the histogram may occur, while the spatial distribution of the grey values may still be affected.

Intensity non-uniformity (INU) due to variations of the magnetic field affected both FOS and textural features. As a matter of fact, INU may introduce variations in both the spatial distribution of the grey values, but also major changes to the histogram.

Effect of image preprocessing on imaging-related variability

The experiments described in Chapter 3 also allowed to evaluate whether image preprocessing could help increasing the stability of radiomic features to variations in the image acquisition parameters. Different preprocessing techniques were investigated: bias-field correction via N4ITK algorithm; image denoising via Gaussian filtering; intensity standardization with different algorithms (Z-score, histogram stretching, histogram matching); voxel size resampling with cubic B-spline. All the aforementioned preprocessing steps seem to improve the stability of radiomic features to variability in the images acquisition conditions and were therefore included in the radiomic workflow.

Intensity standardization had a positive effect on stability of the FOS features to variations in TR/TE. Stability of textural features did not improve after intensity standardization. This behaviour can be explained by the fact that intensity standardization is equivalent to a rigid (linear) registration of the histogram. Such rigid transformation change the values of FOS features but does not affect the textural features, since it does not change the shape of the histogram or the spatial distribution of the discretized grey values. No significant differences were found among the intensity standardization algorithms (histogram matching, histogram stretching and Z-score normalization). This is a positive aspect because it means that other researchers that want to use intensity standardization may use the method that they prefer. For the future analyses of this thesis it was decided to use the default standardization method within pyradiomics (Z-score).

Spatial resampling with isotropic voxel size improved the stability of the features, in particular for the SS and textural features. SS features only depend on the geometrical properties of the ROI and therefore they stay equal if the grid on which the ROI is interpolated (i.e. the image grid) does not change. This explain why the ICC values of the features after the resampling were 1. In case of textural features, ICC values increased after resampling, probably because using the same resolution in all the 3 direction allowed to normalize the computation of the textural matrices from which the textural features derive.

Gaussian filtering was used to reduce image noise and had a slight but consistently positive effect on features stability, especially on textural features. The N4ITK algorithm used for bias-field correction also had a positive effect on both FOS and textural features. This result highlights how important it is to handle those type of noise prior to any image analysis.

Evaluation of features stability to ROI-related variability

Chapter 4 addresses the issue of stability of radiomic features to variations in the ROI. In particular two different sources of variability in the ROI were considered: 1) multiple manual segmentations (by two different radiologist); 2) geometrical transformation of the ROI. The two analyses were performed on a dataset of patients affected by either HNC and STS. For the analyses T1w, T2w and ADC images were considered. The stability analyses were performed for HNC and STS separately.

Features stability to variations in the ROI was dependent on the district considered (HNC or STS). Features extracted from STS were more stable compared to features extracted from HNC. This result may depend from the fact that the volume of STS tumor is much larger compared to the one

of HNC, meaning that a smaller portion of the ROI is modified, leading to less variability in the features and higher ICC values.

Image type was also a factor influencing features stability, with T1w images showing lower ICC values compared to T2w images and ADC maps. This may depend on the lower inter-tumor variability in T1w images. Since it is reasonable to expect the same amount of ROI-related variability in all the images, and since the values of ICC are proportional to the inter-tumor variability, it is understandable that lower ICC may be observed.

From the analysis in Chapter 4 it was possible to see that there is a high correlation, in terms of ICC values, between the results of multiple segmentations and ROI geometrical transformations, with the latter being more restrictive than the former. This is a very important finding that suggests a potential use of ROI geometrical transformation as a possible surrogate for the test of multiple segmentations. Stability to multiple segmentation is currently one of the most widely used method to asses the stability of radiomic features, but is time consuming since requires the work of multiple radiologists. On the other hand, ROI translations can be easily automatized and implemented within a radiomic workflow, making the assessment of stability to ROI-variations easier.

Selection of a set of stable radiomic features

After the analyses of Chapters 3 and 4 it was also possible to classify the radiomic features in stable and unstable using an ICC of 0.75 as threshold.

For the imaging-related variability, only T1w and T2w images were considered and a set 536 features of different categories (FOS, SS, textural and wavelet) was extracted for each image type, for a total of 1072 features. In total, 550 features were stable to variations in image acquisition parameters (266 T1w and 284 T2w). SS features were the most stable and wavelet features were the least stable. The stability of the features was mainly independent from the image type with 229 images being stable for both T1w and T2w images. This common features set was used as a surrogate of ADC stable features, since no analysis on ADC could be performed in Brainweb.

Of the 1608 features considered (536 features for T1w, T2w and ADC), 701 and 1057 were stable to ROI-variability for HNC and STS respectively. SS were the most stable features and wavelets were the least stable. T1w images presented a significantly lower number of stable features compared to T2w and ADC images.

A total of 410 and 617 features were stable to both imaging-related and ROI-related variability in HNC and STS respectively. These features were the ones used for all the features analyses performed in the thesis.

Optimization of the radiomic postprocessing

Chapter 5 focused on the features postprocessing, i.e. the steps that are performed after the features extraction. In particular, features normalization and features selection were optimized in order to maximize the performance of a Cox proportional hazard regression model for OS in HNC. Combinations of four different features normalization methods and two different feature selection pipeline were compared.

According to a 2-way ANOVA for repeated measures, both normalization algorithm and features selection pipeline had a significant effect on the prognostic performance of the survival model, measured by the Harrel's C-index. Moreover, a significant interaction between the two factor was identified meaning that is not possible to select one step of post-processing without taking into account all the others.

The best postprocessing pipeline included Z-score normalization and three feature selections steps based on features pairwise correlation and prognostic performance, assessed using both univariate and multivariate Cox regression models.

Training and validation of radiomic-based survival models

In Chapters 6 and 7, all the workflow designed in the first part of the thesis was applied to two different datasets containing images acquired from multiple centers and/or with non-standardized protocols in order to develop prognostic models for OS for HNC and STS. For each problem, a prognostic radiomic signature was trained and a validation (either internal or external) was performed to asses the ability of the radiomic features to be prognostic for OS for unseen patients.

In the case of HNC, a five-features radiomic signature was trained on 262 patients. The signature performed well (Harrel's C-index>0.6) on both internal cross-validation (C-index 0.67) and external validation of 232 patients (C-index 0.63). The Kaplan-Meier curves of the high and low risk groups as defined by radiomics were also significantly different in both types of validation.

For STS, a 1-feature signature was trained on a dataset of 91 patients. The cross-validation showed that the signature had a good prognostic performance for OS (C-index 0.74).

The results of both Chapter 6 and 7 showed that the developed radiomic signature had a prognostic performance that was similar to the one of signatures trained on standardized datasets. This seems to prove that, with the adequate pre- and post-processing, radiomics could be build prognostic

model for OS. It is also reasonable to think that a similar workflow could be used to train prognostic models for other types of outcome (e.g. disease-free survival).

Added prognostic value of radiomic signatures

Since the identified radiomic signature are correlated to other prognostic clinical variables, different analyses were performed to ensure that the signatures included significant and independent prognostic information.

In HNC, multivariate Cox analysis confirmed that the radiomic signature was a significantly prognostic factor and that was independent from other variables like stage TNM, HPV status or tumor subsite. Similar results was found for STS, for which the multivariate Cox model was built using radiomic signature and tumor grade as prognostic factors.

In HNC, multiple Kaplan-Meier analyses were performed in the different subgroups stratified by clinical variables such as stage TNM, HPV status and sub-site. The risk groups classified by radiomics corresponded to significantly different curves for Stage IV patients and for HPV- patients. In the other subgroups the differences were not significant but this could be related to the low number of events. Similar results were found for the STS dataset in which the radiomic signature could significantly separate Kaplan-Meier curves in Grade 3 patients, while no significant differences were found for grade 1-2. The two previous results indicate that radiomics is particularly effective on more aggressive tumors.

For both STS and HNC, prognostic models trained using both clinical and radiomic variables led to improved performance compared (higher C-index) to the models obtained using the clinical variables alone (STS: 0.78 vs 0.74; HNC-cross-validation: 0.69 vs 0.67; HNC-validation: 0.72 vs 0.69).

All the previous results highlight how radiomics could provide additional prognostic information that is not obtainable by a traditional clinical approach, even though the entity of this added value may vary depending on the characteristics of the tumor (e.g. stage, grade sub-site).

8.2 Impact, limitations and future developments

Each of the analysis performed in the thesis had its own limitation that could be the starting points for future analyses.

The stability to variations in the image acquisition parameters were performed on simulated MRI using virtual phantoms (BrainWeb). On one hand, virtual simulations are a powerful tool to simulate a large number of

image acquisitions in a way that would not be feasible in the real world. On the other hand, a few limitations are also present: Brainweb does not allow to simulate some types of variability, such as test-retest or scanner-related variability; also, DWI cannot be simulated using BrainWeb and therefore no direct stability analysis could be performed for those type of images; last, since there is no clear tumor in BrainWeb images, the selection of the ROI was done somewhat arbitrarily. To address the aforementioned limitations, simulations on real phantom with specific ROI could be performed, analyzing different types of imaging-related variability (imaging parameters variations; inter-scanner variability; test-retest) on all the image sequences of interest.

In the thesis, stability was used as a criteria for preliminary features selection, using an ICC>0.75 for both imaging-related and ROI-related variability as a requirement. Although this preliminary selection tends to improve the performance of the downstream classification/prognostic analysis, the removal of the features may cause a loss of information leading to a sub-optimal model [115]. An alternative approach to overcome this issue could be to use the stability analysis to estimate the generated variability (standard deviation) of each radiomic features and use this information to perform data augmentation of the clinical datasets used to train the prognostic models. Recent studies show the potential improvement that could be provided by such approach [115]. Future analysis may test the robustness of this kind of approach.

The textural features considered for most of the analyses were extracted from specific textural matrices (GLCM and GLRLM). This was done since GLCM and GLRLM are the ones that are used the most in radiomics studies and because they are available in any software for radiomic features extraction. However, all the methodologies applied in this thesis (from stability analysis of the features to the development of prognostic models) could also involve other textural matrices (such as GLCM, GLSZM and NGTDM). Also, other image transform, such as Laplacian of Gaussian or Gabor filtering could be consider to provide a more detailed description of the ROI.

In the thesis a preliminary optimization was performed for feature selection and normalization methods. It was found that there was an interaction effect between the features selection pipeline and the features normalization. In the thesis 4 features normalization algorithm and 2 features selection pipeline were analyzed for a total of 8 combinations. The analysis performed in the thesis provide some information to guide the design of features post-processing but the analysis was not exhaustive. The same can be

said for the survival model, since alternative to the traditional Cox proportional hazard regression used in this thesis (such as LASSO-Cox regression or survival random forest) could be considered in future studies. Another possible future developments is to automatize the optimization of features postprocessing. Some frameworks to perform such task already exist, like the Workflow for Optimized Radiomic Classification (WORC) [149], but they are focused on classification. Unfortunately, to the knowledge of the authors no such framework was developed for survival analysis.

