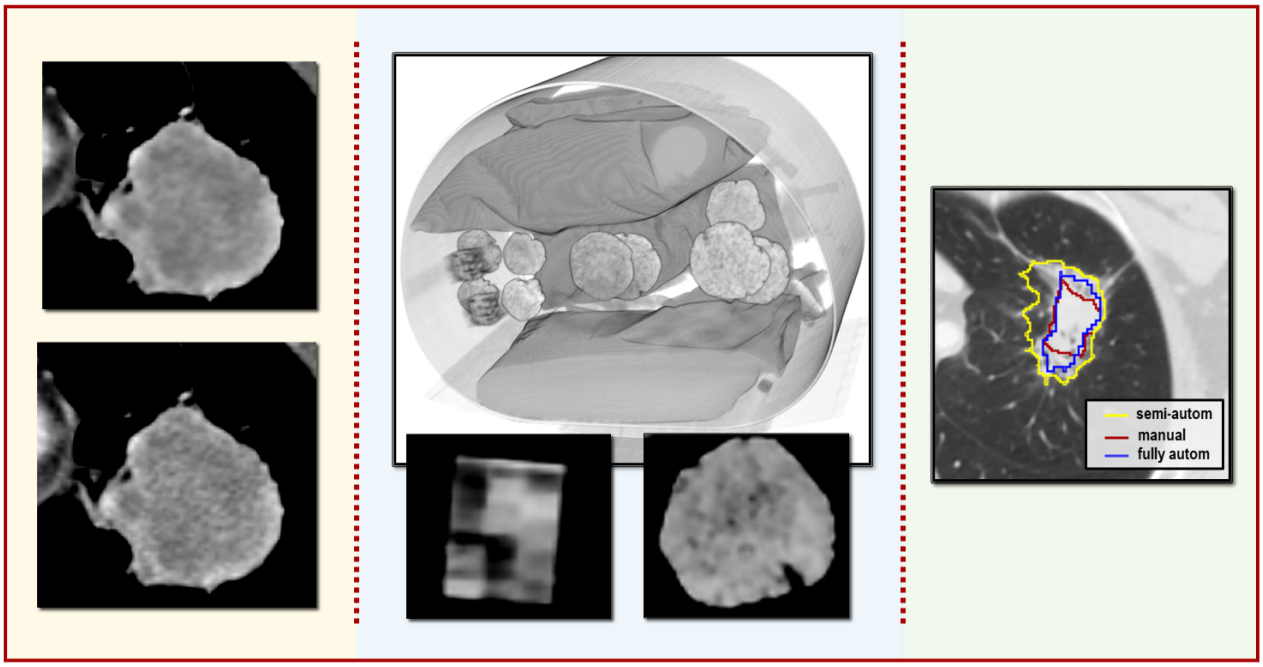
UNIVERSITÀ DEGLI STUDI DI PAVIA
DOTTORATO DI RICERCA IN FISICA - XXXIV CICLO

From grey-levels to numbers

Investigation of radiomic feature robustness in CT images of lung tumours

Lisa Rinaldi



Tesi per il conseguimento del titolo

UNIVERSITÀ DEGLI STUDI DI PAVIA
DOTTORATO DI RICERCA IN FISICA - XXXIV CICLO

FROM GREY-LEVELS TO NUMBERS

INVESTIGATION OF RADIOMIC FEATURE ROBUSTNESS IN CT IMAGES OF LUNG TUMOURS

LISA RINALDI

Supervisor: Botta Francesca

Submitted to the Graduate School of Physics
in partial fulfillment of the requirements for the degree of
DOTTORE DI RICERCA IN FISICA
DOCTOR OF PHILOSOPHY IN PHYSICS
at the
University of Pavia

Cover:

Representative pictures of the three main topics covered in this thesis: parameter influence on radiomic features, CT phantom fabrication and segmentation. *Left*: CT axial images of the same lung lesion but reconstructed with a different algorithm (iterative on the top and FBP on the bottom) *Centre*: schematic representation of the phantom developed for radiomic purposes in this thesis along with the CT axial images of two inserts, taken as examples. *Right*: illustrative comparison of three different contours (manual, semi-automatic and fully automatic) of a lung lesion.

From grey-levels to numbers: investigation of radiomic feature robustness in CT images of lung tumours
Lisa Rinaldi
PhD Thesis – University of Pavia
Pavia, Italy, July 2022

What we would see, instead, would be a matter of complementation. It could be that human and computer might form a symbiotic intelligence that would be far greater than either could develop alone, a symbiotic intelligence that would open new horizons and make it possible to achieve new heights.

— Isaac Asimov, *The Roving Mind*, 1983

Contents

Introduction

Radiomics is a technique that is spreading more and more in the field of medical images. The interest in radiomics from the scientific community arises from the fact that it is an intuitive and non-expensive discipline, which aims at characterising and predicting the evolution of the disease. The key idea underlying radiomics is to view the medical images not just as pictures to be analysed visually, but rather as raw data. After all, the medical images are matrices of numbers representing grey-level intensities. Showing them as pictures certainly unveils a great wealth of visual information, but is that all? Radiomics aims at uncovering and exploiting the entire information contained in the medical images, possibly going beyond what can be appreciated by the human eye. For this purpose radiomics makes use of quantitative parameters, named *radiomic features*, which are extracted directly from the medical images. These descriptors capture various properties of the lesion, such as the shape and the local heterogeneity of its texture, constituting a sort of fingerprint of the lesion, built in the medical image. Usually, hundreds or even thousands of these parameters can be defined and extracted. An essential part of the radiomic analysis is the identification of the radiomic features which are able to detect the underlying biological characteristics of the lesion. Radiomic data, for instance, may be useful to classify the pathology of interest (i.e. malignant or benign), to predict an outcome (overall survival, response to a therapy, etc.) or to identify the genetic mutation profile. The main advantage of this approach is that the description of the lesions is performed using images that are already acquired in the clinical routine for the staging and the monitoring of the disease. Radiomics is therefore a non-invasive technique, which may become a precious support for the physicians in the choice of the best treatment for each patient.

However, radiomics is not yet a robust enough technique. Some challenges still remain, preventing the introduction of radiomics in the clinical practice. One of main issues, which is not yet fully under control, is the radiomic feature stability. For example, datasets of images collected in different periods of time or acquired in multiple institutes are characterised by different acquisition protocols and scanners. Moreover, there is not a strict consensus in how to delineate the volume of interest, from which the features are extracted. This variability may introduce a significant level of noise in the texture and in the resulting features, obscuring the underlying phenotype and giving spurious

results in the final model.

For this reason, in recent years several methodological studies have been performed to explore feature variability, for different imaging modalities — mainly computed tomography (CT), positron emission tomography (PET) and magnetic resonance imaging (MRI) — and different anatomical districts. However, even though nowadays there exist some guidelines which are commonly adopted by researchers when performing radiomic studies — such as the recommendations provided by the Image Biomarker Standardisation Initiative (IBSI) [1] — the lack of standardisation in the radiomic studies makes it difficult to generalise and validate the results.

In this thesis the robustness of radiomic features in CT imaging was investigated, focusing specifically on the oncological field. In particular, the anatomical target considered in all the presented analyses was the lung tumour. The main objective of the thesis was the characterisation of the feature behaviour in different acquisition settings, by performing methodological studies both in patients and in phantom. To this aim, the CT images of patients affected by lung tumours were collected to investigate feature reproducibility when several parameters are varied. In particular, the impact of the scanner model, of the X-ray tube voltage, of the reconstruction blending levels and of the segmentation was evaluated. Phantoms are, instead, inanimate objects, fabricated to reproduce the human shape and/or signal in the medical image, and therefore useful as patient surrogates. An important part of this project was dedicated precisely to the identification of the suitable materials and the design of a phantom for radiomic purposes. The phantom was assembled so as to encompass inserts specifically developed to mimic the CT signal of lung tumours and to have a heterogeneous texture. The main reason why we need a phantom was to assess feature repeatability, by repeating the same acquisition multiple times in fixed conditions. This scenario is in fact difficult to achieve in patients because of the limits imposed by the exposure to radiation. Beyond that, the phantom was also used to compare the features among various acquisition/reconstruction conditions (different voltage peaks, scanners and reconstruction algorithms).

All the experiments and analyses were performed in collaboration with the European Institute of Oncology (IEO) in Milano.

Thesis overview

Chapter 1 introduces radiomics as a tool developed to capture the tumour heterogeneity. The radiomic workflow is presented, from image acquisition to data analysis. Special attention is given to the pre-processing techniques, which are the procedures usually applied before the feature extraction, and to the categories of features used in this thesis. After a brief introduction to CT images, where the meaning of their numerical content and the main parameters involved in their formation are illustrated, the chapter covers the

main limitations and open challenges of radiomics for this imaging modality. The importance of methodological investigations to identify robust features and the need for common strategies in performing radiomic studies are particularly emphasised. Finally, a quick overview of the literature of radiomics in the lung oncology is provided, with examples of applications for various clinical outcomes.

In **Chapter 2** a methodological study on feature reproducibility is presented, considering for the analysis the diagnostic CT images of patients with lung tumours. The impact of the CT scanner, of the voltage peak and of the reconstruction algorithm on the radiomic feature reproducibility is evaluated. Two different types of analysis are performed, one based on a univariate approach and the other on a multivariable one. In this way, the impact of each separate parameter as well as the simultaneous interactions among the different parameters is analysed.

The use of phantoms in radiomic studies is an alternative approach to investigate the feature behaviour. **Chapter 3** is dedicated to this topic. First of all, the importance of phantoms in the medical field is highlighted, and the state of the art of the phantoms for radiomic purposes in CT imaging is shown. After this brief introduction, the chapter is devoted to the description of the design and fabrication of inserts mimicking lung tumours. The first part of the phantom development consists in the identification of the target grey-level intensities for the lung lesions and the lung itself, using the CT images of patients affected by lung tumour. The study of the CT signal of the anatomical structures of interest was necessary to find the most suitable materials for the phantom. After a preliminary study, we developed and characterised two prototypes of lung lesions. To evaluate the similarity in the texture between inserts and real tumours, the radiomic features extracted from the phantom CT images are compared with those extracted from the patient ones. The second part of this chapter is concerned with the analysis of the repeatability and reproducibility of the radiomic features using the fabricated inserts, showing the usefulness of such objects for radiomic studies. As in Chapter 2, the feature variability due to changes in voltage peak, scanner model and reconstruction algorithm is assessed. Finally, the advantages and disadvantages of each of the proposed insert types are discussed.

Last but not least, **Chapter 4** develops another key aspect of the radiomic workflow, which comes into play even before the feature extraction: the segmentation of the volume of interest. The chapter consists of two main parts. In the first part a semi-automatic and a fully automatic approach for lung lesion contouring are described, and their application on a group of patients affected by lung tumour is discussed. The results of these two segmentation approaches are compared to manual contours in terms of overlapping volume and radiomic feature variability. The second part, instead, focuses on the fully automatic algorithm and aims to understand the impact of segmentation on the performance of a predictive model. For this purpose, an overall survival

model is built using separately the features extracted from the manual and the automatic contours, and the performances of the two models are then compared.

Chapter 1

Radiomics

Solid tumours are heterogeneous, not only among patients but also within the same lesion. Radiological images can provide a global representation of the lesion heterogeneity, and for this reason they can be a precious tool for the development of personalised medicine. *Radiomics* is an emerging technique that uses the numerical content of the medical images to achieve this goal by extracting quantitative descriptors, called *radiomic features*, from the image of the lesion. Radiomic features capture information that is complementary to the clinical and genetic one, and this collection of data can be used to build new predictive models, or improve those already used in the clinics, with the ultimate aim of supporting the clinicians in the choice of the most suitable treatment for each patient.

Chapter 1 presents radiomics, starting with a gentle introduction of this technique, where the main fields of applicability are illustrated, along with its main advantages and drawbacks. Then, the radiomic workflow will be described in detail, from image acquisition to statistical analysis. Finally, the chapter will focus on computed tomography (CT) images, especially on the CT parameters which may have an impact on the feature robustness. Special attention is given to the lung district, which is the anatomical region of interest for the studies presented in Chapters 2, 3 and 4.

1.1 A gentle introduction to radiomics

"The science on which it (medicine) is based is accurate and definite enough; the physics of a man's circulation are the physics of the waterworks of the town in which he lives, but once out of gear, you cannot apply the same rules for the repair of the one as of the other. Variability is the law of life, and as no two faces are the same, so no two bodies are alike, and no two individuals react alike and behave alike under the abnormal conditions which we know as disease." [2] With these words William Osler, one of the fathers of modern medicine, in 1904 described the difficulties for a physician in the art of medicine.

Diversity among tumours of different types and among patients with the same tumour type (inter-tumour heterogeneity) is well established as concerns its genetic and phenotypic profile. Indeed, this variability is among the causes which gave rise to personalised medicine, namely the idea of tailoring the medical treatment to the specific characteristics of each patient. But this is not

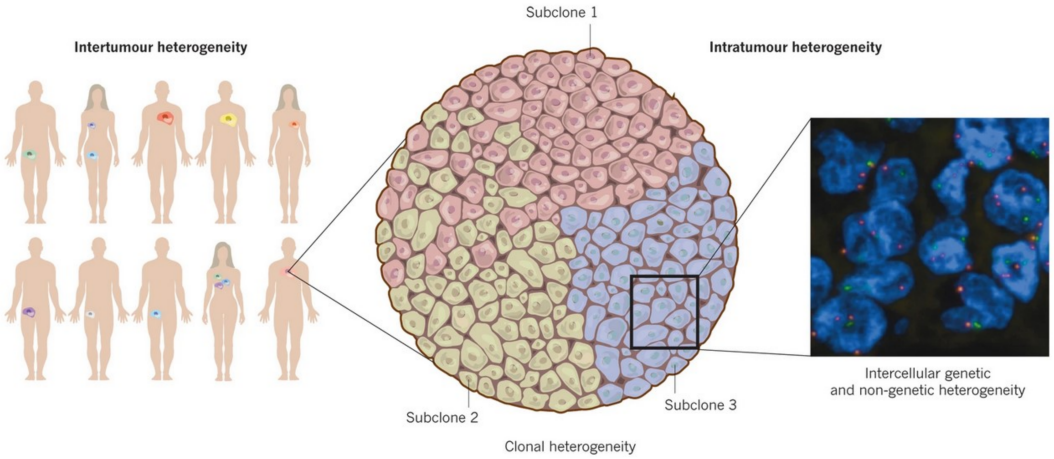


Figure 1.1: Illustrative representation of the tumour complexity: variability was observed among different tumour types, among patients with the same pathology and even within the same lesion. Inside the tumour, sub-population of cells with a different underlying genetic and epigenetic package can grow and mix, increasing the tumour heterogeneity, as illustrated for subclone 2 in the figure. On the right, a picture of fluorescence *in situ* hybridisation of cancerous cells is reported, showing the heterogeneity present among cells of the same tumour sub-clone. The centromeres of two chromosomes are depicted in red and in green, the DNA in blue. Image from ref. [5].

all. An intra-tumour heterogeneity is added to this variability as well (e.g. see **Figure 1.1**). Spatial and temporal heterogeneity, in fact, is observed inside tumoral lesions and is considered as one of the main factors of cancer resistance to oncological therapies [3,4]. The histological and/or genetic characteristics of the tumour are typically identified by carrying out a solid biopsy of the lesion. However, biopsies can capture this information only from a small portion of the lesion volume. Moreover, because of their invasiveness, they are performed typically once per patient, usually at the beginning of the therapeutic treatment, and — even if in some cases the biopsies are performed more than once — it is not feasible to repeat them routinely over time. Therefore, the spatial and time information about the tumour heterogeneity is lost, and monitoring of the disease during the treatment turns out to be unfeasible.

In order to overcome these limitations several new strategies have been proposed in the recent years. The liquid biopsy-based technique, for example, studies the genetic profile of the tumour from blood samples, thus analysing the circulating material belonging to the entire tumour and metastasis, if present. Moreover, thanks to their reduced invasivity, liquid biopsies can be performed multiple times, this way capturing the time evolution of the disease. A completely different approach is the quantitative description of the tissue properties through the analysis of medical images. This is the basic idea of radiomics.

Radiomics is an emerging field of research which aims to support the physicians in the choice of the best clinical strategies by providing quantitative information about the tumour through the extraction of numerical descriptors from the medical images, named radiomic features. Radiomics makes use of the

fact that the medical images are matrices of numbers, whose content depends on the physical properties of the diagnostic modality involved. This content, for example, represents the absorption coefficient of the X-rays in computed tomography (CT) and radiography, the radiotracer uptake from the detection of annihilation photons in positron emission tomography (PET), nuclei density and characteristic relaxation times in magnetic resonance imaging (MRI), and the acoustic impedance produced at the interfaces among different tissues in ultrasound imaging (US). It thus captures the details of the lesion on the voxel resolution. The voxel is the volumetric picture element and is usually of the order of the millimetre. This means that the heterogeneity produced by the underlying phenotype at lower scale is described macroscopically, catching the variability due to the metabolism, vascularisation, oxygenation and to the presence of necrotic area [6]. In the clinical practice some generic information about lesion size (such as the largest diameter) and shape (spiculated, round, irregular) or about the texture (ground-glass opacity, subsolid, solid, presence of cavitation or necrosis) are typically indicated in the radiology report. Nevertheless, radiomics can provide information about the morphology and the texture of the pathology in an objective and systematic way. This kind of analysis can go beyond the limitations of the human eye in texture discrimination by identifying complex patterns in the images and quantifying the spatial relations among the grey-level intensities [7, 8].

One of the first times that the word "radiomics" was used in the literature was in 2010 by Gillies et al. [9]. From that moment on, the interest in radiomics has been rapidly growing, as shown by the exponential increase in the publication numbers over the last years [10]. Its suffix "-omics" recalls disciplines such as genomics, transcriptomics, proteomics, metabolomics and pathomics, which study the DNA, RNA, proteins, metabolites and digital pathology, respectively, as summarised in **Figure 1.2**.

Radiomics has two main advantages over solid biopsies. First of all, radiomics is a non-invasive technique and does not require any additional exam on the patient, since it works on radiological images that typically are already acquired in the clinical practice for diagnostic and therapeutic purposes. Additionally, considering that images in the clinical routine are often acquired steadily over time to monitor the disease evolution, it is possible to assess the feature changes during the treatment and correlate them with the lesion modifications (*delta-radiomics*) [12]. Secondly, this technique is able to describe the pathology as a whole, by capturing the entire three-dimensional shape and heterogeneity of the lesion.

There are two different ways of formulating radiomics. Both of them start from medical images and try to associate the properties of the image texture to the selected clinical endpoints. The difference lies in how the features are engineered. In the first approach — the traditional one — predefined mathematical descriptors, referred to as *hand-crafted features*, are manually implemented and extracted from the images. The second approach is instead

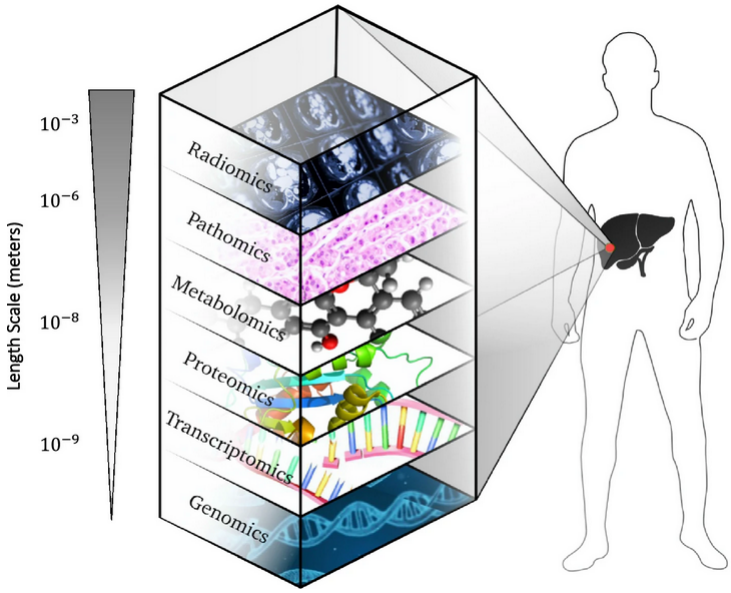


Figure 1.2: A summary of the "-omics" disciplines: from the DNA level (genomics) to radiological images (radiomics). Image from ref. [11].

based on deep learning, which means that the features, referred to as *deep features*, are learned automatically by the algorithm without human intervention in their definition [13]. Therefore, if on the one hand deep features are able to capture the heterogeneity profile of the lesion going beyond human intuition and knowledge, on the other they are more difficult to interpret because of their abstract nature and require a larger amount of labelled data in input.

How can radiomics be a supportive tool for the physicians? Once radiomic data have been collected from a sufficiently wide population of patients, possibly in combination with other information such as the genetic and clinical one, the final goal is to develop models which may improve diagnostic, prognostic, and predictive accuracy. The combination of radiomic data with the genetic analysis is known as *radiogenomics*. A typical objective in a radiomic analysis can be the prediction of the treatment response. Common targets in this sense are the distinction between responder and non-responder, progression-free survival, overall survival, risk of distant metastasis development, and toxicity to normal tissues after radiotherapy. Other clinical outcomes have been investigated as well, such as the classification among stages, correlation with gene mutations and the discrimination among different pathologies [14, 15]. Various anatomical districts and various imaging modalities have been investigated. For instance, radiomics was applied to evaluate the treatment response and overall survival, to distinguish between benign and malignant lesions, and to investigate the association with gene expression in lung [16–22], in prostate [23–25], in breast [26–29], pancreatic [30–35], and in head and neck [16, 36–39] cancer.

Oncology is the most well studied field of application of radiomics, and it

will be the focus of this thesis in its traditional version based on the extraction of hand-crafted features. However, in the literature other fields of application has been investigated as well. For example, in neurodegenerative diseases, like Alzheimer's and Parkinson's diseases, radiomics was used to evaluate the progression of the pathology [40, 41] or to classify the cognitive impairment [42]. Radiomics was also applied to non-oncological pathologies in the thoracic district, like Covid-19 inflammation [43–45], fibrosis [46] or chronic obstructive pulmonary disease [47]. Other fields of interest are, for example, cardiac [48,49] and liver [50] diseases.

However, since radiomics is a novel technique, some aspects have not been completely explored yet. For instance, the relation between the macroscopic appearance of the tumour through images and the underlying biologic indicators is not fully understood. Another criticism is the lack of standardisation in the image acquisition, reconstruction, and processing settings. Indeed, the study of their influence on features are one of the most important topics in radiomic research, and it will be thoroughly investigated in this thesis project. These aspects — with a special attention to CT images — will be introduced in Section 1.3.2, after a detailed description of the radiomic workflow.

1.2 The radiomic workflow

In this section all the necessary phases to perform a radiomic study based on traditional feature extraction are presented, from the image collection to the model development.

The traditional radiomic workflow includes four main steps [51–53]: 1) image acquisition and reconstruction, 2) tumour segmentation, 3) feature extraction and 4) statistical analysis. **Figure 1.3** summaries this chain of operations.

1) Acquisition and reconstruction

Medical images are acquired routinely in the clinical practice for diagnosis, to assess the response to a treatment and for follow-up, by collecting anatomical and/or structural properties of the disease. Depending on the type of pathology to be investigated and the type of information that has to be collected about the disease itself, one or a combination of the following imaging modalities can be adopted: radiography, CT, PET, MRI and US. Medical images are acquired using dedicated acquisition protocols and, in case of tomographic imaging, then reconstructed with proper mathematical algorithms. The CT modality is one of the first and most investigated in radiomic studies. The key aspect of CT images is that their numeric content is associated to a universal scale where the tissue attenuation coefficient is normalised to the water signal. This scale is named Hounsfield unit (HU) scale, and will be discussed more thoroughly in section 1.3.1. Radiomic studies with nuclear medicine and MRI imaging are rapidly spreading too. Since in MRI imaging the signal

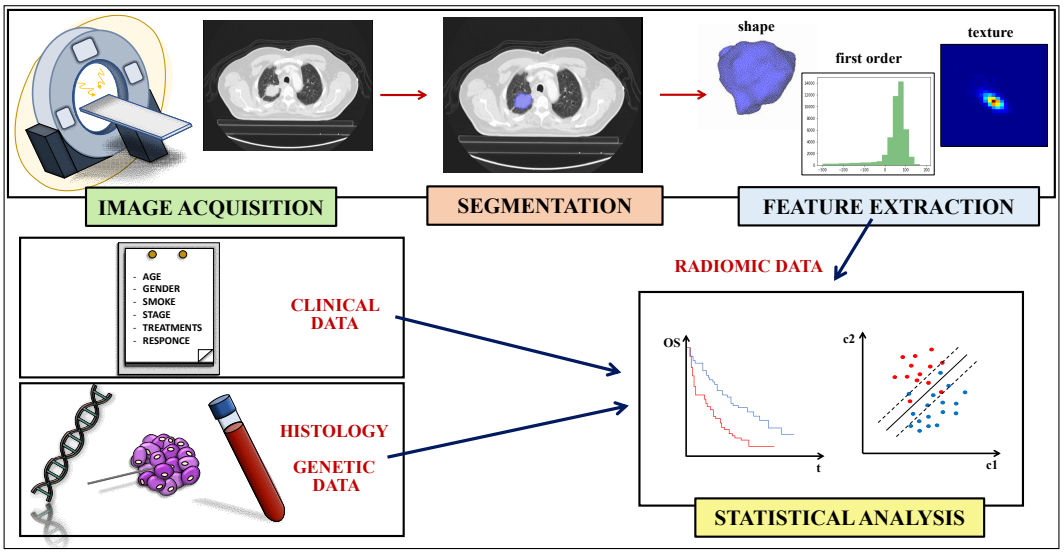


Figure 1.3: Schematic description of the workflow of traditional radiomics. First, medical images are acquired and reconstructed. The lesion inside the image is then segmented using specialised tools. From the identified VOI the radiomic hand-crafted features are extracted with dedicated tools, after the application of processing techniques such as filtering, grey-level discretisation, and resampling. Finally, the radiomic features along with clinical and genetic information — collected from solid or liquid biopsies — are given in input of the statistical model to predict the chosen endpoint.

is not standardised and is arbitrarily assigned, one of the main challenges is the standardisation of the grey-level intensities, in order to make the inter- and intra-patient images comparable. Radiography is used quite commonly in the clinical practice, particularly in screening examination, and it can thus be useful for those studies where a control group is required (for example in the discrimination between malignant and benign lesions). Unfortunately, this modality is not tomographic and the information about the lesion properties in the 3D space is lost. Finally, US imaging is the least involved in the radiomic studies and began to be investigated more deeply only in recent years [54]. Similarly to radiography, it is often used for screening and has the great advantage of being based on non-ionizing radiation, therefore avoiding harmful radiation exposure. However, there are still relevant drawbacks that make its use for radiomic purposes difficult: first, the information comes from a single slice of the lesion and not from the entire volume; second, the positioning of the probe is not reproducible.

Medical images are normally stored in the PACS (Picture archiving and communication system) in DICOM (Digital Imaging and COmmunications in Medicine) format (*.dcm*), which contains the image as a matrix of numbers along with a collection of tags with the information about the patient and the image acquisition.

2) Segmentation

One or more volumes of interest (VOIs) are identified and segmented. The VOI corresponds to the anatomical region whose properties have to be quantified. It could be a lesion, such as a tumour or a benign nodule, or a portion of healthy tissue. The task of segmentation is usually performed manually by one or more physicians, who delineate the borders of the tissue under investigation slice by slice in the axial view (which is normally the one with highest spatial resolution). The contouring of the VOI thus defines two distinct regions: one inside the borders, corresponding to the volume to be investigated, and the background, which is instead excluded from the subsequent analysis. Since this operation is extremely time-consuming and achieving an identical segmentation result twice is in practice impossible (even for the same operator), semi- and fully automatic approaches are being investigated in the recent years.

In some radiomic studies, only one slice of the tumour was analysed, usually the slice associated to the maximal cross-sectional area [55–57]. While in this approach the time spent on segmentation is reduced considerably compared to the segmentation of the entire volume, almost all the information inside the lesion is not taken into account, thus limiting the potentiality of radiomics in capturing the heterogeneity of the lesion as a whole, as previously described.

The segmentation is typically stored in a DICOM-RT Structure Set file, which contains the coordinates of the border points. Unfortunately, the available tool to read these files may produce a different contouring result, since they may perform different interpolation operations on the coordinates. An alternative option is to save the segmentation in a mask form. The mask is a binary matrix of the same size of the reference medical image, containing values equal to 1 in the cells corresponding to the inside of the VOI and 0 outside (the background). Other examples of formats typically used in the medical field to store segmentation as a binary matrix are the Neuroimaging Informatics Technology Initiative (NIfTI) and the Nearly Raw Raster Data (NRRD). Both of them were used in this work. Even though the segmentation masks have the advantage of being univocally interpretable, all the information about the image acquisition/reconstruction that is usually stored in the DICOM header is lost, except few spatial information such as the voxel size.

3) Feature extraction

Radiomic features are extracted from the segmented volume to describe its grey-level pattern with different degrees of complexity: simple descriptors are able to capture the fine details of the shape (*shape* features), others the grey-level distribution of the individual pixels (*firstorder features*), while more complex ones evaluate the spatial relationship between two or more voxels at a fixed distance and direction (*textural features*). Moreover, the features can be extracted after the application of processing operations, such as filtering, resampling of the voxel size and discretisation of the grey-level intensities. Guidelines and recommendations on feature extraction are suggested by the

Image Biomarker Standardization Initiative (IBSI) [1]. This research group reported the nomenclatures and the definition of some features. Moreover, they provided the reference values of these features for different publicly available datasets in order to make the new software IBSI compliant. In this way the results become comparable and can be validated. Details and updates can be found online in the reference manual (https://arxiv.org/abs/2006.05470).

In the following sections a brief description of the most typical processing techniques is given. Image processing techniques are applied before the feature extraction, acting directly on the grey-level pattern of the image, and are often already implemented in the software used for the extraction.

3.1) Image processing operations

Filtering is useful to reduce the noise inside the image or to enhance structures like edges, which are rapid changes in the grey-level intensities. For instance, let us consider a 2D image of size M $\times$ N. The filtered image in output is again a M $\times$ N matrix but with different grey-level intensities. The final effect of the filtering procedure depends on the definition of the filtering operator (or *kernel*), which is a matrix of size k $\times$ l, usually much smaller than the original image size. Various types of filter are conventionally considered in radiomic studies. Most of them act in the spatial domain, meaning that the operation is applied directly on the content of the voxel, and the filtered image F is simply computed as a convolution between the original image I and the kernel matrix h:

$$F(x, y) = (h \otimes I)(x, y). \tag{1.1}$$

The filtering process can be roughly described as the calculation of the filtering response at each pixel, given by the definition of h, followed by shifting the kernel matrix from one pixel to the other until the entire image has been spanned. At each step of this process only the pixels in the proximity of the one that has to be filtered out are considered, and the number of the neighbouring pixels is related to the chosen kernel size.
In the case of linear filters, the convolution can be written as

$$F(x, y) = \sum_{i=-\frac{k-1}{2}}^{\frac{k-1}{2}} \sum_{j=-\frac{l-1}{2}}^{\frac{l-1}{2}} h(i, j)\, I(x + i, y + j). \tag{1.2}$$

The Gaussian filter is an example of smoothing operator which can be used in order to reduce the noise of the image. In the Gaussian filter the kernel weights correspond to a discrete approximation of a normalised Gaussian function with the peak in the centre of the kernel matrix. The final effect depends on the choice of the standard deviation parameter (σ) of the Gaussian profile: a larger value corresponds to a stronger smoothing effect.

In addition to filters based on integral (or sum in the case of a discrete space) operators, there are also filters based on derivatives. This kind of filters facilitates the identification of discontinuities and fine details inside the image and, for this reason, are named *enhancement* filters. The *Laplacian* filter is a second-derivative filter with the peculiarity of being isotropic. In 2D it can be defined as

$$\begin{aligned} \nabla^2 I = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) I \approx & [I(x+1,y) + I(x-1,y) - 2\,I(x,y)] + \\ & + [I(x,y+1) + I(x,y-1) - 2\,I(x,y)] = \\ & = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \otimes\ I(x,y). \end{aligned} \tag{1.3}$$

Finally, the Laplacian of Gaussian (LoG) filter (or Mexican hat kernel) is the combination of a Gaussian and a Laplacian filters in sequence: the Gaussian operation is first applied to remove the noise of the image, followed by the Laplacian to sharpen the edges.

A completely different group of filters is the one acting in the frequency space: the transform-based filters. The Fourier transform is probably the most famous in this category, but it has the disadvantage of losing the spatial information when passing from the spatial domain to the frequency one. It is thus not useful when the frequencies are not constant in space. *Wavelet* filters are designed to overcome this limitation [58]. The crucial components of the wavelet transform are the *wavelets*, which are wave functions $\Psi_{a,b}$ with a limited and variable window. The limited window enables to cover a small section of the signal in space at a time, and it is thus useful to restore the spatial localisation of the frequency information. The size of the window, instead, defines the resolution at which the details can be seen. Using a variable window, details at different scales can be detected and highlighted (multi-resolution analysis). In the continuous space, the wavelet transform is the convolution of the original signal i with the wavelet functions $\Psi_{a,b}$,

$$f(a,b) = \int_{-\infty}^{\infty} i(s)\psi_{a,b}(s)ds. \tag{1.4}$$

The wavelet functions are obtained by contracting/dilating (by a scale factor a) and shifting (by a value of b) a *mother wavelet* Ψ,

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right), \quad a,b \in \mathbb{R}, \quad a \neq 0. \tag{1.5}$$

The mother wavelet has to satisfy a number of analytic requirements, such as to oscillate at least once and to go to zero very quickly as its argument goes to infinity. See ref. [59] for a thorough discussion of the wavelet

transform. For our purposes it suffices to say that smaller values of the scale factor a correspond to more compressed wavelet functions, and consequently to higher space-detail views. In this case, the wavelet transform captures the high frequency signal. Conversely, for large values of a the wavelet transform analyses the low frequency signal.

The wavelet filters are typically applied in their discrete form (*discrete wavelet transform*, DWT). This means that the parameters a and b are not continuous, but they are sampled in a discrete space: $a = a_0^m$ and $b = n\, b_0 a_0^m$ with $m, n \in \mathbb{Z}$, $a_0 > 1$ and $b_0 > 0$.

Working at different scales, the original image is divided into *sub-bands*, each of them obtained by applying a band-pass filter. A 2D image can be decomposed into four components using both a low- (L) and a high- (H) pass filter, separately, along the rows and the columns of the image matrix: LL to approximate the original image, and HL, LH and HH to highlight vertical, horizontal and diagonal details, respectively. In 3D the number of components becomes eight, since the high- and low-pass filters are applied along the three spatial directions: HHH, LLL, HHL, LLH, HLH, LHL, LHH and HLL.
The choice of the discrete wavelet family to be used is unfortunately not univocal. Haar, symlets (or symmetrical wavelets), biorthogonal and coiflets are examples of functions, each with a different shape and properties. Details can be found in [60].

Figure 1.4 and **Figure 1.5** illustrate examples of application of the LoG and the wavelet filters on a CT image of a lung lesion.

Resampling, discretisation and normalisation are instead used to harmonise among them the images of the dataset.
The **resampling** procedure consists in modifying the voxel size to a new fixed value, keeping the total image size constant. The resampling operation is useful to homogenise the voxel size among the CT images of a dataset, since it is a parameter that varies often, particularly in the axial plane. Resampling is usually applied along the three spatial directions or only in the two axial ones, depending on the variability observed in the image dataset for these parameters. Different interpolation functions can be used to achieve this goal, such as the nearest neighbour, linear, and cubic spline interpolation. However, there is no common agreement on the choice of the function for radiomic studies. It is worth noting that this operation has to be applied both on the image and on the segmentation mask in order for them to overlap even after the resampling.
Discretisation (or quantisation) of the grey-level intensities is the grouping of the intensities with similar values into ranges of the same size (*bins*), thus reducing the number of intensities in the image. This technique can be useful to reduce the noise inside the image, but its effect should not be too extreme in order not to remove also some relevant information. There

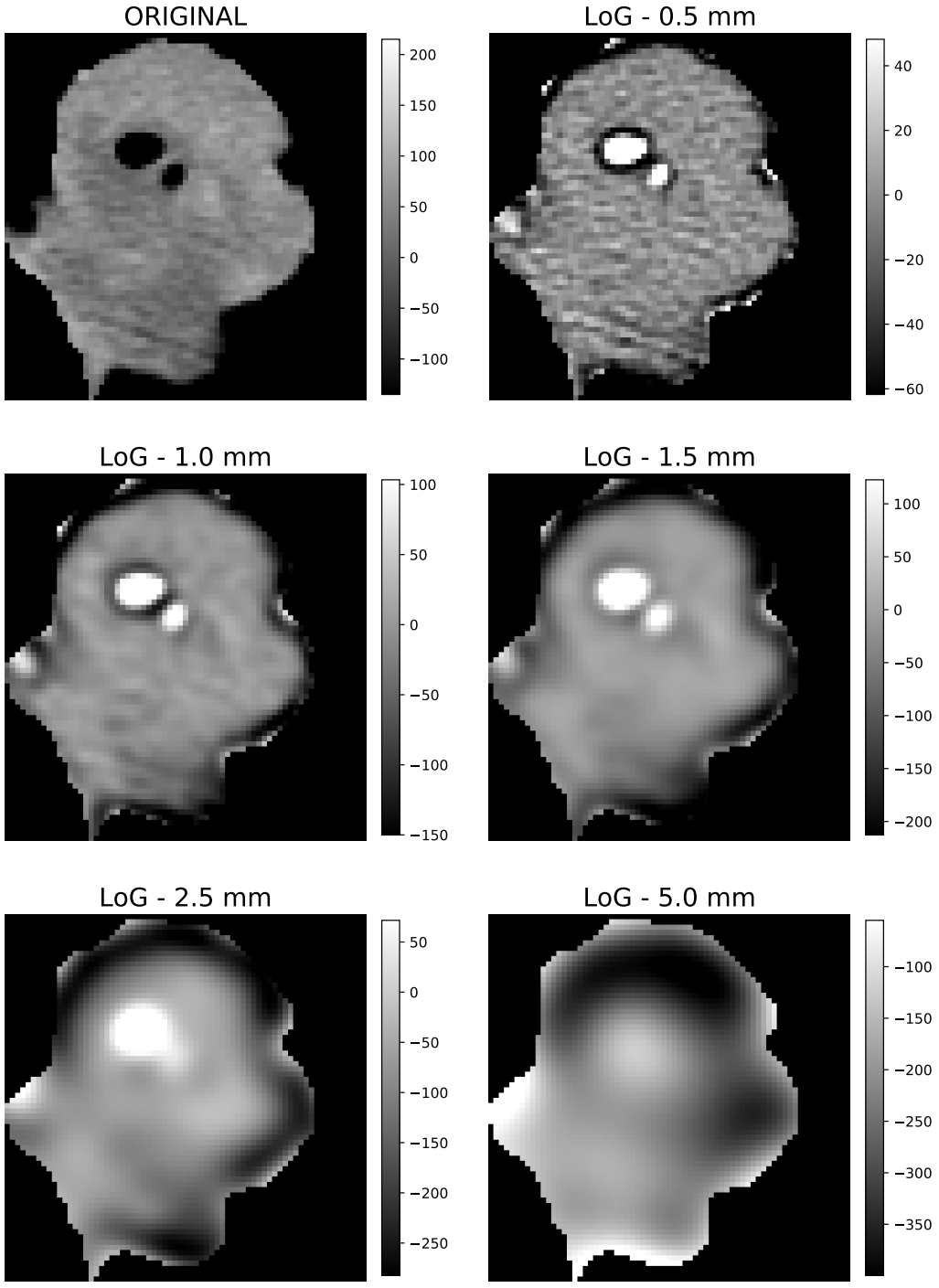


Figure 1.4: Examples of LoG filter application, using a different *sigma* value. Increasing the *sigma* value, the effect of the Gaussian filter can be easily appreciated compared to the original texture, which causes a loss of fine details for larger values. Images are displayed in the range [mean signal - 2 × standard deviation, mean signal + 2 × standard deviation].

are two alternative approaches to perform this operation, by fixing the bin

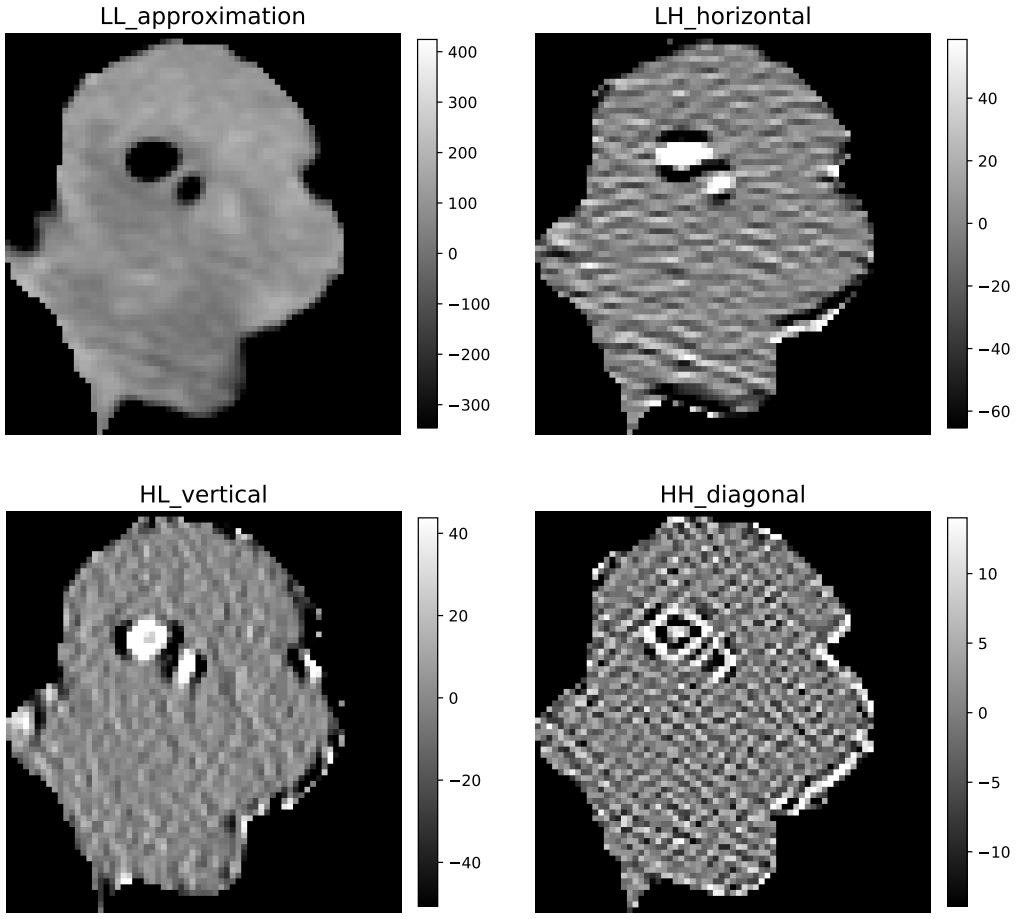


Figure 1.5: Examples of wavelet filter application to a CT image of a lung lesion. The four sub-bands are shown — LL, LH, HL and HH — which correspond to a smoothed form of the original image and to versions with highlighted horizontal, vertical and diagonal details, respectively. The original image can be found in **Figure 1.4**. For the wavelet calculation, we applied the Coif1 function. Images are displayed in the range [mean signal - 2 × the standard deviation, mean signal + 2 × the standard deviation].

width or by fixing the number of bins. For example, a CT image of a NSCLC lesion is show in **Figure 1.6** before and after the discretisation of the grey-level intensities, fixing the bin width to 25 HU.

Finally, intensity **normalisation** is a further processing procedure which can be applied directly on the medical images. It is usually useful in MRI, where images are not directly comparable for the lack of a standard intensity scale. Various normalisation techniques have been applied to medical images, such as the histogram-matching normalisation or the scaling-shifting method [61–64]. At the moment, finding the best algorithm able to make the images comparable without losing any of the relevant information is still an open field of research. For the purposes of this study, we did not apply any normalisation procedure.

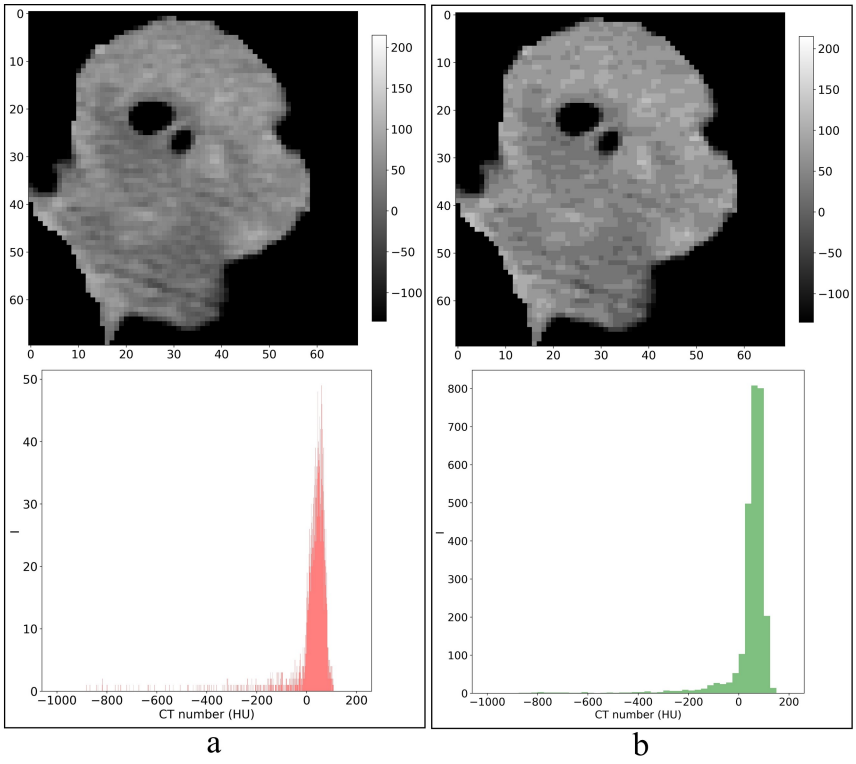


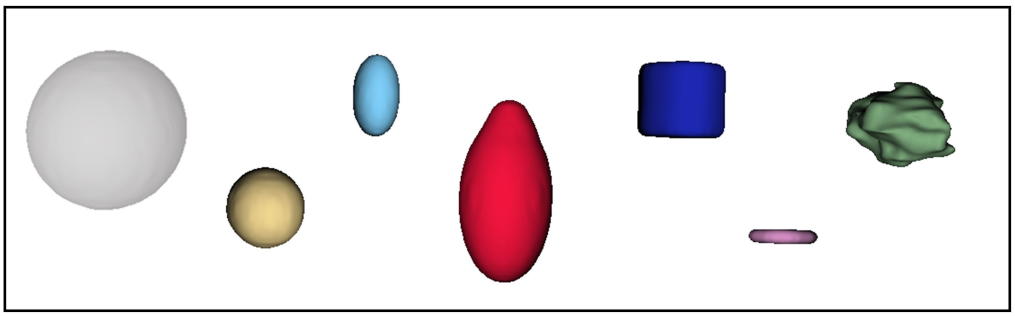
Figure 1.6: Example of the application of the discretisation of the grey-level intensities. The figure illustrates on the top a CT axial slice of a lung tumour before (**Figure 1.6 a**) and after (**Figure 1.6 b**) the discretisation process. This operation clusters the intensities into groups with a bin width equal to 25 HU. The two histograms on the bottom display the distribution of the intensities inside the lesion. The histogram in (**Figure 1.6 a**) is displayed with a bin width of 1 HU, while the one in (**Figure 1.6 b**) has a bin width of 25 HU, and they refer to the image before and after the discretisation, respectively. A reduction of the noise can be observed after the discretisation, but it is still possible to identify the heterogeneous texture inside the lesion.

3.2) Feature categories

The quantitative descriptors which can be extracted from the identified VOI can be divided into three main categories: morphological, first- order or histogram-based, and textural features.

Morphological features describe the shape of the VOI, capturing its volumetric and superficial properties. Some simple examples of such descriptors are the volume size, the diameter length, the sphericity and the elongation of the VOI. **Figure 1.7** shows some examples of morphological feature values for different shapes and sizes.

Histogram-based (or first-order) features, instead, quantify the characteristics of the histogram of the grey-level intensities inside the VOI. Thanks to their simplicity, first-order features can be calculated easily, and their meaning can be interpreted quite intuitively. Some examples are the mean and the variance of the intensities, as well as the skewness (asymmetry of



Feature Name	VoxelVolume (mm^3)	SurfaceVolume Ratio	Sphericity	Elongation	Flatness
Big sphere (gray)	45823.91	0.15	0.90	1.00	1.00
Small sphere (yellow)	2283.44	0.42	0.88	0.99	0.98
Ellipsoid (light blue)	1208.77	0.52	0.87	0.57	0.57
Assymetric Ellipsoid (red)	20639.95	0.20	0.86	0.53	0.52
Big cylinder (blue)	5857.22	0.31	0.86	1.00	0.98
Small cylinder (pink)	1255.12	0.72	0.63	1.00	0.20
Irregular shape (green)	12787.88	0.31	0.66	0.70	0.56

Figure 1.7: Examples of morphological features extracted from different geometrical shapes. The features were computed with the Pyradiomics package (v. 2.1.2) without any image processing. *VoxelVolume* is the volume of the 3D object computed from the number of voxels inside. The *SurfaceVolumeRatio* is the ratio between the surface and the volume of the object: the smaller it is, the more compact is the shape. *Sphericity*, *Elongation* and *Flatness* describe the characteristics of the object shape: the higher is the value of the sphericity, the more compact (similar to a sphere) is the shape; the values of the elongation and flatness features are between 0 and 1, where 1 is for non-elongated shape and non-flat (sphere-like), and 0 for a line and single-slice, respectively.

the intensity distribution) and the kurtosis (tendency to outlier of the intensity distribution). This group of features however takes each single voxel separately, this way neglecting its surroundings.

Textural features are higher-order features that, in contrast to the first-order ones, take into account the spatial relationship between a reference pixel and its neighbours. The order identifies the degree of relation between the investigated voxels: for second-order features the relationship is evaluated between two voxels at once, while for order higher than two it is computed among more than two voxels. Textural features are calculated along a fixed direction (θ) and at a predefined distance (d) in voxel number from specific matrices built from the grey-level intensities inside the VOI. They are 2D descriptors if they are extracted from a 2D slice, or 3D (or 2.5D) if they are obtained from a 3D volume. 2.5D means that the feature is calculated on each axial slice in 2D, and then averaged among all the slices in the VOI. According to the IBSI recommendations, the 3D extraction modality should be used only in the case of isotropic voxels. Alternatively, the 2.5D method can be adopted. Since these features consider a more complex relationship among voxels, it is much more difficult for the

Matrix name	Acronym
Gray-Level Co-occurrence Matrix [65]	glcm
Gray-Level Run Lenght Matrix [66]	glrlm
Gray-Level Size Zone Matrix [67]	glszm
Gray-Level Dependence Matrix [68]	gldm
Neighbourhood Gray-Tone Difference Matrix [69]	ngtdm

Table 1.1: List of the matrices from which the textural features are extracted.

human eye to identify the corresponding textural properties and hence to provide a straightforward interpretation.
The matrices proposed for radiomic studies are: Gray-Level Co-occurrence Matrix (glcm), Gray-Level Run Lenght Matrix (glrlm), Gray-Level Size Zone Matrix (glszm), Gray-Level Dependence Matrix (gldm) and Neighbourhood Gray-Tone Difference Matrix (ngtdm). Their definitions, together with useful references, are collected in **Table 1.1**.
Features of the second order are extracted from the glcm, while those of higher order from the other matrices. Each element of the glcm ($glcm(i,j)$) counts how many times the voxel with intensity i and the voxel with intensity j appear at a distance d along a direction θ. It should be noted that d must be greater than 0, since for $d = 0$ the matrix is equivalent to computing a histogram of grey-level intensities. The glrlm, instead, evaluates the number of *runs* in the VOI. A run is an array of contiguous and collinear voxels with the same intensity i. Therefore, each entry of the glrlm ($glrlm(i,j)$) quantifies the number of runs of intensity i and length j inside the VOI. Glszm and gldm are very similar to the glrlm, but instead of the run they count the number of *zones* and *dependencies*, respectively. A zone is a connected area of voxels with the same intensity i and with size j. The dependency, instead, is the number of voxels with intensity j, that are both connected and "dependent" to the reference voxel with intensity i within a distance d. "Dependent" means that the following relation has to exist between the two voxels: $|i - j| \leq \alpha$, where α is a parameter set by the user. Finally, the ngtdm is a 1D array built by computing at each voxel the difference between its intensity i and the average intensity of the neighbouring voxels. The element $ngtdm(i)$ is given by the sum of these differences for all the pixels with intensity i.
The features extracted from the glcm and glrlm matrices can be computed for each direction separately, or a unique feature value can be computed by averaging the outputs among a chosen set of directions. The directions are typically 4 when the neighbours are identified in the 2D slice, and 13 for the 3D extraction.

There exist many algorithms to extract the features. In many studies, in-house tools were developed but not always made publicly available, this way

making the validation and replication of the results impossible. Fortunately, several programs have been developed and made available for other users in the field of radiomics. Some of them are free, such as MaZda [70], IBEX [71], LIFEx [72] and Pyradiomics [73]. Others are commercial, such as RadiomiX Research Toolbox (OncoRadiomics SA, Liège, Belgium) and TexRad (TexRAD Ltd, www.texrad.com, part of Feedback Plc, Cambridge, UK).

4) Statistical analysis

The final step of the workflow is the feature analysis, where the radiomic data are correlated to additional information — such as clinical, histological and genetic information — and a model is developed to predict the clinical outcome under investigation. The statistical model learns from a set of data, named for this reason training dataset, and then infers the hypothesis on an unseen dataset to evaluate the performance of the prediction [74].
To achieve this goal it is first necessary to perform *feature selection*: only the non-redundant and robust features should be included for the model development. Features that are associated with the acquisition/reconstruction parameters should be rejected because they can introduce a spurious signal which may distort the results of the model. Section 1.3.2 in this chapter will describe in more detail the issue of the robustness of the radiomic features. As regards the feature redundancy, excluding the not useful or correlated features is extremely important, since their inclusion in the model may increase its complexity without adding any relevant information. Possible selection techniques are correlation analysis, filter-based models, the wrapper models, or the embedded ones, such as the Least Absolute Shrinkage and Selection Operator (LASSO) model [75]. Among the filter-based algorithms it is worth noting the RELevance In Estimating Features (RELIEF) approach or those based on mutual information as the Minimum Redundancy and Maximum Relevance (mRMR) algorithm.

The next operation is the *model building* and the identification of the group or combination of features — called the *radiomic signature* — which are clinically-relevant for the specific endpoint of the study. Different machine learning algorithms can be adopted for this purpose, such as the support vector machine, linear or logistic regression, random forest, decision tree and neural network. Logistic regression is the most commonly used for classification in the medical field, for its simplicity in the training and in the interpretability of the results. Linear regression is the equivalent algorithm for continuous endpoints (e.g. overall survival). The LASSO algorithm is a linear regression model that performs also a feature selection. The simplicity of linear and logistic regressions has however the drawback that they may not perform well on complex data.

Two aspects that should be taken into account during the development of the model are the *overfitting* and the *class imbalance*. Overfitting occurs when the model memorises the training data too well (together with the eventual

noise inside) and is not able to generalise to a new dataset. The larger is the number of features compared to the input data (number of patients), the more likely it is to find erroneous correlations, and the performance of the model therefore gets worse. Imbalance, instead, is observed when one class is more represented than the others in terms of numerosity of the input data. One approach to reduce the imbalance is to simulate or inputs new data, as it is done for example in the Synthetic Minority Over-sampling TEchnique (SMOTE) [76], where new samples are obtained from the k-nearest neighbours of a randomly chosen element of the minority class.

Once the model is built on the input dataset, it should be *validated* on an independent cohort (for example on a dataset acquired in a different institute with different protocols) to verify whether the features are really storing a predictive value and that the model is not overfitting the input data. If the external validation is not feasible, an alternative approach is the split of the entire dataset into two groups, one for the training and the other for the validation. Otherwise, a k-fold cross-validation [77] can be applied, particularly when the input dataset has a small sample size. In this technique, the training set in randomly divided into k sub-datasets of the same size: $k-1$ of these subsets are then used for the training, and the remaining one is used to test the model performance. This procedure of training and testing is repeated k times, changing at each time the subset used for the training and testing from the input dataset. A final estimation of the model is given by averaging the performance results of the k repeated steps.

1.3 Radiomics in computed Tomography

1.3.1 A brief introduction to CT imaging

A CT image is a volumetric representation of tissue density. An X-rays tube coupled with a detector row rotates in the x–y plane (the axial plane) all around the table where the object of interest is laying, and at the same time the table translates along the z-axis. When acquiring in helical mode, the tube follows a spiral movement around the table, collecting the transmission profiles of the scanned object from different views. These profiles are then reconstructed using proper algorithms to recover the spatial information about the tissue density. Each CT slice corresponds to a x–y plane (cross-sectional image) at a certain point along the z axis direction (longitudinal to the table direction).

A fan beam of photons is generated by the impact of an electron current against an anode. The electrons are produced through thermionic emission by heating a filament (cathode) and are then accelerated by a potential difference between cathode and anode. When the electrons hit the cathode, they transfer their kinetic energy to the atoms of the target, producing heat and photons. Both characteristic X-rays and bremsstrahlung photons are produced and constitute the photon spectrum in output. The maximum energy of the photon

spectrum corresponds to the kinetic energy of the electrons (*end point energy*), which is commonly referred to as kilovoltage peak, kVp.

The signal in a CT image is given by a collection of voxels with various grey-level intensities, which corresponds to the **linear attenuation coefficient** (μ) of the tissues hit by the beam. The primary beam of intensity I_0 is attenuated exponentially following the Beer-Lambert law for a monoenergetic beam. Since the beam may cross various materials with different attenuation properties μ is a function of the position (x) inside the material. The intensity attenuation I for a polyenergetic beam is given by

$$I = \int_0^{E_{max}} I_0(E)\, e^{-\int_0^d \mu(E,x)\, dx}\, dE\,, \tag{1.6}$$

where d is the total thickness crossed by the beam. The beam intensity before (I_0) and after (I) the absorption is known from measurements, while the matrix of the attenuation coefficients, and therefore the CT images itself, can be recovered through dedicated reconstruction algorithms.

Strictly speaking, the numerical content of a CT image is given by the *CT number* which is derived from the μ of the tissue, normalised by the linear attenuation coefficient in water (μ_{water}) and in air (μ_{air}) at standard pressure and temperature:

$$CTnumber = \frac{(\mu - \mu_{water}) \times 1000}{\mu_{water} - \mu_{air}}\,. \tag{1.7}$$

Its value is equal to 0 HU for distilled water and to -1000 HU for air. Therefore, low values of the CT number correspond to less dense materials, which appear darker in the CT image. Brighter areas, instead, are indicative of denser tissues. For this reason, lungs appear black and bones white.

The transmitted photons are collected by the detectors placed on the opposite side of the tube. Conventionally, the CT device is equipped with multiple rows of detectors next to each other along the z axis (multiple-slice CT), with anti-scatter collimators to protect the detectors from scatter radiation. The width of each detector row determines the tissue width sampled during the acquisition. For each row and at each rotational angle θ, the transmitted radiation is detected and converted to the attenuation coefficients, by inverting eq. (1.6). These attenuation coefficients are the sum of the attenuation coefficients of all the tissues crossed by the radiation at that angle and are named *projections* (see **Figure 1.8**). The various projections are stored in the form of a sinogram (alias raw data). The sinogram $p(r, \theta)$ gives the projection at each angle θ and detector position r. The 3D image can then be reconstructed from the sinogram using ad-hoc algorithms. The two main algorithms used for this purpose are the Filtered Backprojection (FBP) and the Iterative Reconstruction (IR) algorithms [78, 79].

The FBP algorithm consists of two mathematical operations: the Radon and the Fourier transform. The Radon transform $\mathcal{R}$ maps the image in the

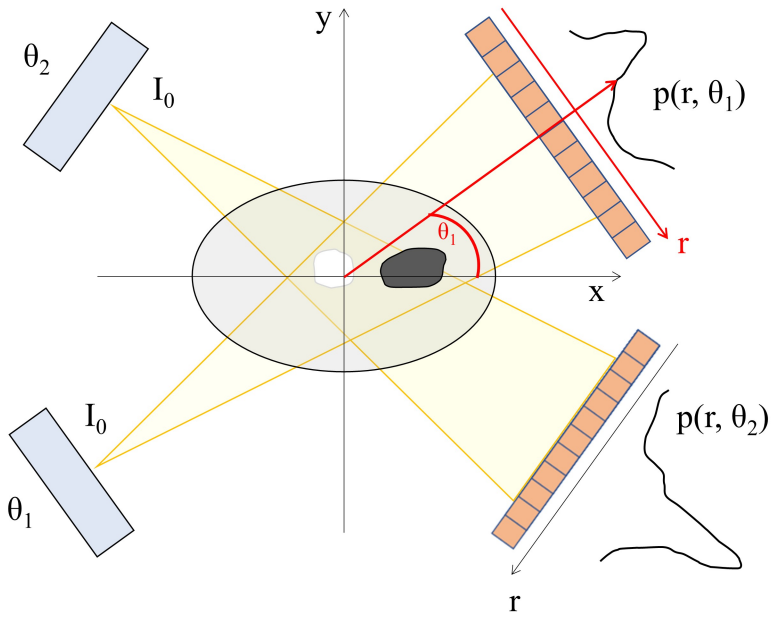


Figure 1.8: Illustrative example of projections $p(r, \theta)$ collected by the detectors at two different rotation angles around the scanned object, while r is the position along the detector, perpendicular to the beam direction.

object space f to the sinogram in the projection space (named the Radon space),

$$p(r, \theta) = \mathcal{R}(r, \theta)\,[f(x, y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)\,\delta(r - x\,\cos\theta - y\,\sin\theta)\,dx\,dy\,, \tag{1.8}$$

where (x, y) are the Cartesian coordinates on the 2D slice, and δ is Dirac's delta function. A Fourier transform is then performed in the Radon space to apply suitable kernel functions in the frequency space. The kernel is chosen so as to filter out the noise in the data or to increase the spatial resolution and/or the contrast resolution, according to the desired final result. With an inverse Fourier transform, the system goes back to the Radon space, where the projections are now filtered. Finally, the filtered projections are "back projected" into the CT image domain with the inverse Radon transform to obtain the spatial map of the attenuation coefficients, which is the reconstructed image.

