UNIVERSITÀ DEGLI STUDI DI PADOVA

University of Padua

Department of Information Engineering

Ph.D. Course in Information Engineering

Curriculum: Bioengineering

SERIES: XXXV

Radiomics: from medical imaging to precision medicine. Development of advanced tools for texture analysis and implementation of AI techniques for prognostic modelling in oncology

Thesis written with the financial contribution of Veneto Institute of Oncology – IOV IRCCS

Coordinator: Prof. Andrea Neviani

Supervisor: Prof. Alessandra Bertoldo

Co-Supervisor: Dott.ssa Marta Paiusco

Ph.D. candidate: Andrea Bettinelli

Thesis Abstract

In the biomedical field, radiomics is a highly promising and prolific discipline that explores the quantification, by means of *radiomic features*, of biomedical imaging for predictive and prognostic purposes. The identification of radiomic features that could act as imaging biomarkers and their integration with clinical and molecular data will lead to the creation of innovative tools for precision medicine. The main hamper to its clinical translation consists in the low reproducibility of radiomic studies, linked both to imaging data and software aspects. Shared standardization guidelines, references, and benchmarking tools are essential to overcome this issue.

In this field, the aim of this PhD research project was threefold, namely the development of a feature-extraction tool focused on standardization and reproducibility aspects, its validation and comparison with literature's alternatives, and its employment for radiomic clinical research.

After an introduction to the topic, in Chapter 2 the work of development and standardization of the tool, called S-IBEX, will be presented, together with my own personal contributions to a global standardization effort of convolutional filters' implementation and application within radiomic.

Subsequently, Chapter 3 will describe the comparison work between S-IBEX and other standardized software programs available in the literature, an analysis that required the design of a novel benchmarking approach based on new specific tools. The investigation confirmed the high standardization level achieved across programs but also identified discrepancies that should be addressed to ensure a transparent interchangeability of the radiomic software.

In Chapter 4, the usage of S-IBEX in three clinical research studies will be discussed. S-IBEX-derived radiomic features allowed to capture prognostic information contained within biomedical images in the settings of 1) locally advanced breast cancer for the prediction of 5-year overall survival, 2) prostate cancer for the prediction of biochemical recurrence and 3) hepatocellular carcinoma for the non-invasive assessment of vascular invasion.

Sommario

In ambito biomedico, la radiomica è un'area di ricerca altamente promettente e prolifica, dedita alla quantificazione, per mezzo delle cosiddette 'feature radiomiche', dell'imaging biomedicale allo scopo di creare modelli predittivi e prognostici di supporto alla decisione clinica. L'identificazione di feature che fungeranno da biomarcatori e la loro integrazione con dati clinici e molecolari porterà alla creazione dei futuri strumenti utili alla medicina di precisione. Ad oggi, il principale ostacolo a questa traslazione è costituito dalla bassa riproducibilità degli studi di radiomica, riproducibilità legata sia al dato di imaging che al software di analisi impiegato. Linee guida standardizzate, riferimenti condivisi e strumenti di benchmark sono elementi essenziali per superare questo scoglio.

In quest'ambito, l'obbiettivo del progetto di dottorato è stato triplice, ossia lo sviluppo di un software per l'estrazione di feature radiomiche che puntasse sulla standardizzazione e riproducibilità dei risultati, quindi la sua validazione e confronto con le alternative presenti in letteratura e infine e il suo impiego per progetti di ricerca clinica.

Dopo un'introduzione all'argomento, nel secondo capitolo verrà presentato il lavoro di realizzazione e standardizzazione dello strumento, chiamato S-IBEX, insieme ai contributi dell'autore ad un'iniziativa globale di standardizzazione riguardante l'implementazione e l'applicazione del filtraggio convoluzionale in radiomica.

Successivamente, nel terzo capitolo, verrà esposto il lavoro di confronto tra S-IBEX e altri strumenti presenti in letteratura, confronto che ha richiesto la progettazione di un nuovo approccio di benchmark basato su nuovi strumenti creati appositamente. L'indagine ha evidenziato l'alto livello di standardizzazione raggiunto tra i software, ma ha altresì identificato discrepanze che dovranno essere affrontate per garantire l'intercambiabilità del software radiomico.

Nel quarto capitolo sarà infine presentato l'impiego di S-IBEX in tre studi di ricerca clinica. Le feature radiomiche da esso ricavate hanno permesso di quantificare l'informazione prognostica contenuta all'interno delle immagini biomedicali negli ambiti 1) del cancro della mammella in stadio localmente avanzato per la predizione della sopravvivenza a 5 anni, 2) del cancro alla prostata per la predizione della ricorrenza biochimica e 3) del carcinoma epatocellulare per la valutazione dell'invasione vascolare.

TABLE OF CONTENTS

Introduction and summary

In the medical fields, and more specifically in the oncological field, biomedical imaging plays a fundamental role in the daily clinical practice, from screening to diagnosis, staging, treatment planning, and follow-up[1].

Biomedical images can be acquired in different modalities, such as Computed Tomography (CT), Positron Emission Tomography (PET), Magnetic Resonance Imaging (MRI), Ultrasounds, Mammography, Tomosynthesis. All these modalities have been proven essential in one or more of the steps of the clinical routine and each one comes with its unique advantages[2]. Depending on the aspect we are considering, we can categorize biomedical images based on their structure (e.g., two- or three-dimensional imaging, time dependent imaging), their associated risk (usage of ionising radiation, contrast agents), or the information they convey (e.g., morphological, functional, etc). Despite this variety, most biomedical images are stored using a single digital format, the DICOM (Digital Imaging and Communications in Medicine) format, in the Picture Archive and Communication system (PACS). Thanks to the structured storage, biomedical images can be retrieved from the PACS and accessed, representing an unprecedented amount of data that can be mined to extract information and extrapolate new data-driven evidences.

In the oncological field, the conversion of biomedical images into a high-dimensional space of minable quantitative features as been defined in 2012 as "*Radiomics" by Lambin et al.*[3] while its breakthrough arrived in 2014 with the work of Aerts et al.[4], which, for the first time, showed the capabilities of radiomics with large datasets. Radiomics finds its roots in many decades of texture analysis and adds to that a high-throughput feature extraction procedure coupled with advance machine learning models. Lambin at al. proposed it as a promising potential alternative to invasive biopsy. The assumption is that the branched evolution of different clonal populations of cancer cells contributes to the intra-tumour heterogeneity seen in solid tumour[5] and is linked to physiologically different subregions (called habitats). Thus, genomic/proteomic patterns of the tumour could translate to the macroscopic scale in the phenotype captured by biomedical images. This spatial and temporal heterogeneity could hinder the accuracy of biopsies[6], as 1) biopsy outcomes are subject to the location in which the sample is collected (i.e., sampling bias) and 2) increasing the number of biopsy samples is not always a viable

option do to the invasiveness of the procedure. Instead, radiological imaging allows to sample the whole tumour volume, repeatedly over time, with little to no invasiveness. Subsequently, radiomic features can translates the information contained within biomedical images into quantitative characteristics describing the morphology, distribution of intensities values (specific of each image modality), and textural characteristic of the entire tumour volume, thus overcoming the main limitations of biopsy.

The promise of mining new prognostic information from routinary acquired diagnostic imaging, combined with its broad applicability to various imaging modalities, led radiomics to its success and drove the exponential growth in the number of radiomic publications that we currently appreciate (Figure 1.1).

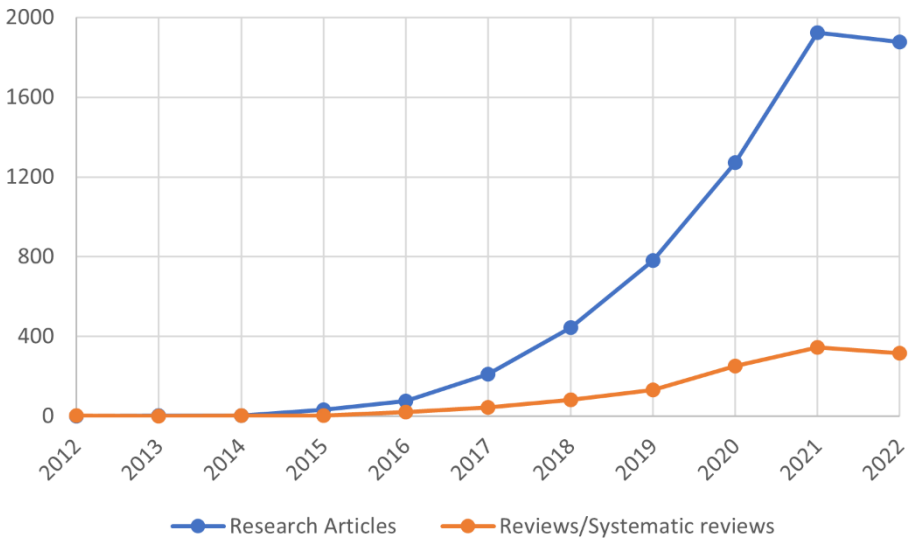

Figure 1.1. Number of publications per year retrieved on PubMed.gov with the query «radiomic OR radiomics» as of September 2022.

The figure highlights the ongoing interest in the field, and, as imaging will be more and more the fulcrum of the clinical routines, tools that will allow to automatically extract useful information from images, such as radiomics, will be needed.

The ultimate goal of radiomic is to provide physicians with robust decision support systems (e.g., for optimal therapy/treatment planning, early risk assessment), or with computer-aided detection or diagnosis (CAD) systems to be used in the clinical practice for tailoring therapy to each individual, overcoming the "one-size-fits-all" paradigm and moving towards precision medicine. Those tools do not necessarily need to only rely on image analysis, in fact the quantitative integration of multi-source data

appears to be the preferred way to build future decision support systems. Combining traditional methods (e.g., family history, environment, and lifestyle) and omics data (e.g., molecular, genomics, radiomics profiling) will allow to tailor for each patient the optimal treatment, ultimately improving healthcare and lowering costs[7].

Nevertheless, over time, caveats and limitations of radiomics also emerged[8–11], together with the awareness of a publication bias for positive results[12]. Some of the challenges have already been addressed, such as the need of a standardization of radiomic feature extraction[13], while others, such as repeatability and reproducibility have been investigated in both digital and physical phantoms, but their translation into clinical studies is difficult, since it is not straightforward to know if results of previous studies are appliable to the specific case, due to the amount of parameters that could differ.

In this thesis, I will present my contributions to the field of radiomics, which constitute the backbone of my three-year PhD project.

In Chapter 1, I will introduce radiomics by presenting the typical workflow of a clinical radiomic study, from data selection to outcome modelling, with a focus on the feature extraction part. Subsequently, I will provide the readers with a summary of the major guidelines we have in radiomics (e.g., the Radiomic Quality Score - RQS, the Image Biomarker Standardization initiative - IBSI), fundamental for a deeper understanding of this thesis's work.

In Chapter 2, I will present the feature-extraction software I have developed, called S-IBEX, and I will report the standardization steps I followed to achieve its IBSI compliance. Moreover, since S-IBEX was included in the latest IBSI chapter (for convolutional filtering standardization) I will discuss my personal contribution to the initiative itself.

In Chapter 3, I will report the work I made to compare S-IBEX to other software tools available in literature, both commercial and open-source ones. This study, that required the creation of two custom digital phantoms, allowed a more detailed comprehension of software discrepancies and showed the high standardization level achieved by S-IBEX.

Eventually, in Chapter 4, I will present the clinical works that relied on S-IBEX radiomic feature extraction. Those studies covered a wide range of cancer types and image modalities (e.g., CT, PET, Mammography), proving the high versatility of S-IBEX. Three investigations will be discussed in high

details (i.e., breast, prostate, and liver cancer), while other four, which are worth mentioning, will only be briefly presented to the reader.

Chapter 1: Introduction to Radiomics

Radiomics has been defined as the high-throughput extraction of quantitative features from medical images and their subsequent analysis[14]. The aim is to build models predictive of clinical endpoints. However, to this goal, technical studies are necessary to define the reliability boundaries of radiomic applications and define a solid knowledge base (e.g., investigating the repeatability/reproducibility of radiomic features or the generalizability of radiomic studies) on top of which investigate clinical outcomes. Therefore, it comes naturally to partition radiomic literature into two main areas: 1) one investigating technical aspect and 2) one more clinically-oriented. In the former literature branch, we find:

- Standardization studies: which aim to make the clinical studies more efficiently comparable among each other (e.g., propose benchmarks phantoms to validate software, provides guidelines on feature extraction or manuals with mathematical formulation of radiomic features, define a consensus nomenclature to remove ambiguities)[13,15,16].
- Repeatability/reproducibility studies: which aim to make the clinical studies more generalizable by studying feature repeatability/reproducibility[17–26]. The term 'repeatable' identifies features that do not change when imaged multiple times in the same subject, be that a human person (e.g., test-retest) or a suitable phantom, while the term '*reproducible'* refers to features that are stable when imaged with different image acquisition settings or by different operators (e.g., different centres), when computed with different software, or, more in general, when one or more steps prior to feature computation is changed, be that in the same subject or in different subjects[17,27].

On the other side we find the clinically oriented investigations where usually some kind of prognostication is done. The most convenient way for further categorization is to stratify clinical studies by the district of interest (e.g., prostate or breast cancer). As an example, lung tumours are the most extensively studied malignancies in the radiomic field[28–30]. Subcategorization can also be achieved when specifying the considered image modality (e.g., CT, MRI, PET) and the outcome of interest (e.g., overall survival, disease-free survival).

Despite the evident variety of clinical study, they all follow a conceptually simple workflow, made of a series of discrete steps. In Figure 1.1 we can appreciate the general workflow that constitute the backbone of a clinical radiomic study[14]: data selection, acquisition of medical imaging, feature extraction, exploratory analysis, and the modelling part.

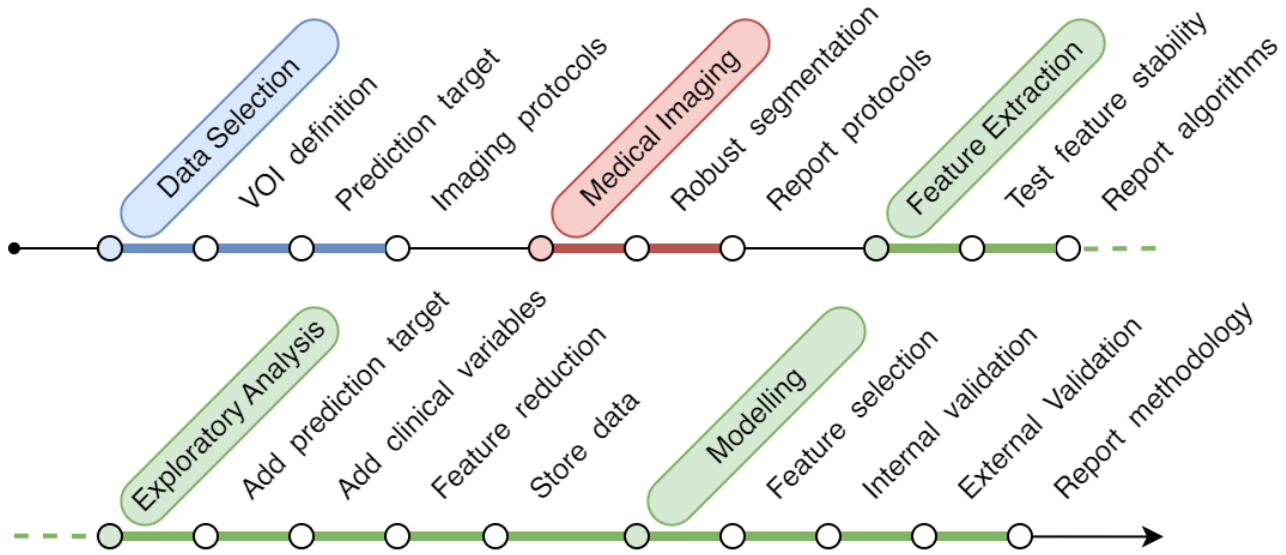

Figure 1.1. The general radiomic workflow, for clinical studies, is intended to be a reference guideline for radiomic study-design, nevertheless, it should be tailored to each investigation setting. Adapted from Lambin et al. (2017).

The scheme can be regarded as a best practise workflow for radiomic studies; however, data availability might impede the implementation of some steps (e.g., the assessment of test-retest feature stability or the presence of an external validation).

We will discuss in better details the main steps in the following section.

1.1 The radiomic workflow

1.1.1 Data Selection

At first, the nature of the study must be defined by selecting an image protocol, the volume of interest and the prediction target. In oncology, usually the gross tumour volume (GTV), delineated for radiotherapy purposes, is chosen as the region of interest (ROI) or volume of interest (VOI) but other choices can also be found in the radiomic literature, such as tumour subregions (habitats), the peritumoral area around the GTV, tumour metastases, lymph nodes as well as normal tissue. Although

radiomic analysis was firstly introduced in oncology, it may find applications in other medical research fields, as the workflow is generalizable to any biomedical image.

The usage of standardized acquisition and reconstruction protocols is advised to eliminate unnecessary confounding factors, given that most radiomic features have been found to be dependent on different acquisition modes, matrix sizes, post-filtering widths, reconstruction algorithms and iteration numbers[8]. In practise, non-standardized protocols are commonly found in radiomic literature, as a number of retrospective studies collected non standardized imaging. Therefore, repeatability and reproducibility of those studies is nontrivial.

1.1.2 Medical Imaging

1.1.2.1 Segmentation

ROI segmentation is a crucial step of the radiomic workflow, as delineations will determine on which voxels features will be computed on. Thus, differences in delineations may reflect in a variation of feature values. Delineations can be performed manually, semi-automatically or in a fully automated fashion. Manual segmentations drawn by expert radiologists are regarded as the gold standard and can be performed with a variety of tools (e.g., by drawing polygons in a slice-by-slice manner, or by using a 2D or 3D brush) and software (e.g., ITK-snap, 3D Slicer, Eclipse). Nevertheless, segmentations performed by two or more expert radiologists (or by the same radiologist at different time points) will slightly differ, due to inter-reader (or intra-reader) variability. On the contrary, semi-automatic or automatic segmentation appears to be more reproducible, with less to no dependence on the reader, but should be, in any case, approved by an experienced physician.

Three-dimensional manual delineations are the most time consuming to obtain and can make the delineation step the most burdensome, in terms of time costs, of the entire radiomic workflow. Because of this, some study come to a compromise by only segmenting the largest tumour slice, reducing the entire analysis to a 2-dimensional analysis. Nevertheless, this methodology is regarded as suboptimal, as it does not allow to fully capture the information contained in the lesion volume[31].

In order to derive radiomic features that robustly characterize the ROI (i.e., irrespectively of delineation differences) and that minimize the evaluation bias (e.g., of radiomic features derived from a slightly incorrect segmentation) the method of multiple-segmentation can be employed. Multiple segmentations

can be derived in many ways: by multiple physicians, by the same physician at different time points, by multiple algorithms, or by systematically perturbing the ROI (e.g., translations, rotations). Depending on the segmentation methods, a feature may or may not be robust, and this is one of the reasons why it is advised to perform such analysis whenever possible instead of relying on previous literature results (which may be obtained with different segmentation methods, different image protocols and/or for different body district).

1.1.2.2 Inter-scanner, inter-vendor, and test-retest variability

The assessment of inter-scanner or inter-vendor variability of radiomic features[32] is of primary importance and allows to mitigate the confounding factors found in retrospective or multicentre studies. If those factors are not considered, they can jeopardize the generalizability of a study's results (e.g., the proposed model might not work on an external validation dataset). Since imaging a patient on different scanners is not always feasible (for safety, ethical and/or practical reasons) human data are scarce and phantom studies might be a mean to identify and discharge vendor-dependent features[32].

Similarly, variability of radiomic features due to organ motion or shrinkage of the target volume over time may also limit the applicability of predictive models. To avoid critical dependencies on those factors, test-retest data can be used when possible (e.g., the RIDER dataset, where two chest CT are acquired 15 minutes apart with the same protocol).

1.1.3 Feature Extraction

The core of radiomic studies is the conversion of biomedical images into minable features. This process, called feature extraction, is carried in a high-throughput fashion, and allows to quantitively characterize the ROIs. Initially, a lack of a consistent nomenclature across studies, as well as variations in the mathematical definitions and software implementations of radiomic features hindered their inter-operability. The impact of those factors has been greatly reduced since the publication, by the Image Biomarker Standardization Initiative (IBSI), of consensus guidelines. In 2020 IBSI provided a definitive manual that reported, not only the general workflow of radiomic feature extraction, but also a consistent and unambiguous nomenclature, the actual mathematical definition of all features, together with some tools to verify software compliance. All these aspects will be covered in greater detail in section 1.3.

1.1.4 Exploratory Analysis and feature reduction

After the extraction step, features are usually stored in a tabular format. Typically, each row corresponds to different patient/subject whereas columns contain different features' expression. In other words, if we focus on a row, we find the 'signature' of a patient, namely its coordinates in the feature space, and if we observe a column, we will see how that specific feature changes across subjects.

Subsequently, the table is combined with clinical data and prediction targets to create a single dataset. Clustering techniques can be applied to features for the identification of groups that share similar expressions across patients. Clustering can also be applied to patients/subjects to find those who are closer together in the features space, namely those whose feature representation is similar. Patient clusters can then be checked against clinical data, for example by means of a chi-squared test, to assess whether the grouping conveys significant clinical information.

Since the number of features that can be extracted from a single ROI is virtually unlimited, we run the risk of facing the so-called 'curse of dimensionality', meaning that, as the number of features grows, the amount of data we need to accurately train a prediction model that generalize grows exponentially. However, in the clinical settings, collecting a large amount of standardized data is not always feasible and the datasets often do not comprehend more than a few hundreds of subjects. Moreover, if we thoughtless increase the number of features with respect to samples, chances are that one feature can perfectly correlate with the prediction target even if it does not carry any relevant information, resulting in overfitting. To face both this issues, feature reduction should also be performed in the exploratory analysis by accounting for feature inter-correlation, feature repeatability and feature reproducibility. Thanks to dimensionality-reduction techniques such as principal component analysis or clustering, the dataset can be analysed to decrease redundant information, for example by identifying clusters of highly correlated features and by only keeping a representative feature for the cluster (e.g., the cluster's centroid, the feature that has the higher association with the prediction target, the most stable feature). Features that lack of robustness against sources of variability (e.g., intra/inter-reader, intra/inter scanner, intra-inter vendor) should also be removed, for example by computing the Intraclass Correlation Coefficient – ICC (that assess the consistency or reproducibility of quantitative measurements) and by only keeping for further analysis features that have an ICC value higher than a certain threshold (e.g., 0.9).

1.1.5 Modelling

1.1.5.1 Feature Selection

Feature selection is the step where few features are chosen for model building. In literature, feature selections methods are mainly divided into three categories: filter methods, wrapper methods and embedded methods, in which the selection step is respectively separate, partially, or fully integrated with respect to the modelling part.

Filter methods constitute the simplest approach to feature selection and work by ranking features accordingly to a heuristic scoring criterion. The number of features to be selected, among the top-scoring ones, is left to the modelling step as hyper-parameter optimization. Filters methods, being classifier independent, have high generalizability and scalability and are usually computational efficient. If the scoring criteria only depend on the prediction target, filter methods are called univariate (e.g., Fisher score, T-test, mutual information maximization - MIM) while, if they also consider interactions within radiomic features, they are called multivariate methods (e.g., mutual information feature selection - MIFS, minimum redundancy maximum relevance - MRMR).

Wrapper methods are basically '*greedy search*' approaches that evaluate all the possible combinations of features against the evaluation criterion (i.e., the performance measure). The search through the whole feature space allows the identification of a relevant and non-redundant feature subset. However, those methods are computationally expensive and prone to overfitting, since the selected features might be overly specific to the evaluated classifiers, resulting in lower generalization capability than filters methods. Well-known techniques in this category are forward selection, backward elimination, and stepwise selection.

Eventually, embedded methods do perform feature selection as a part of the model training process, being more computationally efficient than wrapper methods but still less generalizable than filter methods. The most widespread methods that belongs to this category is the least absolute shrinkage and selection operator (LASSO)[33].

1.1.5.2 Modelling

In the radiomic field, we mainly find two types of modelling strategies: binary classification (e.g., whether the patient will respond to a certain treatment, have a recurrence, or be alive beyond certain

time threshold, etc) and survival analysis. Depending on the strategy we may resort to different machine learning techniques. Logistic regression is a standard approach for binary classification, where the model coefficients are tuned so to predict a logit transformation of the probability of the prediction target. Other possibilities exist, such as support vector machines (SVM), ensemble methods (e.g., random forests, extreme gradient boosting), feed-forward neural networks, K-Nearest Neighbours, and others[34,35].

For survival analysis, methods are available to investigate not only the presence of an event but also the timing it takes to occur. Those methods deal with the peculiar structure of survival data: for example, if a patient is lost to follow-up, does not present the event in the follow-up period, or dies before the prediction target occurs (e.g., onset of side effects), the data for that patient is right-censored. Moreover, since follow-up visits are schedule periodically (e.g., every three months) survival data is also interval-censored, as we might not know the exact time in which the event occurred, but only a broad time interval. The most widespread techniques are Kaplan-Meier method and the log-rank test for univariate survival analysis and the multivariate Cox proportional-hazards model, that allows to simultaneously evaluate the effect of several factors on survival.

1.1.6 Validation

Once the model coefficients have been fitted and the model built with the selected features, it is important to quantifying the predictive ability of the model, in other words, its validation. Based on the "Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis" (TRIPOD) statement[36], there are 4 types of validation for model-based approaches (Table 1.1).

Table 1.1. Types of validations of multivariable prediction models as defined by the TRIPOD statement.

Type 1a	Predictive performance is then directly **evaluated on the same training data** (apparent performance).
Type 1b	**P**redictive performance is evaluated using **resampling** (e.g., bootstrapping or cross-validation) techniques. Type 1b is generally referred to as "internal validation" and is recommended if data are limited.
Type 2a	The data are **randomly** split into 2 groups: one to develop the prediction model and one to evaluate its predictive performance.
Type 2b	The data are **nonrandomly** split (e.g., by location or time) into 2 groups: one to develop the prediction model and one to evaluate its predictive performance.
Type 3	Predictive performance is evaluated using a **separate dataset** (e.g., from a different study).
Type 4	The evaluation of the predictive performance of an existing (published) prediction model on separate data.

In the radiomic field, most preliminary clinical studies follow a Type 1a-b validation scheme, as the number of patients is not sufficient to split the data into separate training and validation cohorts. Following in increasing level of evidence, we find validation types 2a-b, that do require a larger population than types 1, and validation type 3 that requires, at least, two different datasets (e.g., from different studies/institutions).

As validation is fundamental for future integration of radiomic models into clinical practise, it is also evaluated by the Radiomic Quality Score – RQS, which will be presented in the next section, where validation is one of the aspects that impact the most (up to a 10-point difference) on the score.

1.2 Radiomic Quality Score (RQS)

To overcome the lack of a homogeneous evaluation criteria to specifically evaluate clinical radiomic studies, Lambin et al.[14] proposed in 2017 the first version of the Radiomics Quality Score (RQS), an ad-hoc metric to assess both past and future studies, so that researchers can easily determine whether a study follows the best-practices. The RQS 1.0 is a rating scales that goes up to 36 points, evaluating many aspects of radiomic clinical studies (Table 1.2), that can be used both in a prospective manner, as a guide to design new radiomic studies that achieve the highest level of scientific evidence, and retrospectively, as a review metric to weight past studies. Up to September 2022, there has been 75 reviews, systematic reviews and meta-analysis that employed RQS to clearly summarize the quality of the considered studies (PubMed query: *«(RQS AND (radiomic OR radiomics)) OR "radiomic quality score" OR "radiomics quality score"»*).

In the same work, Lambin provided extensive reporting guidelines, integrating the TRIPOD statement, illustrating the necessary level of detail that should be given as supplementary material in any future radiomic publication.

RQS 1.0 is the criteria that will be later referenced in Chapter 4, however, it is worth mentioning that the RQS 2.0 is currently under development (www.radiomics.world) and will consists of 36 checkpoints to encourage best practice in both radiomics studies based on handcrafted features, such as the ones considered in this thesis (i.e., feature for which there is an explicit formulation), and on deep learning.

Table 1.2. The radiomic quality score (RQS) 1.0.

DOMAIN 1: PROTOCOL QUALITY AND STABILITY IN IMAGE AND SEGMENTATION
+1 if the **image protocols** are well-documented and/or public image protocols are used.
+1 if feature robustness to **multiple segmentation** is considered (e.g., multiple readers/software).
+1 if feature robustness to **different scanners/vendor**s is considered.
+1 if feature robustness to **temporal variabilities** is considered (e.g., organ movement, organ shrinkage).
DOMAIN 2: FEATURE SELECTION AND VALIDATION
+3 if **feature reduction** or adjustment for multiple testing is employed (decreases the risk of overfitting), -3 otherwise
-5 if **validation** (performed without retraining and without adaptation of the cut-off value) is missing +2/+3/+4/+5 if validation is based on a dataset from the same institute, a dataset from another institute, two datasets from two distinct institutes (or if the study validates a previously published signature) and on three or more datasets from distinct institutes, respectively.
DOMAIN 3: BIOLOGIC/CLINICAL VALIDATION AND UTILITY
+1 if **non-radiomic** features are included in the analysis (leads to more holistic models and allows correlation analysis between radiomics and non-radiomics features).
+1 if **biological correlates** are discussed (deepens understanding of radiomics and biology).
+2 if comparison to '**gold standard**' is performed (shows the added value of radiomics)
+2 if the **potential clinical utility** of the model in a clinical setting is reported (e.g., decision curve analysis).
DOMAIN 4: MODEL PERFORMANCE INDEX
+1 if **discrimination statistics** are reported together with their statistical significance (e.g., C-statistic, ROC curve, AUC). +1 if resampling method are employed (e.g., bootstrapping, cross-validation).
+1 if **calibration statistics** are reported together with their statistical significance (e.g., calibration plots). +1 if resampling method are employed (e.g., bootstrapping, cross-validation).
+1 if **cut-off analyses** are performed (reduces the risk of reporting overly optimistic results).
DOMAIN 5: HIGH LEVEL OF EVIDENCE
+7 if the study protocol is **prospective** and registered in a trial database.
+1 **Cost-effectiveness** analysis - report on the cost-effectiveness of the clinical application.
DOMAIN 6: OPEN SCIENCE AND DATA
+1 if scans/images are open source. +1 if ROI segmentations are open source. +1 if code is open source. +1 if radiomics features are calculated on a set of representative ROIs and the calculated features and representative ROIs are open source.

1.3 The IBSI Initiative

Other fundamental guidelines for radiomics come from the "Image Biomarker Standardisation Initiative' (IBSI), an international collaboration born in 2016, which aims to standardize the image processing used in the radiomic field. In its first chapter, IBSI-1[15], concluded in 2020[13], IBSI standardised the general image processing workflow, feature definitions, nomenclature and reporting guidelines, and shared tools for the verification of radiomics software implementations (section 1.3.1). In its second chapter, IBSI-2[16], which is currently near completion, IBSI focused on the standardization of biomedical imaging pre-processing, and more specifically on convolutional filtering (section 1.3.3).

Both IBSI chapters are deeply linked to the work presented in this thesis, which consisted both in the standardization, upgrade, and validation of a radiomic feature-extraction software, S-IBEX, following the IBSI standard but also in the contribution to the actual standardization effort of convolutional filtering (Chapter 2). For this reason, a summary of both IBSI chapters will be provided in the next sections, as background material for the reader.

1.3.1 IBSI-1

IBSI chapter 1 comprised 25 research teams worldwide and spanned from 2016 to 2020[13,16]. The work was organized in three stages: two standardization phases followed by a last validation one (Figure 1.2).

1.3.1.1 IBSI-1: Phase 1

The first phase aimed to standardize radiomic feature definitions and to establish reference benchmark values, without considering any image processing. To this goal, a reference phantom, the *IBSI digital phantom*, was created and shared, together with a feature extraction workflow, among IBSI participants so to iteratively reach a consensus on computed feature values.

In general, a digital phantom is typically made of a pair: a 3D stack of 2D images describing the spatial distribution of intensity values, and a corresponding binary mask that defines the area of interest on which to extract radiomic features. In the case of the *IBSI digital phantom* the image was artificially

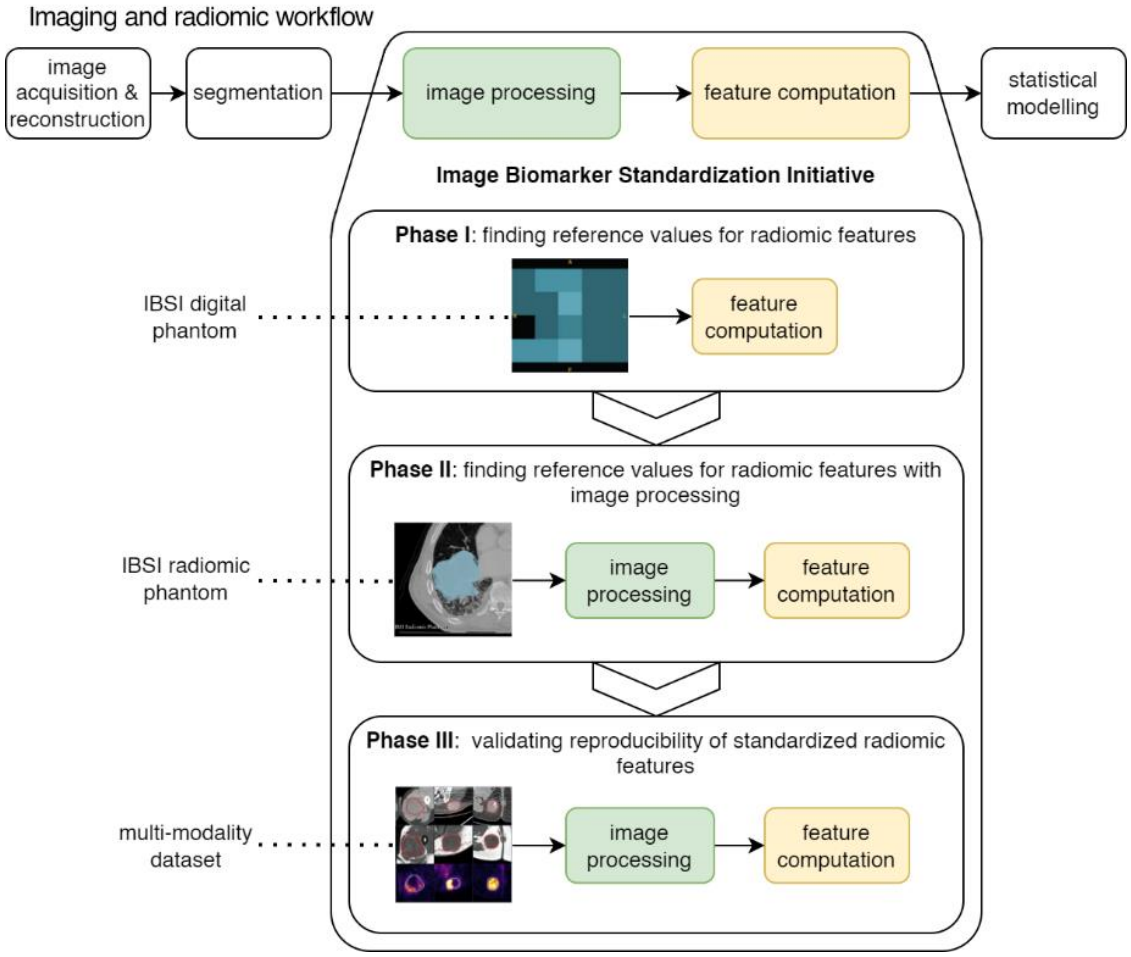

Figure 1.2 The three-phase setup of IBSI-1. Adapted from Zwanenburg et al (2020).

created by randomly combining intensity values in the range [1,9] and is composed of four slices of 5x4 voxels (for a total of 80 voxels), with a voxel dimension of 2.0x2.0x2.0 mm. The ROI mask comprises 74 voxels, excluding 5 peripheral voxels and one internal voxel. The entire digital phantom is represented in Figure 1.3.

The phantom was designed not to require any pre-processing before feature calculation, as to ease the standardization of feature implementation (appendix Table A.1).

1.3.1.2 IBSI-1: Phase 2

In the second phase, IBSI focused on the standardization of image-processing, defining the general feature-extraction workflow for radiomics. To this aim, a different phantom, the *IBSI radiomic phantom* was required, on which mimicking the image processing of a typical radiomic study. Thanks to this phase IBSI achieved a consensus for reference values under different image processing configurations.

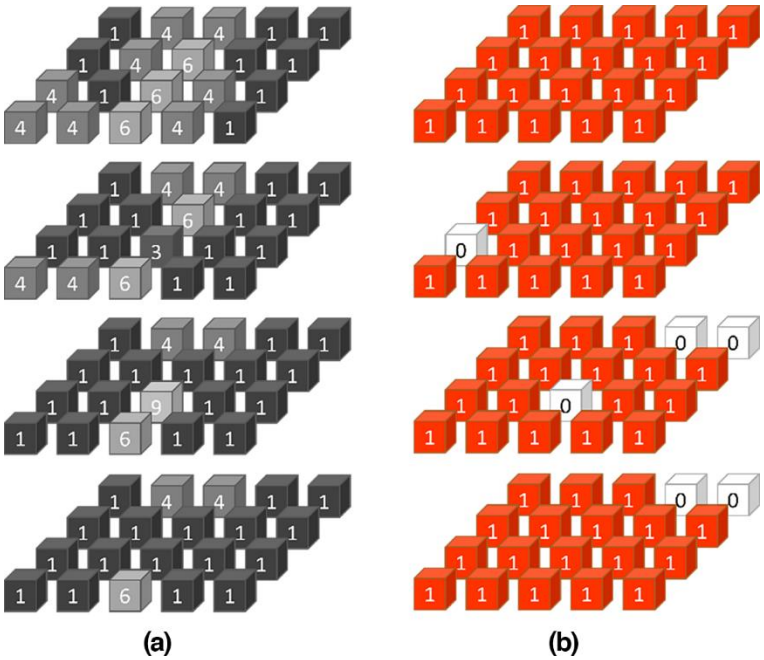

Figure 1.3. Image and mask of the IBSI digital phantom. (a) Spatial distribution of the intensity values inside the image and (b) morphological mask associated to the image.

The *IBSI radiomic phantom* is more clinically-oriented than the *IBSI digital phantom* and was derived from the CT scan of a non-small cell lung cancer patient (Figure 1.4).

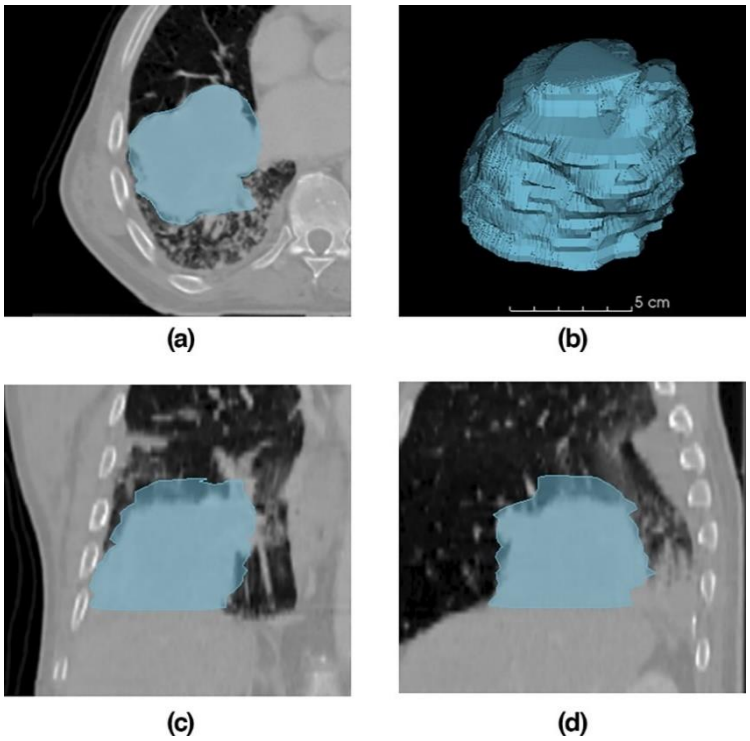

Figure 1.4.(a) axial, (c) coronal and (d) sagittal view of the IBSI radiomic phantom. (b) volumetric representation of the gross tumour volume ROI.

To mimic the operations of a real radiomic study, it is to be used in combination of several pre-processing (configurations A-E reported in appendix Table A.1). The CT image was anonymized, cropped, and centred around the lesion and the provided segmentation represents the gross tumour volume (GTV). The image has dimensions of 204x201x60, with a voxel size of 0.9x0.9x3mm.

1.3.1.3 IBSI-1: Phase 3

The third phase consisted in the prospective validation of the standardization results obtained in the first two phases of the Initiative. Validation was carried on a multimodality dataset (i.e., CT, PET, and MRI), publicly available on the Cancer Imaging Archive[37], comprising 51 patients with soft-tissue sarcoma[38]. For reference, the first three patients of the dataset are presented in Figure 1.5.

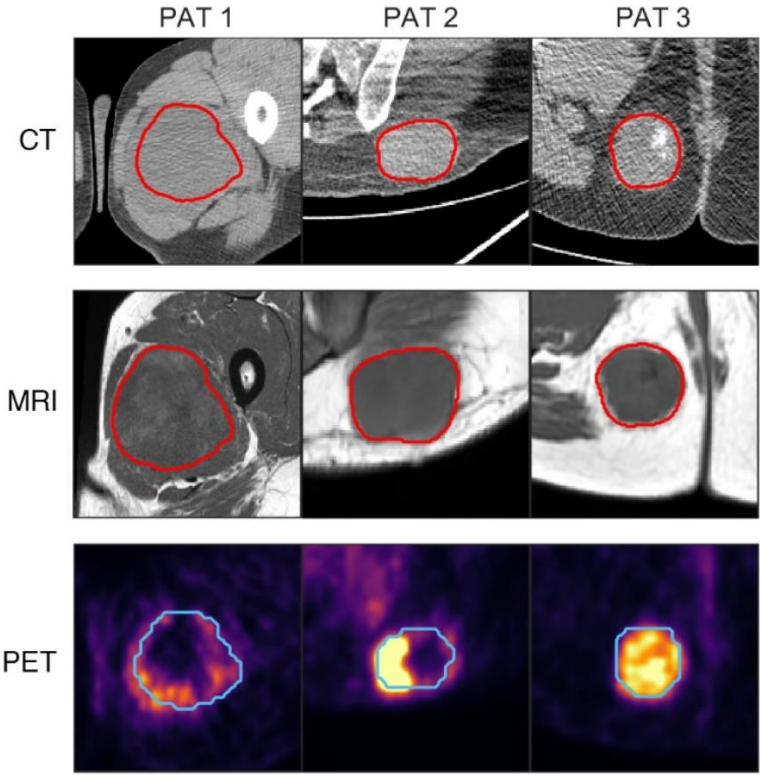

Figure 1.5. Representative axial slices of the first three patients of the soft-tissue sarcoma dataset. (top) CT, (middle) MRI and (bottom) PET imaging.

1.3.2 IBSI-1 results: the guidelines

The image processing workflow that was standardized by IBSI-1 is reported in Figure 1.6 and will be described in better details in the next subsections.

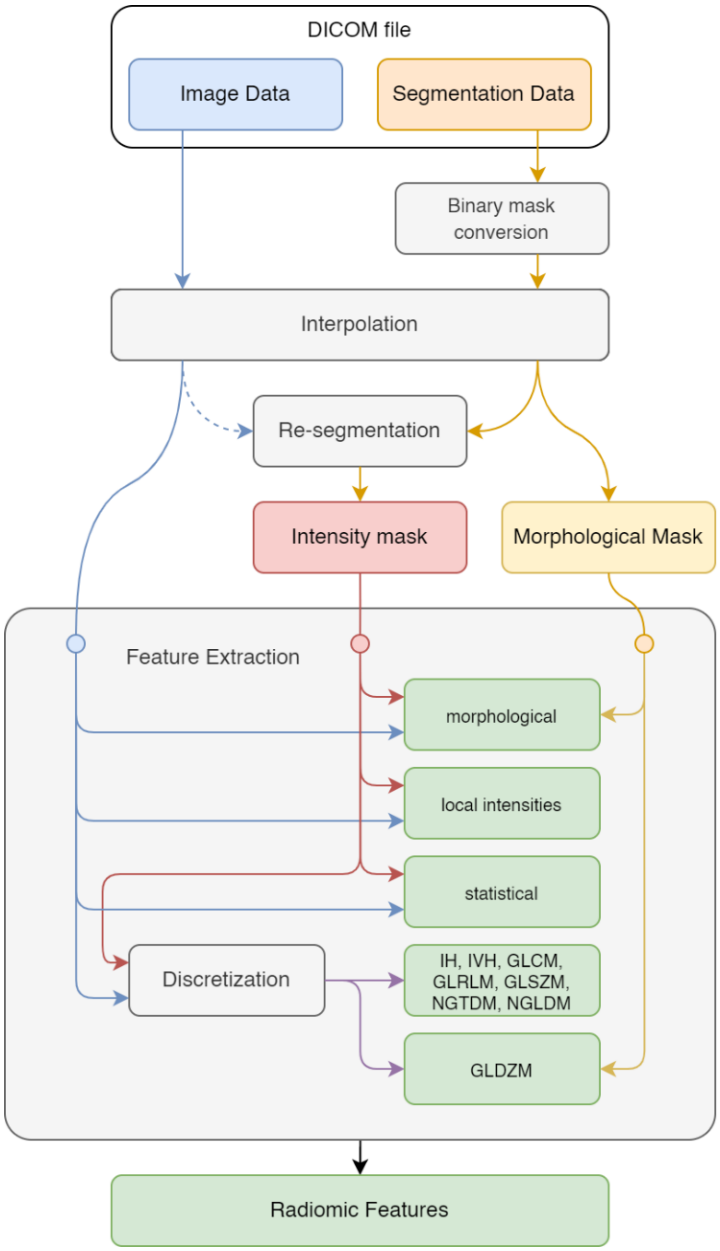

Figure 1.6. Image processing scheme for radiomic feature computation.