Last, performing more and more external validations for the developed radiomic signatures is a fundamental next steps in further ensuring their prognostic performance in a more generalized context such as the clinical practice. Also, by the results of Chapters 6 and 7, it was possible to see that the signature has more prognostic power on specific subgroups (e.g. patients with high stage and grade). In order to enhance the prognostic power of radiomics also for the other categories of patients, new data could be collected and new, more-specific signatures could be created with a workflow like the one developed in this thesis.

Radiomics may potentially have a huge impact in the clinical practice because it could provide a non-invasive tool to provide additional information that may help tumor prognosis, especially in those subgroups were no clinical prognostic variables are available. Another possible application, not treated in this thesis, could be to use radiomics to provide clinical information, such as determining the grade of a tumor or its HPV status, in a non-invasive and cheaper way, with huge advantage for both the patients (more comfort) and the healthcare system (lower costs). However, the application of radiomics to the clinical practice is still a far goal. To achieve such goal, it is necessary to test the consistency of results such as the one presented in the thesis for more and more independent validation sets, and gradually increasing the variability in terms of both imaging protocols (more and more centers and scanners should be considered) and biological (early stage tumors as well as advanced stage tumors). In case the results will not hold, the new validation set will be included in the train set and new models will be iteratively updated and improved. Also, in this thesis only STS and HNC were considered but other types of rare cancers exists (e.g. bone tumors) and radiomic analyses for those cancers should be performed as well.

Despite the results of this thesis were far from exhaustive. The results shown were definitely a further step towards a successful application of radiomics in the clinical practice, that will result in a better and cheaper managment of cancer.

Bibliography

[1] Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England journal of medicine*, 372(9):793–795, 2015.

[2] Ewa Gubb and Rune Matthiesen. Introduction to omics. In *Bioinformatics Methods in Clinical Research*. Humana Press, 2010.

[3] Freddie Bray, Jaques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[4] Loris De Cecco, Paolo Bossi, Laura Locati, Silvana Canevari, and Lisa Licitra. Comprehensive gene expression meta-analysis of head and neck squamous cell carcinoma microarray data defines a robust survival predictor. *Annals of Oncology*, 25(8):1628–1635, 2014.

[5] Hugo J W L Aerts, Emmanuel Rios-Velazquez, Ralph T H Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5:4006, 2014.

[6] Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A Eschrich, Matthew B Schabath, Kenneth Forster, Hugo J W L Aerts, Andre Dekker, David Fenstermacher, Dmitry B Goldgof, Lawrence O Hall, Philippe Lambin, Yoganand Balagurunathan, Robert A Gatenby, and Robert J Gillies. Radiomics: the process and the challenges. *Magnetic resonance imaging*, 30(9):1234–1248, 2012.

[7] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud G P M van Stiphout, Patrick Granton, Catharina M.L. Zegers, Robert Gillies, Ronald Boellard, Andre Dekker, and Hugo J W L Aerts. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*, 48(4):441–446, 2012.

[8] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. Radiomics: Images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2015.

Bibliography

[9] Zhenyu Liu, Shuo Wang, Di Dong, Jingwei Wei, Cheng Fang, Xuezhi Zhou, Kai Sun, Longfei Li, Bo Li, Meiyun Wang, and Jie Tian. The applications of radiomics in precision diagnosis and treatment of oncology: Opportunities and challenges. *Theranostics*, 9(5):1303–1322, 2019.

[10] Marta Bogowicz, Oliver Riesterer, Luisa Sabrina Stark, Gabriela Studer, Jan Unkelbach, Matthias Guckenberger, and Stephanie Tanadini-Lang. Comparison of PET and CT radiomics for prediction of local tumor control in head and neck squamous cell carcinoma. *Acta Oncologica*, 56(11):1531–1536, 2017.

[11] Bin Zhang, Jie Tian, Di Dong, Dongsheng Gu, Yuhao Dong, Lu Zhang, Zhouyang Lian, Jing Liu, Xiaoning Luo, Shufang Pei, Xiaokai Mo, Wenhui Huang, Fusheng Ouyang, Baoliang Guo, Long Liang, Wenbo Chen, Changhong Liang, and Shuixing Zhang. Radiomics features of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma. *Clinical Cancer Research*, 23(11):4259–4269, 2017.

[12] Eran Segal, Claude B Sirlin, Clara Ooi, Adam S Adler, Jeremy Gollub, Xin Chen, Bryan K Chan, George R Matcuk, Christopher T Barry, Howard Y Chang, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nature biotechnology*, 25(6):675–680, 2007.

[13] Maximilian Diehn, Christine Nardini, David S Wang, Susan McGovern, Mahesh Jayaraman, Yu Liang, Kenneth Aldape, Soonmee Cha, and Michael D Kuo. Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *Proceedings of the National Academy of Sciences*, 105(13):5213–5218, 2008.

[14] Ruben T H M Larue, Gilles Defranes, Dirk De Ruysscher, Philippe Lambin, and Wouter van Elmpt. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. *British journal of radiology*, 90(1070):20160665, 2017.

[15] Chintan Parmar, Emmanuel Rios-Velazquez, Ralph Leijenaar, Mohammed Jermoumi, Sara Carvalho, Raymond H Mak, Sushmita Mitra, Uma B Shankar, Ron Kikinis, Benjamin Haibe-Kains, Philippe Lambin, and Hugo J W L Aerts. Robust radiomics feature quantification using semiautomatic volumetric segmentation. *Plos One*, 9(7):e102107, 2014.

[16] Ruben T H M Larue, Janna E van Timmeren, Evelyn E C de Jong, Giacomo Feliciani, Ralph T H Leijenaar, Wendy M J Schreurs, Meindert N Sosef, Frank H P J Raat, Frans H R van der Zande, Marco Das, Wouter van Elmpt, and Philippe Lambin. Influence of gray level discretization on radiomic feature stability for different ct scanners, tube currents and slice thicknesses: a comprehensive phantom study. *Acta oncologica*, 56 (11):1544–1553, 2017.

[17] Loïc Duron, Daniel Balvay, Saskia Vande Perre, Afef Bouchouicha, Julien Savatovsky, Jean-Claude Sadik, Isabelle Thomassin-Naggara, Laure Fournier, and Augustin Lecler. Gray-level discretization impacts reproducible mri radiomics texture features. *PloS one*, 14(3)(3), 2019.

[18] Elaine Limkin, Roger Sun, Laurent Dercle, Evangelia Zacharaki, Charlotte Robert, Sylvain ReuzÃ©, Nikos Schernberg, Antoine Paragios, Eric Deutsch, and Charles FertÃ©. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. *Annals of Oncology*, 28(6):1191–1206, 2017.

[19] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, 6:610–621, 1973.

[20] Xiaoou Tang. Texture information in run-length matrices. *IEEE transactions on image processing*, 7(11):1602–1609, 1998.

[21] Guillaume Thibault, Jesus Angulo, and Fernand Meyer. Advanced statistical matrices for texture characterization: application to cell classification. *IEEE Transactions on Biomedical Engineering*, 61(3):630–637, 2013.

[22] Chengjun Sun and William G Wee. Neighboring gray level dependence matrix for texture classification. *Sarcoma*, 2019:1–18, 2019.

[23] Moses Amadasun and Robert King. Textural features corresponding to textural properties. *IEEE Transactions on systems, man, and Cybernetics*, 19(5):1264–1273, 1989.

[24] Rafael C Gonzalez and Richard E Woods. *Digital image processing*. Prentice hall Upper Saddle River, 2002.

[25] Jon C Aster, Abul K Abbas, and Vinay Kumar. *Robbins Basic Pathology, 9th edition*. Elsevier Saunders, 2012.

[26] National cancer institute (NCI). What is cancer? `https://www.cancer.gov/about-cancer/understanding/what-is-cancer`.

[27] Lisa Fayed. Differences between a malignant and benign tumor. `https://www.verywellhealth.com/what-does-malignant-and-benign-mean-514240`.

[28] World Health Organization (WHO). Cancer. `https://www.who.int/news-room/fact-sheets/detailcancer`.

[29] World Health Organization (WHO). Italy country profile: Italy. `https://www.who.int/cancer/country-profiles/ita_en.pdf?ua=1`.

[30] Associazione Italiana di Oncologia Medica (AIOM). L'incidenza dei tumori in italia. `http://www.registri-tumori.it/PDF/AIOM2016allegato1015426.pdf`.

[31] National cancer institute (NCI). Head and neck cancers. `https://www.cancer.gov/about-cancer/understanding/what-is-cancer`.

[32] Kerstin M Stenson, Bruce E Brockstein, and Sonali Shah. Epidemiology and risk factors for head and neck cancer. `https://www.uptodate.com/contents/epidemiology-and-risk-factors-for-head-and-neck-cancer`.

[33] World Heatlh Organization (WHO). *World Cancer Report*. The International Agency for Research on Cancer, 2014.

[34] Anil K Chaturvedi, Eric A Engels, Ruth M Pfeiffer, Brenda Y Hernandez, Weihong Xiao, Esther Kim, Bo Jiang, Marc T Goodman, Maria Sibug-Saber, Wendy Cozen, Lihua Liu, Charles F Lynch, Nicolas Wentzensen, Richard C Jordan, Sean Altekruse, William F Anderson, Philip S Rosenberg, and Maura L Gillison. Human papillomavirus and rising oropharyngeal cancer incidence in the united states. *Journal of Clinical Oncology*, 29(32):4294–4301, 2011.

[35] Manuela Quaresma, Michel P Coleman, and Bernard Rachet. 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in england and wales, 1971-2011: a population-based study. *The Lancet*, 385(9974):1206–1208, 2015.

[36] Sian Taylor. The psychological and psychosocial effects of head and neck cancer. *Cancer Nursing Practice*, 15(9), 2016.

[37] Domagoj Ante Vodanovich and Peter F M Choong. Soft-tissue sarcomas. *Indian journal of orthopedics*, 51(1):35–44, 2018.

[38] Vittoria Colia, Paolo Casali, Silvia Stachiotti, and Salvatore Provenzano. Soft tissue sarcoma: a guide for patients.

[39] Nicolas Penel, Jessica Grosjean, Yves Marie Robin, Luc Vanseymortier, Stephanie Clisant, and Antoine Adenis. Frequency of certain established risk factors in soft tissue sarcomas in adults: A prospective descriptive study of 658 cases. *Sarcoma*, 2008:2–6, 2008.

Bibliography

[40] Lesley Storey, Lorna Fern, Ana Martins, Mary Wells, Lindsey Bennister, Craig Gerrand, Maria Onasanya, Jeremy S Whelan, Rachael Windsor, Julie Woodford, and Rachel M Taylor. A critical review of the impact of sarcoma on psychosocial wellbeing. *Sarcoma*, 2019:1–18, 2019.