The IR algorithm was introduced more recently in the image-reconstruction field, partly thanks to the increase in computational power. The main advantage of this technique compared to the FBP algorithm is the noise reduction. The reconstruction is an iterative procedure. At each iteration, an artificial image is generated, and its sinogram is compared against the measured one. The difference between the two defines a loss function which is minimised in the iteration. The iteration is initiated with an arbitrary prior image, which constitutes a snapshot of all the information on the image in the real space that is available immediately after the acquisition in the Radon space. For instance,

the prior image may be just white, meaning that we assume no knowledge of the image in the real space, or the result of the FBP reconstruction. Clearly the choice of the prior image is crucial, since the more similar it is to the real image, the faster the algorithm converges. This iteration stops when the difference between the real and the computed projections falls under a chosen threshold. Alternatively, a fixed number of iterations can be set to stop the iterations.

The reconstruction algorithm considered in this thesis is the Adaptive Statistical Iterative Reconstruction (ASiR), implemented in the GE scanners (GE Healthcare, Waukesha, WI). The starting point of this algorithm is the FBP reconstructed image. The percentage of FBP involved for the final result is named *blending level*, and it goes from 0% (pure FBP) to 100% (pure IR).

Parameters such as the X-rays tube voltage and current, the pitch, the voxel size (field of view and slice thickness) and the reconstruction algorithm may affect the signal-to-noise ratio and therefore may have an impact on the image texture. More in detail, the current and the voltage are two parameters that affect the image quality and the absorbed radiation dose. Higher voltage means more penetrating photons, while higher current values correspond to more photons generated at the anode: in both these conditions a larger number of photons is detected and the signal-to-noise ratio increases, but at the same time also the patient radiation exposure increases. Optimising these parameters based on patient size and weight is nowadays a diffused approach to balance the noise and the radiation dose. The Automatic Tube Current Modulation, for example, is a system used to change the tube current according to the anatomical region scanned: in regions with lower attenuation the current is reduced, while it is increased where the tissues attenuate more the radiation. The pitch is the ratio of the table shift during a 360° rotation to the beam width: larger values are associated to less counting statistics (more noise). Finally, the voxel size is given by its dimension in the axial plane (reconstructed field of view/matrix size) and the slice thickness (along the z axis). Larger thickness values and larger axial sampling are associate to a higher signal and thus to a less noise, but to a worse spatial resolution.

1.3.2 Feature robustness

The main weakness of radiomics is the lack of robustness of some features against various factors: each step of the radiomic workflow, in fact, may introduce variability that may have relevant effects on the feature behaviour. This limitation of radiomics has been known since its origin [80] and several methodological studies have been performed to quantify the degree of influence of these parameters on the texture, with the final aim of avoiding misleading interpretation of the model results [14, 81–83]. A high level of standardisation during image acquisition and processing would be the ideal condition to foster the radiomic research. This may be a major requirement when design-

ing a prospective study and when implementing novel tools, but it is hardly achievable in retrospective studies, be they monocentric and even more when multiple institutions are involved. These latter scenarios are however extremely frequent for different reasons: first, because retrospective studies are generally preferred to test the clinical hypothesis and thus justify the efforts of dedicated prospective studies; second, because retrospective and multi-centre studies are often the only practicable solution for collecting the large amount of data required to construct reliable models. For this reason, it is very common to face image databases that are intrinsically heterogeneous due to the use of different scanners, acquisition protocols and post-acquisition techniques, including reconstruction algorithms and settings. In these cases, dedicated efforts must be oriented towards the identification of radiomic features which remain stable with respect to the specific database heterogeneity. Conversely, features whose value is affected importantly by the use of different image acquisition settings should be either rejected — to avoid the risk of introducing confounding elements during data analysis — or, whenever possible, corrected and harmonised in order to make the values obtained from different images comparable.

Two important aspects should be assessed to evaluate feature robustness: **repeatability** and **reproducibility**. The repeatability is a measure of precision among measurements repeated multiple times on the same object by the same operator with the same procedure and the same experimental apparatus in a short time. Reproducibility, instead, is a measure of precision among repeated measurements, where the measurements and/or the investigated object change between a repetition and the other [84]. This kind of investigation can be performed both with phantoms and with dataset of patients.

Methodological investigations using directly the patient images are useful for radiomic studies because these images are more representative of the intra- and inter-heterogeneity and of the biological characteristics of the tissue than phantoms. Patient databases have to be properly collected, meaning that they have to include an adequate number of patients, they have to be clinically compatible with the pathology of interest and all the parameters to be investigated have to be well represented. However, this approach has usually the drawback of not providing repeated acquisitions for each patient. In CT imaging, the RIDER (Reference Image Database to Evaluate Therapy Response) dataset [85] has been often used for the evaluation of feature stability [16, 86–89]. This dataset is publicly available online at The Cancer Imaging Archive (TCIA) repository [90] (https://wiki.cancerimagingarchive.net/display /Public/RIDER+Collections), and it consists of *test-retest* CT images of 31 patients with NSCLC. Each patient underwent two CT examinations 15 minutes apart on the same scanner with the same protocol. Unfortunately, the patients got up from the table between the two acquisitions, introducing inevitably a variability in the positioning. This change between scans collides the hypotheses defining the repeatability conditions.

Different CT acquisition parameters have been analysed on patient im-

ages for their influence on the radiomic features, including tube current [56], slice thickness [91–94], and tube voltage peak [56]. The reproducibility of the features related to the use of contrast medium [95, 96] and to different reconstruction techniques was investigated as well [91, 92, 94, 97–100]. Finally, a few studies in the literature worked on the impact on features of the software for the extraction [101] and of post-processing techniques, such as the grey-level binning, resampling and filters [102–105]. In this sense, Fornacon-Wood et al. [101] compared different extraction platforms — CERR, Pyradiomics, LIFEx and IBEX (only the last one was not IBSI-compliant) — and they obtained a good agreement among the features extracted from the three tools following the IBSI indications, but only after a proper match of the extraction settings. This means that the default values do not correspond among the software and that they should be properly set if the results from different platforms have to be compared.

Using phantoms can compensate the limitations of patient dataset and can give complementary information for the robustness analysis. Phantoms, in fact, are inanimate objects that are built to reproduce some properties of human body, such as shape and tissue attenuation, mainly for quality assurance purposes. The main advantage of phantoms is that they can be used for multiple acquisitions in fixed conditions of measurement or by changing these conditions in a controlled way. Testing the robustness and the reliability of the radiomic features with phantoms is therefore extremely convenient, since they allow us to bypass the typical obstacles which we encounter with patients: radiation exposure, movement artefacts and patient discomfort due to multiple acquisitions. These peculiarities have been exploited in radiomics to investigate both repeatability and reproducibility of the features.

The most used phantom for radiomic purposes is the Credence Cartridge Radiomics (CCR) phantom, presented for the first time in ref. [106]. It consists of ten cartridges, each of them with a size of $10.1 \times 10.1 \times 3.2$ cm^3 and fabricated with various materials to create a wide range of image texture (20%, 30%, 40%, and 50% acrylonitrile butadiene styrene, natural wood, standard and high-density cork, rubber of tires, zp®150 powder bonded with Colorbond^{TM} and solid polymethyl methacrylate). This phantom was used to compare different scanners by Mackin et al. [106] and the CT images of this study are available on the TCIA platform (https://wiki.cancerimagingarchive.net/display/Public/Credence+Cartridge+ Radiomics+Phantom+CT+Scans). The main finding of this analysis was that the variation of the features among different CT vendors is comparable with the variability of the same descriptors in a cohort of NSCLC patients, indicating a non-negligible impact of the CT scanner. The effect of the tube current was also investigated using the CCR in ref. [107]. They changed the Exposure Time Product (expressed in mAs) between 25 and 300 mAs and did not observe a relevant impact on features for the materials more similar to the real tissues (cork and rubber).
The CCR phantom was also adopted to study the impact of post-processing

operations. For example, Shafiq-ul-Hassan et al. [108] evaluated both the impact of different voxel size before and after the resampling to a uniform value and of the grey-level discretisation, and proposed correction in feature definition to reduce their variability. Larue et al. [109] considered for their analysis multiple parameters, both acquisition ones (scanner, slice thickness and exposure) and processing ones (bin width and resampling). The scanner and the slice thickness are those parameters that impact the most on the features, but for the latter a reduction in feature variability was observed after resampling, suggesting linear and cubic interpolators over the nearest neighbour one. Finally, they performed a test-retest analysis without changing anything and found that some of them are not repeatable and recommended to eliminate them. Furthermore, Mackin et al. [110] considered both patient images with lung cancer and the CCR phantom to evaluate the impact of resampling and filtering. They showed that in both the cases, the application of the two operations in sequence increased the feature robustness.
More recently, Ger et al. [111] proposed an updated version of the CCR phantom to compare features among protocols adopted on different CT scanners in different centres. This new version was made of six round cartridges with a diameter of 10.8 cm (50% + 25% acrylic beads + 25% polyvinyl chloride, 50% acrylonitrile butadiene styrene + 50% PVC, 50% acrylonitrile butadiene styrene and 50% acrylic beads, hempseeds in polyurethane, rubber, cork), embedded in a polystyrene body ($28 \times 21 \times 22$ cm^3).
Besides the CCR phantom, other types of phantoms, more or less anthropomorphic, have been fabricated specifically for radiomic study to investigate feature repeatability, to compare scanners and protocols, and to evaluate the impact of post-processing [112–114]. This topic will be further investigated in Chapter 3.

In order to reduce the variability among different acquisition protocols, new approaches have been proposed in the recent literature. The ComBat method, for instance, was proposed to harmonise heterogeneous databases by acting directly on the feature value [115–118]. It was developed in genomics to remove the "batch effects" (non-biological variables that impact on the data) in gene expression microarray analysis [119], and it was investigated in radiomics to correct site effects in CT [120, 121], MRI [122] and PET [123] images. Other research groups, instead, adopted different strategies. They applied deep learning-based algorithms to generate synthetic harmonise images before the feature extraction. For example, a convolutional neural network was developed by Park and colleagues [93] to generate CT images with a lower slice thickness compared to the original one, and by Choe et al. [124] and Yoon et al. [125] to convert the CT image into a new one reproducing the effect of a different reconstruction kernel without using the sinogram.

The segmentation procedure was also investigated as a possible confounding factor. Manual segmentation is prone to inter- and intra-variability, and semi-automatic approaches can impact on the feature value because of the sub-

jective human intervention [126–130]. Huang et al. [131] evaluated the impact of the inter-reader variability on the model performance in 46 patients with NSCLC undergoing chemotherapy to predict the mutational status. Three radiologists with different experience contoured independently the lesion with a semi-automatic tool and they found a significant difference in the accuracy of the model due to the variability introduced by the different segmentations. Automatic software base on deep learning algorithms have been recently introduced in the radiomic analysis [132, 133], but further investigations on their output reliability should be assessed against physician segmentation. More details on this topic can be found in Chapter 4.

In general, the slice thickness and the reconstruction algorithms are the most studied parameters — particularly in patients —, since they can be varied *a posteriori* if the raw data are available, without the need of additional scans of the patient. Moreover, many studies found that these parameters have the strongest impact on the feature stability.

Finally, the impact of the application of different selection methods and machine learning algorithms on the model performance was also evaluated [134–140]. There is not a consensus in the choice of the best methodology for the model development, even when similar cohort of patients and the same outcome are selected for the study [134, 137]. Corso et al. [140], in fact, showed as the performance of the models depends on the characteristics of the dataset, such as the size, the association strength to the outcome and the class imbalance. Considering radiomic features simulated from those extracted from the CT images of 270 NSCLC patients, Random Forest and Extreme Gradient Boosting achieved the best performance to find a correlation between features and lymph node status in all the investigated characteristics of the dataset.

In conclusion, several parameters that may impact on image texture have been investigated in the literature in the last years. Methodological studies — developed for the anatomical district of interest and according to the dataset complexity – are a useful and necessary tool to identify possible sources of noise and confounding factors. By properly removing or correcting unstable features may help in avoiding the introduction of bias in the analysis. However, a standard procedure in development of this kind of study has not been established yet. The variability observed in this methodological investigation makes difficult the comparison and thereby the generalisation of the results. In addition, papers in the literature sometimes do not report all the details about how the study was carried out, from the feature extraction to the analysis, and this lack of information makes even more complicate the replication and validation of the results [141–143]. In order to identify a general consensus in the evaluation of feature stability, a greater international collaboration should be mandatory. Common guidelines, in fact, should be assessed, taking as reference model the IBSI work in the standardisation of feature definition and, for the next future, of filtering operations [144].

1.3.3 State of the art in radiomics of lung oncology

Cancer is one of the leading cause of dead in the word, along with cardiovascular diseases. According to the World Health Organization (WHO) in 2019 the 16.8% and the 23.0% of the deaths were due to malignant tumours in the world and in Europe, respectively (https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death). The estimates of cancer incidence and mortality provided by The International Agency for Research on Cancer (https://gco.iarc.fr/) for 2020 indicated that lung tumours were the most frequent type of cancer in the male population worldwide (14.3%) and the second one in Europe (13.2%) after prostate tumours (20.2%). Moreover, mortality due to lung cancer is the highest among males (21.5% worldwide, 24.0% in Europe). In the female population, instead, lung cancer is the second tumoral cause of death (13.7% worldwide, 14.3% in Europe) after breast cancer (15.5% worldwide, 16.3% in Europe), and the third tumour (8.4% worldwide, 7.9% in Europe) for incidence after breast (24.5% worldwide, 25.8% in Europe) and colorectum tumours (9.4% worldwide, 11.6% in Europe) [145].

Non-Small Cell Lung Cancer (NSCLC) is the most commonly diagnosed lung tumour, identified in about the 84% of the lung cases according to the American Cancer Society (https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html), and it encompasses various sub-types of tumours, the most frequent are adenocarcinoma and squamous cell carcinoma.

CT and PET images are conventionally used to diagnose and stage lung tumours, before the start of the treatment. The clinical stage of the disease is the main factor in the definition of the treatment options, along with presence of comorbidity. For early-stage lung cancer (I and II), surgery is the main choice, sometimes followed by adjuvant chemotherapy to remove cancer cells that may have left and therefore reduce the risk of recurrence. Alternately, radiotherapy on the lung lesion can be performed. For higher stages, a combination of treatments is usually proposed (surgery, chemotherapy and radiotherapy) for stage III, while a systemic therapy (conventional chemotherapy, target therapy and immunotherapy) is adopted for stage IV patients [146, 147]

Furthermore, information from histologic and genetic examinations through solid biopsies are essential to assess the therapeutic path of the individual patient (*personalised treatment*). The choice of a chemotherapy drug, for instance, varies according to the type of NSCLC and the type of genetic mutations [148]. Epidermal growth factor (EGFR), anaplastic lymphoma kinase (ALK), BRAF and ROS1 mutations are examples of oncogenic drivers that can be inhibited by specific molecular targeted agents [149, 150].

However, some challenges still remain in the development of a personalised and efficient treatment. For example, solid biopsies are invasive and not able to capture the entire tumour heterogeneity, as previously described in this chapter. Moreover, traditional image features describing lesion shape and appearance are prone to intra- and inter-observer variability. Radiomics can be

useful to help the physicians to overcome these limitations by extracting additional information from the whole lesion in an objective manner. In this direction, different tasks have been recently faced in the field of lung oncology: classification of lesion malignancy [151], prediction of the treatment response or of the overall survival [152–154]. Another field of analysis is the identification of a correlation between radiomic features and genetic mutations [155]. The aim of this type of analysis is to understand whether medical images can be used as a "digital biopsy", thus overcoming the limitations of real biopsies. Several radiomic studies were performed to predict the mutational status using the data collected from the histopathological samples [156–159]. Furthermore, more recently researchers have introduced in the radiomic models the information provided by the liquid biopsy [160, 161]. In this way the radiomic features can be compared to mutational data which are able to give a more complete description of the lesion and can be collected with less invasiveness during the imaging follow-up. However, deeper investigation is required in this field since the number of investigated patients is still low.

Various studies showed that the performance of the predictive models increases when the radiomic features are added to clinical, genetic, and traditional radiographic information. An increase in the prediction capability (higher area under the curve, AUC) was reported by Choe et al. [162] in the identification of the mutational status (ALK, EGFR and wild-type) from CT images of 503 patients with lung adenocarcinoma when both radiomic and clinical (gender, age, smoke, stage and tumour size) data are used. Similarly, in another study designed to distinguish ALK mutated from ALK wild-type in 335 lung adenocarcinoma patients [163] the integration of clinical (gender, age, smoke, stage, distant metastasis and adenocarcinoma sub-type), conventional (such as maximum diameter, lobe, location, calcification, density, lymph node status and cavity) and radiomic features gave better results (higher AUC) than the model based only on conventional or radiomic features. Tang et al. [164] found enhanced performance (higher mean concordance index) in the combination of clinical (gender, age, smoke, stage and first-line chemotherapy type) and radiomic information for the prediction of the prognosis in patients with NSCLC patients with EGFR mutation, treated with target therapy. Moreover, they obtained that a model based only on conventional CT features (such as lobulation, spiculation, pleural effusion and vascular invasion) were not associated to the investigated outcome. Another interesting result they obtained was that radiomic features extracted from contrast-enhanced CT images had a greater predictive power than those extracted from unenhanced ones. Nevertheless, there are other studies where significant improvements when adding clinical information to the radiomic model are not observed [165, 166]. These contrasting results may be due to the different investigated outcome or to the different study design (i.e. the choice of the predictive model), stressing the need for further standardised studies.

All these results emphasise the potentialities of radiomics as a support

tool in making decisions before and during the treatment. However, despite the increasing number of studies in lung radiomics, a greater validation and standardisation of the models as well as a better biological interpretation of the features are required before its introduction in the clinical practice.

Chapter 2

Feature robustness in patients

One of the questions which is still open in radiomics is: how much heterogeneous can the image dataset be with respect to the acquisition parameters? Increasing the number of patients in a clinical study is essential to achieve the desired statistic power. However, going back in time for retrospective studies or including multiple institutes for multi-centre ones increases the chance to deal with images acquired with different scanners and protocols, and reconstructed with various algorithms.

For this reason, we investigated systematically the heterogeneity observed in the retrospective database of CT images of NSCLC patients at our Institute. It mainly consists of different scanners from the same vendor, different X-ray tube voltages, and different reconstruction algorithms, including filtered back-projection (FBP) and increasing the blending level of the iterative algorithm (IR). The impact of the bin width set during the feature extraction is analysed too. The different metrics adopted to test feature reproducibility will be shown and will be integrated in novel criteria providing indications for feature selection.

2.1 Motivations

The importance of methodological studies to evaluate feature robustness was emphasised by a previous investigation carried out by our group [18]. In the latter we considered 270 NSCLC patients staged up to T3N1M0[1], who did not receive a pre-operative chemotherapy and underwent surgery of the lesion after the CT acquisition. The radiomic features were extracted from the lung lesions to be associated to the lymph node status and the overall survival. A preliminary investigation of the feature robustness was carried out with an analysis of variance test, considering as possible confounding factors the contrast medium, the reconstruction algorithms, the scanner and the exposure. The reconstruction algorithm (IR versus FBP) was the parameter that affected

[1]The TNM classification of the malignant tumours is standardised to classify the malignancy of the disease. T stands for *tumour* and describes the size and the extension of the tumours (it can be a number between 1 and 4, where smaller numbers correspond to smaller tumours). N stands for *nodes* and it classify the involvement of the lymph nodes (0 for no involvement and 3 for the maximum severity). M stands for *metastasis* and can be 0 (no distant metastasis) or 1 (presence of distant metastasis).

the features the most. Therefore, in order to include the largest number of CT images, only the robust features were used for the predictive model.

In a further study we analysed the imaging properties of a second dataset of 226 CT images acquired at the European Institute of Oncology (IEO, IRCCS, Milan). This database consists of patients with NSCLC tested for mutational status and who underwent chemotherapy. They were enrolled to evaluate the correlation between the radiomic features and the gene-expression status, and to develop a radiomic model for the prediction of the overall survival. Since it was a retrospective study, the patients were examined with different acquisition/reconstruction parameters on different CT scanners, with different voltage and reconstruction algorithm.

Therefore, in the present study we decided to investigate the impact of those parameters with larger variability in our dataset, and which may have the strongest influence on the image texture: the scanner, the voltage peak, and the reconstruction algorithm.

2.2 Materials and methods

2.2.1 CT image collection

Patients with CT raw-data available at IEO were enrolled retrospectively between January 2019 and December 2019. Only patients with a histological diagnosis of NSCLC and the availability of the raw-data in the CT scanner were included. Moreover, we enrolled only patients whose CT image was acquired at 100 or 120 kVp. Patients with a CT image acquired at 140 kVp were, instead, not included, because in such a clinical scenario this voltage peak value was rarely used, and a statistically representative sample would not be collected. Cases where the lung lesion was not easily measurable, and therefore cases where the lesion could not be contoured, were not considered in the analysis. Finally, since patients with a too small or too large tumour have usually low occurrence, patients with tumour size smaller than 5 cm^3 or greater than 200 cm^3 were excluded during the selection.

We used the two CT scanners available in the radiologic department: the Discovery CT750 HD and Optima CT660 scanners, both of the same vendor (GE Healthcare, Waukesha, WI). Since the beginning of this study, the institutional standard protocol for diagnostic chest imaging was represented by helical acquisition, slice thickness of 2.5 mm, slice spacing of 2.5 mm, automatic tube current modulation along the z-axis, and automatic angular xy modulation. The X-ray tube voltage of this protocol was selected according to the patient Body Mass Index (BMI) in order to preserve the image quality (higher voltage value for higher BMI). Only the images acquired during the portal phase after iodine-based contrast agent injection were considered for the analysis.

The reconstruction settings are optimised in relation to the acquisition parameters. The typical choice nowadays at the IEO consists in the *Stan-*

dard convolution kernel, and the Adaptive Statistical Iterative Reconstruction (ASIR) algorithm with 60% blending level on the Discovery CT750 HD scanner and 50% blending level on the Optima CT660 scanner. However, in our retrospective study we faced clinical databases made of CT images reconstructed with different blending levels, due to various tests and optimisations carried out after the installation of the scanners. For this reason, we investigated multiple reconstruction levels by applying on the CT image of each included patient the following six IR blending levels: 0% (FBP), 20%, 40%, 50%, 60% and 80% (hereinafter referred to as IR20, IR40, IR50, IR60 and IR80, respectively). The reconstruction of the CT image was performed *a posteriori*, after the CT examination, directly on the CT scanner where the raw-data were stored.

Lesion segmentation

Three operators segmented manually one lung lesion for each patient following common criteria, using the commercial software AWServer 3.2 (Ext. 2.0 tool, GE Healthcare). According to the lesion location and to the contrast with the surrounding tissue, two main level windows were applied: the lung window (width of 1500 HU and level of -600 HU), and the mediastinal one (width of 350 HU and level of 40 HU). It is important to note that, for each patient, the segmentation was performed only on the CT image reconstructed with the standard protocol and not on the other ones. We used, in fact, the same contours among the differently reconstructed images, since they were all intrinsically co-registered. In this way, we prevented the introduction of variability coming from the repetition of the contouring on each separated reconstruction.

2.2.2 Radiomic feature extraction

For each patient we extracted the radiomic features for each of the six reconstructed images with the open-source software PyRadiomics (v. 2.2.0, python v. 3.7). Details on the use of PyRadiomics can be found in Appendix A. We considered all the seven categories of features, and used the 2.5D modality, as recommended in presence of non-isotropic voxels (https://arxiv.org/pdf/1612.07003.pdf). The list of the parameters set in Pyradiomics for the radiomic feature extraction can be found in **Table A.1**.

Following the IBSI recommendations [1], we applied two pre-processing steps before the feature computation. First we performed a resampling in the axial plane, aligning the pixel size among the different patients [108, 110]. To this aim we used PyRadiomics' B-Spline interpolator ("sitkBSpline"). Then we discretised the voxel intensity (in HU) with a fixed bin width. We set the bin width to 25 HU (the default value in PyRadiomics), as often suggested in the literature [73, 96, 167, 168], and to 5HU.

We excluded or modified *a posteriori* some features extracted from PyRadiomics, since they were characterised by a strong correlation with the number

of voxels in the mask. Details on this analysis and the list of features eliminated or corrected are given in Appendix B.

For each patient and for each of the six reconstructions, we calculated the radiomic features from both the original images (without filtering) and the filtered images. The two filters taken in consideration in this study were the wavelet filter (*coif1*, the default setting in PyRadiomics [16, 73, 169]) and the LoG filter (standard deviation, *sigma*, equal to 0.5, 1.0, 1.5, 2.5, and 5.0 mm [73, 92, 170–172]).

The names of the features obtained from the non-filtered and the filtered images will appear in the text including the adjectives "original", "wavelet" and "log", respectively.

2.2.3 Statistical analysis

For each patient we collected the following clinical information: age, volume of the lesion, biological sex (female or male), tumour type from the cytological or histological examination, previous therapy (yes or not), position (upper, middle and/or lower lobe), and side (right or left lobe) of the tumour. These data were used to understand if the patient populations, with CT images acquired with different voltage and scanner, were clinically similar. For this purpose, the Chi-square or Fisher exact test was applied to the categorical variables to verify whether they are equally likely among the populations (null hypothesis). For continuous variables, instead, the Wilcoxon rank-sum test was chosen to assess if the samples of the different populations came from the same distribution (null hypothesis). We performed this analysis using the functions *wilcox.test*, *chisq.test* and *fisher.test* from the *stats* package in R (v. 4.0.0) [173]. A p-value < 0.05 was used to reject the null hypothesis.

The first analysis we performed was the evaluation of the contribution of the three parameters (voltage, scanner and reconstruction algorithm) on the feature reproducibility separately (univariate analysis). The Wilcoxon rank-sum test was chosen to detect feature differences within CT scanners and within tube voltages. The test was computed using the standard reconstruction blending level according to the CT scanner used: IR50 for the Optima CT660 scanner and IR60 for the Discovery CT750 HD.

While the analysis for the scanner and the voltage was performed by comparing populations of different patients, each group corresponding to a different value of voltage and a different CT scanner, the algorithm analysis was instead based on the comparison among multiple CT reconstructions for the same patient. For this reason, we identified reproducible features across the reconstruction blending levels using the *overall concordance correlation coefficient* (OCCC).

OCCC

The OCCC derives from the concordance correlation coefficient (CCC), proposed by Lin in 1989 [174]. The CCC measures the agreement between two samples of continuous data. Let X_1 and X_2 be two vectors of measurements. The CCC is defined as:

$$\mathrm{CCC} = \frac{2\,\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}\,, \tag{2.1}$$

where σ_{12} is the covariance between the X_1 and X_2 samples, while σ_i and μ_i are the standard deviation and the mean of the sample X_i, respectively. The CCC takes values between -1 and +1, where -1 and +1 correspond to perfect reversed agreement and perfect agreement, respectively. A CCC value equal to 0 indicates, instead, a complete disagreement. The CCC definition contains both a measure of precision and accuracy. In fact, it measures the deviation of the measures from the line fitting the data (precision) and at the same time the deviation of this line from the line with a 45° inclination passing through the origin (accuracy). To see this property, we rewrite the coefficient as

$$\mathrm{CCC} = \rho\,\mathrm{C}\,, \tag{2.2}$$

where

$$\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2} \tag{2.3}$$

is the Pearson correlation coefficient (measure of precision), and

$$\mathrm{C} = 2\left[\frac{\sigma_1}{\sigma_2} + \frac{\sigma_2}{\sigma_1} + \frac{(\mu_1 - \mu_2)^2}{\sigma_1\sigma_2}\right]^{-1} \tag{2.4}$$

is the measure of the shift from the the 45° line through the origin (values close to 1 correspond to a less shift).

The OCCC [175] is an extension of the CCC. The OCCC can in fact can be used when more than two sets of measures are present (X_1, ..., X_n). Let be N the number of measures, the OCCC is defined as

$$\mathrm{OCCC} = \frac{2\,\sum_{j=1}^{N-1}\sum_{k=j+1}^{N}\sigma_{jk}}{(N-1)\,\sum_{j=1}^{N}\sigma_j^2 + \sum_{j=1}^{N-1}\sum_{k=j+1}^{N}(\mu_j - \mu_k)^2}\,. \tag{2.5}$$

In our study, the measures X_1, ..., X_n correspond to the multiple reconstruction blending levels investigated for each patient. Therefore, we employed the OCCC to compare each radiomic feature among the six settings of the reconstruction algorithm. We computed it with the *epi.occc* function from the *epiR* library in R, using the default settings ($na.rm = FALSE$, $pairs = FALSE$). There is no universal agreement on the OCCC threshold to assess reproducibility. We chose 0.85 as the value above which we consider the features reproducible, as done previously in the literature [92, 176, 177].

Linear mixed-effects model

The second type of analysis we performed consisted in the application of a linear mixed-effects model to each feature. The purpose was to analyse simultaneously (multivariable analysis) the contribution of each parameter to the feature variability. A linear mixed-effects model can be chosen to investigate non-independent, multi-level or hierarchical data, assuming a linear relationship among the variables (hence the "linear" adjective) and incorporating both fixed and random effects (hence the "mixed" adjective). The fitting model has the following matrix formulation:

$$\boldsymbol{f} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{Z}\,\boldsymbol{u} + \boldsymbol{\epsilon}\,. \tag{2.6}$$

In eq. (2.6) $\boldsymbol{f}$ is the known output response, namely the radiomic feature value. It has the form of a column vector of size $N \times 1$, where N is the number of patients in our study. $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the matrices of the input variables for the p fixed and the q random effects with size $N \times p$ and $N \times q$, respectively. $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are column vectors containing the corresponding regression coefficients for the p fixed and the q random effects with size $p \times 1$ and $q \times 1$, respectively. In our study, the scanner model, the voltage, the lesion volume and the reconstruction algorithm were taken as fixed effects, while the subjects as random effects (with one grouping factor), since only a restricted sample of all the possible patients is taken into account. Since the distribution of the lesion volume may vary among the population of patients included, we added this variable to the model in order to separate any volume influence on the feature variability from impact of the other parameters. Finally, $\boldsymbol{\epsilon}$ in eq. (2.6) is a column vector of size $N \times 1$ containing the residuals (with an expected value $E[\boldsymbol{\epsilon}] = 0$). The FBP was set as the reference algorithm and each IR was compared to it. We used the *lmer* function from the *lmerTest* package in R to compute the linear mixed-effects model.

All the p-values were adjusted by the false discovery rate (FDR) method[178] to correct for multiple comparisons, thus reducing false positives (type I errors). For this purpose we used the *p.adjust* function from the *stats* package in R (with the option *method = "fdr"*). Adjusted p-values < 0.05 were considered as statistically significant. **Table 2.1** summarises the analyses presented so far.

Combination of OCCC and linear mixed-effects model

We combined the results of the OCCC analysis and of the multivariate mixed model, and divided the features into four groups:

- **group 1** OCCC ≥ 0.85 and mixed model p-value < 0.05;
- **group 2** OCCC ≥ 0.85 and mixed model p-value ≥ 0.05;
- **group 3** OCCC < 0.85 and mixed model p-value < 0.05;
- **group 4** OCCC < 0.85 and mixed model p-value ≥ 0.05.

Analysed variable	Statistical test
Clinical similarity among the populations	chi-square test, Fisher exact test, Wilcoxon rank-sum test (p-value < 0.05)
Univariate analysis	
CT scanner and tube voltage	Wilcoxon rank-sum test (p-value < 0.05)
Reconstruction algorithm	Overall concordance correlation coefficient (OCCC < 0.85)
Multivariate analysis	Linear mixed-effects model (p-value < 0.05)

Table 2.1: Summary of the statistical tests. The impact of the acquisition and of the reconstruction parameters was evaluated for the features extracted with 5 HU and 25 HU bin widths, separately.

We performed the analyses on both the filtered (LoG and wavelet ones) and the unfiltered images, and after both the 25 HU and the 5 HU discretisations. All the tests were two-sided.

2.3 Results

The total number of patients whose raw-data were still stored in the two radiological CT scanners under investigation was 163. However, only 103 of them were included in the analysis, since the inclusion criteria were not satisfied for the others. The exclusion criteria for this study, the number of included patients, and the number of excluded patients for each exclusion criterion are shown in **Figure 2.1**.

From the 103 enrolled patients (59 men, mean age 71 years; 44 women, mean age 67 years) we identified four populations:

1. 25 patients acquired on the Optima CT660 scanner at 100 kVp;
2. 25 patients acquired on the Optima CT660 scanner at 120 kVp;
3. 27 patients acquired on the Discovery CT750 HD scanner at 100 kVp;
4. 26 patients acquired on the Discovery CT750 HD scanner at 120 kVp.

This subdivision is summarised in **Figure 2.1** as well.

The clinical information collected for each patient is reported in **Table 2.2**, both for the entire patient population and separately for the four populations. The p-value of the clinical similarity between the two scanners and the two voltage populations is also reported. The comparison between the patient populations according to scanner model (Discovery CT750 HD versus Optima

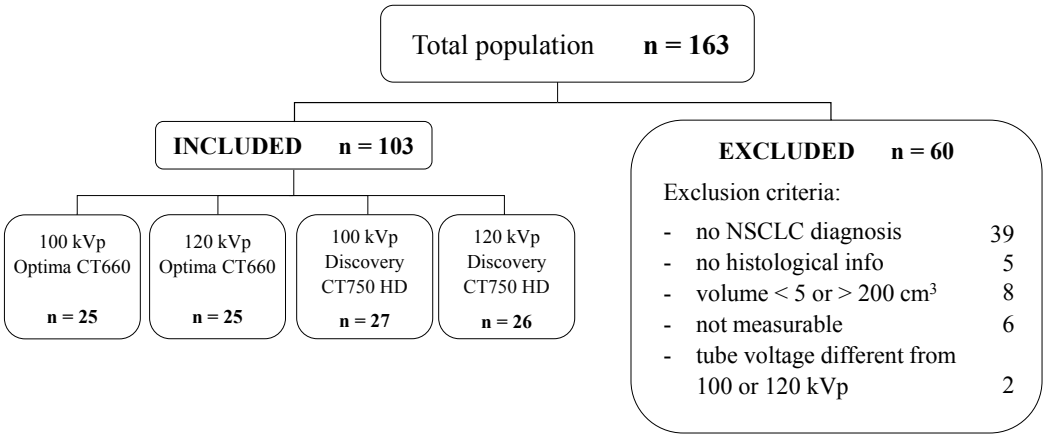


Figure 2.1: Schematic summary of the patient database showing on the left the included patients and their subdivision into four populations, and on the right the excluded patients together with the exclusion criteria.

CT660) and tube voltage (100 versus 120 kVp) showed a compatibility among the groups of patients under investigation (p-value non-significant). Therefore, the clinical characteristics are not a confounding factor for the analysis.

The noise index (NI) of the CT images of the four groups of patients was: 14 and 12 in the 72% and 28% of the cases, respectively, for the Optima CT660 scanner at 100 kVp; 16, 14, 22 and 17 in the 68%, 20%, 8% and 4% of the cases, respectively, for the Optima CT660 scanner at 120 kVp; 19 and 18 in the 93% and 7% of the cases, respectively, for the Discovery CT750 HD scanner at 100 kVp; 18 in the 100% of the cases for the Discovery CT750 HD scanner at 120 kVp. Concerning the volume computed tomography dose index, the values calculated for the four populations — and reported as median (interquartile range or IQR) — were: 10.40 (8.79-13.63) mGy for the Optima CT660 scanner at 100 kVp; 12.34 (10.00-17.24) mGy for the Optima CT660 scanner at 120 kVp; 8.77 (7.93-11.06) mGy for the Discovery CT750 HD scanner at 100 kVp; 12.12 (11.40-14.39) mGy for the Discovery CT750 HD scanner at 120 kVp.

We extracted 1413 radiomic features for each patient: 153 from the original images (13 shape, 17 firstorder, 69 glcm, 14 glrlm, 15 glszm, 13 gldm and 12 ngtdm), 560 from the Wavelet-filtered images (the same firstorder and texture features for each of the four sub-band), and 700 from the LoG-filtered images (the same firstorder and texture features for each of the five *sigma* values). The extraction was repeated for each reconstructed image and each grey-level discretisation value. The list of the feature names and the corresponding category are reported in the Appendix B (see **Table B.3**).

2.3.1 Impact of the tube voltage and of the scanner model

In our study we did not find a relevant impact of the scanner model and of the tube voltage on the feature reproducibility. The majority of the features, in

Variables	Overall cohort (n = 103)	Scanner Optima (n = 50)	Scanner Discovery (n = 53)	p-value scanner	Voltage 120 kVp (n = 51)	Voltage 100 kVp (n = 52)	p-value voltage
Biological sex							
Male	59 (57%)	28 (56%)	31 (58%)	0.798 [a]	33 (65%)	26 (50%)	0.131 [a]
Female	44 (43%)	22 (44%)	22 (42%)		18 (35%)	26 (50%)	
Age							
Mean (median)	69.2 (70)	69.4 (70)	68.9 (69)	0.498 [c]	69.8 (70)	68.6 (68.5)	0.251 [c]
IQR	(64–75)	(65–75.3)	(62–74.5)		(64–76)	(62–74.8)	
Side							
Right	60 (58%)	31 (62%)	29 (55%)	0.454 [a]	31 (61%)	29 (56%)	0.606 [a]
Left	43 (42%)	19 (38%)	24 (45%)		20 (39%)	23 (44%)	
Position							
Upper	63 (64%)	33 (69%)	30 (60%)		30 (61%)	33 (67%)	
Medium	1 (1%)	1 (2%)	0 (0%)	0.360 [b]	1 (2%)	0 (0%)	0.731 [b]
Lower	29 (30%)	13 (27%)	16 (32%)		16 (33%)	13 (27%)	
Mixed	5 (5%)	1 (2%)	4 (8%)		2 (4%)	3 (6%)	
Volume (cm^3)							
Mean (median)	46.4 (39.1)	44.2 (40.6)	48.5 (38.1)	0.843 [c]	52.1 (42)	40.9 (36.7)	0.181 [c]
IQR	(19.1–62.8)	(19–54.7)	(19.5–71.9)		(20.7–67.9)	(18.4–56.2)	
Histological type							
Adenocarcinoma	83 (82%)	38 (78%)	45 (87%)		40 (78%)	43 (86%)	
SCC	16 (16%)	10 (20%)	6 (11%)	0.580 [b]	9 (18%)	7 (14%)	0.380 [b]
Neuroendocrine	2 (2%)	1 (2%)	1 (2%)		2 (4%)	0 (0%)	
Previous therapy							
No	75 (74%)	38 (76%)	37 (73%)	0.692 [a]	33 (66%)	42 (82%)	0.060 [a]
Yes	26 (26%)	12 (24%)	14 (27%)		17 (34%)	9 (18%)	
Scanner							
Optima	50 (49%)	–	–	–	25 (49%)	25 (48%)	0.924 [a]
Discovery	53 (51%)				26 (51%)	27 (52%)	
Voltage (kVp)							
120	51 (50%)	25 (50%)	26 (49%)	0.924 [a]	–	–	–
100	52 (50%)	25 (50%)	27 (51%)				

Table 2.2: Clinical information on the 103 enrolled patients, considering the overall cohort and the four patient populations. The p-value of the clinical similarity test performed among the four populations is also reported, for the scanner and the voltage comparison. The statistical tests used are: [a] chi-square test, [b] Fisher's exact test, and [c] Wilcoxon rank-sum test. Missing data: 2 for histological type; 2 for previous therapy; 5 for position. SCC stands for squamous cell carcinoma.

fact, resulted to be not significantly affected by these parameters at either the univariate or the multivariable analysis. **Table 2.3** summarises the small subset of features which were found to be significantly affected, along the p-value, for the 25 HU discretisation.

More in detail, when the images were discretised with a bin width equal to 25 HU, the number of statistically different features was four for the scanner model and four for the voltage (adjusted p-value < 0.05), in the univariate analysis. According to the multivariable analysis, instead five features showed statistically significant dependence on these parameters (four wavelet features for the voltage and one shape feature for the scanner).

A larger number of features resulted to be significantly affected by the scanner model when the 5 HU discretisation is applied: 14/153 original features, 101/560 wavelet features, and 20/700 LoG features gave an adjusted p-value

Features	Scanner (univar)	Voltage (univar)	Scanner (mixed)	Voltage (mixed)
original_shape_SurfaceArea	0.924	0.714	**0.027**	0.886
original_glrlm_RunVariance	**0.025**	0.714	0.609	0.886
original_gldm_Dependence NonUniformityNormalized	**0.025**	0.873	0.735	0.886
original_gldm_ DependenceVariance	**0.005**	0.714	0.303	0.886
wavelet-HH_glszm_SizeZone NonUniformityNormalized	0.752	**0.038**	0.996	**0.005**
wavelet-HH_glszm_ SmallAreaEmphasis	0.751	**0.038**	0.996	**0.005**
wavelet-HH_glcm1_Correlation	0.434	0.141	0.561	**0.018**
wavelet-HH_glcm1_ InverseVariance	0.066	0.141	0.309	**0.012**
wavelet-HL_glcm1_Correlation	0.886	**0.045**	0.996	0.181
wavelet-HH_gldm_ LargeDependenceEmphasis	0.464	**0.039**	0.996	0.118
log-sigma-0-5-mm-3D_glcm1_ InverseVariance	**0.0002**	1.000	0.973	0.936

Table 2.3: List of the features extracted from the images discretised to 25 HU which resulted significant in either the univariate or the multivariable analysis within the tube voltage and within the scanner model. The significant p-values are highlighted in bold.

< 0.05 in the univariate analysis. For the voltage influence, instead, only two wavelet features yielded a significant difference. In the multivariable analysis, on the other hand, we found results similar to the 25 HU discretisation, with one original feature affected by the scanner (original_shape_SurfaceArea, p-value = 0.027) and one wavelet feature (wavelet-HH_glcm1_Correlation, p-value = 0.001) by the voltage.

This feature affected by the scanner is the same for the two discretisation values and it was the *shape_SurfaceArea*, which provides the size of the surface of the VOI. It is not surprising that the feature is the same for the two discretisation bin widths, because this kind of processing impacts only the texture of the lesion. We expect that the non-reproducibility of this feature between the images of the two scanners may be related to a slight difference, in terms of the lesion shape, between the two populations. Higher values of this features were found in the population scanned on the Discovery CT 750 HD at 120 kVp, indicating thus a larger complexity in the lesion surface of this group. This result, however, is likely sample-specific, and we are not expecting it to be generalisable to different samples. Increasing significantly the number of

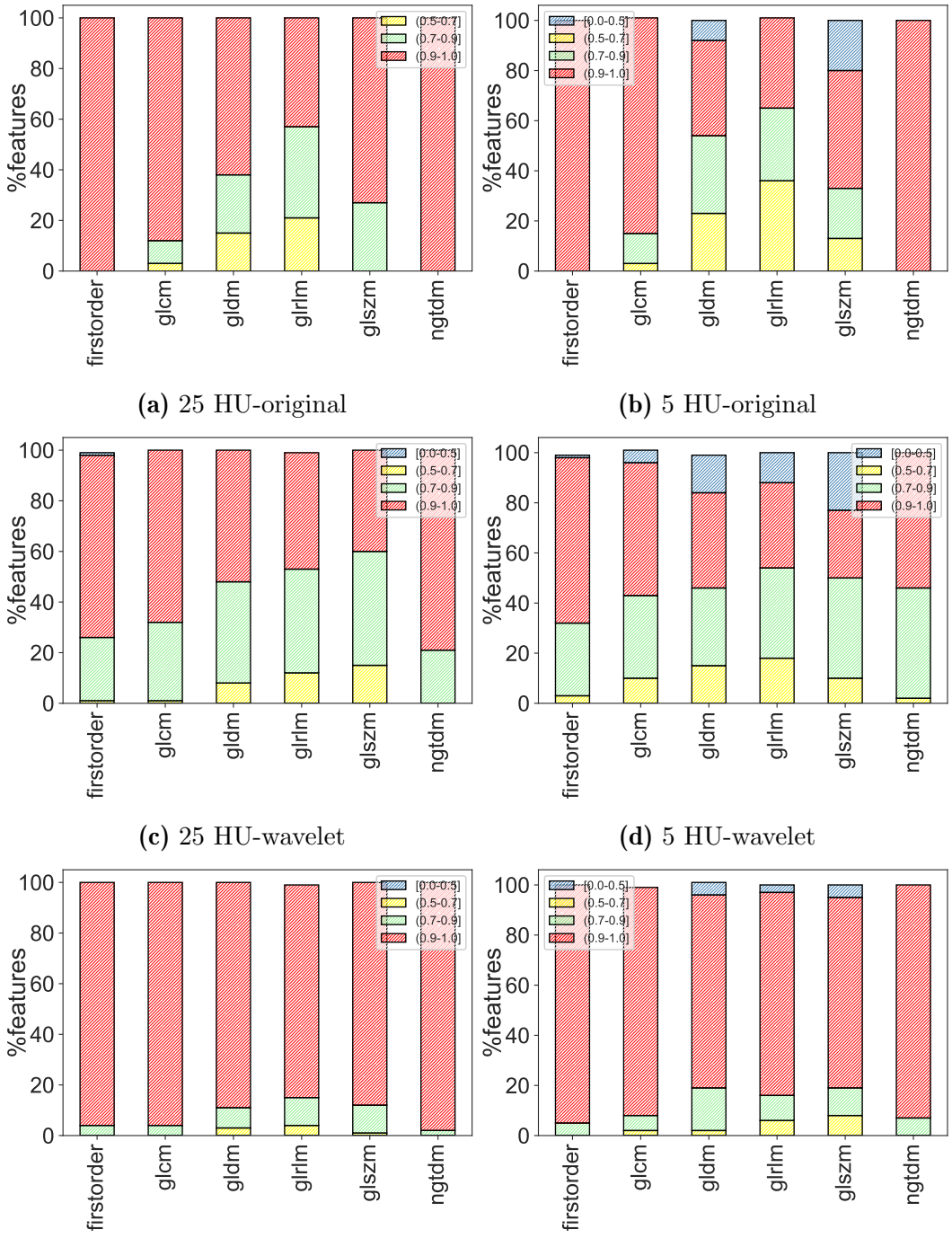
patients may help in supporting this hypothesis.

2.3.2 Impact of the reconstruction algorithm

Figure 2.2 summarises the results for the three image types (original, wavelet with the four sub-bands together and LoG with the five *sigma* values together) and the six feature categories (firstorder, glcm, glrlm, glszm, gldm, ngtdm), for the discretisation at 25 and 5 HU. It must be noted that the *shape* category exhibited a perfect concordance (OCCC = 1), since the same segmentation was applied to the different image reconstructions for each patient. We divided the features into four groups according to the OCCC value: OCCC $\leq$ 0.50, 0.50 $<$ OCCC $\leq$ 0.70, 0.70 $<$ OCCC $\leq$ 0.90 and OCCC $>$ 0.90. We observed that the majority of the features falls in the range (0.9, 1] (red bar), while the group of features with OCCC $\leq$ 0.50 (blue bar) is the least populated, and it is even empty for the original features extracted from the 25 HU discretised images (**Figure 2.2-a**). Moreover, by visually comparing the features from the 25 HU (plots on the left) and the 5 HU (plots on the right) discretisation, it appears that the latter have in general a lower value of the OCCC. Additional information is shown in **Table 2.4**, where the median OCCC and the percentage of feature with OCCC $\geq$ 0.85 are listed for each feature category and image type. In general, the firstorder features are the most robust, while the glrlm, gldm and glszm categories are the least stable. Moreover, the 25 HU discretisation produced a more reproducible texture than the 5 HU one. The LoG-based features are overall the most stable, while the wavalet-based are the least. Focusing on the wavelet sub-groups (HH, HL, LL, LH), most of the non-reproducible features belonged to the HH sub-band (59 and 86 out of 560 features had a OCCC $<$ 0.85 for the 25 HU and the 5 HU discretisation, respectively), while the LL sub-band showed the highest concordance (11 and 13 out of 560 features had a OCCC $<$ 0.85 for the 25 HU and the 5 HU discretisation, respectively). The median OCCC for each sub-band is reported in the boxplot of **Figure 2.3** for the two bin widths.

Considering the different *sigma* values of the LoG-filtered images, we found an excellent reproducibility not only in all the feature categories, but also in each investigated configuration of the *sigma* parameter. In fact, for *sigma* values greater than 1 mm, almost all the features had a OCCC $\geq$ 0.85, both with the 25 HU and the 5 HU discretisation. On the contrary, a few features exhibited a low reproducibility when small values of *sigma* (0.5 mm and 1 mm) were used: 21 (41) and 1 (9) out of the 700 LoG-based features had a OCCC $<$ 0.85 for the 25 HU (5 HU) discretisation and for *sigma* equal to 0.5 mm and 1 mm, respectively. The median OCCC for each sigma value is shown in the boxplot of **Figure 2.4** for the two bin widths.

In the multivariable analysis, the percentages of features significantly different (adjusted p-value $<$ 0.05) when extracted from the IR settings compared to FBP are the following: 110/140 (78.5%) of the original features (*shape* features excluded), 462/560 (82.5%) of the wavelet features (of the 462, 25%,



(a) 25 HU-original **(b)** 5 HU-original

(c) 25 HU-wavelet **(d)** 5 HU-wavelet

(e) 25 HU-LoG **(f)** 5 HU-LoG

Figure 2.2: Percentage of features with an OCCC value in one of the following ranges: [0-0.5], (0.5, 0.7], (0.7, 0.9] and (0.9, 1.0]. The results are reported separately for each feature category (*shape* excluded), for the original features with 25 HU (**a**) and 5 HU discretisation (**b**), for the wavelet–based features with 25 HU (**c**) and 5 HU discretisation (**d**), and for the LoG-based features with 25 HU (**e**) and 5 HU discretisation (**f**).

Filter	Category	**25 HU** median OCCC (% OCCC $\geq$ 0.85)	**5 HU** median OCCC (% OCCC $\geq$ 0.85)
original	firstorder	1.00 (100%)	1.00 (100%)
	glcm	0.99 (96%)	0.99 (88%)
	glrlm	0.88 (57%)	0.86 (50%)
	gldm	0.93 (77%)	0.84 (46%)
	glszm	0.96 (73%)	0.89 (60%)
	ngtdm	1.00 (100%)	1.00 (100%)
	total	**0.98 (89%)**	**0.98 (80%)**
wavelet	firstorder	0.97 (85%)	0.95 (81%)
	glcm	0.95 (87%)	0.92 (70%)
	glrlm	0.88 (66%)	0.85 (52%)
	gldm	0.92 (67%)	0.85 (50%)
	glszm	0.87 (52%)	0.84 (50%)
	ngtdm	0.94 (88%)	0.93 (81%)
	total	**0.93 (79%)**	**0.90 (67%)**
LoG	firstorder	1.00 (99%)	1.00 (98%)
	glcm	1.00 (99%)	1.00 (94%)
	glrlm	0.99 (90%)	0.98 (90%)
	gldm	0.99 (92%)	0.99 (86%)
	glszm	0.99 (95%)	0.99 (85%)
	ngtdm	1.00 (100%)	1.00 (97%)
	total	**1.00 (97%)**	**1.00 (93%)**

Table 2.4: Median OCCC for each category of features and for each type of image (filtered and non-filtered). The percentage of features with OCCC $\geq$ 0.85 is given in parentheses (the percentage is calculated considering the total number of features in each category). The results are given for the two types of discretisation (25 HU and 5 HU).

24%, 26% and 25% for HH, HL, LH and LL, respectively) and 470/700 (67%) of the LoG features (of the 470, 24%, 23%, 20%, 17% and 16% for 0.5 mm, 1.0 mm, 1.5 mm, 2.5 mm and 5.0 mm sigma, respectively). This percentage are obtained considering a p-value < 0.05 for all the configurations of IR (from 20% to 80%). The corresponding percentages are reported for each IR setting in **Table 2.5**, showing an increase in the difference between FBP and IR when the blending level increases, as expected.

The four groups into which we divided the radiomic features, taking into account the results of the two analyses (OCCC-based and multivariable mixed-effects model) on the impact of the reconstruction settings were characterised by peculiar properties, outlined in **Table 2.6**. Features with larger values of OCCC (OCCC $\geq$ 0.85) are characterised by a small variation of the feature for each patient among the reconstruction blending levels compared to the

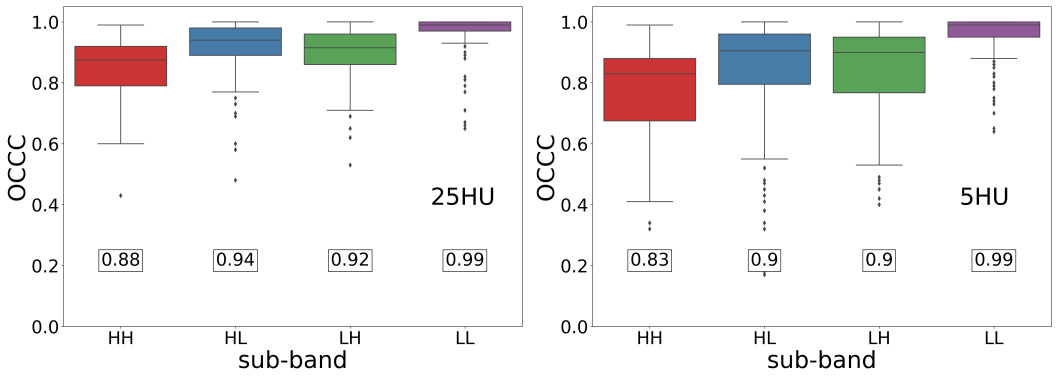


Figure 2.3: Boxplot reporting the OCCC for each sub-bands of the wavelet features for the 25 HU **(a)** and 5 HU **(b)** bin widths.

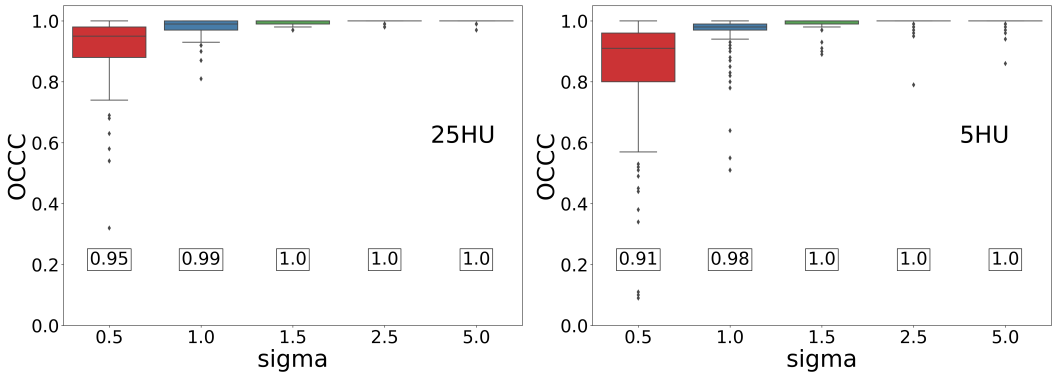


Figure 2.4: Boxplots showing the OCCC for each sigma value of the LoG features for the 25 HU **(a)** and 5 HU **(b)** bin widths.

variation observed among all patients in the dataset. A significant p-value of the multivariate mixed model (p-value < 0.05), instead, indicates that the feature trend, when the reconstruction algorithm changes, is systematic among all patients. In contrast, the features with a non-significant p-value have a trend with the algorithm which is random among the patients. An example of these behaviours is illustrated in **Figure 2.5**, where the absolute value of one feature for each of the four groups is reported as the reconstruction blending level increases (*original-glcm1_ClusterTendency* feature as representative of group 1, *original-glszm_HighGrayLevelZoneEmphasis* feature as representative of group 2, *original-glrm_RunVariance* feature as representative of group 3, *original-glrlm_LongRunLowGrayLevelEmphasis* feature as representative of group 4).

We obtained that the majority of the features fall in group 1 (OCCC $\geq$ 0.85 and adjusted p-value < 0.05), suggesting the capability of the features to capture the gradual smoothing effect of the increasing IR strength on the

Filter	IR	25 HU % p-value < 0.05	5 HU % p-value < 0.05
original	IR20	79%	79%
	IR40	91%	88%
	IR50	91%	89%
	IR60	91%	91%
	IR80	94%	94%
	all IR	79%	79%
wavelet	IR20	83%	82%
	IR40	93%	92%
	IR50	95%	94%
	IR60	96%	97%
	IR80	96%	97%
	all IR	83%	82%
LoG	IR20	67%	72%
	IR40	81%	84%
	IR50	88%	86%
	IR60	89%	88%
	IR80	91%	92%
	all IR	67%	71%

Table 2.5: Percentage of features with a significant adjusted p-value from the linear mixed-effects model for each type of image (filtered and non-filtered) and each IR setting. The percentage is calculated considering the total number of features (*shape* excluded). The results are given for the two types of discretisation (25 HU and 5 HU) separately.

image texture, with a similar trend for all the patients. In contrast, group 4 is the least populated. This result can be observed in **Table 2.7**, in which we reported the percentage of features falling in each of the four groups for the 25 HU and 5 HU bin width. In particular, this table includes only those features that in the multivariable analysis resulted significantly affected by the reconstruction algorithm for all the IR settings (adjusted p-value < 0.05 for all the analysed IR) and those which were not for all the IR settings (adjusted p-value > 0.05 for all the analysed IR), compared to FBP. A very small number of features was found in group 4 for the set of IR blending levels selected for the study (*wavelet-HH_glcm7_Correlation* for the 25 HU bin width, *original_glrlm_LongRunLowGrayLevelEmphasis*, *original_glszm_LargeAreaLowGrayLevelEmphasis*, *original_gldm_LargeDependenceLowGrayLevelEmphasis*, *wavelet-LL_glszm_LargeAreaLowGrayLevelEmphasis*, *wavelet -LL_glcm1_MCC* and *log-sigma-2-5-mm-3D_gldm_LargeDependenceLowGrayLevelEmphasis* for the 5 HU bin width). The remaining features were, instead, partially affected by the algorithm, depending on the intensity of the blending level.

The details of the results for each feature can be found in the supplementary

Group 1	**Group 2**
OCCC $\geq$ 0.85 + mixed p-value $<$ 0.05	**OCCC $\geq$ 0.85 + mixed p-value $\geq$ 0.05**
• *slight* variation among reconstruction settings compared to differences among patients • *systematic* variations among patients	• *slight* variation among reconstruction settings compared to differences among patients • *random* variations among patients
Group 3	**Group 4**
OCCC $<$ 0.85 + mixed p-value $<$ 0.05	**OCCC $<$ 0.85 + mixed p-value $\geq$ 0.05**
• *relevant* variation among reconstruction settings compared to differences among patients • *systematic* variations among patients	• *relevant* variation among reconstruction settings compared to differences among patients • *random* variations among patients

Table 2.6: Schematic description of the four groups into which the radiomic features were classified based on the two analyses performed on their reconstruction algorithm dependence.

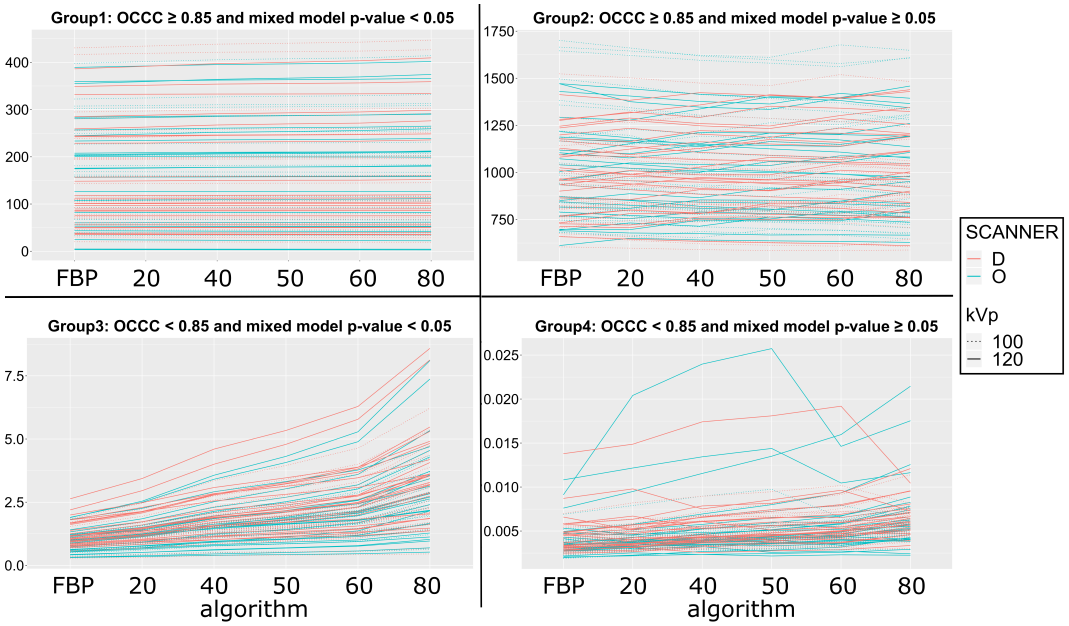


Figure 2.5: Examples of feature behaviour for each of the four groups. The trend of the feature value is plotted as the reconstruction blending level varies, each line representing a different patient. The features plotted are: original-glcm1_ClusterTendency for group 1, original-glszm_HighGrayLevelZoneEmphasis for group 2, original-glrm_RunVariance for group 3 and original-glrlm_LongRunLowGrayLevelEmphasis for group 4.

Filter	IR	25 HU %	5 HU %
original	GROUP1	69%	61%
	GROUP2	5%	4%
	GROUP3	9%	17%
	GROUP4	0%	2%
wavelet	GROUP1	64%	53%
	GROUP2	2%	2%
	GROUP3	18%	29%
	GROUP4	0.2%	0.4%
LoG	GROUP1	64%	66%
	GROUP2	8%	7%
	GROUP3	3%	5%
	GROUP4	0%	0.1%

Table 2.7: Percentage of features in the four groups according to the OCCC and the linear mixed-effects model analysis for each type of image (filtered and non-filtered). The percentage is calculated considering the total number of features (*shape* excluded). The results are given for the two types of discretisation (25 HU and 5 HU) separately.

materials of ref. [179].