1.3.2.1 Binary mask

Within radiomics, segmentation is the step in which the ROI is delineated, identifying the volume on which features are to be computed. If the ROI is directly created as a binary mask (e.g., true/1 if the voxel belongs to the ROI, false/0 otherwise) no additional steps are needed, as the mask has the same format used for later radiomic calculations. However, it is common that ROIs are manually drawn and saved using the DICOM RTSTRUCT format, which stores the ROI as a slice-by-slice set of polygonal curves (set of contours). In this case, IBSI standardize the process of conversion from contour sets to binary masks by advising the usage of the 'crossing number algorithm'[39].

1.3.2.2 Interpolation

Interpolation is considered an essential preliminary step before feature calculation. Having a homogeneous voxel size across different samples is advised for reproducibility purposes, as many features are sensitive to voxel dimensions[40], but is not always possible (e.g., retrospective studies). Interpolation allows to not only address this issue (by choosing the resampled voxel dimension) but also to obtain isotropic voxels (i.e., voxels whose sizes are the same across the three spatial dimensions) which are invariant to rotation and are a requirement for textural features, which consider reciprocal positions of voxels. Interpolation should be applied both to the image and to its binary mask and, depending on the input modality, it can be performed in a 2D (slice by slice) or 3D manner. Whether is still not clear if it is better to choose between up-sampling (introduces artificial information) or down-sampling (aliasing and/or information loss can occur), IBSI standardized the methods without advising any specific configuration.

During interpolation, IBSI identifies voxels by their centre coordinates, which lie on the intersections of a regularly spaced grid, denoted as the original grid. The new interpolation grid must be defined by mean of its size and positioning with respect to the original grid. Several techniques exist for grid positioning, and three are reported by IBSI as "*fit to original grid*", "*align grid origins*", and "*align grid centres*" (Figure 1.7). The usage of the latter method is suggested by IBSI, being it more implementation-independent (i.e., it requires less meta-data which can be not readily available in all radiomic tools).

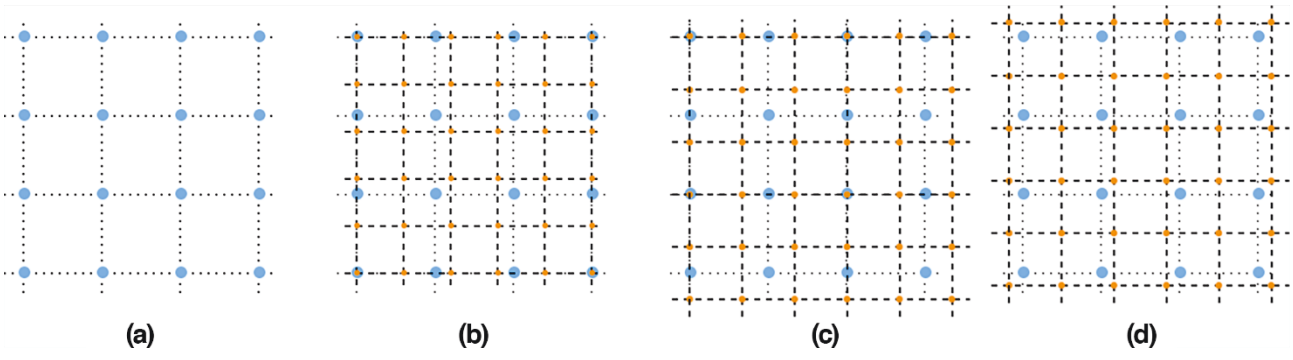

Figure 1.7. Visual examples of the three grid positioning techniques: (a) original grid, (b) fit to grid, (c) align grid origins and (d) align grid centres.

Once the position of the new grid is defined, voxels intensities may be derived with different algorithms, among the most popular IBSI recall nearest neighbour, linear, cubic and spline interpolation (Figure 1.8).

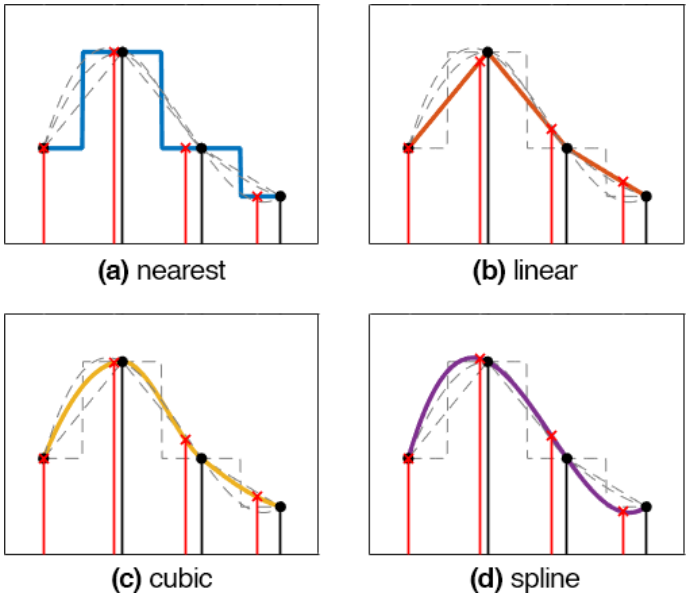

Figure 1.8. (a-d) Examples of different interpolation methods for the 1-dimensional case. Original grid in black, interpolation grid in red.

IBSI guidelines also state that both the interpolated image and binary mask must retain their original data type. In particular, if interpolation does not preserve its Boolean data type, the mask should be converted to logical values by using a threshold (i.e., δ, 0.5 by default).

1.3.2.3 Re-segmentation

As per IBSI guidelines, re-segmentation provides an update of the ROI mask and is typically used, after interpolation, to refine the mask by excluding unwanted regions (e.g., air or bone tissue from the CT ROIs, low uptake regions in PET ROIs). Re-segmentation can create holes in the ROI or split the ROI into non-connected subregions. For these reasons, the unmodified ROI mask is stored as a *morphological mask* and is complemented with the re-segmented mask, the *intensity mask*.

Two techniques are reported by IBSI for re-segmentation. The first one is called *range re-segmentation*, where a range of intensities is defined, so that voxels outside that range are removed from the ROI mask. The interval can be defined both as a closed interval (e.g. [-50, 50] HU) or as a half-open interval (e.g., $[0, \infty)$ HU). The second technique is *outlier filtering*, where only voxels inside the range $[\mu -$

$3\sigma, \mu + 3\sigma]$ are retained in the intensity mask (where μ and σ are the mean and standard deviation of the intensity of voxels inside the ROI).

1.3.2.4 Intensity discretisation

Gray-level discretization is a necessary step for textural matrices creation and performs a beneficial noise-reducing action. The optimal method of choice may be modality dependent and has a significant impact on feature values and reproducibility. Many algorithms exist, making the discretization step a possible source of variation among software. IBSI standardized two discretization methods: Fixed Bin Number (FBN) and Fixed Bin Size (FBS).

In FBN, ROI intensities are discretised to a fixed number N of bins (e.g., 32, 64) using the following formula:

$$X_{k,discretized} = \begin{cases} \left\lfloor N \frac{X_k - X_{min}}{X_{max} - X_{min}} \right\rfloor + 1, & X_k < X_{max} \\ N, & X_k = X_{max} \end{cases}$$

In short, the intensity of voxel k-th is corrected by the lowest occurring intensity in the ROI, X_{min}, normalized by $(X_{max} - X_{min})/N$ and subsequently rounded down to the nearest integer.

On the other hand, using the FBS approach a new bin is created for every intensity interval with width w. The minimum intensity may be user set or data driven as the minimum inside the ROI.

$$X_{k,discretized} = \left\lfloor \frac{X_k - X_{min}}{w} \right\rfloor$$

IBSI does not advise the optimal bin number N or the bin width w to use, as it should be study dependent, however it suggests which method to preferer based in the image type and the re-segmentation step. If the image has calibrated intensity units and the mask has been updated using range re-segmentation, both methods are valid options. Instead, if the image has calibrated intensities but has not been re-segmented, or has arbitrary intensities, only FBS is advised.

1.3.2.5 Feature extraction

Feature calculation is the final step that, after resampling, discretization, and re-segmentation, translates the content of the ROI into a collection of minable features. IBSI standardized the calculation of 174 radiomic features, belonging to 11 feature families, which are summarized in Table 1.3. The interested reader can find the definitions of each feature in the IBSI reference manual[41].

Table 1.3. Feature families standardized by IBSI-1.

Feature Family	*Abbreviation*	*# of features*	*ROI Morph**	*Dis.***	*Description*
Morphological	MF	29	yes	-	Features of the morphological family describe geometric aspects of the ROI, such as surface area and volume.
Local intensity	LI	2	-	-	Neighbouring voxels around a centre spot are used to compute local intensity features, such as the maximum value.
Intensity-based statistics	IS	18	-	-	Features belonging to the intensity-based statistics family describe how intensities are distributed within the ROI (without considering reciprocal positions of voxels).
Intensity histogram	IH	23	-	yes	Features are computed on a histogram that is generated by discretising the original intensity within the ROI into bins.
Intensity-volume histogram	*IVH*	5	-	yes	Features are computed on the cumulative intensity-volume histogram that links each intensity value to the fraction of the volume containing at least that intensity.
Grey-level co-occurrence matrix[42]	*GLCM/ CM*	25	-	yes	Features are computed on the grey level co-occurrence matrix (GLCM), that expresses how combinations of discretised intensities of neighbouring voxels are distributed along one of the image directions.
Grey-level run length matrix[43]	*GLRLM/ RLM*	16	-	yes	Features are computed on the grey level run-length matrix (GLRLM) that counts the runs having a specific run-length and intensity value (a run length is defined as the length of a consecutive sequence of voxel having the same discretized intensity).
Grey-level size zone matrix[44]	*GLSZM/ SZM*	16	-	yes	Features are computed on the grey level size zone matrix (GLSZM), that counts the number of zones having a specific size and intensity (a zone is defined

					as neighbouring voxels having an identical discretised grey level).
Grey-level distance zone matrix[44]	*GLDZM/ DZM*	16	yes	yes	Features are computed on the grey level distance zone matrix (GLDZM), that counts the number of zones of linked voxels which share a specific discretised grey level value and possess the same distance to ROI edge
Neighbourhood grey tone difference matrix[45]	*NGTDM*	5	-	yes	Features are computed on the neighbourhood grey tone difference matrix (NGTDM) which is an alternative to GLCM and reports the average grey-level difference for each grey-level.
Neighbouring grey level dependence matrix[46]	*NGLDM*	17	-	yes	Features are derived from the neighbouring grey level dependence matrix (NGLDM) which is also an alternative to the GLCM. The NGLDM captures the coarseness of the overall texture within the ROI and is rotationally invariant.

* Does the feature family use the morphological masks?

** Is discretization required by the feature family?

1.3.2.6 Feature aggregation

For both grey-level co-occurrence matrix (GLCM) and grey-level run length matrix (GLRLM), namely for directionally sensitive feature families, IBSI standardized six different approaches to derive, from a 3D ROI, a unique feature value. Both this families require the calculation of an intermediate matrix on which features are calculated. Depending on how these matrices (which depends both on a direction parameter and on the slice/volume) are aggregated we can distinguish six different cases, reported in Figure 1.9.

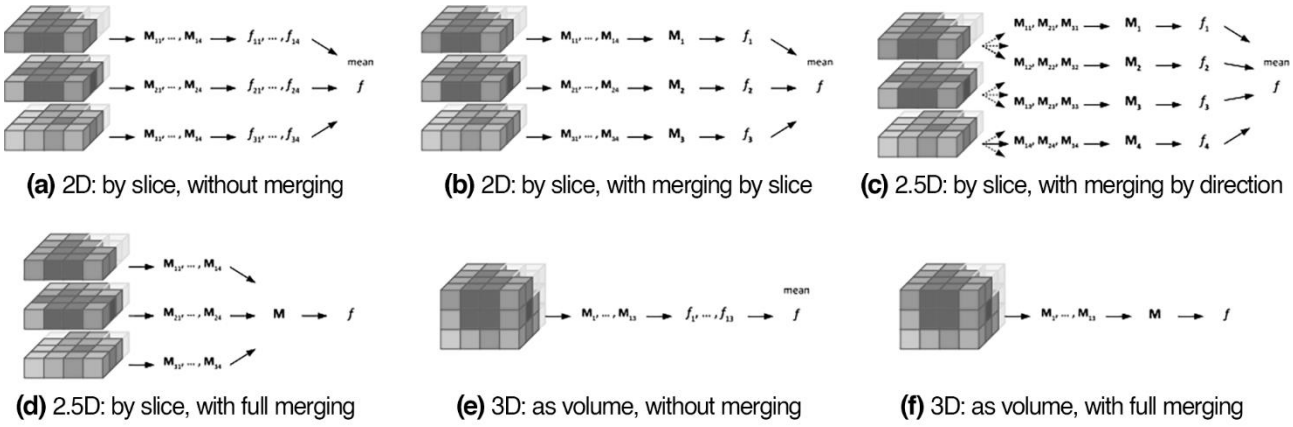

Figure 1.9. (a-f) Aggregation methods for directionally sensitive feature families. Adapted from the IBSI-1 manual.

On the other hand, for grey-level size zone matrix (GLSZM), grey-level distance zone matrix (GLDZM), neighbourhood grey tone difference matrix (NGTDM) and neighbouring grey level dependence matrix (NGLDM), whose matrices do not depend on a direction parameter, and are thus rotationally invariant, IBSI standardized the three aggregation methods reported in Figure 1.10.

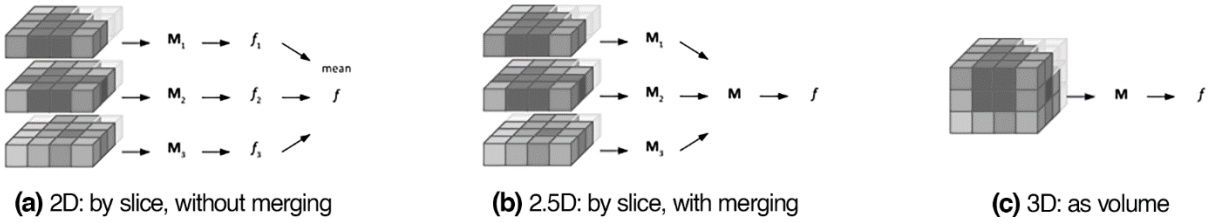

Figure 1.10. (a-c) Rotationally invariant aggregation methods. Adapted from the IBSI-1 manual.

1.3.3 IBSI-2

IBSI chapter 2 is an ongoing project that will end in early 2023 that focuses on the standardization of image filtering and more specifically on convolutional image filters, in order: Mean, Laplacian of Gaussian (LoG), Laws, Gabor, separable (e.g., Daubechies, Coifflet, Haar) and non-separable wavelets, (e.g., Simoncelli), and Riesz transform[16]. In radiomics, image filters can be used to highlight specific characteristics of the ROI, such as edges and contours, at different spatial scales, ultimately leading to radiomic features that may convey more valuable information than the ones computed on the original unfiltered image.

However, being filtering an additional step ahead of feature computation (Figure 1.6), it represents an additional degree of possible variability across software implementations, potentially leading to less

reproducible results. Indeed, if the same filtering technique, implemented in two software tools, results in two different filtered images (also called response maps) radiomics features will in turn differ.

IBSI-2, similarly to the first IBSI chapter, is structured in three main phases, of which the first two are used to establish a reference standard, while the last one is for validation.

Participation to IBSI-2 is conditional upon compliance of three requirements: 1) being the developer of a radiomic software, 2) the software must be compliant with the IBSI 1 standards (for phases I & II) and 3) participation in at least one of the phases of the project.

Thanks to the development of S-IBEX (presented in Chapter 2), the first two criteria were met so I enrolled in the Initiative in August 2021 and joined 13 other teams worldwide.

1.3.3.1 IBSI 2 Phase 1: Technical validation of image filters without additional image processing

In the first phase, a set of newly proposed set of 9 digital phantoms (four of which are visible in Figure 1.11) was used to establish a standard implementation of various convolutional filters.

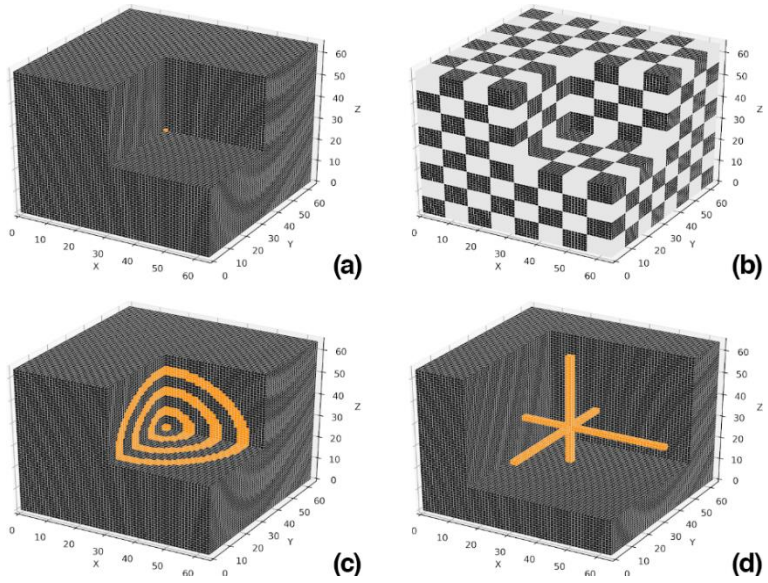

Figure 1.11. (a) impulse, (b) checkerboard, (c) sphere and (d) patter 1 phantoms. Adapted from the IBSI-2 manual.

This phase did not involve any computation of imaging features or image processing other than image filtering itself. Participants were provided with guidelines so to have a common base ground for the implementation of convolutional filters. Thirty-six different filtering configurations were tested on four phantoms of the set (impulse, checkerboard, sphere, and Pattern 1) and the response maps were

collected across centres for comparison. All configurations and all filter parameters are reported in appendix Table A.2.

1.3.3.2 IBSI 2 phase 2: Establishing reference values for features extracted from response maps

In the second phase, the *IBSI radiomic phantom*, presented in subsection 1.3.1.2, was employed to establish reference values for intensity-based statistical features computed from filter response maps. The pre-processing pipelines involved other steps then convolutional filtering and are all reported in Table 1.4. The specific parameters used for convolutional filtering are reported in appendix Table A.3. Like IBSI-1, features were collected across centres to achieve a strong consensus on reference values.

Table 1.4. General pre-processing pipeline ahead of intensity-based statistics feature computation for the IBSI radiomic phantom.

Parameter	***Configuration A***	***Configuration B***
Slice-wise (2D) or as volume (3D)	2D	3D
Interpolation	no	yes
Resampled voxel spacing (mm)	-	1 x 1 x 1
Interpolation method	-	tricubic spline
Intensity rounding	-	nearest integer
ROI interpolation method	-	trilinear
ROI partial mask volume	-	0.5
Re-segmentation		
Range (HU)	[-1000; 400]	[-1000; 400]
Outlier filtering	none	none
Image filters		
Filters	see Table A.3	see Table A.3
Boundary condition	mirror	mirror

1.3.3.3 IBSI 2 phase 3: Validation of filter-derived feature reproducibility

In the third phase, the reproducibility of features obtained from standardised filter implementations is currently under validation using the same multi-modality sarcoma dataset of IBSI 1 phase 3. Modality-specific pre-processing are reported in Table 1.5.

1.3.4 IBSI-2 results: the guidelines

As a results of the IBSI-2 effort, another manual was redacted[47], which extends the first IBSI reference manual by defining and standardising the so-called convolutional filters and by providing the formula-

Table 1.5. Modality-specific pre-processing.

Parameter	***CT***	***MR***	***PET***
Slice-wise (2D) or as volume (3D)	3D	3D	3D
Interpolation	yes	yes	yes
Resampled voxel spacing (mm)	1 x 1 x 1	1 x 1 x 1	3 x 3 x 3
Interpolation method	tricubic spline	tricubic spline	tricubic spline
Intensity rounding	nearest integer	-	-
ROI interpolation method	trilinear	trilinear	trilinear
ROI partial mask volume	0.5	0.5	0.5
Re-segmentation			
Range (HU)	[-200, 200]	$[0, \infty)$	$[0, \infty)$
Outlier filtering	none	none	none
Image filters			
Filters	see Table A.4	see Table A.4	see Table A.4
Boundary condition	mirror	mirror	mirror

tions, parameters and notations that should be used within radiomic studies. In this section, we will briefly cover the major standardization points, which will be necessary to understand the standardisation process presented in section 2.3.

In the manual, convolutional filtering was framed within the general radiomic image processing pipeline as an optional step positioned in-between interpolation and feature extraction (Figure 1.12). Filtering should produce a response map having the same dimensions as the input image, N^D. Eventually, radiomic features can be computed from the response map as well as from the original image.

In the discretized domain of digital images, the convolutional filtering operation can be mathematically expressed as:

$$h[\boldsymbol{k_0}] = (g * f)[\boldsymbol{k_0}] = \sum_{\boldsymbol{k} \in M^D} g[\boldsymbol{k}] f[\boldsymbol{k_0} - \boldsymbol{k}]$$

where $h[k_0]$ denotes the response map at voxel k_0, g the filter, f the original image and M^D the D-dimensional support of the filter. Convolutional filtering can also be computed in the Fourier domain, since the relation $(g * f)(\boldsymbol{x}) \overset{F}{\leftrightarrow} \hat{g}(\boldsymbol{\omega})\hat{f}(\boldsymbol{\omega})$ remains valid in the discrete case, stating that convolution operation can be carried, in the Fourier domain, as the Hadamard product between the image and the filter.

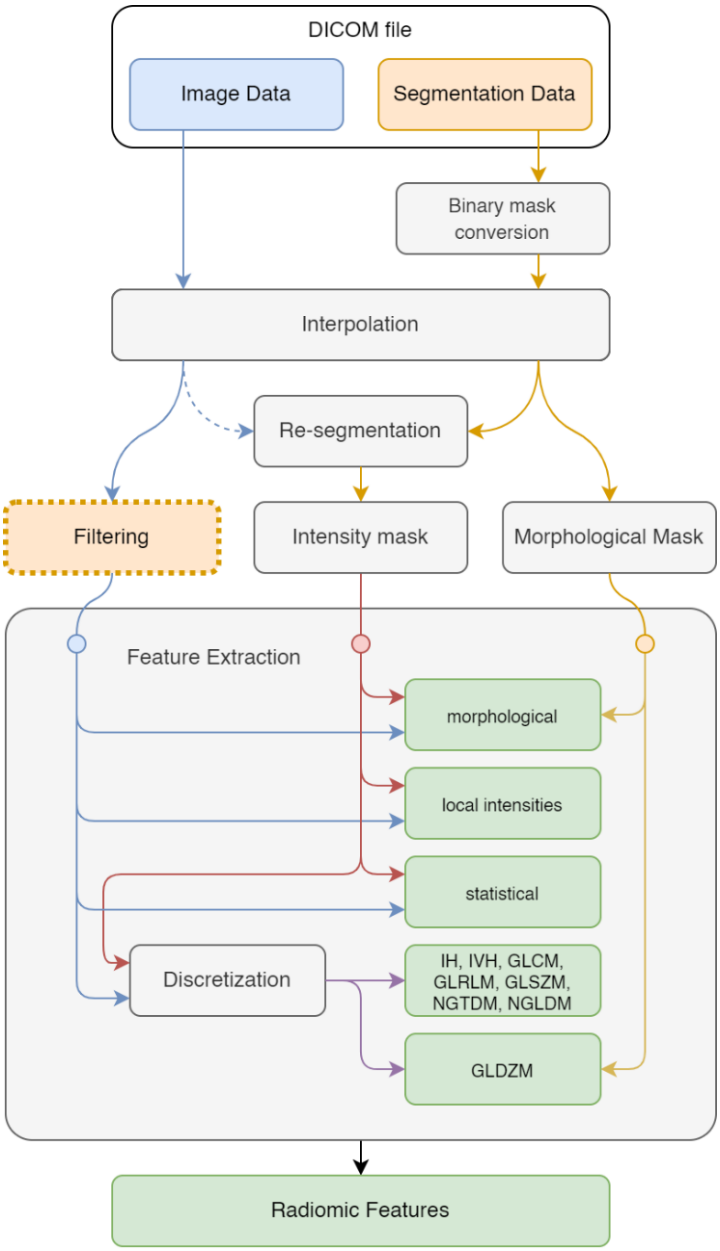

Figure 1.12. Filtering step in the overall feature extraction workflow.

During convolution computation, voxels that are close to image boundaries more than half of the spatial filter's support require accessing voxels that are outside the image support. Those value can be artificially derived with several methods, presented in Figure 1.13 (e.g., constant value padding, nearest value padding, periodisation, and mirroring). Nevertheless, as radiomic features are later derived from a ROI that is usually far from image boundaries, the impact of such conditions on feature values should be minimal.

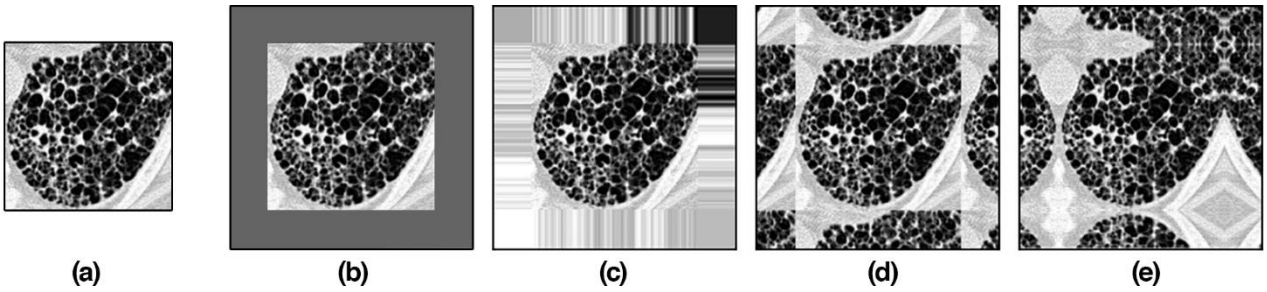

Figure 1.13. Boundary conditions. (a) original image, (b) constant value padding, (c) nearest value padding, (d) periodisation, and (e) mirroring. Adapted from the IBSI-2 manual.

The manual also reminds some desired properties from an image-processing perspective:

- Translation equivariance: by construction, response maps of convolution operations are equivariant to translation.
- Rotation invariance: response maps should be invariant to local rotations since it is important to identify a pattern (e.g., collagen junctions) irrespectively of its local orientation.
- Directional sensitivity: directional sensitive filters (e.g., Gabor filter) are of interest since they can discriminate diverse patterns (e.g., blob and tubular structures) of medical images (e.g., nodules and vessels). Nevertheless, directional sensitive filters are not, by design, rotationally invariant. Only circularly/spherically symmetric filters (e.g., LoG filter) are rotation invariant, which in turn are not directional sensitive. A clever strategy to achieve both invariance to local rotations and directional sensitivity is to employ a directional-sensitive filter and 1) compute a pseudo rotation equivariant representation via a collection of rotated filter responses and 2) perform voxel-wise pooling (e.g., average, or max pooling).

Eventually, the manual reports in details some of the most used convolutional filters and their parameters, which will be briefly presented here.

Mean Filter. This filter is a simple and intuitive method for image smoothing, reducing the high-frequency content by applying the *average* operator (Figure 1.14). The filter can be represented with a homogeneous square/cube, having intensity values all equal to $1/M^D$, with M being the spatial dimension of the filter and D the number of spatial dimensions (e.g., 2 and 3 for 2D and 3D filters, respectively). It is mainly employed for its noise-reducing properties.

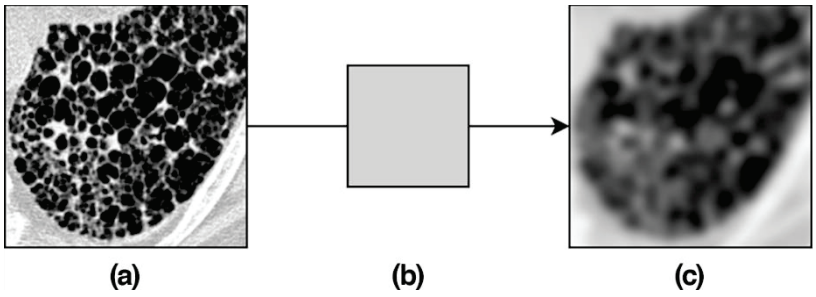

Figure 1.14. (a) original image, (b) Mean filter - $D = 7$, (c) response map. Adapted from the IBSI-2 manual.

Laplacian of Gaussian. The Laplacian of Gaussian is a band-pass filter (Figure 1.15) and its circularly/spherical symmetry makes it invariant to local rotation and directionally insensitive. Its profile is obtained as the radial second-order derivative of a D-dimensional Gaussian filter. Its main parameter, σ^*, controls the scale of the operator. Since its spatial support is virtually infinite, for its implementation in the spatial domain it is necessary to introduce a truncation parameter, d, that, together with σ^*, defines the spatial support size, M, of the filter ($d = 4$ allows to truncate the filter where its values are close to zero).

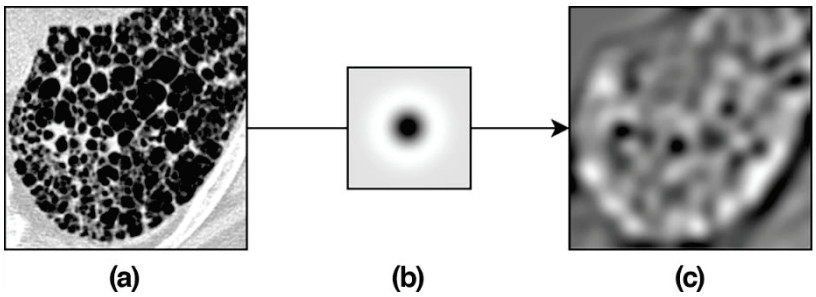

Figure 1.15. (a) original image, (b) LoG filter - $\sigma^* = 10$, $d = 4$, (c) response map. Adapted from the IBSI-2 manual.

Laws Filter. Laws 2D and 3D filters are, by design, separable and can be obtained through the combination of 5 base 1D kernels, therefore response maps are efficiently obtained by sequentially convolving those kernels along each image direction. The five base kernels comprehend one low-pass filter called Level (L) for grey level averaging, and four zero mean kernels for texture detection. The four types of transitions covered are Edges (E), Spots (S), Wave (W) and Ripples (R). L, E, and S come in two scale options (with a spatial support of 3 and 5), while W and R only have one support dimension (support of 5). As an example, the 2D filter identified with the wording "E5R5" is obtained

by convolving the image with the Edge kernel (of support 5) along its first dimension and with the Ripple kernel (of support 5) along its second dimension (second row, last column of Figure 1.16).

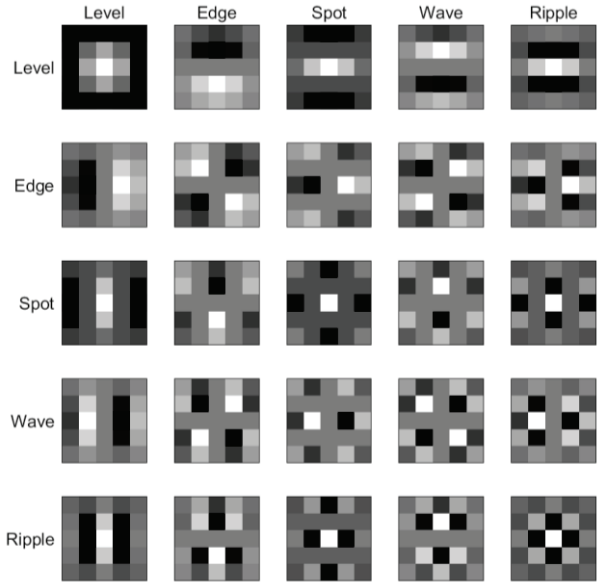

Figure 1.16. The twenty-five 2D Laws kernel, obtained as a combination of the 5 base 1D kernels of support 5.

To derive Laws "*energy maps*", at first the response map is generated as previously described, subsequently the average filter is applied to the response map using a sliding window of a given support δ.

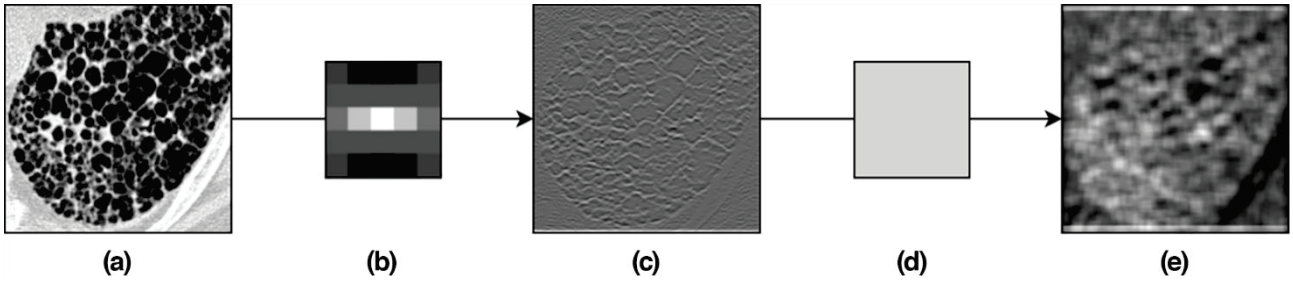

Figure 1.17. (a) original image, (b) Laws L5S5 2D filter, (c) response map, (d) average filter, (e) energy map. Adapted from the IBSI-2 manual.

Gabor Filter. Gabor filters allow the extraction of directional and multi-scale patterns. In the spatial domain the 2D filter is defined by an elliptic Gaussian envelope (parametrized by the scale parameter σ^*, the aspect ratio γ^* and its orientation θ), and an oscillatory function (parametrized by the wavelength λ). In the spatial domain, the kernel is complex and produces complex response maps.

Therefore, usually, the modulus of the response is computed before feature calculation (Figure 1.18). Since the spatial support of the filter is not compact, the kernel is cropped.

Gabor filter is not rotation-invariant, but rotation equivariance/invariance can be approximated by pooling the response maps of a Gabor filter bank (obtained by varying θ with $\Delta\theta$ increments).

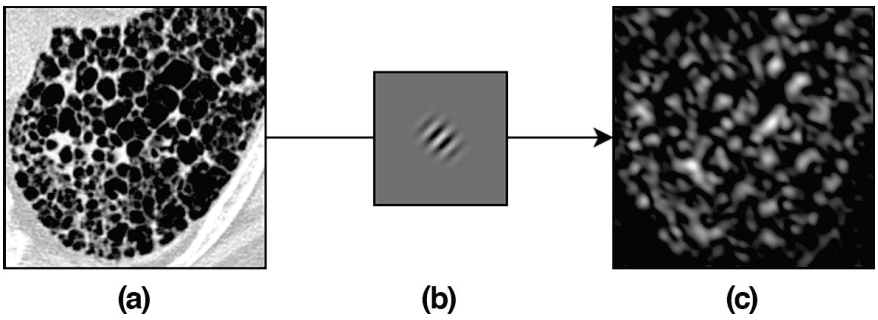

Figure 1.18. (a) original image, (b) Real part of the Gabor filter - σ=5, λ=2/π, γ=3/2, ϑ=π/4, (c) modulus of the response map. Adapted from the IBSI-2 manual.

Separable and non-separable wavelets. Wavelet filtering constitutes a wide group of filters where paired high- and low-pass filters, covering the entire image spectrum, are used in combination. In the radiomic field, where translation invariance/equivariance is a desired property, undecimated filtering is the preferred method, thus response maps always have the same size of the original image. To achieve the typical multiscale decomposition of the decimated transform, the filter itself should be up-sampled with the '*à trous*' algorithm, namely by inserting zeros in-between filter coefficients. Implementation details can be found in the IBSI-2 reference manual.

Separable wavelets can be obtained by the combination of 1D filters, however, none of them is rotationally invariant (except for the Gaussian filter). Among this group we find Haar (Figure 1.19), Daubechies, Coifflet and many other wavelets.

Non-separable wavelets were introduced to achieve an image analysis invariant/equivariant to local rotations. To this aim, filters are constructed in the Fourier domain to be circularly symmetric. Usually, non-separable wavelets are implemented in the Fourier domain, being its frequency support the same of the image to be filtered.

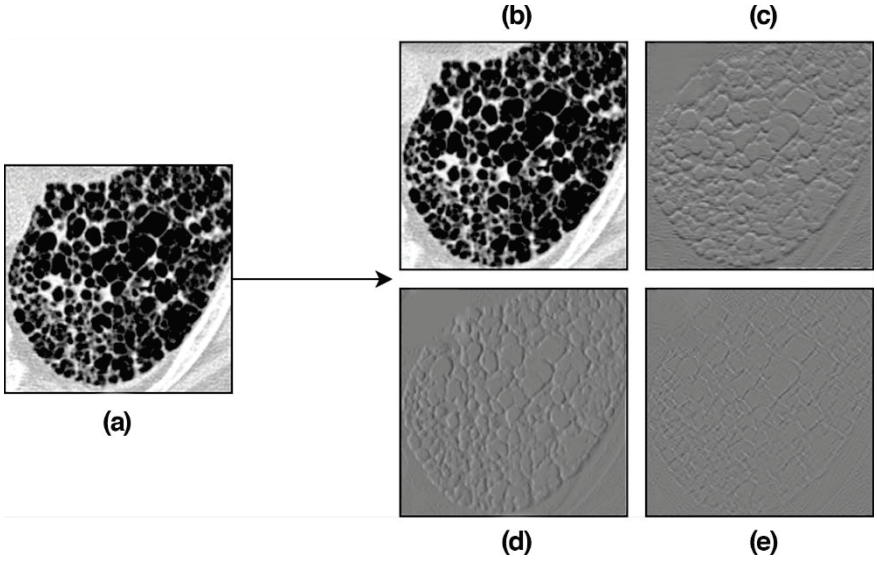

Figure 1.19. (a) original image and response maps (b-e) from the 2D undecimated separable wavelet transform. (b) low-low, (c) high-low, (d) low-high, (e) high-high. Adapted from the IBSI-2 manual.

Riesz Filter. Due to their symmetries, non-separable wavelets cannot characterise directional patterns. An elegant approach to obtain directional sensitivity is to combine them with Riesz filtering, which is made by all-pass normalized image derivatives.

Despite of its definition being standardized, Riesz filtering was not tested in the third validation phase, being the consensus obtained in the previous phases not robust. For the sake of simplicity, details are left in the IBSI-2 manual[47].

Chapter 2: From IBEX to S-IBEX

In early 2015, 1 year before the start of IBSI, Zhang et al., from the University of Texas MD Anderson Cancer Center, publicly released the Image Biomarker Extractor, IBEX, a radiomic software tool for applying, building, and sharing reproducible image analysis algorithms[48], written using both MATLAB 2011a and C/C++. Among its strengths we find multimodality support, a modular architecture that eases the addition of custom algorithms, and a Graphical User Interface (GUI) where parameters for the extraction can be inputted in a straightforward manner. Those are the appealing characteristics that made it a popular tool for radiomic studies[49–57] and that also made us choose this software over others (CGITA, CERR, MAZDA) for our research projects. The IBEX workflow is reported in Figure 2.1.

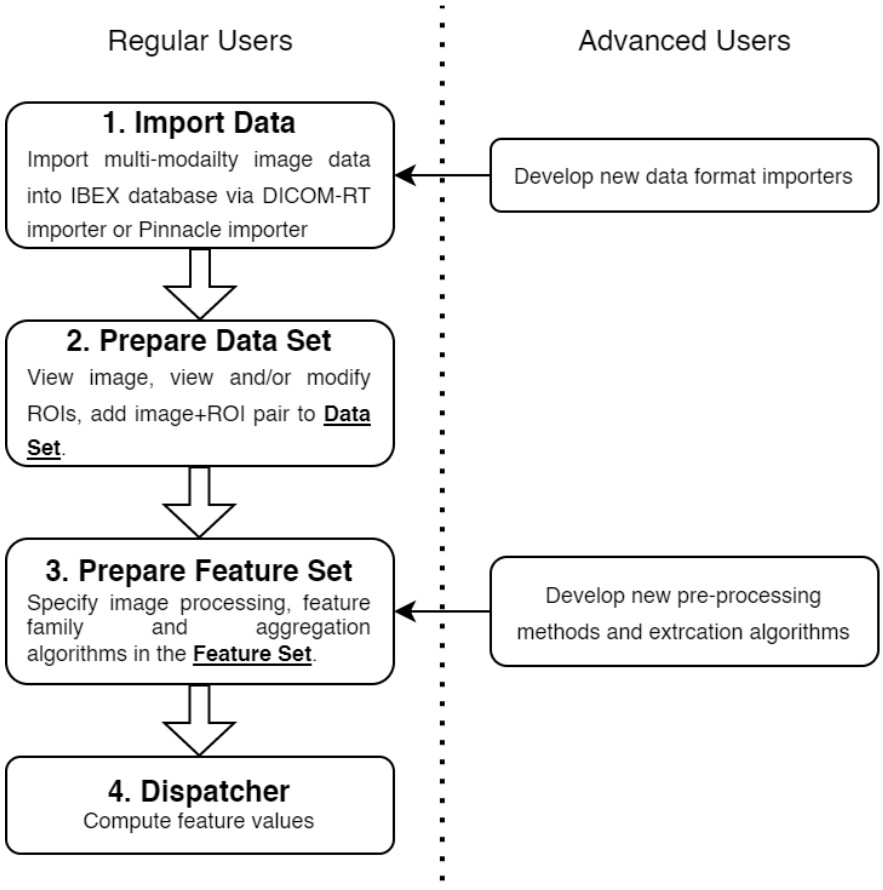

Figure 2.1. The IBEX workflow. Regular users import data, prepare the data set and feature set, and compute radiomic feature values. Advanced users can plug in new data format importers, preprocessing methods, feature algorithms using the IBEX developer studio.

2.1 S-IBEX standardization using IBSI-1 v6 guidelines

After a first usage of the tool by my research group in a technical work investigating feature stability on PET[58], I deemed essential to determine whether IBEX was IBSI-compliant or not. To this end, I identified any nonconformities between IBEX code and IBSI guidelines (version 6)[59], in order to develop a new IBEX version that implemented correctly, or introduced for the first time, all feature families and pre-processing steps that were standardize by IBSI up to that time.

The new version has been called standardized-IBEX, in short S-IBEX (Figure 2.2), and was in turn made publicly available (https://github.com/abettinelli/SIBEX_Source)[60].

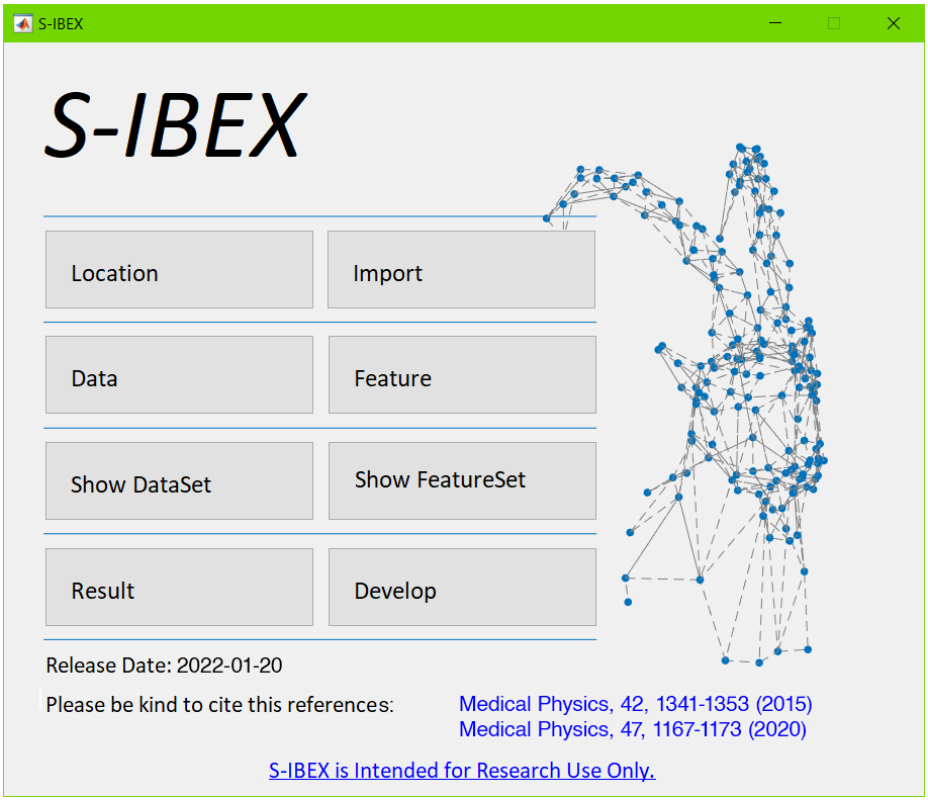

Figure 2.2. S-IBEX main window.

Interventions were implemented following the same ordering of the IBSI-1 manual when presenting the standardized definitions (section 1.3.2) and will now be reported here in the same way.

2.1.1 Contour-to-binary-mask conversion

Since the main input format of the tool is DICOM, contour-to-binary-mask conversion is necessary. IBSI suggests determining the mask by including, slice-by-slice, all the voxels whose centres lie inside

the ROI contour. Differently, IBEX also incorporates most voxels crossed by the ROI contour. Mask differences among the two methods (Figure 2.3) may have an impact on feature values. For contour-to-binary-mask conversion in S-IBEX, the IBSI-advised crossing-number algorithm[39] was integrated using an efficient method based on the point-in-polygon test[61].