[41] American Joint Committee on Cancer. *AJCC Cancer Staging Manual*. Springer, 2010.

[42] National cancer institute (NCI). Cancer staging. `https://www.cancer.gov/about-cancer/diagnosis-staging/staging`.

[43] Christina Neppl, Manuel D Keller, Amina Scherz, Patrick Dorn, Ralph A Schmid, Inti Zlobec, and Sabina Berezowska. Comparison of the 7th and 8th edition of the UICC/AJCC TNM staging system in primary resected squamous cell carcinomas of the lung–a single center analysis of 354 cases. *Frontiers in medicine*, 6:196, 2019.

[44] William Lydiatt, Brian OâSullivan, and Snehal Patel. Major changes in head and neck staging for 2018. *American Society of Clinical Oncology Educational Book*, 38:505–514, 2018.

[45] National cancer institute (NCI). Types of cancer treatment. `https://www.cancer.gov/about-cancer/treatment/types`.

[46] National cancer institute (NCI). Combination treatments. `https://training.seer.cancer.gov/treatment/combination`.

[47] William R Hendee and Russell E Ritenour. *Medical imaging physics: fourth edition*. John Wiley and Sons, 2003.

[48] Nadine Barrie Smith and Andrew Webb. *Introduction to medical imaging: physics, engineering and clinical applications*. Cambridge university press, 2010.

[49] Stephanie N Histed, Maria L Lindenberg, Esther Mena, Baris Turkbey, Peter L Choyke, and Karen A Kurdziel. Review of functional/anatomic imaging in oncology. *Nuclear medicine communications*, 33(4):349, 2012.

[50] Martin J Willemink and Peter B NoÃ«l. The evolution of image reconstruction for ct. from filtered back projection to artificial intelligence. *European radiology*, 29(5):2185–2195, 2019.

[51] Aren van Waarde. Introduction on pet: Description of basics and principles. *Trends On The Role Of Pet In Drug Development*, pages 1–13, 2012.

[52] Paul E Kinahan and James W Fletcher. Pet/ct standardized uptake values (suvs) in clinical practice and assessing response to therapy. *Seminaries in Ultrasound, CT and MR*, 31(6):496–505, 2010.

[53] Frederic H Fahey. Data acquisition in pet imaging. *Journal of nuclear medicine technology*, 30(2):39–49, 2002.

[54] Sven Plein, John P Greenwood, and John P Ridgway. Generating a signal: RF pulses and echoes. In Sven Plein, John P Greenwood, and John P Ridgway, editors, *Cardiovascular MR manual*. Springer, 2010.

[55] MRI questions and answers. Size of T1 vs T2. `http://www.mri-q.com/why-is-t1--t2.html`.

[56] Govind B Chavhan, Paul S Babyn, Bejoy Thomas, Manohar M Shroff, and E Mark Haacke. Principles, techniques, and applications of t2*-based mr imaging and its special applications. *Radiographics*, 29(5):1433–1449, 2009.

[57] Richard Bitar, General Leung, Richard Perng, Sameh Tadros, Alan R Moody, Josee Sarrazin, Caitlin McGregor, Monique Christakis, Sean Symons, Andrew Nelson, et al. MR pulse sequences: what every radiologist wants to know but is afraid to ask. *Radiographics*, 26 (2):513–537, 2006.

[58] Renato Toffanin, Giuseppe Guglielmi, and Maria A Cova. Fast MRI methods for the clinical evaluation of skeletal disorders. *Medical Imaging*, page 239, 2011.

[59] Wolfgang R Nitz and P Reimer. Contrast mechanisms in MR imaging. *European radiology*, 9(6):1032–1046, 1999.

[60] Dow-Mu Koh and David J Collins. Diffusion-weighted MRI in the body: applications and challenges in oncology. *American Journal of Roentgenology*, 188(6):1622–1635, 2007.

[61] Edward O Stejskal and John E Tanner. Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. *The journal of chemical physics*, 42(1):288–292, 1965.

[62] Xiao-Hua Ye, Jia-Yin Gao, Zheng-Han Yang, and Yuan Liu. Apparent diffusion coefficient reproducibility of the pancreas measured at different mr scanners using diffusion-weighted imaging. *Journal of Magnetic Resonance Imaging*, 40 (6):1375–1381, 2014.

[63] Giacomo Belli, Simone Busoni, Antonio Ciccarone, Angela Coniglio, Marco Esposito, Marco Giannelli, Lorenzo N Mazzoni, Luca Nocetti, Roberto Sghedoni, Roberto Tarducci, et al. Quality assurance multicenter comparison of different MR scanners for quantitative diffusion-weighted imaging. *Journal of Magnetic Resonance Imaging*, 43(1):213–219, 2016.

[64] Olaf Dietrich, Andreas Biffar, Andrea Baur-Melnyk, and Maximilian F Reiser. Technical aspects of MR diffusion imaging of the body. *European journal of radiology*, 76(3):314–322, 2010.

[65] XRay Physics. Mri physics: Pulse sequences. `http://xrayphysics.com/sequences.html`.

[66] Pyradiomics community. Pyradiomics documentation, version 2.1.0. `https://pyradiomics.readthedocs.io/en/2.1.0/`.

[67] Alex Zwaneburg, Stefan Leger, Martin Vallieres, and Steffen Lock. Image biomarker standardisation initiative: reference manual. *Arxive Preprint*, arXiv 1612.07003, 2016.

[68] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[69] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[70] Mark Stevenson, Simon Firestore, Anke Wiethoelter, and Caitlin Pfeiffer. *An introduction to survival analysis*. EpiCentre, IVABS, Massey University, 2009.

[71] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53 (282):457–481, 1958.

[72] John P Klein, Hans C Van Houwelingen, Joseph G Ibrahim, and Thomas H Scheike. *Handbook of survival analysis*. CRC Press Boca Raton, FL, 2014.

[73] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[74] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[75] L Latha and S Thangasamy. Efficient approach to normalization of multimodal biometric scores. *International Journal of Computer Applications*, 32(10):57–64, 2011.

[76] Anastasia Chalkidou, Michael J OâDoherty, and Paul K Marsden. False discovery rates in PET and CT studies with texture features: a systematic review. *PloS one*, 10(5):1–18, 2015.

[77] Chintan Parmar, Patrick Grossmann, Johan Bussink, Philippe Lambin, and Hugo JWL Aerts. Machine learning methods for quantitative radiomic biomarkers. *Scientific reports*, 5:13087, 2015.

Bibliography

[78] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[79] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Journal of american medical association*, 247(18):2543–2546, 1982.

[80] Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

[81] Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50 (3):163–170, 1966.

[82] John Ford, Nesrin Dogan, Lori Young, and Fei Yang. Quantitative radiomics: Impact of pulse sequence parameter selection on mri-based textural features of the brain. *Contrast media molecular imaging*, 2018:1–9, 2018.

[83] Xenia Fave, Lifei Zhang, Jinzhong Yang, Dennis Mackin, Peter Balter, Daniel Gomez, David Followill, Kyle A Jones, Francesco Stingo, and Laurence E Court. Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer. *Translational Cancer Research*, 5(4):349–363, 2016.

[84] Dennis Mackin, Xenia Fave, Lifei Zhang, Jinzhong Yang, Kyle A Jones, Chaan S Ng, and Laurence Court. Harmonizing the pixel size in retrospective computed tomography radiomics studies. *PloS one*, 12(9):e0178524, 2017.

[85] Muhammad Shafiq-ul-Hassan, Geoffrey G Zhang, Kujtim Latifi, Ghanim Ullah, Dylan C Hunt, Yoganand Balagurunathan, Mahmoud Abrahem Abdalah, Matthew B Schabath, Dmitry G Goldgof, Dennis Mackin, Laurence Edward Court, Robert James Gillies, and Eduardo Gerardo Moros. Intrinsic dependencies of ct radiomic features on voxel size and number of gray levels. *Medical physics*, 44(3):1050–1062, 2017.

[86] Dennis Mackin, Xenia Fave, Lifei Zhang, David Fried, Jinzhong Yang, Brian Taylor, Edgardo Rodriguez-Rivera, Cristina Dodge, Kyle A Jones, and Laurence Court. Measuring ct scanner variability of radiomics features. *Investigative radiology*, 50(11):757, 2015.

[87] Petros Kalendralis, Alberto Traverso, Zhenwei Shi, Ivan Zhovannik, René Monshouwer, Martijn P A Starmans, Stefan Klein, Elisabeth Pfaehler, Ronald Boellaard, Andre Dekker, and Leonard Wee. Multicenter ct phantoms public dataset for radiomics reproducibility tests. *Medical physics*, 46(3):1512–1518, 2019.

[88] D L Collins, A P Zijdenbos, V Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans. Design and construction of a realistic digital brain phantom. *IEEE transactions on medical imaging*, 17 (3):463–468, 1998.

[89] Attila Forgacs, Hermann Pall Jonsson, Magnus Dahlbom, Freddie Daver, Matthew D DiFranco, Gabor Opposits, Aron K Krizsan, Ildiko Garai, Johannes Czernin, Jozsef Varga, Lajos Tron, and Laszlo Balkay. A study on the basic criteria for selecting heterogeneity parameters of f18-fdg pet images. *PLoS One*, 11(10):e0164113, 2016.

[90] Elisabeth Pfaehler, Roelof J Beukinga, Johan R de Jong, Riemer H J A Slart, Cornelis H Slump, Rudi A J O Dierckx, and Ronald Boellaard. Repeatability of ^{18}F-FDG PET radiomic features: A phantom study to explore sensitivity to image reconstruction settings, noise, and delineation method. *Medical physics*, 46(2):665–678, 2019.

[91] Fei Yang, Nesrin Dogan, Radka Stoyanova, and John Chetley Ford. Evaluation of radiomic texture feature error due to mri acquisition and reconstruction: a simulation study utilizing ground truth. *Physica Medica*, 50:26–36, 2018.

[92] Marco Bologna, Valentina D A Corino, and Luca T Mainardi. Assessment of the effect of intensity standardization on the reliability of T1-weighted mri radiomic features: experiment on a virtual phantom. In *Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, July 2019.

[93] Marco Bologna, Valentina D A Corino, and Luca T Mainardi. Technical note: Virtual phantom analyses for preprocessing evaluation and detection of a robust feature set for mri-radiomics of the brain. *Medical physics*, 46 (11):5116–5123, 2019.

[94] Remi K-S Kwan, Alan C Evans, and G Bruce Pike. An extensible mri simulator for post-processing evaluation. In *International Conference on Visualization in Biomedical Computing*, pages 135–140. Springer, 1996.

[95] Osama Moh'd Alia, Rajeswari Mandava, and Mohd Ezane Aziz. A hybrid harmony search algorithm for mri brain segmentation. *Evolutionary Intelligence*, 4(1):31–49, 2011.