2.4 Discussion

In this study we investigated the influence of the acquisition, reconstruction and post-processing parameters on the radiomic features extracted from a retrospective database of 103 NSCLC patients. Two CT scanners of the same vendor and two tube voltage peaks (100 kVp and 120 kVp) were compared. We analysed various reconstruction blending levels, including FBP and iterative algorithms. Finally, we performed the analysis of reproducibility separately for a discretisation of 25 HU and 5 HU. We performed the analysis with two approaches, univariate and multivariable, in order to evaluate both the impact of each single parameter and the concurrent interactions between the three different settings (scanner, voltage, and reconstruction algorithm).

2.4.1 Impact of the tube voltage and of the scanner model

According to both the univariate and the multivariable analysis, neither the tube voltage nor the scanner model appeared to influence significantly the value of most of the radiomic features in our dataset when a bin width of 25 HU was applied. A larger impact was found for the 5 HU discretisation in the univariate analysis, but this influence vanished with the multivariable analysis.

The weak influence of the tube voltage was also observed in previous radiomic studies [56, 180]. In particular, the study by Fave et al. [56] reached a similar conclusion by simulating different CT textures at 80, 100 and 140 kVp from a 120 kVp acquisition in NSCLC patients. Conversely, a study performed by Berenguer et al. [181], changing peak tube voltage between 80 and 140 kVp during different CT phantom acquisitions, found that 22.7% of the features were not reproducible. This is a larger percentage than what we observed in our study, even though the reproducibility improved when reducing the tube voltage range from 80-140 kVp down to 120-140 kVp. However, it must be noted that Berenguer et al. performed a different statistical analysis and, most importantly, investigated the role of tube voltage while keeping all the other acquisition parameters fixed. In our study, instead, the voltage was modified along with other acquisition parameters (e.g. NI) to maintain a comparable image quality independently on the patient size, mirroring the clinical practice protocol. In this sense, our results are also an indirect confirmation that the chosen optimisation yields comparable image texture as well.

The impact of the scanner was also already investigated in the literature, mainly using phantoms. Berenguer et al. [181] acquired phantom images using five scanners from two vendors, fixing a number of acquisition parameters (voltage, current and slice thickness), and applying the standard reconstruction kernel available for each scanner. In such fixed and controlled conditions, they did not detect a significant influence of the scanner model on the radiomic features, in agreement with our results on NSCLC patient images. It must be noted that the two scanners used in our study are from the same vendor and are calibrated following the same procedure. Moreover, the acquisition protocols are optimised in the same way for each anatomic district. These aspects may explain the reduced impact observed on the radiomic features. The results obtained by Mackin et al. [106] and by Ger et al. [111] supported these hypotheses. The group of Mackin compared the CT scanners of four different manufacturers and found that the features extracted from images of the same vendor clustered together. Ger et al. instead, observed that the feature variability was reduced when a controlled protocol was used for the multi-centre acquisitions of the phantom images, compared to the acquisitions with a distinct protocol for each institute.

2.4.2 Impact of the reconstruction algorithm

Differently from the tube voltage peak and scanner, our study confirmed that the IR blending level has a significant impact on a subset of NSCLC radiomic features, to a different extent in case of the original and the filtered images. Similar analyses on lung cancer CT images were published, including studies which investigated the effect of the kernel (sharp, smooth, and standard) and the algorithm type (FBP and iterative) [92, 97, 98, 182].

In this study we proposed a novel approach to analyse in detail the influence of the algorithm setting by combining two different statistical techniques,

one based on a univariate analysis (OCCC) and the other on a multivariable approach. This allowed us to cluster the feature behaviour into four groups, for which different selection or correction strategies can be adopted. According to the univariate analysis, only 11% of the features (154/1413) resulted non-reproducible yielding an OCCC < 0.85 for the 25 HU discretisation, and 19% (266/1413) for the 5 HU discretisation. Therefore, considering the univariate analysis alone, these features should be rejected for future investigations on a clinical cohort of patients similar to the one included in this study. However, considering the results from the multivariable analysis, we found that the majority of the features were characterised by a systematic dependence on the reconstruction blending levels among the patients (indicated by p-value < 0.05 in the multivariable analysis). Prezzi et al. [99] also observed this trend of the radiomic features with the IR blending levels. They investigated the behaviour of the features extracted from CT images of patients with primary colorectal cancer after the application of the ASIR algorithm by changing the blending level between 0% and 100% at intervals of 20%. A linear relation between the feature value and the IR level was identified in most of the features (wavelet and LoG filters were not investigated in this study).

Taking advantage from this property, features which were excluded in the univariate analysis (OCCC < 0.85) may be reintroduced in the analysis after a suitable correction. The features we are referring to are those belonging to group 3 (OCCC < 0.85 and p-value < 0.05). A methodological study should be designed ad hoc according to the variability observed in the dataset of interest in order to identify the coefficients of correction and harmonise the features extracted from the CT images reconstructed with a different algorithm. For the features in group 1 (OCCC ≥ 0.85 and p-value < 0.05) — which are the vast majority — instead, the real necessity of such a correction may vary on a case-by-case basis depending on the clinical question the radiomic analysis is supposed to answer. For example, if the aim is to discriminate two patient populations for which the difference, in terms of radiomic features, exists but is very small, even the slight feature variation introduced by different IR blending levels may have a relevant impact, confounding the data and impairing the ability of radiomics to reach its goal. In this case, the feature correction is necessary and suggested as for the features in group 3. Conversely, if the difference between the features of the two populations is far larger than the fluctuations due to the different reconstruction settings, such a correction may be irrelevant. This can be investigated in future studies investigating the potential role of radiomics for different clinical endpoints on the NSCLC population. The features belonging to the group 2 (OCCC ≥ 0.85 and p-value ≥ 0.05) for all the IR blending levels can be considered as the most reproducible, but their number is very small. Lastly, the features in group 4 (OCCC < 0.85 and p-value ≥ 0.05) should be rejected.

Thanks to these observations we can also make a number of general con-

siderations about the feature behaviour.

- WAVELET FILTER
  First of all, the wavelet-filtered images had the worst results in terms of feature reproducibility in all the feature categories, confirming the results obtained in previous studies on this topic [124, 182]. When considering each wavelet sub-band separately, the highest reproducibility was observed for features extracted from the LL-wavelet filtered images (see **Figure 2.3**). This filter in fact creates a smoothing effect, this way reducing the noise as compared to the non-filtered images.

- LOG FILTER
  The LoG-filtered images, instead, exhibited the best performance in terms of reproducibility, which increased with the Gaussian sigma value in all the categories of features (see **Figure 2.4** for the univariate analysis). This result agrees with the study from Zhao et al. [92], which compare sharp and smooth reconstructions on 32 images of lung cancer patients. Similarly to the LL-Wavelet filter, this can be explained by the smoothing effect of the Gaussian kernel, which reduces the noise and makes the image texture more similar, with the drawback of possible loss of informative content. **Figure 1.4** shows this increasing effect of the LoG filter on a lung lesion as the *sigma* value increases. The features which belonged to group 1 at low *sigma* value fell into group 2 for high value of *sigma*, because the trend with the IR blending level vanished. We show this progressive flatness with the increasing of *sigma* for the feature *glcm1_idn* in **Figure 2.6**.

  A compromise between reproducibility and preservation of informative content ought to be found, and is likely to be dependent on the feature, the pathology and the clinical endpoint. The present study is helpful in this regard, as it quantifies the effect of different processing settings (LoG sigma value and/or wavelet sub-band) on CT images of NSCLC patients, identifying the optimal processing configuration able to guarantee feature reproducibility. As a next step, dedicated studies on clinical populations should be performed to verify whether the so-obtained features are still informative enough to separate different populations for the clinical endpoints of interest. Notably, the LoG-filtered images with sigma equal to 0.5 mm gave worse reproducibility, not only in comparison to the other LoG-filtered images but also with respect to the original images (see **Figure 2.4** and **Table 2.4**). This could come from the combination of two effects: a limited smoothing effect (small Gaussian sigma with respect to the pixel size), and the enhancement of the remaining noise derived from the subsequent application of a Laplacian filter. The higher noise in the LoG-filtered image with *sigma* equal to 0.5 mm when compared to the original image is manifest in **Figure 1.4**. For this reason, a LoG filtering with a *sigma* smaller than the pixel size should be in general discouraged

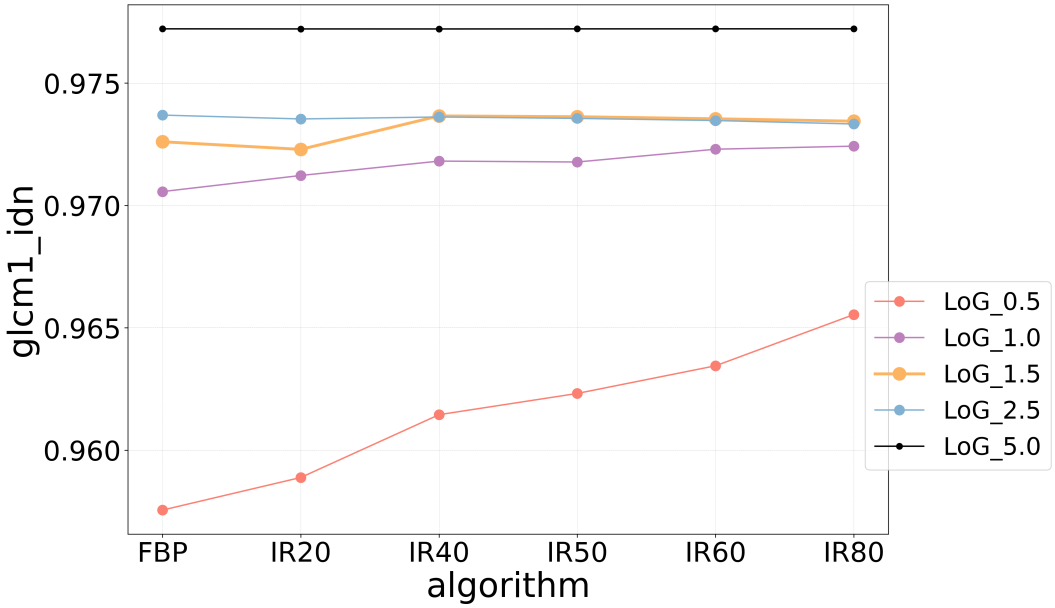


Figure 2.6: Comparison of the absolute value of the *glcm1_idn* feature after the application of the LoG filter with different *sigma* value. The median values among the 103 patients are plotted for the different IR blending levels.

when analysing datasets including different reconstruction algorithms.

- FEATURE CATEGORIES
  For both the original and the filtered images, glrlm, glszm and gldm are the categories exhibiting the largest inter-reconstruction setting variability, whereas the firstorder features are the most reproducible, in agreement with the findings of previous studies [92, 94, 124, 183]. We expect that these results are due to the higher sensitivity of the *texture* features — such as glrlm, glszm and gldm — to small and local changes in the texture, since they measure the presence of adjacent pixels with the same grey-level intensity. When the IR blending level increases, in fact, the smoothness effect on the texture becomes more evident, thus making the texture locally coarser.

- BIN WIDTH
  We observed a larger impact of the scanner and of the reconstruction algorithm in the univariate analysis (see **Figure 2.2**, **Figure 2.4** and **Table 2.4** for the OCCC results). In the multivariable analysis, instead, the two settings turned out to be comparable (see **Table 2.5**). A bin width equal to 25 HU is therefore more stable, as expected for the noise reduction caused by the binning of the grey-level intensities, compared to 5 HU. There exists no agreement for the choice of the bin width in the literature. Even if a lower bin width seems less reproducible, this does not mean that it is less informative. In order to identify the best setting for the feature extraction, a proper study with a clinical outcome should

be performed for each setting, and the performance of the model should be compared.

We did not compare intentionally the features between the 25 HU and the 5 HU discretisation because we were not interested in finding features robust with respect to the bin width. The focus of this study was, in fact, on those parameters that usually vary in the clinical databases. The bin width is a post-processing parameter set during the feature extraction and fixed for all the patients. Therefore, the methodological analysis applied a priori to select robust features should be performed using the same extraction settings used for the clinical investigation.

Limitations

One possible limitation of the present study is the relatively small number of patients enrolled for each tube voltage-scanner combination. In addition, despite the patients were carefully selected to be clinically comparable, it is possible that small differences remained and affected the analysis. New patients should be collected in order to increase the statistical robustness and confirm our results. The inclusion of patients from different centres with the same pathology is an interesting possible development of this study, in view of the generalisation of our results. In this case, in order not to introduce confounding factors from the acquisition and reconstruction procedures — different from scanner, voltage and reconstruction algorithms — a check of the matching among the different clinical protocols is mandatory, for example using the same phantom to evaluate the image quality among the various institutes.

Moreover, the segmentation of the lesion was performed by different persons. It was not possible to evaluate the segmentation impact on the features, because each operator contoured a different lesion. Even though the same segmentation was analysed across the different blending levels for each patient, the impact of the contouring may be present in the scanner/tube voltage analysis, and in general in the multivariable analysis. We believe that, if there were such a difference, it would be small, since all the operators shared the same segmentation criteria, such as the visualisation window, the exclusion of the vessels, and the inclusion of opacity of the lesion edges. The introduction of automatic algorithms may overcome the limits of manual segmentation by reducing operator variability. We will further explore this topic in Chapter 4.

Finally, in this study we did not investigate radiomic feature repeatability, since we did not have repeated acquisitions of the same patient in a short period of time. For this reason, we decided to address this issue in a separate study with ad-hoc phantoms. The design and fabrication of a dedicated phantom for this purpose will be discussed in the Chapter 3.

2.4.3 Conclusions

The results of this study strongly pertain to the characteristics of the image database analysed and cannot be generalised to CT images obtained differently. The proposed methodology can however be exported to each institute in order to guarantee the robustness and reliability of each single-centre study, a fundamental step in order to identify promising radiomic models which may deserve further investigation in a multi-centre setting. Aiming at multi-centre generalisability, instead, a suitable database should be collected to replicate the proposed methodology in a wider image heterogeneity setting, including CT scanners from different vendors, and additional acquisition and reconstruction parameters applied in the clinical practice of different centres.

Alternatively, different methodological approaches can be applied to reduce the feature variability in heterogeneous databases, such as the ComBat method, in which the feature values are adjusted according to the "batch effect" found in the database (see 1.3.2 in Chapter 1), or deep learning algorithms [124, 184], which generate synthetic harmonised images when different imaging protocols are used. Shafiq-ul-Hassan et al. [185] and Mackin et al. [186] corrected the feature values by multiplying them by a factor based on the noise power spectrum peak frequency. All these techniques seem very promising, but a deeper investigation is necessary to understand which of these approaches is the most robust and reliable for a clinical application.

Chapter 3

Feature robustness in phantoms

Phantoms are a valuable tool to investigate radiomic feature behaviour without having to deal with the typical limitations encountered with patients. Such limitations include radiation exposure, which constrains the repetition of image acquisitions, movement artefacts, and tissue changes over time. However, in order to perform this task, objects mimicking the human tissues of interest are required.

In this chapter the design and development of the Heterogeneous Lung Lesion Phantom, HeLLePhant for short, a phantom mimicking the scenario of lung tumour patients to study the feature robustness will be presented. First of all, the importance of phantoms in medical imaging and in particular for radiomics will be emphasised. Next the preliminary investigations to identify the best materials to match the signal of lung tumours in CT images will be presented. The chapter will then focus on the fabrication of the inserts, which are the key elements of the phantom itself. Finally, the radiomic characteristics of the inserts will be compared to a group of patients with lung tumour to evaluate the quality of their fabrication. In the last part of this chapter some applications of the HeLLePhant for radiomic purposes will be presented.

All the CT images of the phantoms were acquired at European Institute of Oncology (IEO, IRCCS, Milan), while the manufacturing of the inserts was partially carried out by the C.I.Ma.I.Na (Centro di Eccellenza Interdisciplinare Materiali e Interfacce Nanostrutturati) group from the University of Milano.

3.1 Motivations

In the following sections the importance of phantoms for radiomic studies is stressed. First of all, the section starts with a brief introduction about what phantoms are and their usage in the clinical practice. Secondly, the advantages of phantoms as substitutes of human beings for methodological studies in radiomics are illustrated. The section ends by empathising the need in radiomics of phantoms as close as possible to the real tissue under investigation.

3.1.1 What are phantoms?

Phantoms are objects introduced in the clinical practice as substitutes of the human body to calibrate the imaging devices, such as scanners, and to control that their performance is stable over time, a necessary requirement for their clinical use [187,188]. For example, in CT imaging, phantoms made of uniform materials, such as water, are used to measure the spatial uniformity of the CT signal and the noise. The same phantoms can be used to check that the scanner is well calibrated. The introduction of small objects with an increasing spatial density inside a uniform background is used to quantify the spatial resolution, while objects with a slightly different electron density are used to quantify the low contrast sensitivity. Other types of phantom incorporate inserts of various tissue-equivalent materials with known electron densities for the correction of tissue heterogeneity in treatment planning systems. For dose distribution measurements, phantoms usually integrate dosimeters to evaluate the radiation energy absorbed at a given position. In order to perform more accurate dose measurements, dosimetric phantoms are often anthropomorphic, which means that they are able to reproduce more closely the geometry of a portion of the body and/or the attenuation of its internal structures. Typical materials used for CT phantoms are polymethyl methacrylate (PMMA), polystyrene, water, epoxy resins, and Teflon. Unfortunately, some of these materials are proprietary. Therefore, the information about their composition is not given, and only some attenuation properties (i.e. the reference electron density) are publicly available.

Similarly to CT imaging, phantoms for quality assurance in nuclear medicine and MRI have been developed too. In nuclear medicine a radionuclide solution is injected, which can be based on ^{18}F-FDG in PET or ^{99m}Tc-based in single photon emission computed tomography (SPECT). Spheres and bars filled with a specific radioactive concentration are used for the calibration and image quality assessment. In MRI, phantoms are typically objects filled with a water-based liquid solution (nickel sulfate, nickel chloride and sodium chloride).

3.1.2 Why do we need phantoms?

Since phantoms are inanimate object, they can be used to perform multiple measurements in a controlled setting, which is unfeasible in-vivo for ethical reasons such as unjustified radiation exposure and patient discomfort. They offer several advantages:

- phantoms enable limitless repeated acquisitions in fixed conditions, i.e. without changing any parameters and without moving the phantom among different scans (**repeatability**);

- phantoms allow us to study thoroughly the impact of the acquisition settings in a controlled way, by changing only the parameters of interest and keeping all the other conditions fixed (**reproducibility**);

- there are not privacy issues with phantoms, as instead in patients, and their images can thus be shared among different centres;

- through phantoms feature variability and stability can be analysed by comparing different scanner models and vendors.

This last point is particularly useful for multi-centre studies, where the patient images are usually acquired in different institutions with different scanners and protocols. For instance, this allows us to select the subset of features which are robust among the different centres for the construction of the models.

3.1.3 Why do we need radiomic phantoms?

The main limitation of methodological studies with phantoms is currently represented by the lack of materials which adequately reproduce the tissue texture for each investigated pathology and encompass the heterogeneity of a clinical population.

Various radiomic studies have been performed with phantoms usually used for quality assurance, which are composed of homogeneous materials. Moreover, these phantoms are commercially available and therefore quite expensive. Lo et al. [189], for instance, used a uniform water phantom to evaluate the feature reproducibility when the dose level and the reconstruction algorithm are modified. Both a uniform water phantom and an anthropomorphic one (ATOM phantoms, CIRS) were scanned in ref. [190], considering as confounding factors the tube current, the noise index, and the reconstruction algorithm. Jin et al. [191], instead, used the acrylic cylinder of the American College of Radiology (ACR) phantom (Gammex) to study the impact of the tube current, the reconstruction kernel, the fields of view and the size of the contour. Similarly, the impact of scanners, voltage and tube current was considered by Nardone et al. [192] using a commercial phantom (Gammex, model 467) with inserts of various materials. The National Electrical Manufacturers Association (NEMA) IQ phantom filled with ^{18}F-FDG was adopted by Jha et al. [193] to perform a test-retest analysis and compare scanners.

However, feature repeatability and reproducibility in general depend on the texture [106, 107, 181], and should thus be assessed specifically for each pathology. For this reason, in parallel to the development of methodological radiomic studies, the need for "radiomic phantom", more appropriate for methodological investigation, has become more and more urgent [188, 194].

Examples of radiomic phantoms in CT imaging

In the CT imaging, the CCR phantom is the most used in radiomic studies, as mentioned in Section 1.3.2. The CCR phantom is a parallelepiped-shape object, composed of ten cartridges. Each cartridge is made of a different material in order to have various densities and textures in a single object. Among

the ten cartridges, the rubber material was found to be the most representative of NSCLC lesions, with a mean CT number of −69 HU [106]. Its simple structure is a strength and a drawback at the same time. It is built by simply juxtaposing ten blocks, half of which 3D printed (based on ABS and plaster), and the other half made of common materials (wood, cork, acrylic). The pile of blocks is then embedded into an acrylic case to keep the entire structure steady. However, the parallelepiped-based form and the absence of materials that simulate the body attenuation make this phantom too basic and far from a real human body. For this reason the CCR phantom manufactured by Ger et al. [111] was characterised by a more patient-like shape, a size close to the average European woman chest dimension (ICRU Report 48), and an increased heterogeneity.

Other types of phantom were developed to study the feature variability. One of the first phantoms proposed for radiomic studies was fabricated by the group of Samei [195] in 2019. Various 3D printed inserts with different shapes, sizes and textures were fabricated and put inside an anthropomorphic thoracic phantom. Starting from the features extracted from the CT images of patients with lung lesions, they developed a genetic algorithm to produce synthetic images of the inserts, which were then used as model for the 3D printing process. For each voxel of the synthetic images ($0.042 \times 0.084 \times 0.03$ mm^3), they ejected two photopolymers in different percentages to reproduce the heterogeneous texture. The influence of some acquisition parameters (slice thickness, dose level and reconstruction algorithm) on the features was investigated. Unfortunately, only *texture* features (glcm and glrlm) were analysed, and basic information — such as mean and standard deviation — was therefore not reported.

A 3D printing approach was also adopted by Varghese et al. [112,114]. They created two similar cylindrical phantoms to be scanned with abdominal CT protocols. The first one [112] consisted of a homogeneous background made of urethane embedding three inserts made of acrylonitrile butadiene styrene (ABS), each with a different amount of printed material (*infill*). In the second [114], instead, six inserts were cast in a homogeneous background of ABS. The inserts were made of ABS, each with a different texture given once again by different infill values. Both phantoms were used to assess the influence of the scanner, the slice thickness, the tube voltage, the tube current and the type of feature extraction (3D versus 2D) on the value of the radiomic features. The first study evaluated also the impact of the reconstruction filters and the repeatability, while the second one the field of view. The major drawback of these two studies is that the inserts were intentionally not created to match the CT signal of anatomical tissue. Nevertheless, their phantoms were useful to create non-homogeneous geometric textures whose fabrication is reproducible.

More recently, Jimenez-del-Toro et al. [113], instead, designed and built a CT phantom with a potassium iodide solution-based paper-printing technique [196] to study the variation of reconstruction parameters (algorithm,

kernel and slice thickness), and the repeatability with and without phantom repositioning, by reproducing the abdominal volume with a metastasis in the liver. The main advantage of this approach is the ability to reproduce the signal voxel by voxel starting from a real CT image. In this regard they stated that "from one voxel to the other, an HU difference of 2 could be reliably achieved." However, they were not able to reproduce a signal below about −100 HU.

With the exception of Samei's work, the other mentioned studies did not report a clear comparison of their phantoms with real tissues, both in terms of the mean signal and of radiomic features.

Examples of radiomic phantoms in MRI and PET imaging

For completeness, let me also mention that similar studies have been carried out to design heterogeneous phantoms in PET and MRI.

In PET, 3D printed inserts were fabricated and filled with a solution based on ^{18}F-FDG to reproduce the shape and the heterogeneous uptake of NSCLC, for example by introducing a necrotic core [197]. The impact of the reconstruction algorithms, matrix size, scan duration, discretisation and segmentation methods was analysed.

MRI phantoms were assembled to reproduce the T_1 and T_2 relaxation times of the female pelvic region, using a solution of $MnCl_2$ to recreate the muscle signal and four cylindrical inserts filled with agar gel mixed with polystyrene spheres to mimic the tumour signal [198]. The feature repeatability and reproducibility were investigated by performing a test-retest analysis, and by changing parameters such as the scanner, the TE and the TR [199]. Rai et al. [200] produced 3D printed inserts made of resins visible in MRI, creating homogeneous and heterogeneous textures, to study the inter- and intra-scanner variability.

3.2 Design and fabrication of the lung phantom

In this section the analysis performed to identify and characterise the materials which will be used to create heterogeneous lung inserts is discussed. Two materials were chosen for this aim: the glycol-modified polyethylene terephthalate (PET-G), and the sodium polyacrylate. In order to estimate how good the final results were, we compared them with real lung lesions.

3.2.1 The choice of the materials

The signal of lung lesions and of lungs in patients

First of all, we selected a group of patients with NSCLC in order to identify the region of the Hounsfield scale that match the lung lesion signal. These

patients belonged to a dataset of CT images of patients with advanced lung adenocarcinoma, whose lesion segmentation was available. Since we wanted to characterise only the solid part of the lesion, we manually modified the contours to be sure that they were inside the lesion borders and that the lung was not involved. We selected 29 patients whose CT images were acquired with the use of the iodinated contrast medium, and 8 without it. The mean CT number was 78.67 ± 13.46 HU (median 79 HU, range 52 – 113 HU) for the first group, and 35.43 ± 2.25 HU (median 36, range 32 – 40 HU) for the second one. The standard deviations inside the VOI were 25.38 ± 4.48 and 16.39 ± 2.80 for the two groups, respectively. In **Figure 3.1** the distribution of the CT signal of one lung lesion is plotted compared to that extracted from a VOI drawn in a homogeneous water phantom with a similar volume. The shift to higher CT numbers and the larger variability of the lesion is clear, with a very small overlay between the two distributions.

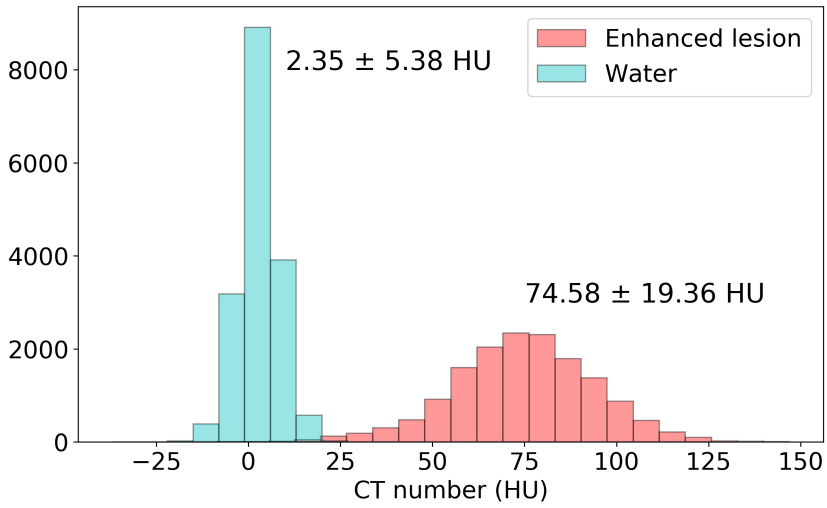
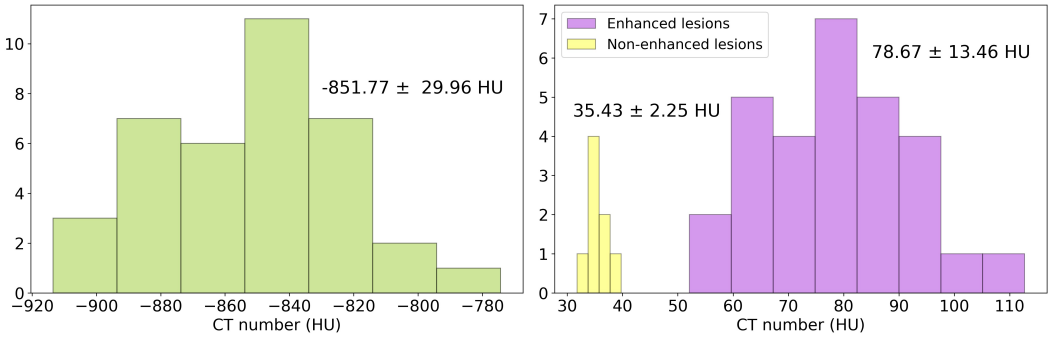


Figure 3.1: Comparison between the CT number distribution of an enhanced lung lesion (pink) and a water phantom (light blue). The mean CT signal and the standard deviation of the two VOIs are reported in the plot. The two CT images were acquired on the same scanner, at 120 kVp with a slice thickness and a space between the slices of 2.5 mm and were reconstructed with the ASIR 60% algorithm. The exposure was 5 mAs for the central axial slice of the lesion and 12 mAs for the phantom, while the pixel size was 0.75 and 0.70 mm, respectively.

Moreover, we drew regions of interest in various parts of the lung tissue on the same CT images of the patients above, avoiding vessels as far as possible, with the aim of identifying the lung CT signal. We found similar results between enhanced and non-enhanced images, with a mean CT number of -847.41 ± 29.43 HU (median -845 HU, range -894 – -774 HU) and a standard deviation of 44.21 ± 7.90 HU in the first case, and a mean CT number of -867.55 ± 28.06 HU (median -858 HU, range -913 – -836 HU) and a standard deviation of 47.53 ± 5.91 HU for the second group of patients.

Figure 3.2 summaries the CT numbers found in patients and used as the

reference values for the following analysis.



(a) **(b)**

Figure 3.2: Distribution of the mean CT numbers corresponding to the lung signal on the left **(a)** and to the lung lesion one on the right **(b)**, among the selected patients.

Material exploration

Starting from this very simple information, we analysed various objects, including food, household items and polymeric materials. We created ad-hoc samples and we scanned them on a CT scanner with an acquisition protocol comparable to the thoracic protocol adopted in patients. A VOI was delineated on the CT image for each material, avoiding the borders of the object, and some simple properties of the CT signal were extracted using the *Segmentation Statistics* tool in 3D Slicer (maximum, minimum, mean, median and standard deviation of the signal). Polystyrene- and cork-based objects and sponges were included in the analysis in order to identify materials that can be useful to reproduce the low CT signal of the lungs. The polymeric materials were fabricated by the C.I.Ma.I.Na (Centro di Eccellenza Interdisciplinare Materiali e Interfacce Nanostrutturati) group from the University of Milano, using additive, subtractive or mould-based techniques. Details about these techniques and the list of all the materials investigated can be found in Appendix C.

The material selected for further analysis was the **PET-G**. PET-G inserts, in fact, were able to cover the desidered CT signal of lung lesions. They were fabricated using the fused filament fabrication technique with a 3Dline 3DiElle Pro (Italy) 3D printer, by fusing the PET-G filament at 320°C and passing it through a nozzle with a diameter of 0.4 mm. The models were created with Autodesk Inventor 3D CAD, and processed with the slicer Raise3D Ideamaker.

In addition to the 3D-printed materials, we created inserts with a completely different approach based on the usage of **sodium polyacrylate**. The latter is a powdered polymer [201] which absorbs the liquid it gets in contact with, turning into gel and expanding considerably in volume.

In order to get familiarised with the sodium polyacrylate, we scanned several plastic containers filled with different quantities of sodium polyacrylate and water. More details on the fabrication procedure are in Appendix C. **Figure 3.3** shows one of these containers filled with sodium polyacrylate after the gelling with water. With this procedure we obtained inserts with a too low CT signal, below 0 HU. For this reason, we added a solution of iodinated contrast medium diluted with water to gel the sodium polyacrylate. The contrast medium used was the Ultravist® 370 mg/mL (Bayer, Germany), the same adopted during the CT examination of the patients.

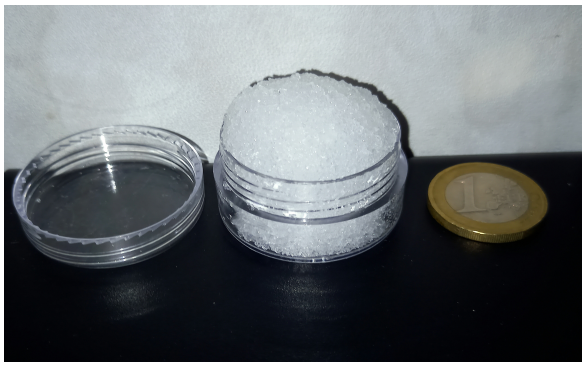


Figure 3.3: Example of a container used to investigate the properties of the sodium polyacrylate. The picture shows the mixture of powder and water at the end of the gelling phase. Each container was 5 mm in height and 20 mm in diameter.

In order to identify the contrast concentration that matched the range of attenuation coefficient of the enhanced lung lesions, we created various inserts with different concentrations of contrast medium. A CT acquisition of these objects is displayed in **Figure 3.4**.

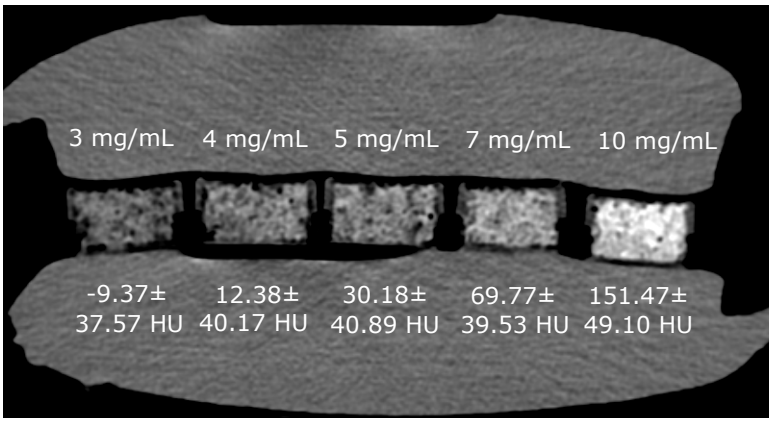


Figure 3.4: CT acquisition of some of the inserts created by changing the concentration of contrast medium. In the picture we also reported the concentration and the mean CT signal ± the standard deviation for each insert. The objects below and above the containers are two saline bags. The CT image is displayed using the mediastinal window (W = 350 HU and L = 40 HU).

As regards the material representing the lungs we chose the **powdered cork**, whose mean CT number was -835.48 ± 14.25 HU (median -836 HU, range $-900 - -769$ HU). Among the materials we investigated which had a CT signal similar to the lung (see Appendix C), powdered cork was in fact the one which required the least effort in the fabrication.

In the next sections, a thorough characterisation of the PET-G and the polyacrylate inserts is presented, which constitutes the starting point for the fabrication of the final inserts.

3.2.2 Characterisation of the materials: calibration curves

In this section the characterisation of the two materials introduced in the previous section (PET-G and sodium polyacrylate) through CT image acquisitions is discussed. The goal is to identify an appropriate configuration of these objects to mimic the texture of lung lesions.

PET-G

One of the most crucial parameters set in the model developed for the 3d printing of the PET-G insert is the infill value. The latter parameter describes the spatial density of the printed pattern: higher values correspond to a larger amount of material inside the volume of the object. In order to better characterise the behaviour of this material when the infill value changes, thirteen new PET-G inserts were fabricated, each with a different infill value (25%, 35%, 45%, 55%, 65%, 75%, 85%, 95%, 100%) and with the infill lines tilted by 45 degrees with respect to the model sides. Each insert had a size of $25 \times 25 \times 5$ mm^3 (width $\times$ length $\times$ height) and a layer height of 0.2 mm. **Figure 3.5** is a picture of the inserts with different infill values fabricated with a layer height of 0.2 mm. For the lower percentages, the infill line pattern is easily discernible.

The texture of the inserts is visible from the CT images[1] in **Figure 3.6** for the different infill percentages.

We drew a VOI within each insert in the axial slice, and collected information about the signal (mean, median and standard deviation), which is reported in **Table 3.1**. For some infill percentages (25%, 45%, 65%, 85%), the inserts were also replicated with a layer height of 0.1 mm, but the results between the two layers heights were comparable. For this reason, in the subsequent studies we will consider only the inserts with layer height of 0.2 mm.

[1]The PET-G inserts were scanned on a BrightSpeed CT Scanner (GE Healthcare, Waukesha, WI) at 120 kVp and 300 mA (exposure equal to 150 mAs), with 512×512 matrix size and a 0.488×0.488 mm^2 pixel size. In order to avoid possible partial volume artifacts due to the small thickness of the inserts (5 mm thick), the slice thickness was set to 0.625 mm.

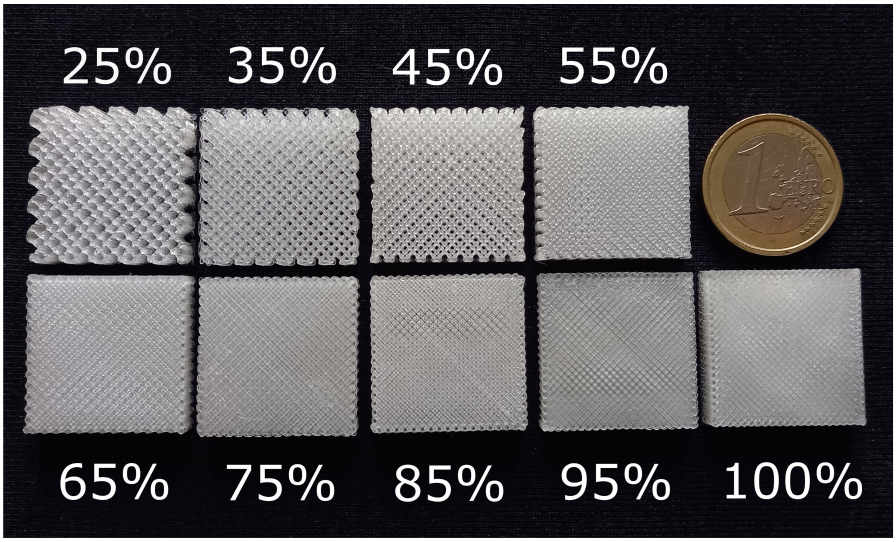


Figure 3.5: Photograph of the PET-G inserts with different infill values, and layer height of 0.2 mm.

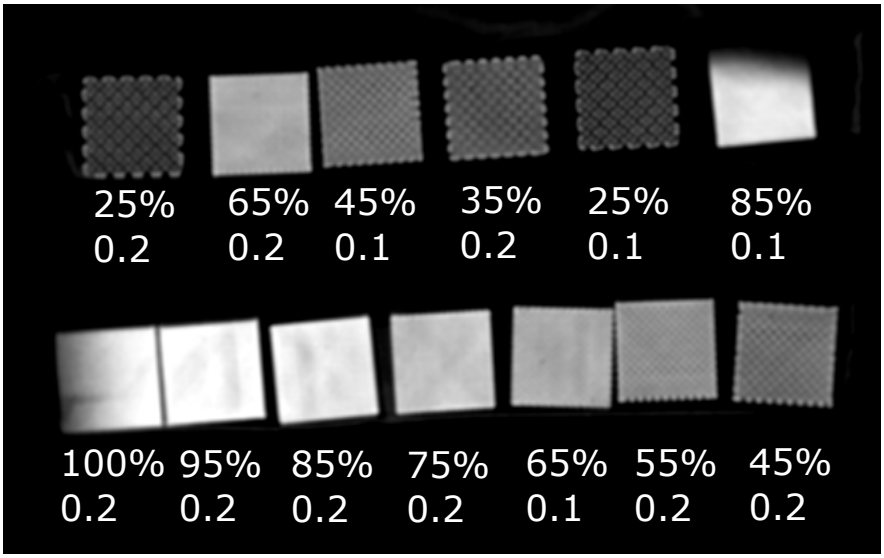


Figure 3.6: CT image of the 13 inserts made of PET-G with different infill percentages and layer heights. The information about the infill percentage and the layer height is reported below each insert. The CT image is displayed using the window W = 1000 HU and L = −426 HU.

The CT number exhibits a linear dependence on the infill percentage, as can be seen in **Figure 3.7**. This follows from the linear reduction of the volume of air within the insert as the infill value (and hence the volume of filament) increases. The interpolated CT number for infill value of 0% (meaning no filament at all) is roughly equal to −1000 HU, as expected since the latter is the CT number of air.

Insert	Min (HU)	Max (HU)	Mean (HU)	Median (HU)	SD (HU)
25%, 0.1 mm	-861	-586	-717.91	-715	45.49
45%, 0.1 mm	-623	-388	-492.07	-492	35.96
65%, 0.1 mm	-354	-196	-256.56	-255	19.46
85%, 0.1 mm	-134	120	-11.18	-10	48.37
25%, 0.2 mm	-836	-602	-717.57	-716	36.20
35%, 0.2 mm	-727	-490	-604.82	-605	37.03
45%, 0.2 mm	-597	-391	-492.22	-492	30.88
55%, 0.2 mm	-447	-293	-367.36	-367	23.12
65%, 0.2 mm	-340	-188	-257.49	-257	22.89
75%, 0.2 mm	-217	-55	-145.48	-145	24.91
85%, 0.2 mm	-140	74	-27.30	-24	35.25
95%, 0.2 mm	-12	177	115.12	122	37.46
100%, 0.2 mm	47	163	119.48	121	18.35

Table 3.1: General information about the CT signal of the 13 PET-G inserts.

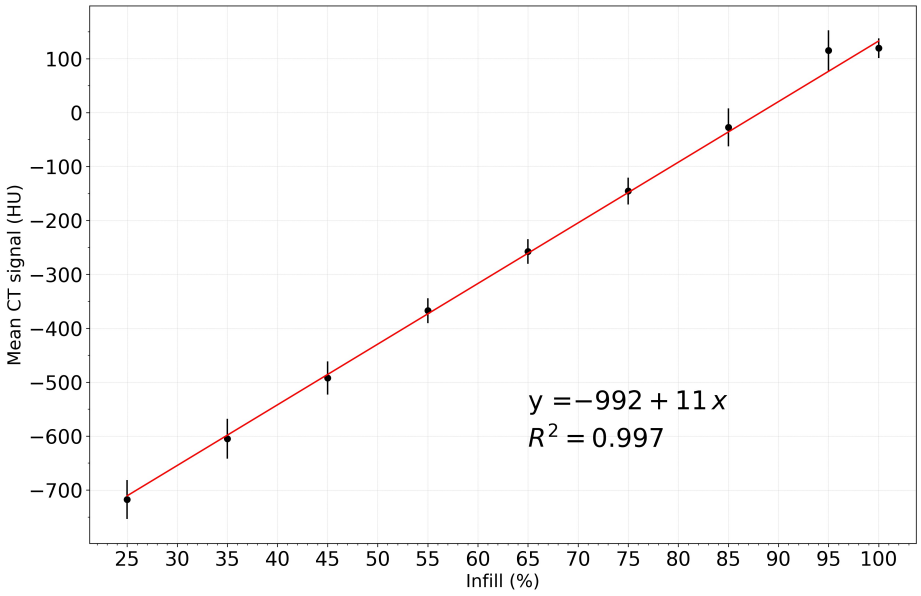
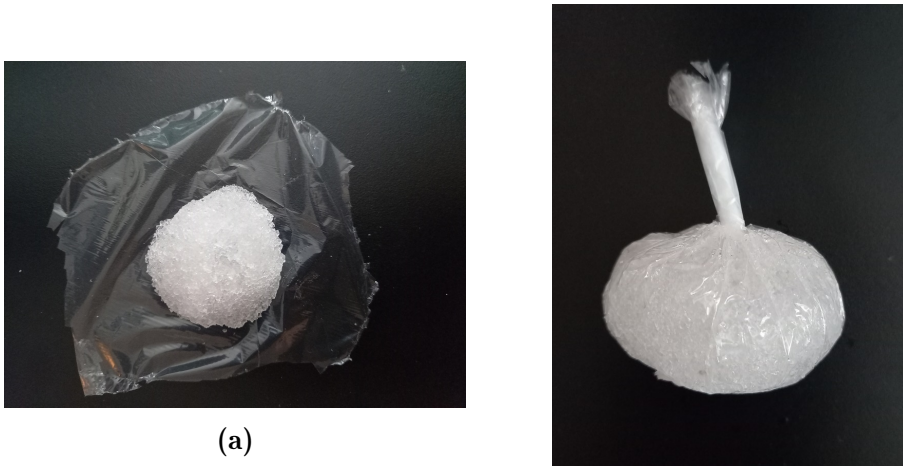


Figure 3.7: Characterisation curve for the preliminary PET-G inserts (for the layer height of 0.2 mm). R^2 is the coefficient of determination and describes how well the regression model fits the data (the higher is the value, the more the model matches the data).

Sodium polyacrylate

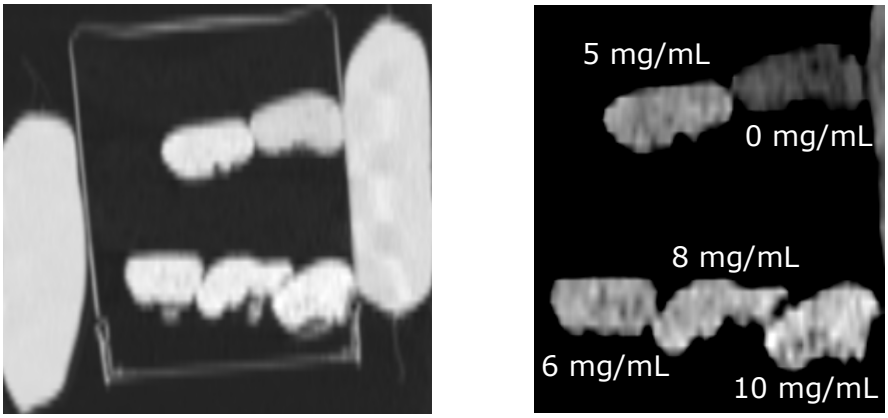
We created new inserts with contrast-medium concentrations of 0, 5, 6, 8, 10 mg/mL, as above. We then wrapped the mixture inside the container in a cling film. The mixture before and after the embedding is shown in **Figure 3.8**.



(a) **(b)**

Figure 3.8: Pictures of the polyacrylate-based compound before (**a**) and after (**b**) the embedding in the cling film.

Finally, we put the inserts inside a container filled with cork. In **Figure 3.9** we give a CT image[2] of the inserts within the container.



(a) **(b)**

Figure 3.9: CT image in the sagittal plane of the polyacrylate inserts with contrast-medium concentrations of 0, 5, 6, 8, 10 mg/mL. Figure (**a**) is displayed with the lung window to show the container, while figure (**b**) with the mediastinal one to show the texture of the inserts. The "white" objects outside the container in (**a**) are saline bags used to attenuate the radiation and simulate soft tissues.

The properties of the texture in terms of CT number, extracted from VOIs drawn in each insert, are listed in **Table 3.2** for the five concentrations.

[2]The container with the five inserts was scanned on a Revolution EVO CT scanner (GE Healthcare, Waukesha, WI) at 120 kVp, with tube current modulation, a slice thickness of 2.5 mm, a pixel size of 0.683 x 0.683 mm^2 and a 512 × 512 matrix size.

Insert	Min (HU)	Max (HU)	Mean (HU)	Median (HU)	SD (HU)
0 mg/mL	-166	12	-51.42	-48	25.37
5 mg/mL	-161	136	53.29	57	29.06
6 mg/mL	-129	156	67.93	75	36.79
8 mg/mL	-103	224	111.71	115	33.88
10 mg/mL	-70	267	138.34	142	46.48

Table 3.2: General information about the CT signal of the five polyacrylate inserts.

As for the infill calibration curve, the mean CT number of the polyacrylate inserts depends linearly on the concentration of the contrast medium, as can be observed from **Figure 3.10**.

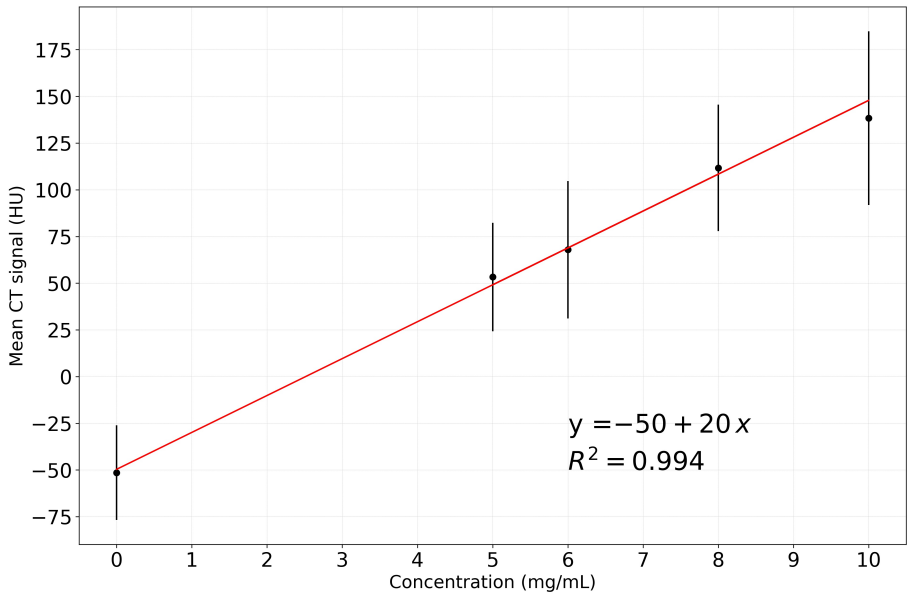
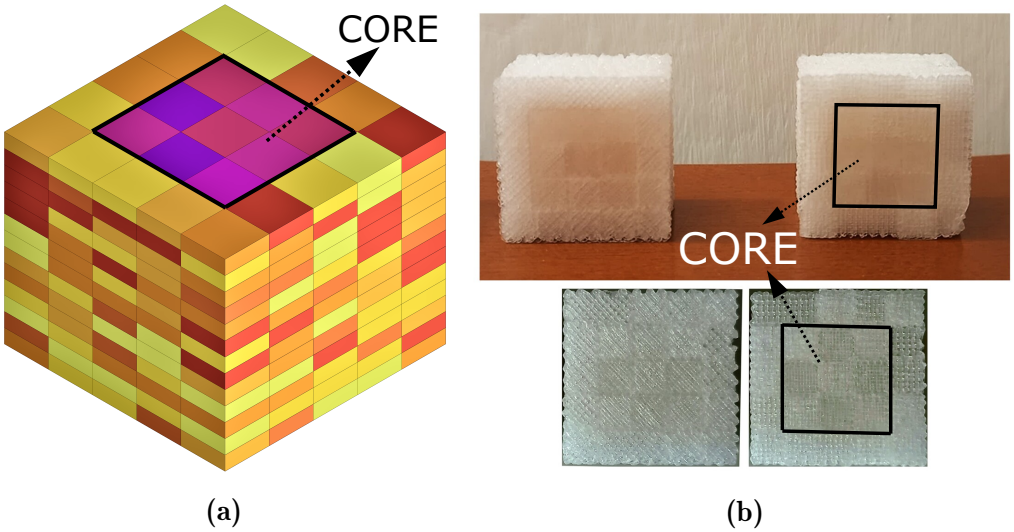


Figure 3.10: Characterisation curve for the preliminary inserts made of polyacrylate mixed with diluted contrast medium. R^2 is the coefficient of determination and describes how well the regression model fits the data (the higher is the value, the more the model matches the data).

In summary, the PET-G inserts allow us to cover a wider range of CT numbers, from about −700 HU to about 120 HU, by varying the infill values from 25% to 100% during the fabrication procedure. The minimum value of the CT number for the inserts made of sodium polyacrylate, instead, is approximately −50 HU, which is reached when no contrast medium is used. However, by increasing the contrast-medium concentration, the sodium polyacrylate reaches a higher CT signal with respect to the PET-G with infill percentage of 100%.

3.2.3 HeLLePhant: the assembly

In order to increase the heterogeneity of the inserts, we fabricated two new PET-G inserts with a size of 25×25×20.8 mm^3. Each insert consisted of 5×5×13 cells with the same size (5×5×1.6 mm^3) but different infill value. The two inserts were identical, except for the infill pattern. One was printed with infill lines tilted by 45 degrees with respect to the model sides (**PET1**), and the other with infill lines parallel to the model sides (**PET2**). More in detail, the inserts were produced by creating a $3 \times 3 \times 13$-cell core with higher infill percentages (between 75% and 100%) in the centre of the insert. The surrounding cells, instead, had infill values ranging between 50% and 75%. The infill value of each cell was selected randomly during the design of the model. Because of their structure made of small cells with different density, we denoted these inserts as "voxelated" PET-G inserts. **Figure 3.11** shows the model of the voxelated PET-G inserts for the 3D printing and the result of the printing.



(a) **(b)**

Figure 3.11: Voxelated PET-G inserts. Figure (**a**) shows the model used for the 3D printing, while figure (**b**) is a picture of the two PET-G inserts (PET1 on the left and PET2 on the right).

As regards the sodium polyacrylate, we created a total of nine inserts. Three inserts were produced, as described above, with contrast concentrations of 5, 6 and 8 mg/mL. We named them **uniform inserts**, because each of them was produced with a unique concentration. We fabricated six additional inserts by mixing manually multiple uniform inserts with a different concentration to increase the heterogeneity of the texture. We referred to them as **non-uniform inserts**. We created the following mixtures: 0 mg/mL, 5 mg/mL and 10 mg/mL; 0 mg/mL, 7 mg/mL and 10 mg/mL; 0 mg/mL, 5 mg/mL and 8 mg/mL; 0 mg/mL, 8 mg/mL and 10 mg/mL; 6 mg/mL with the addition of 2 mL of water; 0 mg/mL, 4.5 mg/mL and 7 mg/mL. We added more diluted

contrast medium to the inserts which, after a preliminary CT scan, had a mean CT signal under the target range.

We named the thoracic phantom proposed in this study *HeLLePhant*, Heterogeneous Lung Lesion Phantom. It comprises eleven inserts: two PET-G and nine polyacrylate-based. The inserts were put inside three plastic containers filled with powdered cork. The containers were put inside an empty body phantom from the NEMA IEC Body Phantom Set™ (Data Spectrum Corporation, Durham, USA), covered by saline bags. The cork is used to simulate the attenuation coefficient of the lung, while the saline bags of the soft tissue. **Figure 3.12** outlines the composition of the phantom.

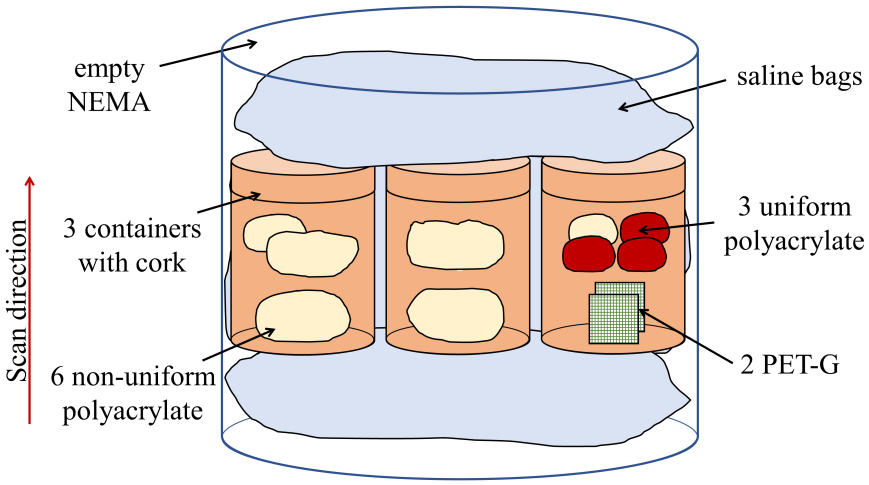


Figure 3.12: Schematic representation of the HeLLePhant.

3.2.4 HeLLePhant: comparison with lung tumours

We collected a set of CT images of patients affected by NSCLC, available at the European Institute on Oncology, in order to evaluate the similarity of the insert texture with that of the real tissue in terms of the radiomic characteristics.

All patients had a histological/cytological evidence of lung adenocarcinoma and underwent the same CT acquisition protocol, which was the one adopted by the IEO for the clinical examination of the chest. It included: acquisition in the portal phase after the injection of contrast medium, automatic tube current modulation, slice thickness and slice spacing of 2.5 mm, 512×512 matrix size, and iterative reconstruction algorithm. The images were selected so that they were acquired on the same CT scanner, the Discovery CT750 HD scanner (GE Healthcare, Waukesha, WI, USA), and with a tube voltage of 120 kVp. Patients with a lung tumour smaller than 3 cm^3 and larger than 60 cm^3 were excluded, in order to focus on a range of volumes sufficiently comparable with the insert ones. A total number of 29 patients were considered for the comparison with the inserts of the HeLLePhant.

In addition to this group of patients we also considered a commercial phantom with homogeneous textures to assess how much this type of texture differs from the real tissue. For this purpose, we used the Catphan® 424, developed by the Phantom Laboratory (Salem, NY, USA). This phantom is a cylindrical object composed of multiple modules, each useful for a different purpose in the CT quality assurance. The following modules were considered in this study: the CTP486 Module, whose material had a CT number within 2% of water density (20 HU), the epoxy background of the CTP404 Module, one acrylic insert of the CTP404 Module, and the plug of the CTP422 Module. The corresponding VOIs are labelled as **Catph1**, **Catph2**, **Catph3** and **Catph4**, respectively. We chose these materials because they were the most similar to the lung lesions in terms of CT number.

We scanned the two phantoms using the same protocol and the same CT scanner of the patients mentioned just above. **Figure 3.13** and **Figure 3.14** show some CT axial slices of the two phantoms, containing the investigated VOIs.

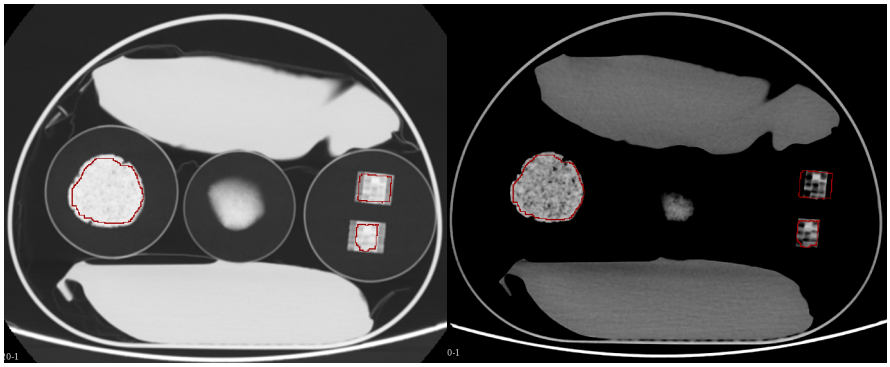


Figure 3.13: CT images in the axial plane of the HeLLePhant phantom using the mediastinal window on the right (W = 350 HU and L = 40 HU), and the lung window on the left (W = 1400 HU and L = −500 HU). Each image shows the three containers surrounded by the saline bags with the two PET-G inserts (on the right), and one sodium polyacrylate insert (on the left).

We then contoured one lesion per patient, and all the inserts of the HeLLePhant. The VOIs were delineated avoiding the border to reduce the partial volume artifacts, and not to include the surrounding tissues. For each of the two PET-G inserts we identified two VOIs: one for the core, and the other for the entire insert (core + surrounding cells). As regards the Catphan®, we performed a cylindrical segmentation in each of the four materials. The CT texture of some lung lesions and some inserts of the two phantoms is displayed in **Figure 3.15**. **Table 3.3** lists the basic CT information about the inserts and the 29 patients.

As can be seen in **Table 3.3**, the PET-G inserts showed a very low signal (mean CT number less than −100 HU) and a very high variability (standard deviation over 150 HU), compared to the patient values, when analysed in

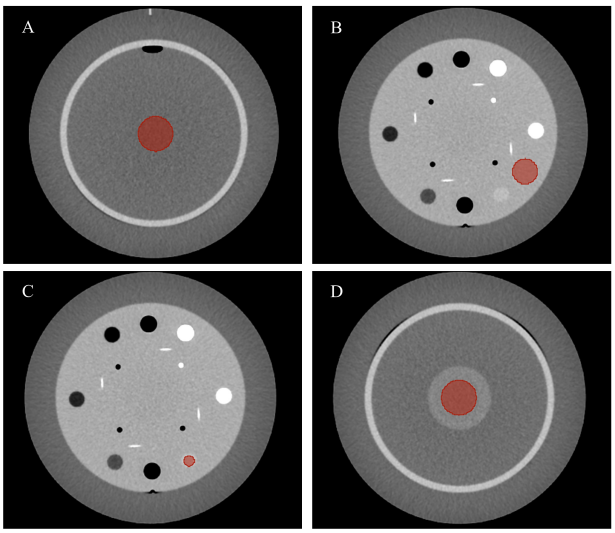


Figure 3.14: CT images in the axial plane of the Catphan phantom. The four VOIs are shown: Catph1 (**A**), Catph2 (**B**), Catph3 (**C**) and Catph4 (**D**). The CT image is displayed using the mediastinal window (W = 350 HU and L = 40 HU).

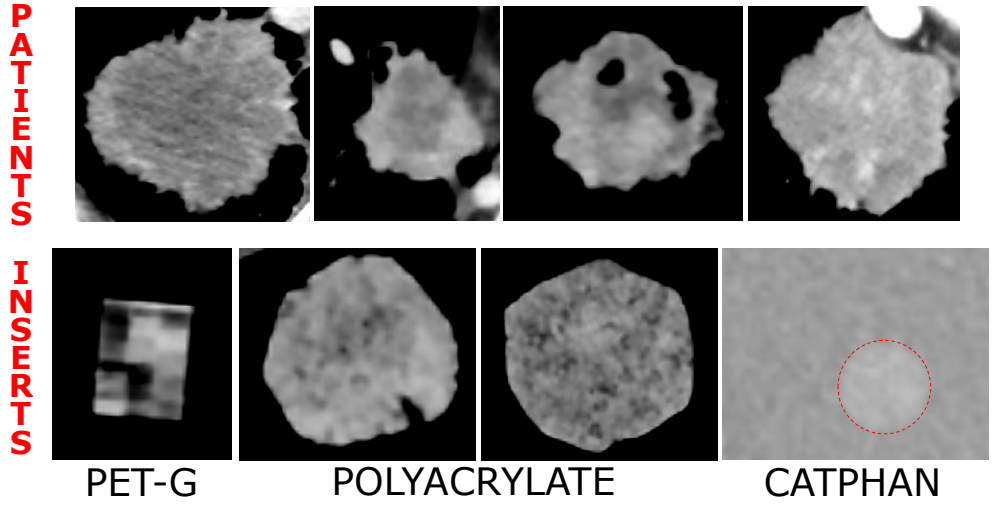


Figure 3.15: Visual comparison of the CT images of some lung tumours (above) and of the inserts (below). One PET-G insert, one uniform and one non-uniform polyacrylate inserts of the HeLLePhant, and the acrylic insert with the epoxy background of the Catphan are shown in the picture. The CT images are displayed using the mediastinal window (W = 350 HU and L = 40 HU). Using this window only the core of the PET-G insert is visible.

their entirety. For this reason, we considered for the subsequent analysis only the data extracted from the core of the inserts.