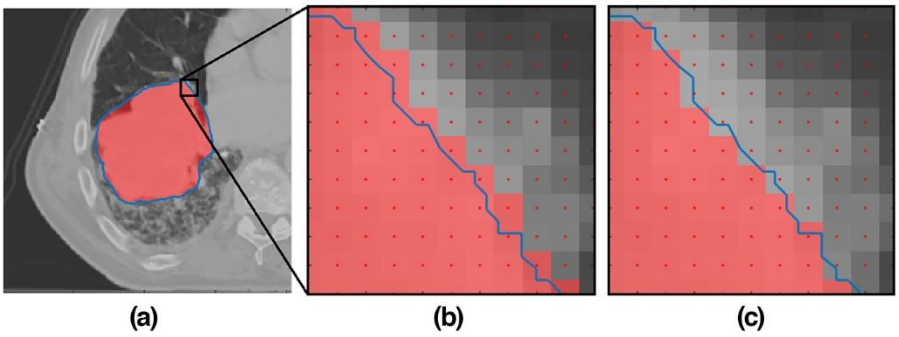

Figure 2.3. (a) A representative slice of the region of interest (ROI) withing the IBSI radiomic phantom. A magnification of image biomarker standardization initiative (b) and image biomarker explorer (IBEX) (c) results of contour-to-binary-mask conversion. S-IBEX results equal the initiative's ones (b). Voxel centres are displayed as dots, the contour as a polygonal curve. ROI voxels are shaded in red. IBEX mask contains a higher number of ROI voxels with respect to IBSI.

2.1.2 Resampling

IBEX implements a grid alignment technique that is different from the aforementioned IBSI-standardized methods and do not explicitly make use of δ, the threshold parameter to recover Boolean masks after interpolation. To solve these non-compliances, interpolation has been re-implemented in S-IBEX: in addition to default IBEX parameters (e.g., voxel sizes and 2D/3D option), S-IBEX gives the user the possibility to choose from several interpolation methods (linear, nearest neighbourhood, cubic, spline, or MATLAB makima) and to set the grid alignment technique ("fit to original grid", "align grid origins", or "align grid centres") and δ.

2.1.3 Re-segmentation

Re-segmentation has been implemented in S-IBEX as two optional preprocessing modules: range re-segmentation and intensity outlier filtering. According to IBSI, the former requires an input range that can be specified both as a closed interval $[a, b]$ or a half-open interval $[a, \infty)$, the latter needs a multiplying factor α so that the range is defined as $[\mu - \alpha * \sigma, \mu + \alpha * \sigma]$, where μ and σ are

respectively the mean and standard deviation of the voxel values inside the ROI. The methods may eventually be used in combination. S-IBEX has been adapted to handle the outcome of re-segmentation by employing the two masks suggested by IBSI: the *morphological mask* retains the original morphology while the *intensity mask* stores the re-segmented mask.

2.1.4 Gray-level discretization

Although IBEX includes discretization both as a preprocessing step and as section of the parent code of feature families, none meets the IBSI requirements. Therefore, discretization has been re-implemented in S-IBEX, as specified by IBSI, in both the form of FBN and FBS.

2.1.5 Feature families & aggregation methods

Except for the morphological family, implementation differences between IBEX and IBSI-1 are mostly due to parent data of feature families and not to feature definitions. Only minor changes have been necessary in feature definitions for variance, kurtosis, median absolute deviation, grey-level nonuniformity, run length nonuniformity, contrast, and texture strength. Because of the PINNACLE format, where water is given a CT number value of 1000, IBEX features extracted from CT are not invariant to intensity offsets will differ from those obtained following IBSI guidelines. Thus, in S-IBEX, the water CT number is reset to zero before any feature extraction. The interventions that were performed on feature families are summarized below:

- MF family: while IBEX handles volumes as a collection of voxels, each with its own volume, IBSI treats volumes both as a set of voxel-centre coordinates and a mesh-based representation of the surface. Furthermore, IBEX uses centimetres instead of millimetres as working units. In S-IBEX, both aspects have been conformed to the standard, and all feature definitions have been re-implemented.
- IS and IH families: water CT number resetting was necessary for both feature families. IBSI compliant discretization has been adopted for IH family.
- GLCM family: the IBEX approach to GLCM matrix calculation has been modified to support the five aggregation methods defined for directionally dependent feature families (2D:avg, 2D:mrg, 2D:vmrg, 3D:avg, and 3D:mrg). Methods identified with 'avg' average features extracted from

different textural matrices, while those identified by 'mrg' merge textural matrices before feature extraction.

- GLRLM family: in IBEX, textural matrix extraction is only implemented in a by-slice manner. In S-IBEX all five aforementioned approaches have been implemented.
- NGTDM family: the IBEX approach to NGTDM matrix calculation has been modified to support the three aggregation methods defined for rotationally invariant feature families (2D, 2.5D, and 3D).
- LI, IVH, GLSZM, GLDZM, and NGLDM: those families are not available in IBEX and have been implemented in S-IBEX following IBSI guidelines. The filtering step required by LI is achieved using a proper ellipsoidal mean convolution filter as suggested by IBSI. As for textural features, matrices can be derived using three aggregation methods (2D, 2.5D, and 3D).

The complete list of all S-IBEX features, grouped by type of intervention and by family, can be found in appendix Table B.1. The detailed definition of each feature and explanation can be found in the IBSI-1 manual.

2.1.6 Validation of S-IBEX

To validate S-IBEX, features were extracted from both the *IBSI digital* and *radiomic phantoms* employing the configurations proposed by IBSI itself in its guidelines. Appendix Table A.1 summaries the parameters of the six tested configurations. Specifically, config. Zero (no preprocessing) is defined for the *IBSI digital phantom* while config. A, config. B, config. C, config. D, and config. E are for the *IBSI radiomic phantom*. As IBEX does not offer the option for tricubic spline interpolation, Config. E was excluded from IBEX testing. The results of feature extraction were rounded to the third significant digit (which is the precision of reported IBSI values) and then compared to their IBSI benchmark. Feature extraction from the *digital phantom* is designed to evaluate the mere implementation of features. In addition, extraction from the *radiomic phantom* allows testing the overall image preprocessing chain (gray-level discretization, re-segmentation, interpolation). Config. Zero, A, B, C, and D were tested for both software implementations and were employed to compare IBEX and S-IBEX values to IBSI reference values. Config. A, B, C, D, and E were used to validate S-IBEX in multiple scenarios.

The number of analysable features depend on the choice of both software and configuration: the software constrains the analysis on the extractable features, while the configuration limits the analysis to the features whose values are reported by IBSI. For config. Zero, where no preprocessing is applied, 75.2% of IBEX feature values are equal to their IBSI benchmark up to the third significant digit, whereas this percentage drops to 1.0%, or below, for config. A, B, C, and D (Table 2.1).

While showing that many IBEX feature definitions are IBSI compliant, these results stress the impact of IBEX preprocessing, which introduces non-negligible non-conformities with respect to the IBSI standard. On the other hand, almost all S-IBEX features are in agreement with the standard for each tested configuration, with and without preprocessing. Indeed, more than 98% of features values is equal to its benchmark and more than 99% lies within IBSI tolerance levels (Table 2.1) with maximum overall absolute percentage error as low as 0.90% (obtained for GLRLM "short run high gray-level emphasis" feature for config. C).

Table 2.1. Statistics describing feature extraction, grouped by software, using IBSI configurations and the two phantoms of IBSI guidelines v6. The number of analysable features depends on the choice of the software as well as the configuration.

	IBEX			**S-IBEX**		
Config.	***# of extractable features***	***% of features equal to IBSI***	***% of features in IBSI tolerance***	***# of extractable features***	***% of features equal to IBSI***	***% of features in IBSI tolerance***
Zero	101/349	75.2%	75.2%	349/349	99.4%	99.4%
A	101/206	1.0%	4.0%	206/206	98.1%	100.0%
B	101/202	0.0%	7.9%	202/202	99.0%	100.0%
C	101/210	1.0%	3.0%	210/210	98.6%	99.0%
D	101/210	1.0%	1.0%	210/210	99.5%	100.0%
E	0/208	0.0%	0.0%	208/208	99.5%	99.5%

2.2 S-IBEX update based on IBSI-1 v11 guidelines

IBSI-1 standardization process required few years for its completion and consequently the IBSI manual was periodically updated and shared on arXiv. So it was until 2020, when the definitive IBSI guidelines were redacted in their 11th version and published alongside the work of Zwanenburg et al.[13]. Reference feature values were also updated and, with respect to v6 benchmarks, v11 included 419 additional benchmarking values since, for each configuration, every appropriate aggregation method was considered. Radiomic software programs that were standardized using non-definitive versions of the guidelines had, by necessity, to be updated following the latest v11 version.

This section presents the interventions that were performed to update S-IBEX compliancy to IBSI guidelines v11.

2.2.1 Differences between IBSI v6 and v11 guidelines

S-IBEX software was originally built and validated using IBSIv6 documentation and benchmark values (section 2.1). Therefore, IBSI v11 documentation[41] has been checked against IBSI v6 to highlight the updates required to conform the software to the latest guideline. Differences emerged both in image pre-processing steps and aggregation methods (Table 2.2): in particular, grey-level discretization methods were updated while 2.5Davg aggregation method was implemented ex-novo.

Table 2.2. Differences between IBSI guidelines v6 and v11.

	IBSI v6 guidelines	***IBSI v11 guidelines***
FBN	$X_{d,k} = \begin{cases} 1, & X_{gl,k} = X_{gl,min} \\ \left\lfloor N_g \frac{X_{gl,k} - X_{gl,min}}{X_{gl,max} - X_{gl,min}} \right\rfloor, & X_{gl,k} > X_{gl,min} \end{cases}$	$X_{d,k} = \begin{cases} \left\lfloor N_g \frac{X_{gl,k} - X_{gl,min}}{X_{gl,max} - X_{gl,min}} \right\rfloor + 1, & X_{gl,k} < X_{gl,max} \\ N_g, & X_{gl,k} = X_{gl,max} \end{cases}$
FBS	$X_{d,k} = \begin{cases} 1, & X_{gl,k} = X_{gl,min} \\ \left\lfloor \frac{X_{gl,k} - X_{gl,min}}{\omega_b} \right\rfloor, & X_{gl,k} > X_{gl,min} \end{cases}$	$X_{d,k} = \left\lfloor \frac{X_{gl,k} - X_{gl,min}}{\omega_b} \right\rfloor$
'2.5D:avg'	not defined	features are computed from a single matrix per slice, obtained by merging 2D directional matrices per direction. Eventually features are averaged.

2.2.2 Validation of the upgraded S-IBEX

To validate the changes on S-IBEX, all features proposed by IBSI v11 guidelines features were extracted from both the *IBSI digital* and *radiomic phantoms* employing the benchmarking configurations. The

parameters of the six tested configurations did not change among IBSI versions and are reported in appendix Table A.1. The results of feature extraction were rounded to the third significant digit and inputted in the benchmark excel sheets provided by IBSI. Config. Zero, A, B, C, and D were tested for both S-IBEX versions (before and after the update) and employed to compare IBEX values to IBSI references. Percentages of features equal to the standard for both S-IBEX version are presented in Table 2.3.

Table 2.3. Statistics describing feature extraction, grouped by software version, using IBSI configurations and the two phantoms of guidelines v11. The number of analysable features depends on the software version as well as the configuration.

	S-IBEX			**S-IBEX** (after update)		
Config.	***# of extractable features***	***% of features equal to IBSI***	***% of features in IBSI tolerance***	***# of extractable features***	***% of features equal to IBSI***	***% of features in IBSI tolerance***
Zero	441/482	90.9%	90.9%	482/482	100%	100%
A	305/346	78.9%	85.0%	346/346	99.4%	99.7%
B	305/346	87.3%	87.3%	346/346	100%	100%
C	210/210	97.6%	99.0%	210/210	100%	100%
D	210/210	100%	100%	210/210	100%	100%
E	210/210	17.1%	19.5%	210/210	100%	100%

The distribution of matching/partial matching and no-matching features can be appreciated, stratified by configuration and feature family, in Figure 2.4. Only two features, "*Moran's I Index*" e "*Geary's C measure*", resulted in a partial- and no-match, respectively, for config. A. Since in other configurations these features resulted in a match with the reference, it is unlikely that the mismatch was due to an implementation error. It is interesting to note that, only for these two features, IBSI suggested that an approximation scheme may be required due to their $O(n^2)$ behaviour, and in turn, confidence intervals may be too strict.

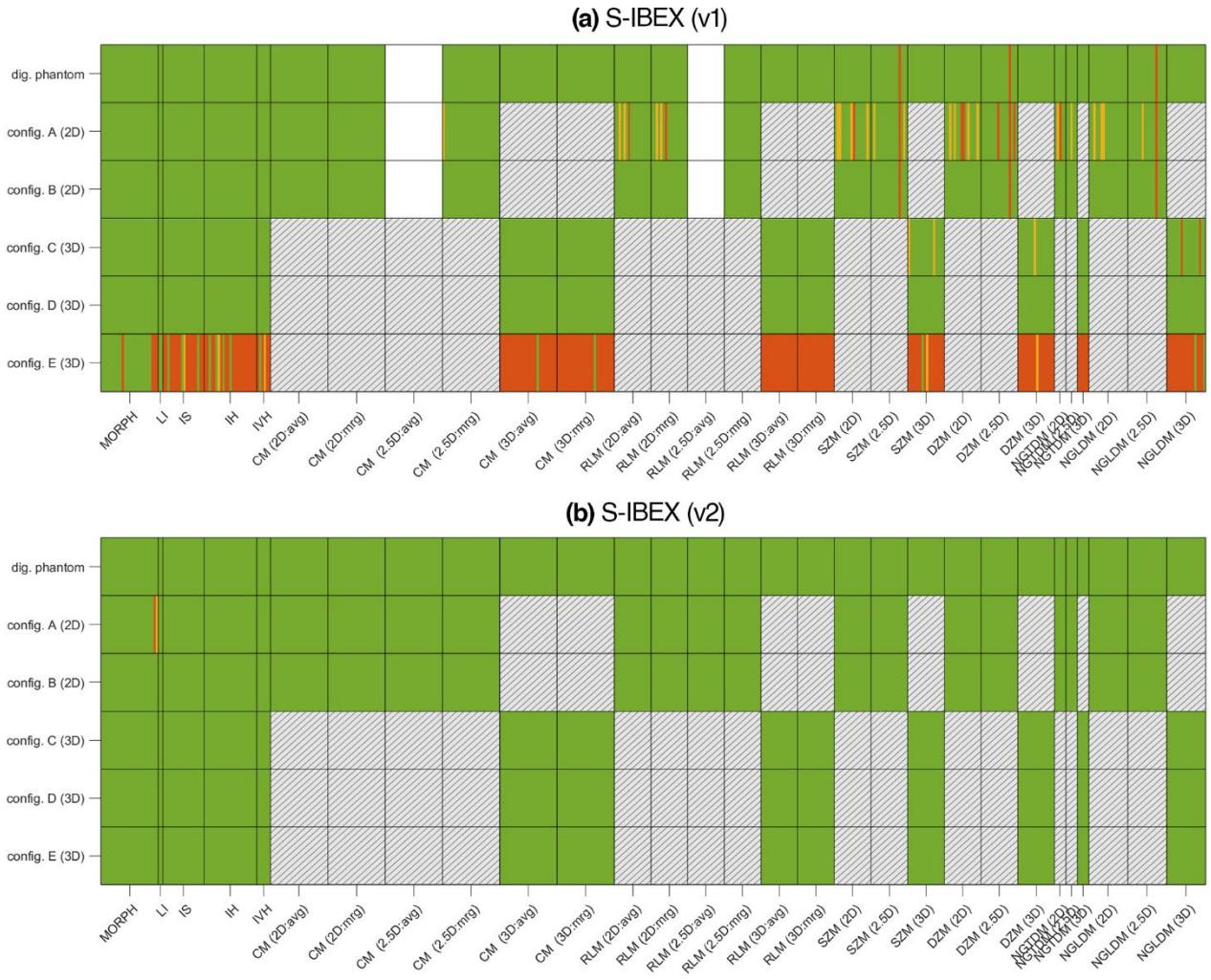

Figure 2.4. (a) Original S-IBEX and (b) updated S-IBEX results using IBSI benchmark values of IBSI v11 guidelines. Green: match, Yellow: partial match, Red: no-match, white: not computable, dashed: not covered by IBSI benchmark values.

2.3 S-IBEX update based on the IBSI-2 initiative

As previously described in section 1.3.3, IBSI-2 is an ongoing effort that will find its conclusion in early 2023. Unlike IBSI-1 guidelines, which I have employed, from a user perspective, to standardize/update my software tool, in IBSI-2 I was directly involved in the standardization process, providing the data obtained with my tool and implementation feedbacks to finalize the guidelines.

Even if IBSI-2 is not yet concluded, major standardization results have already been obtained, and the guidelines have been finalized. In this section, I will present both the S-IBEX update that included convolutional filters and the data I provided for the advancement of IBSI-2.

In the IBSI-2 phase 1, all teams were asked to 1) implement, in their own software, all the convolutional filters that a core group has drawn up in the draft manual and 2) benchmark their implementations using the set of phantoms of Figure 1.11. Eventually, after some iterations, both the manual became more detailed and implementations across teams converged to a consensus. All the convolutional filters and all the methods that were implemented in S-IBEX are reported in Table 2.4.

Table 2.4. Convolutional filters and processing methods implemented in S-IBEX.

METHODS	***PARAMETERS***
Padding	• Type (e.g., constant-value, nearest, periodization, mirror)
Rotation invariance	• Type (e.g., max pooling, average pooling)
Mean Filter	• 2D/3D • Support size M
Laplacian of Gaussian	• 2D/3D • Scale parameter σ^* • Filter size cutoff d: 4 by default
Laws Kernels	• 2D/3D • 1D kernels: Level (3-5), Edges (3-5), Spots (3-5), Wave (5) and Ripples (5). • Rotation invariance: FALSE by default • Energy map: FALSE by default ○ If TRUE mean filter is computed (support $2\delta + 1$ with $\delta = 7$ by default)
Gabor	• 2D/approximate-3D • Scale parameter σ^* • Aspect ratio γ^* • In plane orientation θ • Wavelength λ • Rotation invariance: FALSE by default, ○ if TRUE $\Delta\theta$ and pooling (max/avg) need to be specified
Separable Wavelets (undecimated)	• Type and order: (e.g., Haar, Coifflet 1, Daubechies 2) • 2D/3D • Filter combination (e.g., HLH) • Level of decomposition (e.g., 1st level) • Rotation invariance: FALSE by default ○ if TRUE pooling (max/avg) needs to be specified
Non-separable wavelets (undecimated)	• Method: Simoncelli • 2D/3D • Level of decomposition: e.g., 1st level
Riesz-LoG/Simoncelli	• The same parameters as LoG/Simoncelli • Level l: e.g., (0,2,0) • Aligned by structure tensor: FALSE by default

S-IBEX response maps were computed on the phantoms' set of Figure 1.11 following the 36 filtering configurations of IBSI-2 phase 1 (reported in appendix Table A.2), and subsequently uploaded to the dedicated submission platform (https://ibsi.radiomics.hevs.ch). All response maps are presented in Figure 2.5.

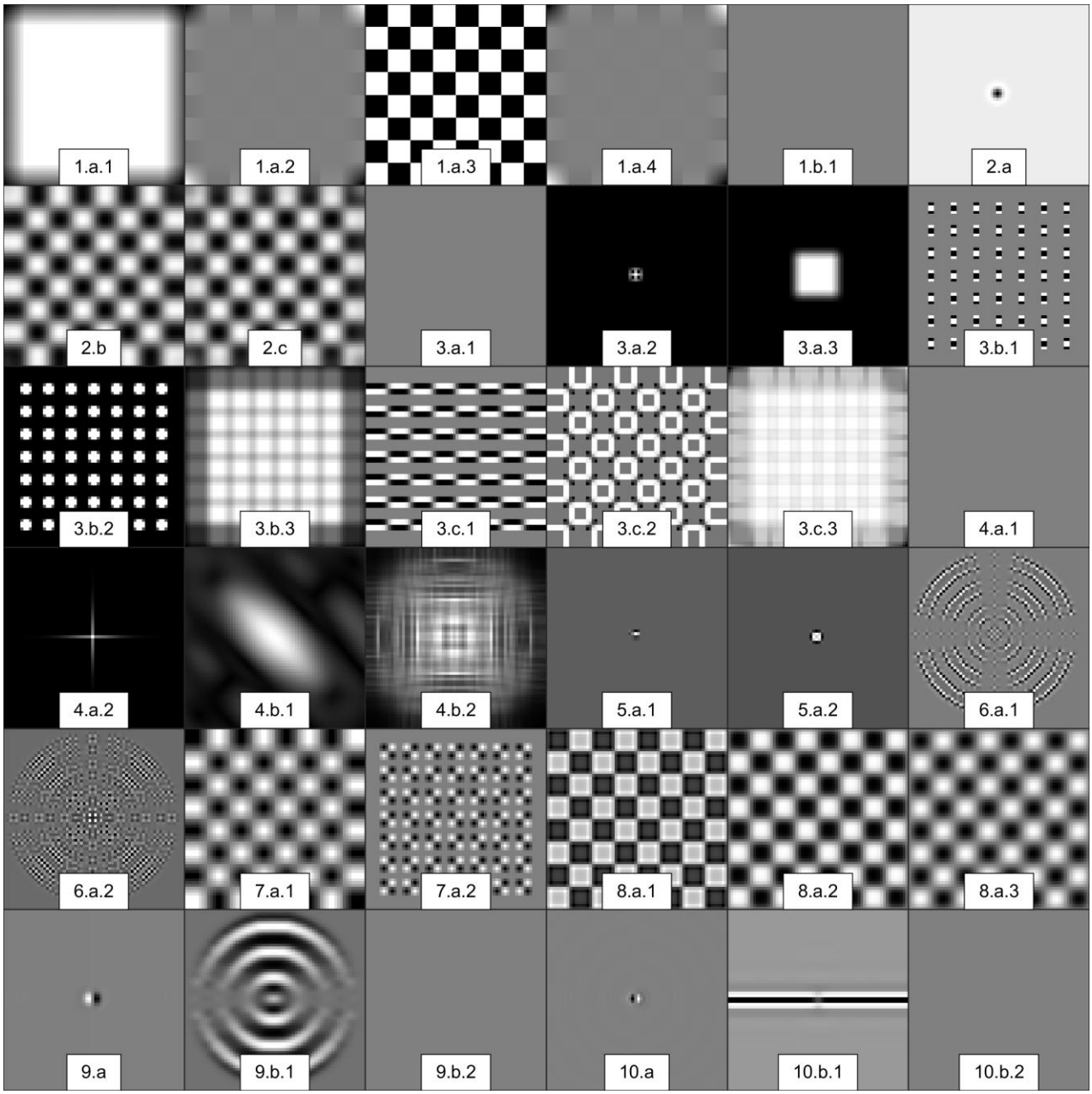

Figure 2.5. S-IBEX filtering results: central slices of the 36 response maps (Table A.2 for reference).

IBSI used Principal Component Analysis (PCA) to assess variations between all team's response maps at once. The first two components were sufficient to summarize the response of each team, allowing to identify outliers (an example is reported in Figure 2.6).

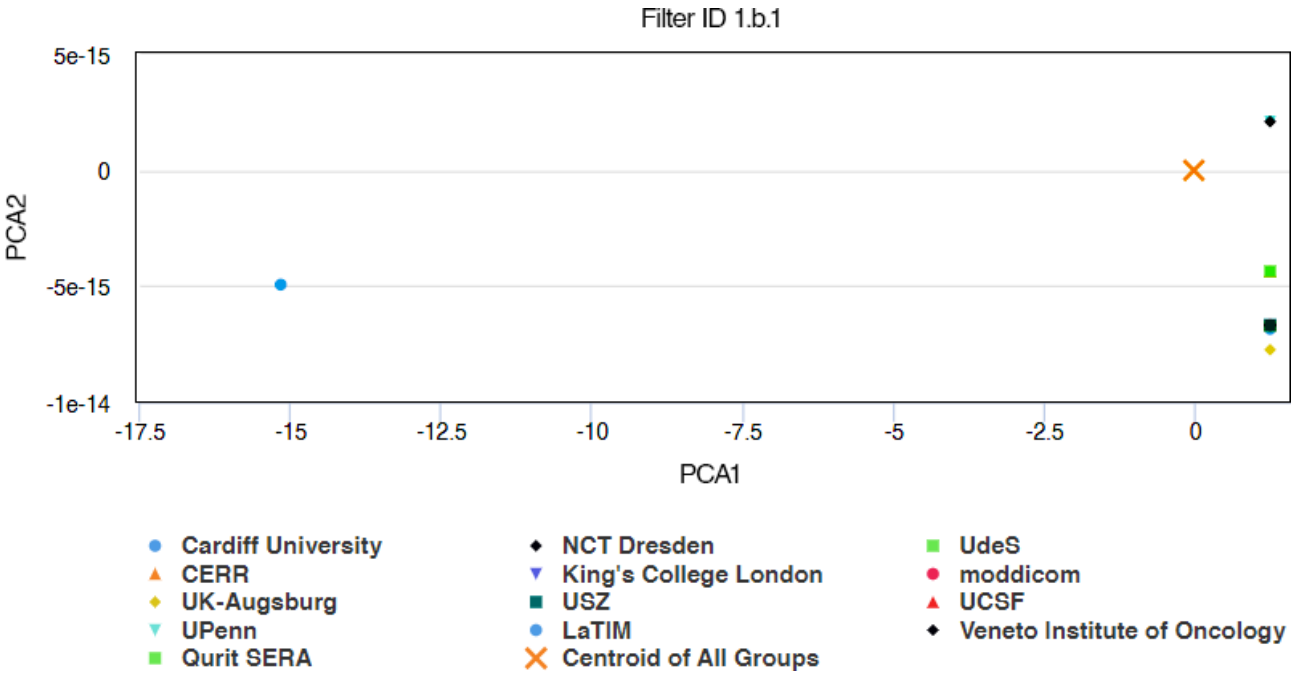

Figure 2.6. PCA plot of the response maps with ID 1.b.1 (Mean Filter on impulse phantom). LaTIM appears to be the only team who do not reached the consensus for this configuration. Please note that differences between other teams (visible on the second component) are negligible (~10^-15).

Eventually, consensus response maps (CRM) were obtained by averaging the response maps whose teams were in agreement, accordingly to PCA plots.

Pairwise comparison between each response map was used to quantify the consensus (less than three teams, weak; three to five, moderate; six to nine, strong; 10 or more, very strong) and to build Figure 2.7. From the figure we can appreciate that S-IBEX was found to have the highest number (together with NCT Dresden) of response maps in consensus (33 over the 34 submitted) with all the other teams while no CRM was available for one filter configuration (ID 10.a).

As for IBSI-2 phases 2 and 3, analysis will be carried out by the initiative in early 2023 to define and validate, based and the team submissions, the acceptable tolerance level and benchmark values for radiomic features extracted after convolutional filtering.

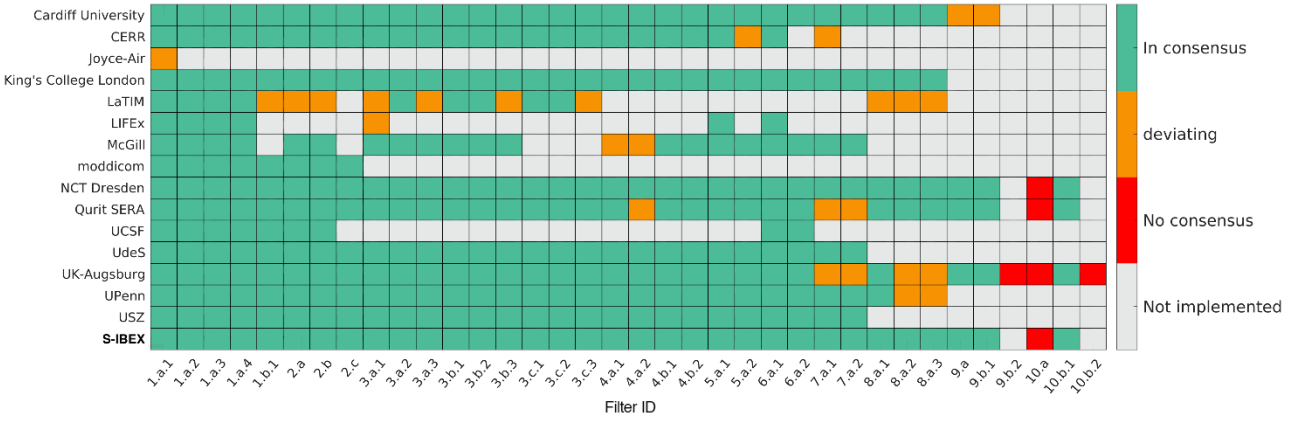

Figure 2.7. Latest consensus map. Consensus was defined if at least three different teams produced the same response map.

My contribution to the latest phases of the initiative was carried out by employing S-IBEX to filter both the *IBSI radiomic phantom* and the multimodal dataset accordingly to appendix Table A.3 and Table A.4, and by extracting intensity-based statistical features. The 9 response maps required by IBSI-2 phase 3 are visible in Figure 2.8, Figure 2.9, and Figure 2.10 for the CT, MR and PET dataset respectively.

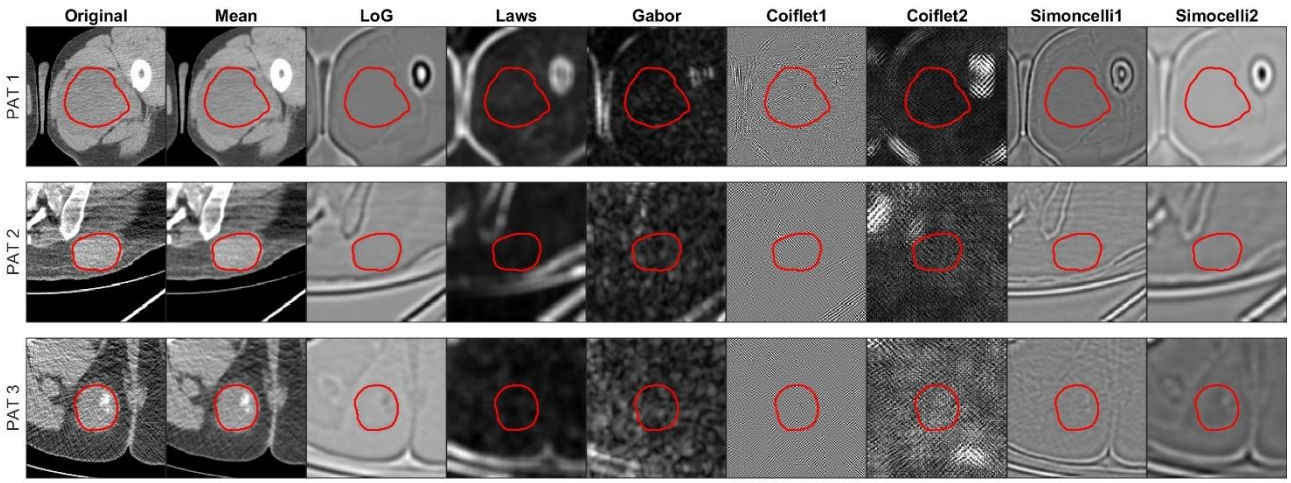

Figure 2.8. Representative CT slice for three patients (of the sarcoma dataset) and their response maps obtained in Phase 3.

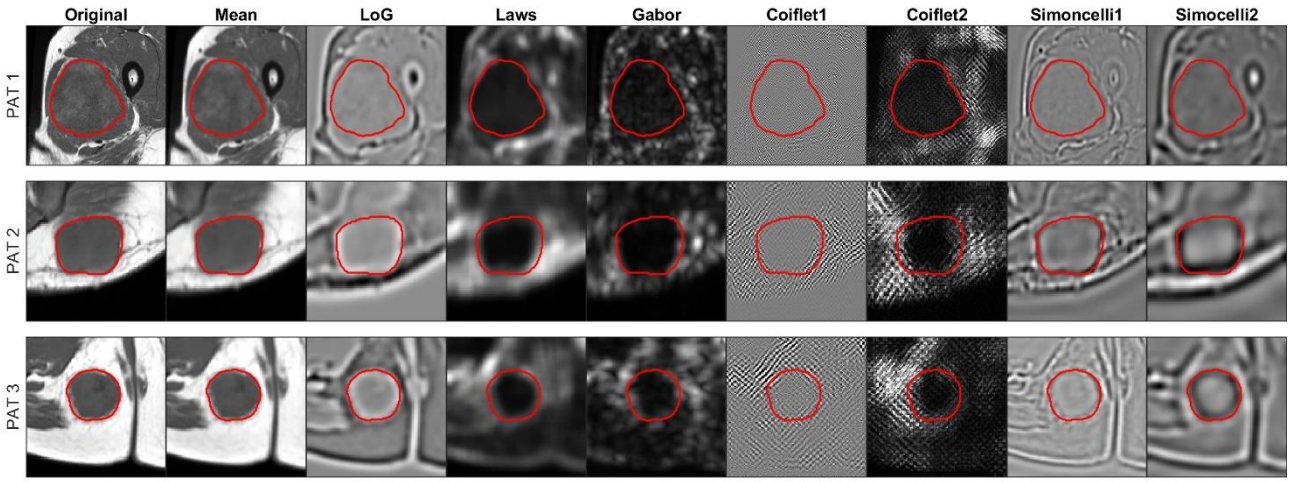

Figure 2.9. Representative MR slice for three patients (of the sarcoma dataset) and their response maps obtained in Phase 3.

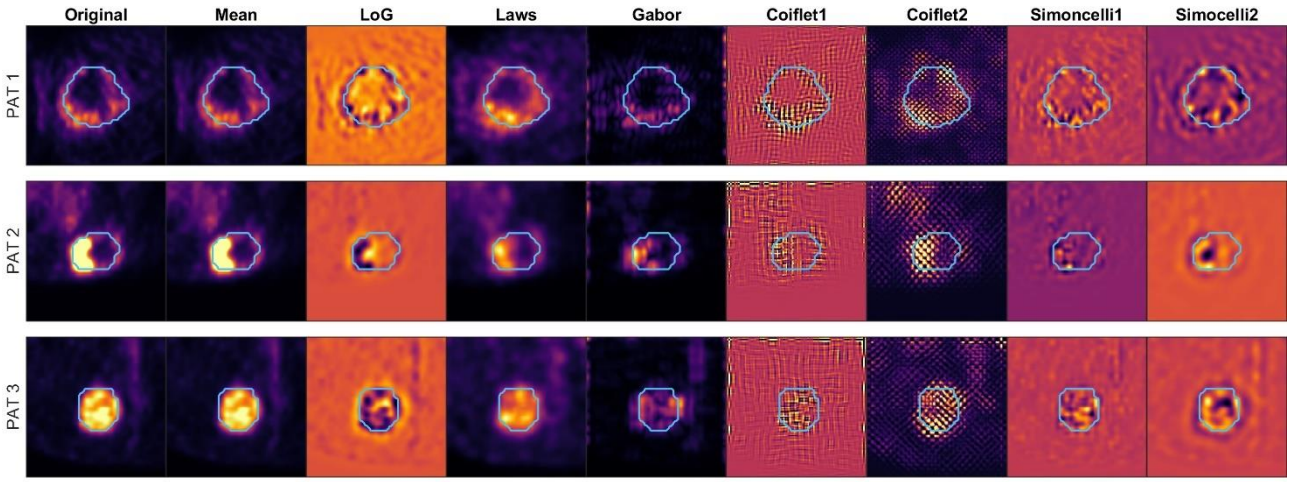

Figure 2.10. Representative PET slice for three patients (of the sarcoma dataset) and their response maps obtained in Phase 3.

Chapter 3: Comparison of S-IBEX to other radiomic extractors

The lack of standardization in the definition and calculation of radiomic features, their ambiguous nomenclature and the limited reproducibility of radiomic studies[8,17,62,63] have all impeded the adoption of radiomics within clinical practice. Some of these concerns were addressed by the IBSI-1[13] which published its reference manual[64] comprising the definition of 169 standardized radiomic features, reporting guidelines on how to perform image pre-processing and the two *IBSI digital* and *radiomic phantoms*[65] with their respective benchmark feature values to assess the accuracy of software tools for radiomic analysis. Because of the raised awareness of the need for standardization, several radiomic tools have begun to conform to the IBSI-1 guidelines[60,66–68]. However, a number of radiomic studies are based on in-house or public software with an unclear level of standardization concerning at least one of the several aspects involved for feature extraction (e.g., pre-processing methods, availability of tuning parameters).

In this context, I started and coordinated a collaboration between four Italian clinical research institutes, the "Italian multicenter Shared Understanding of Radiomic Extractors" (ImSURE) group, with the aim 1) to compare the IBSI-compliance of S-IBEX with some commonly available radiomic software tools and 2) to investigate the causes of possible discrepancy by designing a systematic workflow of radiomic features extraction performed on two custom digital phantoms comprising multiple regions of interest (ROIs) with various volumes and shapes (the ImSURE digital phantoms). For completeness, the entire set of IBSI-1-standardized radiomic features was considered and all possible combinations of pre-processing steps/aggregation methods were explored. Finally, I exhaustively investigated the causes of discrepancy among the programs under consideration and discussed, for each software, whether differences were attributable to the implementation of a limited number of features or to a non-IBSI-compliant implementation of technical aspects. Limited flexibility in parameter setting, as well as software discrepancies, are non-negligible factors for the reproducibility of the radiomic features and for the general validity of the models proposed in the radiomic literature.

3.1 The other radiomic extractors

Within the ImSURE initiative, S-IBEX was compared to six other radiomic tools, of which five were open-source (MIRP[69], RaCaT[68], SERA[66], PyRadiomics[67,70], and RadiomiCRO[71]) and one was commercial

(SOPHiA DDM for Radiomics[72]). For the study, the latest version available of the software tools was used (i.e., MIRP v1.0.2, SOPHiA DDM v2.2.0, RaCaT v1.18, SERA v2.1 and Pyradiomics v3.0.1). The inclusion criteria were: 1) that the software was self-declared IBSI-1 compliant and/or IBSI-1 participant and 2) that there was a consolidated experience in the software tuning by at least one of the four centres participating in the study. Their widespread usage in the radiomic field, and their flexibility in setting various parameters were also considered as selection factors. Table 3.1 reports salient characteristics of these tools. To ensure correctness of their usage, all software programs were picked from the tools readily available at the project's participating centres and, when possible, were assigned to two centres for a consensus feature extraction.

Table 3.1. Software packages included in the study. BO = IRCCS Azienda Ospedaliero-Universitaria di Bologna, CRO = Centro di Riferimento Oncologico di Aviano, IOV = Veneto Institute of Oncology, IRST = Istituto Romagnolo per lo Studio dei Tumori "Dino Amadori".

Software	***IBSI-compliant****	***Version***	***Language***	***Data format***	***Characteristics***	***Documentation***	***Assigned centers***
S-IBEX	self-declared	v2	MATLAB	DICOM	Open source	✓	IOV, IRST
MIRP	IBSI-participant	v1.0.2	Python	DICOM	Open source	✗	CRO, IOV
SOPHiA DDM	self-declared	v2.2.0	-	DICOM	Commercial	✓	IRST, IOV
RaCaT	IBSI-participant	v1.18	C++	DICOM, NRRD, NIfTI	Open source	✓	IOV, BO
SERA	IBSI-participant	v2.1	MATLAB	DICOM	Open source	✓	CRO, IRST
Pyradiomics	IBSI-participant	v3.0.1	Python	NRRD, NIfTI	Open source	✓	BO, IRST
RadiomiCRO	self-declared	-	MATLAB	DICOM	In-house	✗	CRO

Note. — DICOM = Digital Imaging and Communications in Medicine, NIfTI = Neuroimaging Informatics Technology Initiative, NRRD = nearly raw raster data.

** Software that were used in IBSI for the standardization of preprocessing and feature calculation were considered IBSI participants.*

Due to the abovementioned criteria some software programs were not include in the study, such as the commercial tool HealthMyne® that, despite being available and self-declare IBSI-1 compliant, did not allow the user for parameter tuning, while Moddicom and MITK tools were not at the disposal in the involved centres due to unsuccessful installation procedures.

MIRP was the leading software used in the IBSI-1 study; it is based on Python language and specific pre-processing parameters can be set by filling a configuration file. RaCaT is a radiomic calculator written in C++ as a standalone executable. Extraction parameters have to be set by filling a specific configuration file. SERA is a library developed for MATLAB and only works on medical images already imported into MATLAB, thus the importing step is left to the user and requires external functions, while parameters must be set within the SERA main code. Pyradiomics is presently one of the most common radiomic library and similarly to MIRP, it is written in the Python language. The specific parameters for feature extraction can be set both inside the code or with a dedicated configuration file. SOPHiA is a commercial solution by the company SOPHiA Genetics while RadiomiCRO is the in-house developed tool of *Centro di Riferimento Oncologico di Aviano*.

3.2 Assessment on the *IBSI digital* and *radiomic phantoms*

To quantify the level of IBSI-compliance of the six software packages (S-IBEX IBSI-1 compliance assessment is already presented in section 3.3.2) radiomic features were again extracted from two *IBSI digital* and *radiomic phantoms*[65] using the configuration settings of appendix Table A.1: all five IBSI parameter configurations were considered for the *radiomic phantom*, labelled A to E, and characterized by either 2D or 3D feature aggregation and either fixed bin size - FBS or fixed bin number - FBN discretization. In total, 482 feature values were extracted for the *IBSI digital phantom* applying neither interpolation nor discretization, whereas 1322 radiomic feature values were computed using all five configurations for the *IBSI radiomic phantom*. The calculated radiomic features were compared with the corresponding IBSI-1 benchmark values (v11) and classified into 'matching' (differences ≤ IBSI-reported tolerance), 'partial matching' (differences ≤ three times the IBSI-reported tolerance) or 'no matching' (otherwise), accordingly to the evaluation criteria proposed by the initiative[13]. Features that were not implemented within a tool were labelled as 'missing'.

3.2.1 Results of the assessment on the IBSI phantoms

For each software program, the resulting percentage of 'matching', 'partial matching', 'no matching' and 'missing' features, both for the IBSI *digital and radiomic phantoms*, are shown in Figure 3.1a and Figure 3.1b, respectively (for the latter, results have been aggregated over the five parameter configurations).

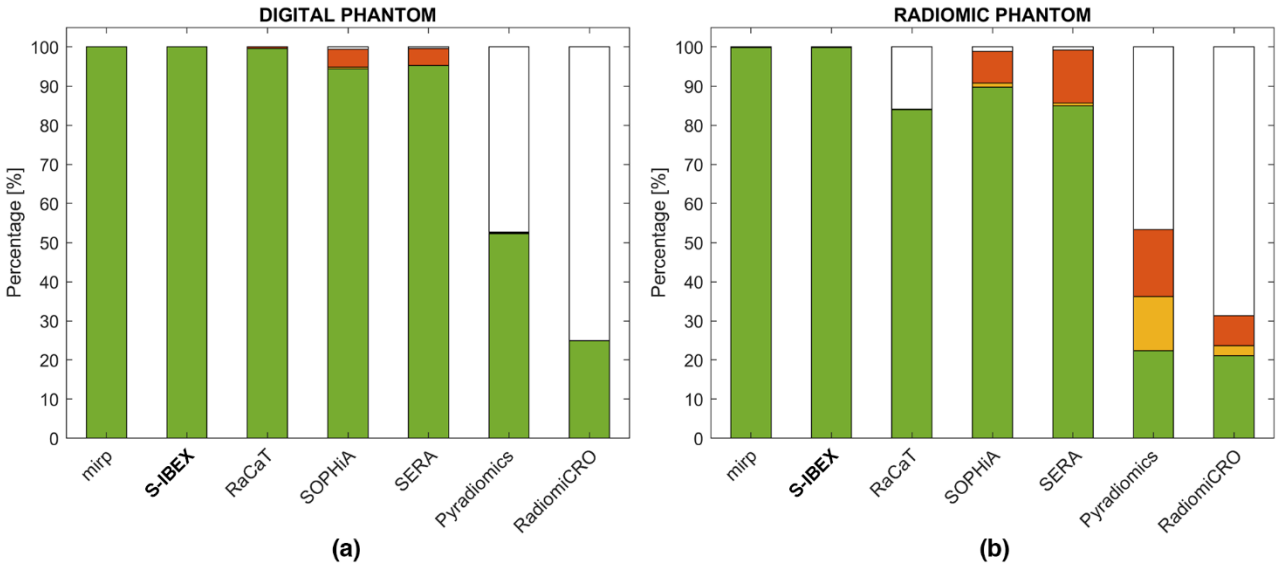

Figure 3.1. Percentages of "matching" (differences below the IBSI-reported tolerance, in green), "partial matching" (differences below three times the IBSI-reported tolerance, in yellow), "no matching" (otherwise, in red) feature values obtained for each software package on the IBSI digital *(a) and* radiomic *(b)* phantoms. *The feature values that could not be calculated within a tool were labelled as 'missing' (white). The percentages for the* radiomic phantom *were averaged across the five IBSI configurations. SOPHiA = SOPHiA DDM for radiomics.*

For the *IBSI digital phantom*, results stratified per feature family are reported in Figure 3.2 while for the *IBSI radiomic phantom*, results stratified by configuration are reported in Figure 3.3.

