Computational Methodologies for Solid Tumor Characterization and Outcome Prediction in Volumetric Medical Images

THIERRY LEFEBVRE

Medical Physics Unit

McGill University, Montreal

August 2020

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science in Medical Radiation Physics

ABSTRACT

Imaging-based quantification and characterization of tumor phenotypes has been the main goal of numerous efforts in recent years for developing and integrating precision oncology in clinical practice. Identifying optimal quantitative image features and machine learning pipelines for computer-aided diagnosis constitute crucial steps towards the development of reproducible, standardized, and clinically relevant imaging biomarkers of cancer phenotypic characteristics. An "image feature" can be understood as an image-derived descriptor of intensity, shape, texture, etc. In radiomics studies, the main hypothesis is that combining many of these quantitative features extracted from tumor regions in medical images can predict underlying genetic or pathological changes occurring in response to disease activity. Given the high variability of processing pipelines in radiomics studies, we first aimed to develop and validate a standardized, IBSI-compliant, and evidence-based processing pipeline for radiomics studies. Second, we aimed to evaluate the diagnostic performance of the well-established robust set of rotationally invariant features from spherical harmonics (SPHARM) decompositions in predicting outcomes from volumetric medical images and compare it to radiomics. Pipelines for these two methods were built and validated on synthetic 3D texture datasets and in two distinct dual-centre diagnostic retrospective studies: i) a study on identifying renal cysts malignancy on contrast-enhanced CT, and ii) a study on identifying histopathological features of endometrial cancer on multi-parametric MRI.

For distinguishing benign from malignant renal cysts, a random forest model based on a set of five most discriminative and reproducible radiomics features resulted in high diagnostic performance (testing area under the receiver operating characteristic curve [AUC] = 0.91). Similarly, for SPHARM decomposition coefficients, a tensor logistic regressor resulted in good diagnostic performance for predicting malignancy of renal cysts (testing AUC = 0.83). For detecting histopathological deep myometrial invasion in endometrial cancer on multi-parametric MRI, a random forest model based on our set of five most discriminative and reproducible radiomics features resulted in good diagnostic performance (testing AUC = 0.81). For SPHARM decomposition coefficients, a tensor logistic regressor resulted in higher diagnostic performance using only dynamic-contrast-enhanced MRI images (testing AUC = 0.86). Furthermore, we

show that in specific situations, approximate spherical tumor segmentations can rival or even outperform painstakingly obtained but accurate tumor segmentations.

Both radiomics features and SPHARM descriptors show promise as reproducible surrogate biomarkers of histopathological features of cancer activity on CT and MRI. Implementing such computational pipelines in clinical practice could improve and accelerate patients' stratification and decision-making for radiologists and radio-oncologists in cancer diagnosis or treatment.

Keywords : Precision oncology; Radiomics; Spherical harmonics; Cancer; Tumor phenotypes; Histopathology; Predictive models; Machine learning; Classification.

RÉSUMÉ

La quantification et la caractérisation des phénotypes tumoraux par l'imagerie médicale sont parmi les principaux objectifs de nombreux efforts de recherche des dernières années pour développer et intégrer l'oncologie de précision dans la pratique clinique. L'identification de descripteurs d'imagerie quantitatifs optimaux et de pipelines d'apprentissage automatique pour le diagnostic assisté par ordinateur constituent des étapes cruciales au développement de biomarqueurs d'imagerie reproductibles, standardisés et cliniquement adaptés aux caractéristiques phénotypiques du cancer. Un « paramètre d'imagerie » peut être défini comme un descripteur d'image dérivé de l'intensité, de la forme, de la texture, etc. Dans les études radiomiques, l'hypothèse principale est que la combinaison de plusieurs de ces paramètres quantitatifs extraits de régions tumorales dans les images médicales cliniques peut prédire des altérations génétiques ou pathologiques sous-jacents se produisant en réponse à l'activité de la maladie. Compte tenu de la grande variabilité des pipelines de pré-traitement d'images dans les études radiomiques, nous avons d'abord cherché à développer et à valider un pipeline standardisé, conforme aux recommandations d'IBSI et fondé sur la littérature la plus récente. Deuxièmement, nous avons cherché à évaluer et à comparer la performance diagnostique des paramètres d'images radiomiques aux ensembles de paramètres robustes, bien établis et invariants sous rotation des décompositions d'harmoniques sphériques (SPHARM) pour prédire des caractéristiques pathologiques à partir d'images médicales volumétriques. Des pipelines pour ces deux méthodes ont donc été développés et validés sur des ensembles de données de textures 3D synthétiques et dans deux études diagnostiques rétrospectives distinctes à deux centres hospitaliers dans le but de i) prédire la malignité de kystes rénaux sur des images de tomodensitométrie avec agent de contraste, et ii) prédire des caractéristiques histopathologiques du cancer de l'endomètre avec une approche d'imagerie par résonance magnétique (IRM) multi-paramétrique. Pour distinguer les kystes rénaux bénins des tumeurs rénales malignes, un modèle de forêt randomisée aléatoire basé sur un ensemble de cinq paramètres radiomiques discriminants et reproductibles a abouti à des performances diagnostiques élevées (aire sous la courbe ROC [AUC] = 0.91). Pour les coefficients de décomposition SPHARM, un modèle de régression logistique tensorielle a donné de bonnes performances diagnostiques (AUC = 0.83). De même, pour détecter l'invasion myopathique profonde prouvée par histologie dans le cancer

de l'endomètre, un modèle de forêt aléatoire basé sur un ensemble de cinq paramètres radiomiques discriminants et reproductibles basées sur toutes les séquences d'IRM a abouti à des performances diagnostiques élevées (AUC = 0.81). Pour les coefficients de décomposition SPHARM, un modèle de régression logistique tensorielle a abouti à des performances diagnostiques plus élevées en utilisant uniquement les images d'IRM à contraste dynamique (AUC = 0.86). De plus, nous montrons que dans des situations spécifiques, des segmentations tumorales sphériques approximatives peuvent rivaliser ou même surpasser les segmentations tumorales obtenues avec précision.

Les caractéristiques radiomiques et les descripteurs SPHARM semblent prometteurs en tant que biomarqueurs reproductibles des caractéristiques histopathologiques de l'activité cancéreuse sur les images de tomodensitométrie et d'IRM. La mise en œuvre de tels pipelines dans la pratique clinique pourrait améliorer et accélérer la prise de décision pour les radiologues et les radio-oncologues dans le diagnostic ou le traitement du cancer.

Mots-clés : Oncologie de précision; Radiomique; Harmoniques sphériques; Cancer; Phénotypes tumoraux; Histopathologie; Models prédictifs; Apprentissage machine; Classification.

ACKNOWLEDGEMENTS

After two years of studies and research at the Medical Physics Unit, I have many people to whom I am grateful. First, I would like to thank my principal supervisor, Dr Peter Savadjiev, for his dedication and guidance throughout every step of this project. The helpful discussions, the troubleshooting sessions, and lines and lines of exchanged emails with Dr Savadjiev have made all the work detailed in this thesis possible. I also want to thank him for his trust and for giving me the opportunity to take part in different research projects with collaborators which have resulted in the many dimensions of my work in this thesis. I am grateful for his great mentorship which has helped me become a better scientist.

This work would not have been possible without clinical collaborators who identified clinical problems which we tried to address with computational solutions. I want to thank my co-supervisor, Dr Caroline Reinhold, for welcoming and trusting me to work on two clinical research projects which she instigated at the McGill University Health Centre and who provided invaluable clinical expertise. I am also grateful to Dr Jeremy Dana and Dr Yoshiko Ueno who gathered all data and performed segmentations on medical images from hundreds of patients – months of work – to allow us to perform analyses of these tumor regions used in our studies. Their expertise and knowledge of clinical needs are what drive research projects such as the ones we have developed in this thesis. I thank Dr Sahir Bhatnagar who provided support on statistical analyses and Ozan Ciga who instigated the SPHARM decomposition project during his MSc degree under the supervision of Dr Savadjiev. I also thank Margery Knewstubb and Tatjana Nisic who provided administrative (and moral) support during these two years.

On a more personal note, I am grateful to my family and friends who supported me in so many ways, in the highs and through the lows of these two years, including François, Carmen, Christyne, Sandryne, Maude, Caroline, Corinne, and especially Julien. I thank my fellow MSc colleagues in whom I found friendship and well-needed support through first-year classes, research activities, and COVID-19-related changes.

Finally, I acknowledge the financial support from the Natural Sciences and Engineering Research Council (NSERC) of Canada, from the Ministère de la santé et des services sociaux (MSSS) of Quebec, and from McGill University's Faculty of Medicine.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GLOSSARY

ADC	Apparent Diffusion Coefficient
AUC	Area Under the Receiver Operating Characteristic Curve
CART	Classification and Regression Tree
CCRL	Complex Cystic Renal Lesion
CT	Computed Tomography
CE-CT	Contrast-Enhanced Computed Tomography
DCE-MRI	Dynamic Contrast-Enhanced Magnetic Resonance Imaging
DWI	Diffusion-Weighted Imaging
EPI	Echo Planar Imaging
FIGO	International Federation of Gynecology and Obstetrics
GLCM	Gray-Level Co-occurrence Matrix
GLDM	Gray-Level Dependence Matrix
GLRLM	Gray-Level Run-Length Matrix
GLSZM	Gray-Level Size-Zone Matrix
GRE	Gradient-echo
HU	Hounsfield Unit
IBSI	Imaging Biomarker Standardization Initiative
ICC	Intraclass Correlation Coefficient
kVp	X-Ray Tube Peak Kilovoltage
LVSI	Lymphovascular Space Invasion
mA	X-Ray Tube Milliamperage
MI	Myometrial Invasion
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
mpMRI	Multi-Parametric Magnetic Resonance Imaging
MSSI	Multi-Scale Structural Similarity Index
NGTDM	Neighbouring Gray-Tone Difference Matrix
NPV	Negative Predictive Value
OOB	Out-of-Bag
PACS	Picture Archive and Communication System
PPV	Positive Predictive Value
RFAI	Reconnaissance de Formes, Analyse d’Images

ROC	Receiver Operating Characteristic
SE	Spin-echo
SPHARM	Spherical Harmonics
TensorReg	Regularized Logistic Tensor Regression
TE	Echo Time
TR	Repetition Time
T1	Longitudinal Magnetization Relaxation Recovery Time
T2	Transverse Magnetization Relaxation Recovery Time
VOI	Volume of Interest
2D	Two-Dimensional
3D	Three-Dimensional
5D	Five-Dimensional

1. INTRODUCTION

1.1. Motivations: Precision Oncology

Precision oncology is a strategy in cancer medicine aiming to identify early in the course of treatment which therapeutic option can lead to best outcomes for a given patient based on diverse types of clinical data (1). Therapeutic options include targeted therapies, blood-typing, immunotherapy, and many others whose efficacy is to be assessed prior to prescribing any specific treatment using risk modeling according to big clinical databases integrating imaging, sequencing, histopathology, blood tests, and/or descriptive patients' data. Hence, prognostic studies in precision oncology aim to provide risk stratification for patients with specific cancer types to predict outcomes such as recurrence after treatment (1). Similar goals are typically defined for diagnostic studies as well, given that the grade of a solid lesion as seen on medical images is generally associated with the prescription of surgeries or adjuvant treatments such as chemotherapy or radiotherapy (2, 3). Implementing precision oncology in clinical practice through the development of precise and accurate response monitoring and diagnostic tools could improve decision-making and reduce patient exposure to ill-adapted treatments and unnecessary toxicity (4, 5).

Monitoring of treatment response and diagnoses in solid tumors is done through the measurement of biomarkers (6). Imaging biomarkers, especially quantitative imaging biomarkers, are of great interest, since they can provide a comprehensive view of the whole lesion while capturing clinically relevant biological predictors such as regional tumor intra-heterogeneity (4, 7-9), thus providing opportunities to tailor treatment decisions based on observed responses (4, 6). Imaging-based quantification and characterization of tumoral phenotypes has been the main goal of numerous efforts in recent years developing and integrating precision oncology in clinical practice (4, 5). Identifying optimal quantitative image features and machine learning pipelines for computer-aided diagnosis constitute crucial steps towards the development of reproducible, standardized, and clinically relevant imaging biomarkers of cancer phenotypic characteristics (10). In recent years, numerous quantitative imaging biomarkers based on different image features have been proposed (4, 6). An image feature can be understood as an image-derived descriptor of intensity, shape, texture, or any other visually assessable or quantitatively measurable characteristics of image appearance. However, clinical acceptance of novel imaging biomarkers is

limited and translation into clinical practice generally takes years if not decades. Currently, tumor response and tumor grading is essentially performed through qualitative measurements or using 1D or 2D descriptors of the size of lesions (11). Subjective visual evaluation of lesions on clinical medical images might not capture histopathological or genetic features of disease activity, including intra-tumoral heterogeneity, an important biomarker of cancer aggressiveness (12). Therefore, improved tumor treatment prescriptions could be achieved with comprehensive quantitative imaging biomarkers, overcoming subjectivity of visual interpretation and over-simplistic assessment of shape markers of pathological structures on medical images (8, 9). Thus, standardized and quantitative computational methods have the potential of improving radiology and oncology workflows in patient screening, decision support, detection, and interpretation of findings to alleviate the current burden on radiologists and radio-oncologists (13-15).

The aim of this thesis was to develop computational methodologies for quantifying tumor appearance in image data, using standard and robust sets of mathematical descriptors. Processing pipelines to extract human-defined features from clinical medical images were built and extracted image features were then classified using machine learning classification techniques. Two classes of human-engineered features were studied in this thesis. The first class was three-dimensional (3D) radiomics features and consisted in the computation of a large number of mathematical descriptors of intensity, shape, and texture (16). The second class, spherical harmonics (SPHARM) decomposition (17), was based on a classical computer vision technique expressing volumetric images into weighted sets of mathematically defined 3D spherical harmonics basis functions.

The main hypothesis of radiomics studies is that combining many quantitative image features extracted from tumor regions on medical images can predict underlying genetic or pathological changes happening in response to disease activity. These features includes first-order statistics, morphological, and second-order textural features (16). On the other hand, the SPHARM decomposition method employed is analogous to Fourier frequency decomposition. It encodes the information in an image region within a 2D matrix of SPHARM coefficients (17). An initial testing and validation of these two descriptors was performed on 3D synthetic dataset in **Chapter 3**. After initial validation, pipelines were tested and validated on two clinical studies: i) predicting renal cysts malignancy on contrast-enhanced computed tomography images (**Chapter 4**), and ii) predicting histopathological features of endometrial cancer on multi-parametric magnetic resonance images (**Chapter 5**).

1.2. Contributions and Outline of Framework

This work consisted in methodological developments from the field of computer vision applied to medical imaging in clinical retrospective studies. Our contributions are both technical and clinical. The main technical contributions were i) adapting spherical harmonics (SPHARM) decomposition to medical imaging and comparing it to radiomics, a technique that is widely used in the literature and that has shown potential for clinical translation, and ii) the comparison of fast spherical segmentations to that of expert radiologists' segmentations in the diagnosis of histology-defined lesions' features. The main clinical contributions are i) the development of a radiomics pipeline based on state-of-the-art radiomics literature and on Image Biomarkers Standardization Initiative (IBSI)'s recommendations, and ii) the implementation of SPHARM and radiomics pipelines in two clinical studies.

In this thesis, **Chapter 2** reviews relevant clinical and technical literature which lead to the development of an IBSI-compliant radiomics pipeline. Every step of the proposed pipeline with a focus on reproducibility is described in detail, along with the random forests models employed for features selection and classification.

Chapter 3 describes the SPHARM decomposition method, including the spherical harmonics mathematical framework and its application to voxelated volumes for the creation of a 2D mathematical image descriptor. The logistic regularized tensor regressor used for descriptors classification is also explained on the basis of regular logistic regression and Lasso regularization. A novel segmentation paradigm using spherical regions around tumors is introduced and discussed as a solution to alleviate the overloaded workload of radiologists. These spherical regions are compared to manual segmentations in the studies detailed below. Finally, this chapter includes preliminary analyses on volumetric texture benchmark datasets to assess the ability of SPHARM descriptors to capture textural information from volumes where shape does not encode any discriminative information.

Chapter 4 and **Chapter 5** – corresponding to retrospective clinical studies – are broader versions of manuscripts in preparation for submission to scientific journals. The former is a study on both volumetric computational approaches, radiomics and SPHARM decomposition, for differentiating benign from malignant renal cysts on contrast-enhanced computed tomographic images. A narrowed-down version of **Chapter 4** including solely radiomic analyses has already

been submitted as an abstract to the Radiological Society of North America (RSNA) Annual Meeting 2020, which I co-authored. These chapters are built upon the work of multiple collaborators who have completed essential tasks for the development of this thesis.

For contributions, I have exclusively written and performed analyses presented in **Chapters 1**, **2**, **3**, and **6** with the support and guidance of Dr Peter Savadjiev and Dr Caroline Reinhold. For **Chapter 4** and **Chapter 5**, my contributions were the hands-on development and implementation of both computational pipelines using Python and MATLAB based on prior literature on the topic, the application of the proposed methods to the labeled and segmented datasets, the preparation and presentation of results including tables and figures, and the drafting of the manuscript with a focus on the methods, results, and discussion. Detailed contributions from each collaborator for **Chapter 4** are provided here:

Dr Jeremy Dana is the first author and co-designed this study. Based on inclusion criteria, Dr Dana retrospectively gathered a list of eligible patients from the McGill University Health Centre to build the training dataset of the study. With collaborators from Assistance Publique des Hôpitaux de Paris – Hôpital Necker in Paris, France, Dr Dana has retrospectively gathered another list of eligible patients to build the external testing dataset. He performed semi-automated segmentations of renal cysts on contrast-enhanced computed tomographic images in both datasets and graded renal cysts according to the Bosniak classification system. Dr Dana also drafted the submitted RSNA abstract and contributed to the manuscript in preparation.

Dr Caroline Reinhold is the principal investigator and co-designed the study. Dr Reinhold built the protocol, obtained funding, and coordinated the study. She also graded renal cysts according to the Bosniak classification system and revised the submitted RSNA abstract and the manuscript in preparation. As my co-supervisor, she provided essential support and clinical expertise on renal cysts management in the clinic and on current needs.

Dr Peter Savadjiev is a co-investigator and co-designed the computational methods pipelines of the study. As my supervisor, he also provided essential support and guidance throughout the development of SPHARM, radiomics, and machine learning pipelines.

In the publication of **Chapter 4** as a journal article, only radiomics analyses will be included. Minor changes might be made to the introduction and methods section, but most of these sections are expected to be included *verbatim*, after removing details on SPHARM analyses. The

use of other classifiers might also be investigated prior to submission, to assess if linear classifiers could outperform the more complex ensemble method (*i.e.* random forests) selected in this thesis. Hence, results might change in the final version of the manuscript along with the Computational Methods subsection of the Material and Methods section. Finally, other clinical aims outside of the scope of this thesis will be included in the manuscript (*i.e.* to compare the diagnostic performance of the 2008 Bosniak diagnostic classification system with the proposed 2019 Silverman's updated Bosniak classification guidelines, to evaluate the interobserver variability of these two radiological classification systems, and to define the most accurate and reproducible visually assessable qualitative set of CE-CT features to predict malignancy of Bosniak cysts using the 2019 Bosniak classification and adapting it to the widely used American College of Radiology Data System). Thus, significant changes are expected to be seen between the manuscript for publication and this thesis.

Chapter 5 is a study on both volumetric computational approaches, radiomics and SPHARM decomposition, for predicting deep myometrial invasion and high grade endometrial tumors on multi-parametric magnetic resonance images. Detailed contributions from each collaborator for **Chapter 5** are provided here:

Dr Yoshiko Ueno was the first author of a preliminary version of this work investigating two-dimensional radiomics, which she co-designed (18). According to the study's inclusion criteria, Dr Ueno retrospectively gathered a list of eligible patients from McGill University Health Centre to build the training dataset of the study. With Dr Anthony Dohan and collaborators from Assistance Publique des Hôpitaux de Paris – Hôpital Lariboisière in Paris, France, Dr Ueno has retrospectively gathered another list of eligible patients to build the external testing dataset. She performed manual expert segmentations of endometrial tumors on images from all magnetic resonance sequences in the training dataset and graded tumors according to the FIGO classification system. Dr Sameh Saif contoured endometrial tumors on images from the testing dataset. Dr Ueno also drafted a preliminary version of the manuscript in preparation.

Dr Caroline Reinhold is the principal investigator and co-designed the study. Dr Reinhold built the protocol, obtained funding, coordinated the study, and revised the manuscript in preparation. As my co-supervisor, she also provided essential support and clinical expertise on

magnetic resonance imaging of endometrial cancer and on current needs for the staging of these lesions.