We extracted the radiomic features from the patient and the phantom images using the same procedure, described in Chapter 2 (Section 2.2.2), with a bin width of 25 HU, and we considered only the features obtained from the non-filtered images (original type).

Insert	**Min** (HU)	**Max** (HU)	**Mean** (HU)	**Median** (HU)	**SD** (HU)	**Volume** (cm^3)
PET1_entire	-550	152	-122	-74	174	6.206
PET1_core	-294	152	-6	10	94	3.057
PET2_entire	-788	159	-101	-49	169	6.206
PET2_core	-258	159	15	31	83	3.057
Pol_unif_1	-274	113	37	41	35	5.645
Pol_unif_2	-193	135	45	50	40	5.873
Pol_unif_3	-148	193	90	94	40	6.595
Pol_non-unif_1	-271	149	62	68	42	24.023
Pol_non-unif_2	-147	135	63	65	31	19.255
Pol_non-unif_3	-185	128	61	63	26	13.384
Pol_non-unif_4	-240	176	51	53	44	22.171
Pol_non-unif_5	-81	128	65	66	25	5.358
Pol_non-unif_6	-115	110	37	38	26	20.873
Catph1	-22	46	10	9	8	12.785
Catph2	76	127	100	99	7	6.763
Catph3	102	144	125	124	7	0.971
Catph4	17	85	47	46	9	7.904
Patients	-374 ± 355	179 ± 121	62 ± 22	65 ± 18	29 ± 11	19.599 ± 14.584

Table 3.3: General information about the CT signal and the volume of the inserts of the HeLLePhant and the Catphan, and of the 29 NSCLC patients.

The radiomic similarity between the inserts and the patients was evaluated by calculating the following quantities. First of all, we calculated the 10^{th} and the 90^{th} percentile from the distribution of the values of the feature i among the 29 patients, for each feature i. Then, we used these percentile values to calculate the *patient range* ($range_{(i,pts)}$), as

$$range_{(i,pts)} = [10^{th}percentile,\ 90^{th}percentile]_{i,pts}\,. \tag{3.1}$$

We compared the inserts (from the HeLLePhant and the Catphan) and the patient lesions by counting the number of features for each insert j (P_j) whose value fell in the corresponding $range_{(i,pts)}$,

$$\begin{aligned} P_j &= \sum_{i=1}^{i=N} P_{ij}\,, \\ P_{ij} &= \begin{cases} 1 & f_{ij} \in range_{(i,pts)} \\ 0 & otherwise \end{cases}\,, \end{aligned} \tag{3.2}$$

where N is the total number of extracted features (153 in case of features from non-filtered images), and f_{ij} is the value of the feature i for the insert j. In **Figure 3.16** we reported the percentages of features that resulted as similar to

those extracted from the lung lesions, for each insert of the HeLLePhant and of the Catphan. The results for each feature category and each insert is shown in **Figure 3.17**.

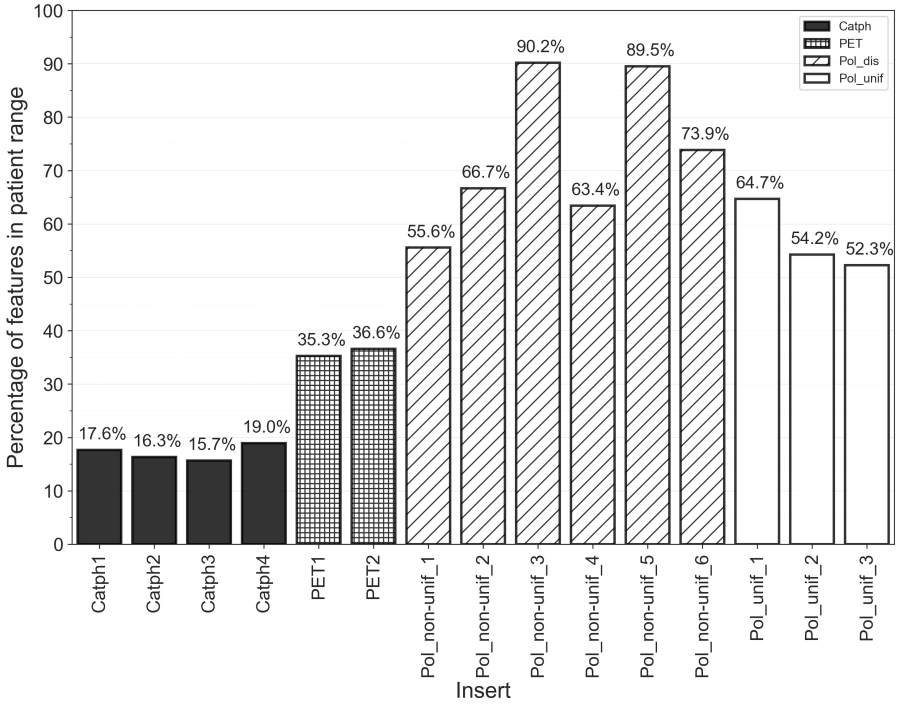


Figure 3.16: Percentage of features falling in the *patient range* for each insert.

3.2.5 Discussion

In the above sections we analysed different materials to reproduce the CT signal of a human thorax with lung tumours. We proposed two materials for the lung lesions, each fabricated with a distinct technique.

The first approach was based on printing 3D objects using the FFF technique and a thermoplastic material — the PET-G — as basic constituent. We used a single material (PET-G) with density varying inside the same object, which we obtained by changing spatially the infill value. Since this was the very first prototype, we printed the PET-G inserts with the simplest shape, namely parallelepiped. Our focus was in fact on the internal structure of the inserts: we wanted to understand whether a structure obtained by combining different infill values would be robust enough, and whether its CT signal was suitable for our purposes. The study of more complex external structures is left for future investigation.

The second technique we investigated, instead, was based on a more handcrafted approach, by creating inserts made of powdered sodium polyacrylate plus diluted contrast medium. Thanks to its gel-like consistency, the compound was easily moulded to take a form resembling tumours.

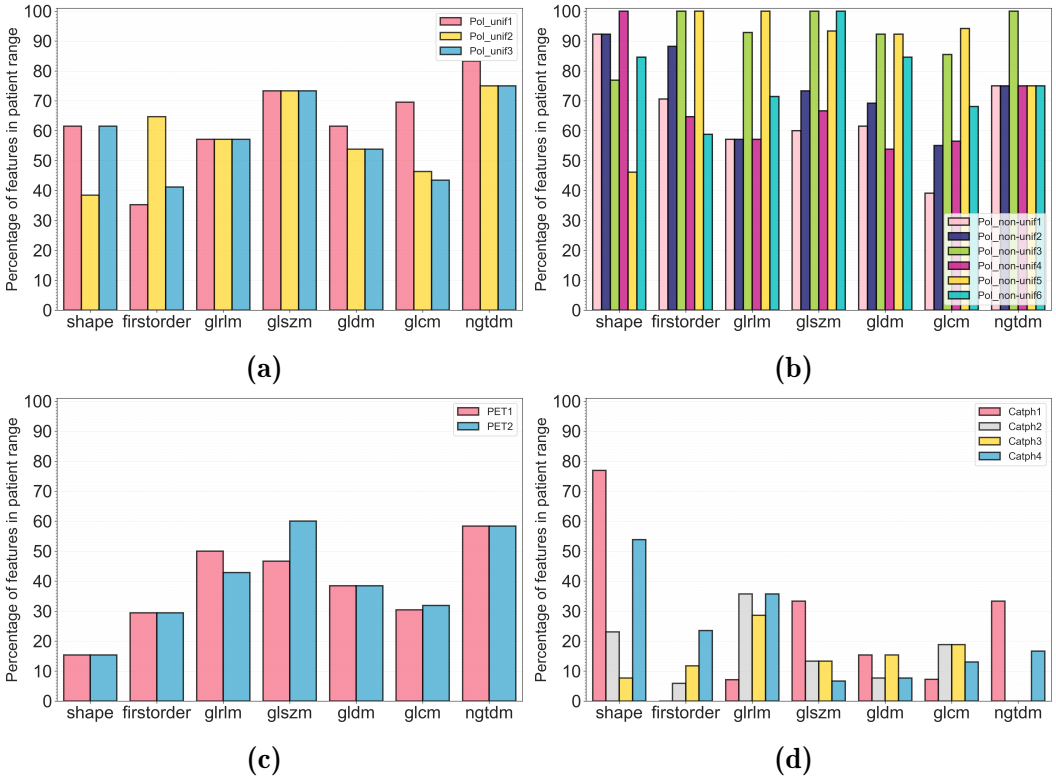


Figure 3.17: Percentage of features falling in the *patient range* for each feature category and for each insert: uniform polyacrylate inserts **(a)**, non-uniform polyacrylate inserts **(b)**, core part of the PET-G inserts **(c)**, and Catphan inserts **(d)**.

In order to compare the impact of a homogeneous and a heterogeneous texture on the feature values, we included in the analysis some homogeneous materials from a commercial phantom (Catphan®).

We assessed the similarity between the 11 inserts (2 PET-G and 9 polyacrylate) and the lung lesions of 29 NSCLC patients, by evaluating if the absolute value of each radiomic feature of the insert fell into the corresponding patient range (see eq. 3.3). In general, we found that the polyacrylate inserts better represented the tumours. The percentage of similar features was, in fact, over 50% for all the polyacrylate inserts. The *Pol_non-unif_3* and *Pol_non-unif_5* inserts were the most similar to the 29 NSCLC lesions (**Figure 3.16**). These two inserts had a median value and a standard deviation of the CT signal inside the VOI closer to the patient values. On the contrary, the *Pol_non-unif_2* showed less similarity with patients, despite the mean and the standard deviation are close to those of real lesions. This means that there are textural characteristics, describing local information, which are not captured by first-order features and are different between these inserts and patients.

From **Figure 3.17-B** we can see that the main limitation of the *Pol_non-unif_5* is its shape. This insert, in fact, had a small volume (5.358 cm^3) compared to patient lesions (19.599 $cm^3 \pm 14.584$), and this impacts on the volume feature and the associated features (i.e. SurfaceArea and diameter).

However, the volume of this type of insert can be increased during the fabrication procedure, and therefore the result obtained with the *Pol_non-unif_5* insert is not an indication of a limitation of this technique.

The Catphan inserts, instead, showed the worst similarity in all the categories of features (see **Figure 3.17-D**). Less than the 20% of all the features, in fact, fell within the lesion range for all the four inserts. We observed a good agreement with patients in the *shape* features of the *Catph1* and *Catph4* inserts thanks to the similarity of the VOI size.

The PET-G inserts exhibited a similarity in between that of the polyacrylate and of the Catphan inserts. For the PET-G inserts, about 35% of the features were compatible with patients. In general, we observed a similar behaviour in the two PET-G inserts, despite the different 3D printing pattern. Low similarity was observed in the *shape* features (**Figure 3.17-C**), for both the inserts. This result was due to the small size of the inserts and to an overly regular shape of the contours we made, far from being tumour-like. For example, in **Figure 3.18** the value of the feature *Sphericity* of each analysed insert is compared to the patient range (red line). The *Sphericity* feature is defined as

$$sphericity = \frac{(36\pi V^2)^{1/3}}{A}\,, \tag{3.3}$$

where V is the volume and S is the surface area of the VOI. It measures the compactness of the VOI, with higher values for shape with smaller surface for a fixed volume (less complex surfaces). From this plot, it can be noticed that the more irregular shape of the tumour was better matched by the polyacrylate inserts.

The strong difference of the Catphan® inserts compared to the patients is probably associated to the high homogeneity of the texture. Conversely, the PET-G inserts exhibited a very high heterogeneity with the presence of areas with similar grey-level patterns. **Figure 3.19** shows the value of the *Uniformity* feature for the inserts and the patients. The low uniformity found in the *Pol_unif* inserts was quite unexpected, since they were created using a unique contrast concentration. A possible explanation for this result may be the formation of cracks in the gel when it was enclosed in the cling film, thus highlighting how difficult it was to control the degree of heterogeneity in the polyacrylate inserts during the fabrication.

One noteworthy property of the PET-G insert is the possibility of producing a cavity inside the lesion, which appears as a "black" area because of its very low density. This characteristic can be evaluated with the *Cluster_Tendency* feature from the glcm category, which measures the presence of clusters (groups of voxels with the same grey-level intensities) in the texture. The mean value of this feature in the 29 patients was 33.87 ± 134.89 (median equal to 4), while for the two PET-G inserts it was much higher (51.23 for the PET1_core and 40.10 for the PET2_core). However, considering for example the texture of one of the lesions with a cavity from the patient group, which corresponds to the

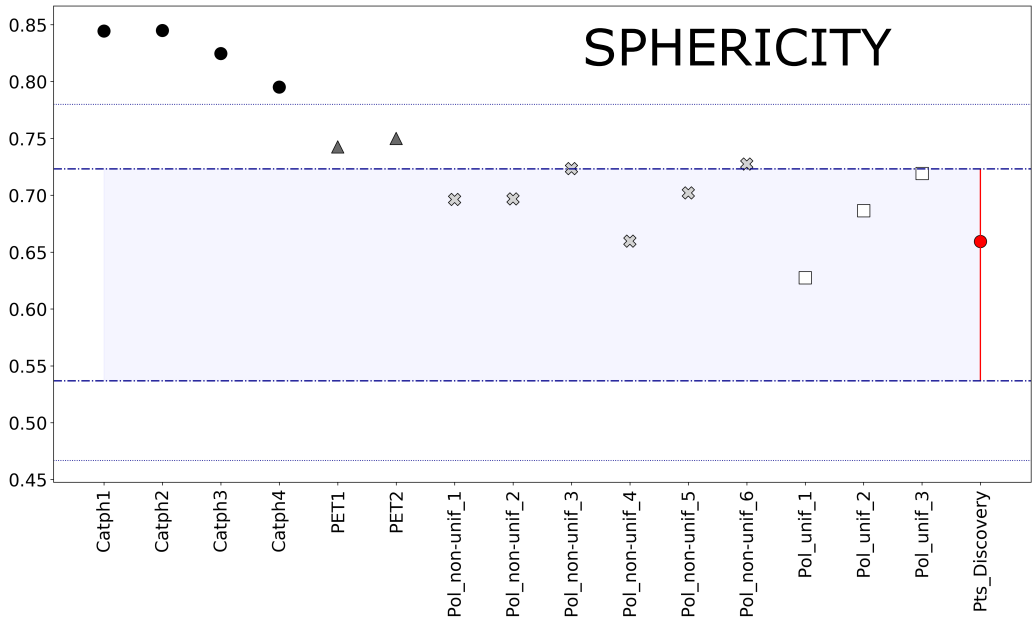


Figure 3.18: Comparison of the absolute value of the *Sphericity* feature from the *shape* category for the inserts and the group of NSCLC patients. The blue area highlights the patient range, while the thinner dotted line delimits the range between the maximum and the minimum values encountered in the patient group.

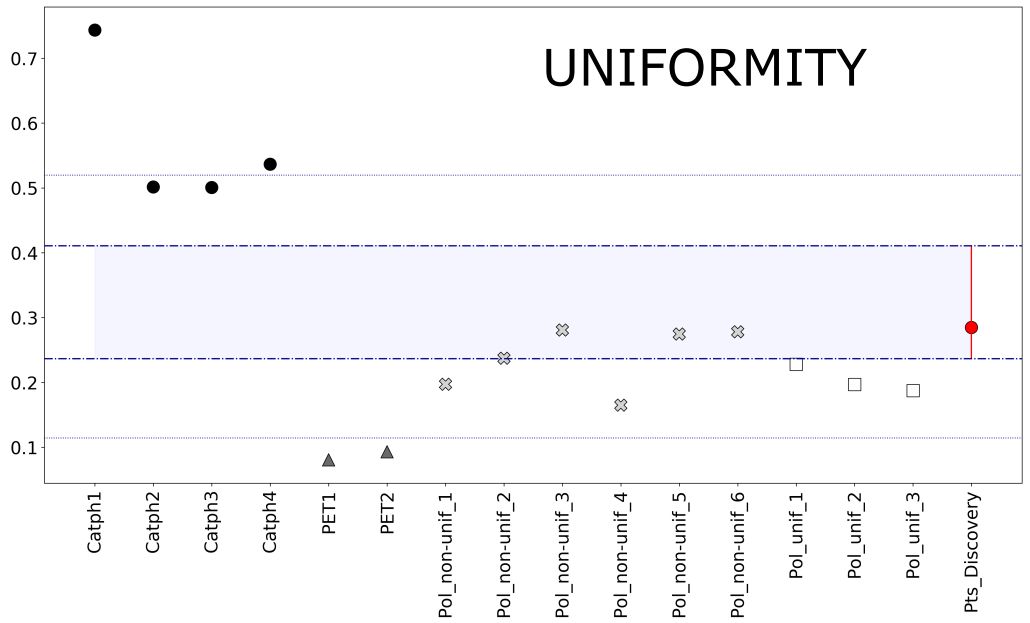


Figure 3.19: Comparison of the absolute value of the *Uniformity* feature from the firstorder category for the inserts and the group of NSCLC patients. The blue area highlights the patient range, while the thinner dotted line delimits the range between the maximum and the minimum values encountered in the patient group.

third lesion from the left in **Figure 3.15**, the value of this feature was equal

to 99.67, much higher than the values mentioned just above. Values similar to typical patient lesions, where no cavities are present, were obtained for the *Cluster_Tendency* feature with the polyacrylate inserts (range between 3.79 and 10.84 among the 9 inserts). The creation of this kind of area was in fact much more difficult with the polyacrylate-based material. A possible solution may be the incorporation of dedicated objects inside the compound during the fabrication of the inserts or the creation of separated compartments with different densities. Values below 1 were, instead, extracted from the Catphan® inserts, due to the uniformity of their texture.

In the next sections some applications of the HeLLePhant for radiomic purposes will be shown, such as the assessment of feature robustness.

3.3 Radiomic analysis with HeLLePhant

This section is dedicated to the discussion on how we used the phantom described in the above sections to investigate the behaviour of the radiomic features in controlled conditions. First of all, we scanned the phantom repeatedly with a fixed acquisition protocol, separately on two CT scanners. In this way we characterised the feature repeatability for each insert and each scanner. This analysis was performed twice, each time with a different tube voltage (100 kVp and 120 kVp). We identified a group of non-repeatable features in all the investigated configurations. Then we compared different acquisition parameters (Scanner 1 versus Scanner 2, 120 kVp versus 100 kVp) to evaluate feature reproducibility by changing one parameter at the time.

3.3.1 Repeatability analysis

Materials and methods

We scanned the HeLLePhant ten times without changing neither the acquisition parameters nor its position on the table of the CT scanner. The acquisition protocol was the same used in the clinical practice for the Discovery CT750 HD scanner (GE Healthcare, Waukesha, WI, USA), which we also adopted in Section 3.2.4 for the comparison between the phantom and the patients (automatic tube current modulation, slice thickness and slice spacing of 2.5 mm, 512×512 matrix size, pixel space of 0.70×0.70 mm^2, IR60, and standard convolution kernel). The repeated acquisitions were performed at 120 kVp and at 100 kVp. We repeated the same procedure on the Optima CT660 scanner (GE Healthcare, Waukesha, WI), both at 120 and 100 kVp. The images acquired on this scanner were reconstructed with the IR50, the IR blending level used in the clinical examination.

The NI of the CT images of the phantom was 14, 17, 18 and 20 for the acquisition on the Optima CT660 scanner at 100 kVp and 120 kVp, and on the Discovery CT750 HD scanner at 100 kVp and 120 kVp, respectively.

All the images acquired on the same scanner were co-registered. For this reason, the same segmentation file was used for all the images. In this way we avoided the introduction of biases due to the contouring procedure. For the acquisition on the Discovery CT750 HD scanner, the same segmentations used in Section 3.2.4 were considered also for the radiomic analysis. As regards the second scanner, instead, the segmentations were obtained from the previous ones through a rigid registration, since the phantom images from the two scanners did not perfectly match, despite the alignment of the phantom with the internal lasers of the scanner. A slight manual correction was required and performed a posteriori by comparing visually the segmentation on the CT images from the two scanners.

For all the inserts and all the image configurations we extracted the features as indicated in Chapter 2 (Section 2.2.2), with a bin width of 25 HU and 5 HU. We included in the analysis only the non-filtered features.

We assessed the repeatability of the features using the **coefficient of variation** (CV) and the **intra-class correlation coefficient** (ICC) [202].
The CV of the ith feature for the jth insert (CV$_{ij}$) for a fixed acquisition protocol is calculated as

$$CV_{ij} = \frac{\text{mean}_{10,\,ij}}{\text{standard deviation}_{10,\,ij}}\,, \tag{3.4}$$

where the mean and the standard deviation are calculated among the features extracted from the ten repeated acquisitions. The ICC for the ith feature is defined as

$$ICC_i = \frac{MS_R - MS_E}{MS_R + (k-1)\,MS_E + \frac{k}{n}(MS_C - MS_E)}\,, \tag{3.5}$$

where MS_R, MS_E and MS_C are the mean squares for the subjects, the error, and the measurements/raters, respectively; k is the number of measurements/raters, and n is the number of subjects. This ICC definition is based on the two-way mixed effect model for absolute agreement for single rater/measurement, which is the recommended model for test-retest measures [203].

The ICC ranges between 0 and 1, indicating no agreement or perfect agreement among raters/measurements, respectively. Conversely, for the CV the best reliability is obtained when its value is 0. For each insert we considered as repeatable the features yielding CV ≤ 0.10 [111]. Excellent repeatability is associated to an ICC greater than 0.90, good for ICC between 0.75 and 0.90, moderate for ICC between 0.50 and 0.75, and poor for ICC less than 0.50 [203]. We calculated the ICC with the *irr* package in R, using the function `icc(dataframe, type = "agreement", model = "twoway")`. In the computation of the ICC, the acquisition agreement is evaluated considering all the subjects — i.e. the inserts — all together for a given feature, while the CV is calculated for each feature and each insert separately (as can be observed from eqs. 3.4 and 3.5). The ICC, in fact, compares the multiple measurements

related to the same subject and at the same time considers the variability among the different subjects for each single measurement. In this preliminary investigation the inserts are, however, not fully able to reproduce the range of heterogeneity found in real lung lesions, especially if the PET-G inserts are included in the analysis. For this reason, there may be an overestimation of the repeatability with the ICC, compared to what we would expect if a more representative population of inserts or even real lesions were used.

The features belonging to the *shape* category were not included in the repeatability analysis since the contours were the same among the repeated acquisitions, as mentioned above. Therefore, we analysed 140 features in total.

For the sake of simplicity, hereinafter the Discovery CT750 HD scanner will be denoted as Scanner 1 and the Optima CT660 scanner as Scanner 2.

Results

In **Table 3.4** and **Table 3.5** the percentage of features with CV ≤ 0.10 is displayed for all the inserts of the HeLLePhant, for the 100 kVp and the 120 kVp, for the 25 HU and the 5 HU bin widths, and for the Scanner 1 and the Scanner 2. This percentage is over 70% in all the configurations, indicating that the majority of the features is repeatable. However, there is not a characteristic trend among the configurations nor among the inserts. It can be observed that the PET-G inserts show almost always a very high repeatability.

Insert	**Features with CV $\leq$ 0.1 (%)**			
	120 kVp 25 HU	**100 kVp 25 HU**	**120 kVp 5 HU**	**100 kVp 5 HU**
PET1_entire	100%	94%	93%	93%
PET1_core	92%	98%	98%	99%
PET2_entire	100%	95%	100%	100%
PET2_core	99%	92%	93%	93%
pol_non-unif_1	92%	98%	94%	100%
pol_non-unif_2	74%	79%	84%	84%
pol_non-unif_3	97%	85%	100%	99%
pol_non-unif_4	99%	91%	99%	93%
pol_non-unif_5	75%	76%	83%	77%
pol_non-unif_6	78%	79%	81%	84%
pol_unif_1	85%	89%	92%	99%
pol_unif_2	88%	86%	86%	94%
pol_unif_3	84%	81%	84%	83%

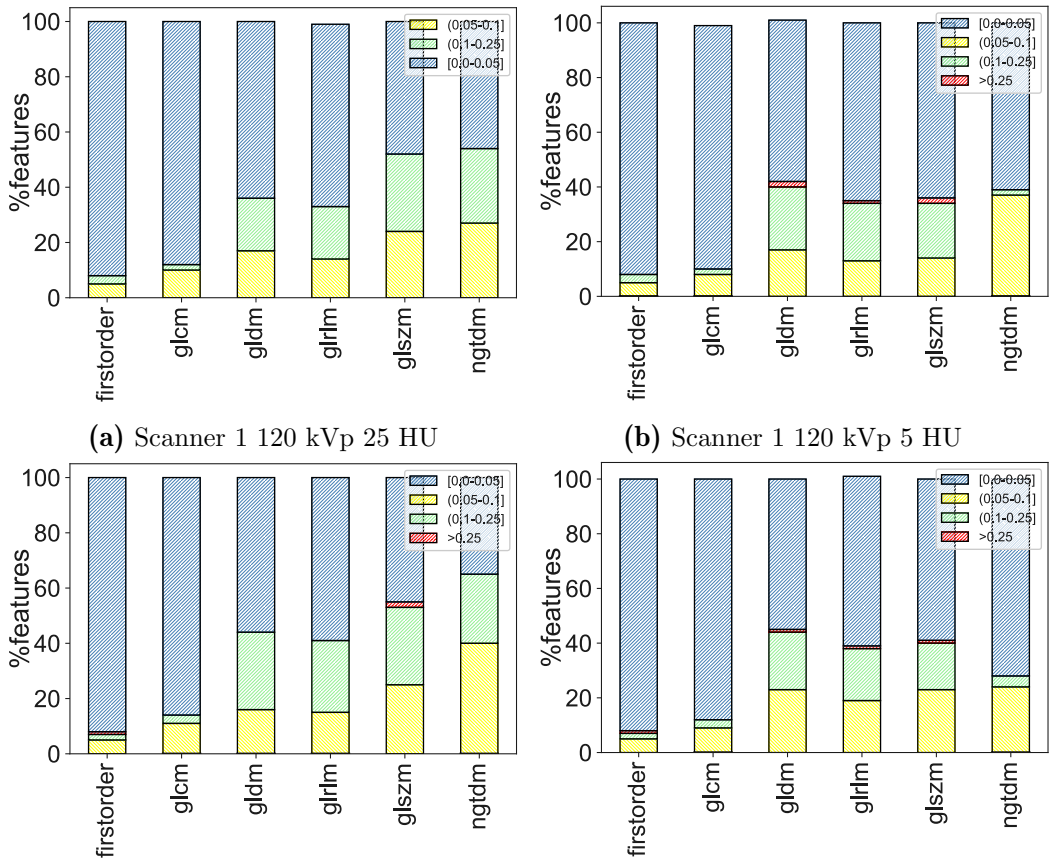
Table 3.4: Percentage of features with CV $\leq$ 0.1 for each insert type and for the four configurations of acquisition/extraction settings on the Scanner 1.

Insert	Features with CV $\leq$ 0.1 (%)			
	120 kVp 25 HU	**100 kVp 25 HU**	**120 kVp 5 HU**	**100 kVp 5 HU**
PET1_entire	100%	94%	93%	94%
PET1_core	93%	92%	91%	93%
PET2_entire	97%	98%	94%	94%
PET2_core	92%	97%	94%	91%
pol_non-unif_1	99%	97%	99%	94%
pol_non-unif_2	75%	76%	99%	86%
pol_non-unif_3	84%	89%	85%	84%
pol_non-unif_4	96%	100%	100%	96%
pol_non-unif_5	74%	76%	80%	80%
pol_non-unif_6	92%	79%	94%	84%
pol_unif_1	78%	78%	84%	91%
pol_unif_2	84%	95%	89%	99%
pol_unif_3	89%	97%	98%	100%

Table 3.5: Percentage of features with CV $\leq$ 0.1 for each insert type and for the four configurations of acquisition/extraction settings on the Scanner 2.

Focusing on the different feature categories, we counted the number of features with the CV less than in 0.05, between 0.05 and 0.1, between 0.1 and 0.25 and finally larger 0.25. In the evaluation of these percentages all the 13 inserts were considered, taking separately the two scanners, voltages and bin widths. **Figure 3.20** and **Figure 3.21** show the results for the Scanner 1 and the Scanner 2. The firstorder and glcm are the categories with the highest number of features with low CV values ($\leq$ 0.05). The CV results for Scanner 1 at 25 HU at 100 kVp and 120 kVp are reported in the supplementary materials of ref. [204].

Moreover, we evaluated the ICC considering all the 13 inserts and the 10 repetitions, for each of the eight configurations (two scanners, two voltages, two bin widths). **Table 3.6** summaries these results, reporting the percentage of features in the only two populated ICC ranges ([0.75, 0.90) and [0.90, 1]). We obtained that almost all the features had an excellent agreement among the repeated measurements ($ICC \geq 0.90$), with a percentage of repeatable features that varies among the configurations. All the features, however, showed a good repeatability, meaning $ICC \leq 0.75$, in all the configurations. When only the polyacrylate inserts were considered in the computation, the ICC of all the features was slightly reduced and the number of features with a $ICC < 0.90$ increased at the expense of those belonging to the range of excellent agreement. Nevertheless, also in this case all the features had a good agreement among repetitions ($ICC \geq 0.75$). Finally, we observed that all the features with $ICC < 0.90$ belonged to texture categories.



(a) Scanner 1 120 kVp 25 HU

(b) Scanner 1 120 kVp 5 HU

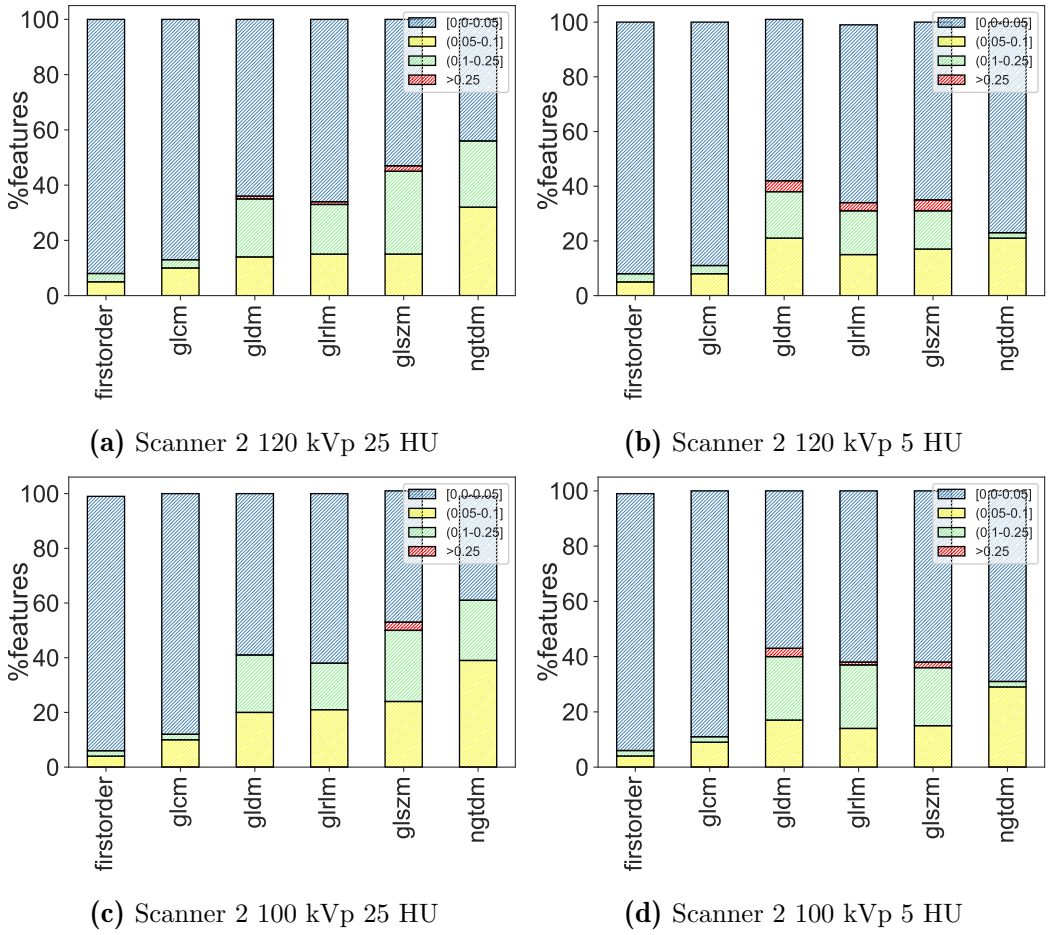
(c) Scanner 1 100 kVp 25 HU

(d) Scanner 1 100 kVp 5 HU

Figure 3.20: Percentage of features for each category (*shape* features excluded) and for Scanner 1 with a CV value in one of the following ranges: [0.0.05], (0.05, 0.10], (0.10, 0.25] and > 0.25.

Configuration	Features (%) with ICC in range:			
	[0.90, 1.00]	**[0.90, 1.00] only poly**	**[0.75, 0.90)**	**[0.75, 0.90) only poly**
S1 - 120 kVp - 25 HU	97%	89%	3%	11%
S1 - 100 kVp - 25 HU	98%	92%	2%	8%
S2 - 120 kVp - 25 HU	97%	90%	3%	10%
S2 - 100 kVp - 25 HU	99%	90%	1%	10%
S1 - 120 kVp - 5 HU	94%	92%	6%	8%
S1 - 100 kVp - 5 HU	94%	94%	6%	6%
S2 - 120 kVp - 5 HU	94%	92%	6%	8%
S2 - 100 kVp - 5 HU	94%	95%	6%	5%

Table 3.6: Percentage of features in the repeatability analysis with ICC between 0.75 and 0.90 and ≥ 0.90, for the indicated configurations. "S1" stands for Scanner 1 and "S2" for Scanner 2. The term "only poly" means that only polyacrylate inserts were included.



(a) Scanner 2 120 kVp 25 HU

(b) Scanner 2 120 kVp 5 HU

(c) Scanner 2 100 kVp 25 HU

(d) Scanner 2 100 kVp 5 HU

Figure 3.21: Percentage of features for each category (*shape* features excluded) and for Scanner 2 with a CV value in one of the following ranges: [0.0.05], (0.05, 0.10], (0.10, 0.25] and > 0.25.

The repeatability measure includes the random fluctuations due to the noise generated by the electronic components of the scanner. The different noise level among repeated acquisitions may impact on the image texture, and this aspect may be captured by the repeatability measurement. For this reason, we also made measures in homogeneous regions of the phantom to estimate the noise level among repeated measurements. Since we observed artefacts inside the saline bags (probably due to the movement of the liquid inside the bag), we performed the analysis inside the cork. Four regions in the axial slice were segmented for each of the ten repeated images at 100 and 120 kVp, on the two scanners. We found that the noise level, measured as the ratio between the standard deviation and the mean inside each ROI, was the same among the repeated acquisitions (with values between 0.5% and 0.7% among all the considered acquisitions). This result suggests that the random noise among repeated acquisitions and in different parts of the scan field is spatially uniform, considering a homogeneous material. An improvement of the HeLLePhant may

be the inclusion of an object made of a homogeneous material to measure the noise in the different positions in the scan field, similarly to the phantoms used for quality assurance. In this way it is possible to evaluate the noise among repeated acquisitions and among protocols on different scanners.

3.3.2 Reproducibility analysis

Materials and methods

After assessing the repeatability, we used the same CT acquisitions to investigate both the intra- and the inter- scanner reproducibility of the features. The list of the comparisons analysed in this study is shown in **Table 3.7**. For each configuration, the table includes the parameter modified — either scanner, voltage or reconstruction algorithm — and thus the fixed parameters. We performed the comparison separately for the bin width at 25 HU (configurations from 1 to 8) and at 5 HU (configurations from 9 to 16).

Config	Comparison	Fixed parameters
1	Scanner 1 vs Scanner 2	120 kVp, 25 HU, IR
2	Scanner 1 vs Scanner 2	100 kVp, 25 HU, IR
3	120 kVp vs 100 kVp	Scanner 1, 25 HU, IR
4	120 kVp vs 100 kVp	Scanner 2, 25 HU, IR
5	IR vs FBP	Scanner1, 120 kVp, 25 HU
6	IR vs FBP	Scanner1, 100 kVp, 25 HU
7	IR vs FBP	Scanner2, 120 kVp, 25 HU
8	IR vs FBP	Scanner2, 100 kVp, 25 HU
9	Scanner 1 vs Scanner 2	120 kVp, 5 HU, IR
10	Scanner 1 vs Scanner 2	100 kVp, 5 HU, IR
11	120 kVp vs 100 kVp	Scanner 1, 5 HU, IR
12	120 kVp vs 100 kVp	Scanner 2, 5 HU, IR
13	IR vs FBP	Scanner1, 120 kVp, 5 HU
14	IR vs FBP	Scanner1, 100 kVp, 5 HU
15	IR vs FBP	Scanner2, 120 kVp, 5 HU
16	IR vs FBP	Scanner2, 100 kVp, 5 HU

Table 3.7: List of the comparisons performed to investigate feature reproducibility.

It is worth recalling that the IR algorithms used for this study were the ones adopted in the clinical routine for the radiological examination of the chest at IEO: IR60 for the acquisitions on the Scanner 1, and IR50 for those on the Scanner 2.

We evaluated the reproducibility with respect to the voltage peaks (configurations 3, 4, 11 and 12) and the reconstruction algorithms (configurations 5, 6, 7, 8, 13, 14, 15 and 16) using the concordance correlation coefficient

(CCC, defined in eq. 2.1). Features with $CCC \geq 0.9$ were selected as robust [190, 205, 206]. We used the function *epi.ccc* from the *epiR* library in R (setting *rep.measure = TRUE* and accordingly with the parameter *subjectid*) to consider the repeated measurements for each subject/insert in the evaluation of the CCC.

Moreover, we performed a second analysis to understand better the feature behaviour for each insert separately. For each feature we evaluated the percentage variation between the two configurations of voltage peaks as

$$PV_{i,voltage} = \frac{f_{i,v1} - f_{i,v0}}{f_{i,v0}}\,, \tag{3.6}$$

where i identifies the feature, $f_{i,v0}$ is the mean value of the feature for the reference voltage (120 kVp) among the 10 repeated measurements, and $f_{i,v1}$ is the corresponding feature value for the compared configuration (100 kVp). We carried out the same analysis for the comparison between the two reconstruction algorithms (IR versus FBP), taking the IR as the reference reconstruction. For the inter-scanner variability (configurations 1, 2, 9 and 10) we considered both the CCC and the PV metrics. For the calculation of the PV, the Scanner 1 was taken as the reference scanner.

Results

1) Tube voltage peak: 120 kVp versus 100 kVp

Table 3.8 summarises the results of the percentage variation analysis performed to compare the two tube voltages for each insert. The percentage of features with a larger variability between the two voltage peaks was lower for the PET-G inserts, indicating a higher voltage reproducibility for this type of objects.

Concerning the CCC results, we found similar results among the four configurations, with more than 90% of the features with a CCC ≥ 0.90 when all the inserts were considered. A slightly lower percentage of features with an excellent agreement between the two voltage peaks was observed when only the polyacrylate inserts were used for the analysis. **Table 3.9** lists the percentage of features with CCC in the ranges [0.50, 0.75), [0.75, 0.90) and [0.90, 1.00] for the voltage agreement, both when all inserts and when only the polyacrylate inserts were included in the analysis. The only features with a CCC < 0.75 belonged to the firstorder category. A relevant shift of the mean of the grey-level histogram was in fact observed between the two voltage peaks.

	PV $\leq$ 0.1			
Insert	**3**	**4**	**11**	**12**
PET1_entire	97%	100%	98%	100%
PET1_core	94%	97%	96%	97%
PET2_entire	97%	99%	97%	99%
PET2_core	91%	94%	91%	96%
pol_non-unif_1	76%	94%	92%	95%
pol_non-unif_2	77%	71%	80%	85%
pol_non-unif_3	63%	84%	64%	87%
pol_non-unif_4	62%	68%	70%	73%
pol_non-unif_5	81%	79%	64%	75%
pol_non-unif_6	92%	92%	94%	95%
pol_unif_1	89%	73%	89%	75%
pol_unif_2	96%	89%	96%	89%
pol_unif_3	84%	74%	91%	71%

Table 3.8: Percentage of features with a percentage variation (PV) smaller than 0.10 in the comparison between the voltage peaks (configurations 3, 4, 11 and 12).

		Features (%) with CCC in range:			
	Config	**[0.90, 1.00]**	**[0.75, 0.90)**	**[0.5, 0.75)**	**[0, 0.5)**
ALL	**3**	95.0%	3.6%	0.7%	0.7%
	4	95.0%	3.6%	1.4%	0%
	11	92.2%	6.4%	0.7%	0.7%
	12	92.2%	6.4%	1.4%	0%
ONLY POLY	**3**	70.0%	24.3%	1.4 %	4.3%
	4	81.4%	12.8%	2.9%	2.9%
	11	80.7%	14.3%	0.7%	4.3%
	12	86.4%	8.6%	2.1%	2.9%

Table 3.9: Percentage of features with CCC between 0 and 0.5, 0.5 and 0.75, 0.75 and 0.90, and greater than 0.90, considering the impact of the voltage peak variation (configurations 3, 4, 11 and 12). The term "ONLY POLY" means that only the polyacrylate inserts were included in the analysis, while "ALL" that both polyacrylate and PET-G inserts were included.

2) Reconstruction algorithm: IR versus FBP

Table 3.10 and **Table 3.11** summarise the results of the comparison between the two different reconstruction algorithms (IR and FBP). A general reduction in the number of the stable features can be observed, compared to the voltage comparison.

Two aspects were similar to what we found in the previous analysis (voltage peak). First of all, we observed a better reproducibility for the PET-G inserts than all the polyacrylate ones from the PV analysis. Secondly, the CCC

	PV $\leq$ 0.1							
Insert	**5**	**6**	**7**	**8**	**13**	**14**	**15**	**16**
PET1_entire	84%	84%	91%	91%	86%	90%	94%	92%
PET1_core	83%	85%	89%	89%	89%	89%	89%	86%
PET2_entire	78%	77%	86%	90%	85%	84%	91%	92%
PET2_core	69%	66%	84%	88%	69%	71%	84%	84%
pol_non-unif_1	49%	43%	66%	61%	47%	41%	68%	63%
pol_non-unif_2	36%	37%	50%	42%	40%	42%	56%	46%
pol_non-unif_3	31%	30%	44%	38%	33%	31%	44%	45%
pol_non-unif_4	44%	42%	56%	45%	46%	45%	60%	47%
pol_non-unif_5	36%	32%	49%	43%	34%	36%	52%	42%
pol_non-unif_6	36%	35%	39%	40%	36%	39%	40%	44%
pol_unif_1	46%	41%	67%	55%	49%	44%	75%	63%
pol_unif_2	62%	56%	68%	67%	66%	60%	71%	71%
pol_unif_3	44%	46%	68%	60%	46%	44%	71%	67%

Table 3.10: Percentage of features with a percentage variation (PV) smaller than 0.10 in the comparison between the reconstruction algorithms (configurations 5, 6, 7, 8, 13, 14, 15 and 16).

analysis gave worse results when only the polyacrylate inserts were included. However, features with a CCC < 0.75 were mainly from the texture categories.

Finally, we found better reproducibility results in the Scanner 2 compared to Scanner 1. However, this dissimilarity may be due to the different blending levels of the IR for the two scanners, closer to the FBP — for which the blending level corresponds to 0% — for the Scanner 2. In order to investigate better this point, the CT images for the Scanner 2 were reconstructed also with a blending level equal to 60% (IR60) and were compared to the FBP. Considering as metric the CCC for the polyacrylate inserts only, we obtained a reduction in the number of features with perfect agreement (120 kVp-25 HU: 46%, 100 kVp-25 HU: 41%, 120 kVp-5 HU: 50%, 100 kVp-5 HU: 45%) with respect to the comparison FBP versus IR50 in the Scanner 2 (see **Table 3.11**, configurations 7, 8, 15 and 16). However, these percentages are still higher than in the comparison FBP versus IR60 in the Scanner 1 (see **Table 3.11**, configurations 6, 7, 13 and 14). This discrepancy may be associated either with the difference in the contours, which may impact the voxels near the borders, or with a difference in the reconstruction procedure between the two scanners. Unfortunately, due to the difference in the segmentation between the scanners for the same insert, it was not possible to identify the actual source of this variability.

3) CT scanner model: Scanner 1 versus Scanner 2

Finally, we compared the acquisitions performed on the two scanners, using the same acquisition protocol (configuration 1, 2, 9 and 10 in **Table 3.7**). In

	Config	Features (%) with CCC in range: [0.90, 1.00]	[0.75, 0.90)	[0.5, 0.75)	[0, 0.5)
ALL	**5**	71.5%	16.4%	6.4%	5.7%
	6	70.7%	13.6%	10.0%	5.7%
	7	78.7%	12.9%	5.7%	2.9%
	8	76.4%	12.9%	5.7%	5.0%
	13	75.0%	12.9%	12.1%	0.0%
	14	75.0%	10.7%	12.9%	1.4%
	15	70.0%	12.9%	12.1%	5.0%
	16	70.8%	11.4%	10.7%	7.1%
ONLY POLY	**5**	32.8%	29.3%	15.0%	22.9%
	6	32.1%	27.1%	17.1%	23.6%
	7	55.7%	14.3%	17.1%	12.9%
	8	47.9%	20.7%	12.1%	19.3%
	13	40.0%	23.5%	17.9%	18.6%
	14	36.4%	26.4%	17.9%	19.3%
	15	57.9%	16.4%	10.7%	15.0%
	16	50.0%	22.9%	10.0%	17.1%

Table 3.11: Percentage of features with CCC between 0 and 0.5, 0.5 and 0.75, 0.75 and 0.90, and greater than 0.90, considering the variation of the reconstruction algorithm (configurations 5, 6, 7, 8, 13, 14, 15 and 16). The term "ONLY POLY" means that only the polyacrylate inserts were included in the analysis, while "ALL" that both polyacrylate and PET-G inserts were considered.

this case, however, the final results included the effects of both the different scanner and the different contours. Since for each insert the segmentations did not match perfectly between the two scanners, we considered also the *shape* features. We found that all the *shape* features extracted from the inserts had a low percentage variation (<5% for all the inserts and all the features). Despite this similarity in the shape, the position of the contours may not be exactly the same inside the insert, potentially including different voxels on the borders. This was particularly evident near the edge of the inserts along the z-direction, where a mismatch — below the slice thickness — of the phantom on the two scanners was present, this way producing different artifact effects.

In **Table 3.12** and **Table 3.13** we list the results for the comparison between the two scanners (configurations 1, 2, 9 and 10), based on the PV and the CCC values, respectively. As in the previous comparisons, the percentages were calculated excluding the *shape* features.

In general we obtained better results than in the comparison between the reconstruction algorithms, but worse than for the impact of the voltage. All the categories of features were affected, both firstorder and texture ones.

It is interesting to note that the PET2 insert was less reproducible than the PET1 and than most of the polyacrylate inserts, in contrast with the previous

	PV $\leq$ 0.1			
Insert	**1**	**2**	**9**	**10**
PET1_entire	89%	93%	91%	88%
PET1_core	86%	88%	79%	81%
PET2_entire	67%	65%	74%	74%
PET2_core	65%	61%	69%	66%
pol_non-unif_1	96%	95%	100%	96%
pol_non-unif_2	94%	76%	99%	84%
pol_non-unif_3	69%	60%	76%	57%
pol_non-unif_4	72%	71%	73%	72%
pol_non-unif_5	59%	56%	60%	59%
pol_non-unif_6	76%	84%	87%	91%
pol_unif_1	70%	63%	77%	67%
pol_unif_2	86%	84%	89%	96%
pol_unif_3	50%	50%	53%	54%

Table 3.12: Percentage of features with a percentage variation (PV) smaller than 0.10 in the comparison between the two scanners (configurations 1, 2, 9 and 10).

		Features (%) with CCC in range:			
	Config	**[0.90, 1.00]**	**[0.75, 0.90)**	**[0.5, 0.75)**	**[0, 0.5)**
ALL	**1**	90.0%	7.1%	2.9%	0.0%
	2	90.0%	7.1%	2.9%	0.0%
	9	88.6%	5.7%	5.7%	0.0%
	10	89.3%	5.7%	5.0%	0.0%
ONLY POLY	**1**	63.6%	15.7%	15.7%	5.0%
	2	60.0%	18.6%	14.3%	7.1%
	9	72.8%	8.6%	13.6%	5.0%
	10	69.3%	10.0%	15.0%	5.7%

Table 3.13: Percentage of features with CCC between 0 and 0.5, 0.5 and 0.75, 0.75 and 0.90 and greater than 0.90, comparing the two scanners (configurations 1, 2, 9 and 10). The term "ONLY POLY" means that only the polyacrylate inserts were included in the analysis, while "ALL" that both polyacrylate and PET-G inserts were included.

robustness analysis (reconstruction algorithm and voltage peak).

Summary

Figure 3.22 and **Figure 3.23** summarise the reproducibility and the repeatability analyses described in detail above.

Figure 3.22 refers to the ICC/CCC analysis, considering the results when only the polyacrylate inserts are used. We observed, in fact, that these metrics are influenced by the population of subjects (in this case the inserts) included

in the analysis. This means that, if this population in phantom is not well representative of the patient population under investigation, the robustness results can be biased. For this reason, in this summary we excluded the ICC/CCC results with also PET-G inserts, which were in many cases out of the patient range (see **Figure 3.17**), increasing the population variability and thus the agreement between the measurements of the same subject.

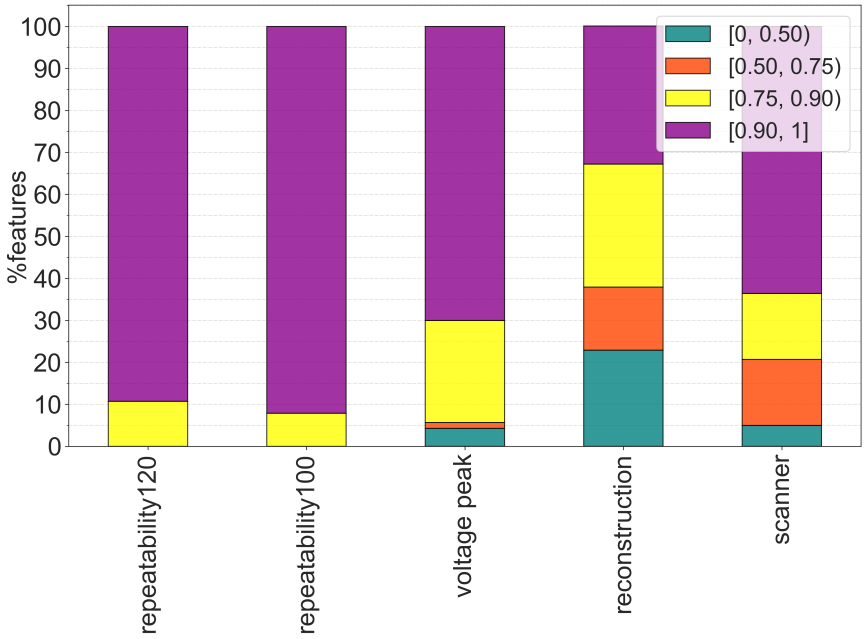
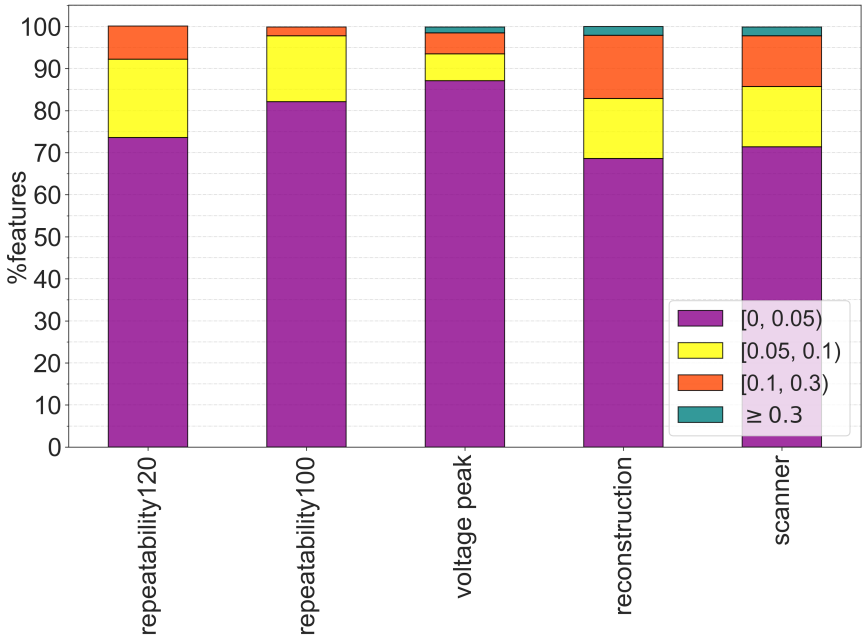
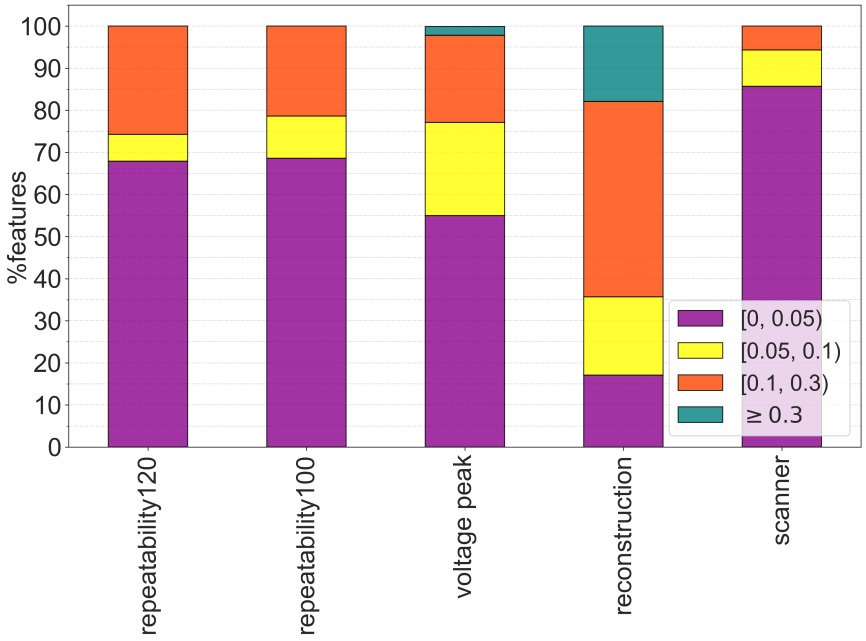


Figure 3.22: Summary of the repeatability and reproducibility **in terms of ICC and CCC**, for the 25 HU bin width. The repeatability results are reported both for the acquisition at 100 kVp and at 120 kVp, performed on Scanner 1 with the IR reconstruction. The reproducibility bar-plot refers to the comparison between the 120 kVp and the 100 kVp acquisitions on Scanner 1 with the IR reconstruction (*voltage peak*), between the IR and the FBP reconstructions on Scanner 1 at 120 kVp (*reconstruction*), and between Scanner 1 and Scanner 2 at 120 kVp with the IR reconstruction (*scanner*).

In **Figure 3.23**, instead, the CV/PV results are shown for two inserts taken as examples: the PET1_core and the pol_non-unif_2. We observed that the PET-G inserts were in general more repeatable and reproducible than the polyacrylate ones, except for the scanner comparison. In the latter, in addition to the direct impact of the scanner model, the effect of the different contours for the same insert must also be taken into account, since the images come from the two scanners not perfectly overlapping. This misalignment may have a greater impact on the PET-G inserts because of their coarser texture. With respect to the polyacrylate inserts, in fact, the PET-G inserts are more characterised by the presence of groups of voxels with similar grey-level intensities next to clusters of voxels with significantly different intensities. This structure makes more likely the inclusion of voxels on the VOI borders belonging to a completely different intensity range when the contours are shifted inside the insert.



(a) PET1_core



(b) pol_non-unif_2

Figure 3.23: Summary of the repeatability and reproducibility in **terms of CV and PV** for the PET1_core and the pol_non-unif_2 inserts, for the 25 HU bin width. The repeatability results are reported both for the acquisition at 100 kVp and at 120 kVp, performed on Scanner 1 with the IR reconstruction. The reproducibility bar-plot refers to the comparison between 120 kVp and 100 kVp acquisitions on Scanner 1 with IR reconstruction (*voltage peak*), between IR and FBP reconstructions on Scanner 1 at 120 kVp (*reconstruction*) and between Scanner 1 and Scanner 2 at 120 kVp with IR reconstruction (*scanner*).

We summarise the behaviour of the features in terms of the CV/PV values in **Figure 3.24**, **Figure 3.25** and **Figure 3.26** for each feature and each insert using the heatmap plot. **Figure 3.24** illustrates the results for the repeatability analysis, where dark red indicates higher repeatability while dark blue lower repeatability. From this plot it can be noted that the texture features from gldm and glrlm, and above all the glzsm and ngtdm categories, are the least reproducible. Moreover, the similarity between the two scanners as well as the two voltage peaks can be appreciated. **Figure 3.25** and **Figure 3.26**, instead, show the reproducibility results, as regards the reconstruction algorithm variability in the first figure, and both the voltage peak and the scanner model in the second figure. The same scale is used in order to highlight the greater impact of the reconstruction parameter. A colour difference can be observed between the polyacrylate inserts (columns A, B, C, D, E, F, G, H, I) and the PET ones (columns J, K, L, M), particularly in the reproducibility analysis.

Finally, the comparison of the feature reproducibility due to the reconstruction algorithm between patient (Chapter 2) and phantom results (Chapter 3) is presented in Appendix D.

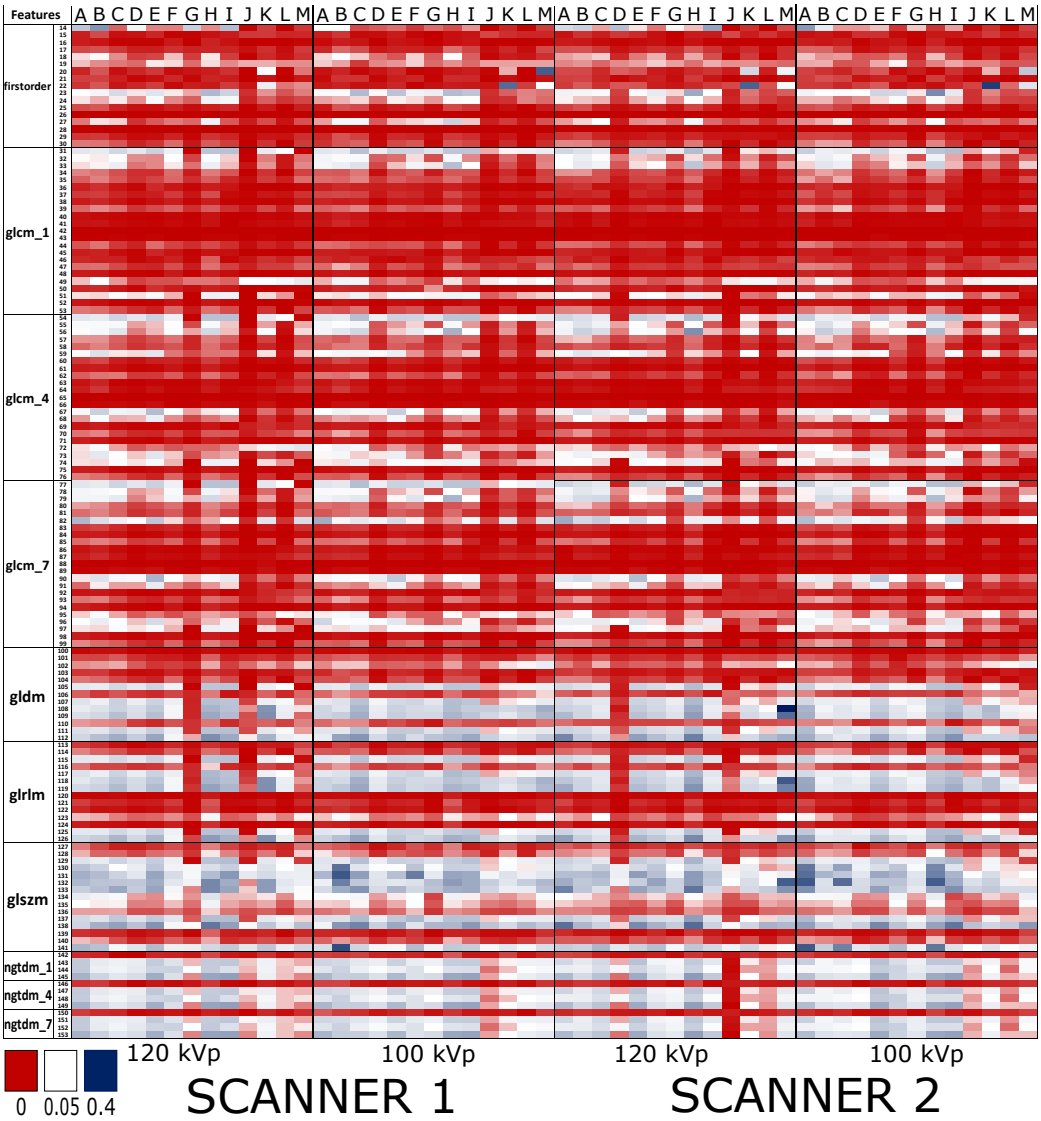


Figure 3.24: Heatmap of the CV values, indicating the **repeatability** for the two scanners and the two voltage peaks. The first column indicates the feature category, and the second the feature number (from 14 to 143). The results are displayed for all the inserts: A=pol_unif_1, B=pol_unif_2, C=pol_unif_3, D=pol_non-unif_1, E=pol_non-unif_2, F=pol_non-unif_3, G=pol_non-unif_4, H=pol_non-unif_5, I=pol_non-unif_6, J=PET1_entire, K=PET1_core, L=PET2_entire, M=PET2_core.