[96] Marc Modat, David M Cash, Pankaj Daga, Gavin P Winston, John S Duncan, and Sébastien Ourselin. Global image registration using a symmetric block-matching approach. *Journal of Medical Imaging*, 1(2):024003, 2014.

[97] Steve Pieper, Michael Halle, and Ron Kikinis. 3D slicer. In *2004 2nd IEEE international symposium on biomedical imaging: nano to macro (IEEE Cat No. 04EX821)*, pages 632–635. IEEE, 2004.

[98] McGill university. Brainweb: Simulated brain database. `https://brainweb.bic.mni.mcgill.ca/`.

[99] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

[100] Valentina DA Corino, Eros Montin, Antonella Messina, Paolo G Casali, Alessandro Gronchi, Alfonso Marchianò, and Luca T Mainardi. Radiomic analysis of soft tissues sarcomas can distinguish intermediate from high-grade lesions. *Journal of Magnetic Resonance Imaging*, 47(3):829–840, 2018.

[101] Kenneth O McGraw and Seok P Wong. Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1):30, 1996.

[102] Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

[103] Guangyi Wang, Lan He, Cai Yuan, Yanqi Huang, Zaiyi Liu, and Changhong Liang. Pretreatment MR imaging radiomics signatures for response prediction to induction chemotherapy in patients with nasopharyngeal carcinoma. *European journal of radiology*, 98:100–106, 2018.

[104] Jiangwei Lao, Yinsheng Chen, Zhi-Cheng Li, Qihua Li, Ji Zhang, Jing Liu, and Guangtao Zhai. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Scientific reports*, 7(1):10353, 2017.

[105] Benjamin M Ellingson, Taryar Zaw, Timothy F Cloughesy, Kourosh M Naeini, Shadi Lalezari, Sandy Mong, Albert Lai, Phioanh L Nghiemphu, and Whitney B Pope. Comparison between intensity normalization techniques for dynamic susceptibility contrast (DSC)-MRI estimates of cerebral blood volume (CBV) in human gliomas. *Journal of Magnetic Resonance Imaging*, 35(6):1472–1477, 2012.

[106] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310, 2010.

Bibliography

[107] John G Sled, Alex P Zijdenbos, and Alan C Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE transactions on medical imaging*, 17(1):87–97, 1998.

[108] Terry S Yoo, Michael J Ackerman, William E Lorensen, Will Schroeder, Vikram Chalana, Stephen Aylward, Dimitris Metaxas, and Ross Whitaker. Engineering and algorithm design for an image processing api: a technical report on itk-the insight toolkit. *Studies in health technology and informatics*, pages 586–592, 2002.

[109] Binsheng Zhao, Yongqiang Tan, Wei Yann Tsai, Lawrence H Schwartz, and Lin Lu. Exploring variability in ct characterization of tumors: a preliminary phantom study. *Translational oncology*, 7(1):88, 2014.

[110] Lan He, Yanqi Huang, Zelan Ma, Cuishan Liang, Changhong Liang, and Zaiyi Liu. Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule. *Scientific reports*, 6:34921, 2016.

[111] Matthew J Nyflot, Fei Yang, Darrin Byrd, Stephen R Bowen, George A Sandison, and Paul E Kinahan. Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards. *Journal of medical imaging*, 2(4):041002, 2015.

[112] Bino A Varghese, Darryl Hwang, Steven Y Cen, Joshua Levy, Derek Liu, Christopher Lau, Marielena Rivas, Bhushan Desai, David J Goodenough, and Vinay A Duddalwar. Reliability of ct-based texture features: Phantom study. *Journal of Applied Clinical Medical Physics*, 2019.

[113] Sandra Fiset, Mattea L Welch, Jessica Weiss, Melania Pintilie, Jessica L Conway, Michael Milosevic, Anthony Fyles, Alberto Traverso, David Jaffray, Ur Metser, et al. Repeatability and reproducibility of mri-based radiomic features in cervical cancer. *Radiotherapy and Oncology*, 135:107–114, 2019.

[114] Sylvain Reuzé, Fanny Orlhac, Cyrus Chargari, Christophe Nioche, Elaine Limkin, François Riet, Alexandre Escande, Christine Haie-Meder, Laurent Dercle, Sébastien Gouy, et al. Prediction of cervical cancer recurrence using textural features extracted from 18f-fdg pet images acquired with different scanners. *Oncotarget*, 8(26)(26):43169, 2017.

[115] Michael Götz and Klaus H Maier-Hein. Optimal statistical incorporation of independent feature stability information into radiomics studies. *Scientific reports*, 10(1):1–10, 2020.

[116] Ralph TH Leijenaar, Sara Carvalho, Emmanuel Rios Velazquez, Wouter JC Van Elmpt, Chintan Parmar, Otto S Hoekstra, Corneline J Hoekstra, Ronald Boellaard, André LAJ Dekker, Robert J Gillies, et al. Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta oncologica*, 52(7):1391–1397, 2013.

[117] Jan C Peeken, Michael Bernhofer, Matthew B Spraker, Daniela Pfeiffer, Michal Devecka, Ahmed Thamer, Mohammed A Shouman, Armin Ott, Fridtjof Nüsslin, Nina A Mayr, et al. CT-based radiomic features predict tumor grading and have prognostic value in patients with soft tissue sarcomas treated with neoadjuvant radiation therapy. *Radiotherapy and Oncology*, 135:187–196, 2019.

[118] Olivier Gevaert, Lex A Mitchell, Achal S Achrol, Jiajing Xu, Sebastian Echegaray, Gary K Steinberg, Samuel H Cheshier, Sandy Napel, Greg Zaharchuk, and Sylvia K Plevritis. Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features. *Radiology*, 273 (1):168–174, 2014.

[119] Marco Bologna, Eros Montin, Valentina DA Corino, and Luca T Mainardi. Stability assessment of first order statistics features computed on ADC maps in soft-tissue sarcoma. In *Proceedings of the 39^{th} Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, July 2017.

[120] Marco Bologna, Valentina DA Corino, Eros Montin, Antonella Messina, Giuseppina Calareso, Francesca G Greco, Silvana Sdao, and Luca T Mainardi. Assessment of stability and discrimination capacity of radiomic features on apparent diffusion coefficient images. *Journal of digital imaging*, 31(6):879–894, 2018.

[121] Alex Zwanenburg, Stefan Leger, Linda Agolli, Karoline Pilz, Esther GC Troost, Christian Richter, and Steffen Löck. Assessing robustness of radiomic features by image perturbation. *Scientific reports*, 9(1):614, 2019.

[122] Ralph TH Leijenaar, Marta Bogowicz, Arthur Jochems, Frank JP Hoebers, Frederik WR Wesseling, Sophie H Huang, Biu Chan, John N Waldron, Brian O'Sullivan, Derek Rietveld, et al. Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: a multicenter study. *The British journal of radiology*, 91(1086):20170498, 2018.

[123] Azim Celik. Effect of imaging parameters on the accuracy of apparent diffusion coefficient and optimization strategies. *Diagnostic and Interventional Radiology*, 22(1):101, 2016.

[124] Janna E van Timmeren, Ralph TH Leijenaar, Wouter van Elmpt, Jiazhou Wang, Zhen Zhang, André Dekker, and Philippe Lambin. Test-retest data for radiomics feature stability analysis: generalizable or study-specific? *Tomography*, 2(4):361, 2016.

[125] Bettina Baeßler, Kilian Weiss, and Daniel Pinto dos Santos. Robustness and reproducibility of radiomics in magnetic resonance imaging: a phantom study. *Investigative radiology*, 54(4)(4):221–228, 2019.

[126] Yucheng Zhang, Anastasia Oikonomou, Alexander Wong, Masoom A Haider, and Farzad Khalvati. Radiomics-based prognosis analysis for non-small cell lung cancer. *Scientific reports*, 7:46349, 2017.

[127] Ji Eun Park, Seo Young Park, Hwa Jung Kim, and Ho Sung Kim. Reproducibility and generalizability in radiomics modeling: possible strategies in radiologic and statistical perspectives. *Korean journal of radiology*, 20(7):1124–1137, 2019.

[128] Chintan Parmar, Ralph TH Leijenaar, Patrick Grossmann, Emmanuel Rios Velazquez, Johan Bussink, Derek Rietveld, Michelle M Rietbergen, Benjamin Haibe-Kains, Philippe Lambin, and Hugo JWL Aerts. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. *Scientific reports*, 5:11044, 2015.

[129] Philipp Kickingereder, Sina Burth, Antje Wick, Michael Götz, Oliver Eidel, Heinz-Peter Schlemmer, Klaus H Maier-Hein, Wolfgang Wick, Martin Bendszus, Alexander Radbruch, et al. Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. *Radiology*, 280(3):880–889, 2016.

[130] BD2Decide consortium. BD2Decide: Big data and models for personalized head and neck cancer decision support. `http://www.bd2decide.eu/`.

[131] Fanny Orlhac, Sarah Boughdad, Cathy Philippe, Hugo Stalla-Bourdillon, Christophe Nioche, Laurence Champion, Michaël Soussan, Frédérique Frouin, Vincent Frouin, and Irène Buvat. A postreconstruction harmonization method for multicenter radiomic studies in pet. *Journal of Nuclear Medicine*, 59(8):1321–1328, 2018.

[132] Fanny Orlhac, Frédérique Frouin, Christophe Nioche, Nicholas Ayache, and Irène Buvat. Validation of a method to compensate multicenter effects affecting ct radiomics. *Radiology*, 291(1):53–59, 2019.

[133] François Lucia, Dimitris Visvikis, Martin Vallières, Marie-Charlotte Desseroit, Omar Miranda, Philippe Robin, Pietro Andrea Bonaffini, Joanne Alfieri, Ingrid Masson, Augustin

Bibliography

Mervoyer, et al. External validation of a combined pet and mri radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. *European journal of nuclear medicine and molecular imaging*, 46(4):864–877, 2019.

[134] Vegard Nygaard, Einar Andreas Rødland, and Eivind Hovig. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, 17(1):29–39, 2016.

[135] J Ferlay, M Colombet, I Soerjomataram, T Dyba, G Randi, M Bettio, A Gavin, O Visser, and F Bray. Cancer incidence and mortality patterns in europe: Estimates for 40 countries and 25 major cancers in 2018. *European Journal of Cancer*, 103:356–387, 2018.

[136] Andrew J Wong, Aasheesh Kanwar, Abdallah S Mohamed, and Clifton D Fuller. Radiomics in head and neck cancer: from exploration to application. *Translational cancer research*, 5(4):371, 2016.

[137] Ralph TH Leijenaar, Sara Carvalho, Frank JP Hoebers, Hugo JWL Aerts, Wouter JC Van Elmpt, Shao Hui Huang, Biu Chan, John N Waldron, Brian Oâsullivan, and Philippe Lambin. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. *Acta oncologica*, 54(9):1423–1429, 2015.