Dr Peter Savadjiev is a co-investigator and co-designed the computational methods pipeline of the study. Again, as my supervisor, he provided essential support throughout the development of SPHARM, radiomics, and machine learning pipelines.

In the publication of **Chapter 5** as a journal article, only radiomics analyses will be included. Minor changes might be made to the introduction and methods section, but most of these sections are expected to be included *verbatim*, after removing details on SPHARM analyses. Finally, aims outside of the scope of this thesis will be included in the manuscript (*i.e.* to compare the reproducibility of radiomics features extracted from MR images segmented by an expert radiologist and extracted from MR images semi-automatically segmented by a less experienced reader, and to evaluate the agreement between radiologists' staging and that of the developed radiomics-based model). Again, significant changes are expected to be seen between this chapter of the thesis and the manuscript to be submitted for publication.

For both studies detailed in **Chapter 4** and **Chapter 5**, Dr Sahir Bhatnagar provided support on statistical considerations especially for the development of the regularized tensor logistic regressor. Prior to submission for journal publication, both studies will be

Discussions on clinical and technical considerations for performing radiomics- and SPHARM-based characterization of tumor volumes are elaborated in both chapters. This thesis concludes with a look forward in what could be expected in future years for such computational methods in the field of oncological imaging.

2. RADIOMICS

2.1. Clinical Background

In the last decade, there has been an increasing interest in quantifying tumor appearance using standard-of-care clinical imaging modalities, such as computed tomography (CT) and magnetic resonance imaging (MRI). These computational methods automatically quantify whole lesions' image-based phenotypes, according to mathematically defined quantitative descriptors (19, 20). This paradigm in medical imaging, called radiomics (8, 10, 19, 21, 22), has been widely explored in the last decade to quantify tumor heterogeneity (12) or predict patient outcomes (23-25). It is also often referred to as radiogenomics when image-extracted mathematical features are associated with genes expression from sequencing data (26). Three radiomics features categories can be defined: first-order statistics, shape-based, and second-order statistics or texture-based features (8, 16) (**Table 2.1**). In radiomics studies, quantitative image features extracted from a 2D slice or from a 3D volume of tumor region on medical images are combined to predict underlying genetic or pathophysiological changes. More specifically, by capturing image intensity level distributions, volumetric shape characteristics, and statistical interrelationships between image contrast values in neighboring voxels, we seek to build reproducible and minimally invasive image-based signatures of tumors. In recent years, numerous studies have explored capacities of these image-based signatures in predicting disease outcomes or grading disease activity (**Figure 2.1**). State-of-the-art biopsies are generally performed to assess histopathological risk factors of tumoral

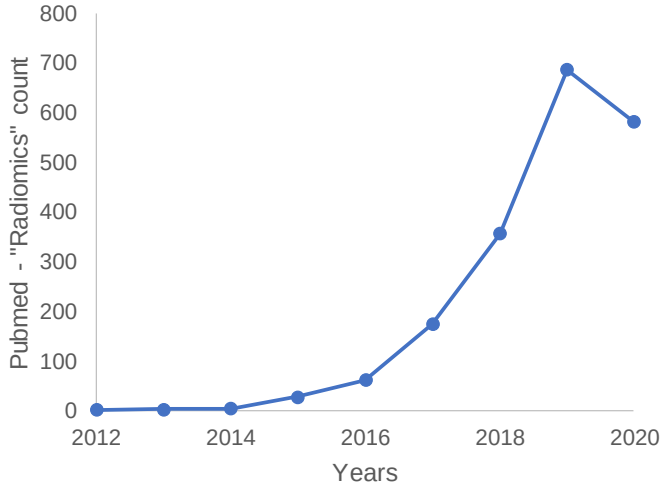


Figure 2.1 Number of scientific publications referring to the term "radiomics" in the current literature (PubMed data, dated 2020-06-16).

activity. Given the invasiveness, the sampling variability of standard-of-care biopsies, and the importance of capturing tumoral heterogeneity (12), the need for noninvasive quantitative whole tumor assessment is often argued, and radiomics could be one solution (27, 28) (**Figure 2.2**). The use of CT- and MRI-based radiomics for decoding tumor phenotypic characteristics in abdominal and pelvic imaging, respectively, are reviewed briefly in the next sections.

Review of Clinical Studies

CT is widely used in oncology clinics especially in the context of radiotherapy since treatment planning and dose distribution assessment are performed based on these images, reproducing the treatment conditions on a flat couch (3). CT is also often used for diagnostic imaging in diagnostic radiology clinics as a frontline imaging modality to diagnose conditions suspected by physicians prior to any treatment prescription. Hence, given the high number of available clinical CT data, CT-based radiomics studies have been widely performed (8, 16, 29-32). In abdominal imaging, radiomics extracted from standard-of-care CT images have been investigated in many different anatomical sites such as the liver, including hepatocellular carcinoma (HCC), either to predict response to treatment (prognostic) such as transcatheter arterial chemoembolization, or to differentiate lesions' grades (diagnostic) such as low grade HCC vs. high grade HCC (29, 33). Since the filtration of contrast material through liver lesions has been shown to be highly predictive of HCC (34), intravenous contrast-enhanced (CE)-CT, typically using iodine (35), is also widely studied qualitatively and quantitatively to predict treatment response or aggressiveness (36). Thus, radiomics analyses of the different enhancement phases of CE-CT has been shown to relate to predictive histopathological features of disease's activity such as microvascular invasion, often associated with bad prognosis (29, 33, 37). Similarly, also in abdominal imaging, renal lesions characterization through radiomics analyses of CT and CE-CT scans has been explored widely but in an unstandardized manner (31, 38). Kidneys are the body's filters and will enhance clearly on CE-CT images while lesions will show either as hyperattenuating or with textured enhancing patterns, which are both characteristics of lesion's grade (39, 40). Therefore, CE-CT is currently the clinically preferred renal imaging technique (39, 40). Radiomics features extracted from enhancement phases of CE-CT, especially from the nephrogenic phase as it is used for small renal cell carcinomas detection, have been included in predictive models to

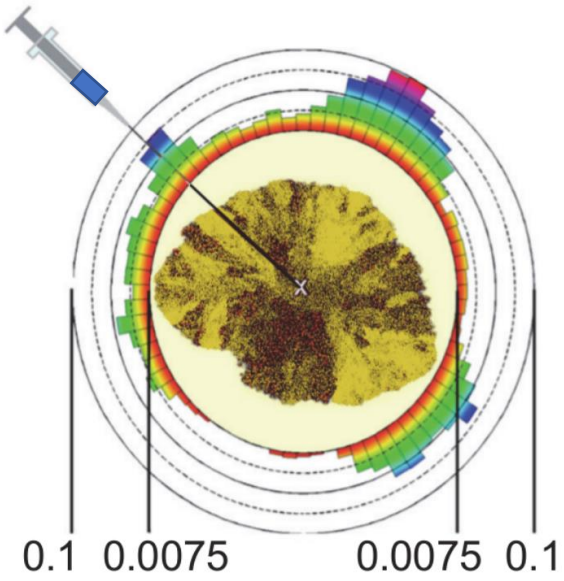


Figure 2.2 Sampling variability of biopsies in capturing tumoral phenotypes. Scale shows angle-dependant fractions of captured phenotypes. Adapted from Poleszczuk J *et al* 2015 (28).

grade disease activity and in particular to differentiate benign from malignant renal lesions for patients' stratification for surgery (31, 38, 39).

In pelvic oncological imaging, MRI is one of the preferred imaging technique as it provides the soft tissue contrast required for tumor detection and delineation (41-44). The clinical use of MRI-based radiomics in prostate cancer for instance has been widely validated in recent years, reporting good features reproducibility and accurate detection of tumor hypoxia or tumor aggressiveness (45-47). MRI has shown high sensitivity for tumor detection and characterization, combining anatomical T2-weighted images with functional imaging such as diffusion-weighted imaging (DWI), derived apparent diffusion coefficient (ADC) maps, or dynamic contrast-enhanced (DCE)-MRI (43). MRI-based radiomics studies had either diagnostic purposes (*e.g.* differentiating healthy from cancerous tissues or grading biological aggressiveness) or prognostic purposes (*e.g.* predicting pathological response or biochemical recurrence) (48). Such studies have been published for cervical, endometrial, prostate, and colorectal cancers (48, 49).

In endometrial cancer, many studies in the last decade have assessed the ability of radiomics features extracted from single sequence or multi-parametric MRI (mpMRI) to diagnose three main histopathological features, *i.e.* lesion's grade, deep myometrial invasion, and the presence of lymphovascular space invasion (18, 50-57). Such noninvasive screening can provide preoperative risk stratification to identify which patients should undergo surgery or adjuvant treatments such as

chemotherapy or radiotherapy (2). In all these cancer sites, the interest in using mpMRI was twofold. First, the need for high soft tissue contrast in pelvic imaging for detection and characterization of lesions has made MRI a clinical standard for grading these tumors. Second, an increase in detection sensitivity and grading accuracy has been achieved thanks to MRI's ability to assess tissue cellularity or perfusion with DWI or DCE-MRI (48). Moreover, just like in the clinical setting where the use of multiple MR sequences helps with patients' stratification through a more comprehensive tumor pathophysiological characterization (41, 43), the use of mpMRI-based radiomics compared to that of single sequence MRI-based radiomics was found to lead to better predictions or diagnoses of histopathological characteristics or treatment response (45, 58, 59).

2.2. Technical Background

While the field of radiomics is attracting a lot of attention, it remains in its infancy and has several challenges to overcome (29, 60-63). Standardization and reproducibility of computational methods are challenging since features extracted through mathematical operations on medical images used in radiomics workflows depend on image acquisition parameters and image preprocessing which vary between scanner hardware and software (63). In fact, research showed variability even in radiomics features themselves when extracted on different platforms (64). All these factors limit greatly the translation of radiomics applications into clinical practice. An international study group recently published 14 recommendations for the translation of quantitative imaging biomarkers such as radiomics into the clinic (6). Among these recommendations were the standardization of the extraction of biomarkers and the reassessment of the precision of biomarkers across institutions (6). Thus, research on reproducibility and repeatability of radiomic features is required to harmonize workflows and to improve radiomics' precision across modalities and across centers. To address these challenges, the Image Biomarker Standardization Initiative (IBSI), an international collaboration of 25 research groups across 8 different countries, was initiated in 2016 to "establish a comprehensive radiomics workflow description, provide verified definitions of commonly used features, and provide benchmarking of features extraction and image processing steps, as well as reporting guidelines", and has recently been published in 2020 (65). Another chapter in IBSI international efforts for radiomics harmonization has also started in June 2020 aiming to improve the reproducibility of image filters (i.e. wavelet filters, Laplacian of Gaussian

filters, etc.) as a preprocessing step in radiomics studies (66). Image filtering, or local image descriptors, was not covered in our work because of the current lack of validation of these filters and since we were interested in non-local image descriptors throughout this project.

Table 2.1 Summary of the 106 radiomics features computed from input images using the Pyradiomics extraction platform and used throughout this thesis.

Category	**Radiomics features**
First-order gray-level statistics: *Global histogram features* (*n* = 18)	Mean, Median, Variance, Skewness, Kurtosis, Entropy, Uniformity, Energy, Total energy, Minimum, 10^{th} percentile, 90^{th} percentile, Maximum, Interquartile range, Range, Mean absolute deviation, Robust mean absolute deviation, Root mean squared
Shape-based statistics: *3D shape features* (*n* = 14)	Mesh volume, Voxel volume, Surface area, Surface area to volume ratio, Sphericity, Spherical disproportion, Maximum 3D diameter, Maximum 3D slice, Maximum 2D column diameter, Maximum 2D row diameter, Major axis length, Minor axis length, Least axis length, Elongation, Flatness
Second-order gray-level statistics: *Gray-Level Co-occurrence Matrix (GLCM) texture features* (67) (*n* = 24)	Contrast, Correlation, Sum average, Autocorrelation, Joint average, Cluster prominence, Cluster shade, Cluster tendency, Contrast, Difference average, Difference entropy, Difference variance, Joint energy, Joint entropy, Informational measure of correlation 1, Informational measure of correlation 2, Inverse difference moment, Maximal correlation coefficient, Inverse difference moment normalized, Inverse difference normalized, Inverse variance, Maximum probability, Sum entropy, Sum of squares, Homogeneity 1 and 2, Dissimilarity, Sum variance
Second-order gray-level statistics: *Gray-Level Run-Length Matrix (GLRLM) features* (68) (*n* = 16)	Short run emphasis, Long run emphasis, Gray-level nonuniformity, Gray-level nonuniformity normalized, Run-length nonuniformity, Run-length nonuniformity normalized, Run percentage, Gray-level variance, Run variance, Run entropy, 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
Second-order gray-level statistics: *Gray-Level Size-Zone Matrix (GLSZM) features* (69) (*n* = 16)	Small area emphasis, Large area emphasis, Gray-level nonuniformity, Gray-level nonuniformity normalized, Size-zone nonuniformity, Size-zone nonuniformity normalized, Zone percentage, Gray-level variance, Zone variance, Zone entropy, Low gray-level zone emphasis, High gray-level zone emphasis, Small area low gray-level emphasis, Small area high gray-level emphasis, Large area low gray-level emphasis, Large area high gray-level emphasis,
Second-order gray-level statistics: *Gray-Level Dependence Matrix (GLDM) features* (70) (*n* = 14)	Small dependence emphasis, Large dependence emphasis, Gray-level nonuniformity, Gray-level nonuniformity normalized, Dependence nonuniformity, Dependence nonuniformity normalized, Gray-level variance, Dependence variance, Dependence percentage, Dependence entropy, Low gray-level zone emphasis, High gray-level zone emphasis, Small dependence low gray-level emphasis, Small dependence high gray-level emphasis, Large dependence low gray-level emphasis, Large dependence high gray-level emphasis
Second-order gray-level statistics: *Neighbouring Gray-Tone Difference Matrix (NGTDM) features* (71) (*n* = 4)	Coarseness, Contrast, Busyness, Complexity

The first category of radiomics features explored in this project was first-order statistics. Histogram-based statistics of the distribution of voxel intensities within the volume of interest (VOI) were extracted directly from preprocessed images (**Table 2.1**). These features included basic descriptive statistics such as the mean of voxel intensities, minimum and maximum values, variance, as well as kurtosis, skewness, and many others (n = 18). The second category was shape- and size-based features. They are descriptors of the 3D topology and morphology of the segmented tumor volume. Features included maximum 3D diameter, surface area to volume ratio, or sphericity, along with many others (n = 14). The third category explored was second-order statistics, or textural features, decrypting gray-level patterns from the spatial distribution of voxels intensities. Five subgroups of texture features were extracted from different matrix representations of images, namely Gray-Level Co-occurrence Matrix (GLCM, n = 24) (67), Gray-Level Run-Length Matrix (GLRLM, n = 16) (68), Gray-Level Size-Zone Matrix (GLSZM, n = 16) (69), Gray-Level Dependence Matrix (GLDM, n = 14) (70), and Neighbouring Gray Tone Difference Matrix (NGTDM, n = 4) (71). Mapping the original image into a textural matrix representation requires gray-level discretization of intensity values across the image. For example, to assess the co-occurrence of a given voxel intensity in its neighborhood, intensities have to be discretized in bins of a preselected gray-level width to count these co-occurrences, or for any other mathematical assessment of inter-relationships of voxel intensities in a VOI (65, 67). Hence, prior to extraction of second-order features, voxel intensities must be resampled into gray-level bins. This discretization reduces image noise and standardizes voxel intensities across patients (65, 72). One textural representation matrix was obtained per texture class from 26-connected voxels considering that voxels in 3D images have direct neighbors in all 13 directions (16). Statistical features are subsequently extracted from these textural representations of the original image, such as contrast, entropy, and many others.

Briefly, GLCM characterizes gray-level co-occurrences by assessing how often pairs of voxels with specific discretized intensity values and in a specified spatial relationship occur along a given angle (67). GLRLM characterizes gray-level runs by assessing lengths in number of consecutive voxels with the same intensity value along a given angle (68). GLSZM characterizes gray-level zones by assessing the number of connected voxels with the same intensity value. GLSZM is rotation invariant since zones are assessed with only one matrix for all directions, unlike GLCM and GLRLM (69). GLDM characterizes gray-level dependencies by assessing the number

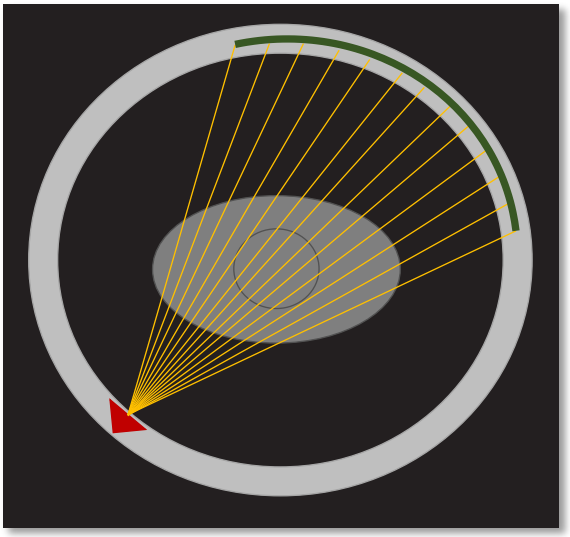


Figure 2.3 Fan beam projection of X-ray source (in red) projected on detector array (in dark green) in CT grantry (in grey).

of connected voxels within a given distance depending on a center voxel (70). NGTDM characterizes differences between an intensity value and the average intensity in its neighbourhood within a given distance, by storing the sum of absolute differences for an intensity (71).

Previous studies have often reported high reproducibility and stability for features in the first two radiomics categories, histogram statistics and morphological features while, on the opposite, texture features tend to provide lower stability under segmentations variations, since they are inherently dependant on regions from which they evaluate voxel intensities spatial distributions (30, 73). IBSI recommendations are modality-specific since each imaging modality depends on several acquisition-specific parameters. Hence, we briefly review hardware and software considerations when performing radiomics studies on CT and MRI.

Radiomics in Computed Tomography

CT is an anatomical transmission-based imaging modality using photon attenuation inside the imaged object to produce an image. It is a tomographic system in the sense that it acquires multiple fan beam projections of the object as represented in **Figure 2.3** in order to reconstruct axial slices typically using iterative algorithms based on filtered backprojection and the Radon transform of all acquired projections (74, 75). Axial slices are obtained sequentially, in a helical fashion while the table supporting the object moves towards the CT gantry. The reconstructed axial

images of the object can be understood as photon mass attenuation coefficients μ maps – mainly from photoelectric and Compton scattering interactions in CT effective energy regions – or electron density maps expressed relative to attenuation in water μ_w in Hounsfield Units (HU) (74):

$$\mathrm{HU} = 1000\frac{\mu - \mu_w}{\mu_w} \tag{2.1}$$

Hence, on CT images, air is represented by -1000 HU, fatty tissues by -120 to -90 HU, bones by +400 to +1900 HU, and fluids (water) by about 0 HU. Radiocontrast such as iodine is radio-opaque and shows as high intensity signal of about 25-30 HU at 120 kV tube voltage (35). Many CT acquisition parameters are adjusted depending on the anatomical region imaged. For better soft tissue contrast, lower peak potential is applied to the x-ray tube (kVp) to adjust the peak photon energy. Increasing tube milliamperage (mA) and the time of exposure is also important to improve the x-ray intensity reaching the detector array. Moreover, different reconstruction kernels can be used to enhance edge detection when a high spatial resolution is needed for instance, or on the opposite, to provide a smoother and a more uniform reconstructed image.

In CT radiomics studies, varying CT acquisition parameters including kVp, section thickness, tube filters, and reconstruction kernel has been shown to decrease greatly the reproducibility of radiomics features when no image preprocessing is performed (30, 76-78). Only standard-of-care variable mA, automatically adapting detector exposure, was found to have little impact on radiomics features, with all other acquisition parameters equal (79). Thus, prior to features extraction, image preprocessing such as voxel size resampling is required to obtain reproducible radiomics features (65, 80).

For reproducibility, IBSI guidelines recommend resampling voxel size to isotropic voxel dimensions and performing either absolute or relative gray-level discretization (65). However, the choice of these preprocessing steps also influences radiomics features extraction and has been shown to impact their reproducibility, especially for second-order statistics or texture features (81). Since CT intensity values (HU) are proportional to electronic density of the imaged tissue, the range of these values for different anatomic sites are well-known and reproducible between two examinations with the same parameters (74). Hence, for similar protocols, CT images have good inter-scan reproducibility and provide “quasi-quantitative” measurements (82). For such a modality, IBSI currently recommends using absolute gray-level discretization by using fixed gray-

level bin widths between the minimum HU and the maximum HU in the image but removing HUs outside of 3 standard deviation range from the mean. This absolute discretization technique enables us to maintain a relationship with the original physical signal, e.g. the mass attenuation coefficients for CT. Even though there still is intrinsic image variability in different manufacturers scanner (82), such preprocessing steps increases reproducibility and harmonizing these steps across radiomics studies will enable the translation of radiomics-based diagnosis or response assessment into the clinic (76, 77).