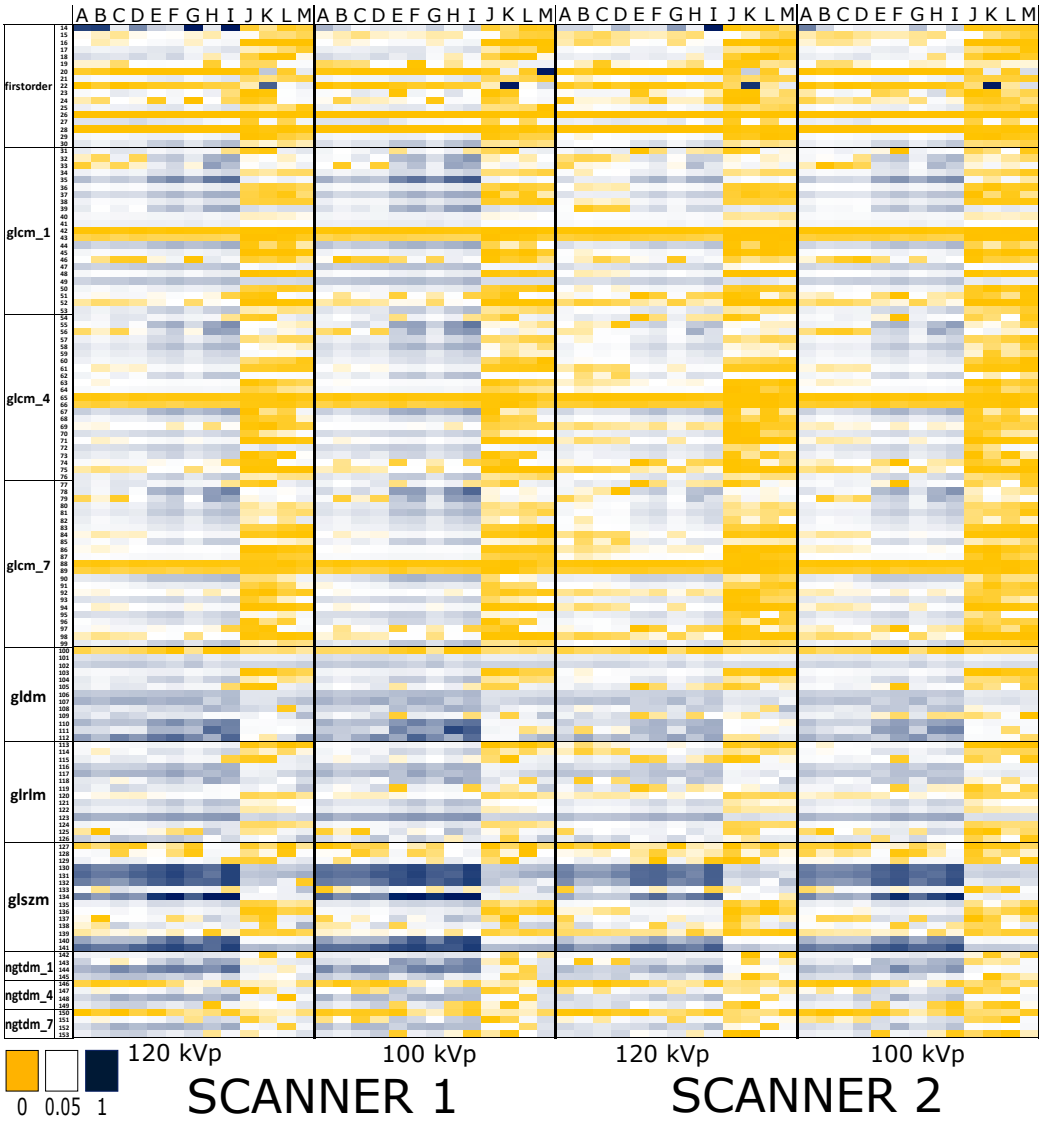


Figure 3.25: Heatmap of the PV values for the **reconstruction algorithm impact (IR versus FBP)** including the two scanners and the two voltage peaks. The first column indicates the feature category, and the second the feature number (from 14 to 143). The results are displayed for all the inserts: A=pol_unif_1, B=pol_unif_2, C=pol_unif_3, D=pol_non-unif_1, E=pol_non-unif_2, F=pol_non-unif_3, G=pol_non-unif_4, H=pol_non-unif_5, I=pol_non-unif_6, J=PET1_entire, K=PET1_core, L=PET2_entire, M=PET2_core.

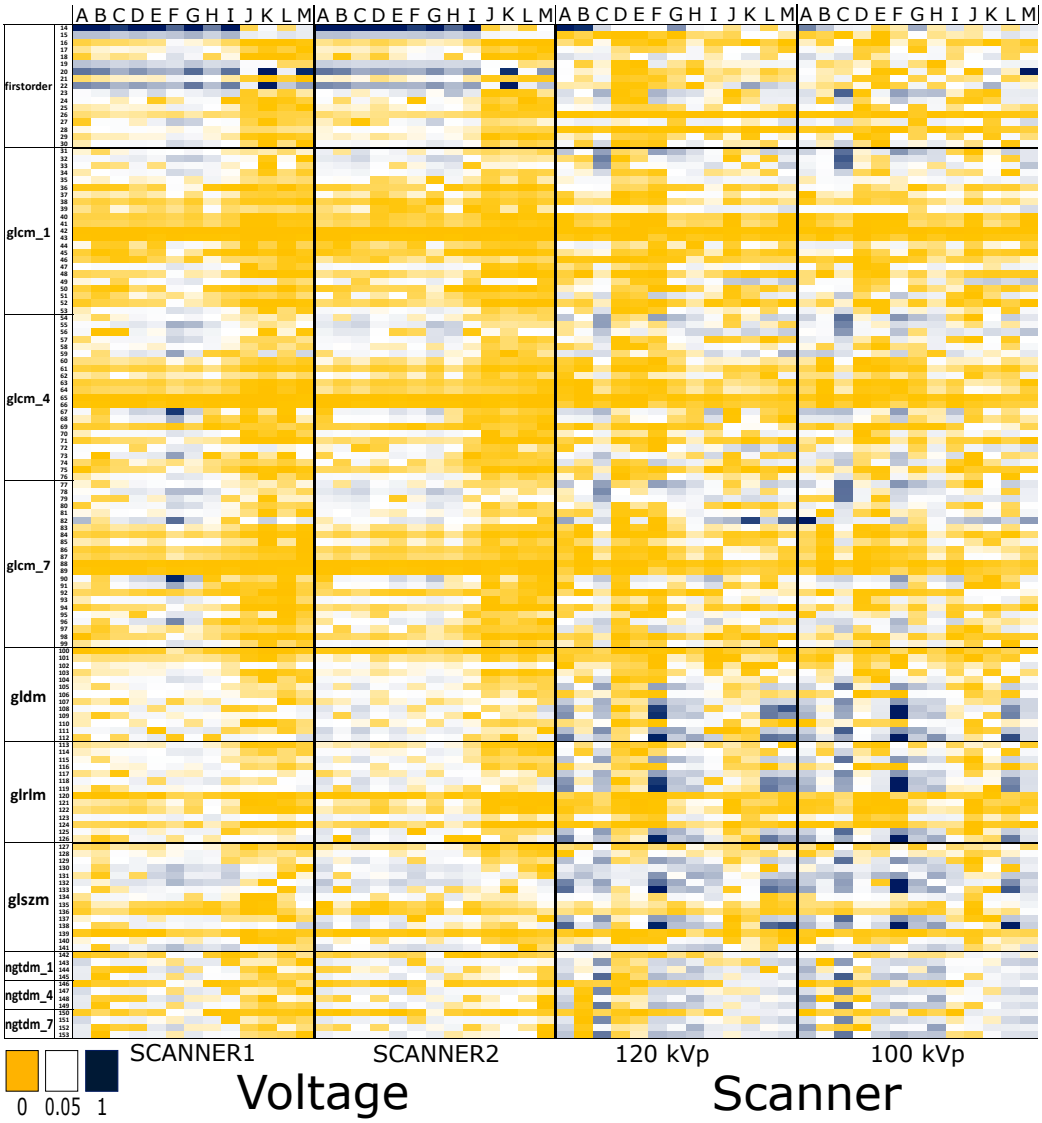


Figure 3.26: Heatmap of the PV values for the **voltage (on the left) and the scanner (on the right) impact**. The first column indicates the feature category, and the second the feature number (from 14 to 143). The results are displayed for all the inserts: A=pol_unif_1, B=pol_unif_2, C=pol_unif_3, D=pol_non-unif_1, E=pol_non-unif_2, F=pol_non-unif_3, G=pol_non-unif_4, H=pol_non-unif_5, I=pol_non-unif_6, J=PET1_entire, K=PET1_core, L=PET2_entire, M=PET2_core. The same colour scale of **Figure 3.25** was used.

3.3.3 Discussion

In this chapter a surrogate of a thorax phantom with inserts simulating NSCLC tumours, made of PET-G or sodium polyacrylate, was presented. In the first part of this chapter, we compared the inserts against real lesions from a radiomic point of view, while in this second part we showed the feasibility of using such a phantom to investigate feature robustness in controlled settings. Feature stability is an important requirement when using these descriptors as imaging biomarkers in clinical models. However, isolating the contribution of each separate confounding factor is not always possible in a patient cohort. Moreover, physical and biological variability in the tissue — for example due to the re-positioning on the scanner table, to the breathing, or to the disease evolution over time — makes it difficult to repeat the acquisitions in fixed conditions. Therefore, various radiomic studies have been carried out using phantoms as substitutes for human tissues.

In this study we considered the impact of ten repeated acquisitions in fixed condition, i.e. without changing any parameters and without moving the phantom among different scans. In addition, we also evaluated the impact of the variation of the voltage peak and the reconstruction algorithm, as well as the scanner-induced variation.

We used two types of metrics to study the repeatability and reproducibility of the features. The ICC/CCC parameters are often used in radiomic studies [109, 114, 181, 193, 206, 207]. However, as we showed in this thesis, their results are strongly affected by the population under analysis. If the phantom materials used in the study are not well-representative of the variability observed in patients, the robustness results should not be generalised to the clinical population of interest. The second type of analysis adopted to estimate the degree of variability for each insert was instead based on two more intuitive coefficients: the coefficient of variation (CV) for the repeatability analysis, and the percentage variation (PV) for the reproducibility one. In general, however, a lack of consensus in the choice of the correct metrics and of a threshold to assess robustness is still present in the literature, obstructing the comparison of the results among different studies.

From our analysis we can highlight four main aspects.

First of all, we found that the majority of the features is repeatable (CV $\leq$ 0.1), independently from the voltage peak and the scanner, with better results for the PET-G inserts (see **Table 3.23**, **Table 3.5** and **Table 3.4**). Features from the firstorder and glcm categories are the most repeatable, as can be observed from **Figure 3.21**. This may be due to the fact that the features from categories of order higher than two (which are features evaluating the spatial relationship among more than two voxels) have a better ability to capture locally uniform clusters of grey levels, and may be more sensitive to the image noise which introduces a random component inside the texture among the repeated acquisitions.

Secondly, we obtained that the reconstruction algorithm is the parameter

which, among those investigated, has the greatest impact on the features. This confirms the results obtained with a cohort of NSCLC patients, presented in Chapter 2. In general, we found a low percentage of features with PV ≤ 0.1 in the polyacrylate inserts (see **Table 3.10**), where this percentage ranges between 30% for the pol_non-unif_3 insert (acquisition on Scanner 1 at 100 kVp, 25 HU) and 75% for the pol_unif_1 insert (acquisition on Scanner 2 at 120 kVp, 5 HU). Instead, the percentage of features with PV ≤ 0.1 is always over 69% for the PET-G inserts. We therefore recommend to pay attention and take the appropriate precautions when the radiomic dataset consists of CT images reconstructed with a different blending level, especially when both the IR and FBP algorithms are present.

The third interesting result concerns the comparison between the two scanners. We found that the results of repeatability and of texture variability due to voltage and reconstruction algorithm are very similar between the two CT scanners. These findings appear to be in contrast with what Varghese et al. found in their study on the feature robustness (repeatability and the impact of exposure, voltage, FOV, reconstruction kernel, and slice thickness) with a dedicated phantom [112]. However, the two CT scanners used in the latter study belonged to two different vendors (Philips and Toshiba) with respect to our study. This confirms the similarity between our scanners, both from a hardware/software point of view and as concerns the calibration and optimisation procedure of the acquisition/reconstruction protocols. Moreover, this corroborates the results of Mackin et al. [106], who showed that features extracted from CT images of the CCR phantom grouped together when acquired on scanners of the same vendor. However, these findings underline the importance of performing further analyses using other CT scanners, even better if manufactured by vendors different from that used in this study (GE Healthcare).

Finally, we observed that the feature behaviour — and hence its robustness — is texture dependent. This variation with the material was observed also by Li et al. [206], who scanned a phantom made of various materials — such as cereal, rice, cork, wood, homogeneous materials and a mini pig — in different settings, test-retest and different acquisition and reconstruction parameters (i.e. voltage, exposure, FOV, slice thickness and kernel). They observed that homogeneous materials, such as solid water and polystyrene foam, were more sensitive to parameter variability. Similar conclusions were reached by Berenguer et al. [181] in the evaluation of inter-scanner comparison, and by Mackin et al. [107] in analysing the exposure effect, both using the CCR phantom.

Our study highlighted that the radiomic features are affected in a different way by some of the acquisition parameters, and that this influence varies with the material under investigation. It is thus important that the methodological studies, whether they are patient or phantom-based, are representative of the anatomical region of interest for the purpose of extending the results to the clinical studies. Furthermore, the phantom we fabricated offered an excellent

opportunity to study the radiomic features behaviour in controlled configurations of acquisition and reconstruction parameters. To do so, we selected a threshold for the chosen metrics to distinguish robust from non-robust features and compare the results among the inserts and the acquisition settings. However, when performing a radiomic analysis on a clinical population, keeping or rejecting these features depends on the final clinical outcome. The impact of feature variation due to acquisition/reconstruction settings, in fact, can be significant or not according to how much the patient populations, which have to be distinguished by the radiomic model, are separate. For this reason, further investigations on a possible impact of a priori selection of the features in the clinical studies is mandatory for different clinical endpoints.

In the next section the advantages and the limitations we found in the proposed phantom will be shown and areas of possible improvement will be pointed out.

Advantages and disadvantages of the two types of insert

The main limitations we encountered with the PET prototype was that its shape was too geometrical and its texture too rough, making it far from what we observe in real lung lesions. Moreover, the mean CT signal achieved with these prototypes is low compared to the contrast enhanced lesions. In order to refine the PET-G inserts, new inserts have to be produced with an increased density and reduced heterogeneity. This could be achieved for example by choosing for the core an infill range smaller than the values used in this study and including the highest percentages (between 90% and 100%). As regards the geometry of the inserts, in order to create more tumour-like inserts, the overall size should be increased, and the shape should be improved by increasing its irregularity and complexity.

On the other hand, the PET-based "voxelated" structure we propose in this work allows us to produce a controlled heterogeneity. And this is not the only advantage of the PET inserts. The fact that they are produced with a 3D printing technique makes them more reproducible as far as shape and texture are concerned. As a result, we can manufacture multiple similar inserts, which is very useful in view of multicentre studies. Further analyses have to be performed to compare multiple PET-G inserts manufactured with the same procedure and printing parameters in order to evaluate the reproducibility in the fabrication.

The inserts made of sodium polyacrylate with diluted contrast medium also have several positive aspects. Thanks to its gel-like consistency, in fact, the polyacrylate can be mouldable, and can thus take shapes which are more similar to those of tumours. However, if not tightly sealed, this consistency makes its structure fragile, prone to cracks and to shape deformation. Moreover, we observed a degradation of the texture over time, with the formation of denser

areas inside. In order to get an idea of the level of the degradation, we acquired again the phantom using the same protocol and scanner (Discovery CT750 HD at 120 kVp) after four months. In this period of time the phantom was kept in a refrigerator, without moving any internal components (saline bags included). We observed that almost all the polyacrylate inserts were characterised by an increase of the mean CT signal inside the insert, with a minimum variation of 2 HU for the Pol_unif_2 and a maximum variation of 18 HU for the Pol_non-unif_4 (the median increase among these inserts was of 10 HU). Only in the Pol_non-unif_5 we observed a decrease of 7 HU. In contrast, for the PET-G inserts we found an increase of only 1 HU. In order to consider the possible effect of the different contours among the two acquisitions, we repeated the segmentation eight times for each insert and calculated the mean CT signal and the standard deviation of the mean among the repeated contours. We obtained that the standard deviation among the eight mean CT numbers was always less or equal to 3 HU for all the inserts (equal to 3 HU for the PET-G inserts and for the Pol_non-unif_4). We envisage that a possible solution to the limited durability of the polyacrylate shape and texture would be to seal the compound inside a rigid shell. Even if we did not observe a variation in the PET-G inserts, both visually and from the mean CT signal calculation, Ger et al. [111] found that a similar material — the ABS — changes over time. Therefore, further analysis on the PET-G material is required, concerning for example its structure stability outside the refrigerator. This promising line of research certainly deserves further investigation.

Despite these drawbacks, the polyacrylate yields the best results in our agreement analysis thanks to the texture heterogeneity and the mean CT signal we achieved. Moreover, the possibility to create various inserts with a mixture of different contrast concentrations allows us to reproduce the variability found in patient datasets.

Future developments

An interesting aspect which should be investigated in order to generalise our results is the repetition of the acquisitions in other institutes, using also CT scanners from different vendors, which are not available at IEO. This would allow us to compare from a radiomic point of view the chest protocols of multiple radiological departments. Moreover, a comparison of the CT scanner quality can be done with the phantom, after fixing and properly matching the acquisition parameters. This may be useful, for instance, to harmonise the reconstruction algorithm kernels developed by different vendors, similarly to what has been done previously by Mackin et al. with the CCR phantom [186].

Another issue we did not address was the impact of the re-positioning of the phantoms on the table of the scanner. In future investigations this analysis should be performed, in order to isolate the impact of the re-positioning factor and estimate more accurately the effect of the scanner on the feature reproducibility.

Furthermore, the analyses presented in this thesis were performed only for the features extracted from original images, without filtering. The same characterisation should be repeated for the features extracted from the filtered images (i.e. LoG and wavelet filters) in order to enable a comparison with the results discussed in Chapter 2.

Finally, improvements in the shape and in the materials mimicking the thorax are necessary in future versions of the HeLLePhant. The configuration used, in fact, was too simple and non-anthropomorphic as far as the internal structures of the thorax are concerned, the lungs in particular. It would be useful, for example, to design a thorax phantom with dedicated spaces to lodge the inserts. 3D printing technology may be a valuable tool for this purpose, as proposed in the literature [208–210]. This kind of structure may in fact increase the reproducibility in the positioning and re-positioning of the inserts inside the "lung".

Chapter 4

Automatic segmentation of lung lesions

Segmentation of medical images is a crucial step for radiomic study based on hand-crafted features. Typically, the physician contours the lesion manually, but this process takes a lot of time and effort. This chapter deals with the comparison of manual contours with two alternative techniques for lung lesion segmentation. The first one is a semi-automatic approach based on a region growing algorithm (*GrowCut*). The second technique, instead, is a fully automatic algorithm based on a deep neural network (*nnU-Net*). The performance of the two techniques will be evaluated by comparing their outputs with the manual segmentation, considering both the spatial overlapping (DICE and HD) and the radiomic feature variability (ICC). Finally, the last part of the chapter will be dedicated to understand whether a difference in the segmentation may significantly impact on the performance of a radiomic model, even in case of a variability detected by the aforementioned metrics. To this aim, an overall survival (OS) model will be built and evaluated separately using manual and automatic contours.

4.1 Introduction to the segmentation task

The presence of different acquisition and reconstruction protocols in the same clinical database of images, discussed so far, is not the only issue that should be taken into account during a radiomic analysis. A further factor that may impact on feature reproducibility is the segmentation of the target volume, from which the radiomic features are extracted.

Segmentation is the delineation of a VOI, by identifying all the voxels that share a common property (such as belonging to the same anatomical structure) and thus separating them from the background. In case of oncologic applications, where the region of interest is most often represented by the tumour, segmentation is usually performed by the physicians, who manually drawn slice by slice the borders of the lesion using ad-hoc software. Unfortunately, manual segmentation suffers from some limitations. First of all, it is a laborious and time-consuming process, particularly when samples with large size are used, as for instance occurs in radiomic studies where a great amount of input images is necessary to increase the statistical power of the models. Secondly, manual

segmentation is prone to intra- and inter-reader variability. This means that not only different radiologists (inter-), each with their own training and experience, but even the same operator (intra-) would not be able to make exactly the same segmentation twice.

To assess the influence of delineation on radiomic features and reduce uncertainty, different segmentation methods were investigated in the recent literature. Both semi-automatic techniques, in which the greatest part of the segmentation is done by the algorithm but with a human initialisation, and fully automatic ones were implemented [211–214]. However, algorithms able to replace the physician's expertise is still a challenging task. Differences in the shape, in the texture and in the location of the lesions or the juxtaposition of different tissues with similar grey-level intensities make difficult, in fact, to generalise the segmentation results. The lack of the *a priori* knowledge and of the physician background is certainly one of the main deficiency in segmentation without human interventions. Therefore, even though in medical imaging a *ground truth* for segmentation of tumours does not exist, the only available reference contour remains the manual one.

A list of evaluation metrics commonly used in the literature to compare contours and adopted for the following analysis is reported in Section 4.1.1: the volumetric Dice similarity coefficient (DICE) [215] and the intersection over union (IoU or Jaccard index) [216] are used to compare the volumes of two VOIs, while the Hausdorff distance (HD) [217] to compare the distance between their surfaces.

4.1.1 Evaluation metrics

DICE

The DICE coefficient is defined as twice the ratio of the intersection between the two contours (A and B) to the sum of the total voxels inside each single contour:

$$\text{DICE(A, B)} = 2 \cdot \frac{n\{A \cap B\}}{n\{A\} + n\{B\}} = 2 \cdot \frac{\text{TP}}{2\ \text{TP} + \text{FP} + \text{FN}}, \tag{4.1}$$

where A and B are the set of voxels in the two contours and $n\{I\}$ is the cardinality — meaning the number of elements — of the set I. TP (true positive), FP (false positive) and FN (false negative) correspond to the correctly segmented voxels, the wrongly included voxels and the wrongly non-included voxels compared to the ground truth VOI, respectively. The DICE values are between 0 (total incompatibility) and 1 (full overlap).

Intersection over union

The IoU, also known as Jaccard index, is quite similar to the DICE, but it is defined as the ratio of the overlapping volume between the two contours (A and B) and their union:

$$\text{IoU(A, B)} = \frac{n\{A \cap B\}}{n\{A \cup B\}} = \frac{\text{TP}}{\text{TP + FP + FN}}\,. \tag{4.2}$$

As for the DICE, the IoU values range between 0 (no overlap) and 1 (full overlap).

It should be noted that in this study both the DICE and the IoU were considered in order to compare more easily our results with the literature, where an agreement in the choice of the metrics has not been reached yet. However, from their formulation it can be observed that these two parameters are correlated:

$$\text{DICE} = \frac{2\ \text{IoU}}{\text{IoU} + 1}\,.$$

Hausdorff distance

The HD measures the largest distance among the closest points from the boundaries of the segmentation A and the segmentation B. First of all, for each point of the surface of the contour A the smallest distance from B is considered. Secondly, among these minimum distances the largest one is taken ($h(A, B)$). The same calculation is repeated by changing the set A with the set B, and the set B with the set A ($h(B, A)$). Finally, the maximum between the two calculations gives the HD. Therefore, the HD is defined as:

$$\text{HD(A, B)} = \max\left(\text{h(A, B)}, \text{h(B, A)}\right), \tag{4.3}$$

where

$$\text{h(A, B)} = \max_{a \in A} \min_{b \in B} ||\text{a-b}||\,,$$

and $||a - b||$ is the distance between the points a and b in the sets A and B, respectively. The smaller is its value, the closer are the two segmentations.

Intra-class correlation coefficient

The three metrics described above are used to compare contours from a geometrical point of view, by quantifying how much the sizes (i.e. volume) and the positions of the two contours match. However, in radiomic studies it is interesting to evaluate the impact of the different contours on the feature stability. To this aim, the ICC is often considered [218]. We used the ICC definition reported in Section 3.5 (two-way mixed effect model, single rater type for absolute agreement estimation). Using this coefficient, two or more contour types can be compared at the level of each single feature among all the patients in a given dataset.

4.1.2 Semi-automatic segmentation: the GrowCut algorithm

In the last decades different segmentation techniques based on semi-automatic approaches were proposed. The key idea behind this type of algorithms is that they need an interaction of the user to guide the segmentation. Typical semi-automatic algorithms are the active contour [219, 220] and the region-based segmentation [221–223].

In this study a semi-automatic algorithm, named *GrowCut* and based on the region growing approach, was investigated using the 3DSlicer software (*Grow from seeds* tool).

Basically, the algorithm that rules this type of segmentation starts from some initial voxels, named *seeds*, which are labelled manually by the user. For each structure which has to be segmented, the operator defines a class of seeds labelled in a different manner (background included). Then, the segmentation iteratively evolves from the seeds, based on a neighbourhood rule. More information about semi-automatic segmentation can be found in Appendix E.

The main drawback of this approach is that the final result depends on where and how many seeds are drawn at the beginning of the segmentation procedure. One of the main advantages, instead, is that this tool is interactive, since the user can add new seeds in real time to improve the result of the segmentation.

4.1.3 Fully automatic segmentation: the nnU-Net

In contrast to semi-automatic approaches, the fully automatic ones do not require any user intervention to generate the segmentation. Thresholding [224], clustering [225], atlas-based [226] and neural network are few examples of automatic techniques. In this study neural network-based approach was addressed, working with a deep architecture named *U-Net*.

Nowadays, fully automatic approaches based on deep learning algorithms are the state-of-the-art in automatic segmentation of medical images [227–230]. Convolutional neural networks (CNNs) are suitable when images are used as input data and fully convolutional neural networks (FCNNs) [231], based on CNNs, are one of the architectures commonly used for segmentation.

The U-Net architecture [232, 233] is a particular FCNN and it is one of the most investigated network for image segmentation [228, 230, 234], both for 2D and 3D contours. A 2D architecture takes in input a single 2D image at once, and the learning procedure is performed considering only the information from that single slice. In case of a 3D domain, instead, each filter of the convolution and deconvolution operations is a 3D matrix, therefore the feature maps are built considering also the information form multiple adjacent slices. The 3D network is particularly useful in medical segmentation, where the majority of the images comes from tomographic acquisition. However, the 3D configuration has to learn more parameters during the training compared to the 2D one, therefore the segmentation is usually more complex and requires more

computational resources [235–238].

The U-Net, which we chose for this study, is the **nnU-Net** ("no-new-net"), recently proposed by Isensee et al. [239]. It is a self-configuring algorithm, which is able to train the model without the need of complex manual intervention, such as hyperparameter optimisation or data preparation, enabling the segmentation for multiple and diversified imaging situations. This architecture was developed using the *PyTorch* library in Python and can be downloaded for free on GitHub (https://github.com/MIC-DKFZ/nnUNet).

The pipeline available online provides the training of a 2D U-Net, a 3D U-Net and a 3D U-Net cascade (for this last, first a 3D U-Net is trained on low resolution images, then the output maps of the first part are refined with another 3D U-Net at full resolution). One or more of these configurations can be selected before the training. If multiple architectures are trained, at the end it is possible to select the best type or a combination of them via cross-validation. Moreover, the probability masks of two trained configurations can be averaged to obtain a combination of them, resulting in the "ensemble" configuration. In our study we built the ensemble output from the 2D and the 3D configurations in order to take advantages from both the architectures, capturing at the same time the in-plane (intra-slice) and the volumetric (inter-slices) information.

Further details on the convolutional neural network, in particular on the U-Net and the nnU-Net, are available in Appendix E.

4.1.4 Related works in lung lesion segmentation

There are some ambiguous situations in lung cancer delineation, which make difficult the identification of the borders of the tumour. Tumours attached to organ with similar grey-level intensities and texture (such as pleura attachment), the presence of atelectasis, vessels or halo of ground-glass around the tumour [80, 131, 240, 241] are some examples of difficulties that can be faced during the segmentation. While the experience of the physician can overcome these issues, inconsistencies among different operators may remain. Moreover, semi-automatic and fully automatic algorithms may be not able to recognise these structures as critical, if not well trained.

In the following sections we reported the results of some studies in the literature, focusing on the variability found in the segmentation of lung lesions.

Segmentation of different tumour types

Pavic et al. [127] analysed the impact of manual inter-reader segmentation considering various tumour sites. Three different cancer types (head and neck squamous cell carcinoma or HNSCC, NSCLC and malignant pleura mesothelioma or MPM) were contoured by three physicians following a common protocol. They found a variability in tumour delineation and, consequently, in radiomic feature reproducibility with different results depending on the tumour

site. They achieved a median DICE of 0.86 (range: $0.57 - 0.90$) for NSCLC, of 0.72 (range: $0.21 - 0.89$) for HNSCC and of 0.26 (range: $0 - 0.90$) for MPM. In the radiomic analysis the best reproducibility was obtained in contouring NSCLC lesions with the 90% of the features with an ICC > 0.80, for HNSCC lesions the 59% of the features was robust, whereas the MPM lesions gave the worst result with only the 36% of the features reproducible. These results emphasised the importance of performing reproducibility analysis ad-hoc for each tumour site.

Semi-automatic segmentation

The feasibility of replacing manual segmentation with a semi-automatic approach was investigated by Gu et al. [242] in 129 CT images of NSCLC patients. They developed the *single click ensemble segmentation* (SCES) algorithm, a region growing-based approach in which the tumour is contoured starting from a single seed. In this study, the results of the SCES segmentation were compared to those of two manual readers and also to two different semi-automatic approaches. The authors achieved an IoU of 79.53% between the two manual segmentations and of 78.29% and 77.72% between each of the two manual segmentations and the SCES algorithm. The proposed tool was more in agreement with the manual VOIs compared to the other two semi-automatic techniques, for which the IoU was always less than 70%. Moreover, they showed that changing several times the starting seed, the segmentation results were consistent among them (IoU = 93%). Despite the good segmentation results obtained with this algorithm, the authors highlighted issues in the delineation of part-solid tumours.

Parmar et al. in 2014 [126] analysed the impact of delineation in a dataset of 20 CT images of NSCLC patients, considering the contours made manually by five independent observers and the contours achieved by three operators twice with the *GrowCut* in 3DSlicer. The radiomic feature reproducibility was evaluated with the ICC. They found that the features were more reproducible when the semi-automatic segmentation was used (ICC = 0.85 ± 0.15) compared to the manual one (ICC = 0.77 ± 0.17). The only exception was the *shape* category for which a significant variation between the two approaches was not observed.

A similar study was performed by Owens et al. [128]. They evaluated the intra- and inter-observer variability, by comparing the manual segmentation —performed twice by three radiation oncologists —with two semi-automatic tools (*Lesion Sizing Toolkit* software and GrowCut in 3DSlicer) in 10 CT images of NSCLC patients. The performance of the segmentation was evaluated with the DICE, the HD and the ICC. Comparing the inter-reader variability in semi-automatic contours, they found a mean DICE of 0.88 ± 0.06 and 0.88 ± 0.08, and a mean HD of 0.48 ± 0.17 cm and 0.43 ± 0.20 cm for LSTK and GrowCut, respectively. From the radiomic feature analysis of the intra-reader variability, they showed better ICC results when the readers used the LSTK

tool, since it required less human interventions. Similarly, the best agreement among readers (inter-reader) was obtained with the LSTK tool. Finally, they compared the results of the different segmentation tools (but manual and semi-automatic) for the same observer (inter-software variability) and obtained a lower robustness of the features compared to intra- and inter-variability, with a mean ICC < 0.8 for all the possible pairwise comparisons.

Deep-learning segmentation

Haarburger et al. [132] investigated the effect of segmentation on radiomic features extracted from CT images of lung nodules, by comparing first the manual segmentation of four readers among them and then these contours with the output of a deep neural network, the PHiSeg [243]. The median DICE coefficient was equal to 0.87 among human operators and was 0.85 between the human readers and the PHiSeg output, therefore obtaining comparable results between manual and automatic approaches. Finally, they evaluated the outcome of the network on two different tumour types (liver and kidney tumours) and found similar ICC results among the three types of lesion. This is in contrast to [127], discussed above. One possible explanation of this discrepancy may be that the group of Haarburger tested only the automatic segmentation on liver and kidney lesions, while Pavic et al. analysed only manual contours. It seems that the automatic segmentation improves the feature reproducibility, thus reducing the inter-readers variability.

A combination of a 3D CNN (V-Net) with a 2D CNN was proposed by Gan et al. [244], using 260 patients with lung tumours. Moreover, the comparison with manual contours was also carried out taking independently the 3D and the 2D CNN results. The best performance was obtained with the hybrid network, achieving a DICE of 0.72 ± 0.10, IoU of 0.58 ± 0.13 and HD of 21.73 ± 13.30 mm (DICE $= 0.52 \pm 0.16$, IoU $= 0.40 \pm 0.17$ and HD $= 70.73 \pm 33.30$ mm for the 2D CNN and DICE $= 0.65 \pm 0.15$, IoU $= 0.53 \pm 0.14$ and HD $= 26.73 \pm 13.30$ mm for the 3D CNN).

Semi-automatic versus deep-learning segmentation

Bianconi et al. [133] compared the segmentation results of 12 semi-automatic algorithms (based on active contour, region growing, clustering, graph, thresholding, etc.) with 12 CNN-based architectures. They considered two datasets of patients with lung lesions: 111 patients from a proprietary dataset (used for training, validation and testing) and 100 from the LIDC-IDRI public one (used as independent test set). The deep learning architectures gave, in general, the best results with a DICE larger than 0.75 (best performance: DICE $= 0.853 \pm 0.082$ with the full-trained *U-Net-ResNet34*, worst performance: DICE $= 0.755 \pm 0.230$ with the fine-tuned *U-Net-MobileNet*). Using semi-automatic algorithms, instead, the DICE was always less than 0.75, except for the *morphological active contours without edges* algorithm (DICE $= 0.761 \pm 0.179$).

4.2 Preliminary investigation on automatic segmentation

In the next sections the segmentation procedures (semi-automatic and deep-learning based) adopted in this thesis to assess feature robustness will be discussed. First of all, the different contours (manual, semi- and fully automatic) will be compared in terms of the simple metrics defined in Section 4.1.1. Then, the variability introduced by segmentation in radiomic features will be evaluated, using the ICC to assess the agreement among the three types of segmentation in terms of radiomic features.

4.2.1 Datasets and CT images

We retrospectively collected two datasets of CT images, acquired at IEO using the institutional standard protocol for diagnostic chest imaging. All the enrolled patients had a histological diagnosis of lung tumour. The first dataset (Dataset A) consists of patients operated between January 2012 and August 2016, with an available pre-surgical CT image and lung cancer staged up to T3N1 [18]. In the second one (Dataset B) only advanced adenocarcinoma patients were included with a mutational status assessed between April 2016 and May 2019. A total of 270 patients and 226 patients were enrolled for the Datasets A and B, respectively.

For the first part of the analysis (semi-automatic segmentation), only a subgroup of the Dataset A (Dataset A_{sub}) and Dataset B (Dataset B_{sub}) was considered, consisting of 50 and 152 images, respectively. While in Dataset B only NSCLC patients with adenocarcinoma were enrolled, in Dataset A 255 out of 270 (94%) patients were affected by NSCLC (181/270 with adenocarcinoma and 50/270 with squamous cell carcinoma), 13 (5%) had a carcinoid tumour, the remaining 3 (1%) had a sarcoma. All the CT images were acquired on a GE scanner (GE Healthcare, Waukesha, WI) after the injection of iodinated contrast medium during the portal venous phase with a slice thickness of 2.5 mm, a slice spacing of 2.5 mm, a *Standard* convolutional kernel, a helical mode and a z-axis current modulation. **Table 4.1** and **Table 4.2** provide an overview of the clinical information, while **Table 4.4** and **Table 4.3** summarise the CT acquisition/reconstruction parameters of the datasets.

4.2.2 Lesion segmentation

Manual segmentation

All the CT images of the Datasets A and B were manually segmented by three radiologists. Each radiologist contoured a different group of patients, following common criteria (same window width and level for visualisation, exclusion of the vessels on the tumour border, inclusion of the opacity). One lesion for each patient was delineated slice by slice with the AWServer 3.2 software (Ext. 2.0

Clinical parameters		Dataset A	Dataset B
Number of patients		270	226
Age		67.4 (61.0 - 72.5)	67.8 (59.8 - 72.2)
Sex	F M	103 (38%) 167 (62%)	88 (39%) 138 (61%)
Tumour size (cm^3)		8.03 (2.39 - 30.95)	17.01 (4.24 - 56.36)
Side	R L	153 (57%) 117 (47%)	126 (56%) 100 (44%)

Table 4.1: Clinical data of the two datasets under investigation (Dataset A and Dataset B). For the continuous variables (age and tumour size), the median value along with the interquartile range (in parentheses) is displayed.

Clinical parameters		Dataset A_{sub}	Dataset B_{sub}
Number of patients		50	152
Age		67.4 (61.49 - 73.0)	68.4 (60.9 - 72.0)
Sex	F M	23 (46%) 27 (54%)	62 (41%) 90 (59%)
Tumour size (cm^3)		14.54 (4.05 - 33.86)	18.26 (5.40 - 57.48)
Side	R L	32 (64%) 18 (36%)	81 (53%) 71 (47%)

Table 4.2: Clinical data of the two sub-populations under investigation. For the continuous variables (age and tumour size), the median value along with the interquartile range (in parentheses) is displayed.

tool, GE Healthcare) and the result of the segmentation was saved in RT Structure format. The segmentation files, along with the CT image, were then converted in NRRD format using 3DSlicer (v. 4.10.0).

The manual segmentations were considered as the ground truth for the semi- and fully automatic techniques described in the next paragraphs.

Semi-automatic segmentation

The study on the performance of the semi-automatic segmentation was carried out on the CT images of the Dataset A_{sub} and Dataset B_{sub}.

CT parameters		Dataset A	Dataset B
Scanner	LightSpeed Ultra	52 (19%)	7 (3%)
	LightSpeed 16	135 (50%)	2 (1%)
	Optima CT660	69 (26%)	117 (52%)
	Discovery CT750 HD	14 (5%)	100 (44%)
Exposure (mAs)		14 (8 - 21)	7 (5 - 11)
Tube Voltage (kVp)	100	0 (0%)	29 (13%)
	120	270 (100%)	185 (82%)
	140	0 (0%)	12 (5%)
Algorithm	FBP	187 (69%)	9 (4%)
	ASIR	83 (31%)	217 (96%)
Pixel size (mm)		0.73 (0.70 - 0.79)	0.77 (0.70 - 0.82)

Table 4.3: List of the CT acquisition and reconstruction parameters of the two populations (Dataset A and Dataset B). For the continuous variables (exposure and pixel size), the median value along with the interquartile range (in parentheses) is displayed.

CT parameters		Dataset A_{sub}	Dataset B_{sub}
Scanner	LightSpeed Ultra	3 (6%)	0 (0%)
	LightSpeed 16	7 (14%)	0 (0%)
	Optima CT660	28 (56%)	85 (56%)
	Discovery CT750 HD	12 (24%)	67 (44%)
Exposure (mAs)		8 (5 - 14)	7 (5 - 10)
Tube Voltage (kVp)	100	0 (0%)	14 (9%)
	120	50 (100%)	130 (86%)
	140	0 (0%)	8 (5%)
Algorithm	FBP	10 (20%)	0 (0%)
	ASIR	40 (80%)	152 (100%)
Pixel size (mm)		0.74 (0.70 - 0.76)	0.77 (0.71 - 0.83)

Table 4.4: List of the CT acquisition and reconstruction parameters of the two sub-populations. For the continuous variables (exposure and pixel size), the median value along with the interquartile range (in parentheses) is displayed.

We produced the semi-automatic segmentations using the GrowCut algorithm available in 3DSlicer (*Grow from seeds*), taking advantages of its graph-

ical user interface (GUI). Each CT image were imported in this software and the lung lesion was identified using the most suitable visualisation window according to the lesion position: the lung window (W= 1400 HU and L = -500 HU) or the mediastinal one (W = 350 HU and L = 40 HU)[1].

After a preliminary test to learn how to use the GrowCut tool, we understood that the best results were achieved when the seeds of the background and of the lesion were placed in the axial slices corresponding to the lesion extremities along the z direction. Moreover, larger lesions as well as lesions attached to tissues with similar grey-level intensities required a higher number of seeds, by using in this case also the sagittal and coronal slices to better identify the 3D structure of the tumour.

The segmentation procedure was performed by one operator with no radiological background. In case of multiple lesions, only the lesion contoured manually by the radiologist was considered. For each patient, we recorded the time between the identification of the lesion and the confirmation of the segmentation suggested by the 3DSlicer tool.

At the end of the segmentation procedure, the resulting delineation was saved as a mask in NIfTI format.

Finally, we applied post-processing techniques to improve the quality of the segmentation, by filling holes inside the VOI and taking the maximum connected component. In our case, the connected component is a region of the binary mask composed by all pixels/voxels with intensity equal to 1 which are physically connected among them. To do so, the masks were imported in Python (v.3.7.3) and processed, by applying the function *BinaryMorphological-ClosingImageFilter()* from the SimpleITK library (v. 1.2.0) and the function *measure.label()* from the scikit-image library (v. 0.14.2), setting *background = 0*, *connectivity = None* and *return_num=True*. The parameter *connectivity = None* means that fully connected voxels are considered, which are 26-connected voxels to each voxel in case of a 3D image.

Fully automatic segmentation

In our study we exploited the nnU-Net architecture in the 2D, 3D and ensemble configurations.

[1]In practice, we collocated the seeds inside the lung lesion using the 2D round brush (*Paint* tool in the *Segment Editor* section), considering various axial slices. Then in the surrounding anatomical structures (the background) — such as the vessels, the bronchi and the lung parenchyma — the user put other seeds using a different segment label. As mentioned in Section 4.1.2, the tool shows a preview of the segmentation, and the proposed contours can be improved by adding new seeds followed by a new run of the algorithm or can be accepted. In order to avoid a too large human intervention, we accepted the proposed segmentation soon after the first run of the algorithm, except when the result was clearly not good enough (for example when a large part of the lesion was not segmented or when part of the parenchyma or of the chest wall were included in the lesion label).

First of all, the Dataset A was divided into training and testing sets: 220 CT images were used to train the network, while 50 samples (Dataset A_{sub}), unseen during the training, were used to evaluate the performance. Moreover, we also used the Dataset B_{sub} as a second testing set. The images were resampled to 256×256 in the axial plane before the training in order to reduce the computational cost of the learning process.

The network was trained on a Nvidia RTX 2080 Ti GPU Card with 11 GB of dedicated RAM memory for a minimum of 400 epochs (number of times that the entire training set is passed to the network), with a batch size (number of training data propagated through the network in each iteration during an epoch) equal to 250.

The file with the weights obtained from the training process was saved and used for the prediction of the segmentation on the testing set. We resampled the predicted masks of the 2D, 3D and ensemble configurations back to 512×512 with the function *ndimage.zoom()* of the scipy library (v. 1.2.1) using a zero spline interpolation, and we saved them in NIfTI format.

Finally, we applied a post-processing procedure. We wrote a script in Python to identify and separate all the components for each output mask and to display all the identified connected components using a GUI. Since there were cases with more than one lesion per patient or cases with pulmonary consolidations mis-classified by the network for their similarity with the target tumours, the GUI was useful to select only the one that matched the ground truth segmentation performed by the physicians. We used the function *measure.label()* from the *scikit-image* library (v. 0.14.2) to select all the connected components and the *PySimpleGUI* package (v. 4.24.0) to create the graphical interface. An example of the GUI is reported in **Figure 4.1**.

4.2.3 Radiomic feature extraction

The radiomic features were extracted with Pyradiomics (v. 2.2.0) following the same procedure and using the same parameters described in Section 2.2.2. However, for this particular investigation we considered only *original* features (meaning not filtered images) with a bin width of 25 HU.

The features were extracted from the manual, the semi-automatic and deep-learning based contours (before and after the post-processing steps).

4.2.4 Data analysis

We calculated the following evaluation metrics using the *seg-metrics* library (v. 0.0.7) in Python (v.3.7.3): DICE, IoU, HD and HD95. The definition of these metrics are in Section 4.1.1. The HD95 is equivalent to the HD coefficient, but it considers only the 95^{th} percentile of the distribution of the minimum distance between the two boundaries in order to eliminate the outliers.

We also performed a correlation analysis between the DICE results for each patient and some shape features (volume, Elongation, Flatness, Spheric-

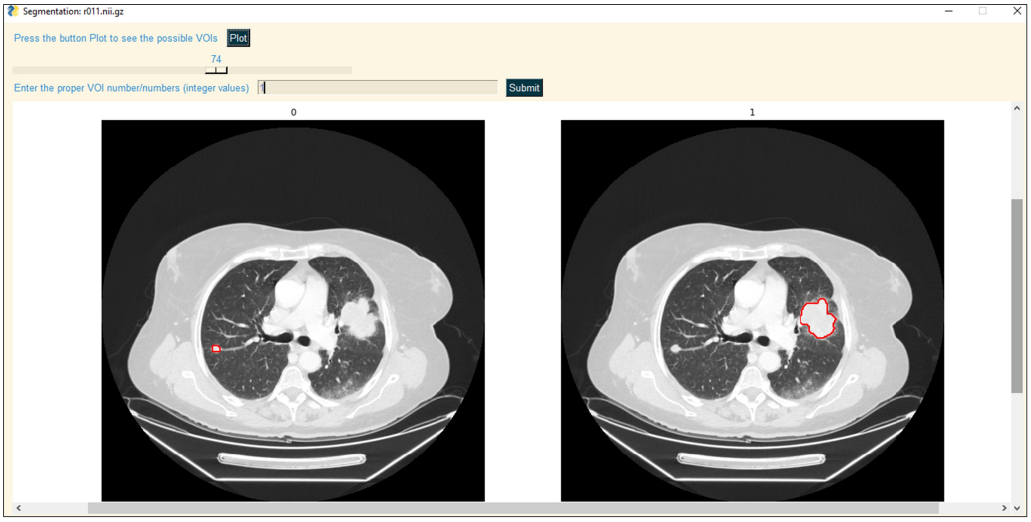


Figure 4.1: Example of the GUI developed to visualise the CT images with the network contours in red. The same CT image is showed multiple times, for each identified connected component. In the picture, two areas were segmented by the algorithm, each of them in a different lung lobe. The GUI helps to select only the correct lesion, which was the one on the right (identified with the number 1) in the displayed example.

ity, SurfaceArea and SurfaceVolumeRatio), and between the DICE and some texture information (mean, median, variance). The same comparison was done considering the HD, instead of the DICE. This information about the lesion was extracted from the manual segmentation in order to understand if the excellent or bad results from the semi- or fully automatic segmentation techniques may be associated to some basic characteristics of the lesion. The *cor.test* function from the *stats* package in R was used, setting *method="spearman"*.

We evaluated the agreement of the features between manual and semi-automatic segmentation and between manual and fully automatic one using the ICC (see Section 3.5 for details).The two segmentation tools (semi- and fully automatic ones) were also compared between them in terms of DICE and ICC. The *irr* package (v.1.0.11) in R was used to calculate the ICC using the function `icc(dataframe, type = "agreement", model = "twoway")`.

4.2.5 Results

Semi-automatic segmentation

A final mask was obtained in 193 out of 202 patients (49 from Dataset A_{sub} and 145 from Dataset B_{sub}) with the GrowCut algorithm in 3DSlicer. For the remaining 9 patients, the segmentation was not performed in 8 cases because the identification of the lesion was too difficult, mainly for the presence of pulmonary atelectasis. Since no mask was produced for these 8 patients, they were excluded from the subsequent analysis. The last of these 9 patients, instead, appeared in both the datasets with the same CT image and manual

segmentation, and therefore it was considered only once. In two cases, instead, the correct lesion was not identified, and the resulting DICE was equal to zero. For this reason, these two CT images were excluded and a total of 191 patients remained for the analysis.

In **Figure 4.2** we reported an example of the application of the GrowCut algorithm in 3DSlicer. From this picture, it appears that the role of the post-processing operations was to make the border less jagged and more similar to the continuous curve of manual segmentation.

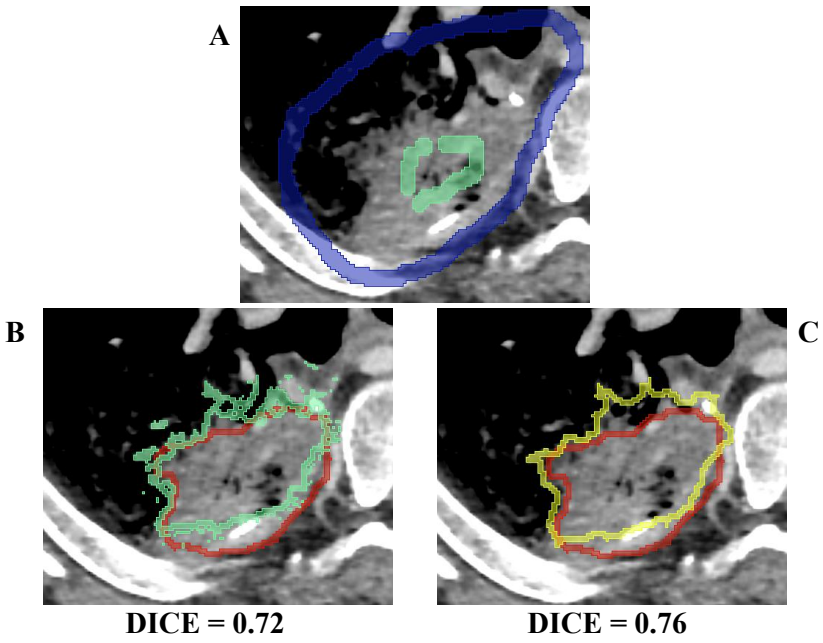


Figure 4.2: Example of the semi-automatic segmentation procedure of one lung lesion in 3DSlicer. **Picture A** shows the scribble used as seeds for the initialisation of the GrowCut algorithm (the background in blue and the lesion in green). **Picture B** and **Picture C** illustrate the resulting segmentation, before and after the post-processing procedures (filling of the holes and maximum connected component), respectively. The manual delineation is reported in red.

The average time for the execution was 66 s (min = 27 s, max = 112 s) for Dataset A_{sub} and 109 s (min = 20 s, max = 361 s) for Dataset B_{sub}. The longer time needed to delineate the lesions in Dataset B_{sub} was due to a larger size of the lesions and a higher number of cases with the lesion attached to other tissues with similar contrast. Both these two situations, in fact, required a larger number of seeds to achieve a satisfactory result.

In **Table 4.5**, the four evaluation metrics calculated for the 191 lesions, after the post-processing procedures, are listed. The metrics achieved using the masks without post-processing were, nevertheless, quite similar: average DICE equal to 0.76 ($\pm$0.14), average IoU equal to 0.63 ($\pm$0.15) and average HD equal to 17.76 mm ($\pm$15.29 mm). We found similar results when the two datasets (Dataset A_{sub} and Dataset B_{sub}) were analysed separately, with

a slightly better performance for the Dataset A_{sub}. We obtained an average DICE equal to 0.79 (±0.10) and 0.76 (±0.15), an average IoU equal to 0.66 (±0.13) and 0.63 (±0.17), an average HD equal to 14.20 mm (±10.51 mm) and 18.04 mm (±16.47 mm), and an average HD95 equal to 6.01 mm (±6.42 mm) and 7.95 mm (±11.28 mm) for Dataset A_{sub} and Dataset B_{sub}, respectively. These values refer to the segmentation after the post-processing phase, but a similar performance was found without post-processing.

	DICE	**IoU**	**HD**	**HD95**
Mean	0.77 ± 0.14	0.64 ± 0.16	17.05 ± 15.23	7.45 ± 10.28
Median	0.81	0.67	13.12	4.39
Min	0.18	0.10	2.50	1.27
Max	0.94	0.88	112.46	71.53

Table 4.5: Evaluation metrics between the semi-automatic and the manual contours, after the application of the two post-processing techniques. The mean, the standard deviation, the median, the maximum and the minimum are evaluated among the 191 segmentations.

Figure 4.3 illustrates two examples of the semi-automatic segmentations (blue), overlapping the corresponding manual contour (red) for a very high and a very low value of the DICE.

A total of 153 radiomic features were extracted from the 191 CT images segmented with the manual and semi-automatic techniques. We found a mean ICC of 0.67±0.18 (median: 0.69, range: 0.24-0.96) between the features extracted from the manual and the semi-automatic approaches with quite identical results between the use or not of the post-processing. The majority of the features (76%) had an ICC < 0.85, indicating a low reproducibility between the two segmentation techniques. This percentage slightly (69%) reduced when only the patients with a DICE > 0.5 were included in the analysis (for this group of 181 patients the mean DICE slightly increased to 0.79±0.09 and the HD reduced to 15.87±13.77).

Analysing each category of features separately, only the *shape* category had a mean CCC over 0.80 (shape: ICC = 0.84 ± 0.14, glcm: ICC = 0.69 ± 0.15, firstoder: ICC = 0.66 ± 0.20, gldm: ICC = 0.65 ± 0.20, glrlm: ICC = 0.64 ± 0.24, ngtdm: ICC = 0.56 ± 0.14, glszm: ICC = 0.59 ± 0.24). **Figure 4.4** shows the ICC results for each extracted feature for the comparison between the manual segmentation and the semi-automatic one, after post-processing. The list of feature names associated to the numbers on the x-axis is reported in **Table B.3** of the Appendix B.

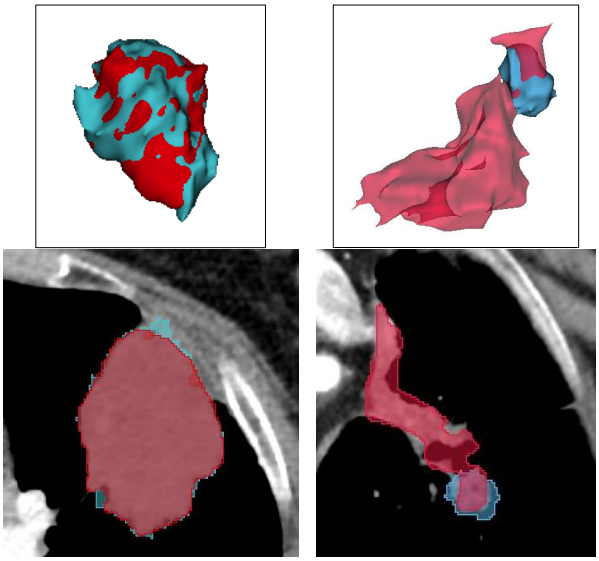


Figure 4.3: Visual comparison of the overlapping between the manual segmentation (in red) and the semi-automatic one (in blue), in case of a high DICE (on the left) and of a low DICE (on the right) value. The comparison between the two delineations is illustrated by overlapping the entire 3D structure of the lesion (on the top) and a single axial slice (on the bottom).

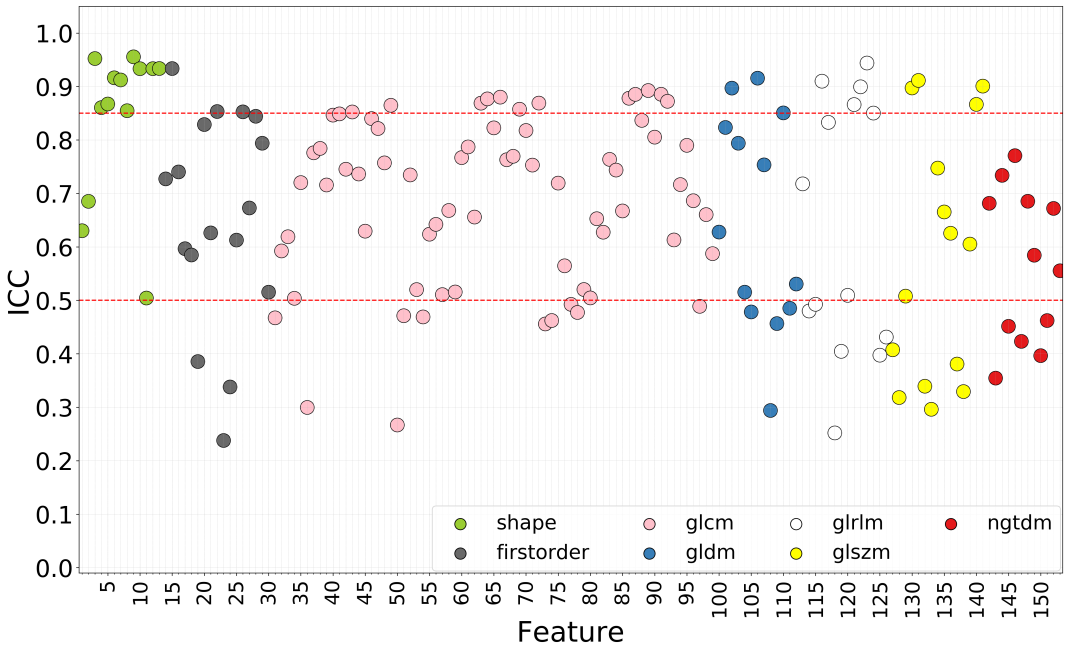


Figure 4.4: ICC value for each feature, indicating the agreement between the manual and the semi-automatic (with post-processing) contours. The different colours correspond to the different categories of features.

The results of the correlation between the DICE/HD and some basic features of the *shape* and of the grey-level histogram did not show any significant association. The highest Spearman correction coefficient was obtained with the SurfaceVolumeRatio (ρ_S = -0.54) for the DICE and with the SurfaceArea (ρ_S = 0.51) for the HD.

Fully automatic segmentation

As for the semi-automatic segmentation, one of the patient appeared identically in the Dataset A_{sub} and Dataset B_{sub}, therefore this case was excluded from Dataset B_{sub}. The remaining 201 patients were therefore analysed.
The automatic algorithm gave an empty output in 11, 2 and 9 out of 201 patients for the 2D, 3D and ensemble configurations, respectively. Moreover, even if the mask was not empty, we obtained a DICE coefficient equal to 0 in 12, 13 and 7 out of 201 patients for the 2D, 3D and ensemble configurations, respectively.

Table 4.6 and **Table 4.7** summarise the results of the comparison between the manual and the fully automatic contours in terms of evaluation metrics, before and after the selection of the connected components with the GUI, respectively. The total number of patients with DICE larger than zero was reported too.

Configuration	n° cases DICE>0		DICE	IoU	HD	HD95
2D	178	**Mean**	0.64 ± 0.27	0.52 ± 0.26	57.15 ± 60.65	31.89 ± 48.93
		Median	0.74	0.58	23.96	7.91
		Min	0.01	0.00	2.69	1.71
		Max	0.94	0.89	238.13	228.43
3D	186	**Mean**	0.69 ± 0.21	0.57± 0.22	127.10 ± 81.44	56.09 ± 76.48
		Median	0.76	0.62	149.67	8.07
		Min	0.02	0.01	2.50	1.48
		Max	0.94	0.89	299.09	278.26
ensemble	185	**Mean**	0.69 ± 0.24	0.57 ± 0.24	41.81 ± 55.47	21.97± 41.31
		Median	0.77	0.62	15.13	5.00
		Min	0.02	0.01	2.41	0.73
		Max	0.95	0.90	296.27	221.74

Table 4.6: Evaluation metrics for the three deep-network configurations (2D, 3D and ensemble) compared to the manual contours. The results refer to the automatic segmentation before the selection of the connected components with the GUI. The metrics are evaluated among the patients with a DICE larger than zero (of a total of 201 patients).

The advantage of using the GUI for the selection of the correct lesion is particularly evident for the 3D configuration, where the multiple lesions present in the lung for some patients were often identified by the deep network.

Configuration	n° cases DICE>0		DICE	IoU	HD	HD95
2D	181	**Mean**	0.66 ± 0.25	0.54 ± 0.25	19.63 ± 22.59	11.26 ± 14.37
		Median	0.75	0.60	12.53	5.53
		Min	0.01	0.00	2.64	1.71
		Max	0.94	0.89	200.73	84.10
3D	186	**Mean**	0.76 ± 0.16	0.63 ± 0.17	14.92 ± 13.69	7.63 ± 9.54
		Median	0.81	0.68	9.95	3.74
		Min	0.04	0.02	2.23	0.74
		Max	0.94	0.89	101.26	60.93
ensemble	185	**Mean**	0.71 ± 0.21	0.59 ± 0.22	15.92± 15.48	8.85 ± 11.56
		Median	0.77	0.63	9.89	4.35
		Min	0.03	0.01	2.41	0.73
		Max	0.95	0.90	91.44	76.15

Table 4.7: Evaluation metrics for the three network configurations (2D, 3D and ensemble) between automatic and manual contours. The results refer to the automatic segmentation after the selection of the connected components with the GUI. The metrics are evaluated among the patients with a DICE larger than zero (of a total of 201 patients).

Since the 2D configuration gave the worst results in terms of DICE, HD and number of empty contours (see **Table 4.6** and **Table 4.7**) and the ensemble configuration included partially the 2D results, we decided not to consider the 2D configuration for the radiomic feature analysis. For the 3D and the ensemble configurations, instead, we extracted the 153 radiomic features. We computed the ICC and compared the two configurations (3D and ensemble) to the manual segmentation, before and after the GUI application. We observed an increase of the ICC when the VOIs were selected with the GUI, both for the ensemble and the 3D configurations. However, the reproducibility of the features is very low, with an ICC < 0.85 for almost all the features (99% for the two ensemble configurations and the 3D one without GUI selection, 97% for the 3D one with GUI selection). **Figure 4.5** displays the ICC value for each feature for the 3D configuration with GUI selection. Even considering only the patients with a DICE > 0.5 (94% with ICC < 0.85 for 3D configuration with GUI selection) or only patients with a HD < 30 mm (88% with ICC < 0.85 for 3D configuration with GUI selection) the agreement did not improve.