[138] Stefan Leger, Alex Zwanenburg, Karoline Pilz, Fabian Lohaus, Annett Linge, Klaus Zöphel, Jörg Kotzerke, Andreas Schreiber, Inge Tinhofer, Volker Budach, et al. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. *Scientific reports*, 7 (1):1–11, 2017.

[139] Tian-Tian Zhai, Lisanne V van Dijk, Bao-Tian Huang, Zhi-Xiong Lin, Cássia O Ribeiro, Charlotte L Brouwer, Sjoukje F Oosting, Gyorgy B Halmos, Max JH Witjes, Johannes A Langendijk, et al. Improving the prediction of overall survival for head and neck cancer patients using image biomarkers in combination with clinical parameters. *Radiotherapy and Oncology*, 124(2):256–262, 2017.

[140] Maarten Lambrecht, Ben Van Calster, Vincent Vandecaveye, Frederik De Keyzer, Ilse Roebben, Robert Hermans, and Sandra Nuyts. Integrating pretreatment diffusion weighted MRI into a multivariable prognostic model for head and neck squamous cell carcinoma. *Radiotherapy and Oncology*, 110(3):429–434, 2014.

[141] Xue Ming, Ronald Wihal Oei, Ruiping Zhai, Fangfang Kong, Chengrun Du, Chaosu Hu, Weigang Hu, Zhen Zhang, Hongmei Ying, and Jiazhou Wang. MRI-based radiomics signature is a quantitative prognostic biomarker for nasopharyngeal carcinoma. *Scientific reports*, 9(1):1–9, 2019.

[142] Kaixuan Yang, Jiangfang Tian, Bin Zhang, Mei Li, Wenji Xie, Yating Zou, Qiaoyue Tan, Lihui Liu, Jinbing Zhu, Arthur Shou, et al. A multidimensional nomogram combining overall stage, dose volume histogram parameters and radiomics to predict progression-free survival in patients with locoregionally advanced nasopharyngeal carcinoma. *Oral Oncology*, 98:85–91, 2019.

[143] En-Hong Zhuo, Wei-Jing Zhang, Hao-Jiang Li, Guo-Yi Zhang, Bing-Zhong Jing, Jian Zhou, Chun-Yan Cui, Ming-Yuan Chen, Ying Sun, Li-Zhi Liu, et al. Radiomics on multi-modalities mr sequences can subtype patients with non-metastatic nasopharyngeal carcinoma (NPC) into distinct survival subgroups. *European radiology*, 29(10):5590–5599, 2019.

[144] Juan C Gutierrez, Eduardo A Perez, Dido Franceschi, Frederick L Moffat Jr, Alan S Livingstone, and Leonidas G Koniaris. Outcomes for soft-tissue sarcoma in 8249 cases from a large state cancer registry. *Journal of surgical research*, 141(1):105–114, 2007.

[145] Martin Vallières, Carolyn R Freeman, Sonia R Skamene, and Issam El Naqa. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Physics in Medicine & Biology*, 60(14):5471, 2015.

[146] Georgios C Manikis, Katerina Nikiforaki, Eleni Lagoudaki, Eelco de Bree, Thomas G Maris, Kostas Marias, and Apostolos H Karantanas. T2-based MRI radiomic features for discriminating tumour grading in soft tissues sarcomas. *Hellenic Journal of Radiology*, 4(3), 2019.

[147] Matthew B Spraker, Landon S Wootton, Daniel S Hippe, Kevin C Ball, Jan C Peeken, Meghan W Macomber, Tobias R Chapman, Michael N Hoff, Edward Y Kim, Seth M Pollack, et al. MRI radiomic features are independently associated with overall survival in soft tissue sarcoma. *Advances in radiation oncology*, 4(2):413–421, 2019.

[148] Amandine Crombé, Cynthia Périer, Michèle Kind, Baudouin Denis De Senneville, François Le Loarer, Antoine Italiano, Xavier Buy, and Olivier Saut. T2-based mri delta-radiomics improve response prediction in soft-tissue sarcomas treated by neoadjuvant chemotherapy. *Journal of Magnetic Resonance Imaging*, 50(2):497–510, 2019.

[149] Melissa Vos, MPA Starmans, MJM Timbergen, SR van der Voort, GA Padmos, W Kessels, WJ Niessen, GJLH van Leenders, DJ Grünhagen, Stefan Sleijfer, et al. Radiomics approach to distinguish between well differentiated liposarcomas and lipomas on MRI. *The British journal of surgery*, 106(13):1800, 2019.

APPENDIX A

List of stable features

In this appendix, the list of stable features, as identified after the experiments described in Chapter 4, is provided. First the features that are stable in both HNC and STS are listed. Then, the list of features that are stable exclusively for HNC is given. Last, the features that are stable only for STS are listed.

The code for the feature is the same as Pyradiomics except for the prefix that has been added, indicating the type of image from which the feature has been extracted.

Stable features for both HNC and STS

- T1w-original-shape-Elongation
- T1w-original-shape-Flatness
- T1w-original-shape-LeastAxisLength
- T1w-original-shape-MajorAxisLength
- T1w-original-shape-Maximum2DDiameterColumn
- T1w-original-shape-Maximum2DDiameterRow
- T1w-original-shape-Maximum2DDiameterSlice
- T1w-original-shape-Maximum3DDiameter
- T1w-original-shape-MeshVolume
- T1w-original-shape-MinorAxisLength
- T1w-original-shape-SurfaceArea

Appendix A. List of stable features

- T1w-original-shape-SurfaceVolumeRatio
- T1w-original-shape-VoxelVolume
- T1w-original-firstorder-Energy
- T1w-original-firstorder-InterquartileRange
- T1w-original-firstorder-Mean
- T1w-original-firstorder-Median
- T1w-original-firstorder-RobustMeanAbsoluteDeviation
- T1w-original-firstorder-RootMeanSquared
- T1w-original-firstorder-TotalEnergy
- T1w-original-glrlm-GrayLevelNonUniformity
- T1w-original-glrlm-RunLengthNonUniformity
- T1w-waveletLLH-firstorder-Entropy
- T1w-waveletLLH-firstorder-Mean
- T1w-waveletLLH-firstorder-Median
- T1w-waveletLLH-firstorder-Uniformity
- T1w-waveletLLH-glcm-ClusterProminence
- T1w-waveletLLH-glcm-ClusterTendency
- T1w-waveletLLH-glcm-Contrast
- T1w-waveletLLH-glcm-DifferenceAverage
- T1w-waveletLLH-glcm-Idmn
- T1w-waveletLLH-glcm-Idn
- T1w-waveletLLH-glcm-JointEntropy
- T1w-waveletLLH-glcm-SumEntropy
- T1w-waveletLLH-glcm-SumSquares
- T1w-waveletLLH-glrlm-GrayLevelNonUniformity
- T1w-waveletLLH-glrlm-GrayLevelNonUniformityNormalized
- T1w-waveletLLH-glrlm-GrayLevelVariance
- T1w-waveletLLH-glrlm-RunLengthNonUniformity
- T1w-waveletLLH-glrlm-RunLengthNonUniformityNormalized
- T1w-waveletLLH-glrlm-RunPercentage
- T1w-waveletLLH-glrlm-ShortRunEmphasis
- T1w-waveletLHL-firstorder-90Percentile
- T1w-waveletLHL-firstorder-Entropy
- T1w-waveletLHL-firstorder-Median
- T1w-waveletLHL-glcm-ClusterTendency
- T1w-waveletLHL-glcm-Contrast
- T1w-waveletLHL-glcm-DifferenceAverage
- T1w-waveletLHL-glcm-Idmn

- T1w-waveletLHL-glcm-Idn
- T1w-waveletLHL-glcm-SumEntropy
- T1w-waveletLHL-glcm-SumSquares
- T1w-waveletLHL-glrlm-GrayLevelNonUniformity
- T1w-waveletLHL-glrlm-RunLengthNonUniformity
- T1w-waveletLHH-glrlm-GrayLevelNonUniformity
- T1w-waveletLHH-glrlm-RunLengthNonUniformity
- T1w-waveletHLL-firstorder-Entropy
- T1w-waveletHLL-firstorder-Median
- T1w-waveletHLL-glcm-DifferenceAverage
- T1w-waveletHLL-glcm-DifferenceEntropy
- T1w-waveletHLL-glcm-Id
- T1w-waveletHLL-glcm-Idm
- T1w-waveletHLL-glcm-Idn
- T1w-waveletHLL-glcm-SumSquares
- T1w-waveletHLL-glrlm-GrayLevelNonUniformity
- T1w-waveletHLL-glrlm-GrayLevelVariance
- T1w-waveletHLL-glrlm-RunLengthNonUniformity
- T1w-waveletHLL-glrlm-RunLengthNonUniformityNormalized
- T1w-waveletHLL-glrlm-ShortRunEmphasis
- T1w-waveletHLH-glrlm-GrayLevelNonUniformity
- T1w-waveletHLH-glrlm-RunLengthNonUniformity
- T1w-waveletHHL-glcm-ClusterTendency
- T1w-waveletHHL-glcm-SumSquares
- T1w-waveletHHL-glrlm-GrayLevelNonUniformity
- T1w-waveletHHL-glrlm-RunLengthNonUniformity
- T1w-waveletHHH-glcm-ClusterProminence
- T1w-waveletHHH-glcm-ClusterTendency
- T1w-waveletHHH-glcm-SumSquares
- T1w-waveletHHH-glrlm-GrayLevelNonUniformity
- T1w-waveletHHH-glrlm-RunLengthNonUniformity
- T1w-waveletLLL-firstorder-Energy
- T1w-waveletLLL-firstorder-InterquartileRange
- T1w-waveletLLL-firstorder-Mean
- T1w-waveletLLL-firstorder-Median
- T1w-waveletLLL-firstorder-RobustMeanAbsoluteDeviation
- T1w-waveletLLL-firstorder-RootMeanSquared
- T1w-waveletLLL-firstorder-TotalEnergy