Radiomics in Magnetic Resonance Imaging

MRI is an anatomical imaging technique which produces volumetric images by manipulating protons' magnetic moments within an object by using magnetic fields to produce signal (74). In the presence of a strong magnetic field, hydrogen nuclei protons will align with the underlying magnetic field $\boldsymbol{B_0}$ due to their spin, an inherent nuclear property arising from quantum mechanics giving each nucleus a small magnetic moment. From this arises a net magnetization $\boldsymbol{M}$ inside the imaged object, which is going to precess at a resonant frequency, the Larmor frequency, due to the nuclear magnetic resonance phenomenon. By tipping this net magnetization with another magnetic field oscillating perpendicularly to the main strong magnetic field $\boldsymbol{B_1}$, we can capture this signal as inductance in a receiver coil as represented in **Figure 2.4** (44). Through slice, phase, and

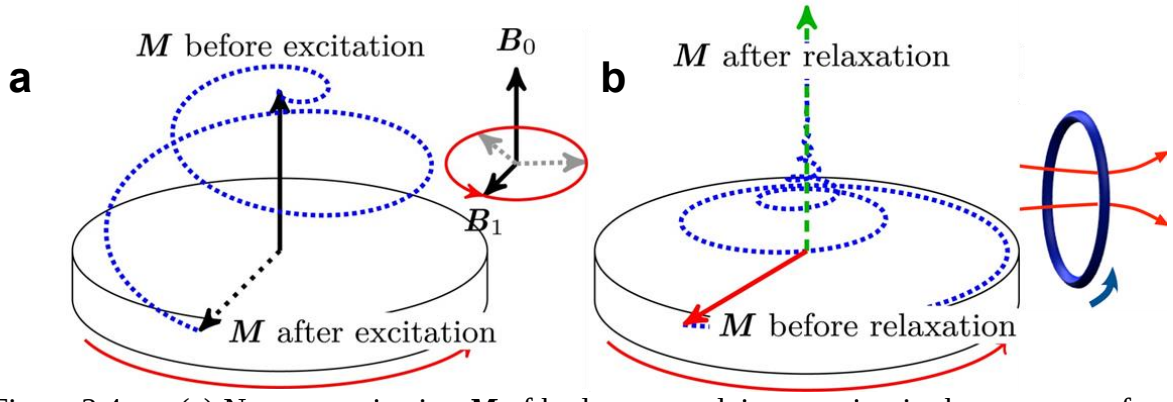


Figure 2.4 (a) Net magnetization $\boldsymbol{M}$ of hydrogen nuclei precessing in the presence of a strong magnetic field $\boldsymbol{B_0}$ tipped in the transverse plane by the radiofrequency oscillating excitation pulse $\boldsymbol{B_1}$ inducing current in inductance loops of MR system's receiver coils. (b) The net magnetization $\boldsymbol{M}$ realigns with the main magnetic field $\boldsymbol{B_0}$ after excitation. Reproduced from Lugauer F *et al* 2018 (44), original figures published under the Creative Commons licence.

frequency encoding of signal with magnetic field gradients, we can spatially localize this signal and acquire echoes after a pre-selected time, or echo time (TE) after the perpendicular excitation magnetic field pulse, while the magnetization recovers its alignment with the main magnetic field, according to the Bloch equations (83). Volumetric images are reconstructed from acquisition of this signal corresponding to data in *k*-space, a spatial frequency space, acquired with different repeated series of magnetic field gradients, or MR sequences, spaced by a preselected time or repetition time (TR) (84). Hence, reconstructions of MR images are based on 3D inverse Fourier transforms, mapping the MR signal from *k*-space to a volumetric image (**Figure 2.5**).

MRI is known for providing high soft tissue contrast and for its ability to provide physiological information about the underlying pathological processes within tissues. In fact, a multitude of MRI protocols are currently clinically available for cancer assessment. Weighting the longitudinal recovery time (T1) of the magnetization with short TR and TE (T1-weighted) will lead to fatty tissues appearing brighter and fluids appearing darker on MR images while weighting the transverse recovery time (T2) of the magnetization with long TR and TE (T2-weighted) will lead fluids appearing brighter on MR images. Numerous MR sequences have been developed to acquire these signals, some including a refocusing pulse known as spin-echo (SE), some without refocusing pulse known as gradient-echo (GRE), some with fast acquisition scheme such as echo planar

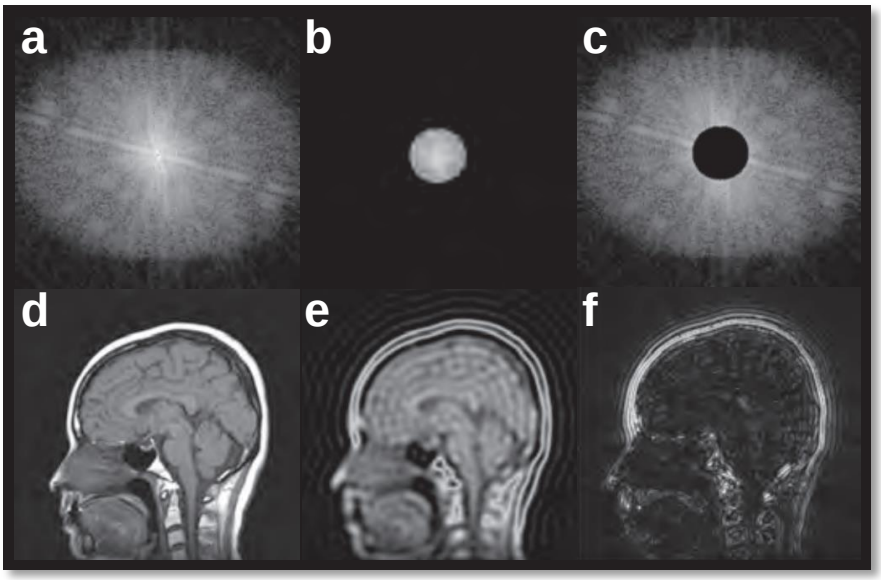


Figure 2.5 MR signal in *k*-space with (a) full acquired sample, (b) low frequencies only, and (c) high frequencies only, and (d-f) their associated reconstructions in spatial domain. Adapted from Bushberg JT *et al* 2012 (74).

imaging (EPI) sequences, and many others with different TE and TR, including or not preparation sequences or acceleration methods. All these possibilities offered by MRI technology also come with some drawbacks: image intensities don't have a fixed meaning.

Other clinical MR sequences can provide physiological information by using contrast agents such as DCE-MRI or without contrast agents such as DWI. These MR contrasts can interrogate tissues' perfusion or diffusion processes as biomarkers of blood volume perfusing through a VOI or by assessing tissue cellularity, respectively. These sequences also come in a variety of implementations to acquire these signals, varying from one implementation to another.

Thus, in MRI-based radiomics studies, one must be even more careful since the measured signal varies between scanners and between repeated examinations. This results in inter-scan and inter-vendor variability often leading to poor reproducibility (85, 86). Since radiomics studies tend to be retrospectively performed on available clinical MRI data, analyses are performed on standard-of-care vendor specific MR sequences with different acquisition parameters which can reduce the reproducibility of extracted radiomics features when acquisition parameters such as TE and TR vary between scans (85, 86). The need for an external testing dataset is therefore even more important for such analyses to be confirmed on other vendor machines across centers (6, 87). Moreover, gray-level discretization and voxel size resampling is required to obtain reproducible radiomics features prior to features extraction according to IBSI recommendations (46, 88).

The use of an absolute discretization algorithm is not recommended by IBSI guidelines for imaging modalities for which image intensity ranges vary from one acquisition to another, like in MRI (65). However, some studies observed that applying normalization to MR images prior to absolute gray-level discretization increased the reproducibility sets of radiomics features on many MR contrasts compared to that of relative discretization with fixed bins number (46, 88). Hence, based on these findings, most recent studies first normalized MRI intensity values prior to absolute discretization of intensities. Most of them either used either z-score normalization of images or rescaled the mean intensity at 300 and the standard deviation at 100 such that gray-level values were within a 0-600 range, with or without a reference tissue (46, 86, 88-94). IBSI's argument on relative discretization for MRI data was not justified in the original report, hence based on recent evidences, absolute discretization can be employed after appropriate image normalization to increase reproducibility (46, 88). Hence, we can conclude that optimizing preprocessing

parameters should be done to identify reproducible features prior to any radiomics study, especially in MRI-based radiomics studies (62).

Radiomics Platforms

A variety of radiomics features extraction platforms have been proposed in the literature. Throughout this thesis, we chose to employ the Pyradiomics platform (Harvard Medical School, Boston, USA, pyradiomics.readthedocs.io) since it is open-source, IBSI-compliant, widely used in prior radiomics studies, and since it enables 3D features extraction (16, 95). In the IBSI report (65), it is indicated that other open-source radiomics platforms have implemented many or all IBSI-covered radiomics features and shown high feature reproducibility following the standardization steps of the IBSI protocol. These platforms include the *Medical Image Radiomics Processor* (MIRP, OncoRay, HZDR, Dresden, Germany), *Medical Imaging ToolKit (MITK) Phenotyping* (DKFZ, Heidelberg, Germany), *RaCat* (University Medical Center Groningen, Groningen, The Netherlands), Dr Martin Vallières's *MATLAB Radiomics Environment* (McGill University, Montreal, Quebec, Canada), and *Standardized Environment for Radiomics Analysis* (SERA, Johns Hopkins University, Baltimore, USA) (96-100). However, most of these platforms have been made available and/or published in 2018 or 2019 and have been limitedly used or referred to since (Google Scholar citations dated 2020-06-16: 14, 9, 9, 143, and 1, respectively). Pyradiomics on the opposite, has been widely used and cited since its original publication in 2014 (Google Scholar citations dated 2020-06-16: *2023*). It is also fully open-source since it is written in the free open-source Python programming language, and not in MATLAB for example, a popular scripting language requiring expensive licenses which has also been widely used to implement radiomics features extraction platforms. Even though using Pyradiomics seems overall advantageous, one of the major drawbacks is that it implements only about 50% of IBSI-defined radiomics features (65).

In the next section, the radiomics pipeline based on the Pyradiomics platform and used throughout this thesis is described in detail. The pipeline is presented step-by-step as a method for the clinical studies presented in **Chapters 4** and **5**.

2.3. Radiomics Pipeline

The radiomics pipeline built in this project consisted of three main steps: i) tumor segmentation, ii) computation of radiomics features within the segmented tumoral VOI, and iii) feature selection, model building, and classification as represented on **Figure 2.6**. **Table 2.1** shows the 106 radiomics features included, as defined and extracted within Pyradiomics. In the development of the pipeline for 3D radiomics features extraction and classification, we followed the recommendations from IBSI and from the most recent literature on the topic. Since recommendations differ for each imaging modality, the pipeline was adapted accordingly.

First, we retrospectively queried for patients' images on the Picture Archive and Communication System (PACS) fitting the inclusion criteria of the study. Clinical and imaging data were stored, and tumors were manually or semi-automatically segmented. Second, in the training dataset, many combinations of pre-processing steps were defined on the Pyradiomics platform to assess the reproducibility of radiomics features. These preprocessing steps included different types of image gray-level normalization (with and without), gray-level discretization, and voxel resampling. For each combination of preprocessing steps, we extracted first-order, shape-based, and textural features. The goal of the fourth step was to define which combination of preprocessing parameters resulted in the highest feature reproducibility as assessed by intra-class correlation coefficients (ICC) (101). To do so, segmentations were automatically dilated and eroded by a single-voxel contour to assess the sensitivity of radiomics features to such changes in the VOI. ICC distributions for all radiomics features classes were then compared to select preprocessing steps resulting in highest reproducibility. More specifically, ICC for each radiomics feature extracted from the three versions of the VOI was assessed as:

$$ICC(1,1) = \frac{BMS - WMS}{BMS + 2WMS} \tag{2.2}$$

where *BMS* is the between-subjects mean squares (*BMS*) estimating feature x_i variance between all *N* subjects, $BMS = \sigma^2_{between-subjects} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}_N)^2$, and *WMS* is the within-subjects mean square (*WMS*) estimating feature x_j variance over all *M* repeated measurements (*M* = 3) in the same subject, $WMS = \sigma^2_{within-subjects} = \frac{1}{M}\sum_{j=1}^{M}(x_j - \bar{x}_M)^2$ (101). Features extracted from the original VOI with the selected normalization, discretization, and resampling

preprocessing steps leading to the highest overall reproducibility (ICC) were further analyzed. Prior to further modeling, each radiomics feature x was normalized according to its mean $\bar{x}$ and standard deviation σ across the population *i.e.* z-score normalized feature $z = \frac{x-\bar{x}}{\sigma}$, in the training and in the testing datasets separately, to enable feature comparison from two different centers (102, 103).

Radiomics features were then included in a final random forest model if they were i) reproducible, ii) uncorrelated as assessed by non-parametric rank correlations, and iii) discriminating, as evaluated by random forests' features selection based on Gini impurity, which is discussed in detail in the next section. More specifically, only features with an ICC greater or equal to 0.80 were included for reproducibility. To handle multicollinear correlated features, we performed hierarchical clustering on features' Spearman's rank correlations (Spearman's $\rho < 0.95$). The feature with the highest variance between classes was kept from each cluster. Finally, up to 5 most discriminative features selected by random forest modeling were included in a final random forest classifier optimized to limit overfitting on our dataset and trained with bootstrapping (1000 bootstraps) and out-of-bag sample validation on the training dataset (104). To account for imbalanced datasets, random undersampling of the majority class in each bootstrapped sample was performed such that the ratio of the number of subjects in the majority class to that in the minority class was less or equal to 75%, using the Python imbalanced-learn toolbox (105). The model was then tested on a testing dataset from another institution to assess its diagnostic performance. Diagnostic performance metrics reported in the testing dataset were used as the real model performance regardless of that in the training which could still be subject to overfitting. Metrics included receiver operating characteristic (ROC) curves with associated areas under the ROC curve

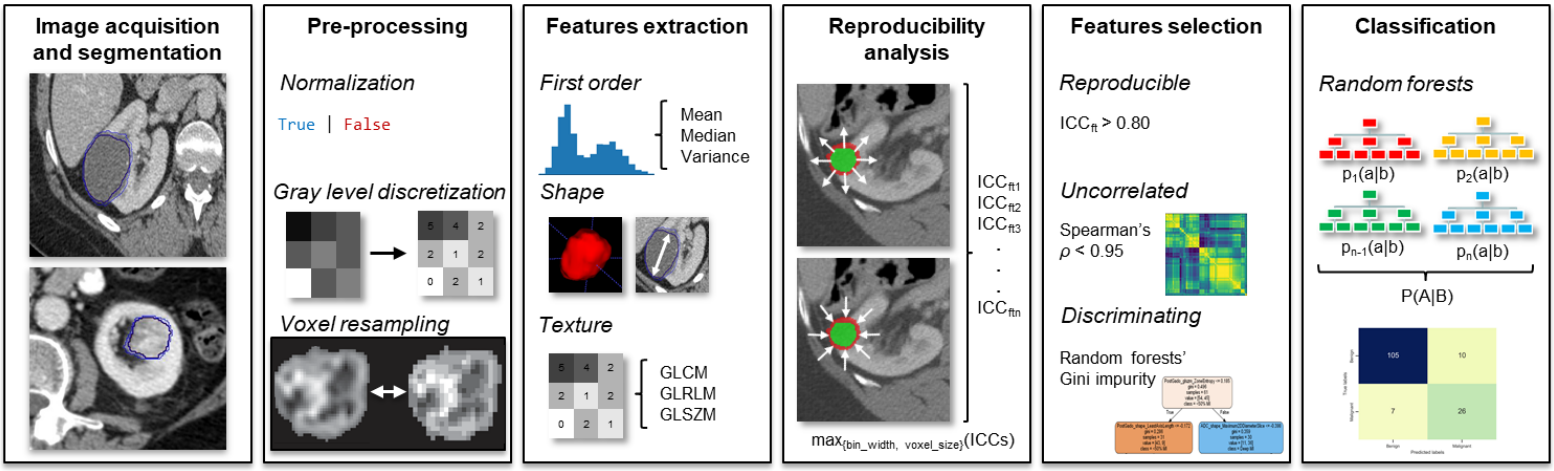


Figure 2.6 Reproducible radiomics pipeline built for this project.

(AUC), sensitivity, specificity, balanced accuracy, positive predictive value, and negative predictive value for the threshold maximizing Youden's index (sensitivity+specificity-1). Understanding of statistical metrics is left to the reader whom will find extensive literature on the topic (106). This pipeline was applied in clinical studies detailed in **Chapter 4** and **Chapter 5**.

2.4. Random Forest Classification

Throughout this project, we were interested in predicting binary categories from labeled data, thus performing supervised classification tasks. Random forest classifiers were selected for their many advantageous properties for features selection and classification. The random forest model is an ensemble technique combining predictions from multiple randomized decision trees which nodes are subsequently split by a subset of randomly selected features from the training dataset (25). Random forest classifiers' advantages include the following: i) ensemble learning with multiple randomized trees allow probabilistic classification and reduced training bias; ii) individual trees are decorrelated because a random proportion of features are used at each split enabling to capture important features even in high-dimensional or noisy data; iii) it is a nonparametric flexible feature selector with few hyperparameters to tune, and iv) it intrinsically evaluates the generalization error model on an out-of-bag sample (107-111). In fact, since random forest classifiers are ensemble-based classifiers using bagging, *i.e.* generating new training sets using sampling with replacement, a part of the training set was always used to assess the generalization error, or the out-of-bag (OOB) error, of each random forest model. The OOB error was calculated as the mean prediction error on each training sample $\boldsymbol{X}$, using only the trees that did not have $\boldsymbol{X}$ in their bootstrap sample. Therefore, there was no need to rely on a separate validation set to perform model selection as we relied on OOB error. To minimize over-fitting, the number of trees, the maximum decision trees' depth, and the number of features to possibly split at each node were minimized, along with the OOB error. We briefly explain the decision tree algorithm used to build random forests.

Each binary decision tree is built in a top-down approach following the Classification and Regression Tree (CART) algorithm (112). The root is built from a labeled bootstrapped sample of the training dataset given as an input to the decision tree. At each decision node of the tree, a

randomly selected set of features is made available. The goal is to find the most discriminating feature from this set, or split feature θ, enabling to separate as many subjects pertaining to a class 0 from the other class 1 for a given feature threshold t_θ, based on Gini impurity criterion i_G (**Figure 2.7**). Hence, at each node, we aim to minimize the mixture of classes passed at each side of the node to further splits such that eventually subjects from the same class are grouped together. Gini impurity i_G, a metric similar to entropy, is used to minimize the probability of misclassification at a given node and measures how accurately a given split based on a selected feature threshold is discriminating between the two classes, 0 and 1, or benign and malignant (112):

$$i_G = 1 - p_0^2 - p_1^2 \tag{2.3}$$

where $p_j = n_j/n$ is the fraction of subjects pertaining to class $j \in 0,1$ out of the total sample at this specific node of the tree. Hence, if at a given node all subjects pertain to class 0 and none to class 1, then $i_G = 0$ and the sample is "pure". If not, the samples are split at this node according to the selected feature threshold and are passed at two other nodes on the *left* and the *right* such that the difference in Gini impurity Δi_G is evaluated as:

$$\Delta i_G = i_G - \frac{n_{left}}{n} i_{G,left} - \frac{n_{right}}{n} i_{G,right} \tag{2.4}$$

To maximize this decrease in Gini impurity ΔI_G, an exhaustive grid search over all features θ available for this node to split and across all their possible threshold values t_θ is efficiently performed as $\theta^* = \text{argmin}_\theta(\Delta i_G(\theta, t_\theta))$. This optimal split $\Delta i_{G,\theta^*}$ is then kept in memory for all nodes and all trees in the random forest for each feature in order to can define Gini feature importance of a given feature as:

$$I_G(\theta) = \sum_{tree=1}^{N_{tree}} \sum_{node=1}^{N_{node}} \Delta i_{G,\theta}(node, tree) \tag{2.5}$$

Decision nodes are split until there is only a single subject on a side of the node, until there are only subjects pertaining to a single class on each side of the node, or until maximal tree depth is reached. This is where the leaves level is reached and where classes are predicted. Each randomized decision tree in the random forest contributes to the final predictions of the forest given outputs at the leaves level. Combining these decision trees by taking the mean of predictions results

in improved classification accuracy, and is known as ensemble learning which reduces model's bias (111).