The correlation analysis between the DICE/HD and the lesion properties extracted from the manual VOIs did not show a strong association, as for the semi-automatic approach. As concerns the correlation with the HD, the Sphericity feature was the most associated (ρ_S = -0.57 for the ensemble, ρ_S = -0.61 for the 3D, ρ_S = -0.42 for the 2D configuration). For the DICE the feature more correlate is the SurfaceVolumeRatio (ρ_S = -0.50 for the ensemble, ρ_S = -0.47 for the 3D, ρ_S = -0.53 for the 2D configuration).

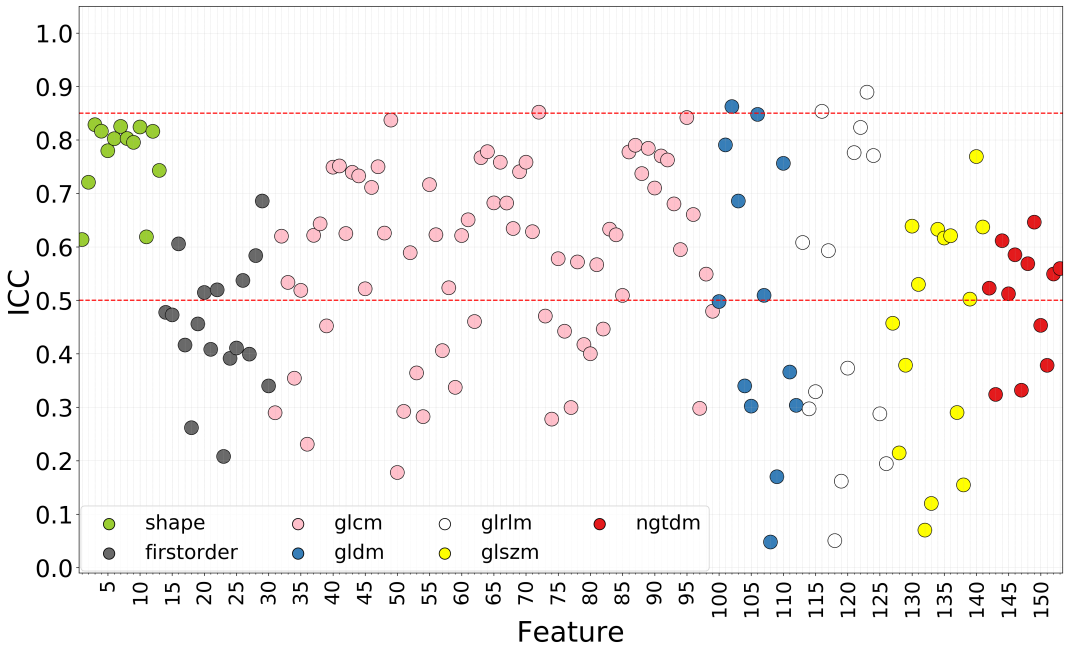


Figure 4.5: ICC value for each feature, indicating the agreement between the manual and the automatic (with GUI selection) contours for the 3D configuration. The different colours correspond to the different categories of features.

Segmentation performance divided by radiologists

We also investigated whether any relevant discrepancy between manual and semi- or deep-based contours could be associated to a different segmentation approach followed by one of the three radiologists who performed the manual contours. As mentioned above, the three radiologists segmented different patients and, therefore, we could not compare the performance among them. However, we can evaluate if one of them contoured in a systematically different way the lesions, by exploiting the objectivity of the algorithms used in this study.

To this aim, we divided the patients into three groups according to the radiologist who made the corresponding contours. For each group we evaluated the DICE and the HD. All the 50 images in dataset A_{sub} were contoured by the same radiologist. The images in dataset B_{sub}, instead, were contoured by two different radiologists, who worked on 82 and 69 images, respectively. The values of these metrics were quite homogeneous among the three groups, both for the semi- and for the deep-based segmentation. The full results are reported in **Table 4.8**.

Semi- versus fully automatic segmentation

For the sake of completeness, we compared the two tools between them (semi- and fully automatic after the post-processing procedures). We found that the

Configuration	Group	n° cases total	n° cases DICE>0	DICE	HD
Semi-automatic	1	50	49	0.79 ± 0.10 (0.79)	14.20 ± 10.51 (14.20)
	2	82	66	0.77 ± 0.14 (0.81)	18.11 ± 13.78 (14.90)
	3	69	76	0.75 ± 0.16 (0.80)	17.95 ± 19.22 (12.72)
Fully automatic ENSAMBLE	1	50	47	0.62 ± 0.24 (0.70)	18.88± 17.16 (12.70)
	2	82	73	0.76 ± 0.20 (0.83)	13.99 ± 16.06 (8.49)
	3	69	65	0.73 ± 0.18 (0.77)	15.95 ± 13.28 (9.63)
Fully automatic 3D	1	50	47	0.72 ± 0.17 (0.76)	16.25 ± 12.45 (12.38)
	2	82	74	0.78± 0.16 (0.83)	13.12 ± 14.90 (8.02)
	3	69	65	0.78 ± 0.13 (0.81)	16.02 ± 13.08 (10.01)

Table 4.8: DICE and HD values for the three groups of patients, each segmented by a different radiologist. Group 1 refers to dataset A_{sub}, groups 2 and 3 refer to the two subgroups of dataset B_{sub}. The metrics are given as mean ± standard deviation (median) and were calculated only for the contours with a DICE > 0 and after the post-processing. .

DICE and the ICC were quite similar to what we obtained before: DICE = 0.79 ± 0.14, ICC = 0.65 ± 0.21 and 180 patients with DICE > 0 when 3D configuration was used, DICE = 0.72 ± 0.22, ICC = 0.49 ± 0.23 and 182 patients with DICE > 0 in case of the ensemble configuration (as above the DICE and ICC values were calculated considering only the contours with DICE > 0). These results indicated that also between the two non-manual techniques there is not a perfect agreement.

Summary

In **Table 4.9**, the results of the ICC analysis are summed up, both for the semi- and the fully automatic segmentation compared to the manual contours.

Figure 4.6 illustrates four lesions segmented using the three types of techniques (manual, semi-automatic and automatic), with different level of agreement among them.

Configuration	DICE	ICC	% ICC<0.85	% ICC<0.50
GrowCut without pp	0.75 ± 0.14	0.67 ± 0.18	76%	22%
GrowCut with pp	0.76 ± 0.14	0.67 ± 0.18	76%	22%
3D without GUI	0.69 ± 0.21	0.45 ± 0.20	99%	57%
Ensemble without GUI	0.69 ± 0.24	0.43 ± 0.18	99%	63%
3D with GUI	0.76 ± 0.16	0.55 ± 0.20	97%	37%
Ensemble with GUI	0.71 ± 0.21	0.48 ± 0.18	99%	57%

Table 4.9: Summary of the DICE and the ICC analysis for the comparison between manual and full semi-automatic contours, and between manual and automatic contour. **pp** = post-processing

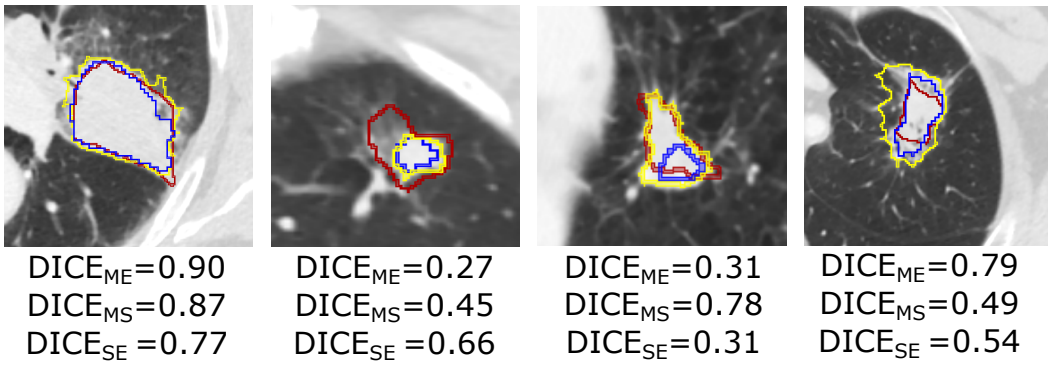


Figure 4.6: Comparison between the manual (**M**, **red**), the ensemble (**E**, **blue**) and semi-automatic (**S**, **yellow**) contours. From the left to the right, the figure shows a case of good overlapping between the three contours, only between fully automatic and semi-automatic, only between the manual and semi-automatic, only between manual and fully automatic.

4.2.6 Discussion

In this preliminary analysis, we used one semi-automatic (GrowCut from 3DSlicer) and one fully automatic tool (nnU-Net) for the segmentation of the lung tumours. The patient population included in the study was quite various, since it consisted of patients with the same tumour type but different stage, resulting in a large variability of tumour size and lung involvement. Atelectasis, GGO and multiple lesions per patient were sometimes found in the datasets.

Practical considerations on GrowCut and nnU-Net tools

The GrowCut algorithm is quite simple to use, and it does not require intense computational resources. However, a priori capability of the user in the identification of the lesion is necessary since the algorithm is not able to automatically detect the lesion. Moreover, the need for initialisation seeds, whose number and position are set by the user, makes it prone to inter- and intra-user variability. In this study we did not evaluate this last point, because our aim was to compare its performance against a deep-learning based algorithm. Using the same algorithm, for example, Parmar et al. showed that the inter-

and intra-observer variability when the GrowCut algorithm was used was lower than the manual inter-rider variability in terms of radiomic features [126], highlighting the best reproducibility of the software compared to a merely human approach. However, further investigations on this aspect should be carried out, if this tool is chosen for segmentation in radiomic studies.

Concerning the automatic segmentation using a deep neural network, the involvement of the user was considerably reduced. In fact, the only task of the user is checking the algorithm output and, if necessary, making the appropriate corrections. For example, in case of multiple lesions inside the lung the network is not able to identify the one of interest, and it instead contours all the structures that it considers similar to the labels learned during the training process. To overcome this issue, in this study we created a GUI in the Python environment to help the user in the identification and selection of the lesion among all the displayed connected components. The main limitation of the developed GUI was the slowness, particularly when many connected components (more than four) were present. An optimisation procedure of the code or the use of proper software for the GUI implementation may improve this tool.

Segmentation performance

In general, we observed a relevant difference between the proposed approaches and the manual segmentation, as concerns both the evaluation metric and the radiomic features. The best results in terms of DICE and HD were obtained using the nnU-Net in the 3D configuration after the connected component selection with the GUI (see **Table 4.7**). The performance of the semi-automatic algorithm with the application of the post-processing (hole filling and maximum connected component) was, however, quite similar (see **Table 4.5**). Moreover, in the semi-automatic segmentation the post-processing did not improve considerably the results (Wilcoxon p-value non-significant between before and after the post-processing for the DICE, HD and ICC). Even if we observed visually a reduction of the noise on the border of the lesion (**Figure 4.2**), the applied post-processing acted probably on a too small number of voxels compared to the entire segmented volume to produce a detectable impact.

The worst results were achieved, instead, with the automatic 2D configuration in terms of DICE and HD values. The output masks of this architecture showed very often a discontinuity along the z direction. This is probably due to how it intrinsically works, since for each patient it looks at each axial slice separately, and the context information along the other planes is lost. For this reason, the ensemble or the 3D configurations should be preferred to the 2D architecture for volumetric segmentation. Gan and colleagues reached a similar conclusion, by comparing a 2D CNN, a 3D CNN (V-Net) and a combination of these two architectures [244].

For the automatic segmentation the presence of atelectasis and pleura effusion confused the network, reducing its performance for these cases. In these

complex situations the human intervention is therefore necessary. For the semi-automatic approach, instead, the part-solid lesions were not properly contoured when the seeds were not put also inside the GGO area. In general, a higher human effort is required with the GrowCut algorithm when the lesions are characterised by wide areas with very different grey-level intensities (such as part-solid lesions or lesions with cavitation).

The agreement analysis with the ICC between manual and semi- or fully automatic segmentation was in general not very encouraging. In all the configurations, the mean ICC among the features was pretty low (see **Table 4.9**), resulting in an influence of the segmentation within all the categories of features (apart from the *shape* category). In fact, we observed a DICE coefficient slightly lower than in the literature (see Section 4.1.4), and this non-excellent overlapping may considerably impact on the feature stability. However, many of these studies compared contours generated multiple times by the same or different persons using the same segmentation software (intra- and inter-operator variability), and less often they compared different tools. Tunali et al. [89] and Owens et al. [128], for example, investigated and compared the impact of various semi-automatic tools on the radiomic feature for lung tumour segmentation. They showed a reduction of the feature reproducibility when different tools were used compared to inter- and intra-reader variability. These results highlight the importance of using a unique segmentation software during the same radiomic analysis, since different approaches can use different a priori knowledge and therefore produce different contours. For future analyses, it would be interesting to evaluate the feature stability among various manual readers in order to understand if the differences observed in this study between manual and non-manual algorithms are comparable or not to the inter-reader variability.

Whereas in the study described above we got familiarised with the segmentation tools and we evaluated their impact using simply an agreement analysis, in the next section the use of different segmentation techniques will be explored in a more clinical application. For this investigation, we will use only the fully automatic tool, based on the nnU-Net.

4.3 Automatic vs manual contours for OS prediction

The main purpose of this second part is to understand whether the difference observed between the radiomic features extracted from the manual contours and those extracted from the deep-learning based ones significantly impacts the performance of radiomic predictive models. To achieve this goal, we developed a model predicting the overall survival (OS) in patients with lung cancer.

4.3.1 Materials and methods

For this analysis we considered the entire Datasets A and B, already described in Section 4.2.1 (see **Table 4.1** and **Table 4.3** for clinical and acquisition information). In order to increase the amount of data for the training procedure, we included a third dataset (Dataset C). This is a publicly available dataset from the MAASTRO Clinic (Maastricht, The Netherlands), composed of 422 patients with a histologically proven malignant lung tumour. This dataset corresponds to the Lung1 dataset of the NSCLC-Radiomics collection and can be downloaded from TCIA (https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics). The available clinical and acquisition information can be found in ref. [16]. Since in some cases the primary lesion was not present and in other cases the CT images were corrupted (and were inconsistent with the corresponding segmentation file), a total of 412 patients of the Lung 1 dataset were included in the analysis. The manual contours of the Dataset C were revised and — when necessary — corrected by a radiation oncologist.

We used Datasets B and C for the training of the segmentation network (638 images in total), while the entire Dataset A (270 images) was used to test independently the performance of the segmentation. In this way, we had a complete and clinically uniform dataset — the Dataset A — with both automatic and manual contours for the next task (aka the OS prediction). We chose the Dataset A for the OS modelling because all the clinical information had already been collected and used in ref. [18].

Automatic segmentation

The nnU-Net was trained, as described in Section 4.2.2. As in the previous study, we considered 2D, 3D and ensemble configurations. Besides the increased number of labelled samples in input, we also trained the network using the images and masks in their full resolution (512×512), instead of 256×256 as before, in order to retrieve back possible spatial information lost during the resampling procedure.

Moreover, instead of the GUI, for the post-processing we selected the connected component from the output masks of the network that overlapped mostly with the corresponding labelled mask of the radiologist. With this post-processing procedure we tried to simulate in a rudimentary way the selection of the correct lesion from the physician. Compared to the usage of a GUI, this approach was quicker, but it was not generalisable because it required necessarily the manual ground truth.

The DICE and the HD coefficients were extracted for the evaluation of the segmentation quality. The same 153 radiomic feature (from non-filtered images) were extracted both from manual and automatic segmentation, and the ICC was calculated.

Survival analysis

The patients with both manual and automatic contours was randomly split into training (70%) and validation (30%) sets for the purpose of OS modeling. The clinical information of the patients used for the OS model is reported in **Table 4.10**. The overall survival (in months) was calculated as the time distance between the date of the CT acquisition, considered for the radiomic study, and the date of the death or of the last follow-up.

The OS was assessed using the LASSO Cox regression model, which includes both the selection of the relevant features and the identification of the features associated to the outcome. A radiomic score was therefore created as the linear combination of the selected features weighted by their respective LASSO coefficients. The accuracy of the prediction obtained with the identified radiomic score was evaluated with the Harrel concordance index (c-index), both in the training and validation datasets. This index measures the goodness of the model: the closer its value is to 1, the better is the performance of the model. The 95% confidence interval (CI) was calculated as well. Starting from the radiomic score, the subjects were divided in two groups, high-risk (high radiomic score) and low-risk (low radiomic score). The third quartile of the radiomic score was used as the threshold to separate the two groups.

A clinical multivariable model, including only those parameters associated to the OS in the univariate analysis, was also developed starting from the collected clinical variables (age, sex, side, site, histological type, grading, pT and pN). As for the radiomic score, the clinical score was calculated as the linear combination of the selected clinical variables weighted by their respective coefficients.

Finally, from the clinical and radiomic scores we created a clinical-radiomic model with a Cox regression multivariable model.

The radiomic and clinical-radiomic models were compared to the clinical one using the likelihood ratio test [245] in order to understand if radiomic features are adding more information compared to the clinical one in the determination of the OS. Moreover, the results of the models obtained from the automatic and manual contours were compared, using the partial likelihood ratio test [246, 247].

4.3.2 Results

Segmentation quality with DICE and ICC

The DICE coefficient is listed in **Table 4.11** for the three automatic configurations, without applying any post-processing. This data refers only to the contours with DICE > 0. The best results in terms of the DICE were achieved with the ensemble configuration, while with the 3D architecture we obtained the smaller number of patients with DICE $= 0$.

In order to increase the quality of the segmentation we replicated the physi-

Characteristics		**All patients** N = 242	**Training set** N = 169	**Validation set** N = 73
Age (years)		67.48 (61.39-72.57)	66.89 (60.78-71.94)	68.70 (63.37-73.64)
Sex				
	F	87(36%)	62 (37%)	25 (34%)
	M	155 (64%)	107 (63%)	48 (66%)
Site				
	upper	142 (59%)	99 (59%)	43 (59%)
	medium	10 (4%)	6 (4%)	4 (5%)
	lower	79 (33%)	57 (34%)	22 (30%)
	mixed	11 (5%)	7 (4%)	4 (5%)
Side				
	L	104 (43%)	76 (45%)	28 (38%)
	R	138 (57%)	93 (55%)	45 (62%)
pT				
	0	2 (1%)	1 (1%)	1 (1%)
	1	78 (32%)	53 (31%)	25 (34%)
	2	117 (48%)	85 (50%)	32 (44%)
	3	45 (19%)	30 (18%)	15 (21%)
pN				
	pN0	176 (73%)	120(71%)	56 (77%)
	pN1	66 (27%)	49 (29%)	17 (23%)
Size (cm^3)		9.67 (2.84-36.81)	11.78 (2.99-40.00)	7.36 (2.64-26.68)
DICE		0.81 (0.73-0.87)	0.82 (0.72-0.87)	0.81 (0.75-0.87)
HD (mm)		8.18 (4.18-14.28)	8.49 (5.00-14.41)	7.55 (4.02-13.13)
Status				
	Alive	170 (70%)	121(72%)	49(67%)
	Deceased	72 (30%)	48 (28%)	24 (33%)
Follow-up (months)		63.0 (34.1-79.2)	63.8 (35.6-77.9)	61.3 (28.4-80.6)

Table 4.10: Clinical data, including follow-up and survival information, for the patients in Dataset A used for the OS analysis. The data are listed for all the patients, only for those used in the training and only for those used in the validation of the model. The follow-up information is given as: median (interquartile range).

cian check in a simplified way. We included in the analysis only the ensemble segmentation. However, for patients with very low ensemble performance compared to the 3D one, the 3D configurations were taken. Then, only the connected component which matched the labelled data was considered. Finally,

Configuration	n° cases DICE>0	DICE	
2D	233	**Mean**	0.68 ± 0.24
		Median	0.77
		Min	0.004
		Max	0.96
3D	260	**Mean**	0.66 ± 0.24
		Median	0.74
		Min	0.02
		Max	0.96
ensemble	246	**Mean**	0.72 ± 0.23
		Median	0.80
		Min	0.001
		Max	0.97

Table 4.11: DICE coefficient for the three configurations (2D, 3D and ensemble), without the application post-processing. The DICE values in the table refer only to the patient with DICE > 0. The number of patients with DICE > 0 (of a total of 270) is also displayed.

the contours with a DICE < 0.3 were excluded. In conclusion, a total of 242 patients out of 270 were included in the final analysis. **Table 4.12** provides a summary of the evaluation metrics (DICE, HD and ICC) obtained among the 242 automatic delineations we did not rejected after the selection during the post-processing procedure on the initial 270 patients.

	DICE	HD	ICC
Mean	0.78 ± 0.12	11.85 ± 11.47	0.66 ±0.19
Median	0.81	8.18	0.67
Min	0.33	2.50	0.02
Max	0.97	79.78	0.97

Table 4.12: Results of the DICE, HD and ICC for the 242 contours remained after the post-processing techniques.

The results of the DICE and HU are quite similar to our best configuration in the previous analysis on a limited number of patients. The ICC, instead, seems improved and more similar to the results achieved with the semi-automatic segmentation. In this case, we obtained that the 83% of the features had a $ICC < 0.85$. **Figure 4.7** provides a graphical overview of the ICC values for all the 153 features.

It can be noted that a very good agreement was obtained with almost all the *shape* features, nevertheless the majority of the firstorder and texture features had a poor or a moderate agreement ($ICC < 0.75$).

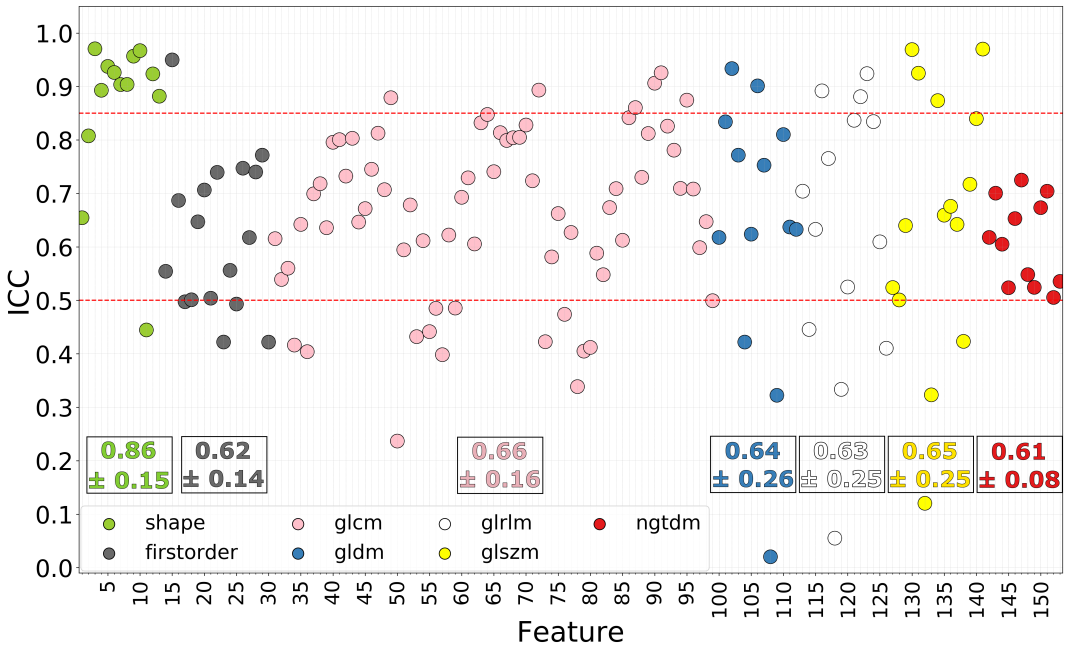


Figure 4.7: ICC value for each feature, indicating the agreement between manual and automatic (with post-processing). The different colours correspond to the different categories of features. The mean and the standard deviation of the ICC for each feature category are reported too.

The next section will show to what extent this poor agreement impacts on the performance of a clinical model, starting from the radiomic features extracted from the manual and the automatic contours of the 242 patients.

Survival analysis

The analysis was performed twice (**Table 4.13**). First of all, the 169 patients with manual contours were used for the training of the OS model, while the same 169 patients with automatic contours, the 73 with manual contours and the same 73 with automatic contours were considered as the test sets (case A). Then, the model was built using the 169 patients with automatic contours and tested on the other three groups of patients (case B).

The Cox regression LASSO model identified five and seven radiomic features for the radiomic score in case of training done with manual (case A) or automatic (case B) contours, respectively. **Figure 4.8** shows the features that mostly contributed in the prediction of the OS in the training set, along with the coefficient of the model for each significant feature.

The features that were in common between the two types of segmentation were the *correlation* from the glcm, which measures the level of correlation between a voxel and its neighbour, and the *SurfaceArea* from the *shape* category. The *Maximum3DDiameter* and the *MajorAxisLength* features, instead, are quite similar and describe the 3D extension of the tumour. The maxi-

	Training	Validation
Case A	Manual contours (169 pts)	Manual (73 pts) Automatic (169 pts) Automatic (73 pts)
Case B	Automatic contours (169 pts)	Automatic (73 pts) Manual (169 pts) Manual (73 pts)

Table 4.13: Schematic description of the datasets used for training and validation in the two analysed cases. The number of patients used in each group of patients is reported in parentheses.

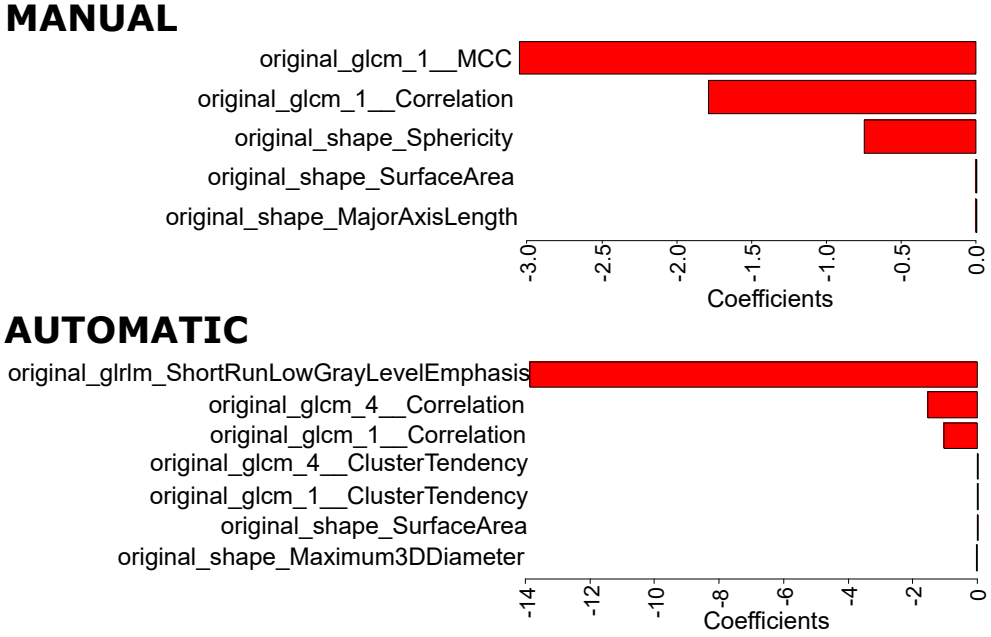


Figure 4.8: LASSO coefficients for the features that mostly contribute to the OS prediction, when the model was trained with manual (top) and when with automatic (bottom) contours.

mum diameter is indeed one of the radiological parameters commonly used in the clinical practice to evaluate the extension and to monitor the evolution of the disease. The physician usually identifies the axial slice with the largest diameter and manually measures it. However, this procedure is performed in only one slice and therefore in 2D, while the *shape* features appearing in the radiomic scores are calculated from the 3D volume, providing a more complete information.

The survival probability curves (Kaplan–Meier curves) for the two groups of patients, the 169 used for the training and the 73 for the validation of the model, are displayed in **Figure 4.9**, along with the percentage of deaths in each

population.

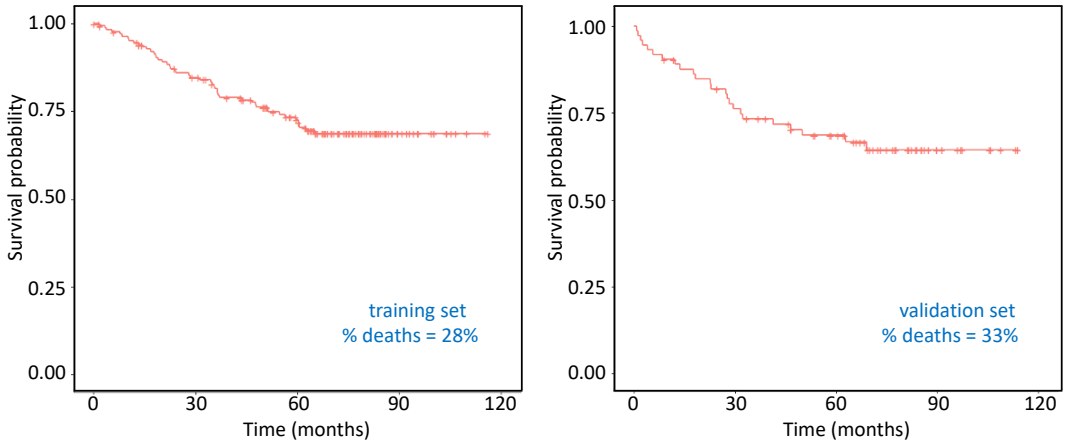


Figure 4.9: Kaplan–Meier curves, in months, for the 169 patients of the training (**Figure a**) and the 73 patients in the validation set (**Figure b**).

The same curves stratified for the high- (over the third quartile of the radiomic score) and low-risk (below the third quartile of the radiomic score) groups are included in **Figure 4.10** for the case A.

The plots show a good separation between the two groups in the 169 patients with manual segmentation used for the training and in the same patients with automatic contours used in this case as a test set (Log Rank test $p-value < 0.05$), with higher survival probability for the group with the lower radiomic score. However, the capability of the model in the generalisation of the prediction result was not satisfactory and, in fact, the separation of the curves in the validation sets was not good (non-significant p-value), when the 73 patients were used for the validation, both with manual ($p - value = 0.15$) and automatic ($p - value = 0.84$) contours.

Similar results were obtained in case B (features from the 169 automatic contours used for the training). The separation was good (Log Rank test $p - value < 0.05$) both in case of the 169 patients used for the training (automatic contours) and the same patients used in the model validation (manual contours). The performance, instead, was not good when the validation was performed on the 73 patients with automatic ($p - value = 0.91$) and manual ($p - value = 0.086$) contours.

It should be noted that the worst separation was achieved with the 73 patients with automatic contours, both when the model was trained with the automatic contours and when with the manual ones.

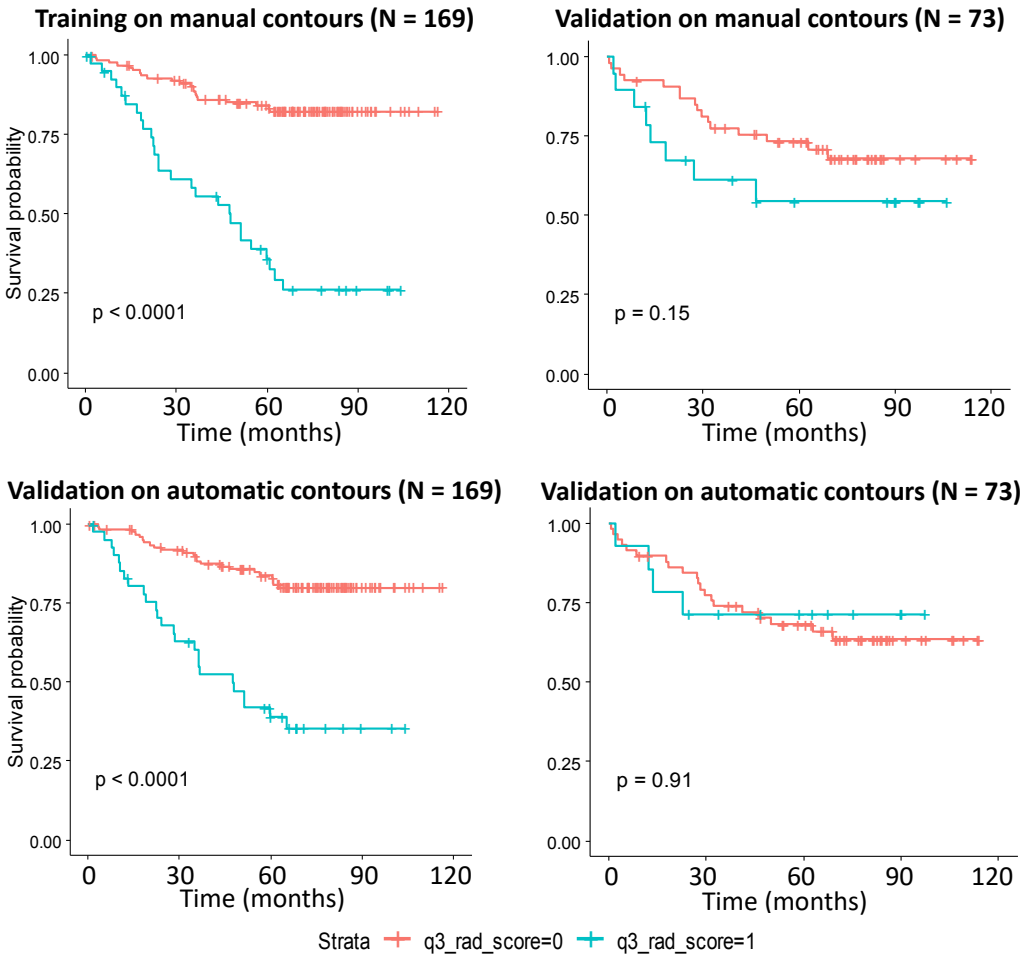


Figure 4.10: Kaplan–Meier curves, in months, for the case A (features from the 169 manual contours used for the training). The curves obtained for the training set is on the top left. The curves for the validation on the 73 patients with manual contours is on the top right, on the 169 patients with automatic contours on the bottom left, on the 73 patients with automatic contours on the bottom right. The pink curves refer to the survival probability curve for the low-risk group ($\leq q3$), while the light-blue ones for the high-risk group ($> q3$).

The results of the OS model — in terms of the c-index — are included in **Table 4.14** for case A and case B, evaluating the performance both in the training and in the validation sets. Moreover, the table displays the c-index values for the model built using only the radiomic features, using only the clinical information, and combining clinical and radiomic parameters. In the clinical model the clinical variables that were more associated to the OS were pT and site.

The most interesting result was that a significant difference between automatic and manual contours was not observed in the performance of the OS model (partial likelihood ratio test $p - value > 0.05$). This behaviour was found both in the training set between automatic and manual contours ($p - value = 0.45$ for the radiomic model, $p - value = 0.44$ for the clinical-

radiomic model), and in the validation set. The latter was evaluated comparing the performance of the model trained with manual contours and validated on manual contours with the performance of the model trained with automatic contours and validated on automatic contours ($p - value = 0.52$ for the radiomic model, $p - value = 0.41$ for the clinical-radiomic model).

As concerns the performance of the predictive model, the best result was obtained with the clinical-radiomic model in the training set, both in case A and case B, with comparable results between the two cases. Moreover, the introduction of radiomic features increased the capability of the model to predict the OS compared to using only the clinical information ($p\text{-}value < 0.01$ between clinical-radiomic and only clinical models). When the model was validated on the group of 73 patients, instead, the performance was in general very low for all the configurations (only radiomic, only clinical, only clinical-radiomic). The best c-index in validation was obtained with the clinical-radiomic model when radiomic features were extracted from the manual segmentation ($c - index = 0.65$), both in case A and in case B.

		c-index (95% CI)		
Model	**Contours used for the training**	**Training *** (N = 169)	**Validation** **manual** (N = 73)	**automatic** (N = 73)
radiomic	manual	0.69 (0.62-0.75)	0.57 (0.47-0.67)	0.52 (0.42-0.61)
clinical-radiomic	manual	0.72 (0.67-0.81)	0.65 (0.52-0.76)	0.60 (0.47-0.71)
radiomic	automatic	0.67 (0.60-0.74)	0.58 (0.49-0.68)	0.50 (0.42-0.59)
clinical-radiomic	automatic	0.71 (0.65-0.80)	0.65 (0.53-0.76)	0.59 (0.47-0.70)
clinical	–	0.64 (0.58-0.73)	0.61 (0.49-0.72)	

Table 4.14: Performance of the model in the training and validation sets, for the automatic and manual contours. The first two rows refer to the case A, while the next two rows to case B. The last row, instead, refers to the clinical model, where the radiomic features were not used, and therefore the results were independent from the segmentation type.
* The performance was evaluated on the 169 patients used for the training of the model (manual in case A, automatic in case B).

4.3.3 Discussion and future improvements

The main result of this segmentation study was that, when two different segmentation techniques were used (one automatic based on a deep-learning algorithm and the other manual) for the prediction of the overall survival in patient with lung cancer, a non-significant different was observed between the two approaches (p-value > 0.05). The performance of the OS model built with the automatic and the manual contours was in fact similar, although the results achieved with the DICE and the ICC metrics were not promising.

As concerns the DICE results, we did not observed a substantial improvement compared to the previous analysis (see Section 4.2.5), despite the increase in the number of input data for the training. In fact, in the analysis of this last section, the DICE was 0.72 ± 0.23 for the best deep-architecture configuration (ensemble) compared to 0.69 ± 0.24 we found previously for the same configuration (see **Table 4.6** and **Table 4.11**). All these values refer to the performance without applying any post-processing, except the exclusion of the cases where the automatic contours did not match the manual ones at all ($DICE = 0$). Also the percentage of contours with $DICE = 0$ was similar in the two automatic segmentation analyses for the ensemble configuration (8% in the first analysis and 9% in the second one), while it slightly decreased for the 3D architecture (8% for the first analysis and 4% in the second one). The 2D configuration gave once again the worst results, in terms of mean DICE value and of number of cases with $DICE = 0$.

After the application of the post-processing and the exclusion of the patients with very bad contours ($DICE < 0.3$), the metrics increased to 0.78 ± 0.12 (**Table 4.12**). Further investigations are required in this field in order to refine the results. For example, a GUI may be developed to allow a physician to identify the connected component of interest, as in the previous analysis, to select the best output among the three configurations (2D, 3D and ensemble) or to neglect a contour which is considered not acceptable.

An increase of the ICC was instead observed in this second training, where the 17% of the features had a $ICC \geq 0.85$ compared to the 1% of the previous analysis. However, the agreement between the features from the manual and the automatic contours is still poor.

Although in this second training we used full resolution images (512×512) and we increased the size of the annotated dataset from 220 to 638 images, we introduced a completely independent dataset (Dataset C), which increased the heterogeneity of the cases used to train the network. This was useful to reduce the risk of overfitting, but it may have impacted negatively on the performance of the network compared to the first segmentation analysis. In the first analysis, in fact, we used CT images acquired at the same institute (IEO) using a similar acquisition protocol both for the training and the validation. Even if the scanner and the acquisition/reconstruction protocols were not exactly the same between training and validation (see **Table 4.3** and **Table 4.4**), we expect that the image quality was more uniform compare to a

dataset from a different institute. For example, not all the CT images from the Dataset C were acquired after the injection of contrast medium, as our datasets, and this may confuse the network during the training. In order to understand whether a dataset so heterogeneous in the image contrast among the anatomical structures can be a possible source of impairment of the network performance, in the future a new training with the exclusion of images without contrast medium should be performed. Another approach to increase the segmentation performance may be the implementation of a different architecture. This aspect has already been investigated in the literature. It is possible that the architecture we used is not able to learn new features, essential for the increase of its performance. For example, the U-net architecture can be maintained but changing some elements of its structures (i.e the size and the number of the convolutional filters, the number of layers and the loss function) or using one of its variants [132,234]. Otherwise, a different architectures from the U-Net may be investigated [248–251]. Another possibility is the combination of images from hybrid acquisition modalities, such as CT/PET, to exploit both the anatomic information from CT images and functional one from the PET images. The two images are inherently co-registered and therefore have the same ground truth. Zhao et al. [252] proposed a deep-network for lung lesion segmentation, which used the information from the PET and the CT images, extracted separately with a V-Net and then fused to combine the information. The combination of these two modalities can be useful for the detection of the pathological lesions, excluding thus the areas that may be wrongly included using only the CT images for the similarity of their texture with the tumours. Moreover, information from the PET and CT images can be combined to improve the performance of predictive models. The main limitation of this approach is that the availability of CT/PET images is lower than radiological CT ones, hindering the creation of dataset with an adequate number of input images for the training.

The fully automatic segmentation is essential to reduce the human intervention and therefore the physician's time and efforts. Moreover, once the architecture is trained and validated, the automatic segmentation become an operator-independent tool and thus more reproducible than manual contours. The delineation of lesions and organs is essential not only for radiomic purposes, but also in the clinical practice, such as in radiotherapy treatment planning, computed-assisted surgery, treatment response evaluation and dose measurements in therapeutic nuclear medicine. However, the algorithms used in this field are at the moment not performing optimally and the supervision by an expert person is still necessary. In our study, for example, the 10% of the patients (28 out of 270) was lost for the poor quality of the segmentation network. In a dedicated radiomic study they can be retrieved back after the corrections of the automatic contours done by the physician.

As concerns the OS prediction, even if a significant difference between the two segmentation types was not found, the results were not completely iden-

tical between them. We obtained, in fact, a different radiomic score when the training was performed with the automatic segmentation instead of the manual one, suggesting that there was a difference in the feature values which impacted on what the model was capturing (see **Figure 4.8**). Moreover, a lower prediction performance was obtained in the validation set with automatic segmentation, which may be associated to a loss of information in the automatic contours compared to the manual ones. In conclusion, we therefore recommend not to mix manual and automatic contours in the same model development until the automatic algorithm is properly improved. However, the automatic and manual models are performing statistically in the same way in terms of OS (p-value > 0.05 between manual and automatic contours). Further studies with an increased number of patients are required to improve these results.

Limitations and future developments

The main limitation of this analysis was the limited number of patients used both for the training of the segmentation network and for the development of the survival model. Secondly, multiple operators got involved in the manual segmentation phase and each of them contoured a different group of CT images. This may introduce a variability in how the tumours were delineated, particularly for the Dataset C where a consensus among operators was not possible to establish. This may have impacted also on what the automatic network learned from the labels during the training part.

Furthermore, in future analysis the survival model needs to be validated on an independent dataset. In this study we aimed at comparing the different segmentation techniques in terms of the overall performance and not to build a definitive model for the prediction of the survival. For this reason, we evaluated the prediction performance only using an internal validation procedure, splitting the dataset in train and validation sets, and other machine learning models different from the LASSO algorithm were not considered. We observed that the model built from the manual contours and the model built from the automatic ones were not significantly different for the prediction of the OS. In contrast to our result, Huang et al. [131] found that the contours performed by three radiologists using a semi-automated tool had a significant impact on the prediction performance. With respect to our study, they analysed a different outcome, which was the prediction of the EGFR mutational status in 46 patients affected by NSCLC. In order to evaluate in more depth the impact of the different contours on the performance of predictive models, other outcomes should be considered in the future.

Finally, the dataset used for the survival prediction (Dataset A) was characterised by a variability in the scanner, acquisition protocols and in the reconstruction algorithms that may be confounding factors for the feature stability, as we deeply discussed in the previous chapters. A selection of the radiomic variables was performed using the LASSO method, which rejects those variables that are introducing noise in the model itself. However, we are going to

perform further analyses, considering only the robust features for the model development and comparing this result to when all features are used.

Conclusions and future perspectives

In this Ph.D. project the robustness of the radiomic features in the field of CT imaging was investigated. The anatomical region of interest of this study was the lung district, focusing in particular on the oncological disease. Among the different factors which potentially affect the robustness of the radiomic features, the role of the CT reconstruction algorithm, of the tube voltage, of the scanner model, of the segmentation methods and of processing techniques (bin width and filters) was investigated. To achieve this goal, along with the CT images of patients affected by lung tumour, we analysed the CT images of a dedicated phantom, we fabricated specifically for CT radiomic analysis.

The phantoms used in the literature are often not suitable for radiomic purposes: either their texture is homogeneous or, whenever they have heterogeneous inserts, the latter are not adequate to reproduce the tumour heterogeneity and the variability observed in a realistic dataset of patients. For this reason, we developed a phantom, named Heterogeneous Lung Lesion Phantom (HeLLePhant), with ad-hoc inserts to reproduce the CT signal and the texture of real contrast-enhanced lung lesions. The novelty of the proposed phantom lies in the two techniques adopted for the fabrication of the inserts, which allowed us to create textures with a high level of heterogeneity. More in detail, we fabricated 11 heterogeneous inserts: 2 3D-printed using a PET-G filament, and 9 made of sodium polyacrylate mixed with a contrast medium diluted with water. These techniques enable the creation of multiple inserts with different heterogeneous textures and mean CT signals.

The main findings of this research project can be summarised in the following points.

- We developed two new methods to fabricate heterogeneous inserts which simulate the lung lesions in CT imaging. They are well suited to study efficiently the repeatability and reproducibility (varying tube voltage, CT scanner and reconstruction algorithm) of the radiomic features.

- Most of the radiomic features has a good repeatability in CT imaging. The analysis of repeatability was performed only with the phantom due to radiation exposure issues. Between 74% and 100% of the original

features (shape features excluded) were found to be repeatable (CV $\leq$ 0.10), depending on the analysed configuration (scanner, voltage, bin width and insert). Excellent repeatability was found particularly for the PET-G inserts, for which the CV was lower than 0.10 for more than the 90% of the features in all the studied cases.

- Feature behaviour is texture dependent, confirming what other studies in the literature have observed. Therefore, in order to generalise the phantom results to a clinical investigation, it is mandatory to reproduce adequately the tissue texture for each investigated pathology and encompass the heterogeneity of a clinical population. The insert fabrication techniques we proposed constitute the first important steps towards these goals.

- Among the investigated parameters (CT reconstruction algorithm, CT scanner, tube voltage and segmentation), the reconstruction algorithm and the segmentation impact the features the most, considering the CT images of patients affected by NSCLC. The larger impact of the reconstruction algorithm was observed also in phantom.

- When the CT images are acquired with different but similar scanners (same vendor, same reconstruction algorithm) and similar acquisition protocols, or when the range of tube voltage used is small (100-120 kVp), the texture is only weakly influenced by these parameters.

- The nnU-Net, a self-adapting framework developed by Isensee [239], was trained and validated on proprietary datasets of CT images of patients with lung tumours. Good results in terms of geometrical evaluation metrics (DICE and HD) were found when compared to the manual ground truth.

- A difference between the various segmentation techniques was observed in terms of ICC results. A poor reproducibility (ICC < 0.5) was in fact found for a large number of *original* features (namely extracted from unfiltered images) in the comparison between manual and automatic contours (22% for the semi-automatic technique, and between 37% and 63% for the deep-based one considering various configurations). However, the statistical analysis did not detect a significant difference in the prediction performance of the overall survival in patients with non-advanced NSCLC between the deep-based segmentation and the manual one in patients affected by NSCLC.

- Neither of the two bin widths we investigated (25 HU and 5 HU) appeared to be more robust, both in patients and in phantom. However, a common choice for this parameter is encouraged in order to generalise and compare the results of different studies.

- The filters impact the feature reproducibility. The wavelet-based features, especially from the HH sub-band, are more prone to variability. LoG features are, instead, more stable, except for a low value of sigma (0.5 mm). These results were found with the CT images of patients. A validation using the phantom images is required for the next steps of the analysis.

- The textural features — in particular glrlm, glszm and gldm — are in general the least robust features. This behviour was found using the CT images of patients and of the phantom.

Future developments

The methodological studies discussed in this thesis provide subsets of robust features. We expect that restricting the input features in predictive models to the robust ones we proposed will improve their performance. Investigating this aspect is the natural next step of this research project. For this purpose, we are working on a dataset of 287 patients with advanced lung adenocarcinoma (which includes the patients of the Dataset B used in the Chapter 4). The general aim of the study is to find an association between the radiomic features and the genetic mutations extracted from a tissue sample for each patient, and between the radiomic features and the overall survival. An additional prospective dataset of patients is currently being collected to validate the results of the model.

Another interesting observation in our study was that different contouring techniques (automatic versus manual) do not impact significantly the prediction of the overall survival in patients affected by NSCLC. Nevertheless, a more accurate validation of this algorithm is required, and other clinical outcomes should be tested to confirm our findings. The ongoing study of the association between radiomic features and genetic mutations mentioned above will offer further insight into this aspect.

Finally, the prototype of phantom we proposed should be improved with a view to multi-centres studies. Our prototype for the PET-G inserts requires further refinement to improve its similarity with the tumours, by increasing the mean CT signal, reducing the heterogeneity, and creating more complex shapes. On the other hand, the PET-G inserts are more suitable for the replication of the experimental settings in different institutes. The PET-G inserts are in fact created starting from a digital 3D-printing model. Sharing the model among the various research groups may enable the replication of the inserts, thus fostering a collaborative effort to harmonise the radiomic procedures and analyses.

Appendix A

Feature extraction: the Pyradiomics package

Pyradiomics [73] is an open-source package for the feature extraction, based on Python language. Pyradiomics software follows in general the IBSI recommendations, and any deviation from the IBSI is reported in the online documentation (https://pyradiomics.readthedocs.io/en/latest/#).

The format of the input image and mask can be chosen among those readable by ITK package, such as NRRD, NIfTI, MHA, MHD, HDR, TIF and PNG. Unfortunately, volumetric images and structures in DICOM format cannot be used as they are but need to be converted to a suitable format. The software 3D slicer for image visualisation and processing is able to read the images in DICOM and the contours in RT Structure format and convert them in one of the formats listed above. A faster alternative is the Plastimatch software (http://plastimatch.org), using the command *convert*. Both these tools are open source.

The following categories of features can be computed: *First-Order Statistics*, *Shape* (3D or 2D), *Gray Level Cooccurence Matrix* (glcm), *Gray Level Run Length Matrix* (glrlm), *Gray Level Size Zone Matrix* (glszm), *Neigbouring Gray Tone Difference Matrix* (ngtdm) and *Gray Level Dependence Matrix* (gldm).

Furthermore, it is possible to extract the features after the application of the subsequent filter: Laplacian of Gaussian (LoG), Wavelet (based on the PyWavelets package), Square, Square Root, Logarithm, Exponential, Gradient, Local Binary Pattern (2D) and Local Binary Pattern (3D).

All the details about the features and the filters, including the source code, can be found in the online documentation.

The first step for the feature extraction is the customisation of the **parameter file**. This file contains all the information about the extraction and the parameters which need to be set, and it has to be a YAML or JSON file. It consists of three main parts:

1. **imageType**
   It contains the list of all the filters that have to be applied to the image before the extraction. Each of the filter name has to be put on a different line and has to end with a colon, followed by the parameter settings, if required. If the features have to be extracted from the images without

applying filter, the image type is indicated by the term "Original". If the parameters of the filter are not specified, the extraction is performed using the default values. For example, if the features have to be extracted from unfiltered images and from LoG filtered images with different value of σ, the following lines have to be added:

```
imageType:
    Original: {}
    LoG: {'sigma': [0.5, 1.0]}
```

2. **featureClass**
   In this part all the categories of features that need to be computed have to be listed on separate lines. All the available features for each category are extracted by default. If only a subgroup of features has to be extracted, their name has to be reported after the category name.

3. **setting**
   This part includes the set of the parameters which need to be customised for the image processing, such as normalisation, resampling and discretisation. The following example shows how to configure the voxel resampling (new size and interpolation function), the bin width for the discretisation of the grey-level intensities, the distance parameter for the glcm e ngtdm categories (the distance is equal to 1 in this example), and the type of extraction (2.5D):

```
setting:
    resampledPixelSpacing: [0.78, 0.78, 0]
    interpolator: 'sitkBSpline'
    binWidth: 25
    distances: [1]
    force2D: true
```

In the *iteractive use* of Pyradiomics — the modality used for the extraction in this thesis — four commands are required to extract the features:

- Import of the library and packages:

```
import radiomics
from radiomics import featureextractor,
   getFeatureClasses
import SimpleITK as sitk
```

- Read the image and mask files in:

```
mask = sitk.ReadImage(os.path.join(directory_mask,
    'namemask.nrrd'), sitk.sitkFloat32)
image = sitk.ReadImage(os.path.join(directory_image,
    'nameimage.nrrd'), sitk.sitkFloat32)
```

- Read the parameter file in:

```
paramsFile =
    os.path.abspath(os.path.join(directory_paramsFile,
    'nameparamsFile.yaml'))
```

- Instantiate the feature extractor class using the information stored in paramsFile and calculate the features:

```
extractor =
    featureextractor.RadiomicsFeatureExtractor(paramsFile)
result = extractor.execute(image, mask)
```

The variable *result* is a dictionary which includes the information about the version of the Python packages, the parameters set in the parameter file, geometric information about the images and the mask, and the extracted features.

The list of the parameters set during the features extraction is shown in **Table A.1**. Features are extracted in the 2.5D modality. To activate this type of feature extraction, the *force2D* parameter was set to TRUE (see Section 1.2, *Feature categories*). The voxels were resampled with the *sitkBSpline* interpolator in the axial plane to 0.78×0.78 mm^2, which was the average value among the 103 images included in the study of Chapter 2 (range 0.52 – 1.38 mm). Both the glcm and the ngtdm matrices were calculated using three different offsets (1, 4 and 7) between the reference and the neighbour voxel. We used the default values for all the remaining parameters which can be customised in PyRadiomics. The only exception is the *voxelArrayShift* parameter, which we fixed to 1000 HU.

Parameter name	**Parameter value**
imageType	• Original: {} • Wavelet: wavelet: coif1 • LoG: sigma: [0.5, 1.0, 1.5, 2.5, 5.0]
featureClass	shape, firstorder, glrlm, glszm, gldm, glcm, ngtdm
resampledPixelSpacing	[0.78, 0.78, 0]
interpolator	'sitkBSpline'
binWidth	25 (5)
distances	[1], [4], [7]
force2D	true
force2Ddimension	0
voxelArrayShift	1000

Table A.1: Parameters set in Pyradiomics for the radiomic feature extraction, for the unfiltered and the wavelet- and LoG-filtered images.

Appendix B

Feature intrinsic dependence on number of voxels

As observed by Shafiq-Ul-Hassan et al. [103] and Fave et al. [102], some radiomic features are intrinsically correlated with the number of voxels inside the VOI. The reason for this correlation, in fact, can be easily identified in the mathematical formula by which these features are defined. Therefore, a proper correction was applied *a posteriori* in order to reduce this correlation.

First of all, all the features (*shape* features excluded) were tested for correlation with the number of voxels, extracted from the original, wavelet-filtered and LoG-filtered images. To do this, the features extracted from the CT images of the 103 NSCLC patients analysed in Chapter 2 were used, taking only the series acquired and reconstructed for clinical purpose.
The correlation between each feature and the feature *interpolated_VoxelNum* was verified by calculating the Spearman's rank correlation coefficient (ρ_S) with the *cor* function in *stats* package in R (v. 3.6.2) [173]. The features with a correlation coefficient $|\rho_S| > 0.85$ are listed in **Table B.1**, as long with the coefficient of correlation for the original images. Since this correlation comes from the feature definition and not from the properties of the texture, the results are quite similar among the three types of images investigated.

For the *glrlm_GrayLevelNonUniformity*, the *glrlm_RunLengthNonUniformity*, the *glszm _GrayLevelNonUniformity* and the *gldm_DependenceNonUniformity* features a normalised version is already implemented in Pyradiomics (indicated as *Normalized*). Since a dependence on the number of voxels was not observed for this version, these four features were excluded and only the already normalised definition was taken into consideration.

The *Busyness* feature from the *ngtdm* category was divided by the number of voxels too. However, the coefficient of correlation was not considerably reduced after the correction ($\rho_S = -0.64$). In fact, the correction proposed in the literature [102] requires a modification of the formula before the extraction. Since the relationship between the *Busyness* feature and the number of voxels, this feature was excluded.

The *TotalEnergy* feature from the *firstorder* category, the *Coarseness* feature from the *ngtdm* category and the *GrayLevelNonUniformity* feature from the *gldm* category, instead, were a posteriori corrected, as indicated in **Table B.2**: the *TotalEnergy* and the *GrayLevelNonUniformity* were divided [102], while the *Coarseness* was multiplied [103] by the number of voxels. Since no reference was

found in the literature, the *GrayLevelNonUniformity* in the *gldm* category was adjusted applying the same correction indicated for the corresponding feature in the *glrlm* category. For all these three features a reduction of the coefficient of correlation was observed (**Table B.2**).

In conclusion, the features *firstorder_TotalEnergy*, *ngtdm_Coarseness* and *gldm_GrayLevelNonUniformity* were properly corrected, while the features *glrlm_GrayLevelNonUniformity*, *glrlm_RunLengthNonUniformity*, *glszm_GrayLevelNonUniformity*, *gldm_DependenceNonUniformity* and *ngtdm_Busyness* were not included in the anlyses. The list of all the features included in the studies presented in this work is reported in **Table B.3**. The suffix "modified" was added to the features that were corrected *a posteriori*. This list of feature names is the same for the "Original", "Wavelet" and "LoG" features.

Feature	ρ_S
original_firstorder_TotalEnergy	0.99
original_glrlm_RunLengthNonUniformity	0.91
original_glszm_GrayLevelNonUniformity	0.96
original_gldm_DependenceNonUniformity	0.97
original_gldm_GrayLevelNonUniformity	0.92
original_ngtdm_1_Busyness	0.92
original_ngtdm_1_Coarseness	-0.96
original_ngtdm_4_Busyness	0.89
original_ngtdm_4_Coarseness	-0.96
original_ngtdm_7_Busyness	0.87
original_ngtdm_7_Coarseness	-0.94

Table B.1: List of features extracted from the original images which are correlated with the number of voxels, along with the Spearman's rank correlation coefficient.

Feature	**Correction**	ρ_S
original_firstorder_TotalEnergy	/	0.41
original_gldm_GrayLevelNonUniformity	/	0.39
original_ngtdm_1_Coarseness	×	-0.02
original_ngtdm_4_Coarseness	×	-0.43
original_ngtdm_7_Coarseness	×	-0.44

Table B.2: Proposed corrections for the original features correlated with the number of voxels that are not removed for the following analysis. The Spearman's rank correlation coefficient after the correction is reported too.

Category	Identification number	Features
shape	1	Elongation
	2	Flatness
	3	LeastAxisLength
	4	MajorAxisLength
	5	Maximum2DDiameterColumn
	6	Maximum2DDiameterRow
	7	Maximum2DDiameterSlice
	8	Maximum3DDiameter
	9	MeshVolume
	10	MinorAxisLength
	11	Sphericity
	12	SurfaceArea
	13	SurfaceVolumeRatio
firstorder	14	10Percentile
	15	90Percentile
	16	Entropy
	17	InterquartileRange
	18	Kurtosis
	19	Maximum
	20	Mean
	21	MeanAbsoluteDeviation
	22	Median
	23	Minimum
	24	Range
	25	RobustMeanAbsoluteDeviation
	26	RootMeanSquared
	27	Skewness
	28	TotalEnergy
	29	Uniformity
	30	Variance
	31	1__Autocorrelation
	32	1__ClusterProminence
	33	1__ClusterShade
	34	1__ClusterTendency
	35	1__Contrast
	36	1__Correlation
	37	1__DifferenceAverage
	38	1__DifferenceEntropy
	39	1__DifferenceVariance
	40	1__Id
	41	1__Idm
	42	1__Idmn
	43	1__Idn
	44	1__Imc1
	45	1__Imc2
	46	1__InverseVariance
	47	1__JointEnergy
	48	1__JointEntropy
	49	1__MaximumProbability
	50	1__MCC

	51	1_SumAverage
	52	1_SumEntropy
	53	1_SumSquares
	54	4_Autocorrelation
	55	4_ClusterProminence
	56	4_ClusterShade
	57	4_ClusterTendency
	58	4_Contrast
	59	4_Correlation
	60	4_DifferenceAverage
	61	4_DifferenceEntropy
glcm	62	4_DifferenceVariance
	63	4_Id
	64	4_Idm
	65	4_Idmn
	66	4_Idn
	67	4_Imc1
	68	4_Imc2
	69	4_InverseVariance
	70	4_JointEnergy
	71	4_JointEntropy
	72	4_MaximumProbability
	73	4_MCC
	74	4_SumAverage
	75	4_SumEntropy
	76	4_SumSquares
	77	7_Autocorrelation
	78	7_ClusterProminence
	79	7_ClusterShade
	80	7_ClusterTendency
	81	7_Contrast
	82	7_Correlation
	83	7_DifferenceAverage
	84	7_DifferenceEntropy
	85	7_DifferenceVariance
	86	7_Id
	87	7_Idm
	88	7_Idmn
	89	7_Idn
	90	7_Imc1
	91	7_Imc2
	92	7_InverseVariance
	93	7_JointEnergy
	94	7_JointEntropy
	95	7_MaximumProbability
	96	7_MCC
	97	7_SumAverage
	98	7_SumEntropy
	99	7_SumSquares
	100	DependenceEntropy
	101	DependenceNonUniformityNormalized
	102	DependenceVariance
	103	GrayLevelNonUniformity

	104	GrayLevelVariance
	105	HighGrayLevelEmphasis
	106	LargeDependenceEmphasis
gldm	107	LargeDependenceHighGrayLevelEmphasis
	108	LargeDependenceLowGrayLevelEmphasis
	109	LowGrayLevelEmphasis
	110	SmallDependenceEmphasis
	111	SmallDependenceHighGrayLevelEmphasis
	112	SmallDependenceLowGrayLevelEmphasis
	113	GrayLevelNonUniformityNormalized
	114	GrayLevelVariance
	115	HighGrayLevelRunEmphasis
	116	LongRunEmphasis
	117	LongRunHighGrayLevelEmphasis
	118	LongRunLowGrayLevelEmphasis
glrlm	119	LowGrayLevelRunEmphasis
	120	RunEntropy
	121	RunLengthNonUniformityNormalized
	122	RunPercentage
	123	RunVariance
	124	ShortRunEmphasis
	125	ShortRunHighGrayLevelEmphasis
	126	ShortRunLowGrayLevelEmphasis
	127	GrayLevelNonUniformityNormalized
	128	GrayLevelVariance
	129	HighGrayLevelZoneEmphasis
	130	LargeAreaEmphasis
	131	LargeAreaHighGrayLevelEmphasis
	132	LargeAreaLowGrayLevelEmphasis
glszm	133	LowGrayLevelZoneEmphasis
	134	SizeZoneNonUniformity
	135	SizeZoneNonUniformityNormalized
	136	SmallAreaEmphasis
	137	SmallAreaHighGrayLevelEmphasis
	138	SmallAreaLowGrayLevelEmphasis
	139	ZoneEntropy
	140	ZonePercentage
	141	ZoneVariance
	142	1_Coarseness
	143	1_Complexity
	144	1_Contrast
	145	1_Strength
	146	4_Coarseness
ngtdm	147	4_Complexity
	148	4_Contrast
	149	4_Strength
	150	7_Coarseness
	151	7_Complexity
	152	7_Contrast
	153	7_Strength

Table B.3: List of Pyradiomics features used in all the studies presented in this work. The prefixes 1_, 4_ and 7_ refer to the offset set for the feature extraction of the glcm and ngtdm categories.