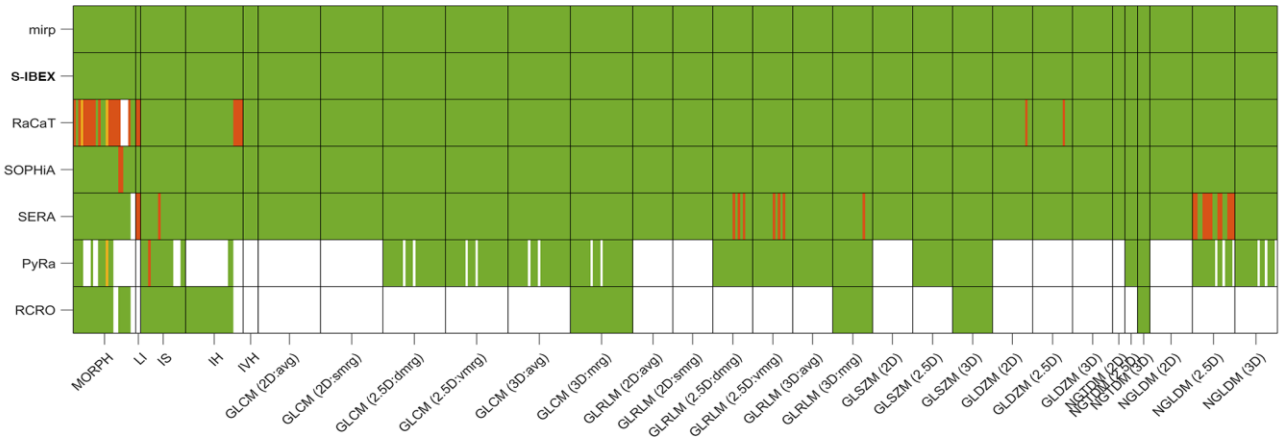

Figure 3.2. Cases of matching (green), partial matching (yellow), no matching (red), and missing (white) features for each software tool, feature family, and type of aggregation method on the IBSI digital phantom. *For each program, only matching and partial matching features were maintained for the analysis on the isotropic and anisotropic phantom. SOPHiA = SOPHiA DDM for radiomics; PyRa = Pyradiomics; RCRO = RadiomiCRO.*

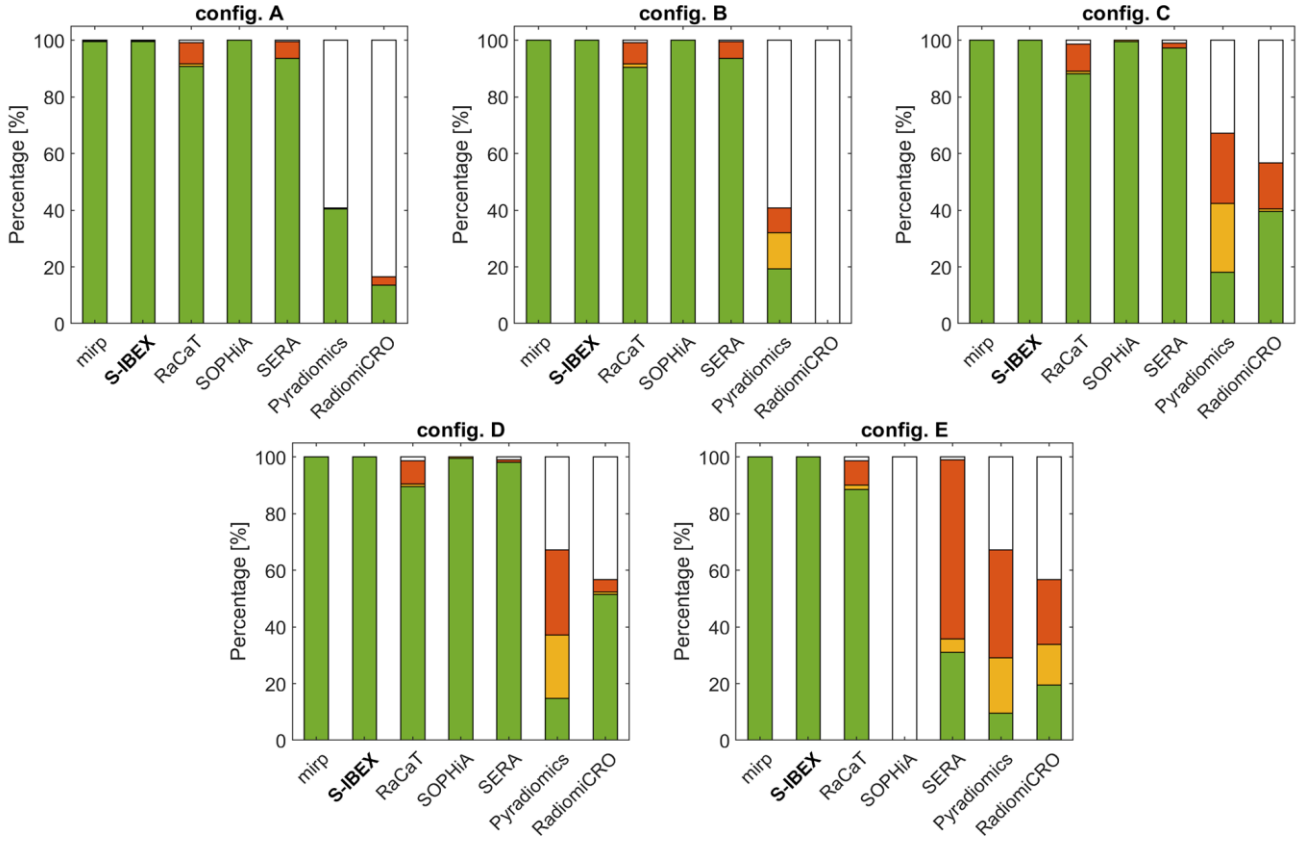

Figure 3.3. Cases of matching (green), partial matching (yellow), no matching (red), and missing (white) features for each software tool on the IBSI radiomic phantom in the five different parameter configurations (A, B, C, D, and E). SOPHiA = SOPHiA DDM for radiomics; PyRa = Pyradiomics; RCRO = RadiomiCRO.

On the *digital phantom* S-IBEX, MIRP, RaCaT, SERA and SOPHiA all achieved percentages of matches above 94%, while PyRadiomics and RadiomiCRO had 52% and 25% of 'matching' features, respectively, due to 'missing' feature values. RaCaT, SERA and RadiomiCRO all exhibited a slight decrease in the percentage of matching features (90%, 85%, and 21%, respectively) on the *radiomic phantom*, while PyRadiomics showed a marked increase in partial matches and no matches. SOPHiA presented 16% missing features as config. E is currently not obtainable.

3.2.2 Discussion

The performance of seven self-declared IBSI-compliant software packages were analysed thanks to IBSI tools. The analysis on the *IBSI digital phantom* revealed that all programs achieved high percentages of 'matching' features, indicating a high standardization level in terms of radiomic feature implementation. However, programs showed different degrees of feature completeness, with PyRadiomics and RadiomiCRO having the highest number of non-computable feature values. The *IBSI radiomic phantom* analysis allowed us to consider the effects of multiple factors, such as image interpolation and intensity discretization, and highlighted the limited flexibility in the parameter settings of some tools. By comparing our results with the ones of the IBSI study[13], we found them in accordance for MIRP and RaCaT, whereas SERA showed a higher percentage of 'matching' features on both the *IBSI digital and radiomic phantoms* in configurations A-D, but only a partial improvement in configuration E. Instead, PyRadiomics presented a lower 'matching' percentage and higher percentages of 'no match' and 'partial match' on the *radiomic phantom*. These differences could be the result of a missing update of either the software documentation or the version used by the IBSI.

3.3 Assessment on the ImSURE digital phantoms

If we consider that the *IBSI digital and radiomic phantoms* have just one region of interest each and hence they only allow a two-sample evaluation of each radiomic feature value, which was an essential requirement to ease the standardisation process, we understand that some implementation differences could go unrecognised, especially on the less complex *IBSI digital phantom*.

IBSI-1 entailed the standardisation of several open- and closed-source radiomic software and the publication of a number of studies aiming to assess feature reproducibility among radiomic tools[73–75]. Some of these works also developed heterogeneous digital phantoms by accounting for different ROI shapes and patterns[75] to overcome the limited casuistry of the two IBSI phantoms. However, the

majority of these works only considered a subset of the 169 IBSI-standardised features, typically those in common to the considered set of radiomic tools, and none of them systematically investigated software capability to set all combinations of pre-processing choices (e.g., interpolation, discretisation), nor their sensitivity to different ROI morphologies (e.g., shape and volume).

This was the reasoning that motivated me in creating the ImSURE phantoms. Having 90-ROIs each, the likelihood of fortuitously obtaining matching values, when the underling implementation is different, is reduced. The phantoms were designed to have both isotropic and anisotropic voxel size and to include ROIs with different textures and morphologies. The phantoms are thought to be used in conjunction with a systematic workflow of feature extraction that allows a meticulous software investigation and comprises the calculation of all the 169 IBSI-standardised radiomic features and includes all the possible combinations of pre-processing and feature aggregation methods, for a total of 919 feature values computed for each ROI. Moreover, the casuistry of ROIs enables the use of statistical tests to examine the effects of novel parameters (such as shape and volume) on software agreement.

Thanks to the ImSURE phantoms and the systematic workflow of feature extraction, it was possible to further investigate S-IBEX agreement with the other six tools, testing the reproducibility of features across software programs for different pre-processing choices (e.g. interpolation and discretization)[76], feature aggregation methods (e.g. 2D, 2.5D, or 3D) and ROI characteristics (e.g. volume and shape).

3.3.1 Design of ImSURE digital phantoms

3.3.1.1 Image retrieval, anonymisation and resample

A CT acquisition from skull base to mid-thigh of a patient randomly selected from a database of patients who signed informed consent was retrieved from the picture archiving and communication System (PACS) and anonymised using the Python library *DicomAnonymizer* (https://github.com/KitwareMedical/dicom-anonymizer). In addition to the standard anonymisation of DICOM fields offered by the routine, '*Instance Creation Date*' and '*Instance Creation Time Attribute*' fields were also overwritten to delete any reference to the date of the examination.

The original CT image had anisotropic voxels with a dimension of 0.98x0.98x3.00 mm and was used to create the '*ImSURE anisotropic phantom*'. An IBSI-compliant trilinear interpolation was then

applied to the CT image to generate a second image with isotropic voxel dimension of 1.00x1.00x1.00 mm, which was used to build the '*ImSURE isotropic phantom*'.

3.3.1.2 Design of the ROIs

Each phantom was designed to contain 10 repetitions of 9 base morphologies, obtained as a combination of 3 shapes (i.e., cube, sphere and bean-like) and 3 volumes (i.e., small - *0.125* cm^3, medium - *1* cm^3, large - *8* cm^3), for a total of 90 ROIs. Cube and sphere shapes were chosen because of their simple and symmetric geometry, while the bean-like shape was introduced to mimic a morphology closer to that of a clinical ROI. The choice of the nine morphologies was made as part of the study design, and the necessity of having enough ROIs (at least thirty) for statistical analyses when stratifying by shape or volume drove the number of repetitions. These decisions reflected a trade-off between the need to conduct a systematic study and to keep an overall manageable number of ROIs.

ROIs were created with pinpoint accuracy down to the single-voxel level following three main steps: 1) design of surface models for each base morphology, 2) conversion of the surface models to binary masks and 3) definition of the ROI contour set (as required by the DICOM standard). The process is depicted in Figure 3.4.

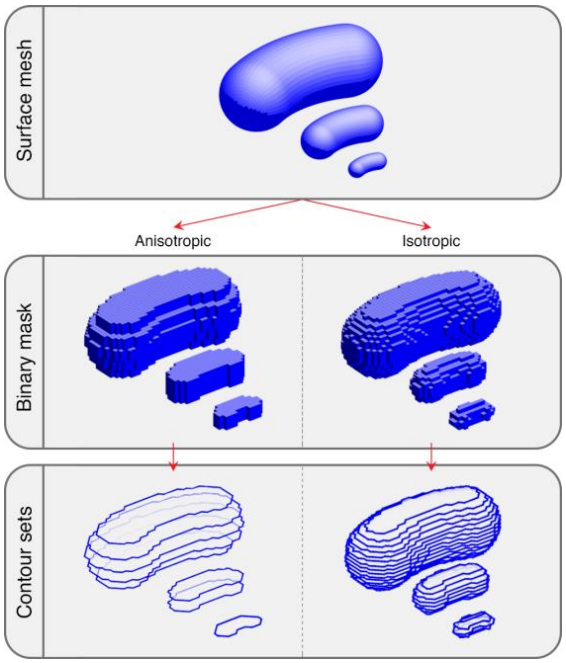

Figure 3.4. Steps of ROI creation for the bean-like shape. (upper) Definition of bean-like 3D surface meshes; (middle) Creation of binary masks for both the anisotropic and the isotropic ROI; (lower) Definition of ROI contour sets.

Surface meshes were designed using Blender software[77]. The Blender built-in "primitive" meshes were used to realise the cubical and spherical ROI morphologies, while for the creation of the bean-like geometry, the extremities of a 50-degree torus section were capped with two half-spheres (Figure 3.5). Each shape was then resized to obtain 3 predefined volumes (i.e. 0.125 cm^3, 1 cm^3, 8 cm^3) and was exported as a separate *.stl* file.

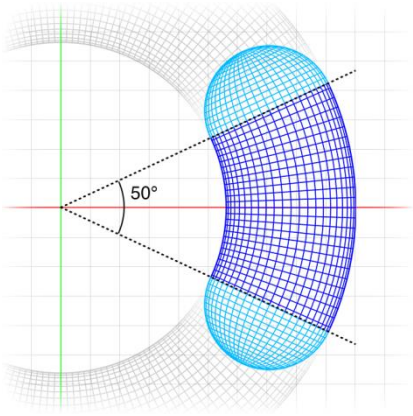

Figure 3.5. Orthographical top view of the bean surface mesh, obtained by capping a 50-degree torus section (blue) with two hemi-spheres (light blue).

The 9 surface models were imported into MATLAB (the MathWorks, Natick, 2020a) and converted into binary masks. For this purpose, two three-dimensional point grids representing both the anisotropic and isotropic voxel centres were built and intersected with the nine surface meshes. Voxels of the grid whose centre fell inside the mesh were set to 1, otherwise to 0. Afterwards, ten repetitions for each of the nine binary masks were positioned over the space of each CT image and were axially arranged to create 18 different groups of five ROIs each (see Figure 3.6).

Eventually, only the texture of the underlying CT image within and around each ROI (with a margin of 6 mm) was kept, while the surrounding voxels were censored by setting their intensity to -1024 Hounsfield Unit. The resulting censored images and the two binary masks constitute the two ImSURE phantoms.

The binary masks of the 90 ROIs were converted to sets of contour points to create the RTSTRUCT file necessary for the DICOM format. For each slice of the binary masks, the external contour of the ROI was traced using a self-developed MATLAB code that drew the contour line between the centre of the last voxel included in the binary mask and that of the first voxel outside the mask.

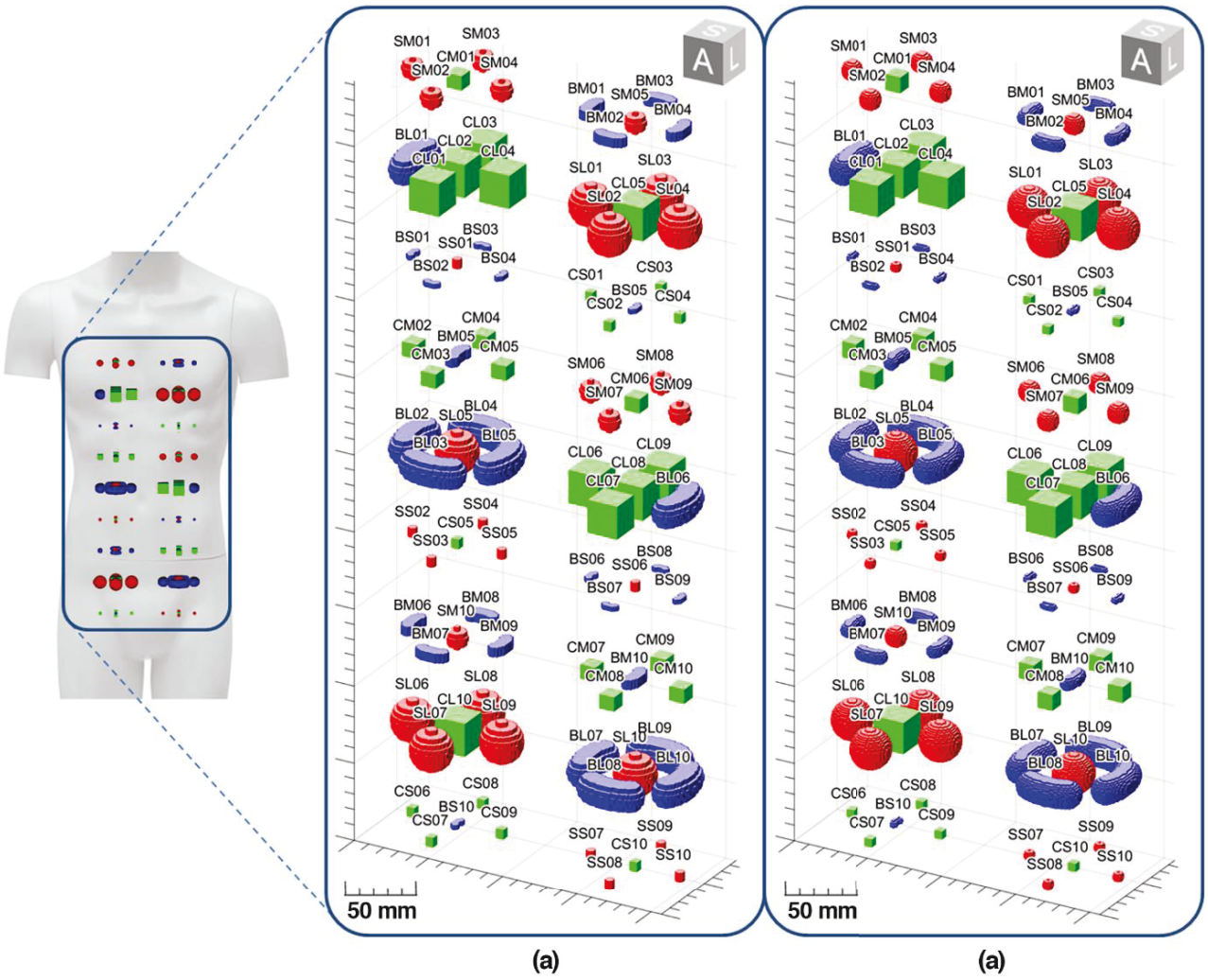

Figure 3.6. Spatial arrangement of the 90 ROIs with respect to the patient's body. (a) ROI binary masks for the anisotropic phantom. (b) ROI binary masks for the isotropic phantom.

Figure 3.7 shows the contour line for the central slices of three representative binary masks corresponding to the small ROI shapes.

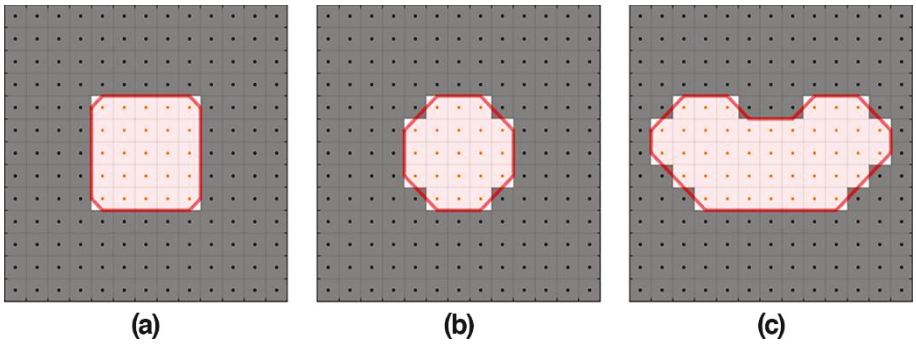

Figure 3.7. Definition of the contour set from the binary masks. Red lines indicate the generated contours for the central slice of the small cube (a), sphere (b) and bean (c) shapes.

The final characteristics of each base morphology are summarised in Table 3.2, while three representative slices of the isotropic phantom are visible in Figure 3.8.

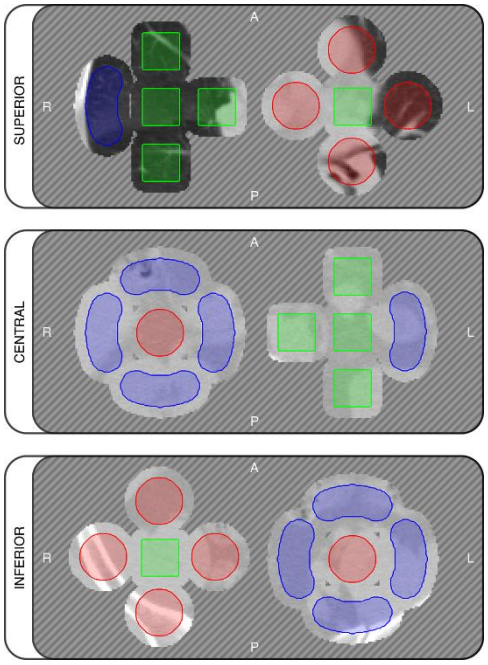

Figure 3.8. Three slices of the isotropic phantom containing the central portions of large ROIs. The segmentation is superimposed on the image textural content. Diagonal lines represent the censored CT data. A: anterior, P: posterior, R: right, L: left.

3.3.1.3 DICOM export and format conversion

The censored CT images and their respective contour sets were saved in DICOM format using MATLAB 2020a. Eventually, 3D Slicer (version 4.10, http://www.slicer.org) software was used to convert the DICOM files to NIfTI and NRRD formats, so as to guarantee ease of use of our phantoms with different software tools. Depending on the file format, ROI unique identifiers (indicating shape, volume and instance number) were stored within the file header, in the mask file name or in a separate file.

3.3.2 Design of the feature-extraction workflow

Our phantoms were ad-hoc designed to test radiomic software in a heterogeneous scenario: the 90 ROIs with different morphological characteristics have been specifically thought for the identification of differences in feature values among radiomic programs due to ROI morphology (i.e., ROI shape and volume) that could not be investigated with previous datasets.

Table 3.2. Characteristics of the 9 different ROI configurations defined for both the isotropic and the anisotropic phantom. The volumes reported in the table correspond to the number of voxels in each ROI multiplied by voxel dimension.

Feature family	**Isotropic phantom**	**Anisotropic phantom**
Total Number of ROI	90	90
Voxel spacing [mm]	1.00x1.00	0.977x0.977
Slice Thickness [mm]	1.00	3.00
Voxel volume [mm³]	1	2.86
ROI Volumes [mm³]		
Bean - Small	120	125.89
Bean - Medium	998	881.20
Bean - Large	7964	7896.43
Sphere - Small	123	120.16
Sphere - Medium	1020	975.61
Sphere - Large	8025	7990.85
Cube - Small	125	143.05
Cube - Medium	1000	858.31
Cube - Large	8000	8010.87

To complement the ImSURE anisotropic and isotropic phantoms, I designed two systematic workflows of feature extraction (with and without interpolation, respectively), which allow to investigate the impact of pre-processing configurations on the reproducibility of features across software programs.

Given a phantom, which may or may not require interpolation, the feature extraction procedure was designed to include all possible combinations of pre-processing steps, namely two intensity discretisation approaches (i.e., FBN or FBS) combined with all feature aggregation methods (i.e., 2D:avg, 2D:mrg, 2.5D:avg, 2.5D:mrg, 3D:avg, 3D:mrg, 2D, 2.5D, 3D). For every method, specific parameters were chosen among the most employed in the radiomic literature (e.g., a bin width of 25 HU). Parameter explanation and details can be found in the IBSI reference manual[41].

From the 169 IBSI-standardised radiomic features, considering all the possible combinations of the extraction parameters, a total of 919 feature values were obtained for each ROI.

Table 3.3 synthesises the proposed extraction settings for the two phantoms while Table 3.4 the considered feature families and corresponding processing requirement. Figure 3.9 presents the scheme of the extracted features in better detail.

Table 3.3. Pre-processing settings used for the isotropic and anisotropic phantoms. (FBN = fixed bin number; FBS = fixed bin size; HU = Hounsfield Unit; IH = Intensity histogram feature family; IVH = Intensity-volume histogram feature family).

Pre-processing step	***isotropic phantom***	***anisotropic phantom***
Trilinear Interpolation		
resampled voxel spacing [mm]	none	1.00x1.00x1.00
Re-segmentation		
range [HU]	[-1000 400]	[-1000 400]
Discretisation		
texture and IH	FBS: 25 HU; FBN: 32 bins	FBS: 25 HU; FBN: 32 bins
IVH	FBS: 2.5 HU; FBN: 1000 bins	FBS: 2.5 HU; FBN: 1000 bins

Table 3.4. Standardized feature families and required settings (MORPH = morphological features; LI = local intensity; IS = Intensity-based statistics; IH = Intensity histogram; IVH = Intensity-volume histogram; GLCM = Grey-level co-occurrence matrix; GLRLM = Grey-level run-length matrix; GLSZM = Grey-level size-zone matrix; GLDZM = Grey-level distance-zone matrix; NGTDM = Neighborhood grey tone difference matrix; NGLDM = Neighboring grey level dependence matrix; FBN = fixed bin number; FBS = fixed bin size; 2D:avg = averaged over slices and directions; 2D:mrg = merged directions per slice and averaged; 2.5D:dmrg = merged per direction and averaged; 2.5D:vmrg = merged over all slices; 3D:avg = averaged over 3D directions; 3D:mrg = merged 3D directions; 2D = averaged over slices; 2.5D = merged over all slices;3D = calculated from single 3D matrix).

Feature family	**Feature Count**	**Discretization**	**Aggregation**
MORPH	25	none	none
LI	2		
IS	18		
IH	23	FBN or FBS	
IVH	6		

GLCM	25		rotation dependent (2D:avg, 2D:mrg, 2.5D:dmrg, 2.5D:vmrg, 3D:avg, 3D:mrg)
GLRLM	16		
GLSZM	16		rotation independent (2D, 2.5D, 3D)
GLDZM	16		
NGTDM	5		
NGLDM	17		

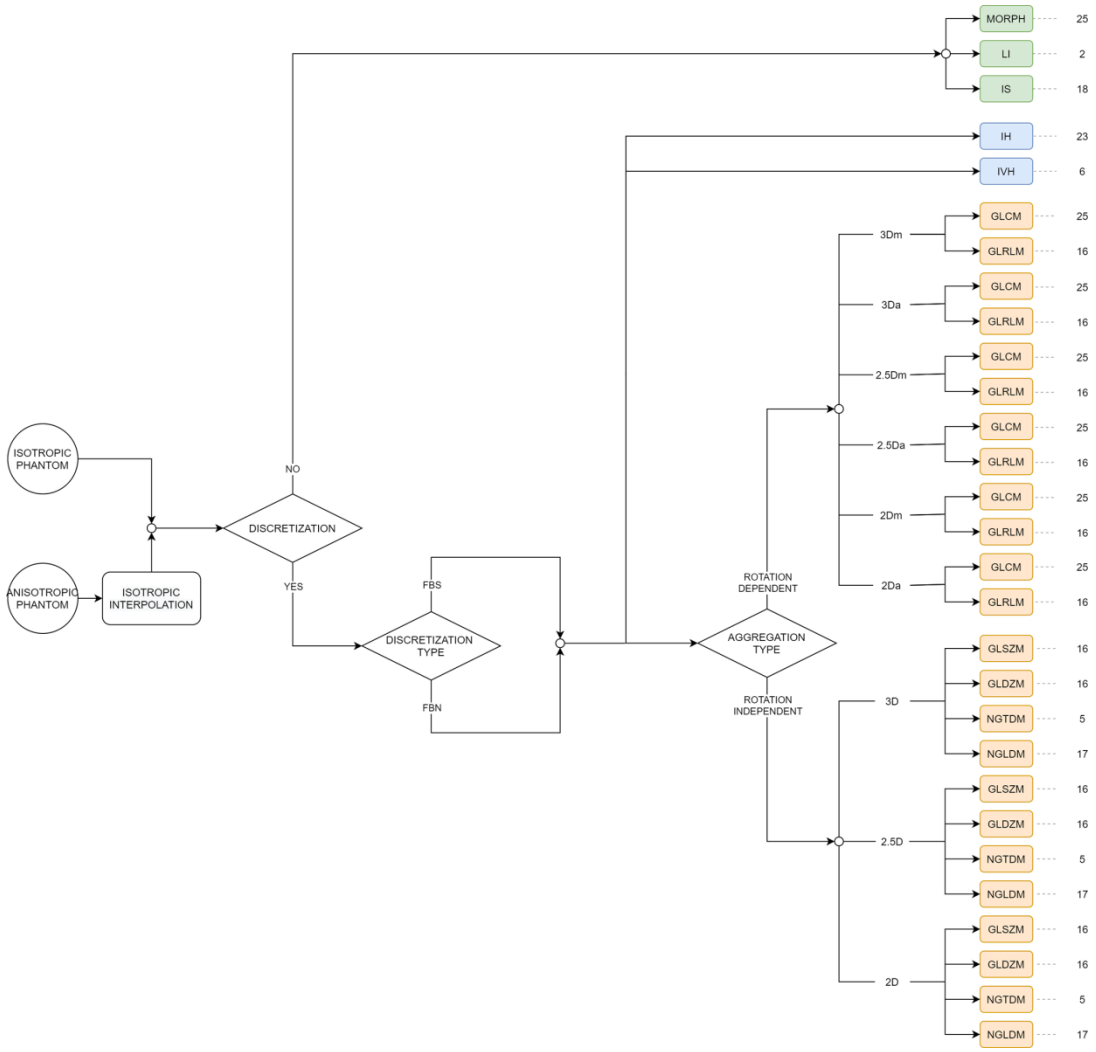

Figure 3.9. Scheme depicting the proposed feature extraction comprehending 919 features values. (FBN = fixed bin number; FBS = fixed bin size; MORPH = morphology; LI = local intensity; IS = intensity-based statistics; IH = Intensity histogram; IVH = Intensity-volume histogram; GLCM = grey level co-occurrence matrix; GLRLM = grey level run length matrix; GLSZM = grey level size zone matrix; GLDZM = grey level distance zone matrix; NGTDM = neighbourhood grey tone difference matrix; NGLDM = neighbouring grey level dependence matrix NGLDM).

3.3.3 Performance metrics and statistical analysis

Feature values were rounded up to the third significant digit prior to any comparison. I used the percentage of matching features between pairs of programs and their level of agreement to assess and compare software performances. For a pair of software $s_i, s_j,$ the percentage of matching features, P, was calculated as:

$$P(s_i, s_j) = P(s_j, s_i) = \frac{\#\ matching\ features\ between\ s_i, s_j}{\#\ of\ comparable\ features\ between\ s_i, s_j}$$

where i and j identify the software tool (with $i \neq j$). The number of comparable features between software i and j was defined as the number of features implemented in both tools.

For each feature f, agreement across all software tools, A, was defined as:

$$A = \frac{1}{\#S} \sum_{S} [f_{s_i} = f_{s_j}]$$

where S is the set of unordered program pairs ($s_i, s_j,$ with $i \neq j$) that can calculate the feature f, and $\#S$ is the dimension of S. Squared brackets represent the Iverson brackets, that is:

$$[f_{s_i} = f_{s_j}] = \begin{cases} 1 & if\ f_{s_i} = f_{s_j} \\ 0 & otherwise \end{cases}$$

The non-parametric Kruskal-Wallis test[78] was used to investigate whether A was significantly influenced by the factors being considered (i.e., discretization, aggregation methods, ROI shape, and ROI volume), under the null hypothesis that all groups came from populations with the same median. The significance level, $\alpha = 0.05$, was corrected with Bonferroni's method (adjusted α of 9e-5). The statistical analysis was performed in MATLAB (version 2018b, The MathWorks, Natick, 2018).

3.3.4 Results of the assessment on the ImSURE phantoms

For each program, 'no matching' features that were found using the *IBSI digital phantom* were excluded for the analysis on the ImSURE phantoms (see Figure 3.2 were "no matching features" are reported grouped by feature family and aggregation method). No features were excluded for MIRP, S-IBEX and RadiomiCRO. One and two 'no matching' features were found for Pyradiomics and SOPHiA,

respectively. A total of 21 and 22 features were eliminated for SERA and RaCaT, respectively, mostly belonging to the NGLDM 2.5D and LI families for the former and MORPH and LI for the latter.

Figure 3.10a shows the percentage of comparable features for each pair of programs out of a total of 919 possible values, while Figure 3.10b and Figure 3.10c compare the percentages of matches of the '*isotropic phantom*' (no program-specific interpolation required) with those of the '*anisotropic phantom*' (interpolated within each program before feature calculation). It should be noted that the reported match percentages were computed with respect to the total number of comparable features shared by each program pair. By comparing Figure 3.10b and Figure 3.10c, we observe that program-based interpolation had an impact on the overall percentage of matching features. When interpolation was applied, the match percentages of PyRadiomics fell below 2.5%, suggesting that the interpolation method used in this program may not be compliant with the IBSI guidelines, while those of SERA presented a marked decrease. SERA behaviour could be ascribed to an erroneous interaction between interpolation and 2D/2.5D aggregation methods for FBS discretization, rather than to a non-compliant interpolation. Instead, program-specific interpolation had no effect on the MIRP, S-IBEX, SOPHiA, RaCaT, and RadiomiCRO values.

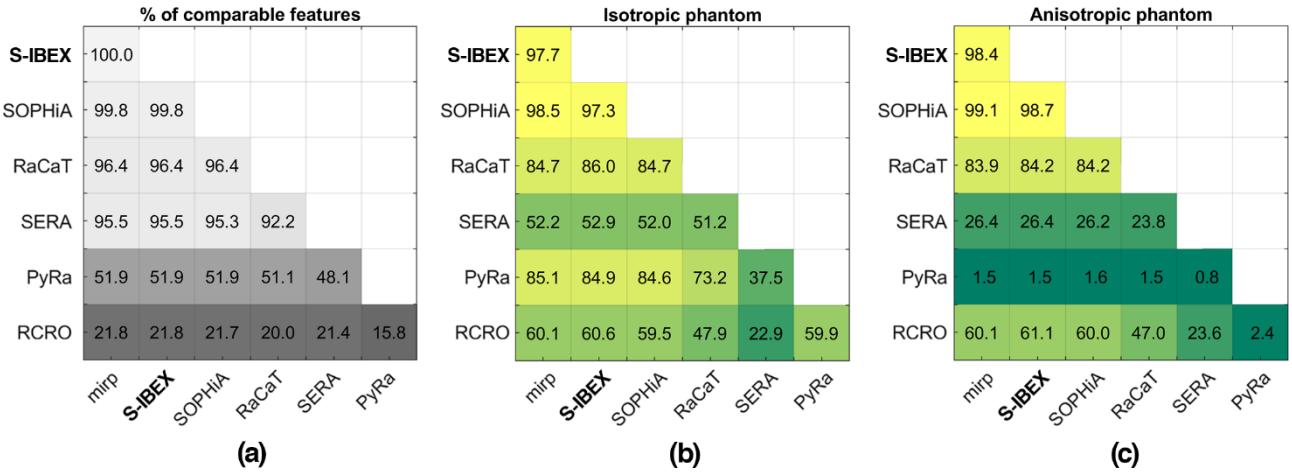

Figure 3.10. Analysis of interpolation effect. (a) Percentage of comparable features between program pairs out of the total of 919 features. (b) Percentage of matches between program pairs for the isotropic phantom (no interpolation required). (c) Percentage of matches for the anisotropic phantom (requiring program-based interpolation). SOPHiA = SOPHiA DDM for radiomics; PyRa = Pyradiomics; RCRO = RadiomiCRO.

Following these results, in the subsequent analyses, I only focused on the *isotropic phantom* to rule out the discrepancies observed between programs caused by software-based interpolation. Consequently, I

investigated the effect of the discretization approach on the percentage of matches between pairs of programs. Figure 3.11 shows the results for the cases of no discretization (left), FBN (middle) and FBS (right). Figure 3.11a,d only includes the MORPH, LI, and IS feature families, which do not require intensity discretization (Table 3.4). In this case, all programs achieved match percentages higher than 80%. Figure 3.11b,e and c,f aggregate the remaining families calculated using the FBN and FBS approaches, respectively, and highlight that SERA FBN discretization and RadiomiCRO FBS discretization are not concordant with the other programs. This suggests that their implementation is not IBSI-compliant for these programs. Notably, the RadiomiCRO discrepancy confirmed the results obtained in Phase I while the SERA discrepancy was not visible on the *IBSI radiomic phantom*. Regardless of discretization, MIRP, S-IBEX, and SOPHiA achieved the highest match percentage.

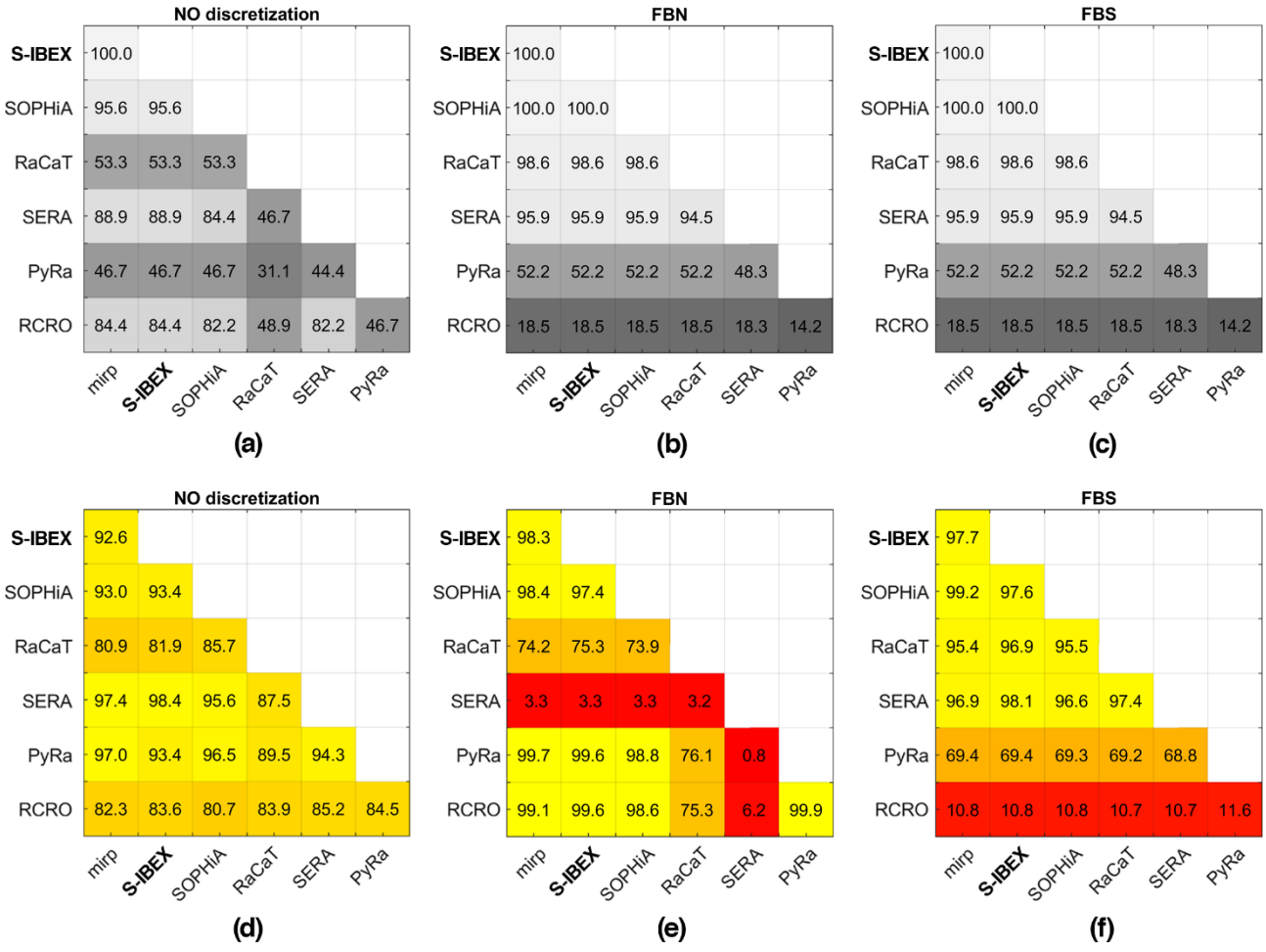

Figure 3.11. Analysis of the discretization effect. (a-c) Percentages of comparable features. (d-f) Percentages of matches between program pairs considering feature families without discretization, with FBN or with FBS discretization, respectively. FBN = fixed bin number; FBS = fixed bin size; SOPHiA = SOPHiA DDM for radiomics; PyRa = Pyradiomics; RCRO = RadiomiCRO.

PyRadiomics showed greater match percentages for FBN discretization than for FBS, while the RaCaT results were complementary, with higher percentages for the FBS method.

The effect of the aggregation method on the percentage of matching features across program pairs, stratified by FBN and FBS approach, was also evaluated. Figure 3.12 and Figure 3.13 illustrate the results in greater detail. PyRadiomics could not calculate the feature values associated with 2D aggregation, while RadiomiCRO was only designed to calculate 3D:mrg aggregation. The match percentages for MIRP were lower in 2D aggregation than in other aggregation methods. This result was observed for some ROI conformations that produced undefined results for the 2D aggregation method in the intermediate steps of feature calculation.

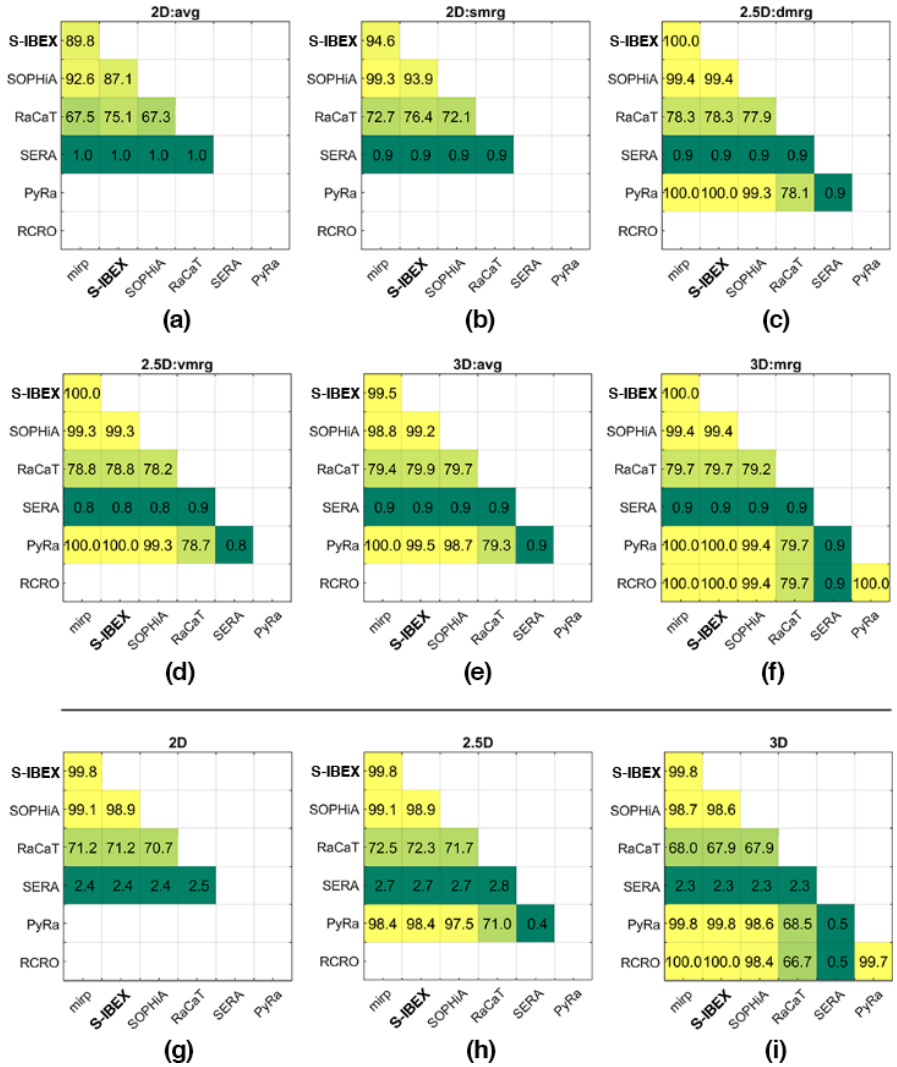

Figure 3.12. Effect of aggregation methods, defined for feature families that are (a-f) and that are not (g-i) based on directional matrices, for the FBN discretization approach. FBN = fixed bin number; SOPHiA = SOPHiA DDM for radiomics; PyRa = Pyradiomics; RCRO = RadiomiCRO.