For example, let's follow one of the branches of the decision tree on **Figure 2.7** built during preliminary analyses of the mpMRI of endometrial cancer project for predicting deep myometrial invasion (MI), presented in **Chapter 5**. The blue color wash is proportional to the purity of the samples associated with histopathological deep MI while orange is associated to purity of samples without deep MI. First, we start at the root with a sample of the training population, 59 lesions without deep MI and 40 lesions with deep MI. Thus, the Gini impurity is equal to $i_G = 1 - \left(\frac{59}{99}\right)^2 - \left(\frac{40}{99}\right)^2 = 0.482$, indicating impure samples. Using a dependence entropy extracted from GLDM on delayed phase DCE-MRI (referred to as *PostGado* in **Figure 2.7** for post-gadolinium) of 0.052, the grid search found that it could decrease significantly the impurity of samples, splitting at two nodes subjects into "purer" samples without deep MI and with deep MI. Following the left branch of subjects mostly without deep MI (48 vs. 5), hence with smaller Gini impurity ($i_G = 1 - \left(\frac{48}{53}\right)^2 - \left(\frac{5}{53}\right)^2 = 0.171$), a dependence nonuniformity threshold, also extracted from GLDM on delayed phase DCE-MRI, of -0.535 resulted in another good node split. Following again the left branch of the node, we now almost have a pure sample with 35 subjects without deep MI and only one subject with deep MI, thus $i_G = 0.054$. We find at this node that a zone entropy threshold extracted from GLSZM on delayed phase DCE-MRI of -1.955 enabled us to split perfectly in a pure sample ($i_G = 0$), on the right side of the node, subjects without deep MI. Thus, if a given lesion without deep MI had radiomics features corresponding to each node split condition which lead to this leaf at the end of the tree, it would be accurately predicted as without deep MI in a binary manner.

Extending from single CART binary decision trees, random forest classifiers enable robust probabilistic classification. Therefore, resulting probabilities, or risks, generally between 0 and 1 can be leveraged to interpret how predictions the random forest model are certain based on included features. Random forests classifiers decision planes can even be visualized to understand probabilistic decision boundaries built during the training process. These boundaries in a *N*-dimensional space (depending on the number *N* of features included in the modeling) are built from the combination of randomized decision trees such as the one represented below in **Figure 2.7**, enabling probabilistic classification.

Reported diagnostic performance metrics on the training dataset are obtained on OOB bootstrapped balanced samples (1000 samples), while that obtained on the testing set is obtained by applying the trained model on bootstrapped samples of the testing dataset.

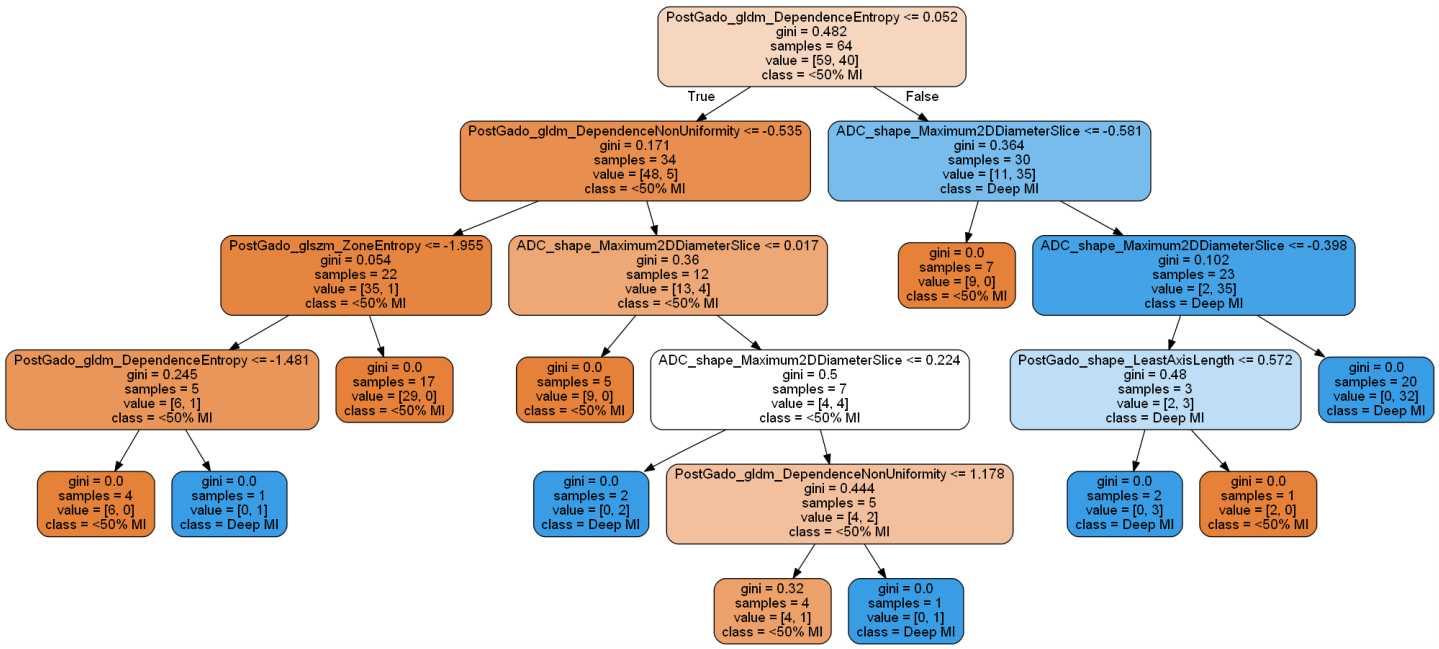


Figure 2.7 Decision tree built in the random forest classifier for differentiating <50% myometrial invasion from deep myometrial invasion in endometrial cancer from radiomics features extracted from multi-parametric MRI.

3. SPHERICAL HARMONICS DECOMPOSITION

3.1. Background

Spherical harmonics functions arise from the angular solution to Laplace's equation in spherical coordinates introduced at the end of the 18th century (113), and also appear as the angular solution to Schrödinger's equation when modeling atomic orbitals in the hydrogen atom introduced at the beginning of the 20th century (**Figure 3.1**) (114-116). Adapting this mathematical formalism for spherical harmonics (SPHARM) decomposition was first introduced to provide compact representation of three-dimensional (3D) models for 3D objects search engines in the early 2000's (17, 117). It has been applied in numerous fields since then, including molecular chemistry (118), evolutionary biology (119), acoustics (116), and even neuroimaging to characterize the brain's cortical surface (120-123). Because of the spherical topology of some solid tumors, SPHARM decomposition is ideally suited for characterizing their harmonic frequency content and therefore their textural appearance.

To alleviate some of the problems encountered in radiomics studies including the need for gray-level discretization and voxel resampling, we propose to compare a standard radiomics approach with the SPHARM decomposition method. This paradigm in medical imaging consists in computing simple transformations on native medical images based on the mathematical framework developed in 1926 to model the hydrogen atom's electronic orbitals (114). The key idea of the SPHARM decomposition approach is to obtain a rotation-invariant representation of tumors storing amplitudes of different SPHARM frequency components at different radii (17). The reason for adopting this paradigm is threefold.

First, SPHARM computation is known to be robust to a large variety of image transformations, such as rotation, translation, or changes in brightness and contrast (17). These invariance properties of SPHARM descriptors make it robust to changes in image signal that are caused not by the underlying pathophysiology, but rather by technical differences in image acquisition, which can be particularly prominent if images are acquired at different scanning sites or institutions or seen in MRI data in general. Hence, it does not require image gray-level resampling or normalization such as what is needed in radiomics studies.

Second, it is a multi-resolution representation as it represents a 3D object in terms of energies at different SPHARM frequencies, with information at lower frequencies representing general smoother shape description of the original object and information at higher frequencies representing rougher textured variations (17, 124). Hence, it is well-suited for medical imaging and does not require image resampling to isotropic voxel sizes such as what needs to be performed in radiomics studies. It is also compact as it reduces the dimensionality of the image while allowing reconstruction of the initial tumor.

Third, the SPHARM technique has been long known to capture structural image information that is relevant both for automated image processing, as well as for biological visual perception of image content (17, 117). Thus, our hypothesis is that a set of SPHARM features can capture predictive tumoral structural and textural intra-heterogeneity and provide similar or better performance for predicting histopathological features of solid tumors than a standard radiomics approach, computed directly on the original anatomical images.

Before further developing on the SPHARM decomposition pipeline, a note is made here to differentiate this work from implemented image filtering in Pyradiomics built upon the spherical harmonics mathematical framework. This filtering process is called 3D local binary pattern (LBP) and uses spherical harmonics basis functions to map locally in a voxel-wise manner surrounding image content in each voxel for a given spherical harmonic degree and radius (125). As mentioned previously, such local filtering methods were not studied throughout this thesis. SPHARM decomposition is significantly different: it is not a local voxel-wise image filtering method, but rather a non-local image transformation providing a compact 2D mathematical image descriptor of a whole volume. Therefore, the SPHARM decomposition pipeline detailed below aimed to provide a quantitative volumetric image descriptor to be analyzed as such, and not to be used to further extract other mathematical descriptive statistics from the resulting descriptor (*e.g.* mean, variance, kurtosis, etc.), which would be done on 3D LBP-filtered images, or other ISBI-defined image filtering processes, in radiomics studies (66).

Second, we also want to mention that preliminary versions of this work included development of more complex SPHARM-inspired volume decompositions. To increase the creativity in SPHARM descriptors and to investigate even more interesting mathematical properties of these descriptors, a weighted Fourier transform on SPHARM coefficients along the

radial dimension in the SPHARM pipeline was tested, for instance, along with other transformations inspired by previous work by Skibbe H *et al* (126, 127). Invariance properties and classification accuracies on 3D shape benchmark datasets of these SPHARM-inspired descriptors are described elsewhere (126, 127). The slightly improved shape characterization accuracy reported with these updated SPHARM descriptors was not worth the major trade-off that had to be done when using such additional transformations, losing the localization and reconstruction properties of SPHARM descriptors, as presented in this thesis. In fact, the interpretability of computer-aided diagnosis methods in diagnostic radiology is of the upmost important for translation of such techniques into the clinical pipeline. Another preliminary version of this work included the use of 3D spherical Bessel functions as a radial basis, combined with angular SPHARM decomposition, as proposed by Galinsky VL and Frank LR (120). This method did not require sampling of shells across the volume to describe its content, but rather selecting the order of Bessel functions to be included in the decomposition, similar to that of SPHARM's maximal degree expansion L_{max}. However, our preliminary implementation was unstable and reconstructions using these descriptors lead to unexplained artifacts. Therefore, a similar SPHARM decomposition approach than that of the original work by Kazhdan M *et al* (17) was chosen and exploratory analyses with other SPHARM-inspired descriptors were not reported in this thesis.

3.2. SPHARM Pipeline

Mathematical Framework

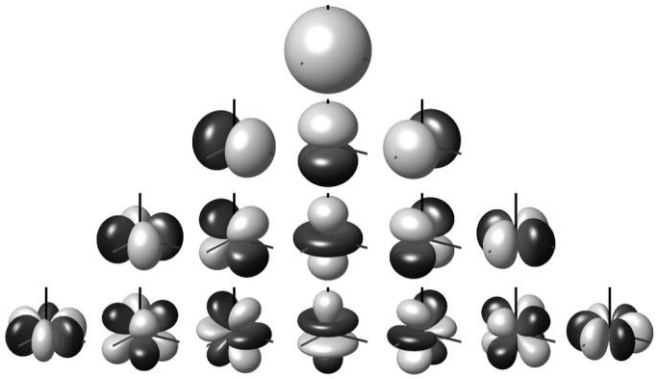


Figure 3.1 Spherical harmonics basis functions $Y_m^l(\theta, \varphi)$ of degree l, from 0 to 3 (top to bottom), and order m, from $-l$ to l (left to right). Reproduced from Zotter F *et al* 2019 (116), original figure published under the Creative Commons licence.

SPHARM decomposition can be viewed as a spherical analogue of Fourier frequency decomposition of any continuous function. A given continuous one-dimensional function can be written out as a Fourier series, or a sine and cosine expansion. This frequency decomposition expresses the initial function as a weighted sum of sine and cosine functions, up to a given expansion degree (128). Similarly, a given 3D spherical function $f(r,\theta,\varphi)$, for example a shell within a tumor volume, can be expressed as a weighted sum of spherical harmonics functions $Y_m^l(\theta,\varphi)$ up to a given expansion degree L_{max}. Hence, spherical harmonics functions in SPHARM decomposition are analogous to sine and cosine functions in Fourier analysis. Since their linear combination represents any arbitrary spherical topology, they can also be understood as eigenfunctions of angular momentum, weighted by coefficients c_{lmr} (129), represented as:

$$f(r,\theta,\varphi) = \sum_{l=0}^{L_{max}} \sum_{m=-l}^{l} c_{lmr} Y_m^l(\theta,\varphi) \tag{3.1}$$

where $f(r,\theta,\varphi) = (x(r,\theta,\varphi), y(r,\theta,\varphi), z(r,\theta,\varphi))$ is the 3D spherical function of the initial voxelated shell in spherical coordinates, $\theta \in [0,\pi]$ is the polar angle, $\varphi \in [0,2\pi]$ is the azimuthal angle, c_{lmr} are SPHARM coefficients, and $Y_m^l(\theta,\varphi)$ are Laplace's spherical harmonics functions. More specifically, SPHARM coefficients c_{lmr} represent the importance or the weight of a spherical harmonic function of degree l and order m. These SPHARM expansion degrees and orders are integer quantum numbers of orbital angular momentum (114, 115). SPHARM eigenfunctions of angular momentum are expressed as a function of associated Legendre polynomials $P_m^l(\cos\theta)$:

$$Y_m^l(\theta,\varphi) = \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)}{(l+m)}} P_m^l(\cos\theta) e^{im\varphi} \tag{3.2}$$

where associated Legendre polynomials $P_m^l(\cos\theta)$ can be written out as:

$$P_m^l(\cos\theta) = (-1)^m 2^l (1-\cos^2\theta)^{m/2} \sum_{k=m}^{l} \frac{k!}{(k-m)!} \cos^{(k-m)}\theta \binom{l}{k} \binom{\frac{l+k-1}{2}}{l} \tag{3.3}$$

where $\binom{l}{k}$ is a binomial coefficient (130). For medical images consisting in voxel grids, SPHARM decompositions are performed after expanding concentric shells at different radii from the center of the volume of interest (VOI), or tumor, to obtain a collection of 3D spherical functions.

Each function defined at a specific radius then has to be mapped separately from a volumetric cartesian coordinate system $f(x, y, z)$ to a parametrized spherical mesh coordinate system $f(r, \theta, \varphi)$ specifically designed for 3D analysis of volumetric images and using uniform icosahedron sampling of shells (131). SPHARM decomposition is performed on these 3D spherical functions to store the amplitude of different spherical harmonics frequency components (degree l and order m) at a given radius and up to a maximal degree L_{max}, by computing:

$$c_{lmr} = \int_0^{2\pi} \int_0^{\pi} \sin\theta \, f(r, \theta, \varphi) \, \overline{Y_m^l}(\theta, \varphi) d\theta d\varphi \tag{3.4}$$

where $\overline{Y_m^l}(\theta, \varphi)$ is the complex conjugate of $Y_m^l(\theta, \varphi)$. Limiting the degree of the SPHARM decomposition L_{max} controls the level of details included in the frequency decomposition of the tumor, avoids modeling the higher frequency noise component of the image, and provides fast computation of compact SPHARM descriptors. Smooth surfaces with lower frequency content can be represented with few SPHARM degrees while complex textured surfaces with protrusions and deformations must be modeled with higher expansion frequencies. This is done in a computationally efficient way by addressing the SPHARM decomposition problem as a non-equispaced discrete spherical Fourier transform for which fast Fourier transforms are already broadly implemented and available (124, 132). Using an implementation in C++ and MATLAB with parallelization with OpenMP enabled us to minimize the complexity of SPHARM descriptors calculation with **Equation 3.4** (computation time < 5s for 64 x 64 x 64 voxelated volumes) (124). After decomposing 3D functions for all radii, we take the L2-norm of each L_{max} x $(2L_{max} + 1)$ matrix along the order m dimension (of size $2L_{max} + 1$) for each radius:

$$|c_{lr}|_2 = \sqrt{\sum_{m=-l}^{l} |c_{lmr}|^2} \tag{3.5}$$

Since any rotation of a spherical function does not change its energy, or L2-norm, resulting $|c_{lr}|_2$ are rotation-invariant – among other things – which explains why there is an interest in summing on the *m* dimension to analyze 2D L_{max} x R_{max} descriptors rather than analyzing a 3D L_{max} x M_{max} x R_{max} descriptors (17). In fact, combining $|c_{lr}|_2$, a 2D representation of the image is obtained in terms of SPHARM coefficients, at each radius and frequency or degree l (**Figure 3.2**):

$$\boldsymbol{C}_{SPHARM} = \begin{bmatrix} |c_{1,1}|_2 & \cdots & |c_{1,R_{max}}|_2 \\ \vdots & \ddots & \vdots \\ |c_{L_{max},1}|_2 & \cdots & |c_{L_{max},R_{max}}|_2 \end{bmatrix} \tag{3.6}$$

The 2D matrix $\boldsymbol{C}_{SPHARM}$ is referred to as SPHARM descriptor throughout this thesis (17). Individual SPHARM coefficient within the 2D matrix corresponds to a relative amount of harmonic filter importance, or spherical harmonics basis function weight, at a given radius and frequency. Each individual feature does not characterize the whole VOI as such. It is only when integrated together in this 2D matrix structure that a complete set of descriptors of the object of interest is obtained which allows classification of the volumetric object.

By construction, the SPHARM representation is translation and rotation invariant, because it does not rely on information depending on the orientation of the tumor. In fact, since we investigate the energy of each frequency band up to a limited expansion (by taking the L2-norm), SPHARM descriptors are also robust to systematic or small gray-level variations and to noise. These invariance properties arise from the fact that any rotation of a spherical function does not change its energy, or L2-norm (17). Moreover, knowing the initial 3D cartesian grid of the tumor region, SPHARM descriptors can be used to reconstruct the initial volumetric tumor through an iterative inverse SPHARM transform (124). SPHARM coefficients as a function of order *l* and degree *m*, c_{lm}, thus prior to taking the L2-norm, and at each radius are simply given as an input in

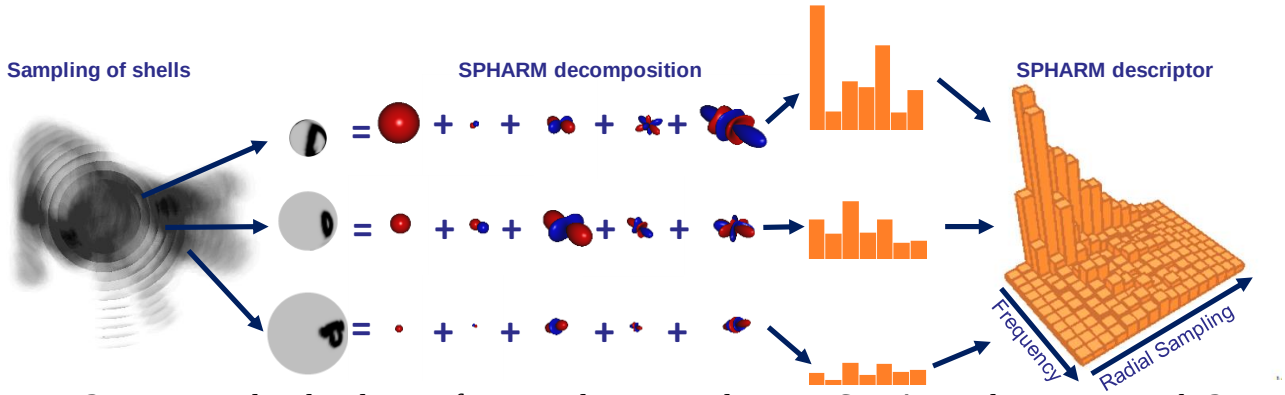


Figure 3.2 Voxelated volume of a cow decomposed into a SPHARM descriptor. Each 3D spherical harmonics function of degree *l* has a specified weight $|c_{lr}|_2$, corresponding to the L2-norm of all orders $m \in [-l, l]$, and describes image content on the shell at radius *r*. The resulting 2D SPHARM descriptor is built from these weights of as a function of degree *l*, or frequency, and shell radius. Adapted from Kazhdan M *et al* 2003 (17).

Equation 3.1 resulting in 3D spherical functions on spherical parametrized meshes which are then mapped back on the known cartesian voxelated grid for each radius.