Note that the features belonging to the *shape* category were included only in the group of features extracted from the original images. The filters we applied, in fact, do not affect the borders of the segmentation but only its texture. The *shape* features extracted from the filtered images are therefore exactly the same as those extracted from the unfiltered ones.

Appendix C

CT number of the investigated materials

The following tables include all the objects scanned and analysed for the identification of the materials suitable for the replication of lung lesions. **Table C.1** lists the food, **Table C.2** the household items and **Table C.3** the polymeric materials. In particular, materials in **Table C.3** were specifically produced by the C.I.Ma.I.Na group from the University of Milano using a subtractive, additive or mould-based technique.

More in details, in the subtractive approach the final product is obtained from a raw object machined by removing the excess material. Conversely, in the additive technique the object is fabricated by adding the material gradually, converting the digital model into its physical representation [253]. In this study the objects were produced with two types of additive techniques: the fused filament fabrication (FFF) and the stereolithography. In brief, stereolithography is based on the polymerisation of a liquid resin. The polymerisation process is triggered by an ultraviolet laser beam, which hits the resin layer by layer, following the indications of a virtual model. Thanks to this chemical reaction, only the point of the resin selectively hit by the UV beam hardens and, after a cleaning process, the final 3D object is obtained. In the FFF, instead, the 3D object is produced by the extrusion of a heated filament trough a small nozzle upon the printing platform. The movable extrusion head — to which the nozzle is attached — moves in the plane transversal to the direction of the extrusion, following computer-controlled instructions. The material hardens very rapidly by cooling, making it possible to print continuously the object layer by layer in the longitudinal direction until the object is complete. The typical materials used for this kind of manufacturing are of the thermoplastic kind, such as acrylonitrile butadiene styrene (ABS), polylactic acid (PLS), thermoplastic polyurethane (TPU), and glycol modified-polyethylene terephthalate (PET-G). During the model design, some parameters can be set to control the fabrication process, such as the nozzle temperature, the height of the layers, the speed of the extrusion head, and the infill percentage.

Material	**Min** (HU)	**Max** (HU)	**Mean** (HU)	**Median** (HU)	**SD** (HU)
coffee beans	-983	-319	-635	-622	117
coffee grounds	-936	-335	-554	-574	89
chickpeas	-901	326	-84	-54	273
beans	-939	402	-114	-81	286
lentils	-596	115	-132	-126	88
rice	-411	76	-116	-116	62
spelt	-712	-59	-331	-328	93
pistacchio nuts in pieces	-739	4	-336	-336	97
peanuts in pieces	-703	46	-316	-323	102
walnuts in pieces	-643	-39	-289	-291	83
cornmeal	-266	-127	-199	-199	17
pepper	-468	-278	-358	-359	30
oregano	-881	-707	-781	-782	29

Table C.1: List of the food materials analysed along with the basic properties of the CT signal calculated inside the VOI.

Material	**Min** (HU)	**Max** (HU)	**Mean** (HU)	**Median** (HU)	**SD** (HU)
floral foam	-974	-937	-961	-961	3
sea sponge	-1011	-910	-974	-974	18
car wash sponge	-1005	-957	-976	-976	4
block of compressed cork	-863	-362	-734	-735	30
powdered cork	-900	-769	-835	-836	14
cork stopper	-753	-576	-705	-709	18
polystyrene sphere	-994	-935	-984	-984	3
eraser	594	979	869	872	48

Table C.2: List of common synthetic objects analysed along with the basic properties of the CT signal calculated inside the VOI.

Material	**Manufacturing technique**	**Min** (HU)	**Max** (HU)	**Mean** (HU)	**Median** (HU)	**SD** (HU)
PMMA	subtractive	92	133	116	116	6
HMWPE	subtractive	-83	-53	-66	-65	5
glass-filled nylon	subtractive	70	98	85	86	6
PVC 1	subtractive	984	1193	1081	1073	44
PVC3	subtractive	-67	-36	-48	-48	5
PVC4	subtractive	873	1115	1023	1022	41
ABS natural 250°C 100% infill	additive (filament)	-97	-41	-69	-70	9
ABS natural 250°C 25% infill	additive (filament)	-837	-495	-741	-748	48
ABS Clear 250°C 100% infill	additive (filament)	-161	3	-52	-45	29
ABS Clear 250°C 25% infill	additive (filament)	-813	-498	-724	-727	48
PET Clear 230°C 100% infill	additive (filament)	52	175	149	154	18
PET Clear 230°C 25% infill	additive (filament)	-831	-410	-667	-678	68
TPU Clear 230°C 100% infill	additive (filament)	-27	106	33	33	25
TPU Clear 230°C 25% infill	additive (filament)	-812	-436	-680	-725	100
Form2 STD GREY 100% infill	additive (stereo-lithography)	82	140	122	123	9
Form2 Tough 100% infill	additive (stereo-lithography)	75	159	132	131	11
polyurethane foam	mould	-982	-859	-919	-917	18

silicone rubber foam	mould	-921	-715	-798	-798	18
epoxy resin foam	mould	-433	-70	-272	-271	74
silicone rubber	mould	273	315	301	302	8
ionogel poli(HEMA-co-AN)	mould	171	203	190	189	5
ionogel poli(HEMA-co-AN)+post-processing with laser at 405 nm	mould	180	204	190	190	5

Table C.3: List of polymeric objects manufactured by the C.I.Ma.I.Na group using one of the technique shown in the table. The information about the CT signal calculated inside the VOI is reported too.

The types of food we investigated had a very low signal, usually much lower than water. Instead, among the polymeric objects, two similar materials matched the desired range: the TPU with infill equal to 100% (mean CT number equal to 33 ± 25 HU) and the PET-G with infill equal to 100% (mean CT number equal to 149 ± 18 HU). Both these materials were fabricated using the FFF technique.

An additional material, deeply investigated in this thesis, is the sodium polyacrylate. It is a non-toxic polymer which appears as a white powder in its dry form. When a liquid is added to the powder, the sodium polyacrylate absorbs it, increases its volume and takes a gel-like consistency. For example, thanks to this property, this material is a key component of diapers. An interesting study making use of the volume expansion of the sodium polyacrylate was performed by Levine et al. [254]. They compared the volume-based measurement with the approach based on the Response Evaluation Criteria in Solid Tumors (RECIST)[1] by scanning repeatedly the inserts with sodium polyacrylate after introducing water in the sample at each scan. In this way they simulated the tumour expansion, with the advantage of knowing the total mass of the target. They showed that using the volume measure better predicts the mass of the insert than the RECIST length.

In the study discussed in Chapter 3, we produced various inserts made of sodium polyacrylate in powder mixed with water or diluted contrast medium. In summary, the fabrication procedure consisted in the placement of the powder in the container using a small spoon, and then of the liquid using a syringe.

[1]RECIST provides guidelines to assess tumour growth over time and thus evaluate the effectiveness of a treatment [255]. The evaluation of changes in tumour size is done by measuring the unidimensional longest diameter.

We determined empirically the amount of powder and liquid solution to be used in order to obtain a homogeneous texture in 0.43 g of sodium polyacrylate and 12 mL of water at room temperature, respectively. We observed that it is better to add all the water at once, since pouring the liquid little by little made the gelling non-continuous, with the consequent formation of less dense areas.

Appendix D

Reconstruction algorithms reproducibility: patients vs phantom

In the following sections the results of the feature reproducibility due to the reconstruction algorithm from Chapter 2 and Chapter 3 are compared. The comparison is done considering the same metrics (PV, percentual of variation) and two different metrics (the two metrics used in the corresponding chapter, CCC and PV).

D.1 Comparison using PV

The PV for the 103 patients of Chapter 2 was calculated for each feature for all the patients between the FBP and the reference IR reconstruction using eq. 3.6. For each feature, the median value was then taken among the patients from the four populations separately. These results were compared visually with the PV value calculated in phantoms, taking the median values among the nine polyacrylate inserts. The plots were created for each of the four protocol (Discovey CT750 HD-120 kVp, Optima CT660-120 kVp, Discovey CT750 HD-100 kVp, Optima CT660-100 kVp) and are shown in **Figure D.1**, where each point is a single feature (original type, 140 total features).

These plots show that the majority of the features had a similar behaviour in the two cases, with in general a worst reproducibility in phantom. This discrepancy may be due to a less numerosity of "subjects " in phantom compared to patients or to a difference in the texture.

D.2 Comparison using CCC and PV

The reproducibility in patients and in phantom was also compared using the metrics adopted in Chapter 2 and Chapter 3, which were the CCC for patients and PV for the phantom. The CCC among the 103 patients was calculated between the FBP and the IR60 reconstructions (eq. 2.1). For the phantom analysis, instead, the median PV among the polyacrylate inserts and the four acquisition protocols was considered. **Figure D.2** plots the value of the PV in

phantom and the CCC in patients for each original feature.

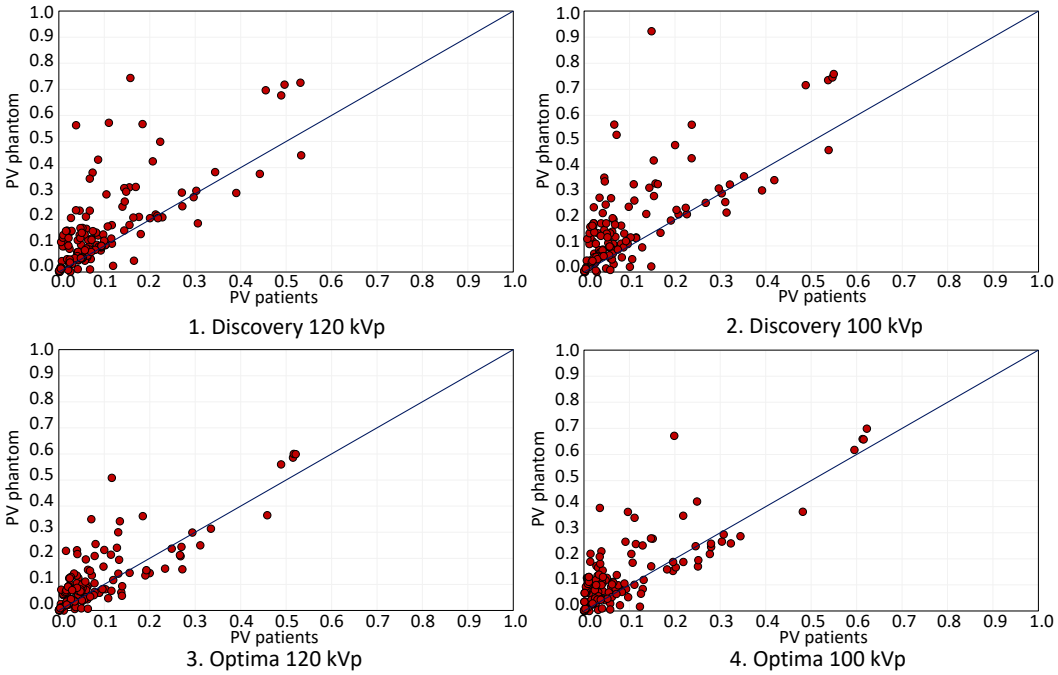


Figure D.1: Comparison of the PV calculated between FBP and the reference IR for the four acquisition protocols. Each point represents an *original* feature.

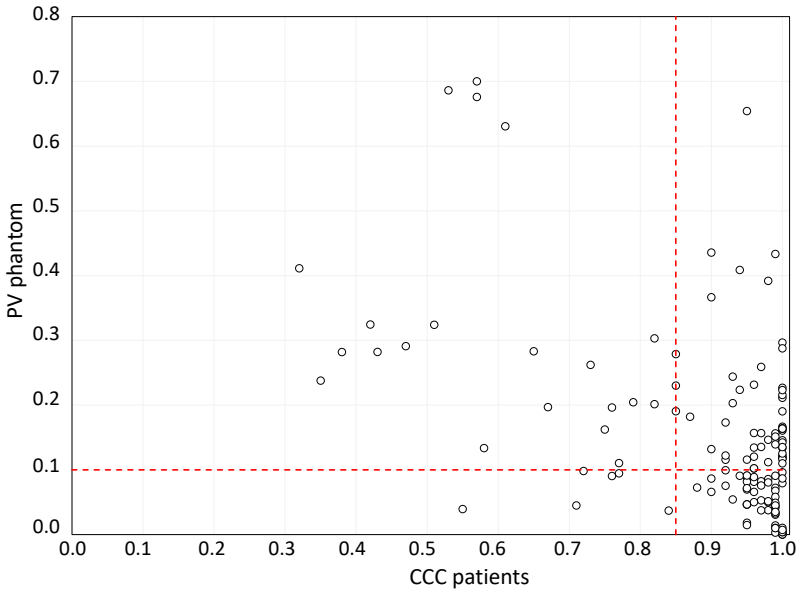


Figure D.2: Comparison of the PV calculated in phantom and the CCC in patients between FBP and the IR. Each red point represents an *original* feature. The two considered thresholds, 0.85 for CCC and 0.10 for PV, are highlighted in red.

The 42% of the features resulted reproducible in both the studies ($CCC \geq 0.85$ and $PV \leq 0.10$), while the 15% were not stable between the two recon-

struction algorithms in both the studies ($CCC < 0.85$ and $PV > 0.10$). The remaining 43% of the features, instead, was not in agreement between the two studies. When 0.20 was chosen as threshold for the phantom analysis, the agreement increased to 78%.

In **Table D.1** we reported the list of features based on the value of the PV and CCC in the two analyses.

CCC≥0.85 PV≤0.10	**CCC≥0.85 0.10<PV≤0.20**	**CCC<0.85 PV>0.10**	**CCC<0.85 PV≤0.10**	**CCC≥0.85 PV>0.20**
firstorder_90 Percentile	firstorder_Interquartile Range	glrlm_LongRun Emphasis	glrlm_ RunEntropy	firstorder_10 Percentile
firstorder_Entropy	firstorder_Kurtosis	glrlm_LongRunHigh GrayLevelEmphasis	glrlm_ShortRun Emphasis	glszm_SizeZone NonUniformity
firstorder_Maximum	firstorder_RobustMean AbsoluteDeviation	glrlm_LongRunLow GrayLevelEmphasis	glszm_Zone Entropy	glszm_Zone Percentage
firstorder_Mean AbsoluteDeviation	firstorder_Skewness	glrlm_RunLength NonUniformity Normalized	glcm1_Id	gldm_Large DependenceLow GrayLevelEmphasis
firstorder_Mean	firstorder_Variance	glrlm_RunPercentage	glcm1_Imc2	gldm_Small Dependence Emphasis
firstorder_Median	glrlm_GrayLevel Variance	glrlm_RunVariance	glcm1_Inverse Variance	gldm_Small DependenceLow GrayLevelEmphasis
firstorder_Minimum	glrlm_ShortRunLow GrayLevelEmphasis	glszm_LargeArea Emphasis		glcm1_Contrast
firstorder_Range	glszm_SizeZoneNon UniformityNormalized	glszm_LargeAreaHigh GrayLevelEmphasis		glcm1_Difference Average
firstorder_ RootMeanSquared	gldm_GrayLevel Variance	glszm_LargeAreaLow GrayLevelEmphasis		glcm1_Difference Variance
firstorder_TotalEnergy	gldm_LowGray LevelEmphasis	glszm_ZoneVariance		ngtdm1_Complexity
firstorder_Uniformity	glcm1_Cluster Prominence	gldm_Dependence NonUniformity Normalized		ngtdm1_Contrast
glrlm_GrayLevelNon UniformityNormalized	glcm1_Cluster Tendency	gldm_Dependence Variance		ngtdm1_Strength
glrlm_HighGray LevelRunEmphasis	glcm1_Difference Entropy	gldm_Large Dependence Emphasis		glcm4_Cluster Prominence
glrlm_LowGray LevelRunEmphasis	glcm1_MCC	gldm_Large DependenceHigh GrayLevelEmphasis		glcm4_Contrast

glrlm_ShortRunHigh GrayLevelEmphasis	glcm1_SumSquares	gldm_Small DependenceHigh GrayLevelEmphasis	glcm4_Imc1
glszm_GrayLevelNon UniformityNormalized	ngtdm1_Coarseness	glcm1_Idm	ngtdm4_Contrast
glszm_GrayLevel Variance	glcm4_Cluster Tendency	glcm1_Imc1	glcm7_Cluster Prominence
glszm_HighGray LevelZoneEmphasis	glcm4_Correlation	glcm1_JointEnergy	glcm7_Imc1
glszm_LowGray LevelZoneEmphasis	glcm4_Difference Average	glcm1_Maximum Probability	ngtdm7_Contrast
glszm_SmallArea Emphasis	glcm4_Difference Variance	glcm4_Maximum Probability	
glszm_SmallAreaHigh GrayLevelEmphasis	glcm4_Imc2	glcm7_Maximum Probability	
glszm_SmallAreaLow GrayLevelEmphasis	glcm4_JointEnergy		
gldm_Dependence Entropy	glcm4_MCC		
gldm_GrayLevel NonUniformity	glcm4_SumSquares		
gldm_HighGray LevelEmphasis	ngtdm4_Strength		
glcm1_Autocorrelation	glcm7_Cluster Tendency		
glcm1_ClusterShade	glcm7_Contrast		
glcm1_Correlation	glcm7_Correlation		
glcm1_Idmn	glcm7_Difference Average		
glcm1_Idn	glcm7_Difference Variance		
glcm1_JointEntropy	glcm7_Imc2		
glcm1_SumAverage	glcm7_JointEnergy		
glcm1_SumEntropy	glcm7_MCC		
glcm4_Autocorrelation	glcm7_SumSquares		
glcm4_ClusterShade	ngtdm7_Strength		
glcm4_Difference Entropy			
glcm4_Id			

glcm4_Idm
glcm4_Idmn
glcm4_Idn
glcm4_InverseVariance
glcm4_JointEntropy
glcm4_SumAverage
glcm4_SumEntropy
ngtdm4_Coarseness
ngtdm4_Complexity
glcm7_Autocorrelation
glcm7_ClusterShade
glcm7_Difference Entropy
glcm7_Id
glcm7_Idm
glcm7_Idmn
glcm7_Idn
glcm7_InverseVariance
glcm7_JointEntropy
glcm7_SumAverage
glcm7_SumEntropy
ngtdm7_Coarseness
ngtdm7_Complexity

Table D.1: List of original features falling in the range of values defined by the threshold 0.85 for the CCC analysis in patients and 0.10 or 0.20 for the PV analysis in phantom. The two columns with the red title correspond to the cases where the features resulted reproducible (column 1) or non-reproducible (column 3) in the analyses of both Chapter 2 and Chapter 3.

Appendix E

Semi- and fully automatic segmentation

E.1 Semi-automatic segmentation: the GrowCut algorithm

Most of the semi-automatic tools for medical imaging segmentation are based for their simplicity on the active contour or the region-based approaches.

The active contour technique, also referred as *Snake*, starts from a parametric curve which is deformed during the algorithm iterations in order to match the boundary of the target object. This deformation is controlled by the minimisation of an energy function, which rules both the contour localisation inside the image and a smooth deformation following the lines and edges.

A region-based model, instead, starts from one or more voxels — named *seeds* — inside the volume of interest and "grows" through the connected voxels following a criterion of similarity among their grey-level intensities. For example, the expansion process from the starting seeds may stop when it collides with an edge. A relevant limitation of this technique is the dependence of the final result on the choice of the similarity criterion and on the input seeds. Moreover, it is possible that due to the noise of the image the criterion of similarity loses its validity, and some holes are generated inside the segmentation. In this thesis work, the *GrowCut* algorithm — based on the region growing technique — was used.

The *GrowCut* algorithm starts from some seeds manually drawn by the operator both in the target objects and in the background (each with a different label). Then, from these inputs the algorithm classifies the voxels iteratively using a weighted similarity criterion [256,257]. Vezhnevets et al. [256] described the growing process using the bacteria simile: at each iteration the segmentation grows from the seeded voxels as the bacteria spread attacking their neighbour cells, until they have available space. Bacteria with greater force defeat the less strong cells, thus imposing their "label" and "strength". In an image, the strength of a voxel is a measure of how likely that voxel is assigned to the correct label. The maximum strength is assigned to the seed points, while unlabelled voxels at the beginning have a strength equal to zero. The force of a voxel (the conqueror) is given by multiplying its strength with the

difference in grey-level intensities with the surrounding voxel (the defender). Convergence is achieved when all voxels are labelled, and their labels do not change iteration after iteration.

The algorithm works both with 2D and 3D images and allows multi-label segmentations. Moreover, it is based on an interactive procedure, which means that the user can revise the output segmentation and guide the algorithm in real time towards the correct solution, by introducing new seeds where the algorithm does not work properly.

The main drawbacks of the *GrowCut* algorithm are that it requires a human intervention, even if minimal compared to manual segmentation, and that the results are not fully reproducible since they depend on the input seeds. Nevertheless, it does not need intensive computational power, and thanks to its interactivity the resulting segmentation can be updated on-the-fly according to the indications of the physician. Moreover, the code has been already implemented in 3DSlicer (*Grow from seeds* tool), and therefore available to everyone, since 3DSlicer is a free software. This last point is important to make the results from different working groups comparable and integrable.

E.2 Fully automatic segmentation: the U-Net

Neural networks are an extremely interesting and complicated topic on which entire theses can be written. Here I content myself with giving the basic notions which are required in order to understand the application of convolutional neural networks discussed in the next sections. For a thorough discussion the interested readers can refer to ref. [258, 259].

Neural networks

A neural network is a collection of operational nodes connected to each other. Each node, or neuron, takes an input and transforms it into an output by performing a certain operation which depends on parameters. The neurons are arranged into layers. The first layer takes as input the data to be analysed, such as an image. The last layer returns the desired output. Each layer feeds the processed data into the consecutive one. The number of layers, the number of neurons per layer, and the neuron's operation distinguish different kinds of neural networks.

The objective is therefore the identification of a function that maps the input $\mathbf{x} \in \mathbb{R}^N$ into the output data $\mathbf{y} \in \mathbb{R}^M$,

$$\mathbf{y} = f^*(\mathbf{x})\,.$$

This function is typically unknown (the symbol * is used to denote this aspect). The only ingredients that are known, and therefore used to build the target function, are the N inputs $\mathbf{x}$ and — in the case of a supervised learning technique — the corresponding outputs $y_i = f^*(x_i)\ \forall\ i$ in $[1, N]$.

The training procedure consists in finding a function $\tilde{y}$, which is an approximator of the real function f. This approximation function is defined using a non-linear function — named *activation function* (g) — which depends on some parameters, named weights ($\boldsymbol{w}$) and bias ($\boldsymbol{b}$). The sigmoid, the hyperbolic tangent and the ReLu (Rectified Linear Unit, defined as $g(x) = max(0, x)$) are examples of activation functions often used in neural networks. The approximation function can then be expressed iteratively as

$$\tilde{\boldsymbol{y}} = \boldsymbol{G}^{[K]} = g^{[K]}(\boldsymbol{w}^{[K]} \cdot \boldsymbol{G}^{[K-1]} + \boldsymbol{b}^{[K]}) \,, \tag{E.1}$$

where K identifies the last layer of the network, and $\boldsymbol{G}^{[l]}$ is the activation of the generic l-th layer (with $1 \leq l \leq K$),

$$\begin{aligned} \boldsymbol{G}^{[l]} &= g^{[l]}(\boldsymbol{z}^{[l]}) \,, \\ \boldsymbol{z}^{[l]} &= \boldsymbol{w}^{[l]} \cdot \boldsymbol{G}^{[l-1]} + \boldsymbol{b}^{[l]} \,. \end{aligned} \tag{E.2}$$

Denoting by $u[l]$ the number of units in the l-th layer, $\boldsymbol{b}^{[l]} \in \mathbb{R}^{u[l]}$ and $\boldsymbol{w}^{[l]} \in \mathbb{R}^{u[l] \times u[l-1]}$. The iteration starts with the input layer, $G^{[0]} = \boldsymbol{x}$.

Constructing the approximation function $\tilde{\boldsymbol{y}}$ means identifying the best values for all the parameters $\boldsymbol{\theta}$ (including both the weights $\boldsymbol{w}$ and the bias $\boldsymbol{b}$) through a training procedure consisting of two steps: the forward- and the back-propagation.

In the forward-propagation, the network is evaluated from input to output, passing through the various hidden layers, mapping the inputs $\boldsymbol{x}$ onto the output data $\boldsymbol{y}$ for the given values of the parameters $\boldsymbol{\theta}$. The learning process is achieved by minimising iteratively a *cost function* which measures the "distance" of the network's output from the known target.

The loss function can be represented in general as

$$L(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \ell(\tilde{y}_i, y_i) + R(\boldsymbol{\theta}) \,, \tag{E.3}$$

where N is the number of points in the training dataset, and ℓ the cost function evaluated on each training point. Typical examples of loss function ℓ are the mean square error and the mean absolute error (also known as L^2 and L^1 norms, respectively) for regression problems, and the cross entropy loss for classification problems.

A regularisation term (R in Eq. (E.3)) is usually added to the loss function as well. It is a function of the parameters which favours smaller values of the latter in the minimisation of the cost L. This reduces the complexity of the network and thus the risk of overfitting. Examples of regularisation techniques are the L^1 and L^2 norms.

The best parameters $\boldsymbol{\theta}^*$ for the network are defined as those for which the loss function is minimal,

$$\boldsymbol{\theta}^* := argmin_{\boldsymbol{\theta}} \, L(\boldsymbol{\theta}). \tag{E.4}$$

The optimisation procedure therefore consists in the minimisation of the loss function. This is achieved with the *gradient descent algorithm*, which consists in updating iteratively each parameter of the network by a small value, proportional to the derivative of the loss function with respect to the parameter which has to be updated. Accordingly, the value of the parameters $\boldsymbol{\theta}$ at each step t of the iteration becomes

$$\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} - \eta \frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta}^{(t)}) \,, \tag{E.5}$$

where $\frac{\partial}{\partial \boldsymbol{\theta}}$ denotes the gradient with respect to the parameters of the network, and η is the *learning rate*, namely how much the gradient points towards the minimum at each step t of the iteration. It is a crucial hyper-parameter, since it affects the training and the achievement of the minimum. It can be either set to a constant throughout the training, or varied step by step in order to improve the speed of convergence of the algorithm.

Each step of the gradient descent therefore requires the computation of the loss function and of its gradient with respect to the parameters of the network. While the loss function is evaluated with the forward propagation as discussed above, its derivatives are computed with the back-propagation algorithm. The latter takes its name from the fact that the loss function is propagated backward in the network from the outputs to the inputs, computing the derivatives layer by layer. Let us consider eqs. E.1 and E.2. For the sake of simplicity, we neglect the regularisation part in the loss function. The derivatives of the loss function can be computed with the chain rule. The derivatives with respect to the parameters of the l-th layer are given by

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{w}^{[l]}} &= \frac{\partial L}{\partial \boldsymbol{z}^{[l]}} \frac{\partial \boldsymbol{z}^{[l]}}{\partial \boldsymbol{w}^{[l]}} = \frac{\partial L}{\partial \boldsymbol{z}^{[l]}} \boldsymbol{G}^{[l-1]} \,, \\ \frac{\partial L}{\partial \boldsymbol{b}^{[l]}} &= \frac{\partial L}{\partial \boldsymbol{z}^{[l]}} \frac{\partial \boldsymbol{z}^{[l]}}{\partial \boldsymbol{b}^{[l]}} = \frac{\partial L}{\partial \boldsymbol{z}^{[l]}} \,. \end{aligned} \tag{E.6}$$

The $\boldsymbol{G}^{[l-1]}$ factor is known for each layer l form the forward-propagation part of the optimisation. The only element which remains to be computed is $\partial L / \partial \boldsymbol{z}^{[l]}$. The latter is obtained during the back-propagation from the $[l+1]$ layer, where the derivatives have already been calculated, as

$$\frac{\partial L}{\partial \boldsymbol{z}^{[l]}} = \frac{\partial L}{\partial \boldsymbol{z}^{[l+1]}} \frac{\partial \boldsymbol{z}^{[l+1]}}{\partial \boldsymbol{z}^{[l]}} = \frac{\partial L}{\partial \boldsymbol{z}^{[l+1]}} \boldsymbol{w}^{[l+1]} \frac{\partial g^{[l]}}{\partial \boldsymbol{z}^{[l]}} \,, \tag{E.7}$$

where in the last equality eq. E.2 is used, exchanging $[l]$ with $[l+1]$ and $[l-1]$ with $[l]$. In this equation, $\partial L / \partial \boldsymbol{z}^{[l+1]}$ and $\boldsymbol{w}^{[l+1]}$ are known from the previous layer $[l+1]$ of the back-propagation, and $\partial g^{[l]} / \partial \boldsymbol{z}^{[l]}$ is fixed by the definition of the activation function g. Therefore, the two derivatives in eq. E.6 are determined from the forward-propagation and from the preceding layer of the back-propagation phase.

This procedure is repeated several times until the algorithm reaches a minimum. Usually, at the first step of the iteration ($t = 1$) the parameters are

set to random values. If $L(\theta)$ is convex, differentiable and its gradient is Lipschitz-continuous with constant $C > 0$ (i.e. $||\frac{\partial L(\boldsymbol{\theta_1})}{\partial \boldsymbol{\theta}} - \frac{\partial L(\boldsymbol{\theta_2})}{\partial \boldsymbol{\theta}}|| \leq C\,||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||$ $\forall\ \boldsymbol{\theta}_1, \boldsymbol{\theta}_2$), the gradient descent converges to an optimum $\boldsymbol{\theta}^*$ for $t \to \infty$ provided that $\eta \leq 1/C$. If instead $L(\boldsymbol{\theta})$ is derivable but non-convex and its gradient is Lipschitz-continuous, the gradient descent algorithm converges to a local minimum for $t \to \infty$ provided that $\eta \leq 1/C$.

Since the gradient descent algorithm requires the sum over the entire training set for each parameter and each step of the iteration, the update of the parameters may take a long time. This issue can be overcome with the *stochastic gradient descent*, which instead relies on the evaluation of the gradient in a single randomly chosen training point, rather than the entire dataset. Picking a single point at each step may however cause the gradient to oscillate during its path towards the minimum. The *mini-batch gradient descent* is a compromise between the two types of gradient descent described above, making use of a subset of the training set — named mini-batch — to update the parameters at each step.

The result of the training is a network with values of the parameters tuned so as to perform the desired task in an optimal way on the training data. The trained network is then tested on another set of labelled data to assess the network performance. Once the network is validated, it can be used on a dataset with unknown targets.

Convolutional neural network

Convolutional neural networks (CNNs) are a particular kind of neural networks. The typical CNN layer architecture is shown in **Figure E.1**. This architecture is particularly suitable for image analysis thanks to layers of neurons which convolve the input data with filters (or kernels). The result of the convolution is then run through the activation function which further filters the output and introduces non-linearity. Each filter generates an abstract image named feature map which is obtained by drawing out a certain feature from the input image. Each layer may involve multiple filters and thus creates various feature maps. The convolutional neurons of the first layer, for instance, usually detect basic patterns such as straight edges. More complex features are extracted as we go deeper into the network.

Convolutional filters can be then followed by a pooling layer. The pooling layer, instead, introduces local shift invariance and reduces the spatial resolution of the feature map (merging local group of neurons into a single unit in the next layer, for example taking the average or the maximum of this group), reducing the number of parameters and therefore the computational complexity. The last layer, instead, is a fully connected layer, by which the final prediction (such as a classification) is made. The main property of this network is the local connectivity, since the input of each unit comes from a limited number of pixels — close to each other and usually correlated to create patterns — of the

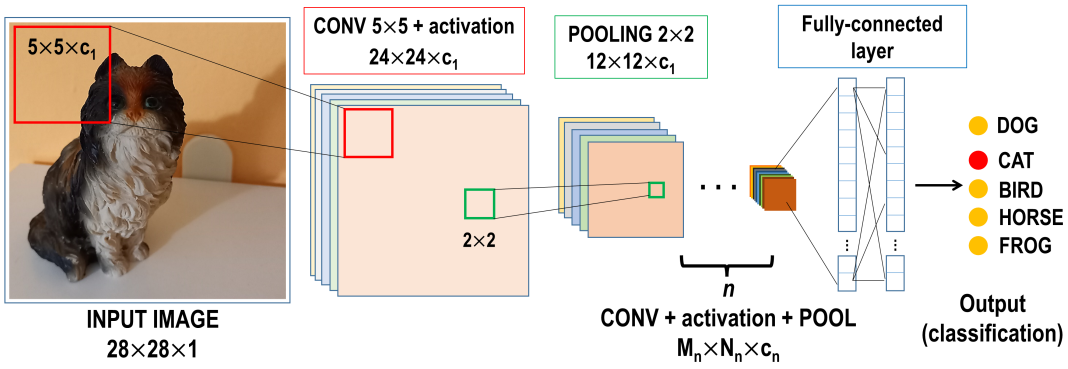


Figure E.1: Schematic representation of the CNN. The input image passes through a series of n convolutional layers and pooling layers. At each step i of this series, the filters can be applied independently multiple times, according to the depth c_i of the filter, called *channel*. The output extracted from each of these filters is thus a different feature map. Then the resulting object is flatted to turn it into an array and to be processed by a fully connected layer (in which each neuron is connected to all the neurons of the next layer) for the classification. For each possible label a probability is associated based on the features extracted in the previous layers.

previous feature map (corresponding to area convolved to the filter, named the *receptive field*). The size of the receptive field defines the size of the feature that can be caught by the network: multi-scale approaches are usually adopted to identified both small and large patterns inside the input image.

The FCNN, instead, is made of two parts: an encoding and decoding structure. The encoding part is the CNN architecture described above, in which the input image is compressed through a series of convolutions, that capture the image features at a lower and lower dimensional space. At the end of the encoding part, instead of the fully connected layer of the standard CNN, there is the decoding path, which applies deconvolutions to restore the original image resolution. With these two processes, the image is down-sampled to extract high level features and its size is more and more reduced, and then it is up-sampled increasing the spatial dimension and carrying the high-level features learned during the decoding phase to the initial resolution. Finally, the last layer is a classification operator, in which each pixel of the output image is assigned to a label. Since the classification is performed pixelwise, the final prediction maintains the information about the localisation. The final outputs of the training are the kernels that must be able to provide a predictive mask once applied to a new image, by properly classify each pixels/voxels of the input image.

U-Net

An improvement of FCNN architecture described above consists in the addiction of *long skip-connections*, which are operations of concatenation between the feature maps of the two side of the network (down-sampling and up-sapling parts) at the same depth level. This is the peculiarity of the U-Net archi-

tecture [232, 233]. The U-Net has the same structure of a FCNN but with skip-connections between the encoding and decoding paths: since each layer of the deconvolution step is receiving information from a lower spatial resolution layer, thanks to the skip connections it recovers the lost spatial localisation directly from the higher resolution layers of the encoding part [260]. In this way, each layer of the decoding architecture has both the high-level information but at lower resolution from the previous layer, and the more detailed information but at lower level from the corresponding layer of the encoding path.

Figure E.2 shows a schematic representation of the U-Net network. The first thing that catches the eye is its u-shaped structure.

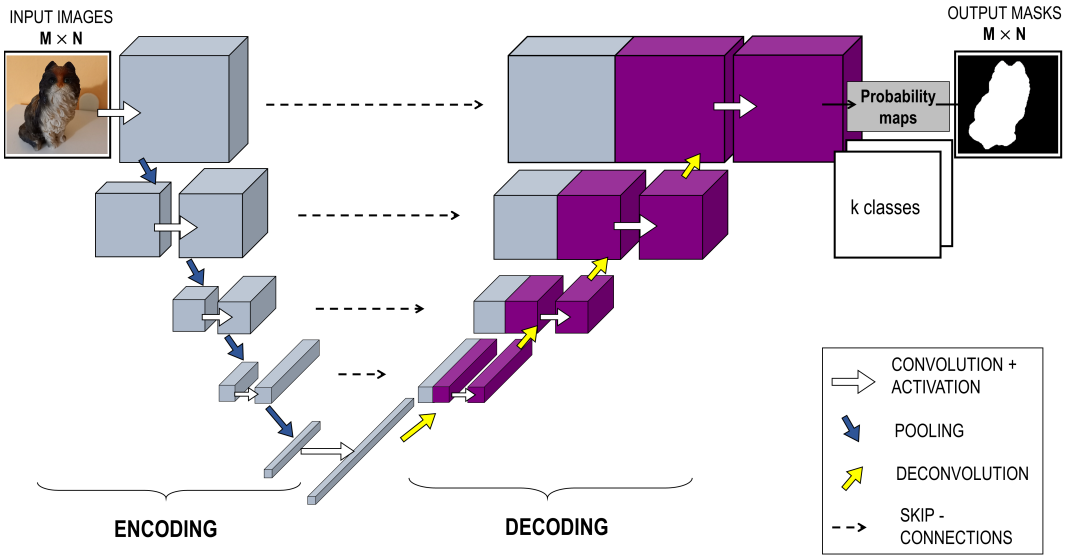


Figure E.2: Schematic representation of the U-Net, with its symmetrical encoder-decoder structure.

At the top left of the **Figure E.2** there is the input image of size $M \times N$. This data is given to the encoding part of the network which is a succession of n convolution blocks. Each block is made of a convolutional layer, an activation function (usually a rectified linear unit function) and a max pooling filter. The number n defines the depth of the network. The higher is the value of n, more features can be captured, but more parameters must be learned during the training and the network complexity increases. From the bottom to the top of the architecture there is the decoding part (on the right part of the chain), which is symmetric to the encoding part except for the deconvolution filters used to restore the original size instead of the convolutional ones. In the centre, the black dotted arrows represent the skip-connections, which connect the left down-sampling part with the corresponding up-sampling layers. At the end of the n deconvolutional blocks, the network generates a map of probability. Each point of the map returns how likely it is that the corresponding pixel belongs to each class. Finally, the class with the highest probability is assigned to each

pixel. In case of the contouring of a single object (i.e. the lesion in medical imaging), there are only two classes: the background (pixels equal to 0, black in the figure) and the target object (pixels equal to 1, white in the figure).

The prediction masks of the training data are compared to the real labels, given in input, and the weights are optimised at each iteration in order to reduce as much as possible the difference between the ground truth and the prediction. The choice of the loss function can impact the performance of the network [261]. Typical functions investigated for the segmentation task are the cross entropy [232] and the DICE coefficient [236].

Since the correct labels are provided to the algorithm at the beginning of the learning process as reference examples and are used during the minimisation operation, the architecture illustrated — and used in this thesis — is based on a supervised learning approach.

nnU-Net

The network used in this thesis was the nnU-Net ('no-new-net"), developed by Isensee and colleagues in 2021 [239]. The nnU-net is a self-configurating framework, meaning that it is able to adapt the hyperparameters and network architecture based on the characteristics of the input dataset of images. The nnU-Net, in fact, is able to select the pre-processing, the structure of the architecture, the training and the post-processing ad-hoc according to the characteristics of the input dataset, by applying some simple heuristic rules. For instance, the algorithm fits the patch size and the batch one according to the memory consumption. More in detail, larger the patch size is and better the global information are seen by the algorithm, nevertheless higher computational resources are requires. Therefore, the network starts considering the median size of all the images after resampling. Then it reduced the size iteratively until the estimated resources are below the GPU limits. If the patch size is reduced the batch size is set to 2, if not it is increased until reaching the available computational memory. Thanks to this automatic configuration of the pipeline, the network is able to adapt itself to completely different targets. The developers of the nnU-Net, for example, showed the feasibility of using this type of self-configuring architecture on a large variety of biomedical datasets, mainly composed of MRI and CT images, succeeding in segmenting different anatomical structures, both tumours (lung, brain, liver and kidney lesions) and organs (heart, liver, and kidneys).

All images are resampled with a third order spline interpolation, while the corresponding masks with linear interpolation. In case of anisotropic voxel (which is defined as maximum axis spacing / minimum axis spacing > 3), both images and masks are resampled with the nearest neighbour for z direction. The grey-level intensities are normalised using the z-score technique (subtracting the mean and then dividing by the standard deviation), but for CT modality the images are clipped between the 0.5 and the 99.5 percentiles of the foreground and then the z-score normalisation is applied. Data augmenta-

tion is also computed, and it includes rotation, scaling, mirroring, addition of Gaussian noise and Gaussian blur, gamma correction, modification of the image brightness and contrast, and reduction of the image resolution. The type of augmentation is chosen randomly according to a predefined probability value. The intensity of each augmentation procedure is assessed by extracting randomly from a predefined range the value of the parameter characteristic of each augmentation type.

Finally, for the training two loss functions are used by taking their average: a DICE and a cross-entropy loss functions.

Limitations

Some challenges still have to be faced in the application of deep architectures in the medical field [262, 263].

One of the main limitations is the difficulty in collecting large and balanced dataset of labelled images. When the number of the parameters learned during the training is large compared to the input data size, the network encounters the *overfitting* problem (see Section 1.2 for the definition). For this reason, it is important to feed the network with a large and heterogeneous training data, otherwise it may be not able to generalised on never seen samples. Testing the network on an independent dataset is often required to be sure that the network is not in overfitting. Data augmentation, the use of the *dropout* in the architecture or the transfer learning [264] may help to reduce this issue.

A second point is the choice of the hyperparameters. Hyperparameters are parameters that are not learned during the training process, but they are set at the beginning by the operator and dictate the entire training process. For example, the number of layers, the filters sizes and the type of optimisation algorithm are typical hyperparameters in CNN. At the moment, fixed values to be used have not been established yet, but case by case it is possible to find in literature advice from other expert developers. Fortunately, there are some tools that optimise the choice of the hyperparameters [265]. However, these algorithms require a long time for the optimisation and the best hyperparameters are obtained ad-hoc for each particular training set, and therefore are not generalisable.

Another important aspect is the computational requests, in particular with complex network such as the 3D architectures. Usually, GPU acceleration hardware is required to speed up and complete the training.

Last but not the least, one of the most mentioned problems of deep networks is the not easy interpretability of its hidden layers. For this reason, these deep algorithms are defined as black box models. In fact, it is not straightforward to figure out what and where the network is learning, making therefore difficult to understand what is happening during the training and the reasons of wrong classifications. However, recent techniques have been developed to help in the interpretation of the results through visualisation approaches. The main idea of such algorithms is to identify the features that are relevant for

the prediction, highlighting the areas inside the image that contribute mostly. Examples of such algorithms are the Local Interpretable Model-agnostic Explanations (LIME) algorithm [266] and the Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm [267].

Bibliography

[1] Alex Zwanenburg, Martin Vallières, Mahmoud A. Abdalah, Hugo J.W.L. Aerts, Vincent Andrearczyk, Aditya Apte, Saeed Ashrafinia, Spyridon Bakas, Roelof J. Beukinga, Ronald Boellaard, Marta Bogowicz, Luca Boldrini, Irène Buvat, Gary J.R. Cook, Christos Davatzikos, Adrien Depeursinge, Marie Charlotte Desseroit, Nicola Dinapoli, Cuong Viet Dinh, Sebastian Echegaray, Issam El Naqa, Andriy Y. Fedorov, Roberto Gatta, Robert J. Gillies, Vicky Goh, Michael Götz, Matthias Guckenberger, Sung Min Ha, Mathieu Hatt, Fabian Isensee, Philippe Lambin, Stefan Leger, Ralph T.H. Leijenaar, Jacopo Lenkowicz, Fiona Lippert, Are Losnegård, Klaus H. Maier-Hein, Olivier Morin, Henning Müller, Sandy Napel, Christophe Nioche, Fanny Orlhac, Sarthak Pati, Elisabeth A.G. Pfaehler, Arman Rahmim, Arvind U.K. Rao, Jonas Scherer, Muhammad Musib Siddique, Nanna M. Sijtsema, Jairo Socarras Fernandez, Emiliano Spezi, Roel J.H.M. Steenbakkers, Stephanie Tanadini-Lang, Daniela Thorwarth, Esther G.C. Troost, Taman Upadhaya, Vincenzo Valentini, Lisanne V. van Dijk, Joost van Griethuysen, Floris H.P. van Velden, Philip Whybra, Christian Richter, and Steffen Löck. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2):328–338, 2020.

[2] William Osler. On the educational value of the medical society. In *Aequanimitas: with other addresses to medical students, nuerses and practitioners of medicine*, pages 345–362. P. Blakiston's Son & Co, 1906.

[3] R. Fisher, L. Pusztai, and C. Swanton. Cancer heterogeneity: Implications for targeted therapeutics. *British Journal of Cancer*, 108(3):479–485, 2013.

[4] Ibiayi Dagogo-Jack and Alice T. Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15(2):81–94, 2018.

[5] Rebecca A. Burrell, Nicholas McGranahan, Jiri Bartek, and Charles Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345, 2013.

[6] Michal R. Tomaszewski and Robert J. Gillies. The biological meaning of radiomic features. *Radiology*, 298(3):505–516, 2021.

[7] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. Radiomics: Images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2016.

[8] Michele Avanzo, Joseph Stancanello, and Issam El Naqa. Beyond imaging: The promise of radiomics. *Physica Medica*, 38:122–139, 2017.

[9] R. J. Gillies, A. R. Anderson, R. A. Gatenby, and D. L. Morse. The biology underlying molecular imaging in oncology: from genome to anatome and back again. *Clinical Radiology*, 65(7):517–521, 2010.

[10] Julien Guiot, Akshayaa Vaidyanathan, Louis Deprez, Fadila Zerka, Denis Danthine, Anne Noelle Frix, Philippe Lambin, Fabio Bottari, Nathan Tsoutzidis, Benjamin Miraglio, Sean Walsh, Wim Vos, Roland Hustinx, Marta Ferreira, Pierre Lovinfosse, and Ralph T.H. Leijenaar. A review in radiomics: Making personalized medicine a reality via routine imaging. *Medicinal Research Reviews*, 42(1):426–440, 2021.

[11] Kyle J. Lafata, Yuqi Wang, Brandon Konkel, Fang Fang Yin, and Mustafa R. Bashir. Radiomics: a primer on high-throughput image phenotyping. *Abdominal Radiology*, 2021.

[12] S. Carvalho, R.T.H. Leijenaar, E.G.C. Troost, W. van Elmpt, J.-P. Muratet, F. Denis, D. De Ruysscher, H.J.W.L. Aerts, and P. Lambin. Early variation of FDG-PET radiomics features in NSCLC is related to overall survival - the "delta radiomics" concept. *Radiotherapy and Oncology*, 118:S20–S21, 2016.

[13] Parnian Afshar, Arash Mohammadi, Konstantinos N. Plataniotis, Anastasia Oikonomou, and Habib Benali. From handcrafted to deep-learning-based cancer radiomics: Challenges and opportunities. *IEEE Signal Processing Magazine*, 36(4):132–160, 2019.

[14] Stephen S.F. Yip and Hugo J.W.L. Aerts. Applications and limitations of radiomics. *Physics in Medicine and Biology*, 61(13):R150–R166, 2016.

[15] Isacco Desideri, Mauro Loi, Giulio Francolini, Carlotta Becherini, Lorenzo Livi, and Pierluigi Bonomo. Application of Radiomics for the Prediction of Radiation-Induced Toxicity in the IMRT Era: Current State-of-the-Art. *Frontiers in Oncology*, 10(October):1–10, 2020.

[16] Hugo J.W.L. Aerts, Emmanuel Rios Velazquez, Ralph T.H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Cavalho, 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, 2014.

[17] Alexandra R. Cunliffe, Samuel G. Armato, Christopher Straus, Renuka Malik, and Hania A. Al-Hallaq. Lung texture in serial thoracic CT scans: Correlation with radiologist-defined severity of acute changes following radiation therapy. *Physics in Medicine and Biology*, 59(18):5387–5398, 2014.

[18] Francesca Botta, Sara Raimondi, Lisa Rinaldi, Federica Bellerba, Federica Corso, Vincenzo Bagnardi, Daniela Origgi, Rocco Minelli, Giovanna Pitoni, Francesco Petrella, Lorenzo Spaggiari, Alessio G. Morganti, Filippo del Grande, Massimo Bellomi, and Stefania Rizzo. Association of a CT-based clinical and radiomics score of non-small cell lung cancer (NSCLC) with lymph node status and overall survival. *Cancers*, 12(6):1432, 2020.

[19] Ilke Tunali, Yan Tan, Jhanelle E Gray, Evangelia Katsoulakis, Steven A Eschrich, James Saller, Hugo J W L Aerts, Theresa Boyle, Jin Qi, Albert Guvenis, Robert J Gillies, and Matthew B Schabath. Hypoxia-Related Radiomics and Immunotherapy Response: A Multicohort Study of Non-Small Cell Lung Cancer. *JNCI Cancer Spectrum*, 5(4):1–11, 2021.

[20] Wei Mu, Lei Jiang, Yu Shi, Ilke Tunali, Jhanelle E. Gray, Evangelia Katsoulakis, Jie Tian, Robert J. Gillies, and Matthew B. Schabath. Non-invasive measurement of PD-L1 status and prediction of immunotherapy response using deep learning of PET/CT images. *Journal for ImmunoTherapy of Cancer*, 9(6):1–15, 2021.

[21] Yan Xu, Lin Lu, E. Lin-Ning, Wei Lian, Hao Yang, Lawrence H. Schwartz, Zheng Han Yang, and Binsheng Zhao. Application of radiomics in predicting the malignancy of pulmonary nodules in different sizes. *American Journal of Roentgenology*, 213(6):1213–1220, 2019.

[22] Emmanuel Rios Velazquez, Chintan Parmar, Ying Liu, Thibaud P. Coroller, Gisele Cruz, Olya Stringfield, Zhaoxiang Ye, Mike Makrigiorgos, Fiona Fennessy, Raymond H. Mak, Robert Gillies, John Quackenbush, and Hugo J.W.L. Aerts. Somatic mutations drive distinct imaging phenotypes in lung cancer. *Cancer Research*, 77(14):3922–3930, 2017.

[23] Catarina Dinis Fernandes, Cuong V. Dinh, Iris Walraven, Stijn W. Heijmink, Milena Smolic, Joost J.M. van Griethuysen, Rita Simões, Are Losnegård, Henk G. van der Poel, Floris J. Pos, and Uulke A. van der Heide. Biochemical recurrence prediction after radiotherapy for prostate cancer with T2w magnetic resonance imaging radiomic features. *Physics and Imaging in Radiation Oncology*, 7:9–15, 2018.

[24] Qiu Zi Zhong, Liu Hua Long, An Liu, Chun Mei Li, Xia Xiu, Xiu Yu Hou, Qin Hong Wu, Hong Gao, Yong Gang Xu, Ting Zhao, Dan Wang, Hai Lei Lin, Xiang Yan Sha, Wei Hu Wang, Min Chen, and Gao Feng Li. Radiomics of Multiparametric MRI to Predict Biochemical Recurrence of Localized Prostate Cancer After Radiation Therapy. *Frontiers in Oncology*, 10:1–8, 2020.

[25] Dong He, Ximing Wang, Chenchao Fu, Xuedong Wei, Jie Bao, Xuefu Ji, Honglin Bai, Wei Xia, Xin Gao, Yuhua Huang, and Jianquan Hou. MRI-based radiomics models to assess prostate cancer, extracapsular extension and positive surgical margins. *Cancer Imaging*, 21(1):1–9, 2021.

[26] Zhenyu Liu, Zhuolin Li, Jinrong Qu, Renzhi Zhang, Xuezhi Zhou, Longfei Li, Kai Sun, Zhenchao Tang, Hui Jiang, Hailiang Li, Qianqian Xiong, Yingying Ding, Xinming Zhao, Kun Wang, Zaiyi Liu, and Jie Tian. Radiomics of multiparametric MRI for pretreatment prediction of pathologic complete response to neoadjuvant chemotherapy in breast cancer: A multicenter study. *Clinical Cancer Research*, 25(12):3538–3547, 2019.

[27] Filippo Pesapane, Anna Rotili, Francesca Botta, Sara Raimondi, Linda Bianchini, Federica Corso, Federica Ferrari, Silvia Penco, Luca Nicosia, Anna Bozzini, Maria Pizzamiglio, Daniela Origgi, Marta Cremonesi, and Enrico Cassano. Radiomics of MRI for the prediction of the pathological response to neoadjuvant chemotherapy in breast cancer patients: A single referral centre analysis. *Cancers*, 13(17):4271, 2021.

[28] Qian Zhang, Yunsong Peng, Wei Liu, Jiayuan Bai, Jian Zheng, Xiaodong Yang, and Lijuan Zhou. Radiomics Based on Multimodal MRI for the Differential Diagnosis of Benign and Malignant Breast Lesions. *Journal of Magnetic Resonance Imaging*, 52(2):596–607, 2020.

[29] Nathaniel Braman, Prateek Prasanna, Jon Whitney, Salendra Singh, Niha Beig, Maryam Etesami, David D.B. Bates, Katherine Gallagher, B. Nicolas Bloch, Manasa Vulchi, Paulette Turk, Kaustav Bera, Jame Abraham, William M. Sikov, George Somlo, Lyndsay N. Harris, Hannah Gilmore, Donna Plecha, Vinay Varadan, and Anant Madabhushi. Association of Peritumoral Radiomics with Tumor Biology and Pathologic Response to Preoperative Targeted Therapy for HER2 (ERBB2) -Positive Breast Cancer. *JAMA Network Open*, 2(4):1–18, 2019.

[30] Elsa Parr, Qian Du, Chi Zhang, Chi Lin, Ahsan Kamal, Josiah McAlister, Xiaoying Liang, Kyle Bavitz, Gerard Rux, Michael Hollingsworth, Michael Baine, and Dandan Zheng. Radiomics-based outcome prediction for pancreatic cancer following stereotactic body radiotherapy. *Cancers*, 12(4):1–12, 2020.

[31] Tian Yu Tang, Xiang Li, Qi Zhang, Cheng Xiang Guo, Xiao Zhen Zhang, Meng Yi Lao, Yi Nan Shen, Wen Bo Xiao, Shi Hong Ying, Ke Sun, Ri Sheng Yu, Shun Liang Gao, Ri Sheng Que, Wei Chen, Da Bing Huang, Pei Pei Pang, Xue Li Bai, and Ting Bo Liang. Development of a Novel Multiparametric MRI Radiomic Nomogram for Preoperative Evaluation of Early Recurrence in Resectable Pancreatic Cancer. *Journal of Magnetic Resonance Imaging*, 52(1):231–245, 2020.

[32] Yucheng Zhang, Edrise M. Lobo-Mueller, Paul Karanicolas, Steven Gallinger, Masoom A. Haider, and Farzad Khalvati. Improving prognostic performance in resectable pancreatic ductal adenocarcinoma using radiomics and deep learning features fusion in CT images. *Scientific Reports*, 11(1):1–11, 2021.

[33] Shadi Ebrahimian, Ramandeep Singh, Arjunlokesh Netaji, Seetharama Kumble Madhusudhan, Fatemeh Homayounieh, Andrew Primak, Felix Lades, Sanjay Saini, Kalra Mannudeep K., and Sharma Sanjay. Characterization of Benign and Malignant Pancreatic Lesions with DECT Quantitative Metrics and Radiomics. *Academic Radiology*, 2021.

[34] Ke Li, Jingjing Xiao, Jiali Yang, Meng Li, Xuanqi Xiong, Yongjian Nian, Linbo Qiao, and Huaizhi Wang. Association of radiomic imaging features and gene expression profile as prognostic factors in pancreatic ductal adenocarcinoma. *Am J Transl Res*, 11(7):4491–4499, 2019.

[35] Yosuke Iwatate, Isamu Hoshino, Hajime Yokota, Fumitaka Ishige, Makiko Itami, Yasukuni Mori, Satoshi Chiba, Hidehito Arimitsu, Hiroo Yanagibashi, Hiroki Nagase, and Wataru Takayama. Radiogenomics for predicting p53 status, PD-L1 expression, and prognosis with machine learning in pancreatic cancer. *British Journal of Cancer*, 123(8):1253–1261, 2020.

[36] Tian Tian Zhai, Johannes A. Langendijk, Lisanne V. van Dijk, Gyorgy B. Halmos, Max J.H. Witjes, Sjoukje F. Oosting, Walter Noordzij, Nanna M. Sijtsema, and Roel J.H.M. Steenbakkers. The prognostic value of CT-based image-biomarkers for head and neck cancer patients treated with definitive (chemo-)radiation. *Oral Oncology*, 95(January):178–186, 2019.

[37] Wenbing Lv, Saeed Ashrafinia, Jianhua Ma, Lijun Lu, and Arman Rahmim. Multi-level multi-modality fusion radiomics: Application to PET and CT imaging for prognostication of head and neck cancer. *IEEE Journal of Biomedical and Health Informatics*, 24(8):2268–2277, 2020.

[38] Rui Yun Chen, Ying Chun Lin, Wei Chih Shen, Te Chun Hsieh, Kuo Yang Yen, Shang Wen Chen, and Chia Hung Kao. Associations

of Tumor PD-1 Ligands, Immunohistochemical Studies, and Textural Features in 18F-FDG PET in Squamous Cell Carcinoma of the Head and Neck. *Scientific Reports*, 8(1):1–10, 2018.

[39] Chao Huang, Murilo Cintra, Kevin Brennan, Mu Zhou, A. Dimitrios Colevas, Nancy Fischbein, Shankuan Zhu, and Olivier Gevaert. Development and validation of radiomic signatures of head and neck squamous cell carcinoma molecular features and subtypes. *EBioMedicine*, 45:70–80, 2019.

[40] Arman Rahmim, Peng Huang, Nikolay Shenkov, Sima Fotouhi, Esmaeil Davoodi-Bojd, Lijun Lu, Zoltan Mari, Hamid Soltanian-Zadeh, and Vesna Sossi. Improved prediction of outcome in Parkinson's disease using radiomics analysis of longitudinal DAT SPECT images. *NeuroImage: Clinical*, 16:539–544, 2017.

[41] Kun Zhao, Yanhui Ding, Ying Han, Yong Fan, Aaron F. Alexander-Bloch, Tong Han, Dan Jin, Bing Liu, Jie Lu, Chengyuan Song, Pan Wang, Dawei Wang, Qing Wang, Kaibin Xu, Hongwei Yang, Hongxiang Yao, Yuanjie Zheng, Chunshui Yu, Bo Zhou, Xinqing Zhang, Yuying Zhou, Tianzi Jiang, Xi Zhang, and Yong Liu. Independent and reproducible hippocampal radiomic biomarkers for multisite Alzheimer's disease: diagnosis, longitudinal progress and biological basis. *Science Bulletin*, 65(13):1103–1113, 2020.

[42] Sara Ranjbar, Stefanie N. Velgos, Amylou C. Dueck, Yonas E. Geda, and J. Ross Mitchell. Brain MR radiomics to differentiate cognitive disorders. *Journal of Neuropsychiatry and Clinical Neurosciences*, 31(3):210–219, 2019.

[43] Qingxia Wu, Shuo Wang, Liang Li, Qingxia Wu, Wei Qian, Yahua Hu, Li Li, Xuezhi Zhou, He Ma, Hongjun Li, Meiyun Wang, Xiaoming Qiu, Yunfei Zha, and Jie Tian. Radiomics analysis of computed tomography helps predict poor prognostic outcome in COVID-19. *Theranostics*, 10(16):7231–7244, 2020.

[44] Fatemeh Homayounieh, Shadi Ebrahimian, Rosa Babaei, Hadi Karimi Mobin, Eric Zhang, Bernardo Canedo Bizzo, Iman Mohseni, Subba R. Digumarthy, and Mannudeep K. Kalra. Ct radiomics, radiologists, and clinical information in predicting outcome of patients with covid-19 pneumonia. *Radiology: Cardiothoracic Imaging*, 2(4), 2020.

[45] Jia Huang, Feihong Wu, Leqing Chen, Jie Yu, Wengang Sun, Zhuang Nie, Huan Liu, Fan Yang, and Chuansheng Zheng. Ct-based radiomics helps to predict residual lung lesions in covid-19 patients at three months after discharge. *Diagnostics*, 11(10), 2021.

[46] Alessandro Stefano, Mauro Gioè, Giorgio Russo, Stefano Palmucci, Sebastiano Emanuele Torrisi, Samuel Bignardi, Antonio Basile, Albert

Comelli, Viviana Benfante, Gianluca Sambataro, Daniele Falsaperla, Alfredo Gaetano Torcitto, Massimo Attanasio, Anthony Yezzi, and Carlo Vancheri. Performance of radiomics features in the quantification of idiopathic pulmonary fibrosis from HRCT. *Diagnostics*, 10(5):306, 2020.

[47] Jihye Yun, Young Hoon Cho, Sang Min Lee, Jeongeun Hwang, Jae Seung Lee, Yeon Mok Oh, Sang Do Lee, Li Cher Loh, Choo Khoon Ong, Joon Beom Seo, and Namkug Kim. Deep radiomics-based survival prediction in patients with chronic obstructive pulmonary disease. *Scientific Reports*, 11(1):1–9, 2021.

[48] Márton Kolossváry, Júlia Karády, Bálint Szilveszter, Pieter Kitslaar, Udo Hoffmann, Béla Merkely, and Pál Maurovich-Horvat. Radiomic Features Are Superior to Conventional Quantitative Computed Tomographic Metrics to Identify Coronary Plaques with Napkin-Ring Sign. *Circulation: Cardiovascular Imaging*, 10(12):1–9, 2017.

[49] Ulf Neisius, Hossam El-Rewaidy, Shiro Nakamori, Jennifer Rodriguez, Warren J. Manning, and Reza Nezafat. Radiomic Analysis of Myocardial Native T1 Imaging Discriminates Between Hypertensive Heart Disease and Hypertrophic Cardiomyopathy. *JACC: Cardiovascular Imaging*, 12(10):1946–1954, 2019.

[50] Aboelyazid Elkilany, Uli Fehrenbach, Timo Alexander Auer, Tobias Müller, Wenzel Schöning, Bernd Hamm, and Dominik Geisel. A radiomics-based model to classify the etiology of liver cirrhosis using gadoxetic acid-enhanced MRI. *Scientific Reports*, 11(1):1–13, 2021.