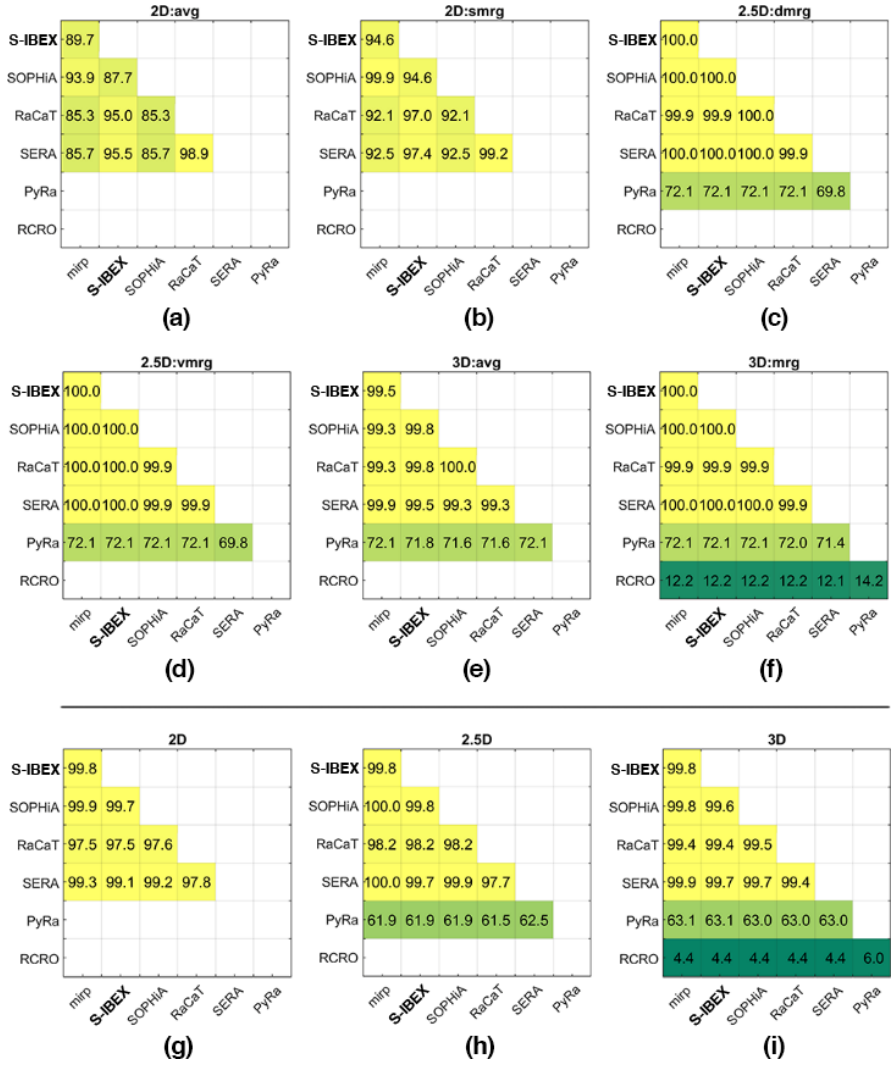

Figure 3.13. Effect of aggregation methods, defined for feature families that are (a-f) and that are not (g-i) based on directional matrices, for the FBS discretization approach. FBS = fixed bin size; SOPHiA = SOPHiA DDM for radiomics; PyRa = Pyradiomics; RCRO = RadiomiCRO.

Multiple ROIs with varied volumes and shapes were included in the two phantoms designed for this study, allowing us to also investigate the differences in program performance due to ROI characteristics. The data were stratified by ROI shape and ROI volume, and match percentages between software pairs were calculated in the two cases. The results are presented in Figure 3.14 and Figure 3.15, respectively. Unlike the other factors, this analysis showed no relevant differences between programs due to ROI shape or ROI volume, meaning that ROI morphology had no discernible impact on match percentages at the whole-feature level.

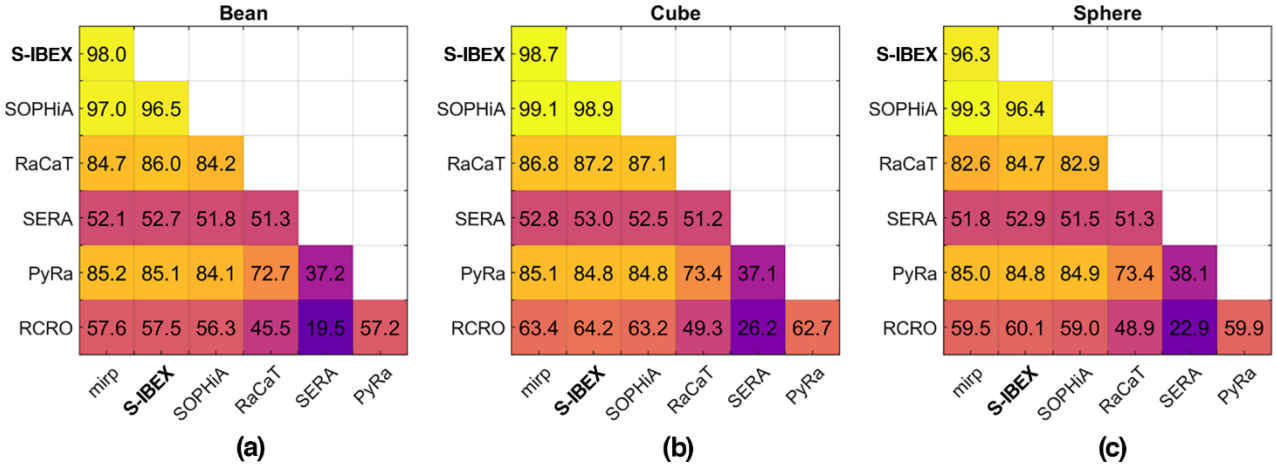

Figure 3.14. Percentages of matches, on comparable features, between program pairs stratified by ROI shape: (a) bean, (b) cube and (c) sphere. SOPHiA = SOPHiA DDM for radiomics; PyRa = Pyradiomics; RCRO = RadiomiCRO.

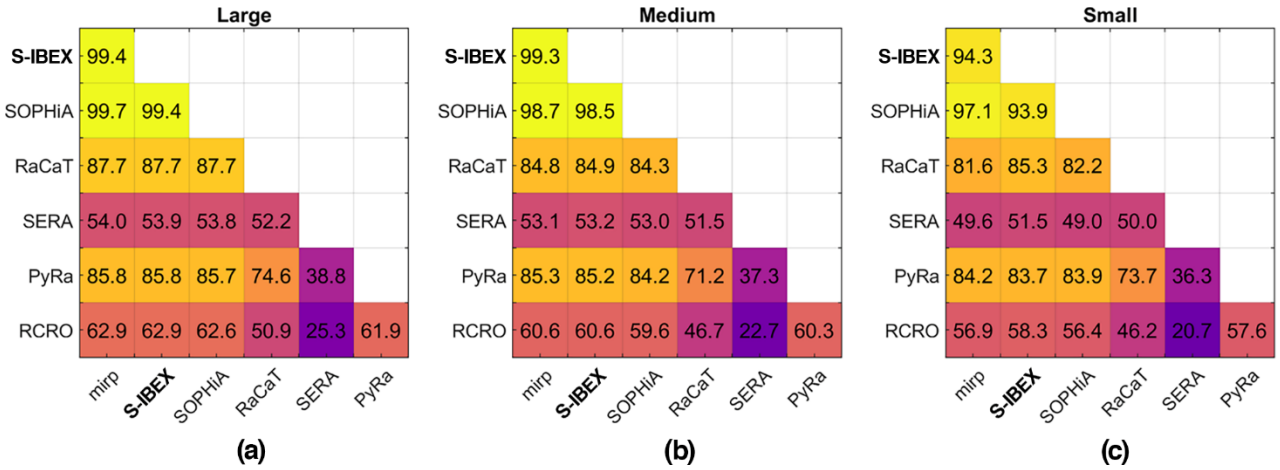

Figure 3.15. Percentages of matches, on comparable features, between program pairs stratified by ROI volume: (a) large, (b) medium and (c) small. SOPHiA = SOPHiA DDM for radiomics; PyRa = Pyradiomics; RCRO = RadiomiCRO.

Finally, the non-parametric Kruskal-Wallis test[78] was applied to agreement values for insights at the single-feature level, distinguishing four main factors: discretization, aggregation methods, ROI shape, and ROI volume. The test results are shown in Figure 3.16. for each factor and feature under examination. This analysis showed that discretization was significant for almost every feature family requiring intensity discretization. The aggregation factor was significant for most of the features belonging to the GLCM and GLRLM classes, as well as for some GLSZM and NGLDM features. The ROI shape was only significant for the features belonging to the MORPH and LI families, while the

ROI volume was significant for almost all the GLCM features, as well as for a portion of the MORPH, IS, IH, and NGTDM features.

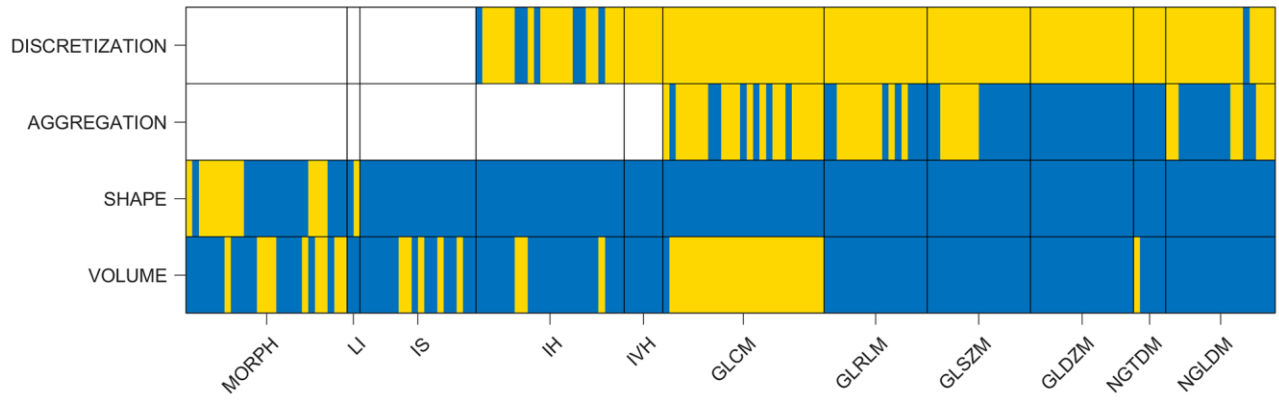

Figure 3.16. The results of the Kruskal-Wallis test applied to the agreement among programs. The test results are presented for each feature family and for four different factors, i.e., discretization, aggregation, ROI shape, and ROI volume. In the figure, the yellow color indicates significant differences after Bonferroni correction ($p \leq 9e\text{-}5$), the blue color denotes non-significant results, and the white cells correspond to non-existing combinations of feature families and factors.

The dependence of the agreement on ROI volume for the GLCM family was further investigated. This dependence was due to the '*bean small*' and '*sphere small*' ROIs, where 2D:avg and 2D:smrg aggregations resulted in discordant feature values among the software applications (Figure 3.16).

It is worth recalling that 2D GLCM features are calculated by aggregating information from four different directional matrices calculated over each slice, that is, along the (x, y) directions (1, 0), (1, 1), (-1, 1), and (0, 1). For 2D:avg aggregation, GLCM features are computed from each 2D directional matrix and averaged over directions and slices; while for 2D:smrg, features are computed from a single matrix after merging the four 2D directional matrices per slice and then averaged over slices.

To explain the discrepancies obtained in the computation of 2D aggregation, we need to look at the top and bottom slices of the ROIs reported in Figure 3.17 (i.e., slices B1, B5 for *'bean small'*, and slices S1, S7 for *'sphere small'*). It is not possible to calculate all four directional matrices on these slices. In the case of *'bean small'* for 3 directions over 4, GLCM matrices cannot be defined as there are no adjacent voxels in the mask along those directions (Figure 3.18a), while for '*sphere small'*, no directional matrices can be calculated as there are no adjacent voxels in the mask for all four directions (Figure 3.18b).

The observed discrepancies across software for 2D aggregation methods are due to the different handling of these undefined matrices in the calculation of GLCM features.

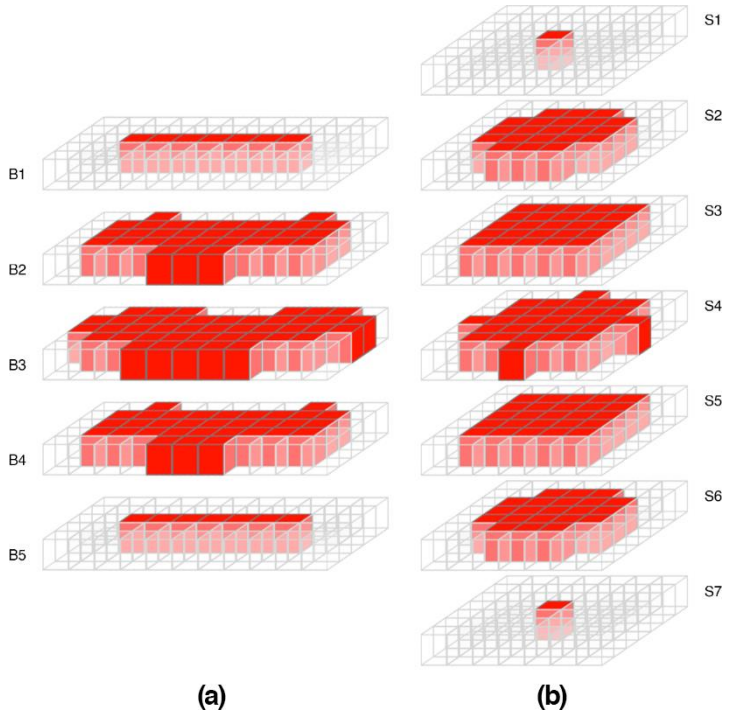

Figure 3.17. Slice per slice visualization of the masks for (a) 'bean small' and (b) 'sphere small' ROIs.

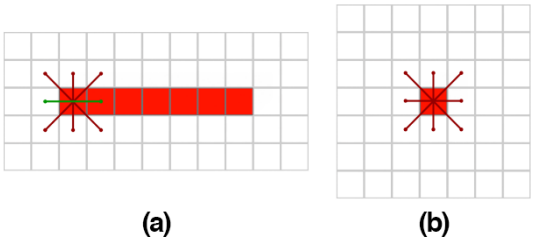

Figure 3.18. (a) 'Bean small' top/bottom slices (B1, B5). Only one direction results in a definite directional matrix for the top/bottom slice of the 'bean small' mask (in green). For the other directions, no adjacent voxels are available to calculate the GLCM matrices. (b) 'Sphere small' top/bottom slices (S1, S7). No voxels are available in all four directions to calculate the GLCM matrices for the top/bottom slice of the 'sphere small' mask.

3.3.5 Discussion

The assessment on the ImSURE phantoms allowed to systematically investigate the effect of factors related to parameter setting (i.e., interpolation, discretization, and aggregation) as well as to ROI characteristics (i.e. volume and shape) on software agreement by employing two custom digital phantoms and a systematic feature extraction. For the calculation of the percentage of matching

features, P, and of software agreement, A, all pairwise comparisons among tools were considered instead of comparing them to a reference one, as it was not possible to justify choosing one tool over the others: even IBSI-1 compliance was not a reasonable criterion, as we explored aspects that were not analysable on the IBSI phantoms.

The interpolation effect was analysed by comparing the match percentages between *isotropic* and *anisotropic* phantoms. The results revealed that for SERA and Pyradiomics, performances were influenced by program-specific interpolation. Notably, interpolation is one of the initial steps in the image processing scheme and has an impact on downstream processes as well as on final feature values. Thus, it should be a priority of standardization for all programs.

Subsequently, the effect of intensity discretization was evaluated focusing on the *isotropic* phantom. Only among S-IBEX, MIRP, and SOPHiA percentages of matching features were not impacted by the discretization method. This pre-processing step is typically applied to the ROI before the calculation of IH, IVH, and textural features. Therefore, a correct implementation is also crucial for the reproducibility of these feature values across tools.

The analysis of the aggregation method allowed the identification of an aspect that still needs to be addressed by the IBSI, which caused the programs to calculate different radiomic feature values as they are currently not aligned in the implementation strategy.

In contrast to previous studies[73–75], the entire set of IBSI-standardized radiomic features was analysed rather than only those that were common to all tools. Secondly, I disregarded *program-default* settings and only considered *harmonized* extraction (i.e., user defined parameter settings) because, in practice, users tweak the software to match a desired parameter configuration.

In literature, digital phantoms range from being purely synthetic (e.g., the *IBSI digital phantom* with artificial texture and arbitrarily-defined ROI) to image-based (e.g., the *IBSI radiomic phantom* with CT-derived pattern and GTV ROI). The ImSURE phantoms were designed with intermediate characteristics (textures derived from a CT image and geometrical ROIs) to allow the assessment, in a single investigation framework, of the impact on the software agreement of factors related to both image pre-processing and ROI morphology. Moreover, by placing multiple ROIs over a patient's image,

different texture patterns were sampled, hence augmenting the casuistry and heterogeneity (different anatomical regions were tested in the same run) of the ROIs that were used in the analysis.

Regarding the limitations of this study, the ImSURE phantoms used for the analysis were made of simplified morphologies arbitrarily positioned on a single image modality (i.e., CT). Nevertheless, the choice of the modality does not affect the overall outcome of the work, which was designed to assess and compare basic aspects of image processing among radiomic tools. In future, phantoms constructed with other modalities will allow for further investigations on modality-specific aspects[79,80]. Concerning the morphologies, the chosen ROI shapes are less complex than clinical ROI, however, this simplification was necessary to systematically study the impact of ROI characteristics. Eventually, ROIs may intersect anatomical structures differently with respect to a ROI defined for clinical studies. However, multiple textural patterns were derived from different anatomical districts, which ensured covering of the feature range obtainable from clinical ROIs imaged with CT. In these terms, I am reasonably confident that the software concordance tested on our phantom could be translated into software concordance calculated for clinical targets in several districts.

It is important to note that the differences observed in extracted feature values might limit radiomic model reproducibility[8,17,63]. Therefore, when building a model, it is recommended that the stability of selected features is checked by comparing the values obtained with at least two different tools. However, future studies are needed to assess the impact of software differences on clinical endpoint prediction[74,81].

As a remark, it is important to note that the initiative itself does not provide a criterion (e.g., a threshold on the percentage of matching features values) to unambiguously identify "IBSI-compliant" software tools. Therefore, in the recent literature, the term "IBSI-compliant" has been used to identify programs that de facto have different degrees of standardization. The clarification of this aspect is a rationale for the set-up of additional studies, such this one, that allow to better clarify the equivalency of standardized radiomic tools.

In conclusion, I designed a new investigation scenario in which I demonstrated that, despite the ongoing efforts of both IBSI and software developers to standardize radiomic tools, additional efforts are needed to achieve full concordance. Nevertheless, S-IBEX resulted in one of the most standardized software tools and achieved the highest agreement with other two tools: MIRP and SOPHiA. These three programs appeared to be practically equivalent from a radiomic extraction point of view and the

interested user can choose between them based on factors such as their computational speed, their cost, their ease of set-up and usage (absence/presence of a GUI) and on his preferences (e.g., programming language).

Eventually, the ImSURE phantoms have also been presented, representing a multi-purpose dataset useful, for example, to compare the agreement of a new radiomic software with those that have been tested in this work or to compare convolutional filtering implementations across radiomic software. Eventually, the methodology proposed for the phantom building, could also be reused to accurately position ROIs of any desired shape and volume inside a medical image to create new phantoms with different characteristics that could be useful for radiomics as well as for other fields. For these reasons both the ImSURE phantoms and the code used to create them were made publicly available in public repositories[82,83].

Chapter 4: S-IBEX for Clinical Studies

In this chapter, I will present the clinical research projects that relied on the S-IBEX tool for radiomic feature computation. These studies bear witness to the intense and prosperous collaboration with the Veneto Institute of Oncology – IOV IRCCS, which provided most of the clinical and imaging data.

The studies covered a wide range of cancer types (i.e., breast, liver, lung, head & neck, prostate cancers) imaged with different modalities (e.g., CT, PET, Mammography), and some works crossed with dosiomics[84] (an extension of radiomics that analysis patients' three-dimensional radiotherapy dose distribution rather than conventional medical imaging). The variety of data that were analysed, each requiring a specific feature-extraction configuration, proved the high versatility of S-IBEX.

In particular, four investigations will only be presented in a concise form, while three will be discussed in higher details (i.e., breast, prostate, and liver cancer), namely the ones to which I contributed the most in terms of study design, data collection, data curation, and data analysis. The RQS (section 1.2) of these three studies ranged from 31% to 36% and was mostly affected by the size of the considered dataset, which did not allow for an external validation of the proposed models. Nevertheless, internal validation was performed by means of repeated cross-validation and hold-out validation, which, according to the TRIPOD statement, are forms of Type 1b validation (subsection 1.1.6). Repetitions were employed to ensure robustness with respect to a random splitting procedure (required for validation Type 2a), which may induce bias in performance estimates especially when dealing with a limited sample size. Other limiting factors were: 1) the retrospective collection of data, that in one case resulted in inhomogeneous image-acquisition protocols, 2) a monocentric perspective that did not allow for an external validation of the results and 3) the limited time that physicians and radiologists were able to dedicate to the segmentation tasks (multi-reader segmentations were obtained for only one study).

Although preliminary, the results of these works showed that radiomic analysis could provide a valuable and cost-effective contribution in many different oncological diseases and could guide the design of future perspective multicentric studies.

The following sections present works that are under preparation (sections 4.1 and 4.3) and under submission (section 4.2) to international journals.

4.1 [^{18}F]FDG PET/CT radiomic features for the prediction of clinical outcomes in high-risk and locally advanced breast cancer

4.1.1 Introduction

Locally advanced breast cancer (LABC) and high-risk breast cancer (BC) are a clinical challenge as the majority of patients with these diagnoses develop distant metastases despite appropriate therapy[85]. Patients with locally advanced disease encompass a wide range of clinical scenarios, including advanced primary tumours, advanced nodal disease, and inflammatory carcinomas. Neoadjuvant chemotherapy (NAC) is the first-choice therapy for this type of patients as well as for patients with high-risk BC[86]. LABC and high-risk breast cancer are usually staged with conventional imaging, such as contrast enhanced computed tomography (CT), magnetic resonance imaging (MRI) and bone scintigraphy. However, these imaging techniques are not able to accurately predict the risk of an early recurrence and of a poor long-term prognosis[87]. [^{18}F]fluorodeoxyglucose ([^{18}F]FDG) positron emission tomography - computed tomography (PET/CT) has been proved to be useful for staging LABC[88] and to provide additional prognostic information in the setting of LABC both qualitatively[89] and semi-quantitatively[90]. Moreover, the advantage of [^{18}F]FDG PET/CT over conventional imaging is that it allows the examination of extra-axillary nodes as well as the chest, abdomen, and bone in a single session. Radiomic metabolic features extracted from [^{18}F]FDG PET/CT may prove to be stronger predictive factors for determining NAC outcomes, capturing the intratumoral heterogeneity which is considered to be strongly associated with treatment response[91,92]. However, the role of these features in predicting overall survival (OS) and disease-free survival (DFS) in this setting of patients is still undefined.

The aim of this study is twofold: first, to investigate the relationship between tumour clusters, as defined by radiomic metabolic features, and histopathologic characteristics, tumour response to NAC and lymph nodal status; second, to assess the additional contribution of radiomic features for the prediction of 5-year overall survival (5Y-OS) and 5-year disease-free survival (5Y-DFS).

4.1.2 Materials and methods

4.1.2.1 Patient and clinical data

Between June 2011 and October 2014, 111 patients with BC who underwent [^{18}F]FDG PET/CT were selected from a single institutional database. Inclusion criteria were: 1) patients with a clinical,

histopathological, and radiological diagnosis of high-risk and LABC; 2) patients candidates for NAC; 3) PET/CT scan before any type of treatment for the primary disease. Patients who received a lumpectomy before PET/CT were excluded.

All patients had given their informed consent before undergoing [^{18}F]FDG PET/CT scan and for their data to be analysed anonymously. The study was performed in accordance with the Declaration of Helsinki.

Clinical characteristics of patients were recovered from clinical charts. In particular, information about type of histology (i.e., ductal or lobular cancer), immunohistochemical analysis (i.e., oestrogen receptor - ER, progesterone receptor - PR, human epidermal growth factor receptor 2 - HER2, proliferation index - Ki67) and molecular subtypes (i.e., luminal A, luminal B, luminal B - HER2 enriched, HER2 positive, triple negative) was collected. The expression of ER and PR was considered positive (ER-pos, PR-pos) if major than 10%; HER2 was considered positive (HER2-pos) when ++ or +++ with Fluorescence in situ hybridization amplification; Ki67 was labelled as high (high-Ki67) when greater than 14%. Molecular subtypes were defined in accordance with the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer). Eventually, all clinical outcomes (i.e., response to NAC, DFS, and OS) were obtained from the archives.

4.1.2.2 Acquisition, reconstruction, and post-processing

Patient images were acquired on the Siemens Biograph 16 PET/CT scanner (Siemens Medical Solutions, Illinois, USA). Supine PET/CT torso examination from skull base to mid-thigh (6-7 bed, 3 min per bed) was performed 60 min after intravenous injection of 3 MBq/kg of [^{18}F]FDG (at least 6 hours of fasting). Data was reconstructed using the iterative ordered subset expectation maximization (OSEM) algorithm with 3 iterations, 21 or 24 subsets and point-spread function (PSF) modelling, with an image matrix size of 168x168 and a voxel size of 4.06x4.06x2 mm. Images were normalized and corrected for scatter and random events, dead-time, attenuation, and decay; a post-reconstruction Gaussian smoothing of 4mm FWHM was applied.

4.1.2.3 Delineation

Patient images were imported in RayStation 6.1.2 (RaySearch Laboratories, Sweden) for segmentation. Primary tumour lesions were drawn semi-automatically on PET images using the seeded region growing

tool based on component trees included in the program[93] and segmentations were revised upon consensus by two experienced nuclear medicine physicians.

4.1.2.4 Feature extraction

S-IBEX[60], an in-house developed radiomic feature extractor compliant with the Image Biomarker Standardization Initiative[13], was employed for the calculation of radiomic features. Following the statistics-based methodology to textural analysis[94], for each region of interest, 168 features were extracted belonging to 11 categories: morphological features (MF), local intensity features (LIF), intensity-based statistical features (IS), intensity histogram features (IH), intensity-volume histogram features (IVH), grey level co-occurrence based features (GLCM), grey level run length based features (GLRLM), grey level size zone based features (GLSZM), grey level distance zone based features (GLDZM), neighbourhood grey tone difference based features (NGTDM), and neighbouring grey level dependence based features (NGLDM). Prior to the extraction of textural features, images were interpolated to a voxel size of 2.03x2.03x2.00mm using slice by slice bilinear interpolation. The parameters for interpolation were set in order to avoid mixing the information of adjacent slices. The desired resolution was chosen as the one that would retain the largest amount of original voxels. This procedure led to a quasi-isotropic voxel size, more suitable than the original voxel size for the calculation of textural features derived with a 3-dimensional approach[13]. Subsequently, range re-segmentation was applied to exclude from textural feature computation voxels with a Standardized Uptake Value (SUV) greater than 25, since SUV values higher than 25 are considered outliers. Images were discretized using Fixed Bin Size (FBS) method with a bin width of 0.4 SUV. Fixed Bin Size was chosen as it showed better feature reproducibility for PET images[58,95] and retains a direct relationship with the original intensity scale[13]. The width was designed to get approximately 64 grey levels in the interval [0-25] SUV. Similar choices are reported in previous works[91,96].

Computed features were normalized by using z-score normalization. Reports on image processing, feature set and parameters used for feature extraction can be found in appendix Table C.1 and Table C.2.

4.1.2.5 Exploratory Analysis

Feature analysis was performed in MATLAB 2018b (MathWorks Inc., Natick, MA, USA) and R software version 3.5.3 (The R Foundation for Statistical Computing, Vienna, Austria).

Firstly, Wilcoxon rank-sum test was used to determine whether features differed significantly between images reconstructed with 21 and 24 subsets.

Consensus clustering (Pearson distance, unsupervised hierarchical clustering algorithm), applied on radiomic feature values, was used to group both features and patients that showed closed association. This technique produces robust clusters without prior knowledge of the number of groups, aggregating results from multiple iterations of hierarchical clustering with sub-sampling. The number of clusters for both features and tumours were determined as the elbow point in the Delta Area plot, which represents the relative change in area under consensus cumulative distribution function curve with respect to the number of clusters[97]. Correlation among features was evaluated by Spearman correlation and summarized as a correlogram. Red-green heat map was built using consensus-clustering ordering. Correlations between clinical endpoint were also evaluated by Spearman correlation. 'MF - volume' (metabolically active tumour volume - MTV, cm^3), 'IS - maximum intensity' (SUVmax), 'IS - mean intensity' (SUVmean), together with other previously reported features that classify the extent of metabolic intratumoral heterogeneity, such as 'IS - skewness', 'IS - coefficient of variation', 'GLCM - joint entropy', 'GLCM - inverse difference moment' (also known as homogeneity) and 'GLSZM - zone percentage', were chosen as representative features[91]. Kruskal-Wallis test was used to compare the values of representative features among unsupervised tumours clusters. Chi-square test was applied in comparing the proportion of ER-pos, PR-pos, HER2-pos, high-Ki67 and pathologic complete response (pCR) and lymph nodal status among clusters. Bonferroni correction was applied to adjust the significance level threshold for all p-values of multiple tests.

4.1.2.6 Modelling

Clinical data and radiomic features were investigated both separately and jointly for prognostic modelling. The outcomes of interest, OS and DFS, were reshaped in a classification framework suitable for logistic regression: OS and DFS were both dichotomized using a 5-year threshold, that is the

standard oncological follow-up time, following a procedure presented in other studies[34,98,99]. Patients censored before 5 years were excluded from subsequent analyses.

Feature collinearity was reduced by removing highly correlated features (Spearman correlation greater than 0.95) from further analysis. Molecular subtypes were not included in the pool of clinical features, in favour of immunohistochemical surrogates, from which they are derived.

Feature selection

Three filter-based methods were employed to obtain different scoring criteria: two of them are univariate (Wilcoxon, Fisher score) and one is multivariate (minimum redundancy maximum relevance - mRMR). We used filter-based methods as they are computationally efficient and less prone to overfitting in relation to wrapper and embedded methods. More specifically, the chosen methods proved to be among the top scoring approaches in terms of both stability and performance in Parmar et al.[34]. For each feature selection method, three models were built: the first model included clinical data only, the second radiomic features only, and the third both clinical and radiomic data.

Clinical and radiomic models

Both for the clinical and the radiomic model, prediction of 5Y-OS and 5Y-DFS were obtained through the use of logistic regression: starting from the null model, one feature at a time (up to a maximum of 7 features) was iteratively included by following the ordering obtained from feature selection methods. At each iteration step, a likelihood ratio test was employed to assess whether the goodness of fit of the two models, consisting of n and $n+1$ covariates respectively, was significantly different.

Model performances were evaluated in terms of the area under the receiver operating characteristic curve (AUC). Since resubstitution AUC (AUC_{RS}), for which the whole dataset is used for both model development and evaluation, could be over-optimistic, cross-validated AUC (AUC_{CV}) was also considered to assess model performances[96]. To compute AUC_{CV} estimates, the model was retrained over 100 repetitions of three-fold cross-validation. In particular, the dataset was randomly split one-hundred times into three balanced sets: two were used for model training and one for the performance estimation through cross-validation methodology.

Repetitions ensure robustness with respect to the random splitting procedure, which may induce bias in AUC estimates especially when dealing with a limited sample size.

Combined models

To derive combined models, features that made up the radiomic models were added to their respective clinical model. Once again, both AUC_{RS} and AUC_{CV} were computed.

4.1.3 Results

4.1.3.1 Clinical Findings

Out of 111 high-risk and LABC subjects, 79 were included in this study and their images were retrieved from the picture archive and communication system (Figure 4.1). Patient characteristics are reported in Table 4.1.

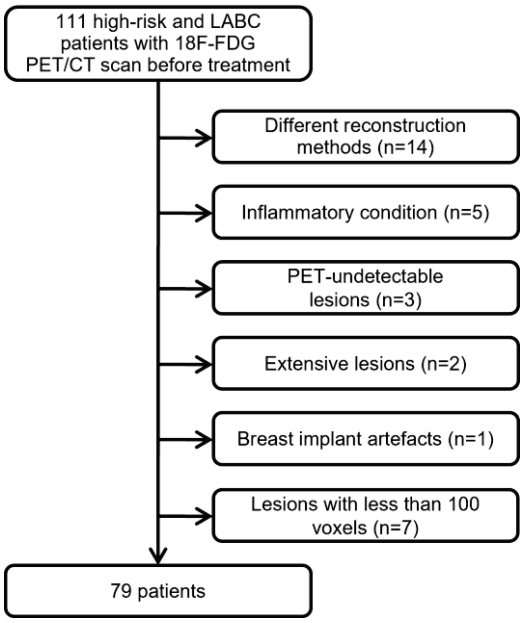

Figure 4.1. Exclusion criteria for the study population. PET-undetectable lesion: breast lesions with an FDG uptake comparable to breast background. Extensive lesion: tumours with a metabolically active tumour volume greater than 400 cm³.

Out of 79 patients, 68 (86%) were diagnosed with invasive ductal carcinoma. Clinical stage II and III occurred in 16 and 47 (20% and 60%) subjects respectively. Sixteen patients (20%) with a clinical stage I BC (20%) had unfavourable immunohistochemistry (i.e., a triple negative or a HER2-pos cancer) and therefore were candidates to NAC. Oestrogen-receptor positive (ER-pos) status was found in 45 (57%) tumours, progesterone-receptor positive (PR-pos) status in 33 (42%), and human epithelial receptor-2 positive (HER2-pos) in 25 (32%) subjects. Nineteen (25%) patients were responders to NAC.

Table 4.1. Patient characteristics.

Number of patients	79
Age, median (range) [years]	51 (27-71)
Clinical stage, n (%)	
I	16 (20%)
II	16 (20%)
III	47 (60%)
Histology, n (%)	
Invasive lobular cancer	8 (10%)
Invasive ductal cancer	68 (86%)
Other	3 (4%)
Grade, n (%)	
2	9 (11%)
3	67 (85%)
Unknown	3 (4%)
Immunohistochemical surrogate, n (%)	
ER-pos	45 (57%)
PR-pos	33 (42%)
HER2-pos	25 (32%)
high-Ki67	57 (72%)
Molecular subtype, n (%)	
Luminal A	8 (10%)
Luminal B	24 (30 %)
Luminal B (HER2 enriched)	12 (15%)
HER2 positive	13 (16%)
Triple negative	22 (28%)

ER-pos, oestrogen-receptor positive (>10%); PR-pos, progesterone-receptor positive (>10%); HER2-pos, human epithelial receptor-2 positive (+++ or ++ and Fluorescence in situ hybridization amplification); high-Ki67, high proliferation index (>14%)

After a median follow-up time of 67 (6-86) months for DFS and 68 (15-88) months for OS, 27 (34%) had a recurrence of disease, and 21 (27%) died.

4.1.3.2 Exploratory analysis

There were no features whose values were significantly associated with the OSEM reconstruction parameters (Wilcoxon rank-sum test, p-value < 0.0003).

Consensus clustering algorithm led to 5 tumour clusters (TCs) and 6 feature clusters (FCs). There were 1 (1%), 30 (38%), 29 (37%), 12 (15%) and 7 (9%) patients in TC I, II, III, IV and V respectively (Figure 4.2). TC I was omitted from subsequent analysis as it only contained one patient.

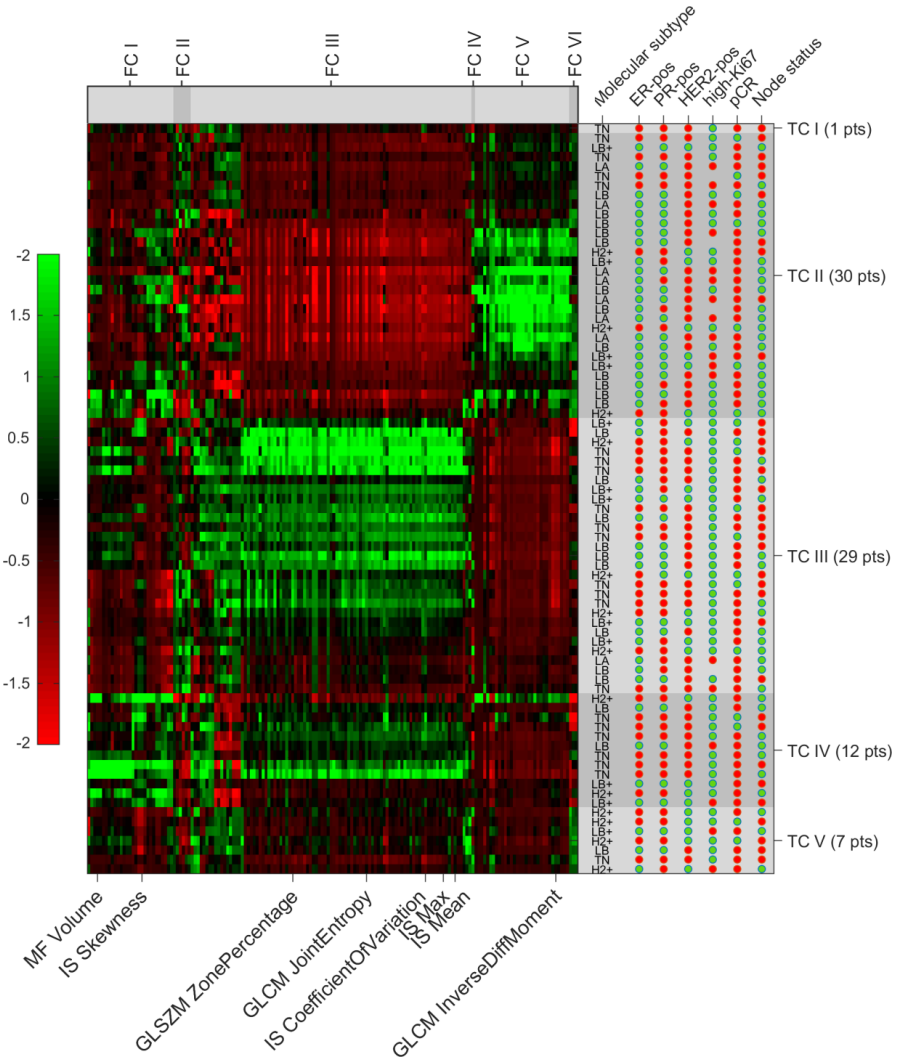

Figure 4.2. Textural features represented as a red-green heat map after z-score normalization. Five tumour clusters and six feature groups were identified with consensus clustering. Notes: row, patients; column, textural features/clinical information; green dot, positive/high-expression; red dot, negative/low-expression; Abbreviations: ER, oestrogen receptor; PR, progesterone receptor; HER2, human epidermal growth receptor 2; high-Ki67, high proliferation index; MF, morphological features; IS, intensity direct; GLSZM, grey level size zone matrix; GLCM, grey level co-occurrence matrix; TC, tumour cluster; pts, patients; FC, feature cluster.

The correlogram of Figure 4.3 shows high correlation among features within the same FCs, and high anti-correlation between FC III and FC V. Metabolically active tumour volume ('MF - volume') and SUVmax ('IS – maximum intensity') were significantly correlated, after Bonferroni correction, with 34 (21%) and 127 (77%) features respectively. Only 17 features (10%) showed no significant correlation with either tumour volume or SUVmax.

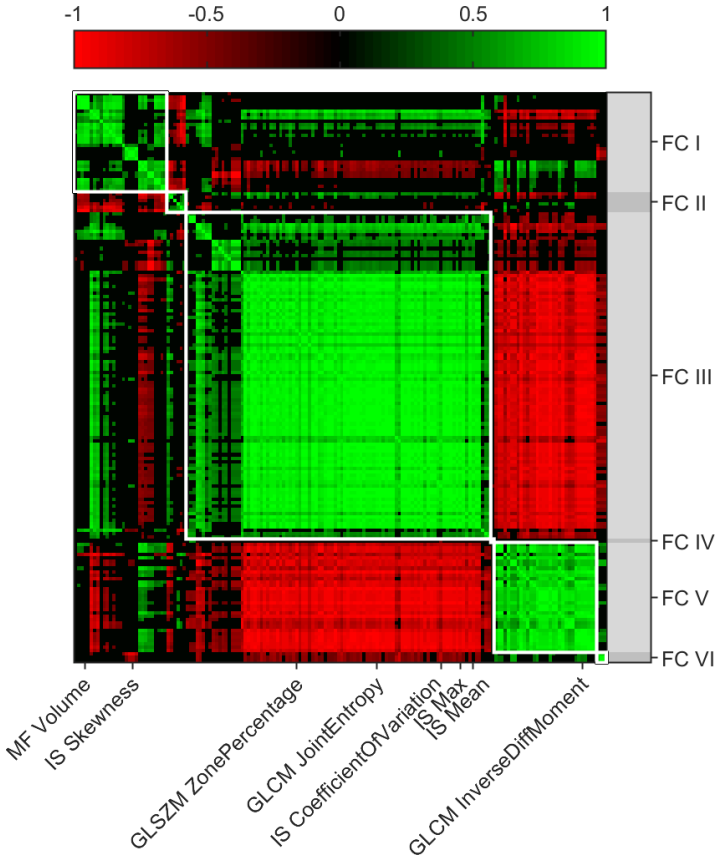

Figure 4.3. Correlogram of radiomic features with Bonferroni correction. Correlogram stresses the correlation structure of clustered radiomic features, where high association among groups are underlined. Spearman correlation coefficients are expressed using a red-green colour scale. Representative features are marked. Abbreviations: MF morphological features; IS, intensity direct; GLCM, grey level co-occurrence matrix; GLSZM, grey level size zone matrix; FC, feature cluster.

ER, Ki67 and high-Ki67 were the only clinical features that proved to be significantly correlated with radiomic features (p-value < 0.0003), with an absolute correlation in the 0.40-0.55 range for 61, 86 and 94 radiomic features respectively, mostly belonging to FC III and V.

The metabolic characteristics of the unsupervised TCs, reported in Table 4.2, were assessed by 8 representative radiomic features (reported in previous clinical studies or considered of interest by

clinicians): 'IS - maximum intensity', 'IS - mean intensity', 'MF - volume', 'IS - coefficient of variation', 'GLCM - joint entropy', 'GLCM - inverse difference moment', 'GLSZM - zone percentage', and 'IS - skewness'.

Table 4.2. Median (range) of radiomic features for each tumour cluster and p-values of the Kruskal-Wallis test.

	TC II	***TC III***	***TC IV***	***TC V***	***p-value***
Total	30	29	12	7	-
IS - maximum (SUVmax)	6.17 (2.60-9.61)	17.41 (8.10-39.09)	16.76 (12.19-29.88)	10.09 (8.38-12.45)	5.28E-12**
IS - mean (SUVmean)	3.06 (1.24-4.97)	6.09 (3.19-11.64)	4.97 (2.57-8.05)	4.98 (4.36-5.71)	3.71E-09**
MF - volume (MTV)	19.92 (5.77–73.91)	14.21 (3.52 – 49.04)	55.46 (10.33 – 187.82)	18.14 (8.95 – 32.20)	6.56E-03*
IS - coefficient of variation	0.30 (0.17-0.43)	0.59 (0.39-0.90)	0.59 (0.42-0.78)	0.36 (0.26-0.42)	5.94E-12**
GLCM - joint entropy	5.52 (3.25-7.13)	9.04 (7.13-11.06)	8.07 (5.76-9.87)	7.44 (6.38-7.97)	4.51E-12**
GLCM - inv. diff. moment	0.54 (0.34-0.77)	0.22 (0.11-0.33)	0.34 (0.22-0.57)	0.35 (0.29-0.45)	1.49E-12**
GLSZM - zone percentage	0.06 (0.01-0.23)	0.38 (0.18-0.68)	0.21 (0.04-0.38)	0.16 (0.08-0.24)	4.60E-12**
IS - skewness	0.70 (0.10-1.39)	0.90 (0.34-1.65)	1.37 (0.64-2.84)	0.35 (0.10-0.66)	8.32E-06**

**p-values less than 0.05; **statistically significant after Bonferroni correction ($p < 0.0003$); TC, tumour cluster; IS, intensity-based statistical features; MS, morphological features; MTV, metabolically active tumour volume; GLCM, grey level co-occurrence-based features; GLSZM, grey level size zone based features.*

Lesions clustered in TC II and TC III had a small to medium volume but opposite characteristics: SUVmax, SUVmean, 'IS - coefficient of variation', 'GLCM - joint entropy', and 'GLSZM - zone percentage' were high in TC III and low in TC II. TC IV includes lesions with higher MTV and SUVmax, while TC V small lesions with lower SUVmax. Moreover, TCs were correlated with histopathological data. Table 4.3 reports the differences found for ER-pos status, and high-Ki67 expression among tumour clusters, however those differences were not significant after Bonferroni correction.

4.1.3.3 Modelling results

Seven patients that were lost to follow-up prior 5 years were excluded from both DFS and OS analysis.

Table 4.3. Relationship between the Tumour Clusters and clinical/histopathological characteristics (chi-squared).

	TC II	***TC III***	***TC IV***	***TC V***	***P-Value***
Total	30	29	12	7	-
ER-pos	23 (29%)	15 (19%)	4 (5.1%)	3 (3.8%)	4.58E-02*
PR-pos	18 (23%)	9 (11%)	4 (5.1%)	2 (2.5%)	1.34E-01
HER2-pos	7 (8.9%)	9 (11%)	4 (5.1%)	5 (6.3%)	1.61E-01
High-Ki67	18 (23%)	27 (34%)	10 (13%)	5 (6.3%)	2.34E-02*
Response	7 (8.9%)	8 (10%)	2 (2.5%)	2 (2.5%)	9.16E-01
Node status	21 (27%)	13 (16%)	7 (8.9%)	3 (3.8%)	2.28E-01

**p-values less than 0.05; TC, tumour cluster; ER-pos: oestrogen-receptor positive; PR-pos: progesterone-receptor positive; HER2-pos: human epithelial receptor-2 positive; high-Ki67: high proliferation index.*

A total of 72 patients were considered for both OS and DFS prognostic modelling, of which 47 (59%) had no progression and 54 (68%) were alive at 5 years. Sixty-one features were retained after the removal of highly correlated features.