Appendix A. List of stable features

- T1w-waveletLLL-glcm-Contrast
- T1w-waveletLLL-glcm-DifferenceAverage
- T1w-waveletLLL-glcm-DifferenceEntropy
- T1w-waveletLLL-glcm-Idmn
- T1w-waveletLLL-glcm-Idn
- T1w-waveletLLL-glrlm-GrayLevelNonUniformity
- T1w-waveletLLL-glrlm-RunLengthNonUniformity
- T2w-original-shape-Elongation
- T2w-original-shape-Flatness
- T2w-original-shape-LeastAxisLength
- T2w-original-shape-MajorAxisLength
- T2w-original-shape-Maximum2DDiameterColumn
- T2w-original-shape-Maximum2DDiameterRow
- T2w-original-shape-Maximum2DDiameterSlice
- T2w-original-shape-Maximum3DDiameter
- T2w-original-shape-MeshVolume
- T2w-original-shape-MinorAxisLength
- T2w-original-shape-SurfaceArea
- T2w-original-shape-SurfaceVolumeRatio
- T2w-original-shape-VoxelVolume
- T2w-original-firstorder-90Percentile
- T2w-original-firstorder-Energy
- T2w-original-firstorder-Entropy
- T2w-original-firstorder-InterquartileRange
- T2w-original-firstorder-Mean
- T2w-original-firstorder-Median
- T2w-original-firstorder-RobustMeanAbsoluteDeviation
- T2w-original-firstorder-RootMeanSquared
- T2w-original-firstorder-Skewness
- T2w-original-firstorder-TotalEnergy
- T2w-original-firstorder-Uniformity
- T2w-original-glcm-Autocorrelation
- T2w-original-glcm-Id
- T2w-original-glcm-Idm
- T2w-original-glcm-JointAverage
- T2w-original-glcm-JointEnergy
- T2w-original-glcm-JointEntropy
- T2w-original-glcm-MaximumProbability

- T2w-original-glcm-SumAverage
- T2w-original-glcm-SumEntropy
- T2w-original-glrlm-GrayLevelNonUniformity
- T2w-original-glrlm-GrayLevelNonUniformityNormalized
- T2w-original-glrlm-HighGrayLevelRunEmphasis
- T2w-original-glrlm-LongRunEmphasis
- T2w-original-glrlm-LongRunHighGrayLevelEmphasis
- T2w-original-glrlm-RunLengthNonUniformity
- T2w-original-glrlm-RunLengthNonUniformityNormalized
- T2w-original-glrlm-RunPercentage
- T2w-original-glrlm-RunVariance
- T2w-original-glrlm-ShortRunEmphasis
- T2w-original-glrlm-ShortRunHighGrayLevelEmphasis
- T2w-waveletLLH-firstorder-Mean
- T2w-waveletLLH-glcm-ClusterProminence
- T2w-waveletLLH-glcm-Imc1
- T2w-waveletLLH-glcm-Imc2
- T2w-waveletLLH-glcm-JointEnergy
- T2w-waveletLLH-glcm-MaximumProbability
- T2w-waveletLLH-glrlm-GrayLevelNonUniformity
- T2w-waveletLLH-glrlm-RunLengthNonUniformity
- T2w-waveletLLH-glrlm-RunLengthNonUniformityNormalized
- T2w-waveletLLH-glrlm-RunPercentage
- T2w-waveletLLH-glrlm-ShortRunEmphasis
- T2w-waveletLHL-firstorder-Entropy
- T2w-waveletLHL-firstorder-Uniformity
- T2w-waveletLHL-glcm-Contrast
- T2w-waveletLHL-glcm-DifferenceAverage
- T2w-waveletLHL-glcm-DifferenceEntropy
- T2w-waveletLHL-glcm-Id
- T2w-waveletLHL-glcm-Idm
- T2w-waveletLHL-glcm-Idmn
- T2w-waveletLHL-glcm-Idn
- T2w-waveletLHL-glcm-JointEnergy
- T2w-waveletLHL-glcm-JointEntropy
- T2w-waveletLHL-glcm-SumEntropy
- T2w-waveletLHL-glrlm-GrayLevelNonUniformity
- T2w-waveletLHL-glrlm-GrayLevelNonUniformityNormalized

Appendix A. List of stable features

- T2w-waveletLHL-glrlm-GrayLevelVariance
- T2w-waveletLHL-glrlm-RunLengthNonUniformity
- T2w-waveletLHL-glrlm-RunLengthNonUniformityNormalized
- T2w-waveletLHL-glrlm-RunPercentage
- T2w-waveletLHL-glrlm-ShortRunEmphasis
- T2w-waveletLHH-glcm-Imc1
- T2w-waveletLHH-glcm-Imc2
- T2w-waveletLHH-glrlm-GrayLevelNonUniformity
- T2w-waveletLHH-glrlm-RunLengthNonUniformity
- T2w-waveletHLL-glcm-Imc1
- T2w-waveletHLL-glcm-Imc2
- T2w-waveletHLL-glrlm-GrayLevelNonUniformity
- T2w-waveletHLL-glrlm-LongRunEmphasis
- T2w-waveletHLL-glrlm-RunLengthNonUniformity
- T2w-waveletHLL-glrlm-RunLengthNonUniformityNormalized
- T2w-waveletHLL-glrlm-RunPercentage
- T2w-waveletHLL-glrlm-ShortRunEmphasis
- T2w-waveletHLH-glcm-Imc1
- T2w-waveletHLH-glcm-Imc2
- T2w-waveletHLH-glrlm-GrayLevelNonUniformity
- T2w-waveletHLH-glrlm-RunLengthNonUniformity
- T2w-waveletHLH-glrlm-RunLengthNonUniformityNormalized
- T2w-waveletHLH-glrlm-RunPercentage
- T2w-waveletHLH-glrlm-ShortRunEmphasis
- T2w-waveletHHL-firstorder-Uniformity
- T2w-waveletHHL-glcm-Id
- T2w-waveletHHL-glcm-Idm
- T2w-waveletHHL-glcm-Idn
- T2w-waveletHHL-glcm-Imc1
- T2w-waveletHHL-glcm-Imc2
- T2w-waveletHHL-glcm-InverseVariance
- T2w-waveletHHL-glrlm-GrayLevelNonUniformity
- T2w-waveletHHL-glrlm-GrayLevelNonUniformityNormalized
- T2w-waveletHHL-glrlm-LongRunEmphasis
- T2w-waveletHHL-glrlm-RunLengthNonUniformity
- T2w-waveletHHL-glrlm-RunLengthNonUniformityNormalized
- T2w-waveletHHL-glrlm-RunPercentage
- T2w-waveletHHL-glrlm-ShortRunEmphasis

- T2w-waveletHHH-firstorder-Entropy
- T2w-waveletHHH-glcm-ClusterTendency
- T2w-waveletHHH-glcm-Contrast
- T2w-waveletHHH-glcm-DifferenceAverage
- T2w-waveletHHH-glcm-Idmn
- T2w-waveletHHH-glcm-Idn
- T2w-waveletHHH-glcm-Imc1
- T2w-waveletHHH-glcm-Imc2
- T2w-waveletHHH-glcm-SumSquares
- T2w-waveletHHH-glrlm-GrayLevelNonUniformity
- T2w-waveletHHH-glrlm-GrayLevelVariance
- T2w-waveletHHH-glrlm-RunLengthNonUniformity
- T2w-waveletHHH-glrlm-RunLengthNonUniformityNormalized
- T2w-waveletHHH-glrlm-RunPercentage
- T2w-waveletHHH-glrlm-ShortRunEmphasis
- T2w-waveletLLL-firstorder-90Percentile
- T2w-waveletLLL-firstorder-Energy
- T2w-waveletLLL-firstorder-Entropy
- T2w-waveletLLL-firstorder-InterquartileRange
- T2w-waveletLLL-firstorder-Mean
- T2w-waveletLLL-firstorder-Median
- T2w-waveletLLL-firstorder-Range
- T2w-waveletLLL-firstorder-RobustMeanAbsoluteDeviation
- T2w-waveletLLL-firstorder-RootMeanSquared
- T2w-waveletLLL-firstorder-Skewness
- T2w-waveletLLL-firstorder-TotalEnergy
- T2w-waveletLLL-firstorder-Uniformity
- T2w-waveletLLL-glcm-Autocorrelation
- T2w-waveletLLL-glcm-DifferenceAverage
- T2w-waveletLLL-glcm-Id
- T2w-waveletLLL-glcm-Idm
- T2w-waveletLLL-glcm-Idn
- T2w-waveletLLL-glcm-JointAverage
- T2w-waveletLLL-glcm-JointEnergy
- T2w-waveletLLL-glcm-JointEntropy
- T2w-waveletLLL-glcm-MaximumProbability
- T2w-waveletLLL-glcm-SumAverage
- T2w-waveletLLL-glcm-SumEntropy

Appendix A. List of stable features

- T2w-waveletLLL-glrlm-GrayLevelNonUniformity
- T2w-waveletLLL-glrlm-GrayLevelNonUniformityNormalized
- T2w-waveletLLL-glrlm-HighGrayLevelRunEmphasis
- T2w-waveletLLL-glrlm-LongRunEmphasis
- T2w-waveletLLL-glrlm-LongRunHighGrayLevelEmphasis
- T2w-waveletLLL-glrlm-RunLengthNonUniformity
- T2w-waveletLLL-glrlm-RunLengthNonUniformityNormalized
- T2w-waveletLLL-glrlm-RunPercentage
- T2w-waveletLLL-glrlm-RunVariance
- T2w-waveletLLL-glrlm-ShortRunEmphasis
- T2w-waveletLLL-glrlm-ShortRunHighGrayLevelEmphasis
- ADC-original-shape-Elongation
- ADC-original-shape-Flatness
- ADC-original-shape-LeastAxisLength
- ADC-original-shape-MajorAxisLength
- ADC-original-shape-Maximum2DDiameterColumn
- ADC-original-shape-Maximum2DDiameterRow
- ADC-original-shape-Maximum2DDiameterSlice
- ADC-original-shape-Maximum3DDiameter
- ADC-original-shape-MeshVolume
- ADC-original-shape-MinorAxisLength
- ADC-original-shape-SurfaceArea
- ADC-original-shape-SurfaceVolumeRatio
- ADC-original-shape-VoxelVolume
- ADC-original-firstorder-10Percentile
- ADC-original-firstorder-90Percentile
- ADC-original-firstorder-Energy
- ADC-original-firstorder-Mean
- ADC-original-firstorder-Median
- ADC-original-firstorder-RootMeanSquared
- ADC-original-firstorder-Skewness
- ADC-original-firstorder-TotalEnergy
- ADC-original-firstorder-Uniformity
- ADC-original-glcm-Autocorrelation
- ADC-original-glcm-Id
- ADC-original-glcm-Idm
- ADC-original-glcm-Imc1
- ADC-original-glcm-JointAverage