For radial sampling, previous studies on 64^3 voxels 3D objects from the Princeton Shape Benchmark dataset (133) used between 25 and 34 radial samples (one shell every 2.5 or 2 voxels) and the most discriminating results were obtained using 25 shells and SPHARM decomposition up to $L_{max} = 25$ (117, 127). Prior to any analyses on clinical data, these expansions terms must be selected. Our aim by selecting maximal radial sampling and frequency expansion is to find the smallest descriptor which is discriminant in different classification tasks, and which enables accurate reconstructions of the original 3D object. For instance, the grade of a solid tumor may be related to smooth low frequency contrast changes while detecting the outline or textured patterns of a tumor may require higher frequency components to represent edges or small gray-level variations. Low frequency components account for the contrast and high amplitude intensity information, whereas high frequency components account for fine high-resolution textured details. This is analogous to spatial frequency data in the Fourier domain, where the origin of Fourier space represents low-resolution spatial information while the periphery represents higher spatial frequency accounting for higher resolution and sharp features in the image, similar to MRI data represented in **Figure 2.5** (84). Hence, selecting these parameters (L_{max} and R_{max}) is a balance between compactness and volume appearance characterization accuracy.

SPHARM Pipeline

Similarly to the pipeline developed for radiomics in **Chapter 2**, a pipeline for SPHARM decomposition is developed in this section in the context of clinical imaging (**Figure 3.3**). Briefly, retrospective patients' images queries were performed on the PACS fitting the study's inclusion criteria. Clinical data was stored, and tumors were manually or semi-automatically segmented. Second, each segmented volumetric tumor was mapped from a cartesian voxelated grid to a spherical mesh using uniform icosahedron sampling of shells (131) on which SPHARM decomposition was performed for a chosen level of radial and frequency expansion. To address the high dimensional nature of SPHARM descriptors and to preserve and leverage spatial information in 2D SPHARM matrices, a regularized tensor logistic regression (TensorReg) technique was selected to classify SPHARM descriptors since it fits regression coefficients directly into a matrix

or tensor structure while identifying most important coefficients and nulling non-discriminative coefficients (134). We performed cross-validated bootstrapped training of the TensorReg model on our training dataset to produce a 2D classification matrix, consisting in regression coefficients from the TensorReg modeling. From this classification matrix, we computed predictions by taking inner scalar product with a given SPHARM descriptor resulting in a risk of pertaining to one of the two classes assessed, between 0 and 1 (135). In the next section, more details are provided on the TensorReg technique and on the reasons for selecting it. The model was then tested on an external testing dataset to evaluate its diagnostic performance. This pipeline was applied in each clinical study described in this thesis.

An additional step can also be taken with SPHARM decomposition by applying an iterative inverse SPHARM transform and reconstructing either the whole original tumor, or a given harmonic filter at a specific radius and frequency which was predictive based on the TensorReg classification matrix, as shown in the last panel of **Figure 3.3**. Visualizing these filters is analogous to the visualization of convolutional filters in convolutional neural networks, since the image content which is predictive of a given class will result in an enhancement of this information by these filters (136). The reconstruction of whole volumes requires SPHARM coefficients as a function of orders *l*, degrees *m*, and radii *r*, and to know the original volume size and resolution. The ability of SPHARM descriptors to reconstruct whole volumetric tumors was assessed with multiscale structural similarity indexes (MSSI), a similarity measurement of image luminance, contrast, and structure (137). The MSSI was chosen since it evaluates the perceived quality between the reconstructed volume and the reference volume similarly to mean squared error-based methods but at different versions of the volume with various resolution scales, thus more robustly. The structural similarity index is first assessed according to intensity mean $\bar{x}$, intensity variance σ^2, and covariance of intensities between each VOI σ_{AB} as:

$$SSI(\mathrm{VOI}_A, \mathrm{VOI}_B) = \frac{(2\bar{x}_A\bar{x}_B + \varepsilon_1)(2\sigma_{AB} + \varepsilon_2)}{(\bar{x}_A^2 + \bar{x}_B^2 + \varepsilon_1)(\sigma_A^2 + \sigma_B^2 + \varepsilon_2)} \tag{3.7}$$

with ε, small stability constants (138). It is then assessed at 5 resolution scales, down-sampling the 3D image resolution by a factor of 2 from one scale to another (*e.g.* 8 x 8 x 8 voxels to 4 x 4 x 4 voxels), such that the MSSI evaluated across this work is finally defined as:

$$MSSI(\mathrm{VOI}_A, \mathrm{VOI}_B) = \prod_{i=1}^{5} SSI_i\left(\mathrm{VOI}_{A,i}, \mathrm{VOI}_{B,i}\right) \tag{3.8}$$

A MSSI close to 1 indicates that the reconstructed VOI is very similar to the original VOI in terms of luminance, contrast, and structure, while a MSSI close to 0 indicates significant differences between the two objects.

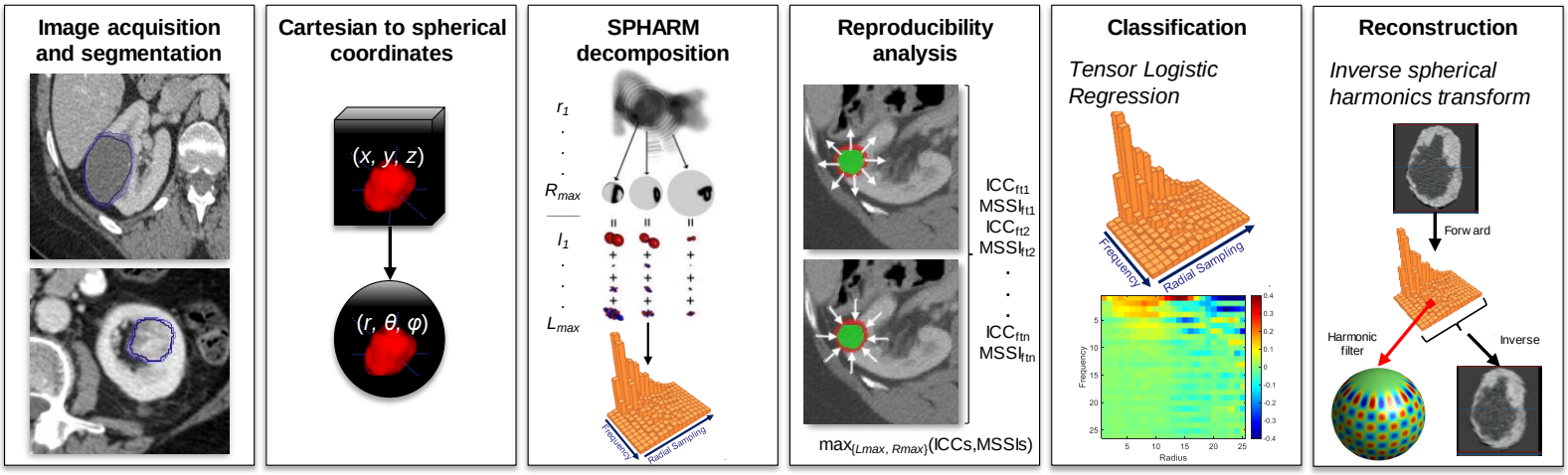


Figure 3.3 SPHARM decomposition pipeline developed for clinical studies. Panel 3 adapted from Kazhdan M *et al* 2003 (17).

3.3. Tensor Logistic Regression

Regularized tensor logistic regression, referred to as TensorReg throughout this thesis, was employed to perform probabilistic classification of SPHARM descriptors in problems with two classes (134, 139). Reasons for choosing TensorReg were twofold. Models such as simple logistic regressors or random forest classifiers require as an input a series of labeled 1D vectors of features to be trained on. Thus, in the case of SPHARM descriptors, we need to flatten the 2D matrix to a 1D descriptor which becomes of exceedingly large size (*e.g.* 25 x 25 = 625), while we might have a sample size of with less than 150 patients to train on (number of features ≫ sample size). Moreover, in this flattening process, we lose all the structural information encoded in the matrix dimensions, *i.e.* each coefficient is related to its neighbors corresponding to a different frequency energy band at a given radius, which is even more critical for reconstruction. Hence, using Hua-Zhou's TensorReg (TensorReg Toolbox v1.0, MATLAB 2020a, Mathworks; hua-

zhou.github.io/tensorreg/) allowed to address the high dimensionality problem and to preserve spatial information in the matrix structure, through robust regularization of matrix-based logistic regression (134). For understanding TensorReg, we first describe regular logistic regression and Lasso regularization.

A logistic regressor is a linear classification model in which we aim to fit the logarithm of the odds (140):

$$\log\left(\frac{\hat{y}}{1-\hat{y}}\right) = \beta_0 + \sum_{i=1}^{N} \beta_i \, x_i \tag{3.9}$$

where $\hat{y}$ is the probability for the estimated subject of pertaining to class 1, β_i are the regression coefficients we aim to find to maximize classification accuracy, and x_i are features on which we fit the model (*N* features). The probability of pertaining to class 1 can be expressed as:

$$\hat{y} = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{N} \beta_i x_i)}} \tag{3.10}$$

In the training process, regression coefficients β_i associated to each feature x_i are adjusted to minimize the error throughout the training dataset. The minimization problem is built from the maximization of the log likelihood, hence the minimization of the negative log likelihood, also known as cross-entropy loss L_{CE}, which indicates how much the estimated class $\hat{y}$ differ from the true class *y*:

$$L_{CE}(\hat{y}, y) = -[y\log(\hat{y}) + (1 - \mathrm{y})\log(1 - \hat{y})] \tag{3.11}$$

Hence given feature vectors x, we perform a stochastic gradient descent optimization (141) on regression coefficients vector β to minimize this loss on all subjects in the population *M* with the following optimization formula:

$$\beta^* = \mathrm{argmin}_\beta \left(\frac{1}{M} \sum_{j=1}^{M} L_{CE}\left(y^{(j)}, x^{(j)}, \beta\right) \right) \tag{3.12}$$

When the optimal set of regression coefficients β_i^* are evaluated in the training phase, we can predict a class for given any vector x on a continuous scale from 0 to 1. However, the logistic regression model will likely overfit the training dataset. To avoid this, we introduce a regularization term on the L1-norm of the regression coefficients, a Least Absolute Shrinkage and Selection

Operation (Lasso) regularization. By adding a regularization term to the minimization problem, our logistic regressor will be less likely to overfit and model noise or random feature fluctuation in the training dataset, leading to better model generalization (140). With Lasso regularization, the minimization problem for fitting the regression coefficients becomes:

$$\beta^* = \text{argmin}_\beta \left(\frac{1}{M} \sum_{j=1}^{M} L_{CE}\left(y^{(j)}, x^{(j)}, \beta\right) \right) - \lambda \sum_{i=1}^{N} |\beta_i| \tag{3.13}$$

with a regularization coefficient λ to be selected. Lasso regularization enables to put some regression coefficients to 0 and tends to assign larger weights to discriminating features, providing sparse representations of large data. During training, cross-validation bootstrapped validation is performed to select regularization coefficient λ minimizing classification error.

Therefore, TensorReg is an expansion of this framework to matrices and tensors (134). The gradient descent algorithm and Lasso regularization are updated to a tensor formalism, integrating spatial interaction of matrix elements, thus providing more accurate classification of matrices or tensors such as electroencephalogram or neuroimaging data (134, 139). Interestingly, after training a TensorReg model on 2D SPHARM descriptors, we obtain the 2D matrix of regression coefficients, weighting each SPHARM coefficient at every radius and expansion frequency. This also tells us which SPHARM coefficient was most predictive for each class over the training dataset and for every subject in the population, as shown on **Figure 3.9** for instance in preliminary experiments. From this 2D classification matrix $\boldsymbol{\beta}_{TensorReg}$, we can compute predictions by taking the Frobenius inner scalar product with a given SPHARM descriptor $\boldsymbol{C}_{SPHARM}$ resulting in a risk of pertaining to one of the two classes assessed, between 0 and 1 (135):

$$\langle \boldsymbol{\beta}_{TensorReg}, \boldsymbol{C}_{SPHARM} \rangle_F = \left\langle \begin{bmatrix} \beta_{1,1} & \cdots & \beta_{1,R_{max}} \\ \vdots & \ddots & \vdots \\ \beta_{L_{max},1} & \cdots & \beta_{L_{max},R_{max}} \end{bmatrix} \begin{bmatrix} |c_{1,1}|_2 & \cdots & |c_{1,R_{max}}|_2 \\ \vdots & \ddots & \vdots \\ |c_{L_{max},1}|_2 & \cdots & |c_{L_{max},R_{max}}|_2 \end{bmatrix} \right\rangle_F = \sum_{l,r=1,1}^{L_{max},R_{max}} \beta_{lr} c_{lr} \tag{3.14}$$

Reported diagnostic performance metrics on the training dataset are obtained on cross-validated folds, by taking a 30% balanced bootstrapped sample of the training dataset (1000 bootstrapped samples) (104). Performance on the testing set was assessed by applying the trained TensorReg model on the testing dataset with bootstrapping to produce 95% confidence intervals.

3.4. Novel Segmentation Paradigm: Spherical Volumes

Lesion segmentation is one of the most contentious and critical steps in radiomics workflows since subsequently extracted features rely on these regions (63). Recent studies showed that tumor delineation in radiomics studies had a large impact on the reproducibility of extracted features (142, 143). Hence, the need for expert manual or standardized semi-automatic segmentations is even more important for accurately capturing radiomics-based tumor phenotypes in tumor regions. However, performing these segmentations corresponds to hours of work which would burden the healthcare system and the already overloaded radiologists' workload (14, 15).

Given the spherical morphology generally observed in solid tumors emerging from different body regions, the SPHARM decomposition method appeared to be an interesting and appealing approach to describe structural and textural properties from these tumor regions. We hypothesized that SPHARM descriptors would not be confounded by segmentations variability as much as radiomics and that faster delineation approaches could be developed to segment these tumors. Hence, using our SPHARM pipeline, we were interested in simply creating spherical volumes around lesions to extract imaging descriptors from these spherical VOIs and to compare the accuracy of our method built in regular segmentations to that in spheres. We also aimed to assess the impact on radiomics signatures built in regular segmentations when tested in these new VOIs. As can be seen in results in **Chapter 5** for instance, SPHARM descriptors in these spherical regions performed better than expected, even in the presence of endometrial tumors with morphologies that do not tend to be spherical.

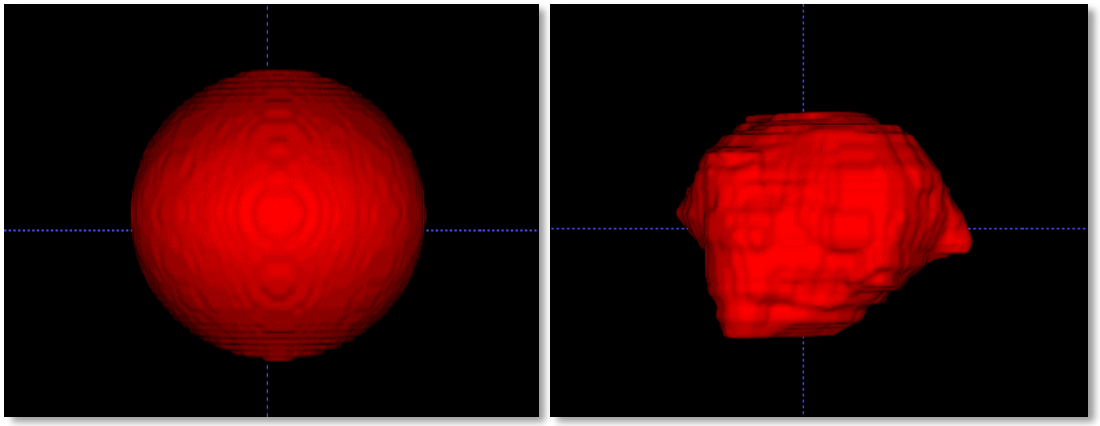


Figure 3.4 Spherical volume of interest extended from an original segmentation on computed tomographic images.

Spherical VOIs are obtained by selecting the center of the lesion on a central slice and to extend a radius on the largest dimension. In cases for which regular segmentations were already performed, the center of mass from these segmented regions were obtained along with the radius of the maximal dimension and a sphere was generated around the original delineation (**Figure 3.4**). To assess the spatial change in VOI in 3D when moving to spherical VOIs, Dice similarity coefficients, also known as the Sørensen-Dice index (144), were used. This metric assesses the number of voxel in the first VOI intersecting with the number of voxel in the other VOI, over the total number of voxels in both VOI, multiplied by 2:

$$I_{Dice} = 2\frac{\text{num}(\text{VOI}_A) \cap \text{num}(\text{VOI}_B)}{\text{num}(\text{VOI}_A) + \text{num}(\text{VOI}_B)} \tag{3.15}$$

A Dice coefficient close to 1 indicates that both binary masks have most of their respective structure spatially overlapping, while a Dice coefficient close to 0 indicates that there is almost no overlap between segmented VOIs.

3.5. Preliminary Experiments with Volumetric Synthetic Textures

Preliminary Experiments Overview

Since previous reports assessed the performance of SPHARM descriptors mainly for characterizing 3D shapes (117, 127), we first explored and adjusted our SPHARM pipeline on a

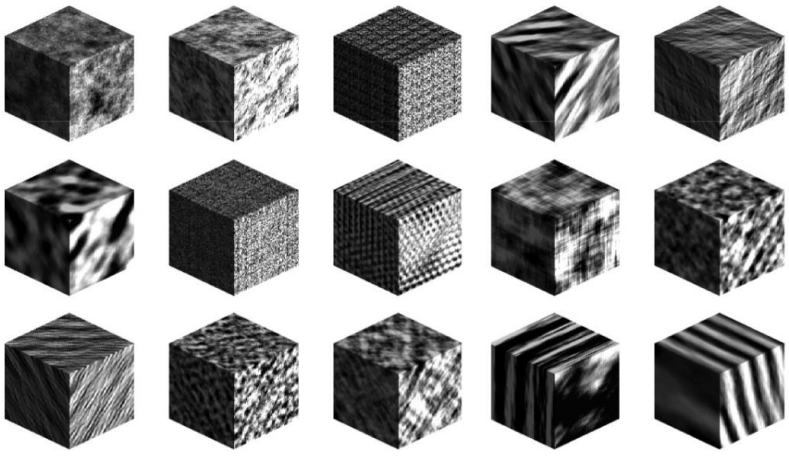


Figure 3.5 Fourier synthetic textures classes from the RFAI dataset from Paulhac L *et al* 2009 (145).

volumetric synthetic texture benchmark dataset, the RFAI (*Reconnaissance de Formes, Analyse d'Images*) database (145). The set of textures used consisted in fifteen classes of 3D patterns of 64^3 voxels in size, with 256 gray-levels, reconstructed from insertions of points in Fourier space, and with ten examples per class, as shown in **Figure 3.5**.

To characterize the ability of SPHARM descriptors to encode these complex textures, we first performed reconstruction tasks on decompositions at different radial and frequency levels and compared them to original 3D objects. The quality of reconstructions was assessed with the MSSI, an index ranging between 0 and 1 and assessing similarity in luminance, contrast, and structure integrated at 5 different resolution scales of the original and reconstructed objects (137, 138). This experiment was repeated on provided noisy versions of the Fourier textures with added Gaussian noise with signal to noise ratio of 10 (145) and compared to the original objects with the MSSI. This enabled us to identify the combination of radial and frequency sampling maximizing mean MSSI and minimizing the MSSI standard deviation across texture classes, while minimizing radial and frequency expansions *e.g.* L_{max} and R_{max} (to obtain a discriminative descriptor of minimal size).

With these SPHARM decomposition parameters, an evaluation of the classification performance of SPHARM descriptors was performed, according to the pipeline developed and discussed in previous sections.

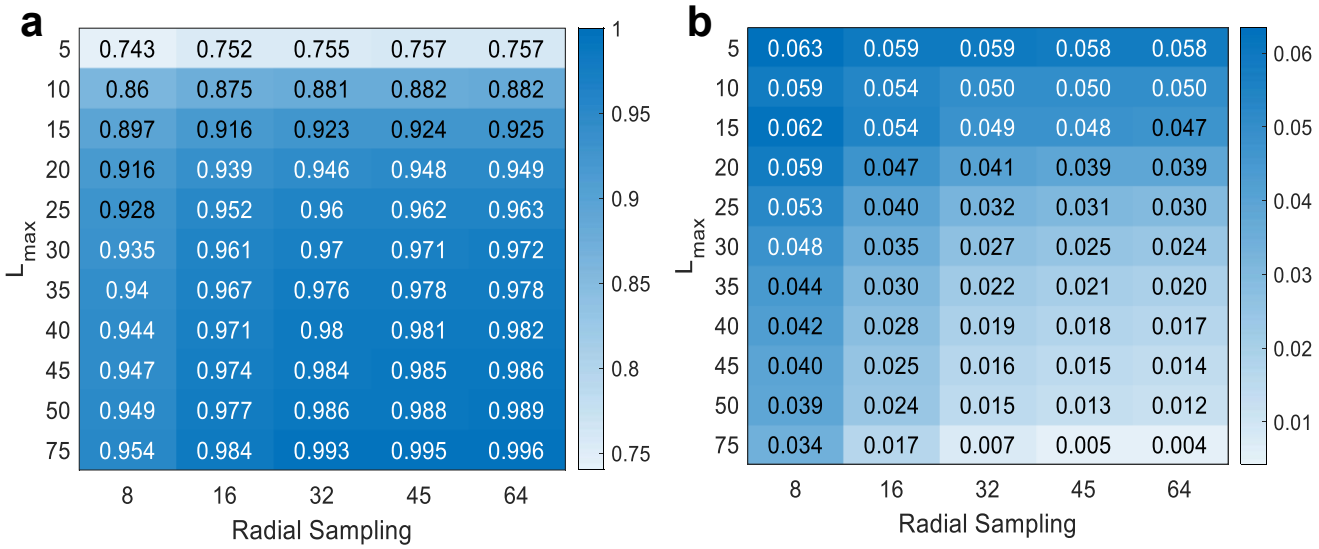


Figure 3.6 (a) Mean and (b) standard deviation of multi-scale structural similarity indexes across different SPHARM expansion and radial sampling in the 3D Fourier texture dataset.