Clinical models

Based on the likelihood-ratio test and the AUC obtained through re-substitution technique, the optimal number of features to be included in the clinical model for OS and DFS predictions were two when considering Wilcoxon and Fisher Score selections methods (appendix Figure C.1a). In particular, the ordering obtained with those two methods were identical up to the fifth feature for OS and DFS. Therefore, when selecting the first 2 features, lymph nodes and ER-pos, the same model was built for OS with an AUC_{RS} of 0.77 and an AUC_{CV} of 0.75. Similarly, for DFS prediction Wilcoxon and Fisher methods converged to the same two-feature model with lymph nodes and high-Ki67 as independent predictors (AUC_{RS}=0.75, AUC_{CV}=0.73). Conversely, likelihood-ratio test and AUC_{RS} suggested the inclusion of one feature for mRMR, lymph nodes, for both OS (AUC_{RS} 0.69, AUC_{CV} 0.69) and DFS (AUC_{RS} 0.70, AUC_{CV} 0.70). Table 4.4 reports selected features and performances of clinical models.

Radiomic models

Using the same procedure, radiomic models for OS and DFS were built with one (Wilcoxon-Fisher-mRMR OS, Wilcoxon-Fisher DFS) and two (mRMR DFS) radiomic features (Table 4.5 and appendix Figure C.1b). A single model was built for OS (AUC_{RS}= 0.64, AUC_{CV}=0.64) by including one feature,

Table 4.4. Clinical models for OS and DFS obtained with Wilcoxon, Fisher Score and mRMR selection methods.

Outcome	***Feature selection method***	***Feature***	***AUC_{RS}***	***AUC_{CV} [95% CI]***
OS	Wilcoxon - Fisher	Lymph nodes ER-pos	0.77	0.75 [0.68-0.81]
OS	mRMR	Lymph nodes	0.69	0.69 [0.66-0.72]
DFS	Wilcoxon - Fisher	Lymph nodes High-Ki67	0.75	0.73 [0.69-0.77]
DFS	mRMR	Lymph nodes	0.70	0.70 [0.69-0.71]

OS, overall survival; DFS, disease-free survival; AUCRS, re-substitution AUC; AUCCV, cross-validated AUC; CI, confidence intervals; mRMR, minimum redundancy maximum relevance; ER-pos, oestrogen-receptor positive; high-Ki67, high proliferation index.

'IVH - area under the IVH curve', which had the highest rank for all feature selection methods. For DFS prediction, Wilcoxon and Fisher methods converged to the same univariate model with 'IH - quartile coefficient of dispersion' (AUC_{RS}=0.66, AUC_{CV}=0.66), while for mRMR, likelihood-ratio test and AUCRS suggested the inclusion of two features, 'IH - quartile coefficient of dispersion' and 'IVH - volume at intensity fraction 90' (AUC_{RS} 0.70, AUC_{CV} 0.67).

Table 4.5. Radiomic models for OS and DFS obtained with Wilcoxon, Fisher Score and mRMR selection methods.

Outcome	***Feature selection method***	***Feature***	***AUCRS***	***AUCCV [95% CI]***
OS	Wilcoxon - Fisher - mRMR	IVH - area under IVH curve	0.64	0.64 [0.60-0.68]
DFS	Wilcoxon - Fisher	IH - quartile coefficient of dispersion	0.66	0.66 [0.59-0.73]
DFS	mRMR	IH - quartile coefficient of dispersion IVH - volume at intensity fraction 90	0.70	0.67 [0.61-0.73]

OS, overall survival; DFS, disease-free survival; AUCRS, re-substitution AUC; AUCCV, cross-validated AUC; CI, confidence intervals; mRMR, minimum redundancy maximum relevance; IH, intensity-histogram; IVH, intensity-volume histogram.

Spearman correlation coefficient between selected radiomic features and MTV was -0.051, -0.583 and -0.036 for 'IVH - area under IVH curve', 'IVH - volume at intensity fraction 90' and 'IH - quartile coefficient of dispersion' respectively.

Combined models

Following the results of clinical and radiomic models, combined models derived from Wilcoxon and Fisher Score converged to a unique three-feature model both for OS and DFS: lymph nodes, ER-pos and 'Area under IVH curve' were combined for OS prediction (AUC_{RS}= 0.83, AUC_{CV}= 0.79); conversely, lymph nodes, high-Ki67 and 'IH - quartile coefficient of dispersion' were selected for DFS prediction (AUC_{RS}=0.79, AUC_{CV}=0.77). mRMR selection method led to poorer combined model performances for both OS and DFS (AUC_{RS}=0.75, AUC_{CV}=0.75 and AUC_{RS}=0.76, AUC_{CV}=0.72 respectively). Results are reported in Table 4.6.

4.1.4 Discussion

4.1.4.1 Exploratory Analysis

The unsupervised consensus clustering algorithm allowed the exploration of any potential relationship between PET radiomic features and BC histological characteristics. In a population of 79 women affected by high-risk BC and LABC, both oestrogen receptor expression and Ki67 status were associated with the TCs. Indeed, a positive ER expression was more frequent in the TC II (29%); high-Ki67 status was prevalent in the TC III (34%). Therefore, TC II more often comprised tumours with low SUVs, ER-pos expression and medium-high MTV. Conversely, BC with low ER expression, high Ki67 and low MTV were found in TC III.

Unsupervised clustering is reported in LABC PET radiomic literature. Huang et al.[98] performed unsupervised consensus clustering of 113 cases of BC based on [^{18}F]FDG PET- and MRI-derived parameters and three TCs were identified, that had a significant correlation with recurrence free survival. Ha et al.[91] likewise considered three individual TCs with distinctive metabolic radiomic patterns. The authors found that TC I, which had a high MTV, high SUVmax, and high intratumoral heterogeneity, was identified as an independent risk factor for recurrence when compared to the established parameters of high stage (TC III) and non-pCR. Notably, in accordance with Ha et al., in

Table 4.6. Combined Models.

Outcome	***Feature selection method***	***Feature***	***AUC_{RS}***	***OR***	***95% CI***	***p-values***	***AUC_{CV} [95% CI]***
OS	Wilcoxon - Fisher	Lymph nodes	0.83	0.09	0.02-0.43	0.0027	0.79 [0.73-0.85]
		ER-pos		7.60	1.69-34.2	0.0082	
		IVH - area under IVH curve		0.44	0.21-0.92	0.0286	
OS	mRMR	Lymph nodes	0.75	0.18	0.05-0.70	0.0137	0.75 [0.70-0.79]
		IVH - area under IVH curve		0.60	0.32-1.13	0.1122	
DFS	Wilcoxon - Fisher	Lymph nodes	0.79	0.15	0.04-0.50	0.0023	0.77 [0.71-0.83]
		High-Ki67		0.17	0.03-0.80	0.0257	
		IH - quartile coefficient of dispersion		1.81	1.00-3.27	0.0506	
DFS	mRMR	Lymph nodes	0.76	0.26	0.08-0.86	0.0278	0.72 [0.66-0.79]
		IH - quartile coefficient of dispersion		1.72	0.94-3.15	0.0791	
		IVH - volume at intensity fraction 90		0.67	0.38-1.20	0.1758	

OS, overall survival; DFS, disease-free survival; AUCRS, re-substitution AUC; AUCCV, cross-validated AUC; CI, confidence intervals; mRMR, minimum redundancy maximum relevance; IH, intensity-histogram; IVH, intensity-volume histogram, ER-pos, oestrogen-receptor positive; high-Ki67, high proliferation index.

our study, high-Ki67 and ER-pos expressions were found to be mildly associated with unsupervised clusters, regardless of slightly different extraction parameters and a different feature pool.

4.1.4.2 Modelling

Logistic regression was previously used in [^{18}F]FDG PET/CT breast radiomic studies for distinguishing breast carcinoma from breast lymphoma[100], decoding cancer phenotyping[98], predicting the pathological response to NAC[96], exploring the relationship between tumour heterogeneity and prognostic factors[101], predicting invasive component[102] and recurrence-free survival[98].

In the present study, logistic regression was employed to predict 5Y-OS and 5Y-DFS, and to evaluate the additional prognostic power gained by combining clinical and radiomic features. We found that the 'IVH - area under IVH curve' was negatively correlated with OS, showing an odd ratio below 1, meaning that patients with higher values had a worse prognosis. Van Velden et al.[103] showed that this feature is a quantitative index of tumour uptake heterogeneity, both in a simulation and in a clinical set with non-small cell lung cancer. In breast cancer, mainly for high FDG uptake values, heterogeneity is correlated with an aggressive histotype[104] and therefore with a high probability of an early recurrence.

Conversely, 'IH - quartile coefficient of dispersion' showed a protective role as a predictive variable for DFS (higher values were associated with a higher disease-free survival rate). This latter radiomic feature is associated with the histogram distribution of FDG uptakes values in the primary tumour. The presence of a tumour heterogeneity for low FDG uptake values would be linked with a lower tumour aggressiveness.

Wilcoxon and Fisher Score selection methods behaved similarly, leading to the same combined models for both OS and DFS; mRMR achieved lower performances.

For all combined models, based on both clinical and radiomic features, correlations among covariates were low (absolute correlation < 0.3), suggesting the possible complementarity of the selected clinical and radiomic covariates. Moreover, the inclusion of radiomic parameters in the clinical model (Wilcoxon - Fisher) increased the AUC_{RS} from 0.77 to 0.83 (AUC_{CV} 0.75 to 0.79) for 5Y-OS prediction and from 0.75 to 0.79 (AUC_{CV} 0.73 to 0.77) for 5Y-DFS prediction, thus contributing to a better stratification of patient prognosis. Therefore, the complementary role of clinical and radiomic information was shown in patients with high-risk and LABC. Similar result was recently reported by Antunovic et al.[96] in a cohort of 79 LABC patients who underwent [^{18}F]FDG PET/CT before NAC. They found that age and molecular subtype proved to be the sole, albeit weak, predictors of outcome in the simple model, with slight improvement in the model's predictive power upon addition of the textural PET-derived features.

In our study, age feature was not retained in any of the clinical models by the feature selection methods, and it was thus missing from combined models. However, Boughdad et al. suggested that age should be included in models involving textural features, to counteract the dependence of features on age[105]. The authors also reported that IH category features did not vary between age groups. In our study, clinical models were extended with the addition of first-order statistical features (belonging to IH and

IVH categories), which may be less prone to age dependence compared to textural features. Nevertheless, we evaluated whether the model performances improved by adding age in the combined models and no significant improvements on AUC were found after the inclusion (appendix Table C.3).

The tumoral heterogeneity in its various components (intratumoral, intertumoral and temporal) contributes to the resistance to treatment and recurrence of the disease[106]. Based on the study by Yoon et al., patients with elevated high-intensity short-zone emphasis (HISZE) and high-intensity zone emphasis (HIZE) in the primary breast tumours showed worse prognosis than patients with low HISZE and HIZE[107]. In the present work, we found that some radiomic variables indicative of tumour heterogeneity, such as the 'IVH - area under IVH curve' and the 'IH - quartile coefficient of dispersion', were complementary to clinical data for long-term prognosis in high risk and locally advanced BC. In accordance with Yoon et al., we confirmed the importance of regional heterogeneity on PET as a predictor of disease progression in LABC.

To deal with the limited sample size, AUC was assessed through both re-substitution and cross-validation approaches following the methodology proposed by Antunovic et al.[96], which, however, may be subject to information leakage since feature selection is performed ahead of cross-validation. When larger datasets are available, the recommended procedure is to split the population into separate train and test sets to perform an internal validation of the model[14], or better, confirm the result on a separate dataset (external validation).

Another aspect to consider is that the study is based on retrospective images acquired at a single institute with a single scanner. However, the resulting homogeneity of the collected data enabled to remove feature dependence on unwanted effects, such as acquisition and reconstruction parameters, which have been shown to have an impact on feature values[58,108–110].

Furthermore, PET examinations were acquired in the supine position rather than the prone one, as per clinical practice. Several papers have reported the advantages of prone position for PET images in patients with breast cancer[111,112], mainly for quantitative analysis[113]. Only Huang et al.[98] used the prone position for radiomic analysis, while other studies do not specify the patient position used for the acquisition[91,92,96,114].

Eventually, data balancing techniques were not considered due to the small cohort size.

4.1.5 Conclusions

Radiomic features improved the clinical data predictive performance for 5Y-DFS and 5Y-OS. Our study suggests that integrating clinical features and radiomics can increase the performance of prognostic models for high-risk cancer and LABC. Therefore, radiomic features should be further investigated in a large prospective study, to finally define their contribution for the prediction of 5Y-OS and 5Y-DFS in patients with LABC.

4.2 Role of radiomics analysis of [^{18}F]choline PET/CT in predicting biochemical recurrence in a cohort of intermediate and high-risk prostate cancer patients at initial staging

4.2.1 Introduction

Prostate cancer (PCa) is the most frequently diagnosed cancer in men and the fifth leading cause of death worldwide[115,116], even though its mortality rates have decreased in most high-income countries since the mid-1990s thanks to improvement in earlier stage detection and therapeutic options[117].

In PCa, risk stratification at staging is crucial to determine the optimal treatment strategies and, therefore, prognosis. The 5-year risk stratification in patients with primary PCa is mainly based on clinical stage, baseline prostate specific antigen (PSA) level and Gleason score (GS)[117]. However, biopsy sampling is prone to incorrectly grade PCa, often resulting in undergrading[118], and can determine side-effects[119]. The recent introduction of magnetic resonance imaging (MRI) fusion-guided biopsy has significantly improved the detection of primary tumours, although the agreement between MRI and biopsy is sub-optimal and the entire whole gland cannot still be assessed before the radical prostatectomy (RP). Although primary treatments, either RP or curative radiotherapy (RT), 20-50% of patients experience a biochemical recurrence (BCR) within 10 years from therapy[120–122]. Therefore, the pre-treatment assessment of BCR's risk would be essential to plan the appropriate treatment approach.

Positron emission tomography (PET) combined with computed tomography (CT) or magnetic resonance imaging (MRI) using several prostate-specific radiotracers (i.e., choline labelled with either ^{18}F and ^{11}C, [^{18}F]fluciclovine and prostate specific membrane antigen-PSMA ligands labelled with ^{68}Ga or ^{18}F) can help localize suspicious lesions in the prostate gland, providing a valuable tool for the detection of cancer and thus to guide biopsies and treatment. Since its introduction, PET with prostate-specific radiotracers has been proved to be a fundamental examination at the initial staging of disease[123,124] and it is currently recommended by several guidelines, especially in case of high-risk PCa[22,117,125]. A major advantage of imaging relies on the possibility to non-invasively and repeatedly sample an entire volume (whole tumour and/or any metastases), revealing its phenotypic characteristics over time, thus overcoming the invasiveness and sampling errors of biopsy[126].

In this context, artificial intelligence (AI) offers a promising adjunct to assist physicians in the analysis and interpretation of biomedical images, by performing tasks that would typically require human intelligence[127]. In oncology, AI-based models are often fed with features extracted from biomedical images, e.g., the radiomic features, combined with other clinical, demographic and/or histopathological parameters, to build predictive or prognostic mathematical models of clinical outcomes, such as overall survival, recurrence, risk factor and others. Specifically, radiomics is an evolving field in which large amounts of quantitative features are extracted from diagnostic medical images. These features may provide information linked to the underlying molecular and genetic characteristics, and thereby could be used to improve treatment response prediction and prognostication and potentially to allow personalisation of cancer treatment[126]. In particular, there is increasing interest in extracting additional characteristics from PET images that describe the heterogeneity of voxel intensities, that might be only subjectively measured or even missed by an expert eye, thereby providing additional, potentially relevant diagnostic information for clinical decision-making in a non-invasive manner[128,129].

A recent review about PET radiomics shows that, although some published studies have limited robustness and reproducibility because of small amount of data available (less than 50 of patients for the 30% of the works) and miss validation on external datasets or in an independent subsample of the initial dataset (for 28%), the interest in PET radiomics is increasing exponentially[130]. The majority of these studies have concentrated mostly on lung, head and neck, and gynaecological cancers, likely as a consequence of their diffusion, while data about PCa and PET radiomics are still limited[130,131]. To the best of our knowledge, in PCa patients, radiomics analysis has been investigated at initial staging, for recurrence detection and in case of metastatic disease by using mainly MRI[132]. However, PET radiomics has been shown to hold great potential in the assessment of tumour characterization, diagnosis and prognosis[130] .

The aim of the present study is to perform a radiomics analysis of [^{18}F]choline PET/CT images in a cohort of intermediate and high risk PCa patients, in order to predict BCR. Since radiomics is demanding in terms of data and large cohorts of subjects are not always available, we implemented a robust internal validation pipeline.

4.2.2 Materials and Methods

4.2.2.1 Patient population

For the study, we retrospectively selected patients with an intermediate and high risk PCa (according to the National Comprehensive Cancer Network-NCCN classification[133]) who underwent [^{18}F]choline PET/CT for initial staging of disease at the Veneto Institute of Oncology (Padua, Italy) from March 2013 to October 2019. The following inclusion criteria were used: 1) a confirmed intermediate- to high-risk PCa; 2) age $>$ 18 years; 3) patients who were candidates for radical prostatectomy and lymphadenectomy or radical radiotherapy; 4) accessible follow-up information and 5) no visible CT artifacts due to implants. Conversely, patients with a previous history of cancer and/or patients who were pre-treated with hormone therapy were excluded. The final database included 74 patients (median age: 73 years, range [43 - 86]). Baseline clinical, demographic, and biological data, such as age, PSA, histological subtype, pre-surgery GS, clinical stage and BCR were retrieved from medical records. Of these, baseline PSA, pre-surgery GS and clinical stage were considered in the analysis. Missing clinical values were imputed with the k-Nearest Neighbours algorithm. To be included in the model, GS and clinical stage were dichotomized: GS $\leq$ 7 versus GS $>$ 7 and T1/T2 versus T3, respectively.

All patients gave their informed consent for the use of their personal and clinical data; moreover, the retrospective use of data from clinical routine was performed according to institutional rules. All procedures performed were in accordance with the ethical standards defined by the 1964 Helsinki Declaration and its later amendments.

4.2.2.2 PET/CT acquisition, reconstruction, and interpretation

A whole-body PET/CT was acquired from skull vertex to proximal femur, with 6-7 beds, 2-3 min per bed, 60 min after intravenous administration of the tracer (3 MBq/kg of [^{18}F]choline). A low-dose whole-body CT scan (with no contrast enhancement; 140 kV, 80-120 mA) was used for attenuation correction and for the anatomical localization of the sites of disease. The PET data were reconstructed with an in-plane voxel size of 4 mm and a slice thickness of 2 or 4 mm. The processed images were displayed in coronal, transverse, and sagittal planes. [^{18}F]choline PET/CT images were jointly interpreted by two specialists trained to perform PET/CT imaging. The primary tumour was assessed by analysing the whole prostate gland and identifying areas with a focal tracer uptake.

4.2.2.3 Prostate delineations

The whole prostate gland (PG_{whole}) was delineated by two expert physicians on the CT data of the hybrid imaging. Delineations were subsequently transferred to the PET data and, whenever necessary, they were refined to exclude from the segmentation the uptake due to the spill-out effect of the tracer accumulated in the bladder. Two additional segmentations were obtained by applying two conventional thresholds to the SUV values inside the prostate gland: i.e., 41% of the maximum SUV value inside the prostate ($PG_{41\%}$), and SUV > 2.5 ($PG_{2.5}$). The resulting regions (i.e., PG_{whole}, $PG_{2.5}$, $PG_{41\%}$) were considered separately for the analysis. Figure 4.4 shows the three segmentations approaches, for a representative patient.

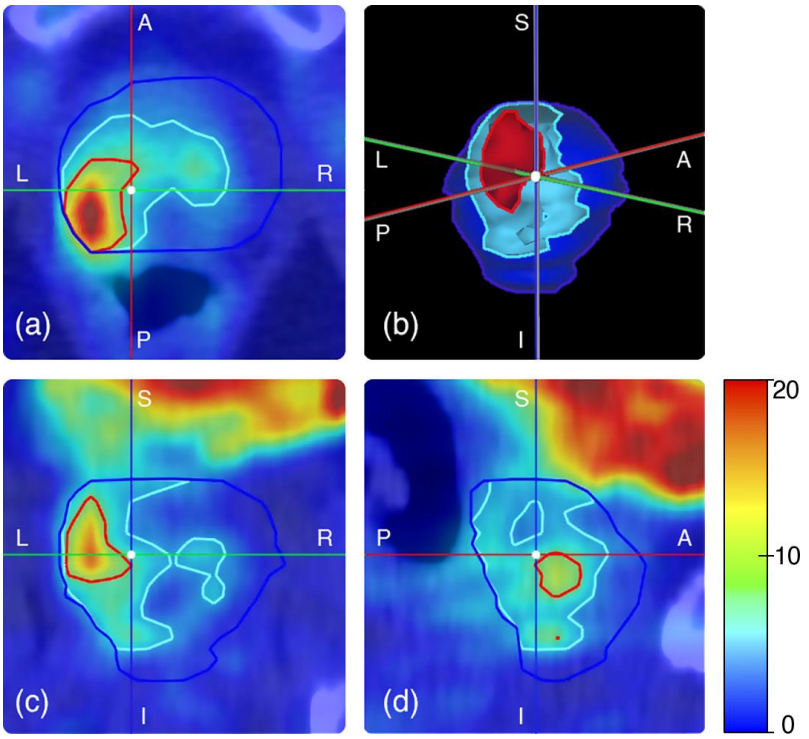

Figure 4.4. Fused PET and CT images of a representative patient in axial (a), coronal (c) and sagittal (d) views. High bladder uptake is visible in the upper portion of panels (c) and (d) as an inhomogeneous red-orange-yellow area. Panel (b) shows the corresponding 3D VOIs. In the figure, the blue line represents PGwhole, the light blue represents PG2.5 and the red one PG41% (S=superior, I=inferior, P=posterior, A=anterior, R=right, L=left).

4.2.2.4 Radiomic features

Radiomic feature extraction was separately performed on the three different PG segmentations using the open-source and IBSI-compliant software S-IBEX[60,134]. PET images were linearly interpolated to

obtain an isotropic voxel size of 2 mm and re-segmented in [0-20] SUV range. To compute features requiring SUV discretization, the fixed bin size (FBS) method was chosen using bin widths of 0.2, 0.4 and 0.6 SUV, which resulted in 3 different feature sets for each PG segmentation method, for a total of 9 combinations. Further details regarding feature extraction are reported in appendix Table C.1 and Table C.2.

Each combination of PG delineations/bin widths included 172 radiomic features, belonging to 11 feature families[13], describing the shape, intensity distribution and textural characteristics of the volume of interest (VOI). Finally, each patient had 9 different radiomic features sets.

4.2.2.5 Logistic regression models

A single standard pipeline was designed for the training of the prediction model and is depicted in Figure 4.5. At first, we considered only clinical data to train the baseline model and assess whether the available clinical parameters alone enclosed predictive information for BCR. Subsequently, clinical data were integrated with each of the nine radiomic feature sets, given by PG delineation/bin width. Eventually, the performances of models trained with radiomic features alone have been also assessed for comparison.

At first, features with absolute Spearman's correlation coefficient greater than 0.95 were removed from the dataset to reduce redundancy among predictors. Among highly correlated features, the one to keep was chosen at random, since they were considered equivalent in terms of informativeness. The remaining features were fed to a logistic regression model to predict the BCR in a 30-repeated hold-out validation procedure, with a train-test ratio of 3:1. For each training phase, we repeated the following steps:

1. Data balancing with the synthetic minority oversampling technique (SMOTE)[135].
2. Feature normalization using Z-score.
3. Feature selection was performed with "SelectFromModel" method of scikit-learn Python library on training set, which selected the features whose coefficients, calculated from a dedicated model (i.e., a linear support vector machine model with l1 penalty), were greater or equal to 1e-5.

4. Training of a logistic regression model combined with the least absolute shrinkage and selection operator (LASSO) to further selected the most informative parameters and predict BCR. A 5-fold cross-validation procedure was employed on the training set to optimize the regularization parameter lambda.
5. Model retraining on the whole training set using the optimal lambda.
6. Evaluation of prediction results through the area under the receiver-operating characteristic (ROC) curve (AUC), as well as balanced accuracy (ACC), specificity (SPEC), and sensitivity (SENS).

For all metrics, median and 5th-95th percentiles values on the 30 test sets of the validation procedure were derived. The entire analysis was implemented in Python programming using Scikit Learn and SciPy libraries (version 3.7).

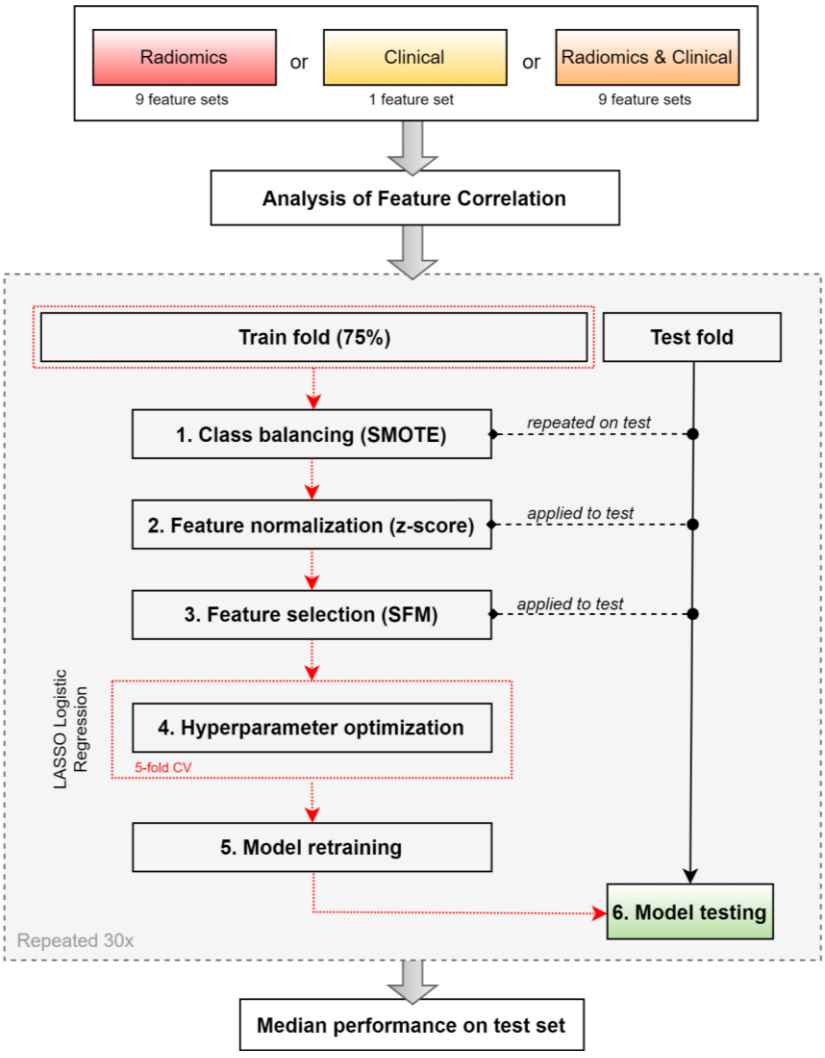

Figure 4.5. Training scheme for the LASSO logistic regression model.

4.2.3 Results

For our cohort, median PSA (that was missing for 2 patients) was 11 ng/ml, GS was ≤ 7 for 34 (46%) and GS > 7 for 40 (54%) patients, clinical stage was T1 or T2 for 66 (89%), T3 for 7 (9%) patients and missing for 1 (1%). Thirty-nine patients (53%) were treated with radical prostatectomy (with or without pelvic lymphadenectomy), while 35 (47%) underwent definitive radiotherapy. The BCR occurred in 28 (38%) patients. Median follow-up was 35.5 months (range 3.8-94 months). The PG_{whole} dataset included all 74 patients. Instead, 2 and 4 patients for $PG_{2.5}$ and $PG_{41\%}$, respectively, were discarded because their VOIs presented a volume smaller than 0.5 cm^3, which was not sufficient for a robust texture characterization of the volume of interest.

4.2.3.1 Biochemical recurrence prediction

For each feature set, defined by segmentation approach and bin width, the median and [5th - 95th] percentiles of the performance metrics of the LASSO logistic regression model are summarized in Table 4.7 for the baseline clinical model and for the models trained with both clinical and radiomic features and in appendix Table C.4 for the radiomic features alone.

Table 4.7. Medians [5th - 95th percentile] of the prediction results on the 30 test set folds for each segmentation/bin size feature set considering clinical or radiomic & clinical features (AUC = area under the ROC curve; ACC = balanced accuracy; SPEC = specificity; SENS = sensitivity; FBS = Fixed Bin Size; PG = Prostate Gland).

		AUC	***ACC***	***SPEC***	***SENS***
Clinical data only		0.73 [0.47 - 0.84]	0.69 [0.5 - 0.81]	0.69 [0.38 - 1]	0.69 [0.11 - 0.92]
PG_{whole}	***FBS 0.2***	0.67 [0.47 - 0.92]	0.65 [0.5 - 0.88]	0.77 [0.31 - 1]	0.69 [0.27 - 0.85]
	FBS 0.4	0.66 [0.51 - 0.89]	0.65 [0.56 - 0.83]	0.77 [0.46 - 1]	0.65 [0.15 - 0.85]
	FBS 0.6	0.62 [0.43 - 0.9]	0.65 [0.56 - 0.87]	0.69 [0.34 - 1]	0.69 [0.22 - 0.92]
$PG_{2.5}$	***FBS 0.2***	0.66 [0.42- 0.91]	0.67 [0.54 - 0.86]	0.83 [0.54 - 1]	0.58 [0.16 - 0.83]
	FBS 0.4	0.78 [0.62 - 0.88]	0.75 [0.62 - 0.86]	0.83 [0.50 - 1]	0.75 [0.37 - 0.83]
	FBS 0.6	0.69 [0.35 - 0.9]	0.67 [0.5 - 0.86]	0.92 [0.54 - 1]	0.58 [0.08 - 0.80]
$PG_{41\%}$	***FBS 0.2***	0.69 [0.47 - 0.88]	0.69 [0.56 - 0.83]	0.88 [0.46 - 1]	0.65 [0.19 - 0.85]
	FBS 0.4	0.72 [0.41 - 0.89]	0.69 [0.54 - 0.83]	0.92 [0.46 - 1]	0.62 [0.08 - 0.89]
	FBS 0.6	0.72 [0.47 - 0.91]	0.71 [0.56 - 0.85]	0.92 [0.45 - 1]	0.58 [0.15 - 0.85]

The baseline clinical model achieved good performances with a median AUC of 0.73 (Figure 4.6a). The highest performance scores were obtained by the model trained on $PG_{2.5}$ volumes using a bin size of 0.4 SUV. For the model, median AUC, ACC, SPEC, SENS on the 30 test folds were 0.78, 0.75, 0.83, 0.75 (Figure 4.6b).

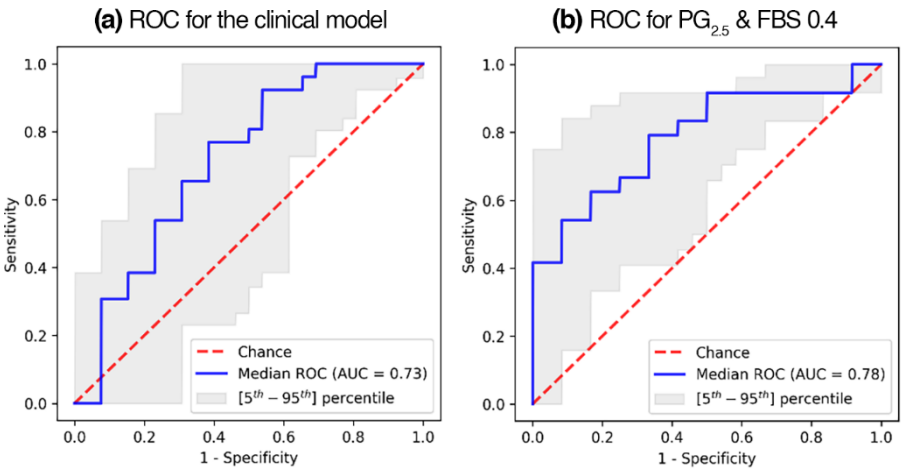

Figure 4.6. (a) Median ROC curve of the baseline clinical model and (b) for the model trained on $PG_{2.5}$ 0.4 SUV considering radiomics and with clinical features (ROC = Receiver Operating Characteristics curve; AUC = area under the ROC curve; FBS = Fixed Bin Size; PG = Prostate Gland).

PG_{whole} had the lowest performances: for all bin sizes median AUC was lower than 0.70 and all other metrics obtained similar scores. $PG_{41\%}$ segmentation reached AUC values of 0.72 for 0.4 and 0.6 bin widths. For this approach, all bin widths obtained good scores in terms of specificity, but at a price of a reduced sensitivity. Overall, considering all ten feature sets, sensitivity was the metric with the lowest median and percentile scores. The highest value for the sensitivity, equal to 0.75, was obtained for the model $PG_{2.5}$ & FBS 0.4 SUV, followed by the baseline clinical model, where sensitivity was equal to 0.69. On the contrary, specificity achieved the highest median scores (up to 0.92) for almost every feature set.

Compared to the other models (except for $PG_{2.5}$ & FBS 0.4 SUV), clinical data alone obtained good prediction results, with improved median AUC (0.73) but worse results for specificity.

Models trained with radiomic features only confirmed that the best combination was the one formed by $PG_{2.5}$ & FBS 0.4 SUV for almost all metrics, and that radiomic features alone contain predictive information for the BCR (appendix Table C.4). Nevertheless, the importance of including clinical features in the model is supported by the fact that, for all models, GS and PSA were the most frequently

selected features, being included in the model more than 20 out of 30 times in the model validation procedure. Furthermore, the resulting best model was able to point out the radiomic features that mostly contributed to the BCR prediction. Besides the GS selected 30/30 times, the model identified the 'centre of mass shift' of the morphological feature family, and the 'maximum histogram gradient intensity' of the intensity histogram feature family, as equally important predictors of BCR, both included in the model 30/30 and 28/30 times.

4.2.4 Discussion

In the present study, we tested the utility of radiomic analysis for the prediction of BCR in a cohort of intermediate and high-risk PCa patients undergoing [^{18}F]choline PET/CT at the initial staging of disease.

The results of our analysis show that, with respect to the baseline clinical model, based on PSA, GS and Clinical stage, BCR prediction performance further increases when clinical data are complemented with radiomic features. In particular, we found that the combination of a specific PG segmentation ($PG_{2.5}$) with a 0.4 SUV discretization approach is the best way to process the original PET image in view of the prediction of BCR. This means that discarding low SUV values inside the prostate by setting a 2.5 SUV threshold is beneficial for the analysis: more precise with respect to considering the whole prostate gland and more conservative than the 41% threshold. Similarly, in the study by Tu et al.[136], the authors went beyond the traditional tumour-centric view of radiomic analysis and divided the whole prostate organ of 77 patients in three radiomic zones: the metabolic tumour zone, the proximal peripheral tumour zone, and the extended peripheral tumour zone inside the imaging boundaries of the organ). The authors found that these zones have different predicting strengths in classifying risk groups. Their study supports the hypothesis that radiomics features extracted from Choline PET images can be predictive of several clinical endpoints, and shows that, depending on the outcome, the useful information might be confined in specific areas of the gland.

As for the bin width used for SUV discretization, it resulted that the trade-off between the investigated bin sizes was the most successful. This may be due to the fact that using smaller steps may reduce the beneficial noise-suppressing property of discretization, while larger steps may determine an information

loss, with different intensity values being condensed within the same bin, thus becoming undistinguishable.

To the best of our knowledge, our study is the first that correlates radiomics features to BCR events using [^{18}F]choline PET/CT. Indeed, some papers are now available about the use of radiomic models to predict the aggressiveness of PCa by using both radiolabelled choline and PSMA PET/CT or PET/MRI, while few data about outcome are at disposal[131,136–138].

However, growing evidence supports the use of risk stratification tools that combine clinical parameters, genomic biomarkers, morphological and functional features able to either optimize health care and predict BCR in PCa patients[139–142]. Nevertheless, the lack of validation of these predictive tools in prospective randomized clinical trials represents the main limitation for their introduction in clinical practice. Methodology standardization, data sharing, and software accessibility are deemed additional important factors to increase the applicability and reuse of published studies. In this work, we adopted an open-source and highly standardized radiomic software, S-IBEX[134], to perform a complete and reproducible radiomic feature extraction.

This study has some limitations. First, data were retrospectively collected. Second, it is built on a single-centre cohort, as other studies in the field[137,143], and an external validation, that would have allowed to assess the robustness of the findings, is missing. Nevertheless, the approaches we used for data preparation (i.e., redundancy reduction through correlation analysis, feature selection, class imbalance correction) as well as the cross-validation scheme implemented, minimized the chances of biased results, increasing generalizability and allowing to handle the relatively small sample size. The cross-validated prediction results indicate that our model was able to identify patients at risk of BCR in independent data.

Patients' management included two diverse types of treatments (i.e., surgery or radiotherapy). However, for the purpose of this study, treatment did not affect the validity of prediction results. Moreover, some authors have demonstrated similar outcomes for patients with high-risk PCa, independently from the curative treatments[144,145].

4.2.5 Conclusions

This study demonstrates the feasibility of radiomic analysis of PET imaging to extrapolate the useful information for the stratification of patients at risk of BCR, especially when PG uptake is thresholded to discard low (i.e., SUV < 2.5) and non-specific SUV values that could be imputed to imaging limitations or poor tracer specificity.

In future studies we aim to investigate the validity of proposed methods with novel tracers and imaging approaches such as ^{68}Ga-PSMA PET/MRI. However, prospective, multicentric studies are needed to investigate the clinical application of our findings and to fully explore the role of PET radiomics in clinical practice. Integration of clinical data, biochemical parameters, and radiomic features may greatly act as a multi-modal system to add prognostic information at initial staging of PCa with the final purpose of addressing a tailored treatment strategy.

4.3 CECT-based radiomic prediction of vascular invasion for resected hepatocellular carcinoma (HCC) patients

4.3.1 Introduction

Worldwide, liver cancer is one of the leading causes of cancer-related death[146] and ranks second among most lethal tumours with a 5-year survival of 18%. Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer, mostly occurring in patients with underlying liver disease (e.g., hepatitis B or C virus infection, non-alcoholic fatty liver disease) or alcohol use disorder.

Multiphase CT imaging with contrast agents (contrast enhanced computed tomography - CECT) is suggested by many clinical practice guidelines[147,148] as one of the first-line modalities for the diagnosis and staging of HCC. CECT imaging should include at least the arterial and portal venous phases, being the typical HCC hallmark its hyper vascularisation in the arterial phase and its washout in the portal venous phase.

HCC tumour staging accounts for the number and size of nodules and the presence of distant metastasis or vascular invasion. In particular, vascular invasion refers to an invasive manifestation of the tumour and is of interest since HCC commonly involves local branches of the portal and/or hepatic veins, potentially leading to tumour thrombus even at early stages and to a consequent spreading of the disease.

As for treatment options, nodule resection is the preferred intervention for patients with an early-stage solitary tumour, irrespectively of tumour size, if the patient's performance status is good and liver function is preserved[149]. Radiofrequency ablation is recommended for patients not eligible for surgery (lower local control with respect to nodule resection)[148], while transarterial therapies (e.g., transarterial chemoembolization - TACE) are advised for intermediate-stage tumours and ultimately systemic therapies for advanced diseases. Eventually, liver transplantation is an option for patients with a limited disease who are not candidate for resection. However, the presence of vascular invasion is a contraindication for transplantation being an independent prognostic factor linked to a higher risk of tumour recurrence after surgery.

Vascular invasion can be classified into *macroscopic* vascular invasion, that is the tumour involvement of large to medium vessels, and *microvascular invasion* (MVI), which refers to the microscopic

involvement of small vessels. While macrovascular invasion can be assessed with conventional imaging (such as CT, MRI or ultrasounds), the presence of MVI is only assessed at the tissue level through histopathology[150]. Many studies underline a relationship between MVI and poor outcomes, with a higher risk of recurrence in patients with MVI[151,152].

At present, MVI has been the target prediction of several studies focused on its non-invasive assessment through imaging since, determining whether MVI exists before surgery, would allow to better tailor the treatment. One study, by Kornberg et al.[153], investigated the expression of a [^{18}F]FDG positron emission tomography (PET) metric (the maximum standardized uptake value - SUVmax) and reported that its increase could predict MVI, while other studies[49,154–161] applied texture-analysis techniques to several image modalities, with a preponderance for CECT.

In recent years, the idea that the information perceivable by the naked eye is only a fraction of that contained in the imaging, has become increasingly acknowledged[162] and texture analysis has proven its capability to accurately convert radiological images into minable quantitative characteristics. Radiomics is the non-invasive approach that combines a high-throughput extraction of those image-derived features with modern machine learning techniques for the development of prediction models able to improve over the unaided and subjective assessment of images carried by physicians[14]. On top of that, image convolutional filtering, that enhance specific characteristic (e.g., blobs, tubular structures, edges) of the biomedical image, allows to further extend the amount of extractable information. In the last years, filtering techniques were not always reproducible across different radiomic software tools, hampering the reproducibility of the studies that used them. Currently, the Image Biomarker Standardization Initiative (IBSI) is finalising the standardization of convolutional filtering in radiomics and provides guidelines, definitions and implementation details which will results in a further harmonisation of radiomics tools.

In this study, we investigated the application of radiomics analysis for the prediction of vascular invasion on resected HCC patients imaged with CECT. In particular, the analysis considered the employment of a Laplacian of Gaussian (LoG) filter to understand whether model predictive performances could improve by integrating clinical data with radiomics features extracted from both original and filtered images. LoG filter was chosen mainly for two reasons: it was previously employed

in several radiomic HCC studies[163], and, when this study was designed, it was one of the most standardized filters according to IBSI-2.

4.3.2 Materials and methods

4.3.2.1 Population and clinical data

Patients with a diagnosed HCC who underwent liver resection between March 2013 and November 2020 at Padua University Hospital were retrospectively collected for this study. The inclusion criteria were as follow (1) availability of both the arterial and portal phase of a contrast-enhanced computed tomography CECT prior to liver resection, (2) histopathology confirmed HCC, (3) presence of naïve HCC nodules (locoregional untreated disease), (4) age $>$ 18 years. Patients with evident imaging artefacts or poor image quality were discarded. For each patient, each untreated nodule was included in the analysis. As per institutional practice, all surgery specimens underwent histopathology examination for the assessment of MVI.

Clinical data were retrieved from medical charts and comprehended age, sex, cirrhosis status, Child-Pugh score, MELD score, alpha-fetoprotein levels, and the number of nodules visible on pre-operative imaging.

This single-centre retrospective study was approved by the institutional review board, and written informed consent was waived.

4.3.2.2 Volume of interest delineation

All the images were retrieved from the PACS and anonymized. Three-dimensional segmentations of nodules were manually performed on both the arterial and portal phases of the CECT imaging using Eclipse (Varian Medical Systems, Palo Alto, CA) software. Readers used the default liver window of [-25, 125] Hounsfield Unites (HU) to increase reproducibility. An adaptive 3D brush or a slice-by-slice contour delineation were the tools at the disposal for the segmentation task and their usage was at the discretion of the reader. Thirty-one nodules were randomly chosen to be independently and blindly delineated three times by three expert radiologists in abdominal imaging review for the purpose of inter-reader variability assessment. The remaining nodules were segmented by one of three radiologists. Eventually, all segmentations were reviewed and approved for the study.

4.3.2.3 Radiomic feature extraction

From each nodule and for both the arterial and portal phases, 174 radiomic features were extracted using the open-source, IBSI-compliant[134] software S-IBEX[48,60]. The images were resampled with linear interpolation to get an isotropic voxel size of 1mm and were re-segmented in the range [-400, 400] HU. The features described the nodules' morphology, intensity distribution and textural pattern. Whenever discretisation was necessary (see Table 1.3), fixed bin size (FBS) method was employed with a bin width of 25 HU. For each of the 174 features we additionally computed their percentage difference, f_{DIF}, between the arterial and portal phases as follow:

$$f_{DIF}^{i} = \frac{f_{ART}^{i} - f_{POR}^{i}}{|f_{ART}^{i}|}$$

with i identifying the specific feature, f_{ART} the feature value computed in the arterial phase and f_{POR} that of the portal phase.

Moreover, to both the arterial and portal phases eight Laplacian of Gaussian (LoG) filters were applied by varying their scale parameter, σ^*, in the range [0.5, 4] cm with increasing steps of 0.5 cm. From each of the eight filtered images, 145 (all but the morphological) radiomic features were extracted. The feature-extraction workflow is depicted in Figure 4.7.

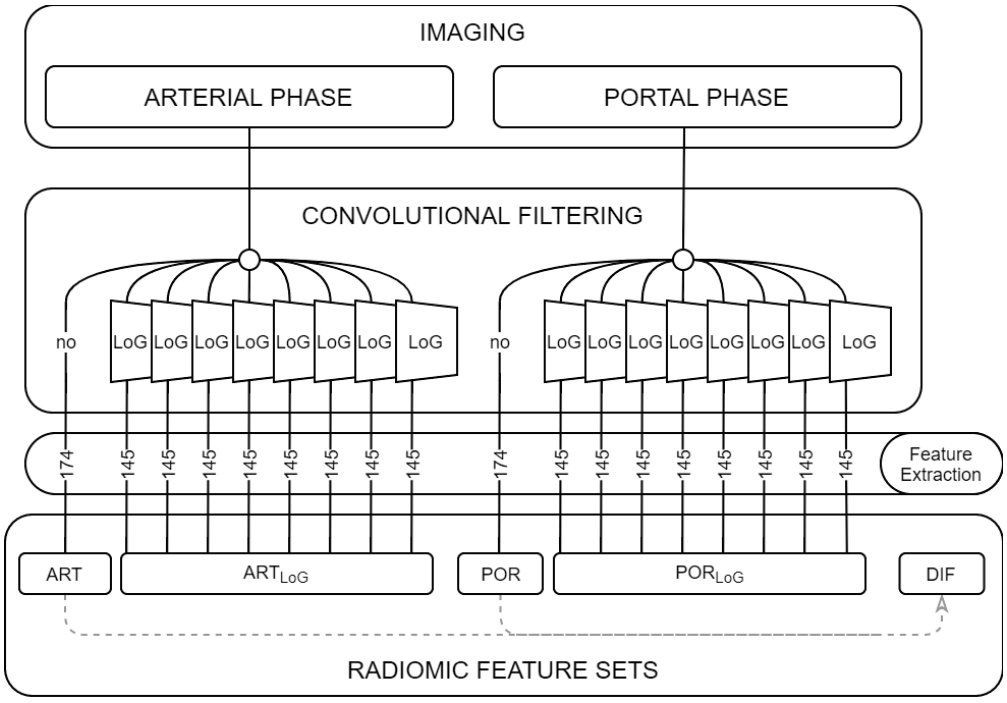

Figure 4.7. Radiomic feature extraction scheme.