- ADC-original-glcm-SumAverage
- ADC-original-glrlm-GrayLevelNonUniformity
- ADC-original-glrlm-HighGrayLevelRunEmphasis
- ADC-original-glrlm-LongRunEmphasis
- ADC-original-glrlm-LongRunHighGrayLevelEmphasis
- ADC-original-glrlm-LowGrayLevelRunEmphasis
- ADC-original-glrlm-RunLengthNonUniformity
- ADC-original-glrlm-RunLengthNonUniformityNormalized
- ADC-original-glrlm-RunPercentage
- ADC-original-glrlm-RunVariance
- ADC-original-glrlm-ShortRunEmphasis
- ADC-original-glrlm-ShortRunHighGrayLevelEmphasis
- ADC-original-glrlm-ShortRunLowGrayLevelEmphasis
- ADC-waveletLLH-firstorder-Mean
- ADC-waveletLLH-glcm-Imc1
- ADC-waveletLLH-glrlm-GrayLevelNonUniformity
- ADC-waveletLLH-glrlm-LongRunEmphasis
- ADC-waveletLLH-glrlm-RunLengthNonUniformity
- ADC-waveletLLH-glrlm-RunLengthNonUniformityNormalized
- ADC-waveletLLH-glrlm-RunPercentage
- ADC-waveletLLH-glrlm-RunVariance
- ADC-waveletLLH-glrlm-ShortRunEmphasis
- ADC-waveletLHL-firstorder-Entropy
- ADC-waveletLHL-glcm-Contrast
- ADC-waveletLHL-glcm-DifferenceAverage
- ADC-waveletLHL-glcm-DifferenceEntropy
- ADC-waveletLHL-glcm-Id
- ADC-waveletLHL-glcm-Idm
- ADC-waveletLHL-glcm-Idmn
- ADC-waveletLHL-glcm-Idn
- ADC-waveletLHL-glcm-InverseVariance
- ADC-waveletLHL-glrlm-GrayLevelNonUniformity
- ADC-waveletLHL-glrlm-GrayLevelVariance
- ADC-waveletLHL-glrlm-RunLengthNonUniformity
- ADC-waveletLHL-glrlm-RunLengthNonUniformityNormalized
- ADC-waveletLHL-glrlm-RunPercentage
- ADC-waveletLHL-glrlm-ShortRunEmphasis
- ADC-waveletLHH-glcm-ClusterTendency

Appendix A. List of stable features

- ADC-waveletLHH-glcm-Contrast
- ADC-waveletLHH-glcm-DifferenceAverage
- ADC-waveletLHH-glcm-Idmn
- ADC-waveletLHH-glcm-Idn
- ADC-waveletLHH-glcm-Imc1
- ADC-waveletLHH-glcm-Imc2
- ADC-waveletLHH-glcm-SumSquares
- ADC-waveletLHH-glrlm-GrayLevelNonUniformity
- ADC-waveletLHH-glrlm-GrayLevelVariance
- ADC-waveletLHH-glrlm-RunLengthNonUniformity
- ADC-waveletLHH-glrlm-RunLengthNonUniformityNormalized
- ADC-waveletLHH-glrlm-RunPercentage
- ADC-waveletLHH-glrlm-ShortRunEmphasis
- ADC-waveletHLL-glcm-Imc1
- ADC-waveletHLL-glrlm-GrayLevelNonUniformity
- ADC-waveletHLL-glrlm-GrayLevelVariance
- ADC-waveletHLL-glrlm-RunLengthNonUniformity
- ADC-waveletHLL-glrlm-RunLengthNonUniformityNormalized
- ADC-waveletHLL-glrlm-RunPercentage
- ADC-waveletHLL-glrlm-ShortRunEmphasis
- ADC-waveletHLH-glrlm-GrayLevelNonUniformity
- ADC-waveletHLH-glrlm-RunLengthNonUniformity
- ADC-waveletHHL-firstorder-Entropy
- ADC-waveletHHL-glcm-ClusterTendency
- ADC-waveletHHL-glcm-DifferenceAverage
- ADC-waveletHHL-glcm-Idn
- ADC-waveletHHL-glcm-JointEntropy
- ADC-waveletHHL-glcm-SumSquares
- ADC-waveletHHL-glrlm-GrayLevelNonUniformity
- ADC-waveletHHL-glrlm-RunLengthNonUniformity
- ADC-waveletHHL-glrlm-RunLengthNonUniformityNormalized
- ADC-waveletHHH-glcm-ClusterTendency
- ADC-waveletHHH-glcm-SumSquares
- ADC-waveletHHH-glrlm-GrayLevelNonUniformity
- ADC-waveletHHH-glrlm-RunLengthNonUniformity
- ADC-waveletLLL-firstorder-90Percentile
- ADC-waveletLLL-firstorder-Energy
- ADC-waveletLLL-firstorder-Mean

- ADC-waveletLLL-firstorder-Median
- ADC-waveletLLL-firstorder-RootMeanSquared
- ADC-waveletLLL-firstorder-Skewness
- ADC-waveletLLL-firstorder-TotalEnergy
- ADC-waveletLLL-firstorder-Uniformity
- ADC-waveletLLL-glcm-Autocorrelation
- ADC-waveletLLL-glcm-DifferenceAverage
- ADC-waveletLLL-glcm-Id
- ADC-waveletLLL-glcm-Idm
- ADC-waveletLLL-glcm-Idn
- ADC-waveletLLL-glcm-JointAverage
- ADC-waveletLLL-glcm-JointEnergy
- ADC-waveletLLL-glcm-MaximumProbability
- ADC-waveletLLL-glcm-SumAverage
- ADC-waveletLLL-glrlm-GrayLevelNonUniformity
- ADC-waveletLLL-glrlm-HighGrayLevelRunEmphasis
- ADC-waveletLLL-glrlm-LongRunEmphasis
- ADC-waveletLLL-glrlm-LongRunHighGrayLevelEmphasis
- ADC-waveletLLL-glrlm-RunLengthNonUniformity
- ADC-waveletLLL-glrlm-RunLengthNonUniformityNormalized
- ADC-waveletLLL-glrlm-RunPercentage
- ADC-waveletLLL-glrlm-RunVariance
- ADC-waveletLLL-glrlm-ShortRunEmphasis
- ADC-waveletLLL-glrlm-ShortRunHighGrayLevelEmphasis

Stable features for HNC only

- T1w-original-glcm-InverseVariance
- T1w-waveletLLH-glrlm-LongRunEmphasis
- T1w-waveletLLH-glrlm-RunVariance
- T1w-waveletLHL-firstorder-10Percentile
- T1w-waveletLHL-glcm-ClusterProminence
- T1w-waveletHHH-glcm-Imc2
- T2w-original-firstorder-MeanAbsoluteDeviation
- T2w-original-glrlm-GrayLevelVariance
- T2w-waveletLLH-firstorder-Median
- T2w-waveletLLH-glrlm-LongRunEmphasis
- T2w-waveletLLH-glrlm-RunVariance
- T2w-waveletLHL-glrlm-LongRunEmphasis

Appendix A. List of stable features

- T2w-waveletLHL-glrlm-RunVariance
- T2w-waveletLLL-firstorder-MeanAbsoluteDeviation
- T2w-waveletLLL-glcm-ClusterShade
- ADC-waveletLLH-firstorder-Median
- ADC-waveletLHL-glcm-DifferenceVariance
- ADC-waveletLHL-glcm-Imc1
- ADC-waveletLHL-glcm-Imc2
- ADC-waveletHLL-glcm-Imc2
- ADC-waveletHLH-glcm-Imc1
- ADC-waveletHLH-glcm-Imc2
- ADC-waveletHLH-glcm-MCC
- ADC-waveletHHL-glcm-ClusterProminence
- ADC-waveletHHL-glcm-Imc1
- ADC-waveletHHL-glcm-Imc2
- ADC-waveletHHL-glcm-InverseVariance
- ADC-waveletHHL-glrlm-GrayLevelVariance
- ADC-waveletHHH-glcm-ClusterProminence
- ADC-waveletHHH-glcm-Imc1
- ADC-waveletHHH-glcm-Imc2
- ADC-waveletLLL-firstorder-10Percentile
- ADC-waveletLLL-glcm-ClusterShade

Stable features for STS only

- T1w-original-firstorder-10Percentile
- T1w-original-firstorder-90Percentile
- T1w-original-firstorder-Entropy
- T1w-original-firstorder-Maximum
- T1w-original-firstorder-Uniformity
- T1w-original-glcm-DifferenceAverage
- T1w-original-glcm-DifferenceEntropy
- T1w-original-glcm-Id
- T1w-original-glcm-Idm
- T1w-original-glcm-Idn
- T1w-original-glcm-Imc1
- T1w-original-glcm-Imc2
- T1w-original-glcm-JointAverage
- T1w-original-glcm-JointEntropy
- T1w-original-glcm-SumAverage

- T1w-original-glrlm-LongRunEmphasis
- T1w-original-glrlm-LongRunHighGrayLevelEmphasis
- T1w-original-glrlm-LongRunLowGrayLevelEmphasis
- T1w-original-glrlm-RunLengthNonUniformityNormalized
- T1w-original-glrlm-RunPercentage
- T1w-original-glrlm-ShortRunEmphasis
- T1w-waveletLLH-glcm-Imc1
- T1w-waveletLHL-firstorder-Uniformity
- T1w-waveletLHL-glcm-DifferenceEntropy
- T1w-waveletLHL-glcm-DifferenceVariance
- T1w-waveletLHL-glcm-Id
- T1w-waveletLHL-glcm-Idm
- T1w-waveletLHL-glcm-JointEnergy
- T1w-waveletLHL-glcm-JointEntropy
- T1w-waveletLHL-glrlm-GrayLevelNonUniformityNormalized
- T1w-waveletLHL-glrlm-GrayLevelVariance
- T1w-waveletLHL-glrlm-RunLengthNonUniformityNormalized
- T1w-waveletLHL-glrlm-RunPercentage
- T1w-waveletLHL-glrlm-ShortRunEmphasis
- T1w-waveletLHH-firstorder-Entropy
- T1w-waveletLHH-glcm-ClusterTendency
- T1w-waveletLHH-glcm-Contrast
- T1w-waveletLHH-glcm-DifferenceAverage
- T1w-waveletLHH-glcm-Idmn
- T1w-waveletLHH-glcm-Idn
- T1w-waveletLHH-glcm-SumSquares
- T1w-waveletLHH-glrlm-GrayLevelVariance
- T1w-waveletLHH-glrlm-RunLengthNonUniformityNormalized
- T1w-waveletLHH-glrlm-RunPercentage
- T1w-waveletLHH-glrlm-ShortRunEmphasis
- T1w-waveletHLL-firstorder-Uniformity
- T1w-waveletHLL-glcm-ClusterTendency
- T1w-waveletHLL-glcm-DifferenceVariance
- T1w-waveletHLL-glcm-JointEnergy
- T1w-waveletHLL-glcm-JointEntropy
- T1w-waveletHLL-glcm-MaximumProbability
- T1w-waveletHLL-glcm-SumEntropy
- T1w-waveletHLL-glrlm-GrayLevelNonUniformityNormalized