[51] Elisa Scalco and Giovanna Rizzo. Texture analysis of medical images for radiotherapy applications. *British Journal of Radiology*, 90(1070), 2017.

[52] Michele Avanzo, Lise Wei, Joseph Stancanello, Martin Vallières, Arvind Rao, Olivier Morin, Sarah A. Mattonen, and Issam El Naqa. Machine and deep learning methods for radiomics. *Medical Physics*, 47(5):e185–e202, 2020.

[53] Nikolaos Papanikolaou, Celso Matos, and Dow Mu Koh. How to develop a meaningful radiomic signature for clinical use in oncologic patients. *Cancer Imaging*, 20(1):1–10, 2020.

[54] Yingying Jia, Jun Yang, Yangyang Zhu, Fang Nie, HaoAO Wu, Ying Duan, and Kundi Chen. Ultrasound-based radiomics: current status, challenges and future opportunities. *Medical Ultrasonography*, 2021.

[55] Francesca Ng, Robert Kozarski, Balaji Ganeshan, and Vicky Goh. Assessment of tumor heterogeneity by CT texture analysis: Can the largest cross-sectional area be used as an alternative to whole tumor analysis? *European Journal of Radiology*, 82(2):342–348, 2013.

[56] Xenia Fave, Molly Cook, Amy Frederick, Lifei Zhang, Jinzhong Yang, David Fried, Francesco Stingo, and Laurence Court. Preliminary investigation into sources of uncertainty in quantitative imaging features. *Computerized Medical Imaging and Graphics*, 44:54–61, 2015.

[57] Lifeng Yang, Jingbo Yang, Xiaobo Zhou, Liyu Huang, Weiling Zhao, Tao Wang, Jian Zhuang, and Jie Tian. Development of a radiomics nomogram based on the 2D and 3D CT features to predict the survival of non-small cell lung cancer patients. *European Radiology*, 29(5):2196–2206, 2019.

[58] A.N. Akansu, M.V. Tazebay, M.J. Medley, and P.K. Das. Wavelet and subband transforms: fundamentals and communication applications. *IEEE Communications MagazinIEEE Communications Magazine*, 35(12):104–115, 1997.

[59] Rafael C. Gonzalez and Richard E. Wood. *Digital Image Processing (Second Edition)*. Prentice Hall, 2002.

[60] Paul S. Addison. The discrete wavelet transform. In *The Illustrated Wavelet Transform Handbook*. CRC Press, 2002.

[61] László G. Nyú and Jayaram K. Udupa. On standardizing the MR image intensity scale. *Magnetic Resonance in Medicine*, 42(6):1072–1081, 1999.

[62] Michael Schwier, Joost van Griethuysen, Mark G. Vangel, Steve Pieper, Sharon Peled, Clare Tempany, Hugo J.W.L. Aerts, Ron Kikinis, Fiona M. Fennessy, and Andriy Fedorov. Repeatability of Multiparametric Prostate MRI Radiomics Features. *Scientific Reports*, 9(1):1–16, 2019.

[63] Lars J. Isaksson, Sara Raimondi, Francesca Botta, Matteo Pepa, Simone G. Gugliandolo, Simone P. De Angelis, Giulia Marvaso, Giuseppe Petralia, Ottavio De Cobelli, Sara Gandini, Marta Cremonesi, Federica Cattani, Paul Summers, and Barbara A. Jereczek-Fossa. Effects of MRI image normalization techniques in prostate cancer radiomics. *Physica Medica*, 71:7–13, 2020.

[64] Elisa Scalco, Antonella Belfatto, Alfonso Mastropietro, Tiziana Rancati, Barbara Avuzzi, Antonella Messina, Riccardo Valdagni, and Giovanna Rizzo. T2w-MRI signal normalization affects radiomics features reproducibility. *Medical Physics*, 47(4):1680–1691, 2020.

[65] Robert M. Haralick, Its'hak Dinstein, and K. Shanmugam. Textural Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3(6):610–621, 1973.

[66] Mary M. Galloway. Texture analysis using grey level run lengths. *Computer Graphics and Image Processing*, 4(2):172–179, 1975.

[67] Guillaume Thibault, Bernard Fertil, Claire Navarro, Sandrine Pereira, Pierre Cau, Nicolas Levy, Jean Sequeira, and Jean Luc Mari. Shape and texture indexes application to cell nuclei classification. *International Journal of Pattern Recognition and Artificial Intelligence*, 27(1), 2013.

[68] Chengjun Sun and William G Wee. Neighboring gray level dependence matrix for texture classification. *Computer Vision, Graphics, and Image Processing*, 23(3):341–352, 1983.

[69] Moses Amadasun and Robert King. Texural Features Corresponding to Texural Properties. *IEEE Transactions on Systems, Man and Cybernetics*, 19(5):1264–1274, 1989.

[70] Piotr M. Szczypiński, Michał Strzelecki, Andrzej Materka, and Artur Klepaczko. MaZda-A software package for image texture analysis. *Computer Methods and Programs in Biomedicine*, 94(1):66–76, 2009.

[71] Lifei Zhang, David V. Fried, Xenia J. Fave, Luke A. Hunter, Jinzhong Yang, and Laurence E. Court. Ibex: An open infrastructure software platform to facilitate collaborative work in radiomics. *Medical Physics*, 42(3), 2015.

[72] Christophe Nioche, Fanny Orlhac, Sarah Boughdad, Sylvain Reuze, Jessica Goya-Outi, Charlotte Robert, Claire Pellot-Barakat, Michael Soussan, Fred erique Frouin, and Irene Buvat. Lifex: A freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. *Cancer Research*, 78(16):4786–4789, 2018.

[73] Joost J.M. Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21):e104–e107, 2017.

[74] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer New York, New York, 2013.

[75] Jilian Tang, Salem Aleyani, and Huan Liu. Feature selection for classification: a review. In *Data Classification*, pages 37–64. CRC Press, 2015.

[76] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[77] Juan Diego Rodríguez, Aritz Pérez, and Jose Antonio Lozano. Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):569–575, 2010.

[78] Marcel Beister, Daniel Kolditz, and Willi A. Kalender. Iterative reconstruction methods in X-ray CT. *Physica Medica*, 28(2):94–108, 2012.

[79] J. Geleijns. Computed Tomography. In *Diagnostic radiology physics: A handbook for teachers and students*, pages 257–290. INTERNATIONAL ATOMIC ENERGY AGENCY, Vienna, 2014.

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

[81] Isabella Fornacon-Wood, Corinne Faivre-Finn, James P.B. O'Connor, and Gareth J. Price. Radiomics as a personalized medicine tool in lung cancer: Separating the hope from the hype. *Lung Cancer*, 146:197–208, 2020.

[82] Reza Reiazi, Engy Abbas, Petra Famiyeh, Aria Rezaie, Jennifer Y.Y. Kwan, Tirth Patel, Scott V. Bratman, Tony Tadic, Fei Fei Liu, and Benjamin Haibe-Kains. The impact of the variation of imaging parameters on the robustness of Computed Tomography radiomic features: A review. *Computers in Biology and Medicine*, 133:104400, 2021.

[83] Binsheng Zhao. Understanding Sources of Variation to Improve the Reproducibility of Radiomics. *Frontiers in Oncology*, 11:1–21, 2021.

[84] ISO. International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM) - 3rd ed. *Vocabulary*, (VIM):1–150, 2007.

[85] Binsheng Zhao, Leonard P. James, Chaya S. Moskowitz, Pingzhen Guo, Michelle S. Ginsberg, Robert A. Lefkowitz, Yilin Qin, Gregory J. Riely, Mark G. Kris, and Lawrence H. Schwartz. Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. *Radiology*, 252(1):263–272, 2009.

[86] Tai H. Dou, Thibaud P. Coroller, Joost J.M. van Griethuysen, Raymond H. Mak, and Hugo J.W.L. Aerts. Peritumoral radiomics features predict distant metastasis in locally advanced NSCLC. *PLoS ONE*, 13(11):1–15, 2018.

[87] Mohammadhadi Khorrami, Monica Khunger, Alexia Zagouras, Pradnya Patil, Rajat Thawani, Kaustav Bera, Prabhakar Rajiah, Pingfu Fu, Vamsidhar Velcheti, and Anant Madabhushi. Combination of Peri- and Intratumoral Radiomic Features on Baseline CT Scans Predicts Response to Chemotherapy in Lung Adenocarcinoma. *Radiology: Artificial Intelligence*, 1(2):180012, 2019.

[88] Shohei Tanaka, Noriyuki Kadoya, Tomohiro Kajikawa, Shohei Matsuda, Suguru Dobashi, Ken Takeda, and Keiichi Jingu. Investigation of thoracic four-dimensional CT-based dimension reduction technique for extracting the robust radiomic features. *Physica Medica*, 58:141–148, 2019.

[89] Ilke Tunali, Lawrence O. Hall, Sandy Napel, Dmitry Cherezov, Albert Guvenis, Robert J. Gillies, and Matthew B. Schabath. Stability and reproducibility of computed tomography radiomic features extracted from peritumoral regions of lung cancer lesions. *Medical Physics*, 46(11):5075–5085, 2019.

[90] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, 2013.

[91] Lin Lu, Ross C. Ehmke, Lawrence H. Schwartz, and Binsheng Zhao. Assessing agreement between radiomic features computed for multiple CT imaging settings. *PLoS ONE*, 11(12):1–12, 2016.

[92] Binsheng Zhao, Yongqiang Tan, Wei Yann Tsai, Jing Qi, Chuanmiao Xie, Lin Lu, and Lawrence H. Schwartz. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Scientific Reports*, 6:1–7, 2016.

[93] Sohee Park, Sang Min Lee, Kyung Hyun Do, June Goo Lee, Woong Bae, Hyunho Park, Kyu Hwan Jung, and Joon Beom Seo. Deep learning algorithm for reducing ct slice thickness: Effect on reproducibility of radiomic features in lung cancer. *Korean Journal of Radiology*, 20(10):1431–1440, 2019.

[94] Barbaros S. Erdal, Mutlu Demirer, Kevin J. Little, Chiemezie C. Amadi, Gehan F.M. Ibrahim, Thomas P. O'Donnell, Rainer Grimmer, Vikash Gupta, Luciano M. Prevedello, and Richard D. White. Are quantitative features of lung nodules reproducible at different CT acquisition and reconstruction parameters? *PLoS ONE*, 15(10):1–9, 2020.

[95] Jinzhong Yang, Lifei Zhang, Xenia J. Fave, David V. Fried, Francesco C. Stingo, Chaan S. Ng, and Laurence E. Court. Uncertainty analysis of quantitative imaging features extracted from contrast-enhanced CT in lung tumors. *Computerized Medical Imaging and Graphics*, 48:1–8, 2016.

[96] Ryo Kakino, Mitsuhiro Nakamura, Takamasa Mitsuyoshi, Takashi Shintani, Hideaki Hirashima, Yukinori Matsuo, and Takashi Mizowaki. Comparison of radiomic features in diagnostic CT images with and without contrast enhancement in the delayed phase for NSCLC patients. *Physica Medica*, 69:176–182, 2020.

[97] Hyungjin Kim, Chang Min Park, Myunghee Lee, Sang Joon Park, Yong Sub Song, Jong Hyuk Lee, Eui Jin Hwang, and Jin Mo Goo. Impact of reconstruction algorithms on CT radiomic features of pulmonary tumors: Analysis of intra- and inter-reader variability and inter-reconstruction algorithm variability. *PLoS ONE*, 11(10):1–11, 2016.

[98] Justin Solomon, Achille Mileto, Rendon C. Nelson, Kingshuk Roy Choudhury, and Ehsan Samei. Quantitative features of liver lesions, lung nodules, and renal stones at multi-detector row CT examinations: Dependency on radiation dose and reconstruction algorithm. *Radiology*, 279(1):185–194, 2016.

[99] Davide Prezzi, Katarzyna Owczarczyk, Paul Bassett, Muhammad Siddique, David J. Breen, Gary J.R. Cook, and Vicky Goh. Adaptive statistical iterative reconstruction (ASIR) affects CT radiomics quantification in primary colorectal cancer. *European Radiology*, 29(10):5227–5235, 2019.

[100] Pamela Sung, Jeong Min Lee, Ijin Joo, Sanghyup Lee, Tae Hyung Kim, and Balaji Ganeshan. Evaluation of the impact of iterative reconstruction algorithms on computed tomography texture features of the liver parenchyma using the filtration-histogram method. *Korean Journal of Radiology*, 20(4):558–568, 2019.

[101] Isabella Fornacon-Wood, Hitesh Mistry, Christoph J. Ackermann, Fiona Blackhall, Andrew McPartlin, Corinne Faivre-Finn, Gareth J. Price, and James P.B. O'Connor. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. *European Radiology*, 30(11):6241–6250, 2020.

[102] Xenia Fave, Lifei Zhang, Jinzhong Yang, Dennis Mackin, Peter Balter, Daniel Gomez, David Followill, A. Kyle 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.

[103] Muhammad Shafiq-Ul-Hassan, Kujtim Latifi, Geoffrey Zhang, Ghanim Ullah, Robert Gillies, and Eduardo Moros. Voxel size and gray level normalization of CT radiomic features in lung cancer. *Scientific Reports*, 8(1):1–9, 2018.

[104] H. Y.C. Wang, E. M. Donovan, A. Nisbet, C. P. South, S. Alobaidli, V. Ezhil, I. Phillips, V. Prakash, M. Ferreira, P. Webster, and P. M. Evans. The stability of imaging biomarkers in radiomics: A framework for evaluation. *Physics in Medicine and Biology*, 64(16), 2019.

[105] Bum Woo Park, Jeong Kon Kim, Changhoe Heo, and Kye Jin Park. Reliability of CT radiomic features reflecting tumour heterogeneity ac-

cording to image quality and image processing parameters. *Scientific Reports*, 10(1):1–13, 2020.

[106] Dennis Mackin, Xenia Fave, Lifei Zhang, David Fried, Jinzhong Yang, Brian Taylor, Edgardo Rodriguez-Rivera, Cristina Dodge, Aaron Kyle Jones, and Laurence Court. Measuring computed tomography scanner variability of radiomics features. *Investigative Radiology*, 50(11):757–765, 2015.

[107] Dennis MacKin, Rachel Ger, Cristina Dodge, Xenia Fave, Pai Chun Chi, Lifei Zhang, Jinzhong Yang, Steve Bache, Charles Dodge, A. Kyle Jones, and Laurence Court. Effect of tube current on computed tomography radiomic features. *Scientific Reports*, 8(1):1–10, 2018.

[108] 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.

[109] 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.

[110] Dennis Mackin, Xenia Fave, Lifei Zhang, Jinzhong Yang, A. Kyle Jones, Chaan S. Ng, and Laurence Court. Harmonizing the pixel size in retrospective computed tomography radiomics studies. *PLoS ONE*, 12(9):1–17, 2017.

[111] Rachel B. Ger, Shouhao Zhou, Pai Chun Melinda Chi, Hannah J. Lee, Rick R. Layman, A. Kyle Jones, David L. Goff, Clifton D. Fuller, Rebecca M. Howell, Heng Li, R. Jason Stafford, Laurence E. Court, and Dennis S. Mackin. Comprehensive Investigation on Controlling for CT Imaging Variabilities in Radiomics Studies. *Scientific Reports*, 8(1):1–14, 2018.

[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*, 20(8):155–163, 2019.

[113] Oscar Jimenez-del Toro, Christoph Aberle, Michael Bach, Roger Schaer, Markus M. Obmann, Kyriakos Flouris, Ender Konukoglu, Bram Stieltjes, Henning Müller, and Adrien Depeursinge. The Discriminative Power and Stability of Radiomics Features With Computed Tomography Variations. *Investigative Radiology*, 56(12):820–825, 2021.

[114] Bino A. Varghese, Darryl Hwang, Steven Y. Cen, Xiaomeng Lei, Joshua Levy, Bhushan Desai, David J. Goodenough, and Vinay A. Duddalwar. Identification of robust and reproducible CT-texture metrics using a customized 3D-printed texture phantom. *Journal of Applied Clinical Medical Physics*, 22(2):98–107, 2021.

[115] R. Da-ano, I. Masson, F. Lucia, M. Doré, P. Robin, J. Alfieri, C. Rousseau, A. Mervoyer, C. Reinhold, J. Castelli, R. De Crevoisier, J. F. Rameé, O. Pradier, U. Schick, D. Visvikis, and M. Hatt. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. *Scientific Reports*, 10(1):1–12, 2020.

[116] R. Da-Ano, D. Visvikis, and M. Hatt. Harmonization strategies for multicenter radiomics investigations. *Physics in Medicine and Biology*, 65(24), 2020.

[117] Fanny Orlhac, Jakoba J Eertink, Anne-Segolene Cottereau, Josee M. Zijlstra, Catherine Thieblemont, Michel A. Meignan, Ronald Boellaard, and Irene Buvat. A guide to ComBat harmonization of imaging biomarkers in multicenter studies. *Journal of Nuclear Medicine*, 63(2):172–179, 2022.

[118] Shruti Atul Mali, Abdalla Ibrahim, Henry C. Woodruff, Vincent Andrearczyk, Henning Müller, Sergey Primakov, Zohaib Salahuddin, Avishek Chatterjee, and Philippe Lambin. Making radiomics more reproducible across scanner and imaging protocol variations: A review of harmonization methods. *Journal of Personalized Medicine*, 11(9):842, 2021.

[119] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[120] 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.

[121] Marta Ligero, Olivia Jordi-Ollero, Kinga Bernatowicz, Alonso Garcia-Ruiz, Eric Delgado-Muñoz, David Leiva, Richard Mast, Cristina Suarez, Roser Sala-Llonch, Nahum Calvo, Manuel Escobar, Arturo Navarro-Martin, Guillermo Villacampa, Rodrigo Dienstmann, and Raquel Perez-Lopez. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. *European Radiology*, 31(3):1460–1470, 2021.

[122] Fanny Orlhac, Augustin Lecler, Julien Savatovski, Jessica Goya-Outi, Christophe Nioche, Frédérique Charbonneau, Nicholas Ayache, Frédérique Frouin, Loïc Duron, and Irène Buvat. How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. *European Radiology*, 31(4):2272–2280, 2021.

[123] Fanny Orlhac, Sarah Boughdad, Cathy Philippe, Hugo Stalla-Bourdillon, Christophe Nioche, Laurence Champion, Michael Soussan, Frederique Frouin, Vincent Frouin, and Irene Buvat. A postreconstruction harmonization method for multicenter radiomic studies in PET. *Journal of Nuclear Medicine*, 59(8):1321–1328, 2018.

[124] Jooae Choe, Sang Min Lee, Kyung Hyun Do, Gaeun Lee, June Goo Lee, Sang Min Lee, and Joon Beom Seo. Deep Learning–based Image Conversion of CT Reconstruction Kernels Improves Radiomics Reproducibility for Pulmonary Nodules or Masses. *Radiology*, 292(2):365–373, 2019.

[125] Jin H. Yoon, Shawn H. Sun, Manjun Xiao, Hao Yang, Lin Lu, Yajun Li, Lawrence H. Schwartz, and Binsheng Zhao. Convolutional neural network addresses the confounding impact of CT reconstruction kernels on radiomics studies. *Tomography*, 7(4):877–892, 2021.

[126] Chintan Parmar, Emmanuel Rios Velazquez, Ralph Leijenaar, Mohammed Jermoumi, Sara Carvalho, Raymond H. Mak, Sushmita Mitra, B. Uma 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):1–8, 2014.

[127] Matea Pavic, Marta Bogowicz, Xaver Würms, Stefan Glatz, Tobias Finazzi, Oliver Riesterer, Johannes Roesch, Leonie Rudofsky, Martina Friess, Martin Huellner, Isabelle Opitz, Walter Weder, Matthias Guckenberger, Stephanie Tanadini-lang, Matea Pavic, Marta Bogowicz, Xaver Würms, Stefan Glatz, Tobias Finazzi, Oliver Riesterer, Johannes Roesch, Leonie Rudofsky, Martina Friess, Patrick Veit-haibach, Isabelle Opitz, Walter Weder, Thomas Frauenfelder, and Matthias Guckenberger. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. *Acta Oncologica*, 57(8):1070–1074, 2018.

[128] Constance A. Owens, Christine B. Peterson, Chad Tang, Eugene J. Koay, Wen Yu, Dennis S. Mackin, Jing Li, Mohammad R. Salehpour, David T. Fuentes, Laurence E. Court, and Jinzhong Yang. Lung tumor segmentation methods: Impact on the uncertainty of radiomics features for non-small cell lung cancer. *PLoS ONE*, 13(10):1–22, 2018.

[129] Jeffrey Wong, Michael Baine, Sarah Wisnoskie, Nathan Bennion, Dechun Zheng, Lei Yu, Vipin Dalal, Michael A. Hollingsworth, Chi Lin, and Dandan Zheng. Effects of interobserver and interdisciplinary segmentation

variabilities on CT-based radiomics for pancreatic cancer. *Scientific Reports*, 11(1):1–12, 2021.

[130] Francesco Bianconi, Mario Luca Fravolini, Isabella Palumbo, Giulia Pascoletti, Susanna Nuvoli, Maria Rondini, Angela Spanu, and Barbara Palumbo. Impact of lesion delineation and intensity quantisation on the stability of texture features from lung nodules on ct: A reproducible study. *Diagnostics*, 11(7):1–17, 2021.

[131] Qiao Huang, Lin Lu, Laurent Dercle, Philip Lichtenstein, Yajun Li, Qian Yin, Min Zong, Lawrence Schwartz, and Binsheng Zhao. Interobserver variability in tumor contouring affects the use of radiomics to predict mutational status. *Journal of Medical Imaging*, 5(01):1, 2017.

[132] Christoph Haarburger, Gustav Müller-Franzes, Leon Weninger, Christiane Kuhl, Daniel Truhn, and Dorit Merhof. Radiomics feature reproducibility under inter-rater variability in segmentations of CT images. *Scientific Reports*, 10(1):1–10, 2020.

[133] Francesco Bianconi, Mario Luca Fravolini, Sofia Pizzoli, Isabella Palumbo, Matteo Minestrini, Maria Rondini, Susanna Nuvoli, Angela Spanu, and Barbara Palumbo. Comparative evaluation of conventional and deep learning methods for semi-automated segmentation of pulmonary nodules on CT. *Quantitative Imaging in Medicine and Surgery*, 11(7):3286–3305, 2021.

[134] Chintan Parmar, Patrick Grossmann, Johan Bussink, Philippe Lambin, and Hugo J.W.L. Aerts. Machine Learning methods for Quantitative Radiomic Biomarkers. *Scientific Reports*, 5:1–11, 2015.

[135] Yucheng Zhang, Anastasia Oikonomou, Alexander Wong, Masoom A. Haider, and Farzad Khalvati. Radiomics-based Prognosis Analysis for Non-Small Cell Lung Cancer. *Scientific Reports*, 7:1–8, 2017.

[136] Stefan Leger, Alex Zwanenburg, Karoline Pilz, Fabian Lohaus, Annett Linge, Klaus Zöphel, Jörg Kotzerke, Andreas Schreiber, Inge Tinhofer, Volker Budach, Ali Sak, Martin Stuschke, Panagiotis Balermpas, Claus Rödel, Ute Ganswindt, Claus Belka, Steffi Pigorsch, Stephanie E. Combs, David Mönnich, Daniel Zips, Mechthild Krause, Michael Baumann, Esther G.C. Troost, Steffen Löck, and Christian Richter. A comparative study of machine learning methods for time-To-event survival data for radiomics risk modelling. *Scientific Reports*, 7(1):1–11, 2017.

[137] Wenzheng Sun, Mingyan Jiang, Jun Dang, Panchun Chang, and Fang Fang Yin. Effect of machine learning methods on predicting NSCLC overall survival time based on Radiomics analysis. *Radiation Oncology*, 13(1):1–8, 2018.

[138] Mengmeng Yan and Weidong Wang. A Non-invasive Method to Diagnose Lung Adenocarcinoma. *Frontiers in Oncology*, 10, 2020.

[139] Darcie A.P. Delzell, Sara Magnuson, Tabitha Peter, Michelle Smith, and Brian J. Smith. Machine Learning and Feature Selection Methods for Disease Classification With Application to Lung Cancer Screening Image Data. *Frontiers in Oncology*, 9:1–8, 2019.

[140] Federica Corso, Giulia Tini, Giuliana Lo Presti, Noemi Garau, Simone Pietro De Angelis, Federica Bellerba, Lisa Rinaldi, Francesca Botta, Stefania Rizzo, Daniela Origgi, Chiara Paganelli, Marta Cremonesi, Cristiano Rampinelli, Massimo Bellomi, Luca Mazzarella, Pier Giuseppe Pelicci, Sara Gandini, and Sara Raimondi. The challenge of choosing the best classification method in radiomic analyses: Recommendations and applications to lung cancer CT images. *Cancers*, 13(12), 2021.

[141] Philippe Lambin, Ralph T.H. Leijenaar, Timo M. Deist, Jurgen Peerlings, Evelyn E.C. De Jong, Janita Van Timmeren, Sebastian Sanduleanu, Ruben T.H.M. Larue, Aniek J.G. Even, Arthur Jochems, Yvonka Van Wijk, Henry Woodruff, Johan Van Soest, Tim Lustberg, Erik Roelofs, Wouter Van Elmpt, Andre Dekker, Felix M. Mottaghy, Joachim E. Wildberger, and Sean Walsh. Radiomics: The bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*, 14(12):749–762, 2017.

[142] Alberto Traverso, Leonard Wee, Andre Dekker, and Robert Gillies. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *International Journal of Radiation Oncology Biology Physics*, 102(4):1143–1158, 2018.

[143] Elisabeth Pfaehler, Ivan Zhovannik, Lise Wei, Ronald Boellaard, Andre Dekker, René Monshouwer, Issam El Naqa, Jan Bussink, Robert Gillies, Leonard Wee, and Alberto Traverso. A systematic review and quality of reporting checklist for repeatability and reproducibility of radiomic features. *Physics and Imaging in Radiation Oncology*, 20(November):69–75, 2021.

[144] Adrien Depeursinge, Vincent Andrearczyk, Philip Whybra, Joost van Griethuysen, Henning Müller, Roger Schaer, Martin Vallières, and Alex Zwanenburg. Standardised convolutional filtering for radiomics. 2020.

[145] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021.

[146] P. E. Postmus, K. M. Kerr, M. Oudkerk, S. Senan, D. A. Waller, J. Vansteenkiste, C. Escriu, and S. Peters. Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 28(Supplement 4):iv1–iv21, 2017.

[147] Kathryn C. Arbour and Gregory J. Riely. Systemic therapy for locally advanced and metastatic non-small cell lung cancer: A review. *JAMA - Journal of the American Medical Association*, 322(8):764–774, 2019.

[148] William D. Travis, Elisabeth Brambilla, Andrew G. Nicholson, Yasushi Yatabe, John H.M. Austin, Mary Beth Beasley, Lucian R. Chirieac, Sanja Dacic, Edwina Duhig, Douglas B. Flieder, Kim Geisinger, Fred R. Hirsch, Yuichi Ishikawa, Keith M. Kerr, Masayuki Noguchi, Giuseppe Pelosi, Charles A. Powell, Ming Sound Tsao, and Ignacio Wistuba. The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances since the 2004 Classification. *Journal of Thoracic Oncology*, 10(9):1243–1260, 2015.

[149] Roy S. Herbst, Daniel Morgensztern, and Chris Boshoff. The biology and management of non-small cell lung cancer. *Nature*, 553(7689):446–454, 2018.

[150] D. Planchard, S. Popat, K. Kerr, S. Novello, E. F. Smit, C. Faivre-Finn, T. S. Mok, M. Reck, P. E. Van Schil, M. D. Hellmann, and S. Peters. Metastatic non-small cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 29:iv192–iv237, 2018.

[151] Ilke Tunali, Robert J. Gillies, and Matthew B. Schabath. Application of radiomics and artificial intelligence for lung cancer precision medicine. *Cold Spring Harbor Perspectives in Medicine*, 11(8), 2021.

[152] Michele Avanzo, Joseph Stancanello, Giovanni Pirrone, and Giovanna Sartor. Radiomics and deep learning in lung cancer. *Strahlentherapie und Onkologie*, 196(10):879–887, 2020.

[153] Madhurima R. Chetan and Fergus V. Gleeson. Radiomics in predicting treatment response in non-small-cell lung cancer: current status, challenges and future perspectives. *European Radiology*, 31(2):1049–1058, 2021.

[154] G. M. Walls, S. O.S. Osman, K. H. Brown, K. T. Butterworth, G. G. Hanna, A. R. Hounsell, C. K. McGarry, R. T.H. Leijenaar, P. Lambin, A. J. Cole, and S. Jain. Radiomics for Predicting Lung Cancer Outcomes Following Radiotherapy: A Systematic Review. *Clinical Oncology*, 34(3):e107–e122, 2022.

[155] Gaia Ninatti, Margarita Kirienko, Emanuele Neri, Martina Sollini, and Arturo Chiti. Imaging-based prediction of molecular therapy targets in NSCLC by radiogenomics and AI approaches: A systematic review. *Diagnostics*, 10(6), 2020.

[156] Dongdong Mei, Yan Luo, Yan Wang, and Jingshan Gong. CT texture analysis of lung adenocarcinoma: Can Radiomic features be surrogate biomarkers for EGFR mutation statuses. *Cancer Imaging*, 18(1):1–9, 2018.

[157] Wei Zhao, Yuzhi Wu, Ya'nan Xu, Yingli Sun, Pan Gao, Mingyu Tan, Weiling Ma, Cheng Li, Liang Jin, Yanqing Hua, Jun Liu, and Ming Li. The Potential of Radiomics Nomogram in Non-invasively Prediction of Epidermal Growth Factor Receptor Mutation Status and Subtypes in Lung Adenocarcinoma. *Frontiers in Oncology*, 9:1485, 2020.

[158] Nguyen Quoc Khanh Le, Quang Hien Kha, Van Hiep Nguyen, Yung Chieh Chen, Sho Jen Cheng, and Cheng Yu Chen. Machine learning-based radiomics signatures for egfr and kras mutations prediction in non-small-cell lung cancer. *International Journal of Molecular Sciences*, 22(17):9254, 2021.

[159] Guojin Zhang, Yuntai Cao, Jing Zhang, Jialiang Ren, Zhiyong Zhao, Xiaodi Zhang, Shenglin Li, Liangna Deng, and Junlin Zhou. Predicting EGFR mutation status in lung adenocarcinoma: development and validation of a computed tomography-based radiomics signature. *American journal of cancer research*, 11(2):546–560, 2021.

[160] Federico Cucchiara, Marzia Del Re, Simona Valleggi, Chiara Romei, Iacopo Petrini, Maurizio Lucchesi, Stefania Crucitta, Eleonora Rofi, Annalisa De Liperi, Antonio Chella, Antonio Russo, and Romano Danesi. Integrating Liquid Biopsy and Radiomics to Monitor Clonal Heterogeneity of EGFR-Positive Non-Small Cell Lung Cancer. *Frontiers in Oncology*, 10:1–8, 2020.

[161] Bardia Yousefi, Michael J. LaRiviere, Eric A. Cohen, Thomas H. Buckingham, Stephanie S. Yee, Taylor A. Black, Austin L. Chien, Peter Noël, Wei Ting Hwang, Sharyn I. Katz, Charu Aggarwal, Jeffrey C. Thompson, Erica L. Carpenter, and Despina Kontos. Combining radiomic phenotypes of non-small cell lung cancer with liquid biopsy data may improve prediction of response to EGFR inhibitors. *Scientific Reports*, 11(1):1–13, 2021.

[162] Jooae Choe, Sang Min Lee, Wooil Kim, Kyung Hyun Do, Seonok Kim, Sehoon Choi, and Joon Beom Seo. CT radiomics-based prediction of anaplastic lymphoma kinase and epidermal growth factor receptor mutations in lung adenocarcinoma. *European Journal of Radiology*, 139:109710, 2021.

[163] Lan Song, Zhenchen Zhu, Li Mao, Xiuli Li, Wei Han, Huayang Du, Huanwen Wu, Wei Song, and Zhengyu Jin. Clinical, Conventional CT and Radiomic Feature-Based Machine Learning Models for Predicting ALK Rearrangement Status in Lung Adenocarcinoma Patients. *Frontiers in Oncology*, 10:1–14, 2020.

[164] Xin Tang, Yuan Li, Wei Feng Yan, Wen Lei Qian, Tong Pang, You Ling Gong, and Zhi Gang Yang. Machine Learning-Based CT Radiomics Analysis for Prognostic Prediction in Metastatic Non-Small Cell Lung Cancer Patients With EGFR-T790M Mutation Receiving Third-Generation EGFR-TKI Osimertinib Treatment. *Frontiers in Oncology*, 11:1–10, 2021.

[165] Xinguan Yang, Xiao Dong, Jiao Wang, Weiwei Li, Zhuoran Gu, Dashan Gao, Nanshan Zhong, and Yubao Guan. Computed Tomography-Based Radiomics Signature: A Potential Indicator of Epidermal Growth Factor Receptor Mutation in Pulmonary Adenocarcinoma Appearing as a Subsolid Nodule. *The Oncologist*, 24(11):e1156–e1164, 2019.

[166] Hailin Li, Rui Zhang, Siwen Wang, Mengjie Fang, Yongbei Zhu, Zhenhua Hu, Di Dong, Jingyun Shi, and Jie Tian. CT-Based Radiomic Signature as a Prognostic Factor in Stage IV ALK-Positive Non-small-cell Lung Cancer Treated With TKI Crizotinib: A Proof-of-Concept Study. *Frontiers in Oncology*, 10:1–9, 2020.

[167] Mattea L. Welch, Chris McIntosh, Benjamin Haibe-Kains, Michael F. Milosevic, Leonard Wee, Andre Dekker, Shao Hui Huang, Thomas G. Purdie, Brian O'Sullivan, Hugo J.W.L. Aerts, and David A. Jaffray. Vulnerabilities of radiomic signature development: The need for safeguards. *Radiotherapy and Oncology*, 130:2–9, 2019.

[168] Ivan Zhovannik, J. Bussink, Alberto Traverso, Zhenwei Shi, Petros Kalendralis, Leonard Wee, A. Dekker, Rianne Fijten, and René Monshouwer. Learning from scanners: Bias reduction and feature correction in radiomics. *Clinical and Translational Radiation Oncology*, 19:33–38, 2019.

[169] Gregory R. Lee, Ralf Gommers, Filip Waselewski, Kai Wohlfahrt, and Aaron O'Leary. PyWavelets: A Python package for wavelet analysis. *The Journal of Open Source Software*, 4(36):1237, 2019.

[170] Thibaud P. Coroller, Patrick Grossmann, Ying Hou, Emmanuel Rios Velazquez, Ralph T.H. Leijenaar, Gretchen Hermann, Philippe Lambin, Benjamin Haibe-Kains, Raymond H. Mak, and Hugo J.W.L. Aerts. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiotherapy and Oncology*, 114(3):345–350, 2015.

[171] 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*, 6:1–10, 2016.

[172] José Raniery Ferreira Junior, Marcel Koenigkam-Santos, Federico Enrique Garcia Cipriano, Alexandre Todorovic Fabro, and Paulo Mazzoncini de Azevedo-Marques. Radiomics-based features for pattern recognition of lung cancer histopathology and metastases. *Computer Methods and Programs in Biomedicine*, 159:23–30, 2018.

[173] R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, pages https://www.R–project.org, 2019.

[174] Lawrence I-kuei Lin. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, 45(1):255–268, 1989.

[175] Huiman X. Barnhart, Michael Haber, and Jingli Song. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics*, 58(4):1020–1027, 2002.

[176] Janna E. van Timmeren, Ralph T.H. 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):361–365, 2016.

[177] Ruben T.H.M. Larue, Lien Van De Voorde, Janna E. van Timmeren, Ralph T.H. Leijenaar, Maaike Berbée, Meindert N. Sosef, Wendy M.J. Schreurs, Wouter van Elmpt, and Philippe Lambin. 4DCT imaging to assess radiomics feature stability: An investigation for thoracic cancers. *Radiotherapy and Oncology*, 125(1):147–153, 2017.

[178] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995.

[179] Lisa Rinaldi, Simone P. De Angelis, Sara Raimondi, Stefania Rizzo, Cristiana Fanciullo, Cristiano Rampinelli, Manuel Mariani, Alessandro Lascialfari, Marta Cremonesi, Roberto Orecchia, Daniela Origgi, and Francesca Botta. Reproducibility of radiomic features in CT images of NSCLC patients: an integrative analysis on the impact of acquisition and reconstruction parameters. *European Radiology Experimental*, 6(1), 2022.

[180] K. Buch, B. Li, M. M. Qureshi, H. Kuno, S. W. Anderson, and O. Sakai. Quantitative assessment of variation in CT parameters on texture features: Pilot study using a nonanatomic phantom. *American Journal of Neuroradiology*, 38(5):981–985, 2017.

[181] Roberto Berenguer, María Del Rosario Pastor-Juan, Jesús Canales-Vázquez, Miguel Castro-García, María Victoria Villas, Francisco Mansilla Legorburo, and Sebastià Sabater. Radiomics of CT features may be nonreproducible and redundant: Influence of CT acquisition parameters. *Radiology*, 288(2):407–415, 2018.

[182] Nastaran Emaminejad, Muhammad Wasil Wahi-Anwar, Grace Hyun J. Kim, William Hsu, Matthew Brown, and Michael McNitt-Gray. Reproducibility of lung nodule radiomic features: Multivariable and univariable investigations that account for interactions between CT acquisition and reconstruction parameters. *Medical Physics*, 48(6):2906–2919, 2021.

[183] Mathias Meyer, James Ronald, Federica Vernuccio, Rendon C. Nelson, Juan Carlos Ramirez-Giraldo, Justin Solomon, Bhavik N. Patel, Ehsan Samei, and Daniele Marin. Reproducibility of CT radiomic features within the same patient: Influence of radiation dose and CT reconstruction settings. *Radiology*, 293(3):583–591, 2019.

[184] Gongbo Liang, Sajjad Fouladvand, Jie Zhang, Michael A. Brooks, Nathan Jacobs, and Jin Chen. GANai: Standardizing CT Images using Generative Adversarial Network with Alternative Improvement. *2019 IEEE International Conference on Healthcare Informatics, ICHI 2019*, 2019.

[185] Muhammad Shafiq-ul Hassan, Geoffrey G. Zhang, Dylan C. Hunt, Kujtim Latifi, Ghanim Ullah, Robert J. Gillies, and Eduardo G. Moros. Accounting for reconstruction kernel-induced variability in CT radiomic features using noise power spectra. *Journal of Medical Imaging*, 5(01):1, 2017.

[186] Dennis Mackin, Rachel Ger, Skylar Gay, Cristina Dodge, Lifei Zhang, Jinzhong Yang, Aaron Kyle Jones, and Laurence Court. Matching and Homogenizing Convolution Kernels for Quantitative Studies in Computed Tomography. *Investigative Radiology*, 54(5):288–295, 2019.

[187] Larry A Dewerd and Michael Kissick. *The Phantoms of Medical and Health Physics: Devices for Research and Development*. Springer New York, 2014.

[188] Conor K. McGarry, Lesley J. Grattan, Aoife M. Ivory, Francesca Leek, Gary P. Liney, Yang Liu, Piero Miloro, Robba Rai, Andrew P. Robinson, Albert J. Shih, Bajram Zeqiri, and Catharine H. Clark. Tissue mimicking materials for imaging and therapy phantoms: A review. *Physics in Medicine and Biology*, 65(23), 2020.

[189] P. Lo, S. Young, H. J. Kim, M. S. Brown, and M. F. McNitt-Gray. Variability in CT lung-nodule quantification: Effects of dose reduction and reconstruction methods on density and texture based features: Effects. *Medical Physics*, 43(8):4854–4865, 2016.

[190] Abhishek Midya, Jayasree Chakraborty, Mithat Gönen, Richard K. G. Do, and Amber L. Simpson. Influence of CT acquisition and reconstruction parameters on radiomic feature reproducibility. *Journal of Medical Imaging*, 5(01):1, 2018.

[191] Hyeongmin Jin and Jong Hyo Kim. Evaluation of Feature Robustness Against Technical Parameters in CT Radiomics: Verification of Phantom Study with Patient Dataset. *Journal of Signal Processing Systems*, 92(3):277–287, 2020.

[192] Valerio Nardone, Alfonso Reginelli, Cesare Guida, Maria Paola Belfiore, Michelangelo Biondi, Maria Mormile, Fabrizio Banci Buonamici, Eugenio Di Giorgio, Marco Spadafora, Paolo Tini, Roberta Grassi, Luigi Pirtoli, Pierpaolo Correale, Salvatore Cappabianca, and Roberto Grassi. Delta-radiomics increases multicentre reproducibility: a phantom study. *Medical Oncology*, 37(5):1–7, 2020.

[193] A. K. Jha, S. Mithun, V. Jaiswar, U. B. Sherkhane, N. C. Purandare, K. Prabhash, V. Rangarajan, A. Dekker, L. Wee, and A. Traverso. Repeatability and reproducibility study of radiomic features on a phantom and human cohort. *Scientific Reports*, 11(1):1–12, 2021.

[194] Alejandra Valladares, Thomas Beyer, and Ivo Rausch. Physical imaging phantoms for simulation of tumor heterogeneity in PET, CT, and MRI: An overview of existing designs. *Medical Physics*, 47(4):2023–2037, 2020.

[195] Ehsan Samei, Jocelyn Hoye, Yuese Zheng, Justin B Solomon, and Daniele Marin. Design and fabrication of heterogeneous lung nodule phantoms for assessing the accuracy and variability of measured texture radiomics features in CT. *Journal of medica*, 6(2):021606, 2019.

[196] Paul Jahnke, Felix R.P. Limberg, Andreas Gerbl, Gracia L.Ardila Pardo, Victor P.B. Braun, Bernd Hamm, and Michael Scheel. Radiopaque three-dimensional printing: A method to create realistic CT phantoms. *Radiology*, 282(2):569–575, 2017.

[197] 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 18F-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.

[198] Linda Bianchini, Francesca Botta, Daniela Origgi, Stefania Rizzo, Manuel Mariani, Paul Summers, Pablo García-Polo, Marta Cremonesi, and Alessandro Lascialfari. PETER PHAN: An MRI phantom for the optimisation of radiomic studies of the female pelvis. *Physica Medica*, 71:71–81, 2020.

[199] Linda Bianchini, João Santinha, Nuno Loução, Mário Figueiredo, Francesca Botta, Daniela Origgi, Marta Cremonesi, Enrico Cassano, Nikolaos Papanikolaou, and Alessandro Lascialfari. A multicenter study on radiomic features from T2-weighted images of a customized MR pelvic phantom setting the basis for robust radiomic models in clinics. *Magnetic Resonance in Medicine*, 85(3):1713–1726, 2021.

[200] Robba Rai, Lois C. Holloway, Carsten Brink, Matthew Field, Rasmus L. Christiansen, Yu Sun, Michael B. Barton, and Gary P. Liney. Multicenter evaluation of MRI-based radiomic features: A phantom study. *Medical Physics*, 47(7):3054–3063, 2020.

[201] Lawrence R. Parks. Cross-linked sodium polyacrylate absorbent. *US Patent Document*, (19), 1981.

[202] David L. Raunig, Lisa M. McShane, Gene Pennello, Constantine Gatsonis, Paul L. Carson, James T. Voyvodic, Richard L. Wahl, Brenda F. Kurland, Adam J. Schwarz, Mithat Gönen, Gudrun Zahlmann, Marina V. Kondratovich, Kevin O'Donnell, Nicholas Petrick, Patricia E. Cole, Brian Garra, and Daniel C. Sullivan. Quantitative imaging biomarkers: A review of statistical methods for technical performance assessment. *Statistical Methods in Medical Research*, 24(1):27–67, 2015.

[203] 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.

[204] Lisa Rinaldi, Federico Pezzotta, Tommaso Santaniello, Paolo De Marco, Linda Bianchini, Daniela Origgi, Marta Cremonesi, Paolo Milani, Manuel Mariani, and Francesca Botta. HeLLePhant: A phantom mimicking non-small cell lung cancer for texture analysis in CT images. *Physica Medica*, 97(March):13–24, 2022.

[205] Simon Lennartz, Aileen O'Shea, Anushri Parakh, Thorsten Persigehl, Bettina Baessler, and Avinash Kambadakone. Robustness of dual-energy CT-derived radiomic features across three different scanner types. *European Radiology*, 32(3):1959–1970, 2022.

[206] Yanjing Li, Meral Reyhan, Yin Zhang, Xiao Wang, Jinghao Zhou, Yang Zhang, Ning J Yue, and Ke Nie. The impact of phantom design and material-dependence on repeatability and reproducibility of CT-based radiomics features. *Medical Physics*, 49(3):1648–1659, 2022.

[207] Xueqing Peng, Shuyi Yang, Lingxiao Zhou, Yu Mei, Lili Shi, Rengyin Zhang, Fei Shan, and Lei Liu. Repeatability and Reproducibility of Computed Tomography Radiomics for Pulmonary Nodules: A Multicenter Phantom Study. *Investigative Radiology*, 57(4):242–253, 2022.

[208] Colien Hazelaar, Maureen Van Eijnatten, Max Dahele, Jan Wolff, Tymour Forouzanfar, Ben Slotman, and Wilko F.A.R. Verbakel. Using 3D printing techniques to create an anthropomorphic thorax phantom for medical imaging purposes. *Medical Physics*, 45(1):92–100, 2018.

[209] Dayeong Hong, Sangwook Lee, Guk Bae Kim, Sang Min Lee, Namkug Kim, and Joon Beom Seo. Development of a CT imaging phantom of anthromorphic lung using fused deposition modeling 3D printing. *Medicine (United States)*, 99(1), 2020.

[210] Kai Mei, Michael Geagan, Leonid Roshkovan, Harold I. Litt, Grace J. Gang, Nadav Shapira, J. Webster Stayman, and Peter B. Noël. Three-dimensional printing of patient-specific lung phantoms for CT imaging: Emulating lung tissue with accurate attenuation profiles and textures. *Medical Physics*, 49(2):825–835, 2022.

[211] Dzung L Pham, Chenyang Xu, and Jerry L Prince. Current methods in medical image segmentation. *Annu Rev Biomed*, 2:315–337, 2000.

[212] Erik Smistad, Thomas L. Falch, Mohammadmehdi Bozorgi, Anne C. Elster, and Frank Lindseth. Medical image segmentation on GPUs - A comprehensive review. *Medical Image Analysis*, 20(1):1–18, 2015.

[213] Aarish Shafi Dar and Devanand Padha. Medical Image Segmentation A Review of Recent Techniques, Advancements and a Comprehensive Comparison. *International Journal of Computer Sciences and Engineering*, 7(7):114–124, 2019.

[214] Intisar Rizwan I Haque and Jeremiah Neubert. Deep learning approaches to biomedical image segmentation. *Informatics in Medicine Unlocked*, 18:100297, 2020.

[215] Alex P. Zijdenbos, Benoit M. Dawant, Richard A. Margolin, and Andrew C. Palmer. Morphometric Analysis of White Matter Lesions in MR Images: Method and Validation. *IEEE Transactions on Medical Imaging*, 13(4):716–724, 1994.

[216] Raimundo Real and Juan M. Vargas. The probabilistic basis of Jaccard's index of similarity. *Systematic Biology*, 45(3):380–385, 1996.

[217] D. P. Huttenlocher, W. J. Rucklidge, and G. A. Klanderman. Comparing images using the Hausdorff distance under translation, 1992.

[218] Cindy Xue, Jing Yuan, Gladys G. Lo, Amy T.Y. Chang, Darren M.C. Poon, Oi Lei Wong, Yihang Zhou, and Winnie C.W. Chu. Radiomics feature reliability assessed by intraclass correlation coefficient: A systematic review. *Quantitative Imaging in Medicine and Surgery*, 11(10):4431–4460, 2021.

[219] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[220] Tim McInerney and Demetri Terzopoulos. Deformable models in medical image analysis: A survey. *Medical Image Analysis*, 1(2):91–108, 1996.

[221] Rolf Adams and Leanne Bischof. Seeded Region Growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641–647, 1994.

[222] Regina Pohle and Klaus D. Toennies. Segmentation of medical images using adaptive region growing. *Medical Imaging 2001: Image Processing*, 4322:1337–1346, 2001.

[223] Jamshid Dehmeshki, Hamdan Amin, Manlio Valdivieso, and Xujiong Ye. Segmentation of pulmonary nodules in thoracic CT scans: A region growing approach. *IEEE Transactions on Medical Imaging*, 27(4):467–480, 2008.

[224] P. K. Sahoo, S. Soltani, and A. K.C. Wong. A survey of thresholding techniques. *Computer Vision, Graphics and Image Processing*, 41(2):233–260, 1988.

[225] S. Sivakumar and C. Chandrasekar. Lung nodule segmentation through unsupervised clustering models. *Procedia Engineering*, 38:3064–3073, 2012.

[226] Juan Eugenio Iglesias and Mert R. Sabuncu. Multi-atlas segmentation of biomedical images: A survey. *Medical Image Analysis*, 24(1):205–219, 2015.

[227] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[228] Carlos E. Cardenas, Jinzhong Yang, Brian M. Anderson, Laurence E. Court, and Kristy B. Brock. Advances in Auto-Segmentation. *Seminars in Radiation Oncology*, 29(3):185–197, 2019.

[229] Atsushi Teramoto, Ayumi Yamada, Tetsuya Tsukamoto, Kazuyoshi Imaizumi, Hiroshi Toyama, Kuniaki Saito, and Hiroshi Fujita. *Deep Learning in Medical Image Analysis. Challenges and Applications*, volume 1213. 1 edition, 2020.

[230] Xiangbin Liu, Liping Song, Shuai Liu, and Yudong Zhang. A review of deep-learning-based medical image segmentation methods. *Sustainability (Switzerland)*, 13(3):1–29, 2021.

[231] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2017.

[232] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597*, 2015.

[233] Simon A.A. Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus H. Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A Probabilistic U-Net for Segmentation of Ambiguous Images. *arXiv:1806.05034*, 2019.

[234] Nahian Siddique, Sidike Paheding, Colin P. Elkin, and Vijay Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 9:82031–82057, 2021.

[235] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432. Springer International Publishing, 2016.

[236] Fausto Milletari, Nassir Navab, and Seyed Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, pages 565–571, 2016.

[237] Francesca Lizzi, Abramo Agosti, Francesca Brero, Raffaella Fiamma Cabini, Maria Evelina Fantacci, Silvia Figini, Alessandro Lascialfari, Francesco Laruina, Piernicola Oliva, Stefano Piffer, Ian Postuma, Lisa Rinaldi, Cinzia Talamonti, and Alessandra Retico. Quantification of pulmonary involvement in COVID-19 pneumonia by means of a cascade of two U-nets : training and assessment on multiple datasets using different annotation criteria. *International Journal of Computer Assisted Radiology and Surgery*, 17:229–237, 2022.

[238] Abhishek Shivdeo, Rohit Lokwani, Viraj Kulkarni, Amit Kharat, and Aniruddha Pant. Comparative Evaluation of 3D and 2D Deep Learning Techniques for Semantic Segmentation in CT Scans. *2021 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, pages 1–8, 2021.

[239] Fabian Isensee, Paul F. Jaeger, Simon A.A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[240] Philippe Giraud, Sabine Elles, Sylvie Helfre, Yann De Rycke, Vincent Servois, Marie France Carette, Claude Alzieu, Pierre Yves Bondiau, Bernard Dubray, Emmanuel Touboul, Martin Housset, Jean Claude

Rosenwald, and Jean Marc Cosset. Conformal radiotherapy for lung cancer: Different delineation of the gross tumor volume (GTV) by radiologists and radiation oncologists. *Radiotherapy and Oncology*, 62(1):27–36, 2002.

[241] Ying Wang, Rob J. Van Klaveren, Hester J. Van Der Zaag-Loonen, Geertruida H. De Bock, Hester A. Gietema, Ming Xu Dong, Anne L.M. Leusveld, Harry J. De Koning, Ernst T. Scholten, Johny Verschakelen, Mathias Prokop, and Matthijs Oudkerk. Effect of nodule characteristics on variability of semiautomated volume measurements in pulmonary nodules detected in a lung cancer screening program. *Radiology*, 248(2):625–631, 2008.

[242] Yuhua Gu, Virendra Kumar, Lawrence O. Hall, Dmitry B. Goldgof, Ching Yen Li, René Korn, Claus Bendtsen, Emmanuel Rios Velazquez, Andre Dekker, Hugo Aerts, Philippe Lambin, Xiuli Li, Jie Tian, Robert A. Gatenby, and Robert J. Gillies. Automated delineation of lung tumors from CT images using a single click ensemble segmentation approach. *Pattern Recognition*, 46(3):692–702, 2013.

[243] Christian F. Baumgartner, Kerem C. Tezcan, Krishna Chaitanya, Andreas M. Hötker, Urs J. Muehlematter, Khoschy Schawkat, Anton S. Becker, Olivio Donati, and Ender Konukoglu. PHiSeg: Capturing Uncertainty in Medical Image Segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11765 LNCS:119–127, 2019.

[244] Wutian Gan, Hao Wang, Hengle Gu, Yanhua Duan, Yan Shao, Hua Chen, Aihui Feng, Ying Huang, Xiaolong Fu, Yanchen Ying, Hong Quan, and Zhiyong Xu. Automatic segmentation of lung tumors on CT images based on a 2D & 3D hybrid convolutional neural network. *The British Journal of Radiology*, 94(1126):20210038, 2021.

[245] J. Neyman and E. S. Pearson. On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I. *Biometrika*, 20A(1/2):175–240, 1928.

[246] Quang H. Vuong. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57(2):307–333, 1989.

[247] J. P. Fine. Comparing nonnested Cox models. *Biometrika*, 89(3):635–647, 2002.

[248] Jue Jiang, Yu Chi Hu, Chia Ju Liu, Darragh Halpenny, Matthew D. Hellmann, Joseph O. Deasy, Gig Mageras, and Harini Veeraraghavan. Multiple resolution residually connected feature streams for automatic lung tumor segmentation from CT images. *IEEE Transactions on Medical Imaging*, 38(1):134–144, 2019.

[249] Shuchao Pang, Anan Du, Mehmet A. Orgun, Zhenmei Yu, Yunyun Wang, Yan Wang, and Guanfeng Liu. CTumorGAN: a unified framework for automatic computed tomography tumor segmentation. *European Journal of Nuclear Medicine and Molecular Imaging*, 47(10):2248–2268, 2020.

[250] Fuli Zhang, Qiusheng Wang, and Haipeng Li. Automatic Segmentation of the Gross Target Volume in Non-Small Cell Lung Cancer Using a Modified Version of ResNet. *Technology in Cancer Research and Treatment*, 19, 2020.

[251] Jiancheng Yang, Xiaoyang Huang, Yi He, Jingwei Xu, Canqian Yang, Guozheng Xu, and Bingbing Ni. Reinventing 2D Convolutions for 3D Images. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3009–3018, 2021.

[252] Xiangming Zhao, Laquan Li, Wei Lu, and Shan Tan. Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network. *Physics in Medicine and Biology*, 64(1):1–35, 2019.

[253] Deepak M. Kalaskar. *3D Printing in Medicine*. Woodhead Publishing, an imprint of Elsevier, 2017.

[254] Zachary H. Levine, H. Heather Chen-Mayer, Adele P. Peskin, and Adam L. Pintar. Comparison of one-dimensional and volumetric computed tomography measurements of injected-water phantoms. *Journal of Research of the National Institute of Standards and Technology*, 122(36):1–9, 2017.

[255] E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer*, 45(2):228–247, 2009.

[256] Vladimir Vezhnevets and Vadim Konouchine. GrowCut- Interactive multi-label N-D image segmentation by cellular automata. *GraphiCon 2005 - International Conference on Computer Graphics and Vision, Proceedings*, 2005.

[257] Linagjia Zhu, Ivan Kolesov, Yi Gao, Ron Kikinis, and Allen Tannenbaum. An Effective Interactive Medical Image Segmentation Method Using Fast GrowCut. *MICCAI Workshop on Interactive Medical Image Computing*, 2014.

[258] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016.

[259] Charu C. Aggarwal. *Neural Networks and Deep learning*. Springer, Cham, 2018.

[260] Michal Drozdzal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*, pages 179–187. Springer International Publishing, 2016.

[261] Shruti Jadon. A survey of loss functions for semantic segmentation. *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2020*, 2020.

[262] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[263] Daniele Ravi, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang Zhong Yang. Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1):4–21, 2017.

[264] Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1), 2019.

[265] Matthias Feurer and Frank Hutter. *Hyperparameter Optimization. In: Hutter F., Kotthoff L., Vanschoren J. (eds) Automated Machine Learning*. Springer International Publishing, 2019.

[266] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[267] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2):336–359, 2020.

List of publications

- Lizzi Francesca, Postuma Ian, Brero Francesca, Cabini Raffaella Fiamma, Fantacci Maria Evelina, Lascialfari Alessandro, Oliva Piernicola, Rinaldi Lisa, and Alessandra Retico. **Quantification of pulmonary involvment in COVID-19 pneumonia: an upgrade of the LungQuant software for lung CT segmentation**. *Submitted to The European Physical Journal - Plus.*

- Lisa Rinaldi, Federico Pezzotta, Tommaso Santaniello, Paolo De Marco, Linda Bianchini, Daniela Origgi, Marta Cremonesi, Paolo Milani, Manuel Mariani, and Francesca Botta. **HeLLePhant: a phantom mimicking non-small cell lung cancer for texture analysis in CT images**. *Physica Medica*, 97:13-24, 2022. https://doi.org/10.1016/j.ejmp.2022.03.010.

- Lisa Rinaldi, Simone P. De Angelis, Sara Raimondi, Stefania Rizzo, Cristiana Fanciullo, Cristiano Rampinelli, Manuel Mariani, Alessandro Lascialfari, Marta Cremonesi, Roberto Orecchia, Daniela Origgi, and Francesca Botta. **Reproducibility of radiomic features in CT images of NSCLC patients: an integrative analysis on the impact of acquisition and reconstruction parameters**. *European Radiology Experimental*, 6(1):2, 2022. https://doi.org/10.1186/s41747-021-00258-6.

- Lizzi Francesca, Agosti Abramo, Brero Francesca, Cabini Raffaella Fiamma, Fantacci Maria Evelina, Figini Silvia, Lascialfari Alessandro, Laruina Francesco, Oliva Piernicola, Piffer Stefano, Postuma Ian, Rinaldi Lisa, Talamonti Cinzia, and Alessandra Retico. **Quantification of pulmonary involvement in COVID-19 pneumonia by means of a cascade of two U-nets: training and assessment on multiple datasets using different annotation criteria**. *International Journal of Computer Assisted Radiology and Surgery*, 17(2): 229-237, 2021. https://doi.org/10.1007/s11548-021-02501-2.

- Lizzi Francesca, Brero Francesca, Cabini Raffaella, Fantacci Maria Evelina, Piffer Stefano, Postuma Ian, Rinaldi Lisa and Retico Alessandra. **Making Data Big for a Deep-learning Analysis: Aggregation of Public COVID-19 Datasets of Lung Computed Tomography Scans**. *In Proceedings of the 10th International Conference on Data Science, Technology and Applications – DATA*, pages 316-321. 2021. https://10.5220/0010584403160321.

- Corso Federica, Giulia Tini, Giuliana Lo Presti, Noemi Garau, Simone P. De Angelis, Federica Bellerba, Lisa Rinaldi, Francesca Botta, Stefania Rizzo, Daniela Origgi, Chiara Paganelli, Marta Cremonesi, Cristiano Rampinelli, Massimo Bellomi, Luca Mazzarella, Pier G. Pelicci, Sara Gandini, and Sara Raimondi. **The Challenge of Choosing the Best Classification Method in Radiomic Analyses: Recommendations and Applications to Lung Cancer CT Images**. *Cancers*,13 (12): 3088, 202. https://doi.org/10.3390/cancers13123088.

- Botta Francesca, Sara Raimondi, Lisa Rinaldi, Federica Bellerba, Federica Corso, Vincenzo Bagnardi, Daniela Origgi, Rocco Minelli, Giovanna Pitoni, Francesco Petrella, Lorenzo Spaggiari, Alessio G. Morganti, Filippo Del Grande, Massimo Bellomi, and Stefania Rizzo. **Association of a CT-Based Clinical and Radiomics Score of Non-Small Cell Lung Cancer (NSCLC) with Lymph Node Status and Overall Survival**. *Cancers*, 12 (6): 1432, 2020. https://10.3390/cancers12061432.

Published abstracts and posters:

- Ferrante Matteo, Rinaldi Lisa, Hu Xiaobin, Lo Presti Giuliana, Botta Francesca, Rizzo Stefania, Volpe Stefania, Shi Kuangyu, Origgi Daniela. **C-15419 - Automatic segmentation of lung cancer on CT images based on a self-adapting deep neural network**. Poster for the *European Congress of Radiology, 2022*.

- Lisa Rinaldi, Simone P. De Angelis, Sara Raimondi, Daniela Origgi, Stefania Rizzo, Cristiana Fanciullo, Cristiano Rampinelli, Manuel Mariani, Alessandro Lascialfari, Massimo Bellomi, Marta Cremonesi, Francesca Botta. **OD126 - Reproducibility of radiomic features in CT images of NSCLC patients**. *Physica Medica*, Volume 92, Supplement, p. S115, 2021. ISSN 1120-1797, https://doi.org/10.1016/S1120-1797(22)00243-5.

- Lisa Rinaldi, Federico Pezzotta, Tommaso Santaniello, Paolo De Marco, Daniela Origgi, Marta Cremonesi, Manuel Mariani, Alessandro Lascialfari, Paolo Milani, Francesca Botta. **OD125 - A prototype of heterogenous insert simulating lung lesions for quantitative texture analysis in CT acquisitions**. *Physica Medica*, Volume 92, Supplement, p. S114-S115, 2021. ISSN 1120-1797, https://doi.org/10.1016/S1120-1797(22)00242-3.

- Lisa Rinaldi, Simone P. De Angelis, Sara Raimondi, Daniela Origgi, Alessandro Lascialfari, Manuel Mariani, Marta Cremonesi, Massimo Bellomi, Francesca Botta. **RPS 113-5 Integrated investigation of radiomic features reproducibility in NSCLC patients: the impact of scanner, x-ray tube voltage and strength of reconstruction algorithms in contrast-enhanced CT images**. ECR 2021 Book of Abstracts. Insights Imaging 12, 75, p. A22, 2021. https://doi.org/10.1186/s13244-021-01014-5