When filtering was employed, re-segmentation was skipped, and the fixed bin number approach (FBN) was preferred for discretization (with N equal to 32). In summary, for each nodule segmentation, a total of 2842 features were extracted belonging to 5 feature sets: the arterial (ART), portal (POR), percentage differences (DIF), filtered arterial (ART_{LoG}) and filtered portal (POR_{LoG}). The feature extraction parameters for each feature set are reported in Table 4.8. Further reporting of extraction parameters is available in appendix Table C.1 and Table C.2.

Table 4.8. High-level extraction parameters and the number of extracted features from the arterial and portal phases. LoG: Laplacian of Gaussian, applied with $\sigma^* = [0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4]$ cm.

FEATURE SET	INTERPOLATION	RE-SEGMENTATION	DISCRETIZATION	# OF FEATURES
ART	*1x1x1 mm*	*[-400, 400] HU*	*FBS: 25 HU*	*174*
POR	*1x1x1 mm*	*[-400, 400] HU*	*FBS: 25 HU*	*174*
DIF	*not applicable*	*not applicable*	*not applicable*	*174*
ART_{LoG}	*1x1x1 mm*	*no re-segmentation*	*FBN: 32*	*1160 (145x8)*
POR_{LoG}	*1x1x1 mm*	*no re-segmentation*	*FBN: 32*	*1160 (145x8)*

4.3.2.4 Statistical analysis

Feature reduction

To reduce the number of features considered in the analysis, and thus lowering the risk of overfitting, the assessment of feature reproducibility was employed, based on the available multiple segmentations. At first, segmentations' concordance was evaluated by mean of pairwise DICE score, with the intent of potentially excluding, from feature reproducibility assessment, segmentations whose DICE was lower than 0.5. Subsequently, the intraclass correlation coefficients – ICC (two-way mixed effect single-rater absolute agreement) was computed on all the remaining ROIs for all the features set (i.e., ART, POR, DIF, ART_{LoG} and POR_{LoG}) to evaluate the reproducibility of radiomic features to multiple readers (inter-observer variability). Only features presenting an ICC value greater than 0.9 were considered reproducible and were selected for further analysis.

Model building

This study employed a binary classification framework for predicting the presence/absence of vascular

invasion (both macrovascular and MVI). As an input to the modelling part, 13 combinations of feature sets were considered (reported in Table 4.9) with and without the inclusion of clinical data.

Table 4.9. The 13 considered combinations of feature sets. All the combinations were tested both with and without the inclusion of clinical data.

Combination ID	*ART*	*POR*	*ART LoG*	*POR LoG*	*DIF*
A--	✓				
-P-		✓			
--D					✓
AP-	✓	✓			
-PD		✓			✓
A-D	✓				✓
APD	✓	✓			✓
A--LoG	✓		✓		
-P- LoG		✓		✓	
AP- LoG	✓	✓	✓	✓	
-PD LoG		✓		✓	✓
A-D LoG	✓		✓		✓
APD LoG	✓	✓	✓	✓	✓

For each feature-set combination, the same pipeline was used (visible in Figure 4.8). The modelling, training and evaluation part were implemented in R using the open-source package "Familiar" (version 1.2.0).

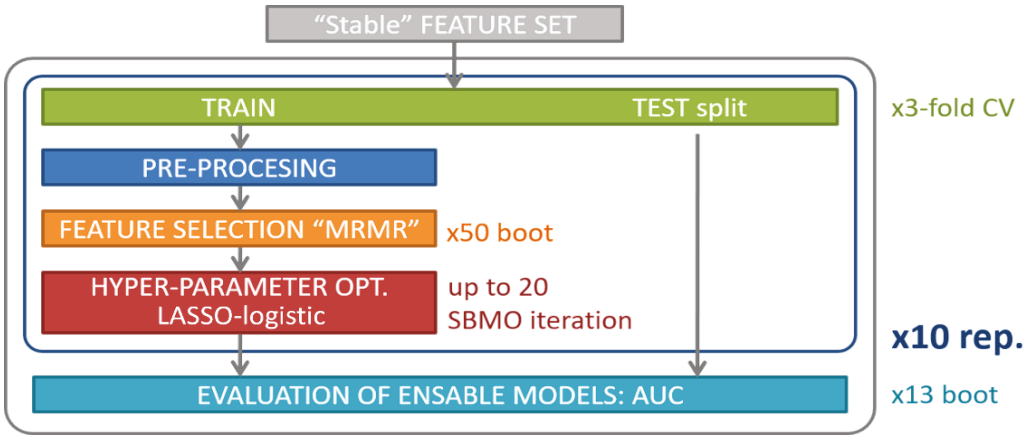

Figure 4.8. Training scheme for the LASSO logistic model. CV: cross-validation; boot: bootstrap; SBMO: Bayesian Sequential Model-Based Optimization; mRMR: minimum redundancy maximum relevance; LASSO: lasso least absolute shrinkage and selection operator.

In a 10-repeated 3-fold cross-validation scheme a 3-step training procedure was performed:

1. Preprocessing: feature transformation (i.e., Yeo-Johnson) and data imputation for missing values (with the median value of the feature) were applied and highly correlated features were discarded. Hierarchical-clustering based on Spearman correlation was used and, for each cluster, only the feature with the highest importance according to univariate regression with the outcome was retained.
2. Feature selection: the mRMR univariate filtering method was used. The mRMR ranking was obtained on 50 bootstraps and aggregated with BORDA technique.
3. Hyperparameter optimization: thanks to the Bayesian Sequential Model-Based Optimization (SBMO) the best hyperparameter of the LASSO logistic regression model were found (i.e., the regularization parameter, lambda, and the number of features to be included in the model, in the range 1-5).

Performances of the model were evaluated in terms of area under the receiver-operating characteristic (ROC) curve (AUC). Bias-corrected estimates of AUC, accuracy (ACC), sensitivity (SENS), and specificity (SPEC) and their confidence level (at 0.95) were obtained with bootstrap[164].

4.3.3 Results

4.3.3.1 Population

A total of 80 patients were enrolled in this study and a total of 89 naïve nodules were segmented. Table 4.10 details the cohort demographics and tumour characteristics. Thirty-one of these nodules were delineated three times by different radiologists, while the remaining 58 nodules (belonging to 49 patients) just once.

Table 4.10. Demographics and clinical characteristics of the cohort.

Variable	**N (percentage) or median [range]**
Number of patients	80 (100%)
Sex: *male/female*	66 (82.5%)/14 (17.5%)
Age, median [range]	67 [17 - 82]
Death	25 (31.3%)
Survival time (days), median [range]	836 [24 - 3244]
Cirrhosis	57 (71.3%)

Child-Pugh:	
A	74 (92.5%)
B	6 (7.5%)
MELD, median [range]	8 [6 - 42]
Alpha-fetoprotein:	
not available	4 (5%)
median [range] (ng⁄mL)	9.81 [1 - 26687]
Number of nodules per patient:	
1	50 (62.5%)
2	11 (13.8%)
3	10 (12.5%)
>3	9 (11.3%)
Vascular Invasion	*(over 89 nodules)*
not present	50 (56.2%)
microvascular	25 (31.3%)
macrovascular	14 (17.5%)

4.3.3.2 Feature reduction

Dice score coefficients were calculated pairwise for the three readers on the 31 nodule delineations. Figure 4.9 shows examples of high and low overlap (based on DICE score) between reader delineations.

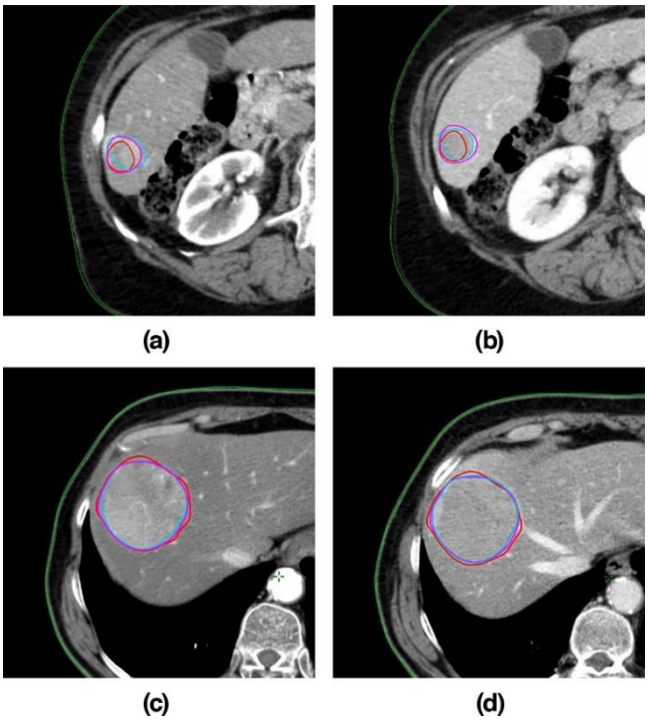

Figure 4.9. Example delineations for two patients by three different readers with different overlaps for arterial (left) and portal (right) images. The dice similarity coefficient ranged from ~0.62 (a-b) to ~0.9 (c-d).

All DICE coefficients are reported in Figure 4.10.

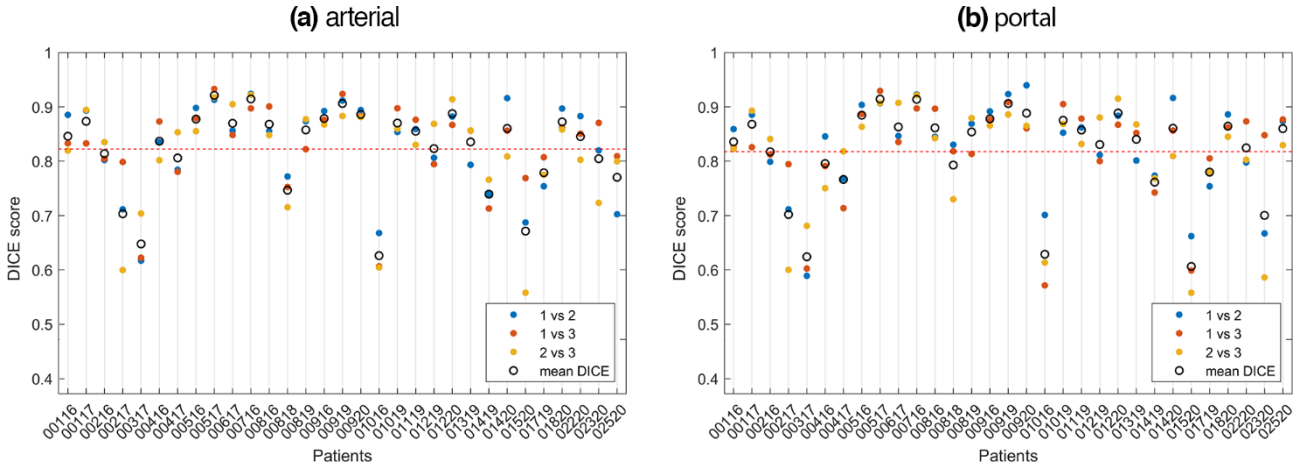

Figure 4.10. Pairwise DICE score for the 31 nodules with multiple segmentations for the arterial (a) and portal (b) phases.

For both the arterial and portal phases, the median DICE coefficient was above 0.8. From the figure, we can notice that few patients had DICE coefficients lower than others. Upon inspection, those patients presented less defined HCC boundaries which possibly resulted in a lower inter-reader agreement. For the calculation of ICC values, no nodules were discarded because of low DICE. The percentages of stable features, stratified by feature sets are visible in Figure 4.11. Globally, from the total of 2842 features, 719 had an ICC greater than 0.9. The reproducibility of features derived from unfiltered images were in accordance across the arterial and portal phases. The ICC values of those features are reported in Figure 4.12, stratified by feature family.

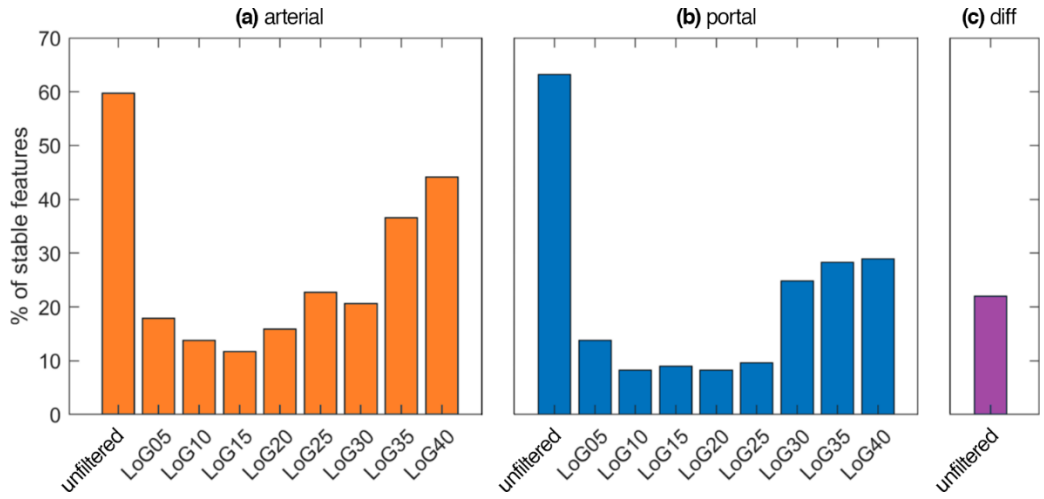

Figure 4.11. Percentages of reproducible features (ICC higher than 0.9 assessed through multiple-reader segmentation) per feature sets: (a) arterial, (b) portal and (c) difference. LoG: Laplacian of Gaussian.

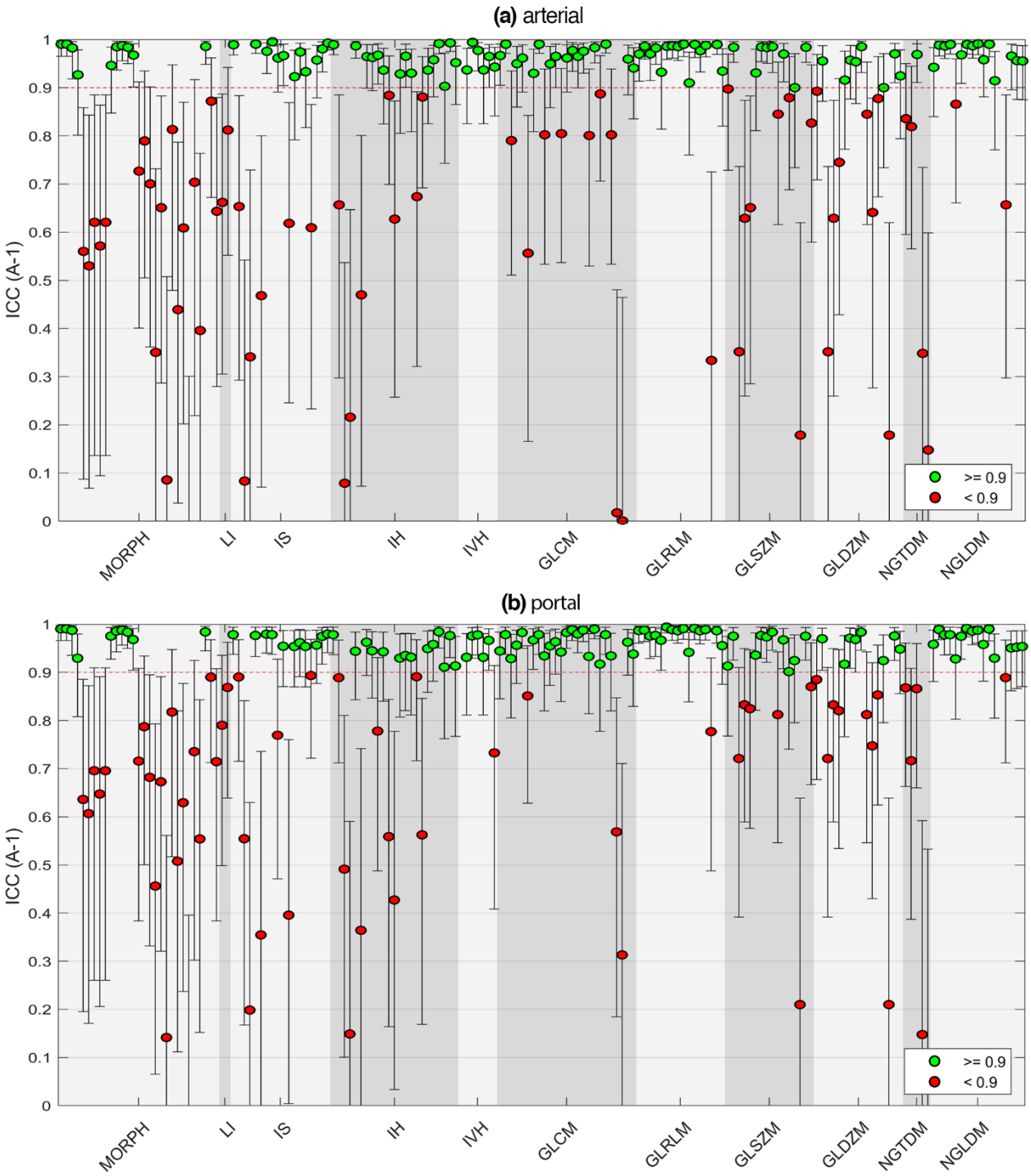

Figure 4.12. ICC values for the 174 features were computed on the unfiltered arterial (a) and portal (b) phases.

4.3.3.3 Modelling

Performances of all tested models are reported in Figure 4.13 and Figure 4.14, without and with the inclusion of ART_{LoG}/POR_{LoG} features, respectively. The baseline clinical model was among the worst performing with a median AUC of 0.55 (Figure 4.13a). The highest performance scores were obtained by the model trained on POR, without the inclusion of clinical data, which achieved

median AUC, ACC, SENS, SPEC of 0.78, 0.74, 0.64, 0.82, respectively (Figure 4.13b). In general, when clinical features were included in the feature set (Figure 4.13c), they did not improve model performances (e.g., the model comprising both POR & clinical features achieved an AUC, ACC, SENS, SPEC of 0.75, 0.70, 0.62 and 0.76). This, together with the fact that clinical model performances were low, strengthen the idea that the collected clinical variables did not carry, in this setting, relevant information for the prediction of vascular invasion.

When LoG-derived features were considered, performances were slightly lower than models relying on features extracted from unfiltered images (Figure 4.14a). Once again, the highest performances were achieved when the model only comprised POR features (AUC, ACC, SENS, SPEC of 0.77, 0.74, 0.65 and 0.82) and lowered when clinical features were added (Figure 4.14b).

Overall, considering all feature sets, sensitivity was the metric with the lowest median and percentile scores. The highest value for the sensitivity, equal to 0.67, was obtained for the model based on ART_{LoG}, POR_{LoG} and DIFF features, without clinical features.

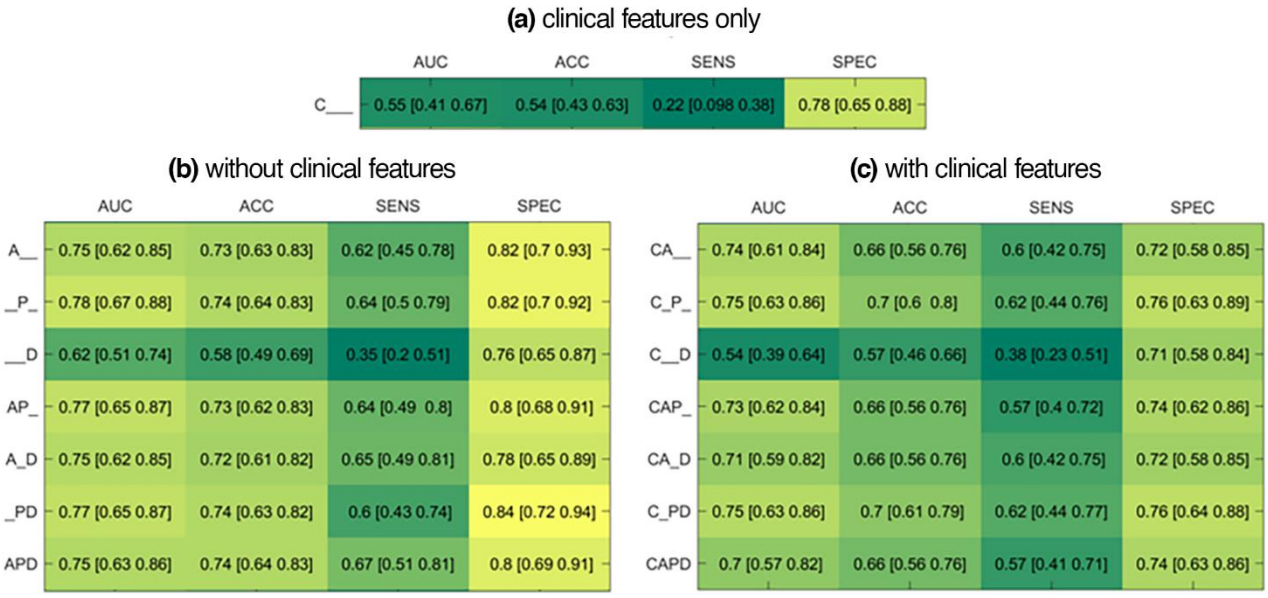

Figure 4.13. Performance metrics (median value and 95th bootstrap confidence intervals) of the tested models based on (a) clinical features, (b) radiomic features and (c) clinical and radiomic features from unfiltered images.

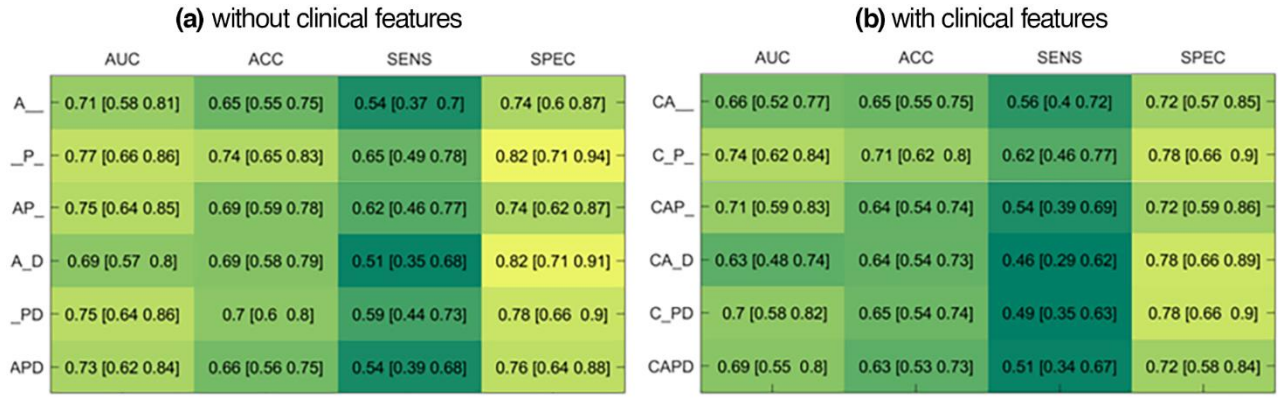

Figure 4.14. Performance metrics (median value and 95th bootstrap confidence intervals) of the tested models based on (a) radiomic features and (b) clinical and radiomic features from LoG-filtered images.

4.3.4 Discussion

The presence of vascular invasion, macrovascular or MVI, is one of the main, independent factor for the prediction of early recurrence in patients undergoing surgical resection or liver transplantation[165].

In this work, we investigated radiomics for the prediction of vascular invasion in patients with resected HCC patients imaged with CECT. Moreover, we evaluated the impact of LoG image filtering on model performances.

The results of our analysis show that, with respect to the baseline clinical model, vascular invasion prediction performances were higher when models were built using radiomic features. In particular, we found that the usage of portal features led to slightly higher performances than arterial ones or their combinations. This performance difference was observable and consistent both when clinical data were taken into account, and when LoG was considered.

Feature reduction based on multiple-segmentations played a fundamental role in dealing with the high number of features that were extracted from filtered images. As an example, up to 90% of features were not reproducible when the LoG filter was utilized with the scale parameter $\sigma^* = 1.5$ cm in both phases. Non-reproducible features vary with minimal delineation perturbations; thus, they do not reliably characterize the ROI content. Since delineations differences arbitrarily arise due to inter-reader subjectivity, features variations are unpredictable and features values possess a great noise-driven component. In general, lower feature reproducibility was observed on filtered images, with increasing stability at higher spatial scales. Since LASSO feature selection do account for feature reproducibility,

without a proper feature reduction step based on multiple-segmentation, it could have selected unreproducible features, thus hampering the generalizability of the models.

The analysis suggested that, from a radiomic perspective, information contained within the GTV on the portal phase might be more prognostic than the one on the arterial one: the typical HCC washout on the portal phase[166], which makes the lesion appear as hypo-dense as compared to the surrounding liver parenchyma (specificity of 95–96% for HCC diagnosis) might be the driver of the radiomic-extracted information. Moreover, image filtering did not seem to provide any additional information for vascular invasion prediction.

The present study has some limitations. The collected data belonged to a small retrospective single-institution cohort of patients, which might not fully capture the variability of HCC disease, and an external validation is missing. In addition, only naïve nodules that underwent surgical resection were considered for the analysis. In future studies, larger cohorts, including liver transplant patients, will be needed to validate our findings.

Moreover, our model was developed on CT imaging data having heterogeneous acquisition and reconstruction protocols, which, although it might represent the present de facto situation of several institutions, might also introduce confounding factors in the analysis. Nevertheless, constrains on imaging protocols were introduced (e.g., ranges of acceptable slice thickness) as well as image-processing steps that aimed at image harmonization (e.g., resampling to single isotropic voxel size). To avoid overfitting we implemented several strategies, from a bootstrapped feature selection to a cross-validated model building.

Currently, vascular invasion is one of the main, independent factors for the prediction of early recurrence after surgery. Still, MVI can only be diagnosed post-treatment through the histopathological examination of the resected sample. A non-invasive pre-operative vascular invasion assessment would allow to choose more consciously the best resection procedure (if MVI is present, higher 5-year survival rate was found for anatomical resection with respect to wedge resection[157]) and to better plan the hepatectomy extension.

4.3.5 Conclusions

In conclusion, several models were evaluated and compared for the pre-operative diagnosis of vascular invasion for HCC, based both on clinical data and radiomic features extracted from original and filtered CECT phases. Results showed the feasibility of the analysis and suggested that highest performances are to be obtained when analysing the HCC nodules on the portal phase. However, given the additional computational cost and the slightly lower performances we could not advise employing LoG filter in this setting. The IBSI-standardized filtering and feature extraction will ease the design of future studies, needed for the validation of the current results, fastening the translation of these models into the clinical practice.

4.4 Other studies

In addition to the above, S-IBEX was employed in other studies for texture characterization of diverse image modalities (e.g., digital mammographs, syntenic mammographs) and on dose maps of radiotherapy plans (i.e., a newly developed branch of radiomic called "dosiomic").

4.4.1 Breast Density prediction on digital and synthetic mammograms

For this study, S-IBEX was employed on screening digital mammograms as well as on synthetic mammograms (derived from the digital breast tomosynthesis) to characterize breast parenchyma. The aim was to build a predicative model that could automatically assess breast density (low vs high) that is one of the major risk factors considered in the Breast Imaging-Reporting and Data System (BI-RADS) evaluation. Breast density is usually determined through visual inspection by an expert physician, but radiomic could provide a more efficient and reproducible way to assess it.

This work was a collaboration with "Istituto Tumori della Romagna IRST - IRCCS" where 93 subjects were acquired for screening purposes and S-IBEX was employed to perform a 2-dimensional image analysis (e.g., 2D bilinear interpolation, 2D feature-aggregation method). A square ROI of 5 cm^3 was placed 3 cm behind the nipple on the mediolateral oblique (MLO) view of both digital and synthetic mammograms. Feature reproducibility was assessed by considering, for each ROI, 9 additional segmentations with a random maximum displacement of 5mm from the original ROI. After feature reduction based on ICC and Pearson's correlation coefficient, a LASSO-penalized logistic regression model was trained to predict breast density classes and a 3-fold cross validation scheme was used for hyperparameter tuning.

As a result, digital mammograms resulted superior to synthetic mammograms for radiomic characterization of breast density (median AUC of 0.76 versus 0.68).

4.4.2 Prediction of dysgeusia based on dose maps in the setting of head & neck cancer

Radiotherapy is one the most consolidated treatments for head & neck cancer, which however is also associated with acute and late dysgeusia (taste alteration/loss). In this study, we wanted to investigate the relationship between the heterogenicity of the dose delivered to the tongue and dysgeusia.

Eighty head & neck cancer patients treated with radiotherapy at Veneto Institute of Oncology, having at least a 24-month follow-up, were selected. Information about dysgeusia was acquired at each follow-up time. The whole tongue was manually delineated on the planning CT scan of each patient and ported to the dose map of his treatment plan. Besides the whole tongue volume, its surface, as well as the posterior, central and anterior regions were separately considered for the analysis. From every region, 145 radiomic features were extracted from the dose map with S-IBEX and were fed to a LASSO-penalized logistic regression model to predict the presence/absence of dysgeusia at each follow-up time.

As a results, the AUC of the models that considered the central and anterior 2/3 regions were higher than 0.85, while lower performances were observed at later times, with AUC values below 0.6 for most regions. As far as we know, this is the first study demonstrating the correlation between radiomic features extracted from dose maps in the tongue region and dysgeusia and future large-population prospective trials could further support these findings.

4.4.3 Assessing Aperture Shape Controller (ASC) and Monitor Units limit (MU) impact on dose maps (lung, prostate, and head & neck sites)

In radiotherapy, between treatment planning and patient treatment, the plan undergoes quality assurance (QA) controls on a physical phantom, to verify that the machine-delivered plan matches with the calculated one. If the plan does not pass the QA, it must be re-planned, requiring additional time costs. The discrepancies between the calculated and delivered plan might arises because of physical movement constrains of the treatment machine (i.e., the linear accelerator) and/or because of the accuracy of the computed plan. The Aperture Shape Controller (ASC) and Monitor Units limit (MU) are two parameters that allow the planner to reduce the modulation of the calculated plan, producing simpler plans that have a higher probability of passing the QA.

The primary aim of this investigation was to assess if changes in the treatment plan due to ASC and MU constrains were captured by radiomic features, so that, in case of a positive confirmation, prediction models of QA results could be built. The cohort comprehended 30 patients, equally distributed across lung, prostate, and head & neck sites, each one replanned with different combinations of ASC and MU. Radiomic features were extracted using S-IBEX on several ROIs (i.e., low-dose and high-dose areas, target planning volume).

Results of this study did not show any significant differences (Bonferroni-corrected Wilcoxon signed-rank test) between radiomic features computed on dose maps with/without the application of ASC and MU constrains, suggesting that radiomics might not be able to capture changes in the dose maps due to a plan complexity reduction.

4.4.4 Sinograms of TomoTherapy Plans: Patient-specific quality assurance

The aim of this study was to predict the QA results of Helical TomoTherapy plans. Helical TomoTherapy is a technology which allows to deliver intensity-modulated radiotherapy plans through a fan-beam whose modulation at each delivery angle can be summarized in a sinogram.

For the study, 881 plans were collected at Veneto Institute of Oncology. S-IBEX was employed to extract 174 radiomics features from plan sinograms and integrated with other 65 indicators (e.g., typical delivery parameters and plan-complexity metrics). XGBoost regression models were trained using a 50-repeated hold-out validation scheme (4:1) and performances were assessed in terms of AUC.

The integration of delivery parameters and complexity metrics with sinogram-derived radiomics features allowed for a robust and reliable prediction of QA passing rate with a 100% specificity detection. To the best of our knowledge, this is the first investigation on the development of a prediction models for the QA of helical TomoTherapy plans, which, translated into a real clinical scenario, would decrease the QA workload by approximately 35%.

Chapter 5: Conclusions and future developments

5.1 Summary of the main thesis' achievements

The aim of this thesis' project was threefold: to 1) develop and standardize a software tool for radiomic feature extraction following the guidelines of an internationally recognised initiative, the IBSI, 2) compare the tool to other popular software programs available in the literature and 3) employ the developed tool for clinical research within the collaboration between Padua University and Veneto Institute of Oncology.

The program, S-IBEX, was firstly standardized to achieve the IBSI-1 compliance and was validated using the benchmarking and phantoms provided by the initiative. Thanks to this effort, I joined the Initiative itself, for its second chapter, the standardization of convolutional filtering. The enrolment has meant a paradigm shift: from a guideline's user to an active player that brought his own contribution for the creation of the IBSI-2 standard. This work comprehended the implementation and integration in S-IBEX of 10 convolutional filters. The response maps obtained by filtering 4 IBSI benchmark phantoms with 34 different configurations were provided to the initiative to determine a consensus-base reference.

As for the second aim, I designed a multicentre study that allowed a comprehensive comparison of radiomic software tools, built an ad-hoc digital dataset, the ImSURE phantoms, and employed it to identify software discrepancies that were not reported, yet, and that could affect software agreement in an undesirable way. The ImSURE phantoms constitute a comprehensive but compact dataset that, together with a systematic feature extraction, eases software comparison, and proposes itself as an additional benchmark reference tool in the radiomic field.

Eventually, I designed and performed the analysis for several clinical studies based on data acquired locally both at Veneto Institute of Oncology and Padua University Hospital. S-IBEX was the feature-extraction tool of choice for these investigations, proving itself as a versatile and complete tool.

Three major investigations are reported in this thesis with greater details, which suggest promising preliminary results for the application of radiomics for: 1) the prediction of the 5-year overall survival

in locally-advanced breast cancer, 2) the prediction of biochemical recurrence in the setting of prostate cancer and 3) the non-invasive assessment of vascular invasion for hepatocellular carcinoma.

5.2 Data Records and Repositories

In accordance with the FAIR principles[167] of the Findability, Accessibility, Interoperability, and Reuse of digital assets, I published S-IBEX and other developed tools (ImSURE phantoms) of this thesis' work.

S-IBEX source code was uploaded to GitHub (https://github.com/abettinelli/SIBEX_Source) and made publicly available for free download. The ImSURE phantoms were uploaded to the public "*ImSURE Phantoms*" repository on Figshare[83], which contains:

I. The ImSURE isotropic and anisotropic phantoms. To ease their usage with different radiomic tools, phantoms are available in DICOM, NIfTI, NRRD and MATLAB formats.

II. A document reporting the parameter settings necessary to reproduce the proposed feature extraction on the ImSURE phantoms.

The MATLAB code used to design the two phantoms presented in this work is available in the Figshare "*Phantom-Creator*" repository[82]. The MATLAB version 2020a and the Image Processing Toolbox are required to run the code.

5.3 Future developments

One of the most crucial upcoming developments is the finalization of the IBSI chapter 2 for the standardization of convolutional filtering. Once done, future investigations could be done to determine the most appropriate image filtering method and parameter configurations for different image modalities and cancer settings. The same studies would also be relevant for IBSI-1-related aspects, where still optimal pre-processing configurations and impact of different extraction settings remain unclear (e.g., up-sampling versus down-sampling, feature-family-specific optimal discretization levels). The creation and adoption of these additional guidelines would further ease the comparison of radiomic studies, even more if backed by recommendations of the best practise for radiomic model building.

In Chapter 3, I explored the hypothesis that implementation discrepancies might still exist across IBSI-1 compliant software and might play a role in radiomic reproducibility and found, using the proposed ImSURE dataset, that this was the case. I believe it will be of interest to verify in the future if radiomic software, which is constantly evolving and upgraded following the latest guidelines, will reach higher levels of agreement. Similarly, an assessment of convolutional filtering implementation standardization, as per IBSI-2, would be of interest, and could also be performed on the ImSURE phantoms, in both terms of response maps and feature values.

As discussed in Chapter 4, S-IBEX software was employed for clinical studies, where several radiomic models were evaluated for the breast, prostate, and liver cancer. Nevertheless, due to the small casuistry of the datasets, it was not possible to propose a single model nor to externally validate the findings. Nevertheless, those preliminary results could be considered as a starting point for future radiomic studies, where larger multi-institutional datasets will allow to directly identify any potential imaging biomarkers. Surely S-IBEX will hold a great potential, offering standardised implementations of both radiomic features and filter methods.

APPENDIX A: IBSI

Table A.1. The six IBSI configurations are here reported alongside their parameters for both the IBSI digital and radiomic phantoms.

	config. Zero	***config. A***	***config. B***	***config. C***	***config. D***	***config. E***
IBSI phantom	*digital*	*radiomic*	*radiomic*	*radiomic*	*radiomic*	*radiomic*
ROI name	*'ROI'*	*'GTV-1'*	*'GTV-1'*	*'GTV-1'*	*'GTV-1'*	*'GTV-1'*
Approach	*2D /3D*	*2D*	*2D*	*3D*	*3D*	*3D*
Interpolation	*no*	*no*	*yes*	*yes*	*yes*	*yes*
Voxel dims (mm)	-	-	*2x2x2*	*2x2x2*	*2x2x2*	*2x2x2*
Interpolation method	-	-	*bilinear*	*trilinear*	*trilinear*	*tricubic spline*
Gray Level rounding	-	-	*nearest integer*	*nearest integer*	*nearest integer*	*nearest integer*
ROI interp. method	-	-	*bilinear*	*trilinear*	*trilinear*	*trilinear*
ROI partial volume	-	-	*0.5*	*0.5*	*0.5*	*0.5*
Re-segmentation	*no*	*yes*	*yes*	*yes*	*yes*	*yes*
Range (HU)	-	*[-500, 400]*	*[-500, 400]*	*[-500, 400]*	*[-1000, 400]*	*[-500, 400]*
Outliers	-	*no*	*no*	*no*	3σ	3σ
Discretization	*no*	*yes*	*yes*	*yes*	*yes*	*yes*
Texture and IH	-	*FBS: 25 HU*	*FBN: 32 bins*	*FBS: 25 HU*	*FBN: 32 bins*	*FBN: 32 bins*
IVH	-	*no*	*no*	*FBS: 2.5 HU*	*no*	*FBN: 1000 bins*

Table A.2. Filtering configurations of IBSI-2, phase 1.

ID	***Filter***	***Phantom***	***Filter parameters***
1.a.1	Mean	Checkerboard	• 3D filter, support M = 15, zero padding
1.a.2			• 3D filter, support M = 15, nearest value padding
1.a.3			• 3D filter, support M = 15, periodic padding
1.a.4			• 3D filter, support M = 15, mirror padding
1.b.1		Impulse	• 2D filter, support M = 15, zero padding
2.a	LoG	Impulse	• zero padding • 3D filter, scale σ^* = 3.0 mm, filter size cut-off $4\sigma^*$
2.b		Checkerboard	• mirror padding • 3D filter, scale σ^* = 5.0 mm, filter size cut-off $4\sigma^*$
2.c			• mirror padding • 2D filter, scale σ^* = 5.0 mm, filter size cut-off $4\sigma^*$
3.a.1	Laws	Impulse	• zero padding • 3D filter, E5L5S5 response map
3.a.2			• zero padding • 3D filter, E5L5S5 response map • 3D rotation invariance, max pooling
3.a.3			• zero padding • 3D filter, E5L5S5 response map • 3D rotation invariance, max pooling • energy map, distance δ = 7 voxels
3.b.1		Checkerboard	• mirror padding • 3D filter, E3W5R5 response map
3.b.2			• mirror padding • 3D filter, E3W5R5 response map • 3D rotation invariance, max pooling
3.b.3			• mirror padding • 3D filter, E3W5R5 response map • 3D rotation invariance, max pooling • energy map, distance δ = 7 voxels
3.c.1		Checkerboard	• mirror padding • 2D filter, L5S5 response map
3.c.2			• mirror padding • 2D filter, L5S5 response map • 2D rotation invariance, max pooling
3.c.3			• mirror padding • 2D filter, L5S5 response map • 2D rotation invariance, max pooling • energy map, distance δ = 7 voxels
4.a.1	Gabor	Impulse	• zero padding • 2D modulus response map • σ^* = 10.0 mm, λ^* = 4 mm, $\gamma = 1/2$ • in-plane orientation $\theta = \pi/3$

4.a.2			• zero padding • 2D modulus response map • $\sigma^* = 10.0$ mm, $\lambda^* = 4$ mm, $\gamma = 1/2$ • 2D rotation invariance, $\Delta\theta = \pi/4$, average pooling • average 2D responses over orthogonal planes
4.b.1		Sphere	• mirror padding • 2Dmodulus response map • $\sigma^* = 20.0$ mm, $\lambda^* = 8$ mm, $\gamma = 5/2$ • in-plane orientation $\theta = 5\pi/4$
4.b.2			• mirror padding • 2Dmodulus response map • $\sigma^* = 20.0$ mm, $\lambda^* = 8$ mm, $\gamma = 5/2$ • 2Drotation invariance, $\Delta\theta = \pi/8$, average pooling • average 2D responses over orthogonal planes
5.a.1	Daubechies 2	Impulse	• zero padding • 3D filter, undecimated LHL map - 1st level
5.a.2			• zero padding • 3D filter, undecimated LHL map - 1st level • 3D rotation invariance, average pooling
6.a.1	Coifflet 1	Sphere	• periodic padding • 3D filter, undecimated HHL map - 1st level
6.a.2			• periodic padding • 3D filter, undecimated HHL map - 1st level • 3D rotation invariance, average pooling
7.a.1	Haar	Checkerboard	• mirror padding • 3D filter, undecimated LLL map - 2nd level • 3D rotation invariance, average pooling
7.a.2			• mirror padding • 3D filter, undecimated HHH map - 2nd level • 3D rotation invariance, average pooling
8.a.1	Simoncelli	Checkerboard	• periodic padding • 3D filter, B map - 1st level
8.a.2			• periodic padding • 3D filter, B map - 2nd level
8.a.3			• periodic padding • 3D filter, B map - 3rd level
9.a	Riesz-transformed LoG	Impulse	• zero padding • 3D filter, scale $\sigma = 3.0$ mm, filter size cut-off 4σ • $l = (1; 0; 0)$
9.b.1		Sphere	• zero padding • 3D filter, scale $\sigma = 3.0$ mm, filter size cut-off 4σ • $l = (0; 2; 0)$
9.b.2			• zero padding • scale $\sigma = 3.0$ mm, filter size cut-off 4σ • 3D filter, $l = (0; 2; 0)$ • aligned by structure tensor, $\sigma_{\text{tensor}} = 1$mm

10.a	Riesz-transformed Simoncelli	Impulse	• zero padding • 3D filter, B map - 1st level • l = (1; 0; 0)
10.b.1		Pattern 1	• nearest value padding • 3D filter, B map - 1st level • l = (0; 2; 0)
10.b.2			• nearest value padding • 3D filter, B map - 1st level • l = (0; 2; 0) • aligned by structure tensor, $\sigma_{\text{tensor}} = 1$mm

Table A.3. Filter configurations and filter parameters on the IBSI radiomic phantom.

ID	**Filter**	**A**	**B**	**Filter parameters**
1.A / 1.B	none	x	x	-
2.A / 2.B	Mean	x		• 2D filter, support M = 5 voxels
			x	• 3D filter, support M = 5 voxels
3.A / 3.B	LoG	x		• 2D filter, scale $\sigma^* = 1.5$ mm, filter size cut-off $4\sigma^*$
			x	• 3D filter, scale $\sigma^* = 1.5$ mm, filter size cut-off $4\sigma^*$
4.A / 4.B	Laws	x		• 2D filter, L5E5 energy map, distance $\delta = 7$ voxels • 2D rotation invariance, max pooling
			x	• 3D filter, L5E5E5 response map • 3D rotation invariance, max pooling • energy map, distance $\delta = 7$ voxels
5.A / 5.B	Gabor	x		• 2D modulus response map • $\sigma^* = 5$ mm, $\lambda^* = 2$ mm, $\gamma = 3/2$ • 2D rotation invariance, $\Delta\theta = \pi/8$, average pooling
			x	• 2D modulus response map • $\sigma^* = 5$ mm, $\lambda^* = 2$ mm, $\gamma = 3/2$ • 2D rotation invariance, $\Delta\theta = \pi/8$, average pooling • average 2D responses over orthogonal planes
6.A / 6.B	Daubechies 3	x		• 2D filter, undecimated LH map - 1st level • 2D rotation invariance, average pooling
			x	• 3D filter, undecimated LLH map - 1st level • 3D rotation invariance, average pooling
7.A / 7.B	Daubechies 3	x		• 2D filter, undecimated HH map - 2nd level • 2D rotation invariance, average pooling
			x	• 3D filter, undecimated HHH map - 2nd level • 3D rotation invariance, average pooling
8.A / 8.B	Simoncelli	x		• 2D filter, B map - 1st level
			x	• 3D filter, B map - 1st level
9.A / 9.B	Simoncelli	x		• 2D filter, B map - 2nd level
			x	• 3D filter, B map - 2nd level

10.A / 10.B	Riesz-transformed Simoncelli	x		• 2D filter, B map - 1st level • l = (0; 2)
			x	• 3D filter, B map - 1st level • l = (0; 2; 0)
11.A / 11.B	Riesz-transformed Simoncelli	x		• 2D filter, B map - 1st level • l = (0; 2) • aligned by structure tensor, σ_{tensor} = 1mm
			x	• 3D filter, B map - 1st level • l = (0; 2; 0) • aligned by structure tensor, σ_{tensor} = 1mm

Table A.4. Filter parameter used for the validation phase.