Appendix A. List of stable features

- T1w-waveletHLL-glrlm-RunPercentage
- T1w-waveletHLH-firstorder-Entropy
- T1w-waveletHLH-glrlm-GrayLevelVariance
- T1w-waveletHLH-glrlm-RunLengthNonUniformityNormalized
- T1w-waveletHLH-glrlm-ShortRunEmphasis
- T1w-waveletHHL-firstorder-Entropy
- T1w-waveletHHL-firstorder-Uniformity
- T1w-waveletHHL-glcm-ClusterProminence
- T1w-waveletHHL-glcm-DifferenceAverage
- T1w-waveletHHL-glcm-Idn
- T1w-waveletHHL-glcm-JointEntropy
- T1w-waveletHHL-glrlm-GrayLevelNonUniformityNormalized
- T1w-waveletHHL-glrlm-GrayLevelVariance
- T1w-waveletHHL-glrlm-RunLengthNonUniformityNormalized
- T1w-waveletHHL-glrlm-RunPercentage
- T1w-waveletHHL-glrlm-ShortRunEmphasis
- T1w-waveletLLL-firstorder-10Percentile
- T1w-waveletLLL-firstorder-90Percentile
- T1w-waveletLLL-firstorder-Entropy
- T1w-waveletLLL-firstorder-Maximum
- T1w-waveletLLL-firstorder-Uniformity
- T1w-waveletLLL-glcm-Id
- T1w-waveletLLL-glcm-Idm
- T1w-waveletLLL-glcm-Imc1
- T1w-waveletLLL-glcm-JointAverage
- T1w-waveletLLL-glcm-JointEnergy
- T1w-waveletLLL-glcm-JointEntropy
- T1w-waveletLLL-glcm-MaximumProbability
- T1w-waveletLLL-glcm-SumAverage
- T1w-waveletLLL-glcm-SumEntropy
- T1w-waveletLLL-glrlm-LongRunHighGrayLevelEmphasis
- T1w-waveletLLL-glrlm-LowGrayLevelRunEmphasis
- T1w-waveletLLL-glrlm-RunLengthNonUniformityNormalized
- T1w-waveletLLL-glrlm-RunPercentage
- T1w-waveletLLL-glrlm-ShortRunEmphasis
- T2w-original-firstorder-Maximum
- T2w-original-firstorder-Range
- T2w-original-glcm-Imc1

- T2w-waveletLLH-firstorder-Entropy
- T2w-waveletLLH-firstorder-Kurtosis
- T2w-waveletLLH-firstorder-Uniformity
- T2w-waveletLLH-glcm-ClusterTendency
- T2w-waveletLLH-glcm-DifferenceAverage
- T2w-waveletLLH-glcm-DifferenceEntropy
- T2w-waveletLLH-glcm-Id
- T2w-waveletLLH-glcm-Idm
- T2w-waveletLLH-glcm-Idmn
- T2w-waveletLLH-glcm-Idn
- T2w-waveletLLH-glcm-JointEntropy
- T2w-waveletLLH-glcm-MCC
- T2w-waveletLLH-glcm-SumEntropy
- T2w-waveletLLH-glcm-SumSquares
- T2w-waveletLLH-glrlm-GrayLevelNonUniformityNormalized
- T2w-waveletLLH-glrlm-GrayLevelVariance
- T2w-waveletLLH-glrlm-RunEntropy
- T2w-waveletLHL-firstorder-Kurtosis
- T2w-waveletLHL-glcm-DifferenceVariance
- T2w-waveletLHL-glcm-InverseVariance
- T2w-waveletLHL-glcm-MCC
- T2w-waveletLHL-glcm-MaximumProbability
- T2w-waveletLHH-firstorder-Entropy
- T2w-waveletLHH-firstorder-Kurtosis
- T2w-waveletLHH-firstorder-Uniformity
- T2w-waveletLHH-glcm-ClusterProminence
- T2w-waveletLHH-glcm-ClusterTendency
- T2w-waveletLHH-glcm-Contrast
- T2w-waveletLHH-glcm-DifferenceAverage
- T2w-waveletLHH-glcm-DifferenceEntropy
- T2w-waveletLHH-glcm-DifferenceVariance
- T2w-waveletLHH-glcm-Id
- T2w-waveletLHH-glcm-Idm
- T2w-waveletLHH-glcm-Idmn
- T2w-waveletLHH-glcm-Idn
- T2w-waveletLHH-glcm-JointEnergy
- T2w-waveletLHH-glcm-JointEntropy
- T2w-waveletLHH-glcm-SumEntropy

Appendix A. List of stable features

- T2w-waveletLHH-glcm-SumSquares
- T2w-waveletLHH-glrlm-GrayLevelNonUniformityNormalized
- T2w-waveletLHH-glrlm-GrayLevelVariance
- T2w-waveletLHH-glrlm-RunLengthNonUniformityNormalized
- T2w-waveletLHH-glrlm-RunPercentage
- T2w-waveletLHH-glrlm-ShortRunEmphasis
- T2w-waveletHLL-firstorder-Entropy
- T2w-waveletHLL-firstorder-Kurtosis
- T2w-waveletHLL-firstorder-Uniformity
- T2w-waveletHLL-glcm-ClusterTendency
- T2w-waveletHLL-glcm-Contrast
- T2w-waveletHLL-glcm-DifferenceAverage
- T2w-waveletHLL-glcm-DifferenceEntropy
- T2w-waveletHLL-glcm-DifferenceVariance
- T2w-waveletHLL-glcm-Id
- T2w-waveletHLL-glcm-Idm
- T2w-waveletHLL-glcm-Idmn
- T2w-waveletHLL-glcm-Idn
- T2w-waveletHLL-glcm-JointEnergy
- T2w-waveletHLL-glcm-JointEntropy
- T2w-waveletHLL-glcm-MaximumProbability
- T2w-waveletHLL-glcm-SumEntropy
- T2w-waveletHLL-glcm-SumSquares
- T2w-waveletHLL-glrlm-GrayLevelNonUniformityNormalized
- T2w-waveletHLL-glrlm-GrayLevelVariance
- T2w-waveletHLH-firstorder-Entropy
- T2w-waveletHLH-glcm-MCC
- T2w-waveletHLH-glrlm-GrayLevelVariance
- T2w-waveletHHL-firstorder-Entropy
- T2w-waveletHHL-firstorder-Kurtosis
- T2w-waveletHHL-glcm-ClusterProminence
- T2w-waveletHHL-glcm-ClusterTendency
- T2w-waveletHHL-glcm-Contrast
- T2w-waveletHHL-glcm-DifferenceAverage
- T2w-waveletHHL-glcm-DifferenceEntropy
- T2w-waveletHHL-glcm-Idmn
- T2w-waveletHHL-glcm-JointEntropy
- T2w-waveletHHL-glcm-SumEntropy

- T2w-waveletHHL-glcm-SumSquares
- T2w-waveletHHL-glrlm-GrayLevelVariance
- T2w-waveletHHH-glcm-ClusterProminence
- T2w-waveletHHH-glcm-DifferenceVariance
- T2w-waveletHHH-glcm-MCC
- T2w-waveletLLL-firstorder-Maximum
- T2w-waveletLLL-glcm-Imc1
- ADC-original-firstorder-Entropy
- ADC-original-firstorder-InterquartileRange
- ADC-original-firstorder-Maximum
- ADC-original-firstorder-MeanAbsoluteDeviation
- ADC-original-firstorder-RobustMeanAbsoluteDeviation
- ADC-original-glcm-ClusterTendency
- ADC-original-glcm-JointEntropy
- ADC-original-glcm-SumSquares
- ADC-original-glrlm-GrayLevelNonUniformityNormalized
- ADC-waveletLLH-firstorder-Entropy
- ADC-waveletLLH-firstorder-Uniformity
- ADC-waveletLLH-glcm-ClusterProminence
- ADC-waveletLLH-glcm-ClusterTendency
- ADC-waveletLLH-glcm-DifferenceAverage
- ADC-waveletLLH-glcm-Idmn
- ADC-waveletLLH-glcm-Idn
- ADC-waveletLLH-glcm-JointEntropy
- ADC-waveletLLH-glcm-SumEntropy
- ADC-waveletLLH-glcm-SumSquares
- ADC-waveletLLH-glrlm-GrayLevelNonUniformityNormalized
- ADC-waveletLLH-glrlm-GrayLevelVariance
- ADC-waveletLHL-firstorder-Uniformity
- ADC-waveletLHL-glcm-JointEnergy
- ADC-waveletLHL-glcm-JointEntropy
- ADC-waveletLHL-glcm-MaximumProbability
- ADC-waveletLHL-glcm-SumEntropy
- ADC-waveletLHL-glrlm-GrayLevelNonUniformityNormalized
- ADC-waveletLHL-glrlm-LongRunEmphasis
- ADC-waveletLHH-firstorder-Entropy
- ADC-waveletLHH-glrlm-LongRunEmphasis
- ADC-waveletHLL-firstorder-Entropy

Appendix A. List of stable features

- ADC-waveletHLL-firstorder-Uniformity
- ADC-waveletHLL-glcm-ClusterTendency
- ADC-waveletHLL-glcm-DifferenceAverage
- ADC-waveletHLL-glcm-DifferenceEntropy
- ADC-waveletHLL-glcm-Id
- ADC-waveletHLL-glcm-Idm
- ADC-waveletHLL-glcm-Idn
- ADC-waveletHLL-glcm-JointEnergy
- ADC-waveletHLL-glcm-JointEntropy
- ADC-waveletHLL-glcm-MaximumProbability
- ADC-waveletHLL-glcm-SumEntropy
- ADC-waveletHLL-glcm-SumSquares
- ADC-waveletHLL-glrlm-GrayLevelNonUniformityNormalized
- ADC-waveletHLH-firstorder-Entropy
- ADC-waveletHLH-glrlm-GrayLevelVariance
- ADC-waveletHLH-glrlm-RunLengthNonUniformityNormalized
- ADC-waveletHLH-glrlm-ShortRunEmphasis
- ADC-waveletHHL-firstorder-Uniformity
- ADC-waveletHHL-glrlm-GrayLevelNonUniformityNormalized
- ADC-waveletHHL-glrlm-RunPercentage
- ADC-waveletHHL-glrlm-ShortRunEmphasis
- ADC-waveletLLL-firstorder-Entropy
- ADC-waveletLLL-firstorder-InterquartileRange
- ADC-waveletLLL-firstorder-Maximum
- ADC-waveletLLL-firstorder-MeanAbsoluteDeviation
- ADC-waveletLLL-firstorder-Range
- ADC-waveletLLL-firstorder-RobustMeanAbsoluteDeviation
- ADC-waveletLLL-firstorder-Variance
- ADC-waveletLLL-glcm-ClusterTendency
- ADC-waveletLLL-glcm-Imc1
- ADC-waveletLLL-glcm-Imc2
- ADC-waveletLLL-glcm-JointEntropy
- ADC-waveletLLL-glcm-SumEntropy
- ADC-waveletLLL-glcm-SumSquares
- ADC-waveletLLL-glrlm-GrayLevelNonUniformityNormalized