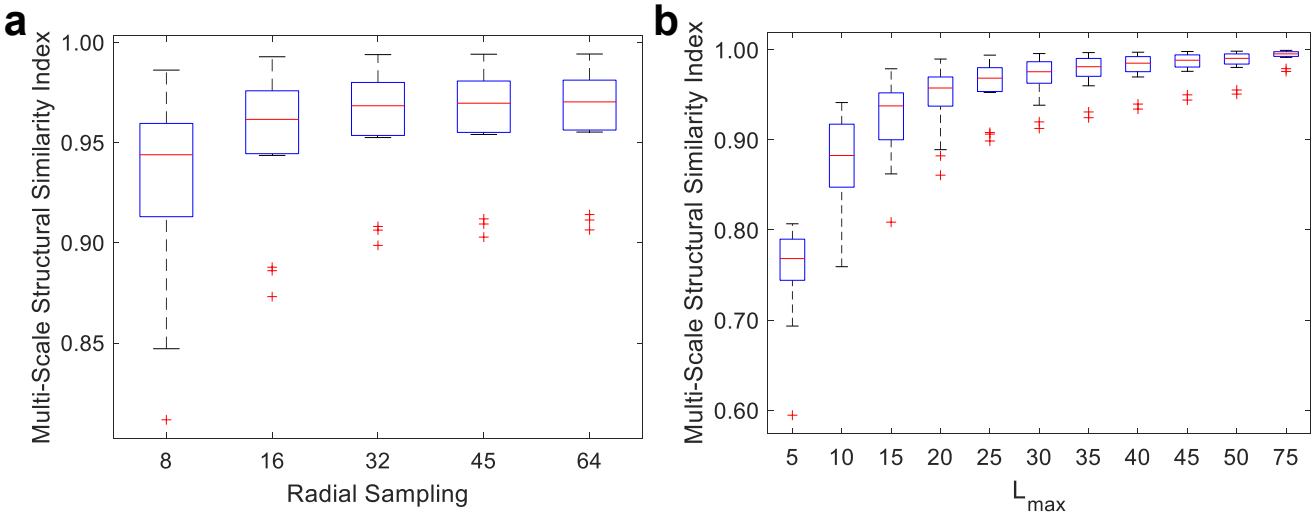


Figure 3.7 Boxplots of multi-scale structural similarity indexes across (a) radial sampling for fixed L_{max} = 25 and (b) SPHARM expansions for fixed radial sampling of 32 in the Fourier dataset.

Preliminary Experiments Results and Discussion

Mean MSSI across all texture classes increased with radial sampling, being at the lowest with 1 sampled shell every 4 voxels (or 8 shells total), and at the highest with 2 sampled shells per voxel (or 64 shells total, which correspond to oversampling). Similarly, MSSI increased with maximal frequency of SPHARM decomposition (L_{max}), being at the lowest with L_{max} = 5 and at the highest with L_{max} = 75 (**Figure 3.6**; **Figure 3.7**). The optimal combination of radial and frequency sampling was found with a radial sampling of 1 shell every voxel (32 shells total) and with frequency decomposition up to L_{max} = 25 (**Figure 3.6**; **Figure 3.7**), since it maximized mean MSSI and minimized the MSSI standard deviation across texture classes, while minimizing L_{max} and R_{max}. Interestingly, the same combination was also identified on another dataset of geometric textures for which results are reported in the **Appendix**. It can be noted on **Figure 3.7** that there is no significant gain in MSSI with radial sampling over 32 shells total and over L_{max} = 25. Examples of a reconstructed volumetric Fourier texture is shown on **Figure 3.8** for different sampling combinations, corresponding to a middle view of the first texture on the left of the one represented on the uppermost right corner of **Figure 3.5**. We can visually observe a significant gain in reconstruction accuracy from radial sampling 8 to 32 and from frequency decomposition 5 to 25, but not from radial sampling 32 to 64 (oversampling) and frequency decomposition 25 to 75. In

fact, complex details are smoothed off the image at low radial and frequency reconstructions, while these complex textures seem adequately captured with the selected optimal decomposition parameters (with 32 shells and with L_{max} = 25; central reconstructed volume on **Figure 3.8**).

The mean MSSI (± standard deviation) for reconstructed volumetric Fourier textures dataset was 0.960 ± 0.032 with 32 sampled shells and with frequency decomposition up to L_{max} = 25. With the same SPHARM decomposition parameters but now reconstructing noisy textures (145), the mean MSSI was 0.946 ± 0.033. Thus, the difference was not significant between reconstructing noisy and normal Fourier textures with this set of SPHARM decomposition parameters. Therefore, we can confirm that this expansion catches just enough textural details through each SPHARM frequency component to model important textural information, without modeling noise. This also indicates that there is a smoothing effect of most important information after SPHARM reconstruction which wipes out noise speckles.

For classification, cross-validation with 3 out of 10 examples was used for measuring validation accuracy after fitting models on the bootstrapped samples of the 7 out of 10 examples from each Fourier texture class (with a total of 15 classes to classify). The validation classification accuracy was very high for SPHARM descriptors using multi-class TensorReg (98% [93%-100%]; **Table 3.1**). The most important harmonics filters of SPHARM descriptors picked up by the TensorReg all modeled high frequency content with $l \geq 6$, with some shown on **Figure 3.9**. The classification 2D matrix also shows which SPHARM coefficients were the most important on **Figure 3.9**. Again, each individual SPHARM descriptor (**Figure 3.9**) corresponds to a relative amount of harmonic filters importance at a given radius and it does not characterize the whole VOI as such. When integrated together, a complete description of the VOI is obtained which enables accurate classification.

The fact that most important coefficients were at higher SPHARM degrees indicates that higher frequency of spherical harmonics basis functions were required to encode rougher textured patterns. However, the characterization of the 3D Fourier texture dataset, seemed overall to require the contribution from most SPHARM coefficients. The fact that the model was highly predictive across texture categories also suggest that SPHARM descriptors could capture relevant information to differentiate complex textures, leveraging the use of comprehensive distributed SPHARM coefficients.

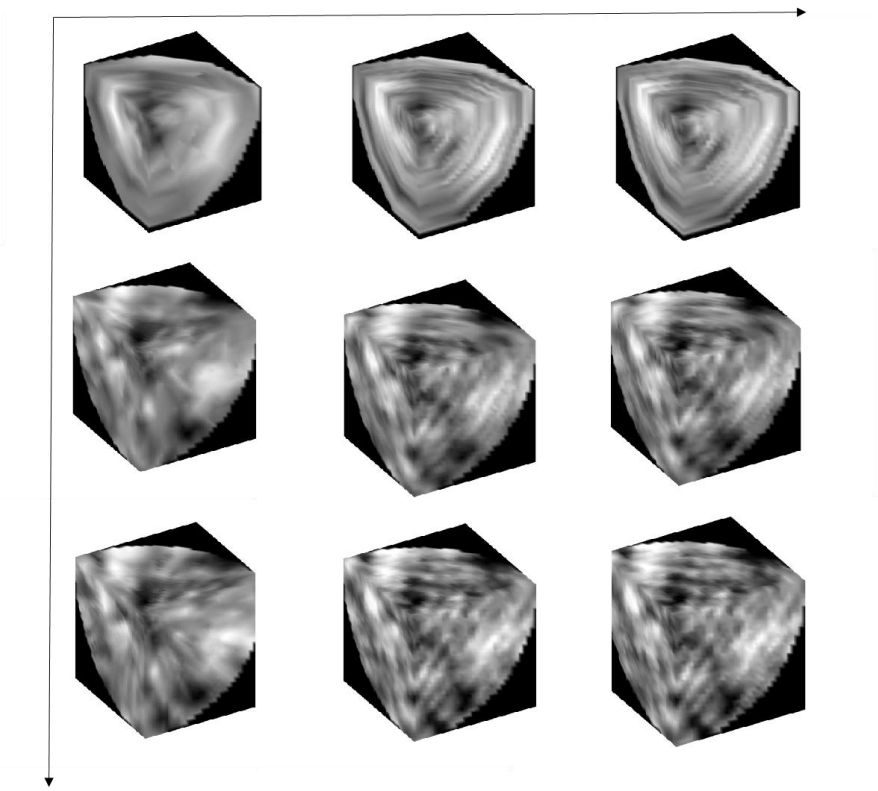


Figure 3.8 SPHARM reconstructions of an example from the first class of the 3D Fourier texture dataset with different radial sampling (8, 32, and 64 from left to right) and SPHARM expansions (5, 25, 50 from top to bottom).

These preliminary results on a synthetic texture dataset suggest that SPHARM descriptors provide accurate encoding of complex volumetric textures. Similar results on the volumetric geometric dataset of the RFAI synthetic textures database can be found in **Appendix I**. Therefore, the proposed SPHARM pipeline was further applied in the two clinical studies detailed in **Chapter 4** and **Chapter 5**.

Table 3.1 Accuracy of SPHARM descriptors for classifying volumetric Fourier textures.

		Accuracy (%)
SPHARM	Training	96 (91–100)
	Validation	98 (93–100)

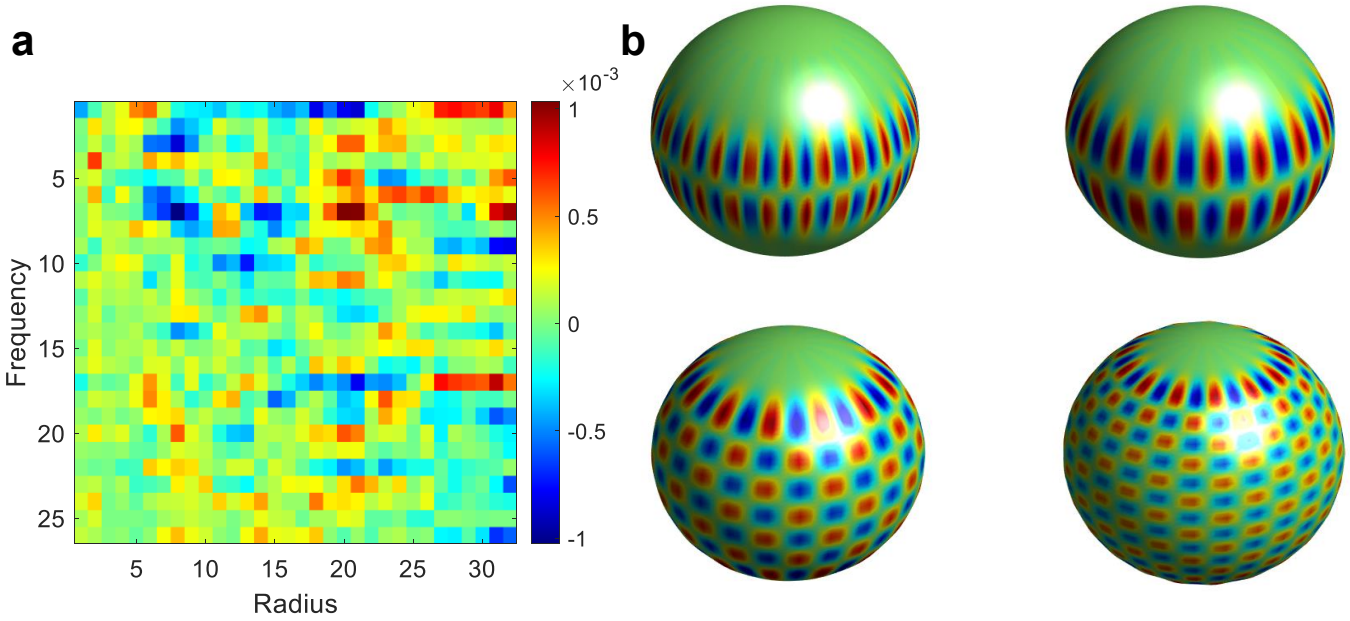


Figure 3.9 (a) TensorReg 2D classification matrix of regularized regression coefficients on SPHARM descriptors of the volumetric Fourier texture dataset. (b) Predictive SPHARM filters with *l*, *m*, *r* = (23,22,21), (17,16,31), (18,12,6), and (26,13,31), respectively.

4. CONTRAST-ENHANCED CT OF RENAL CYSTS

4.1. Background

Management and long-term follow-up of renal cysts is an important burden on the healthcare system in terms of cost and utilization of imaging resources and as a result of unnecessary procedures and related decreased renal function or morbidity (146-150). Fifty percent of adults over 50 years have renal cysts (39, 151, 152). Most cysts appear benign on imaging, but one tenth of renal cell carcinomas are cystic (40). Increased sensitivity is always desirable for predicting malignancy, but false positives may result in incidental morbidity (153). With the associated morbidity in decreased renal function in long-term, strategies sparing nephrons are currently a priority in treating patients with renal lesions (146, 147). Moreover, given the sampling variability of biopsy and the importance of capturing tumoral heterogeneity, quantitative noninvasive techniques assessing the whole tumoral region could improve the grading of renal cysts (27, 28). Hence, validated quantitative imaging-based methods will be required to limit the amount of false positive and accurately stratify renal cysts.

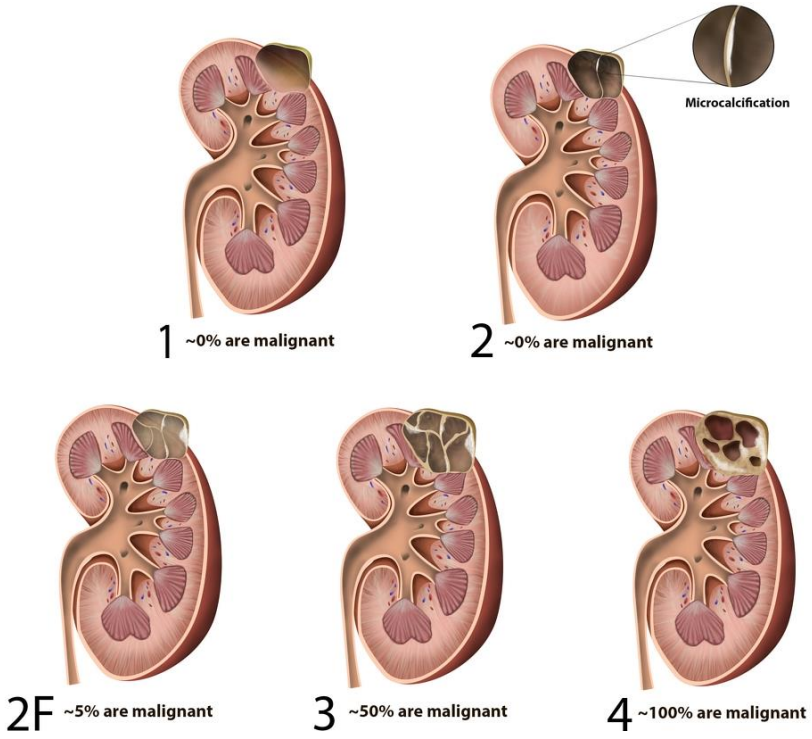


Figure 4.1 Representation of renal cysts' Bosniak categories. Source:radiopaedia.org/articles/bosniak-classification-system-of-renal-cystic-masses. Accessed 2020-07-02.

Qualitative criteria for risk stratification in complex cystic renal lesions (CCRL), known as Bosniak criteria, were introduced over 30 years ago to provide classification guidelines for distinguishing nonsurgical from surgical cystic lesions seen on computed tomographic (CT) images (**Figure 4.1**) (151, 152). More specifically, these are guidelines for radiological assessment of renal cysts based on their morphology as characterized by attenuation, enhancement, calcifications, and septations (39, 40, 154). This CT image-based classification system enables to grade the extent of cysts complexity and malignancy on an ordinal scale ranging from I to IV (39, 151). Lesions that appear benign are typically classified in Bosniak category I or II, while lesions with increased risk of malignancy based on image appearance and potentially subject to surgical removal are classified in Bosniak category III or IV. Over the years, these criteria have helped physicians and radiologists with patient's stratification and have set the basis for predicting malignancy on renal cysts images. After two updates in the late 90's and early 00's, Bosniak criteria now include a fifth category, Bosniak category IIF (follow-up), accounting for numerous misclassifications of benign lesions in category III which should have been classified as Bosniak II according to previous guidelines (**Table 4.1**) (155). According to recent meta-analyses, malignancy rates of renal cysts within these categories are of 3% in Bosniak category I, 6% in Bosniak II, 7% in Bosniak IIF, 55% in Bosniak III, and 91% in Bosniak category IV (153, 156).

For classification of CCRLs in categories I and IV, Bosniak criteria have proven to be highly reliable. However, limitations have been pointed out in numerous previous reports (153, 157, 158), especially for the grading of lesions' complexity in category IIF and III, still including overlapping features leading to bad inter-reader agreement, and complicating the management of surgeries prescriptions. This increases the overall burden of renal cysts on the healthcare system through unnecessary surgeries or follow-up (39).

To improve the specificity of category IIF and III, the 2019 update of Bosniak criteria was introduced with more accurate and discriminative definitions for cysts classification and novel quantitative criteria (**Table 4.2**) (39). These changes have yet to be validated to assess their accuracy for predicting malignancy risk and inter-observer variability. Since previous Bosniak criteria relied mostly on visually assessed anatomical features, the proposition of including non-structural information in the 2019 update, including iodine content and perfusion on contrast-enhanced (CE)-CT, might provide better risk stratification for follow-up and surgery, especially for high-risk categories.

Table 4.1 Bosniak Classification 2008. Adapted from Silverman SG *et al* 2008 (39).

Bosniak class	Classification guidelines
I	-Hairline-thin wall that does not contain septa, calcification or solid components -Water density and attenuation -No enhancement with contrast material
II	Two types: 1. Few hairline-thin septa with or without fine calcification in the wall or septa 2. Uniformly high-attenuation lesions of < 3 cm sharply marginated that do not enhance
IIF	Two types: 1. More hairline-thin septa, minimal enhancement of hairline-thin septum or wall, minimal thickening of septa or wall with or without slightly thickened calcification 2. Intrarenal non-enhancing high-attenuation renal lesions of ≥ 3 cm
III	-Indeterminate cystic masses -Thickened irregular walls or septa in which enhancement can be seen
IV	-Malignant cystic lesions -Enhancing soft-tissue components (*i.e.* nodules)

In this study, we hypothesized that quantitative image analysis using radiomics extracted from CE-CT images and spherical harmonics (SPHARM) decomposition of cystic regions could predict malignancy in CCRLs with high accuracy and with a high degree of reproducibility. Therefore, our aim was to retrospectively determine and compare the diagnostic performance of a radiomics-based random forest model with that of a SPHARM-based regularized tensor regressor (TensorReg) for the differentiation of benign from malignant complex renal cysts using histopathological analyses of resected surgery specimens or biopsy as the reference standard. Our secondary aim was to assess the discriminative ability of both methods in spherical volumes of interest (VOI) compared to semi-automatic segmentations.

4.2. Materials and Methods

This international dual-center retrospective study included patients over 18 years old with complex renal cysts. Institutional review board approval was obtained at both institutions, McGill University Health Centre (Montreal, Canada) and Necker-Enfants Malades Hospital (Paris, France). An institutional review board approval and waiver for informed consent were obtained at both participating institutions. A search for CE-CT studies between January 1998 and December 2018 containing the keywords “renal cyst”, “kidney cyst”, “Bosniak”, “cystic RCC” or “cystic renal cell carcinoma” either in the body or impression of the report, was conducted using the McGill

University Health Centre's Imaging Database (PACS) and the Department of Radiology search engine. The CE-CT report had to indicate the presence of a renal cyst with a minimal diameter of 1 cm, and categorized as Bosniak I, II, IIF, III or IV. Follow-up over 4 years by CE-CT or MRI also had to be available without any changes in the Bosniak classification, in the absence of pathology proof from surgery or biopsy to provide reference standard. CE-CT examinations had to be performed with a renal dedicated protocol including non-enhanced phase, arterial phase, nephrographic phase and delayed phase or including at least a nonenhanced and nephrographic phase between 80 and 120 seconds on CE-CT. More than one cystic lesion could be included for each patient. Patients with cysts nonvisible on CE-CT, with CE-CT images degraded by artifacts, or with any history of interventions on the cyst prior to CT examination were excluded. Finally, if patients had any condition associated with multiple renal cysts, or if a delay in renal enhancement was observed due to obstruction or renal artery stenosis, patients were subsequently excluded. From clinical reports, demographic data, medical history regarding kidneys and pathology reports were collected. Data were deidentified and processed according to European and Canadian Laws. An external independent CE-CT dataset was obtained from Necker-Enfants Malades Hospital for quantitative models testing following the same inclusion and exclusion criteria.