ID	**Filter**	**Filter parameters**
1	none	-
2	Mean	• 3D filter, support M = 3 voxels
3	LoG	• 3D filter, scale σ^* = 3.0 mm, filter size cut-off $4\sigma^*$
4	Laws	• 3D filter, S5E5L5 response map • 3D rotation invariance, max pooling • energy map, distance δ = 5 voxels
5	Gabor	• 2D modulus response map • σ^* = 3 mm, λ^* = 3 mm, γ = 1, $\theta = -5\pi/8$
6	Coifflet 3	• 3D filter, undecimated LHH map - 1st level • 3D rotation invariance, average pooling
7	Coifflet 3	• 3D filter, undecimated HHH map - 2nd level • 3D rotation invariance, max pooling
8	Simoncelli	• 3D filter, B map - 1st level
9	Simoncelli	• 3D filter, B map - 2nd level

APPENDIX B: S-IBEX

Table B.1. S-IBEX features. Retained and modified IBEX features and ex-novo implemented features are listed grouped by category.

MF	
Retained:	none
Modified:	Volume, surface area, compactness 1, compactness 2, spherical disproportion, sphericity, maximum 3d diameter
New:	Approximate volume, surface to volume ratio, asphericity, centre of mass shift, major axis length, minor axis length, least axis length, elongation, flatness, volume density (AABB), area density (AABB), volume density (AEE), area density (AEE), volume density (CH), area density (CH), integrated intensity, Moran's I index, Geary's C measure
LI	
Retained:	none
Modified:	none
New:	Local intensity peak, global intensity peak
IS	
Retained:	Mean, skewness, median, minimum gray level, maximum gray level, interquartile range, range, mean absolute deviation, energy, root mean square
Modified:	Variance, kurtosis, 10^{th} percentile, 90^{th} percentile, median absolute deviation
New:	Robust mean absolute deviation, coefficient of variation, quartile coefficient of dispersion
IH	
Retained:	Intensity histogram skewness, intensity histogram interquartile range, intensity histogram range, intensity histogram mean absolute deviation
Modified:	Intensity histogram kurtosis, intensity histogram 10^{th} percentile, intensity histogram 90^{th} percentile, intensity histogram median absolute deviation
New:	Intensity histogram mean, intensity histogram variance, intensity histogram median, intensity histogram minimum gray level, intensity histogram maximum gray level, intensity histogram mode, intensity histogram robust mean absolute deviation, intensity histogram coefficient of variation, intensity histogram quartile coefficient of dispersion, intensity histogram entropy, intensity histogram uniformity, maximum histogram gradient, maximum histogram gradient gray level, minimum histogram gradient, minimum histogram gradient gray level
IVH	

Retained: none

Modified: none

New: Volume intensity fraction at 10, volume intensity fraction at 90, intensity volume fraction at 10, intensity volume fraction at 90, volume at intensity fraction difference, intensity at volume fraction difference, area under IVH curve

GLCM

Retained: Joint maximum, joint entropy, difference entropy, sum average, sum entropy, angular second moment, contrast, dissimilarity, inverse difference, inverse difference normalized, inverse difference moment, inverse difference moment normalized, inverse variance, correlation, autocorrelation, cluster tendency, cluster shade, cluster prominence, measure of information correlation 1, measure of information correlation 2

Modified: Joint variance, sum variance

New: Joint average, difference average, difference variance

GLRLM

Retained: Short run emphasis, long run emphasis, low gray level run emphasis, high gray level run emphasis, short run low gray level emphasis, short run high gray level emphasis, long run low gray level emphasis, long run high gray level emphasis, run percentage

Modified: Gray level non-uniformity, run length non-uniformity

New: Gray level non-uniformity normalized, run length non-uniformity normalized, gray level variance, run length variance, run entropy

GLSZM

Retained: none

Modified: none

New: Small zone emphasis, large zone emphasis, low gray level zone emphasis, high gray level zone emphasis, small zone low gray level emphasis, small zone high gray level emphasis, large zone low gray level emphasis, large zone high gray level emphasis, gray level non-uniformity, gray level non-uniformity normalized, zone size non-uniformity, zone size non-uniformity normalized, zone percentage, gray level variance, zone size variance, zone size entropy

GLDZM

Retained: none

Modified: none

New: Small distance emphasis, large distance emphasis, low gray level zone emphasis, high gray level zone emphasis, small distance low gray level emphasis, small distance high gray level emphasis, large distance low gray level emphasis, large distance high gray level emphasis, gray level non-uniformity, gray level non-uniformity normalized, zone distance non-uniformity, zone distance non-uniformity normalized, zone percentage, gray level variance, zone distance variance, zone distance entropy

NGTDM	
Retained:	Coarseness, busyness, complexity
Modified:	Contrast, strength
New:	none
NGLDM	
Retained:	none
Modified:	none
New:	Low dependence emphasis, high dependence emphasis, low gray level count emphasis, high gray level count emphasis, low dependence low gray level emphasis, low dependence high gray level emphasis, high dependence low gray level emphasis, high dependence high gray level emphasis, gray level non-uniformity, gray level non-uniformity normalized, dependence count non-uniformity, dependence count non-uniformity normalized, dependence count percentage, gray level variance, dependence count variance, dependence count entropy, dependence count energy

APPENDIX C: Clinical studies

Table C.1. Parameters for the image pre-processing steps.

Feature Extraction	***LABC***	***PROSTATE***	***HCC***	***HCC (filtered)***
Software	S-IBEX	S-IBEX	S-IBEX	S-IBEX
Interpolation	yes	yes	yes	yes
Voxel dimensions	2.03x2.03x2 mm	2x2x2 mm	1x1x1 mm	1x1x1 mm
Image interp. method	2D interpolation	3D interpolation	3D interpolation	3D interpolation
ROI interp. method	2D interpolation	3D interpolation	3D interpolation	3D interpolation
Threshold	0.5	0.5	0.5	0.5
Re-segmentation	yes	yes	yes	no
Range	[0, 25]	[0, 20]	[-400, 400]	-
Convolutional filtering	no	no	no	yes
Method	-	-	-	Laplacian of Gaussian
Parameters	-	-	-	σ^* =(0.5, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40)
Discretization	yes	yes	yes	yes
Method	Fixed Bin Size	Fixed Bin Size	Fixed Bin Size	Fixed Bin Number
Bin width/Bin number	0.4 SUV	0.2, 0.4, 0.6 SUV	25 HU	32

Table C.2. Parameters for the extraction of textural features.

Parameter name	***LABC***	***PROSTATE***	***HCC***	***HCC (filtered)***
GLCM				
Aggregation methods	3D:mrg	3D:avg	3D:mrg	3D:mrg
Direction vectors	13 directions	13 directions	13 directions	13 directions
Symmetry	yes	yes	yes	yes

Distance	1 (Chebyshev)	1 (Chebyshev)	1 (Chebyshev)	1 (Chebyshev)
Distance weighting	1 (default)	1 (default)	1 (default)	1 (default)
GLRLM				
Aggregation methods	3D:mrg	3D:avg	3D:mrg	3D:mrg
Direction vectors	13 directions	13 directions	13 directions	13 directions
Distance weighting	1 (default)	1 (default)	1 (default)	1 (default)
GLSZM				
Aggregation methods	3D:mrg	3D	3D:mrg	3D:mrg
Linkage distance	1 (Chebyshev)	1 (Chebyshev)	1 (Chebyshev)	1 (Chebyshev)
GLDZM				
Aggregation methods	3D:mrg	3D	3D:mrg	3D:mrg
Linkage distance	1 (Chebyshev)	1 (Chebyshev)	1 (Chebyshev)	1 (Chebyshev)
Zone distance norm	default (Manhattan)	default (Manhattan)	default (Manhattan)	default (Manhattan)
NGTDM				
Aggregation methods	3D:mrg	3D	3D:mrg	3D:mrg
Distance	1 (Chebyshev)	1 (Chebyshev)	1 (Chebyshev)	1 (Chebyshev)
Distance weighting	1 (default)	1 (default)	1 (default)	1 (default)
NGLDM				
Aggregation methods	3D:mrg	3D	3D:mrg	3D:mrg
Coarseness	0	0	0	0
Distance	1 (Chebyshev)	1 (Chebyshev)	1 (Chebyshev)	1 (Chebyshev)
Distance weighting	1 (default)	1 (default)	1 (default)	1 (default)

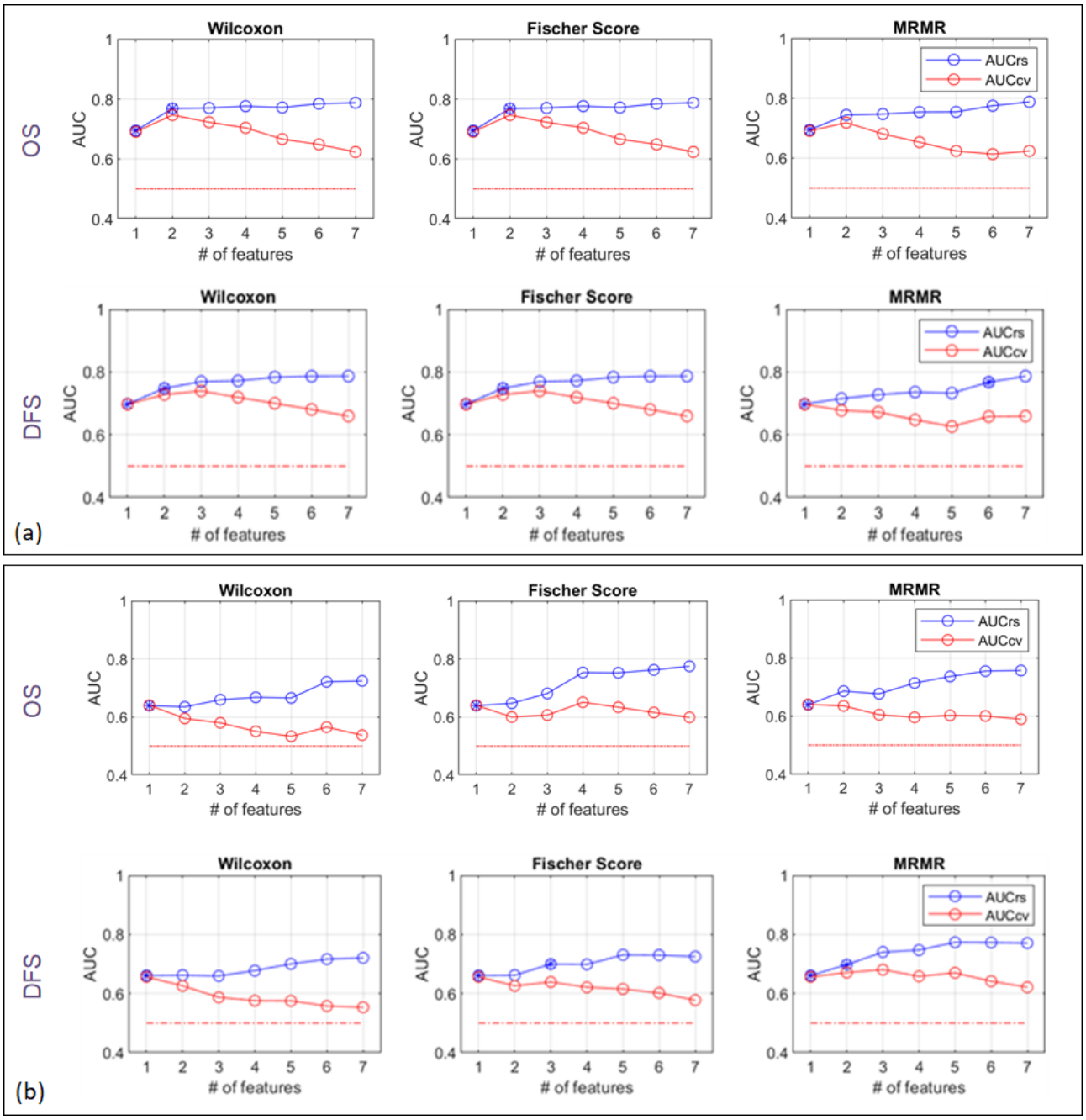

Figure C.1. AUC_{RS} and AUC_{CV} of (a) clinical and (b) radiomic models for the prediction of 5Y-OS and DFS.

Table C.3. Models including age variable.

Model	AUC_{RS}	AUC_{CV}	*Features*	*OR*	*95% CI*	*p-values*
OS: radiomic & clinical & age (Wilcoxon-Fisher)	0.83	0.77	Lymph nodes	0.10	0.019-0.50	0.0053
			ERpos	7.70	1.71-34.60	0.0078
			IVH_AreaUnderIVHCurve	0.43	0.21-0.91	0.0263

			Age	0.63	0.17-2.38	0.5034
OS: radiomic & clinical & age (mRMR)	0.75	0.71	Lymph nodes	0.19	0.05-0.79	0.0216
			IVH_AreaUnderIVHCurve	0.59	0.31-1.12	0.1037
			Age	0.70	0.21-2.35	0.5670
DFS: radiomic & clinical & age (Wilcoxon-Fisher)	0.80	0.75	Lymph nodes	0.15	0.041-0.52	0.0030
			Ki67pos	0.17	0.03-0.81	0.0261
			IH_QuartileCoefficientOfDispersion	1.80	0.99-3.27	0.0521
			Age	0.96	0.3-3.12	0.9512
DFS: radiomic & clinical & age (mRMR)	0.76	0.70	Lymph nodes	0.15	0.04-0.52	0.0030
			IH_QuartileCoefficientOfDispersion	0.17	0.03-0.81	0.0261
			IVH_VolumeIntFract_90	1.80	0.99-3.27	0.0521
			Age	0.96	0.30-3.12	0.9512

Table C.4. Medians [5th - 95th percentile] of the prediction results on the 30 test set folds for each segmentation/bin size feature set considering only radiomic features (AUC = area under the ROC curve; ACC = balanced accuracy; SPEC = specificity; SENS = sensitivity; FBS = Fixed Bin Size; PG = Prostate Gland).

		AUC	***ACC***	***SPEC***	***SENS***
Clinical data only		0,64 [0.39 - 0.86]	0,62 [0.54 - 0.81]	0,77 [0.27 - 1]	0,65 [0.15 - 0.85]
PG_{whole}	***FBS 0.2***	0,56 [0.42 - 0.9]	0,6 [0.5 - 0.87]	0,85 [0.29 - 1]	0,46 [0.08 - 0.85]
	FBS 0.4	0,59 [0.4 - 0.85]	0,62 [0.52 - 0.79]	0,85 [0.34 - 1]	0,42 [0.11 - 0.85]
	FBS 0.6	0,62 [0.41 - 0.87]	0,65 [0.54 - 0.83]	0,92 [0.45 -1]	0,5 [0.08 - 0.88]
$PG_{2.5}$	***FBS 0.2***	0,73 [0.56 - 0.93]	0,71 [0.58 - 0.86]	0,92 [0.67 - 1]	0,58 [0.2 - 0.8]
	FBS 0.4	0,66 [0.4 - 0.83]	0,67 [0.52 - 0.79]	0,92 [0.5 - 1]	0,5 [0.2 - 0.75]
	FBS 0.6	0,63 [0.39 - 0.83]	0,63 [0.52 - 0.77]	0,88 [0.27 - 1]	0,54 [0.07 - 0.92]
$PG_{41\%}$	***FBS 0.2***	0,59 [0.34 - 0.84]	0,65 [0.52 - 0.83]	0,92 [0.52 - 1]	0,54 [0.08 - 0.81]
	FBS 0.4	0,67 [0.43 - 0.89]	0,69 [0.52 - 0.83]	0,88 [0.38 - 1]	0,54 [0.15 - 0.92]
	FBS 0.6	0,56 [0.42 - 0.9]	0,6 [0.5 - 0.87]	0,85 [0.29 - 1]	0,46 [0.08 - 0.85]

BIBLIOGRAPHY

1. Bercovich, E. & Javitt, M. C. Medical Imaging: From Roentgen to the Digital Revolution, and Beyond. *Rambam Maimonides Med. J.* **9**, e0034 (2018).

2. Fass, L. Imaging and cancer: A review. *Mol. Oncol.* **2**, 115–152 (2008).

3. Lambin, P. *et al.* Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48**, 441–446 (2012).

4. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, (2014).

5. Lee, G. *et al.* Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: State of the art. *European Journal of Radiology* **86**, 297–307 (2017).

6. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).

7. Zhang, X. D. Precision Medicine, Personalized Medicine, Omics and Big Data: Concepts and Relationships. *J. Pharmacogenomics Pharmacoproteomics* **06**, 144 (2015).

8. Yip, S. S. F. & Aerts, H. J. W. L. Applications and limitations of radiomics. *Phys. Med. Biol.* **61**, R150 (2016).

9. Welch, M. L. *et al.* Vulnerabilities of radiomic signature development: The need for safeguards. *Radiother. Oncol.* **130**, 2–9 (2019).

10. Timmeren, J. E. van *et al.* Challenges and caveats of a multi-center retrospective radiomics study: an example of early treatment response assessment for NSCLC patients using FDG-PET/CT radiomics. *PLoS One* **14**, e0217536 (2019).

11. Avanzo, M., Stancanello, J. & El Naqa, I. Beyond imaging: The promise of radiomics. *Phys. Medica* **38**, 122–139 (2017).

12. Buvat, I. & Orlhac, F. The dark side of radiomics: On the paramount importance of publishing negative results. *J. Nucl. Med.* **60**, 1543–1544 (2019).

13. Zwanenburg, A. *et al.* The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* **295**, 328–338 (2020).

14. Lambin, P. *et al.* Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* **14**, 749–762 (2017).

15. IBSI Chapter 1 – Image Biomarker Standardisation Initiative. Available at: https://theibsi.github.io/ibsi1/. (Accessed: 30th September 2022)

16. IBSI Chapter 2 – Image Filtering. Available at: https://theibsi.github.io/ibsi2/. (Accessed: 30th September 2022)

17. Traverso, A., Wee, L., Dekker, A. & Gillies, R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *Int. J. Radiat. Oncol. Biol. Phys.* **102**, 1143–1158 (2018).

18. Plautz, T. E., Zheng, C., Noid, G. & Li, X. A. Time stability of delta-radiomics features and the impact on patient analysis in longitudinal CT images. *Med. Phys.* **46**, 1663–1676 (2019).

19. Yang, F., Dogan, N., Stoyanova, R. & Ford, J. C. Evaluation of radiomic texture feature error due to MRI acquisition and reconstruction: A simulation study utilizing ground truth. *Phys. Medica* **50**, 26–36 (2018).

20. Boellaard, R. *et al.* FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *European Journal of Nuclear Medicine and Molecular Imaging* **42**, 328–354 (2015).

21. Bianchini, L. *et al.* PETER PHAN: An MRI phantom for the optimisation of radiomic studies of the female pelvis. *Phys. Medica* **71**, 71–81 (2020).

22. Fendler, W. P. *et al.* 68Ga-PSMA PET/CT: Joint EANM and SNMMI procedure guideline for prostate cancer imaging: version 1.0. *Eur J Nucl Med Mol Imaging* **44**, 1014–1024 (2017).

23. Zhao, B. *et al.* Evaluating Variability in Tumor Measurements from Same-day Repeat CT Scans of Patients with Non–Small Cell Lung Cancer. *Radiology* **252**, 263–272 (2009).

24. Pfaehler, E. *et al.* Repeatability of 18 F-FDG PET radiomic features: A phantom study to explore sensitivity to image reconstruction settings, noise, and delineation method. *Med. Phys.* **46**, 665–678 (2019).

25. van Velden, F. H. P. *et al.* Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [18F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation. *Mol. Imaging Biol.* **18**, 788–795 (2016).

26. Granzier, R. W. Y. *et al.* Test–Retest Data for the Assessment of Breast MRI Radiomic Feature Repeatability. *J. Magn. Reson. Imaging* 1–13 (2021). doi:10.1002/jmri.28027

27. O'Connor, J. P. B. *et al.* Imaging biomarker roadmap for cancer studies. *Nat. Rev. Clin. Oncol. 2016 143* **14**, 169–186 (2016).

28. Avanzo, M., Stancanello, J., Pirrone, G. & Sartor, G. Radiomics and deep learning in lung cancer. *Strahlentherapie und Onkol.* **196**, 879–887 (2020).

29. Hawkins, S. H. *et al.* Predicting Outcomes of Nonsmall Cell Lung Cancer Using CT Image Features. doi:10.1109/ACCESS.2014.2373335

30. Shi, L. *et al.* Radiomics for response and outcome assessment for non-small cell lung cancer. *Technol. Cancer Res. Treat.* **17**, 1–14 (2018).

31. Court, L. E. *et al.* Computational resources for radiomics. *Transl. Cancer Res.* **5**, 340–348 (2016).

32. Mackin, D. *et al.* Measuring computed tomography scanner variability of radiomics features. *Invest. Radiol.* **50**, 757–765 (2015).

33. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1996).

34. Parmar, C., Grossmann, P., Bussink, J., Lambin, P. & Aerts, H. J. W. L. Machine Learning methods for Quantitative Radiomic Biomarkers. *Sci. Rep.* **5**, 13087 (2015).

35. Braghetto, A., Marturano, F., Paiusco, M., Baiesi, M. & Bettinelli, A. Radiomics and deep learning methods for the prediction of 2-year overall survival in LUNG1 dataset. *Sci. Rep.* **12**, 14132 (2022).

36. Moons, K. G. M. *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Ann. Intern. Med.* **162**, W1 (2015).

37. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).

38. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities (Soft-tissue-Sarcoma) - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki. Available at: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=21266533. (Accessed: 10th September 2022)

39. Schirra, S. How reliable are practical point-in-polygon strategies? *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **5193 LNCS**, 744–755 (2008).

40. Shafiq-Ul-Hassan, M. *et al.* Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med. Phys.* **44**, 1050–1062 (2017).

41. Image Biomarker Standardisation Initiative: Reference Manual (v11). (2019). Available at: https://arxiv.org/pdf/1612.07003v11.pdf. (Accessed: 30th September 2022)

42. Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural Features for Image Classification. *IEEE Trans. Syst. Man Cybern.* **SMC-3**, 610–621 (1973).

43. Galloway, M. M. Texture analysis using gray level run lengths. *Comput. Graph. Image Process.* **4**, 172–179 (2008).

44. Thibault, G., Angulo, J. & Meyer, F. Advanced statistical matrices for texture characterization: Application to DNA chromatin and microtubule network classification. in *Proceedings - International Conference on Image Processing, ICIP* 53–56 (IEEE, 2011). doi:10.1109/ICIP.2011.6116401

45. Amadasun, M. & King, R. Texural Features Corresponding to Texural Properties. *IEEE Trans. Syst. Man Cybern.* **19**, 1264–1274 (1989).

46. Sun, C. & Wee, W. G. Neighboring gray level dependence matrix for texture classification. *Comput. Vision, Graph. Image Process.* **23**, 341–352 (1983).

47. Depeursinge, A. *et al.* Standardised convolutional filtering for Radiomics Image Biomarker

Standardisation Initiative (IBSI): Reference Manual. (2022). Available at: https://arxiv.org/pdf/2006.05470.pdf. (Accessed: 14th November 2022)

48. Zhang, L. *et al.* Ibex: An open infrastructure software platform to facilitate collaborative work in radiomics. *Med. Phys.* **42**, 1341–1353 (2015).

49. Peng, J. *et al.* A radiomics nomogram for preoperative prediction of microvascular invasion risk in hepatitis b virus-related hepatocellular carcinoma. *Diagnostic Interv. Radiol.* **24**, 121–127 (2018).

50. Wong, A. J., Kanwar, A., Mohamed, A. S. & Fuller, C. D. Radiomics in head and neck cancer: from exploration to application. *Transl. Cancer Res.* **5**, 371–382 (2016).

51. Gu, J. *et al.* The Feasibility Study of Megavoltage Computed Tomographic (MVCT) Image for Texture Feature Analysis. *Front. Oncol.* **8**, 586 (2018).

52. Berenguer, R. *et al.* Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. *Radiology* **288**, 407–415 (2018).

53. Ger, R. B. *et al.* Effects of alterations in positron emission tomography imaging parameters on radiomics features. *PLoS One* **14**, (2019).

54. Mahmood, U., Apte, A. P., Deasy, J. O., Schmidtlein, C. R. & Shukla-Dave, A. Investigating the Robustness Neighborhood Gray Tone Difference Matrix and Gray Level Co-occurrence Matrix Radiomic Features on Clinical Computed Tomography Systems Using Anthropomorphic Phantoms. *J. Comput. Assist. Tomogr.* **41**, 995–1001 (2017).

55. Mackin, D. *et al.* Harmonizing the pixel size in retrospective computed tomography radiomics studies. *PLoS One* **12**, e0178524 (2017).

56. Ger, R. B. *et al.* Radiomics features of the primary tumor fail to improve prediction of overall survival in large cohorts of CT- And PET-imaged head and neck cancer patients. *PLoS One* **14**, 1–13 (2019).

57. Bibault, J.-E. E. *et al.* Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer. *Sci. Rep.* **8**, 12611 (2018).

58. Branchini, M. *et al.* Impact of acquisition count statistics reduction and SUV discretization on

PET radiomic features in pediatric 18F-FDG-PET/MRI examinations. *Phys. Medica* **59**, 117–126 (2019).

59. Image Biomarker Standardisation Initiative: Reference Manual (v6). (2018). Available at: https://arxiv.org/pdf/1612.07003v6.pdf. (Accessed: 30th September 2022)

60. Bettinelli, A., Branchini, M., De Monte, F., Scaggion, A. & Paiusco, M. Technical Note: An IBEX adaption toward image biomarker standardization. *Med. Phys.* **47**, 1167–1173 (2020).

61. Darren Engwirda. A fast 'point-in-polygon' test for MATLAB / OCTAVE. (2018). Available at: https://github.com/dengwirda/inpoly. (Accessed: 25th June 2019)

62. Bogowicz, M. *et al.* CT radiomics and PET radiomics: Ready for clinical implementation? *Quarterly Journal of Nuclear Medicine and Molecular Imaging* **63**, 355–370 (2019).

63. Zwanenburg, A. & Löck, S. Why validation of prognostic models matters? *Radiother. Oncol.* **127**, 370–373 (2018).

64. Zwanenburg, A., Leger, S., Vallières, M. & Löck, S. The image biomarker standardisation initiative — IBSI 0.0.1dev documentation. (2019). Available at: https://ibsi.readthedocs.io/en/latest/. (Accessed: 24th November 2021)

65. Zwanenburg, A. GitHub - theibsi/data_sets: Data sets used by the IBSI for benchmarking and standardisation.

66. Ashrafinia, S. Quantitative nuclear medicine imaging using advanced Image reconstruction and radiomics. *Jhon Hopkins University* (2019).

67. van Griethuysen, J. J. M. *et al.* Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* **77**, e104–e107 (2017).

68. Pfaehler, E., Zwanenburg, A., de Jong, J. R. & Boellaard, R. RACAT: An open source and easy to use radiomics calculator tool. *PLoS One* **14**, 1–26 (2019).

69. Zwanenburg, A. *et al.* Assessing robustness of radiomic features by image perturbation. *Sci. Rep.* **9**, 1–31 (2019).

70. PyRadiomics from the Computational Imaging & Bioinformatics Lab - Harvard Medical School. (Online). (2017).

71. Avanzo, M. *et al.* Electron Density and Biologically Effective Dose (BED) Radiomics-Based Machine Learning Models to Predict Late Radiation-Induced Subcutaneous Fibrosis. *Front. Oncol.* **10**, 490 (2020).

72. Genetics, Soph. SOPHiA AI Makes Data-Driven Medicine More Valuable by Combining Genomics and Radiomics to Fight Cancer. (2018).

73. Foy, J. J. *et al.* Variation in algorithm implementation across radiomics software. *J. Med. Imaging* **5**, 1 (2018).

74. Fornacon-Wood, I. *et al.* Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. *Eur. Radiol.* **30**, 6241–6250 (2020).

75. McNitt-Gray, M. *et al.* Standardization in quantitative imaging: A multicenter comparison of radiomic features from different software packages on digital reference objects and patient data sets. *Tomography* **6**, 118–128 (2020).

76. Loi, S. *et al.* Robustness of CT radiomic features against image discretization and interpolation in characterizing pancreatic neuroendocrine neoplasms. *Phys. Medica* **76**, 125–133 (2020).

77. Blender Online Community. Blender - a 3D modelling and rendering package. (2018). Available at: http://www.blender.org. (Accessed: 28th September 2021)

78. Bewick, V., Cheek, L. & Ball, J. Statistics review 10: Further nonparametric methods. *Critical Care* **8**, 196–199 (2004).

79. Lei, M. *et al.* Benchmarking features from different radiomics toolkits / toolboxes using Image Biomarkers Standardization Initiative. *arXiv* (2020).

80. Baeßler, B., Weiss, K. & Santos, D. P. Dos. Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. *Invest. Radiol.* **54**, 221–228 (2019).

81. Liang, Z. G. *et al.* Comparison of radiomics tools for image analyses and clinical prediction in nasopharyngeal carcinoma. *Br. J. Radiol.* **92**, (2019).

82. Bettinelli, A. & Marturano, F. Phantom-Creator. *Figshare* (2022). doi:10.6084/m9.figshare.19362119.v1

83. Bettinelli, A. & Marturano, F. ImSURE Phantoms. *Figshare* (2022). doi:10.6084/m9.figshare.c.5625439.v2

84. Placidi, L. *et al.* A multicentre evaluation of dosiomics features reproducibility, stability and sensitivity. *Cancers (Basel).* **13**, (2021).

85. Senkus, E. *et al.* Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **26**, v8–v30 (2015).

86. Thompson, A. M. & Moulder-Thompson, S. L. Neoadjuvant treatment of breast cancer. *Ann. Oncol.* **23**, x231–x236 (2012).

87. Groheux, D. *et al.* 18F-FDG PET/CT for staging and restaging of breast cancer. *J. Nucl. Med.* **57**, 17S-26S (2016).

88. Cardoso, F. *et al.* 4th ESO-ESMO international consensus guidelines for advanced breast cancer (ABC 4). *Ann. Oncol.* **29**, 1634–1657 (2018).

89. Cachin, F., Prince, H. M., Hogg, A., Ware, R. E. & Hicks, R. J. Powerful prognostic stratification by [18F]fluorodeoxyglucose positron emission tomography in patients with metastatic breast cancer treated with high-dose chemotherapy. *J. Clin. Oncol.* **24**, 3026–3031 (2006).

90. Evangelista, L. *et al.* Could semiquantitative FDG analysis add information to the prognosis in patients with stage II/III breast cancer undergoing neoadjuvant treatment? *Eur. J. Nucl. Med. Mol. Imaging* **42**, 1648–1655 (2015).

91. Ha, S., Park, S., Bang, J.-I., Kim, E.-K. & Lee, H.-Y. Metabolic Radiomics for Pretreatment 18F-FDG PET/CT to Characterize Locally Advanced Breast Cancer: Histopathologic Characteristics, Response to Neoadjuvant Chemotherapy, and Prognosis. *Sci. Rep.* **7**, 1556 (2017).

92. Antunovic, L. *et al.* [18F]FDG PET/CT features for the molecular characterization of primary breast tumors. *Eur. J. Nucl. Med. Mol. Imaging* **44**, 1945–1954 (2017).

93. Najman, L. & Couprie, M. Building the component tree in quasi-linear time. *IEEE Trans. Image Process.* **15**, 3531–3539 (2006).

94. Chicklore, S. *et al.* Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis. *European Journal of Nuclear Medicine and Molecular Imaging* **40**, 133–140 (2013).

95. Leijenaar, R. T. H. *et al.* The effect of SUV discretization in quantitative FDG-PET Radiomics: The need for standardized methodology in tumor texture analysis. *Sci. Rep.* **5**, 11075 (2015).

96. Antunovic, L. *et al.* PET/CT radiomics in breast cancer: promising tool for prediction of pathological response to neoadjuvant chemotherapy. *Eur. J. Nucl. Med. Mol. Imaging* **46**, 1468–1477 (2019).

97. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **52**, 91–118 (2003).

98. Huang, S. *et al.* Exploration of PET and MRI radiomic features for decoding breast cancer phenotypes and prognosis. *npj Breast Cancer* **4**, 24 (2018).

99. Hosny, A. *et al.* Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLOS Med.* **15**, e1002711 (2018).

100. Ou, X. *et al.* Ability of 18 F-FDG PET/CT Radiomic Features to Distinguish Breast Carcinoma from Breast Lymphoma. *Contrast Media Mol. Imaging* **2019**, 1–9 (2019).

101. Soussan, M. *et al.* Relationship between Tumor Heterogeneity Measured on FDG-PET/CT and Pathological Prognostic Factors in Invasive Breast Cancer. *PLoS One* **9**, e94017 (2014).

102. Yoon, H.-J., Kim, Y. & Kim, B. S. Intratumoral metabolic heterogeneity predicts invasive components in breast ductal carcinoma in situ. *Eur. Radiol.* **25**, 3648–3658 (2015).

103. van Velden, F. H. P. *et al.* Evaluation of a cumulative SUV-volume histogram method for parameterizing heterogeneous intratumoural FDG uptake in non-small cell lung cancer PET studies. *Eur. J. Nucl. Med. Mol. Imaging* **38**, 1636–1647 (2011).

104. Baba, S. *et al.* Diagnostic and prognostic value of pretreatment SUV in 18F-FDG/ PET in breast cancer: Comparison with apparent diffusion coefficient from diffusion-weighted MR imaging. *J. Nucl. Med.* **55**, 736–742 (2014).

105. Boughdad, S. *et al.* Influence of age on radiomic features in 18F-FDG PET in normal breast

tissue and in breast cancer tumors. *Oncotarget* **9**, 30855–30868 (2018).

106. Lovinfosse, P., Hatt, M., Visvikis, D. & Hustinx, R. Heterogeneity analysis of 18F-FDG PET imaging in oncology: clinical indications and perspectives. *Clinical and Translational Imaging* **6**, 393–410 (2018).

107. Yoon, H. J., Kim, Y., Chung, J. & Kim, B. S. Predicting neo-adjuvant chemotherapy response and progression-free survival of locally advanced breast cancer using textural features of intratumoral heterogeneity on F-18 FDG PET/CT and diffusion-weighted MR imaging. *Breast J.* **25**, 373–380 (2019).

108. Galavis, P. E., Hollensen, C., Jallow, N., Paliwal, B. & Jeraj, R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncol. (Madr).* (2010). doi:10.3109/0284186X.2010.498437

109. Yan, J. *et al.* Impact of image reconstruction settings on texture features in 18F-FDG PET. *J. Nucl. Med.* (2015). doi:10.2967/jnumed.115.156927

110. Aide, N. *et al.* Implications of reconstruction protocol for histo-biological characterisation of breast cancers using FDG-PET radiomics. *EJNMMI Res.* **8**, 114 (2018).

111. Kaida, H. *et al.* Improved breast cancer detection of prone breast fluorodeoxyglucose-PET in 118 patients. *Nucl. Med. Commun.* **29**, 885–893 (2008).

112. Teixeira, S. C. *et al.* Additional prone 18F-FDG PET/CT acquisition to improve the visualization of the primary tumor and regional lymph node metastases in stage II/III breast cancer. *Clin. Nucl. Med.* **41**, e181–e186 (2016).

113. Williams, J. M. *et al.* Comparison of prone versus supine 18F-FDG-PET of locally advanced breast cancer: Phantom and preliminary clinical studies. *Med. Phys.* **42**, 3801–3813 (2015).

114. Garcia-Vicente, A. M. *et al.* Textural features and SUV-based variables assessed by dual time point 18F-FDG PET/CT in locally advanced breast cancer. *Ann. Nucl. Med.* **31**, 726–735 (2017).

115. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **68**, 394–424 (2018).

116. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.* **71**, 209–249 (2021).

117. Parker, C. *et al.* Prostate cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **31**, 1119–1134 (2020).

118. Serefoglu, E. C. *et al.* How reliable is 12-core prostate biopsy procedure in the detection of prostate cancer? *Can. Urol. Assoc. J.* **7**, E293 (2013).

119. Mottet, N. *et al.* EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *Eur. Urol.* **71**, 618–629 (2017).

120. Freedland, S. J. *et al.* Risk of Prostate Cancer–Specific Mortality Following Biochemical Recurrence After Radical Prostatectomy. *JAMA* **294**, 433 (2005).

121. Roehl, K. A., Han, M., Ramos, C. G., Antenor, J. A. V & Catalona, W. J. Cancer progression and survival rates following anatomical radical retropubic prostatectomy in 3,478 consecutive patients: long-term results. *J. Urol.* **172**, 910–4 (2004).

122. Kupelian, P. A., Mahadevan, A., Reddy, C. A., Reuther, A. M. & Klein, E. A. Use of different definitions of biochemical failure after external beam radiotherapy changes conclusions about relative treatment efficacy for localized prostate cancer. *Urology* **68**, 593–598 (2006).

123. Wallitt, K. L. *et al.* Clinical pet imaging in prostate cancer. *Radiographics* **37**, 1512–1536 (2017).

124. Evangelista, L. *et al.* PET/MRI in prostate cancer: a systematic review and meta-analysis. *Eur. J. Nucl. Med. Mol. Imaging* **48**, 859–873 (2021).

125. Mottet, N. *et al.* EAU-EANM-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer—2020 Update. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *Eur. Urol.* **79**, 243–262 (2021).

126. Cook, G. J. R. *et al.* Radiomics in PET: principles and applications. *Clin. Transl. Imaging* **2**, 269–276 (2014).

127. Bi, W. L. *et al.* Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA. Cancer J. Clin.* **69**, (2019).

128. van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H. & Baessler, B. Radiomics in medical imaging-"how-to" guide and critical reflection. *Insights Imaging* **11**, (2020).

129. Zwanenburg, A. Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. *Eur. J. Nucl. Med. Mol. Imaging* **46**, 2638–2655 (2019).

130. Piñeiro-Fiel, M. *et al.* A Systematic Review of PET Textural Analysis and Radiomics in Cancer. *Diagnostics* **11**, (2021).

131. Guglielmo, P. *et al.* Additional Value of PET Radiomic Features for the Initial Staging of Prostate Cancer: A Systematic Review from the Literature. *Cancers (Basel).* **13**, 6026 (2021).

132. Spohn, S. K. B. *et al.* Radiomics in prostate cancer imaging for a personalized treatment approach - current aspects of methodology and a systematic review on validated studies. *Theranostics* **11**, 8027–8042 (2021).

133. Mottet, N. *et al.* EAU-EANM-ESTRO-ESUR-ISUP-SIOG guidelines on prostate cancer. (2022). Available at: https://uroweb.org/guideline/prostate-cancer/#3.

134. Bettinelli, A. *et al.* A Novel Benchmarking Approach to Assess the Agreement among Radiomic Tools. *Radiology* (2022). doi:10.1148/radiol.211604

135. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).

136. Tu, S. J., Tran, V. T., Teo, J. M., Chong, W. C. & Tseng, J. R. Utility of radiomic zones for risk classification and clinical outcome predictions using supervised machine learning during simultaneous 11C-choline PET/MRI acquisition in prostate cancer patients. *Med. Phys.* **48**, 5192–5201 (2021).

137. Papp, L. *et al.* Supervised machine learning enables non-invasive lesion characterization in primary prostate cancer with [68 Ga]Ga-PSMA-11 PET/MRI. *Eur. J. Nucl. Med. Mol. Imaging* **48**, 1795–1805 (2021).

138. Pizzuto, D. A. *et al.* 18F-Fluoroethylcholine PET/CT Radiomic Analysis for Newly Diagnosed Prostate Cancer Patients: A Monocentric Study. *Int. J. Mol. Sci. 2022, Vol. 23, Page 9120* **23**,

9120 (2022).

139. Alarcón-Zendejas, A. P. *et al.* The promising role of new molecular biomarkers in prostate cancer: from coding and non-coding genes to artificial intelligence approaches. *Prostate Cancer Prostatic Dis.* **25**, 431–443 (2022).

140. Farha, M. W. & Salami, S. S. Biomarkers for prostate cancer detection and risk stratification. *Ther. Adv. Urol.* **14**, 175628722211039 (2022).

141. Laditi, F., Nie, J., Jones, T. & Leapman, M. S. Variation and Disparity in the Use of Prostate Cancer Risk Stratification Tools in the United States. *Eur. Urol. Focus* **8**, 910–912 (2022).

142. Sepulcri, M. *et al.* Value of 18F-fluorocholine PET/CT in predicting response to radical radiotherapy in patients with localized prostate cancer. *Clin. Transl. Radiat. Oncol.* **30**, 71–77 (2021).

143. Cysouw, M. C. F. *et al.* Machine learning-based analysis of [18F]DCFPyL PET radiomics for risk stratification in primary prostate cancer. *Eur. J. Nucl. Med. Mol. Imaging* **48**, 340–349 (2021).

144. Boorjian, S. A. *et al.* Long-term survival after radical prostatectomy versus external-beam radiotherapy for patients with high-risk prostate cancer. *Cancer* **117**, 2883–2891 (2011).

145. Klein, E. A., Ciezki, J., Kupelian, P. A. & Mahadevan, A. Outcomes for intermediate risk prostate cancer: are there advantages for surgery, external radiation, or brachytherapy? *Urol. Oncol.* **27**, 67–71 (2009).

146. Villanueva, A. Hepatocellular Carcinoma. *N. Engl. J. Med.* **380**, 1450–1462 (2019).

147. Bruix, J. & Sherman, M. Management of hepatocellular carcinoma: An update. *Hepatology* **53**, 1020–1022 (2011).

148. Galle, P. R. *et al.* EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma. *J. Hepatol.* **69**, 182–236 (2018).

149. Roayaie, S. *et al.* The role of hepatic resection in the treatment of hepatocellular cancer. *Hepatology* **62**, 440–451 (2015).

150. Gouw, A. S. H. *et al.* Markers for microvascular invasion in hepatocellular carcinoma: Where do we stand? *Liver Transplant.* **17**, S72–S80 (2011).

151. Lim, K.-C. *et al.* Microvascular Invasion Is a Better Predictor of Tumor Recurrence and Overall Survival Following Surgical Resection for Hepatocellular Carcinoma Compared to the Milan Criteria. *Ann. Surg.* **254**, 108–113 (2011).

152. Mazzaferro, V. *et al.* Predicting survival after liver transplantation in patients with hepatocellular carcinoma beyond the Milan criteria: a retrospective, exploratory analysis. *Lancet Oncol.* **10**, 35–43 (2009).

153. Kornberg, A. *et al.* 18F-FDG-Uptake of Hepatocellular Carcinoma on PET Predicts Microvascular Tumor Invasion in Liver Transplant Patients. *Am. J. Transplant.* **9**, 592–600 (2009).

154. Ni, M. *et al.* Radiomics models for diagnosing microvascular invasion in hepatocellular carcinoma: Which model is the best model? *Cancer Imaging* **19**, (2019).

155. Xu, X. *et al.* Radiomic analysis of contrast-enhanced CT predicts microvascular invasion and outcome in hepatocellular carcinoma. *J. Hepatol.* **70**, 1133–1144 (2019).

156. Ma, X. *et al.* Preoperative radiomics nomogram for microvascular invasion prediction in hepatocellular carcinoma using contrast-enhanced CT. *Eur. Radiol.* **29**, 3595–3605 (2019).

157. Bakr, S., Echegaray, S., Shah, R., Kamaya, A. & Louie, J. Noninvasive radiomics signature based on quantitative analysis of computed tomography images as a surrogate for microvascular invasion in hepatocellular carcinoma: a pilot study. *J. Med. Imaging* **4**, 1 (2017).

158. Yang, L. *et al.* A Radiomics Nomogram for Preoperative Prediction of Microvascular Invasion in Hepatocellular Carcinoma. *Liver Cancer* **8**, 373–386 (2019).

159. Jiang, Y.-Q. *et al.* Preoperative identification of microvascular invasion in hepatocellular carcinoma by XGBoost and deep learning. *J. Cancer Res. Clin. Oncol.* **147**, 821–833 (2021).

160. Kim, K. A. *et al.* Prediction of microvascular invasion of hepatocellular carcinoma: Usefulness of peritumoral hypointensity seen on gadoxetate disodium-enhanced hepatobiliary phase images. *J. Magn. Reson. Imaging* **35**, 629–634 (2012).

161. Kuo, M. D., Gollub, J., Sirlin, C. B., Ooi, C. & Chen, X. Radiogenomic Analysis to Identify Imaging Phenotypes Associated with Drug Response Gene Expression Programs in Hepatocellular Carcinoma. *J. Vasc. Interv. Radiol.* **18**, 821–830 (2007).

162. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* **278**, 563–577 (2016).

163. Wakabayashi, T. *et al.* Radiomics in hepatocellular carcinoma: a quantitative review. *Hepatol. Int.* **13**, 546–559 (2019).

164. Efron, B. & Hastie, T. *Computer Age Statistical Inference*. (Cambridge University Press., 2016).

165. Imamura, H. *et al.* Risk factors contributing to early and late phase intrahepatic recurrence of hepatocellular carcinoma after hepatectomy. *J. Hepatol.* **38**, 200–207 (2003).

166. Llovet, J. M. *et al.* EASL–EORTC Clinical Practice Guidelines: Management of hepatocellular carcinoma. *J. Hepatol.* **56**, 908–943 (2012).

167. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, (2016).