Table 4.2 Bosniak Classification 2019. Adapted from Silverman SG *et al* 2019 (40).

Bosniak class	Computed tomography classification guidelines
I	-Well-defined, thin (≤ 2 mm) smooth wall -Homogeneous simple fluid (-9 to 20 HU) -No septa or calcifications -Wall may enhance
II	Six types, all well-defined with thin (≤ 2 mm) smooth walls 1. Cystic masses with thin (≤ 2 mm) and few (1–3) septa; septa and wall may enhance; may have calcification of any type 2. Homogeneous hyperattenuating (≥ 70 HU) masses at noncontrast CT 3. Homogeneous nonenhancing masses > 20 HU at renal mass protocol CT, may have calcification of any type 4. Homogeneous masses -9 to 20 HU at noncontrast CT 5. Homogeneous masses 21 to 30 HU at portal venous phase CT 6. Homogeneous low-attenuation masses that are too small to characterize
IIF	-Cystic masses with a smooth minimally thickened (3 mm) enhancing wall, or smooth minimal thickening (3 mm) of one or more enhancing septa, or many (≥ 4) smooth thin (≤ 2 mm) enhancing septa
III	-One or more enhancing thick (≥ 4 mm width) or enhancing irregular (displaying ≤ 3-mm obtusely margined convex protrusion[s]) walls or septa
IV	-One or more enhancing nodule(s) (≥ 4-mm convex protrusion with obtuse margins, or a convex protrusion of any size that has acute margins)

Contrast-enhanced Computed Tomographic Examinations

CE-CT scans were obtained using 16-64 channel on VCT Light Speed systems (General Electric [GE] Healthcare) known for having low image noise and good image uniformity (82). All images were acquired at 120 kVp and with variable adapting tube current (mA) to limit image acquisition variability (79). Pitch varied from 0.8 to 1.5. Standard-of-care CT images were reconstructed in the axial plane and with a section thickness ranging from 1.0 to 3.25 mm. Standard iterative backpropagation-based reconstruction with abdominal reconstruction kernel was used (159). During four-phase renal studies, patients underwent unenhanced scanning first. After intravenous injection of 100-150 mL of nonionic iohexol (Omnipaque 350, GE Healthcare) based on the patient's weight with a power injector at a rate of 3 mL per second as per standard-of-care protocol, a bolus tracking algorithm (SmartPrep, GE Medical Systems) was used to determine the onset of imaging of corticomedullary (40-45s), nephrographic (100-120s), and excretory (≈8min) phases. For bolus tracking, a VOI was manually positioned in the thoracoabdominal aorta junction, with a trigger to start acquisitions at 150 HU.

Radiologists' Visual Assessment of Renal Cysts

Qualitative visual assessment of renal cysts on the nephrographic phase of CE-CT images was performed by two trained readers with over 5 years of experience in renal CT imaging (J.D., C.R.). Readers assessed the presence of multiple morphological features as outlined in the 2019 Bosniak classification system and reported Bosniak categories according to the updated system

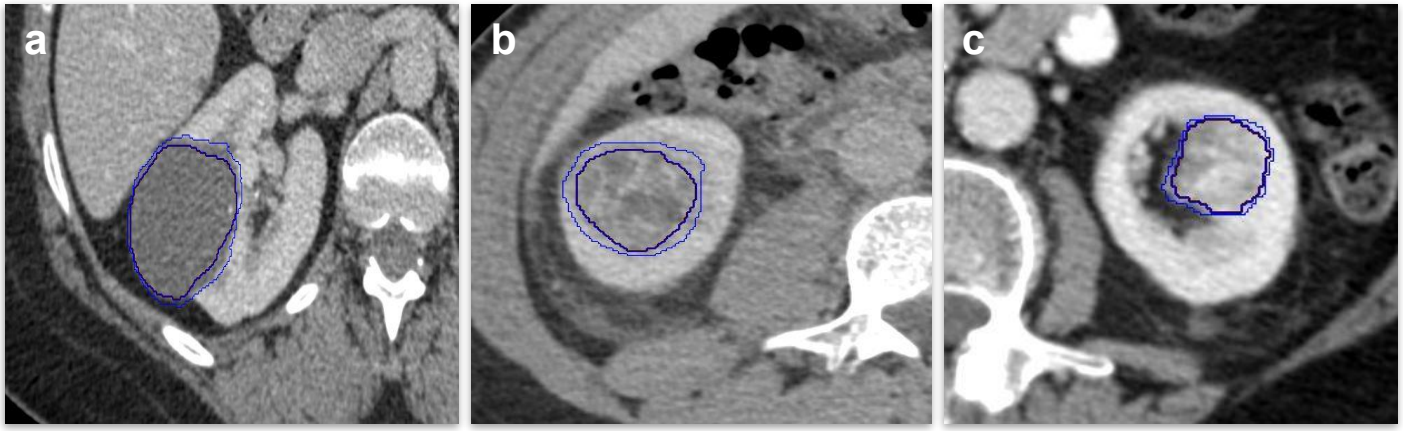


Figure 4.2 CE-CT axial slices of three different cysts with Bosniak category (a) IIF, (b) III, and (c) IV, and their contours: the inner and outer contours.

(**Table 4.2**). Finally, renal cysts were classified as benign or malignant according to pathology reports or according to follow-up of at least 4 years by CE-CT or MRI without any changes or were classified as malignant according to pathology report. Radiologists were blinded to pathology results and imaging follow-up. Pathologists were blinded to radiology results.

Renal Cysts Segmentation

CCRLs were segmented semi-automatically on CE-CT images at the nephrographic phase. The nephrographic phase was defined as the time point at which the cortex and the medulla both enhanced uniformly. Lesions segmentation was performed semi-automatically using a commercial research software (Myrian Intrasense, Montpellier, France) with an implemented algorithm developed at McGill University. The algorithm delineated the outer and inner edges of the entire cyst on every axial section showing the tumor as represented on **Figure 4.2**. After initial automatic segmentations, VOIs were manually corrected around the gross cystic volume by an experienced radiologist in renal CT imaging (J.D.).

Computational Methods

Both computational pipelines described in **Chapter 2** and **Chapter 3** were applied in this clinical study to extract radiomics features and SPHARM descriptors as surrogate biomarkers of malignancy on CE-CT. Prior to radiomics features extraction, images and VOIs were resampled to achieve isotropic voxels and gray levels were discretized. Each feature was calculated with 12 different sets of extraction parameters (4 gray-level bin widths x 3 isotropic voxel sizes). Gray-level bin width sizes were 15, 20, 25, and 30 HU; and resampled voxel sizes were 0.5, 1, and 2 mm^3. Gray-levels resampling in bins of fixed width (absolute gray-level discretization) was selected to increase reproducibility of radiomics features and the set of pre-processing parameters which led to the highest radiomics features reproducibility, based on intraclass correlation coefficients (ICC), was identified. ICCs were obtained by comparing features extracted from three VOIs: inner, outer, and eroded inner VOIs. Inner and outer VOIs were obtained directly from semi-automatic segmentations while the eroded VOI was obtained by removing a one voxel-thick surface from the delineated inner volume. B-spline interpolation was used for image resampling to isotropic voxel size in Pyradiomics (160). After selecting the set of preprocessing parameters

leading to the highest features' reproducibility and robustness to these VOI variations, 106 radiomics features were extracted from both inner and outer segmented VOIs separately (as shown on **Figure 4.2**), to evaluate if including features from outer rims of cysts could provide increased performance. From this set of 212 features, unstable (ICC > 0.80) and multicollinear correlated (Spearman's ρ > 0.95) radiomics features were removed prior to random forest feature selection. All random forest modeling was conducted in Python 3.7.4 using the Scikit-learn machine learning package (109). The 5 most important radiomics features for the diagnosis of malignancy with a random forest model were identified after minimizing trees' depth, number of trees, and maximal features possibly splitting at each node. Training and validation were performed using bootstrapped out-of-bag samples with random undersampling of majority class to balance training dataset (such that $N_{malignant}/N_{benign} = 0.75$). 95% confidence intervals were reported.

SPHARM decomposition was performed by sampling 25 shells at equispaced radii which were decomposed up to a maximal SPHARM degree of $L_{max} = 25$, resulting in 25 x 25 matrices. The ability of SPHARM descriptors to encode and reconstruct renal cysts' volumes was assessed by comparing original volumes to inverse SPHARM-reconstructed volumes with the multi-scale structural similarity index (MSSI). A tensor logistic regressor (TensorReg) was used to fit a classification matrix of SPHARM coefficients for classifying malignancy. All SPHARM and TensorReg analyses were performed in MATLAB (2020a, Mathworks, Natick, MA, USA). Training and validation were performed using stratified 5-fold cross-validation and using bootstrapping of training folds to produce 95% confidence intervals and randomly undersampling majority class to balance training dataset ($N_{malignant}/N_{benign} = 0.75$) while adjusting the regularization parameter to minimize both training and validation error. Both models were independently tested with an external testing dataset from the Necker-Enfants Malades Hospital. Developed models were applied on descriptors extracted from spherical VOIs extended from semi-automatic segmentations. Original segmentations were compared to spherical VOIs with Dice coefficients.

Receiver operating characteristics (ROC) curves were reported based on models' predictions on the training and testing datasets. Thresholds maximizing Youden's index (sensitivity+specificity-1) were identified and associated diagnostic performance metrics of models were reported, *i.e.* sensitivity, specificity, balanced accuracy, positive predictive values, and negative predictive values.

4.3. Results

Population characteristics

From the population of patients fitting inclusion criteria between 2005 and 2018, we included 149 CCRL in the training dataset and 50 CCRL in the testing dataset (**Table 4.3**). Twenty three percent (*n* = 33) of the lesions were malignant in the training dataset, and 32% (*n* = 16) in the testing dataset. All lesions categorized as Bosniak I or II were benign. Malignancy rates in each Bosniak categories were similar to those published in a recent meta-analysis including 35 studies with 2578 lesions (153). In the training dataset, average patient age was 63 years (range: 28-89 years) and 34% (*n* = 46) of patients were women. In the testing dataset, mean age was 59 years (28-80 years) and 30% (*n* = 13) of patients were women. All Bosniak IV lesions were malignant in the external testing dataset.

Table 4.3 Distribution of Bosniak categories and malignancy across the population.

	Bosniak I		**Bosniak II**		**Bosniak IIF**		**Bosniak III**		**Bosniak IV**		**Total**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Total Cysts	32	10	36	9	35	12	23	9	23	10	149	50
Benign Cysts	32 (100%)	10 (100%)	36 (100%)	9 (100%)	32 (91%)	11 (92%)	13 (56%)	4 (44%)	3 (13%)	0 (0%)	115 (77%)	34 (68%)
Malignant Cysts	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (9%)	1 (8%)	10 (44%)	5 (56%)	20 (87%)	10 (100%)	34 (23%)	16 (32%)

4.3.1. Analyses in Segmented Tumors

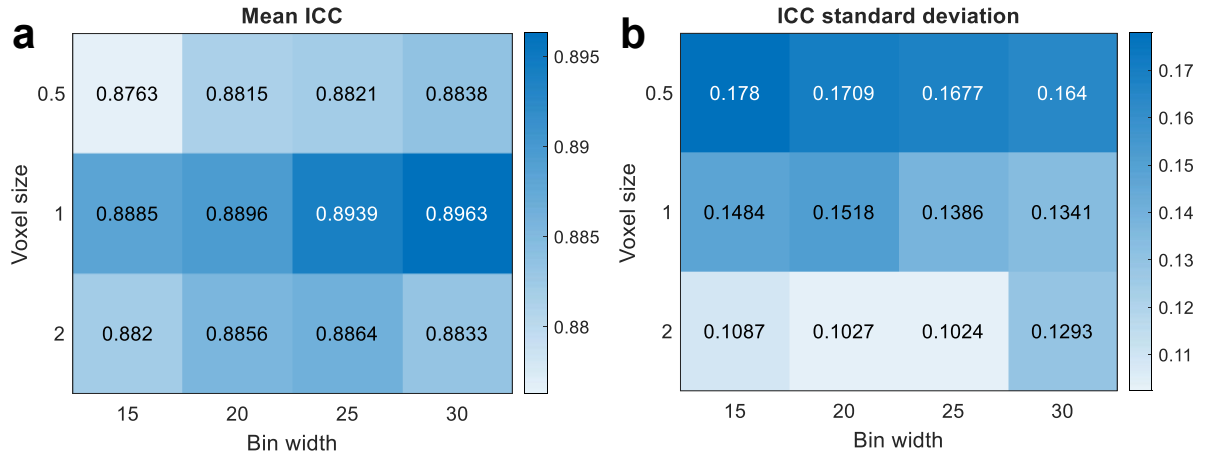


Figure 4.3 (a) Mean and (b) standard deviation of intra-class correlation coefficients for assessing reproducibility of radiomics features evaluated on CT of renal cysts.

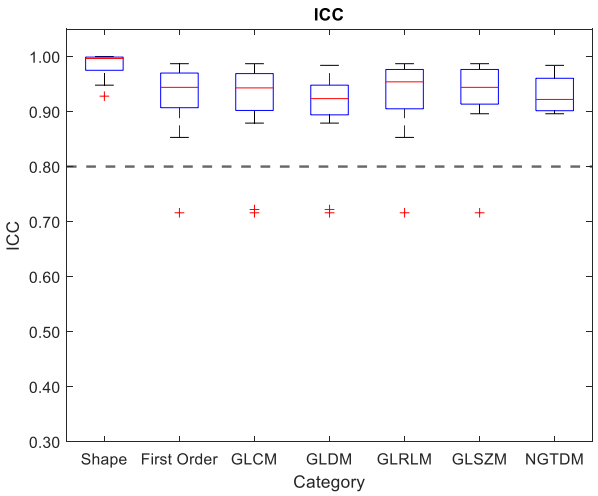


Figure 4.4 Intra-class correlation coefficients for assessing reproducibility of radiomics features for each class extracted from preprocessed CE-CT images with fixed bin width of 30 and resampled isotropic voxels of 1mm^3.

Radiomics Analysis

The reproducibility analysis revealed that the preprocessing parameters which led to the highest overall ICCs were resampled isotropic voxel size of 1 mm^3 and unnormalized absolute discretized CE-CT images with fixed bin size of 30 HU (**Figure 4.3**). Using this set of preprocessing steps, 17 radiomics features were excluded from further analysis since their

associated ICC under VOI variations were below the predefined 0.80 reproducibility threshold as shown on **Figure 4.4**. Among excluded features, 1 was a first-order statistic, none were shape-based features, and 16 were second-order statistics:

- firstorder_Minimum
- glcm_Autocorrelation
- glcm_Idmn
- gldm_HighGrayLevelEmphasis
- gldm_LargeDependenceHighGrayLevelEmphasis
- gldm_LargeDependenceLowGrayLevelEmphasis
- gldm_LowGrayLevelEmphasis
- gldm_SmallDependenceLowGrayLevelEmphasis
- glrlm_LongRunHighGrayLevelEmphasis
- glrlm_LongRunLowGrayLevelEmphasis
- glrlm_LowGrayLevelRunEmphasis
- glrlm_ShortRunLowGrayLevelEmphasis
- glszm_LargeAreaHighGrayLevelEmphasis
- glszm_LargeAreaLowGrayLevelEmphasis
- glszm_LowGrayLevelZoneEmphasis
- glszm_SmallAreaLowGrayLevelEmphasis
- ngtdm_Coarseness

Prior to further feature selection, each reproducible radiomics feature used alone for differentiating benign from malignant renal cysts and those with an area under the ROC curve (AUC) greater than 0.75 in the training dataset were reported (**Table 4.4**). Twelve of these single features were first-order statistics, none were shape-based features, and 13 were texture features. These radiomics features were all evaluated in the inner VOI of the cystic lesions. Interestingly, simple and easily assessed histogram mathematical descriptors such as mean, median, 10^{th} and 90^{th} percentiles, or root mean squared of voxel intensities in the inner VOI led to AUC greater than 0.88. These highly discriminating first-order statistics suggested that simple mathematical analysis of HU distributions in the VOI might lead to improved classification of cystic lesions of the kidneys. After the exclusion of multicollinear correlated features, 112 (112 of 212, 52.8%) features were included in the random forest feature selection. To limit overfitting and to minimize out-of-bag error in the model, the number of trees included in random forest modeling was limited to 20, the depth of each decision tree was limited to 5, and the number of features to possibly split at each

node was limited to the logarithm base 2 of the number features after exploring combinations of these hyperparameters (**Figure 4.6**).

Table 4.4 Most discriminative radiomics features for predicting malignancies on renal cysts contrast-enhanced CT in the training set (AUC ≥ 0.75).

#	**Radiomics features**	**AUC**
1	gldm_DependenceEntropy	0.902
2	firstorder_Median	0.890
3	firstorder_Mean	0.888
4	glcm_Correlation	0.885
5	firstorder_10Percentile	0.883
6	firstorder_90Percentile	0.883
7	firstorder_RootMeanSquared	0.881
8	glcm_Imc1	0.880
9	glcm_SumEntropy	0.835
10	glcm_ClusterTendency	0.833
11	glrlm_RunEntropy	0.823
12	firstorder_TotalEnergy	0.818
13	glszm_ZoneEntropy	0.813
14	firstorder_InterquartileRange	0.811
15	firstorder_RobustMeanAbsoluteDeviation	0.810
16	firstorder_MeanAbsoluteDeviation	0.809
17	glcm_SumSquares	0.806
18	firstorder_Entropy	0.803
19	gldm_GrayLevelVariance	0.801
20	glcm_MCC	0.801
21	firstorder_Variance	0.800
22	glrlm_GrayLevelVariance	0.791
23	glcm_JointEntropy	0.788
24	glcm_ClusterProminence	0.784
25	firstorder_Energy	0.774

The random forest features selection process based on Gini impurity led to the inclusion of a total of 5 radiomics features, over which the diagnostic performance was not significantly increased by the inclusion of more features discriminative. These radiomics features were 3 first-order statistics – median, 10^{th} percentile, and 90^{th} percentile – and 2 texture features – dependence entropy from gray-level dependencies matrices (GLDM) and informational measure of correlation 1 from gray-level co-occurrence matrices (GLCM) – and were all obtained in the inner VOI. In fact, the inclusion of features evaluated in outer rims of the lesions did not impact the diagnostic performance in the final random forest modeling. Interestingly, included features were all in the top 10 most discriminating single features when simply using one feature in ROC analysis (**Table 4.4**). Their distributions against one another are represented in pair plots in **Figure 4.5** where blue dots represent malignant and orange dots represent benign CCRLs.

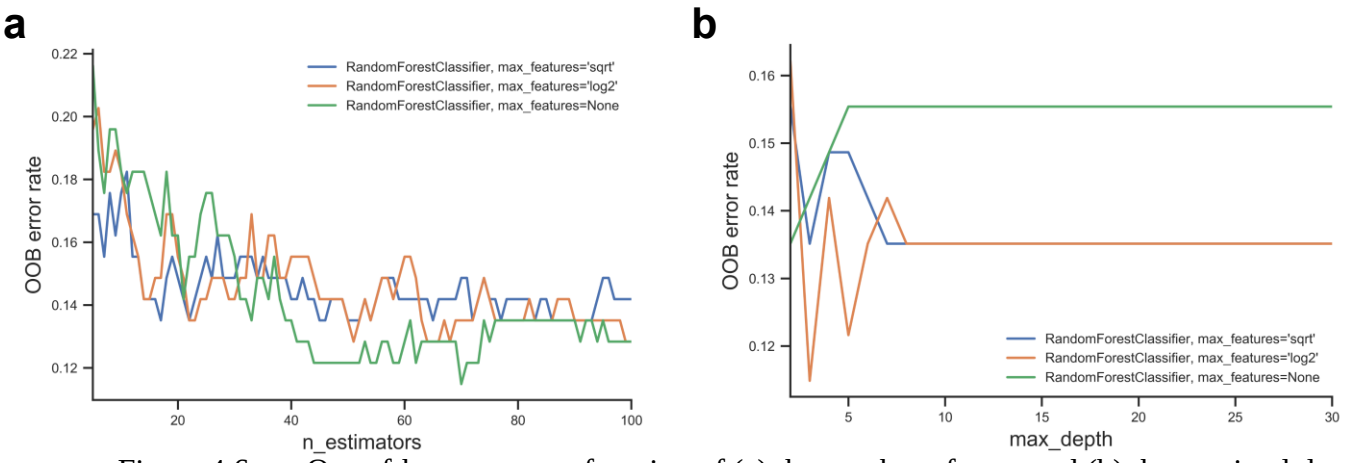


Figure 4.6 Out-of-bag error as a function of (a) the number of trees and (b) the maximal depth of each tree in the random forest. Minimizing these hyperparameters decreases over-fitting.

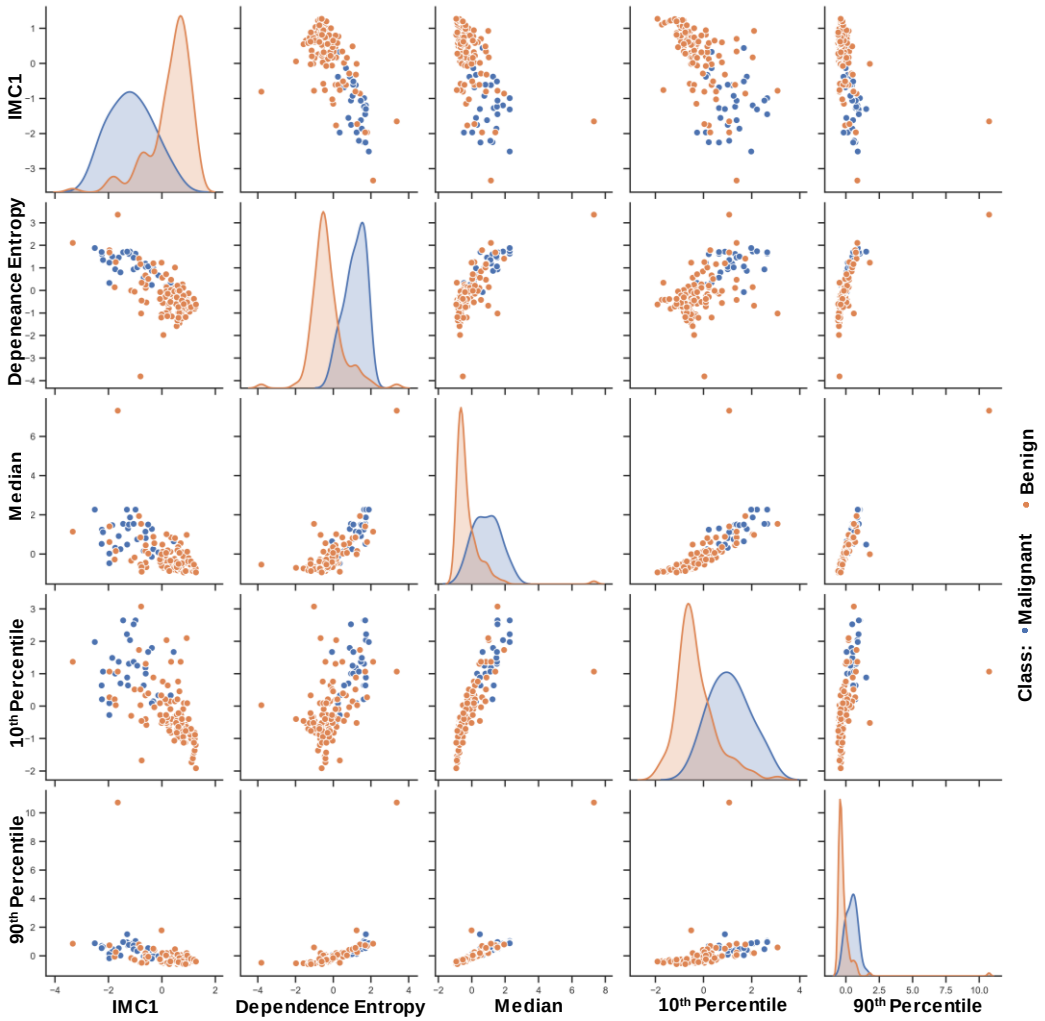


Figure 4.5 Distributions of the 5 most important radiomics features picked up by the random forest model for classifying benign from malignant renal cysts on contrast-enhanced CT.

Bootstrapped training on balanced samples of the training dataset with the random forest classifier resulted in high diagnostic performance for predicting malignancy of CCRL as shown by ROC analysis (**Figure 4.7**; **Table 4.5**). This model was then applied to the external training dataset on which the random forest classifier performed robustly. When using the 5 most reproducible and discriminative features, the random forest model gave consistent diagnostic performance from training to testing with an AUC of 0.91. In the final model, sensitivity, specificity and balanced accuracy were respectively 82%, 94% and 90% when applied on the testing dataset.

Table 4.5 Estimates of diagnostic performance of random forest model for distinguishing benign from malignant renal cysts using 5 radiomics features with 95% confidence intervals.

	AUC	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	**PPV (%)**	**NPV (%)**
Training	0.91 (0.84–0.95)	79 (63–90)	91 (85–96)	89 (82–0.94)	72 (60–84)	94 (88–98)
Testing	0.91 (0.75–0.97)	82 (60–98)	94 (83–100)	90 (79–95)	87 (61–97)	92 (76–98)

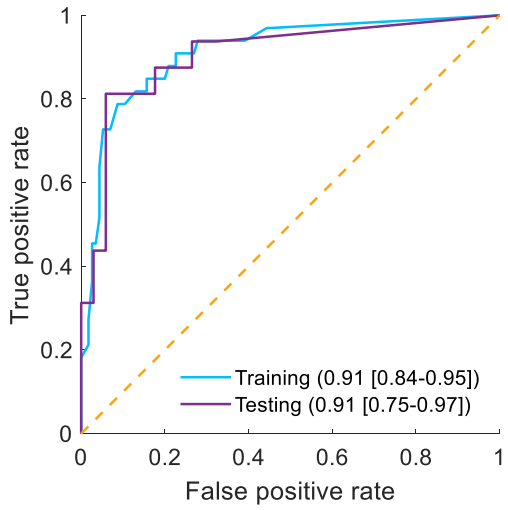


Figure 4.7 Training and testing ROC curves for predicting malignancy in renal cysts with 5 most important features picked up by the random forest classifier on contrast-enhanced CT.

Two-dimensional planes of the random forest classification decision boundaries are represented in **Figure 4.8**, where orange and red are associated with benign cysts and blue with malignant cysts. More specifically, these decision boundaries are obtained from the combination of binary decision trees' classification in the 5D space of selected radiomics features during

training. Since a 5D space cannot be represented by the human eye, we took 2D cross-sections just for the visualization of the training dataset distribution, similar to what was shown in pair plots on **Figure 4.5**, but now including random forest's decision boundaries. In darker red regions, cysts' features are classified as benign with certainty and as malignant in darker blue regions. The overall smooth classification heatmap confirms that overfitting was limited in our random forest modeling.

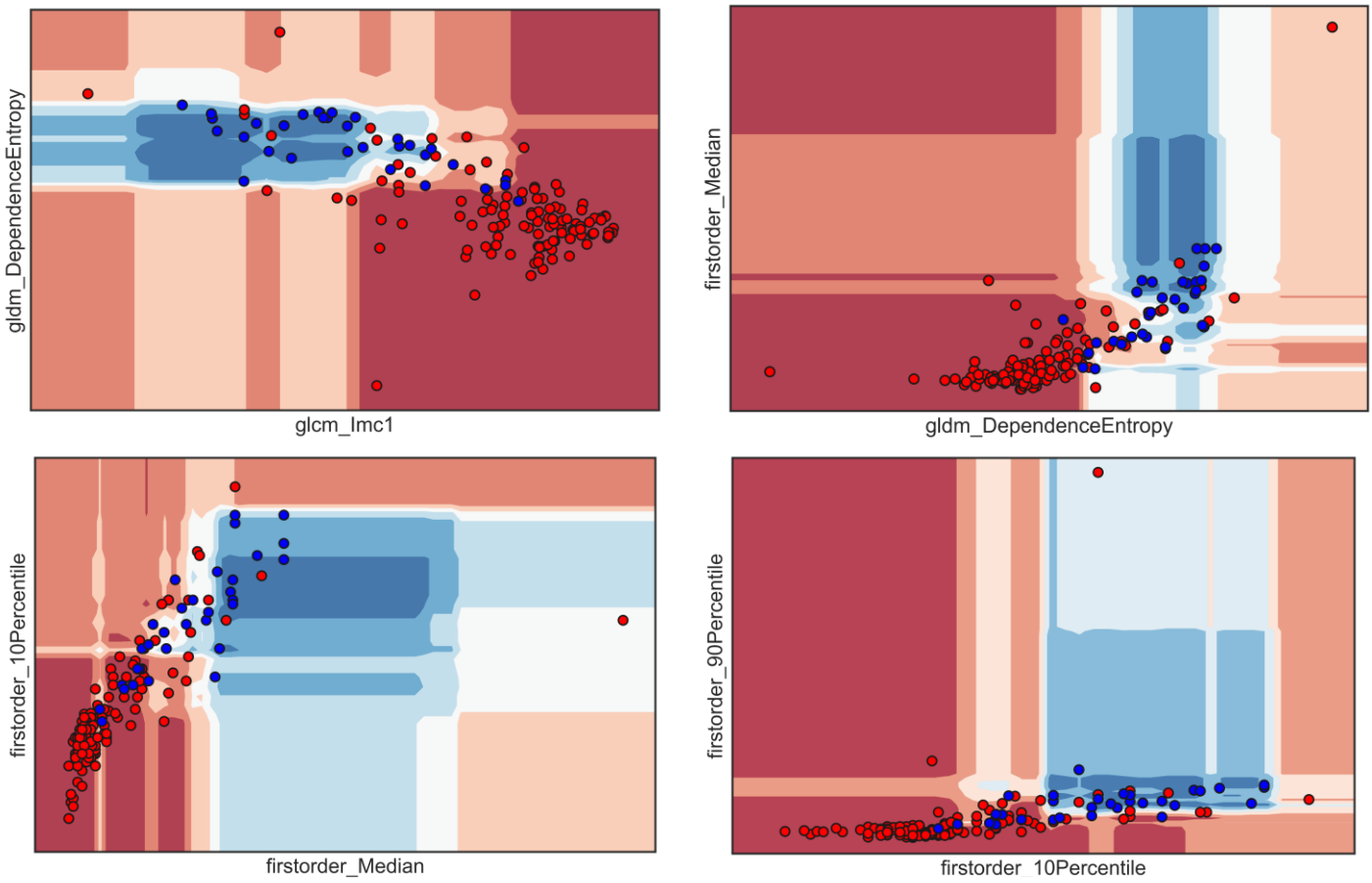
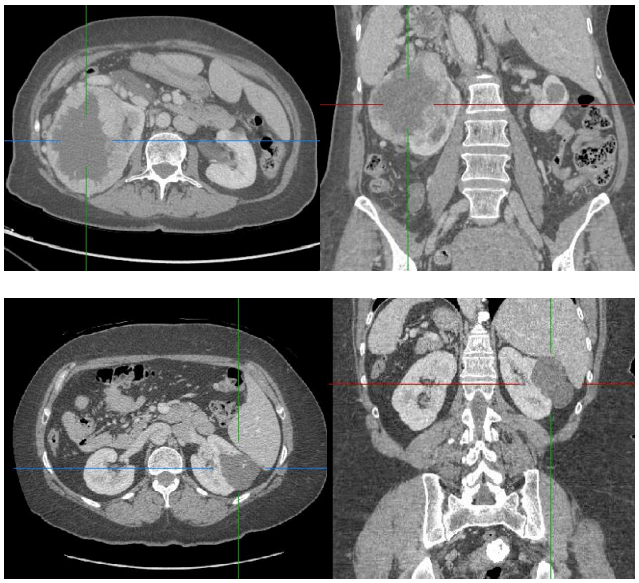


Figure 4.8 Visualization of the random forest classifier's 2D classification planes on the most important features when predicting renal cysts malignancy on CE-CT images. Red decision boundaries are associated with renal cysts classified as benign while blue decision boundaries are associated with cysts classified as malignant.

Based on the final model in the testing dataset, Bosniak I and II cystic lesions were all correctly classified as benign. Benign Bosniak IIF lesions (11 of 12 Bosniak IIF lesions) were all correctly classified with an average risk of 0.11. The only Bosniak IIF malignant lesion in the testing dataset was predicted as benign with a risk of 0.37. Benign Bosniak III lesions (4 of 9

Bosniak III lesions) were all correctly classified with an average risk of 0.17. Three of the five malignant Bosniak III lesions were correctly classified with an average risk of 0.61. The two misclassified Bosniak III lesions had respectively a predicted risk of 0.32 and 0.37. Among Bosniak IV lesions (all malignant), only one lesion was predicted as benign with a risk of 0.21. The average risk of the other Bosniak IV lesions was 0.84.

Examples of well classified cystic lesions with radiomics features are shown in **Figure 4.9**. Notable quantitative features of these cystic lesions on CE-CT were the more enhanced appearance seen in the malignant cyst than in the benign cyst (median [10^{th}-90^{th} percentile], 70 [8-125] HU vs. 16 [-14-47] HU, respectively). For texture features, the informational measure of correlation 1 quantifies the complexity of texture in the image, as assessed by the correlation of intensity probability distributions on GLCMs, between 0 and -1. The dependence entropy quantifies the statistical randomness of neighboring voxels with similar intensities, or of gray-level dependencies, in the VOI (67, 70). More specifically, the former reflects how complex is the texture is, based on the assessment of repeated voxel intensities, while the latter reflects the extent of chaotic textured patterns by assessing if regions of similar or smooth intensity distributions are observed. Hence, the correctly classified benign cyst showed lower texture complexity and less chaotic or random intensity distributions than that of the malignant cyst (IMC1, -0.017 vs. –0.274; dependence



Malignant (Bosniak IV), Predicted Risk = 0.94

Median = 70 HU

10^{th} percentile = 8 HU

90^{th} percentile = 125 HU

Dependence entropy (GLDM) = 6.47

IMC1 (GLCM) = -0.274

Benign (Bosniak IIF), Predicted Risk = 0.04

Median = 16 HU

10^{th} percentile = -14 HU

90^{th} percentile = 47 HU

Dependence entropy (GLDM) = 5.40

IMC1 (GLCM) = -0.017

Figure 4.9 Radiomics features in correctly classified malignant and benign renal cysts with random forests' predicted risks.

entropy, 5.40 vs. 6.47, respectively). This trend in texture features distributions between benign and malignant cysts was also observed in pair plots across the training population (**Figure 4.5**). This indicates that more complex enhancement patterns were seen in the inner segmentations of malignant CCRLs. In fact, increased CE-CT intensity in HU and more complex textured enhancement, which can be related to the development of microcalcifications, were identified as highly discriminative features for predicting malignancy of renal cysts.

SPHARM Analysis

A subpopulation of patients including lesions with minimal radius of 10 mm was analyzed for SPHARM analyses (**Table 4.6**). Therefore, we also reported the diagnostic performance of radiomics features with random forest modeling presented in the previous section to enable comparisons in this subset of the population. This was due to our choice of sampling 25 shells for SPHARM decomposition, which optimized information encoded within descriptors. MSSI of SPHARM reconstructions based on extracted descriptors across the included population was 0.95 ± 0.04 (range: 0.85-0.99) on average, indicating high and robust encoding of original volumetric image content of renal cysts from native nephrographic phase CE-CT images.

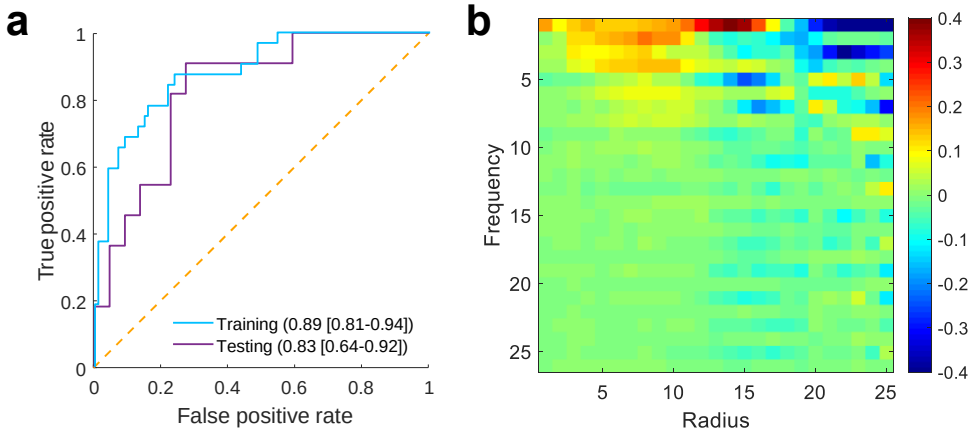


Figure 4.10 (a) Training and testing ROC curves for predicting malignancy in renal cysts with SPHARM descriptors on contrast-enhanced CT, and (b) classification matrix picked up by the TensorReg model.

Table 4.6 Distribution of Bosniak categories and malignancy in a subgroup with renal cysts with minimal radius of 10 mm.

	Bosniak I		**Bosniak II**		**Bosniak IIF**		**Bosniak III**		**Bosniak IV**		**Total**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Total Cysts	32	4	32	8	33	12	23	9	23	10	133	43
Benign Cysts	32 (100%)	4 (100%)	32 (100%)	8 (100%)	32 (97%)	11 (92%)	13 (56%)	4 (44%)	3 (13%)	0 (0%)	101 (76%)	27 (63%)
Malignant Cysts	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (3%)	1 (8%)	10 (44%)	5 (56%)	20 (87%)	10 (100%)	32 (24%)	16 (37%)

Cross-validated training with bootstrapping on balanced samples of the training dataset with a TensorReg model resulted in high diagnostic performance for predicting malignancy of CCRL as shown by ROC analysis (**Figure 4.10**; **Table 4.7**). This model was then applied to the external testing dataset on which the resulting AUC was of 0.83 with TensorReg modeling. In the final model, sensitivity, specificity and balanced accuracy were respectively 91%, 73% and 81% in the testing dataset. The Lasso-regularized TensorReg estimator (2D classification matrix) used for classification is shown on **Figure 4.10**.

Table 4.7 Detailed diagnostic performance of SPHARM descriptors and radiomics features for predicting malignancy of renal cysts on CE-CT.

	AUC	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	**PPV (%)**	**NPV (%)**
SPHARM, Training	0.89 (0.81–0.94)	88 (67–96)	85 (75–93)	86 (77–92)	75 (58–87)	93 (84–98)
SPHARM, Testing	0.83 (0.64–0.92)	91 (67–100)	73 (55–88)	81 (64–91)	65 (33–83)	94 (65–100)
Radiomics, Training	0.92 (0.83–1.00)	85 (71–98)	87 (73–99)	86 (66–1.00)	65 (61–69)	95 (91–99)
Radiomics, Testing	0.92 (0.78–0.99)	88 (60–98)	92 (82–100)	90 (82–98)	82 (61–97)	94 (81–100)

On this 2D classification matrix, positive regression coefficients were used for predicting malignancy while negative coefficients were predictors for benign lesions. Lower frequency content located in the cysts' center seemed to capture contrast material enhancement related to malignancy. On the opposite, lower frequency content on the wall of the CCRL appeared to relate to benign characteristics of these lesions, as captured by SPHARM descriptors on the training dataset. The frequency content describing the center of renal cysts in SPHARM descriptors must

have been associated with the development of microcalcifications and enhanced nodules as represented in **Figure 4.1** and **Figure 4.2.** These are both characteristics of malignancy (**Table 4.1**; **Table 4.2**). Enhancement on the wall, on the opposite, is a characteristic of low grade benign cysts, which seemed to be captured by SPHARM descriptors on outer shells, as shown by negative blue regression coefficients of low frequency ($l = 1$).

On **Figure 4.11** and **Figure 4.12**, two examples of correctly classified malignant and benign cysts, respectively, are represented with slices through their 3D SPHARM reconstructions, and representative harmonic filters for predicting malignancy are also reported. A large nodule in the first represented malignant cyst is observed and was captured by harmonic filters (yellow-orange overlay) on **Figure 4.11**. In fact, recalling the analogy to convolutional neural networks, SPHARM's harmonic filters enhanced regions of the VOI which led to increased prediction accuracy during the training process. This cyst was accurately classified as malignant. On the opposite, enhancement in the wall is observed on **Figure 4.12** and seemed also to be captured by harmonic filters (blue overlay) and correctly classified as benign.

Based on the final SPHARM model in the subset of the testing dataset, Bosniak I and II cystic lesions were all correctly classified as benign. Seven of the benign Bosniak IIF lesions (11 of 12 Bosniak IIF lesions) were correctly classified with an average risk of 0.31. The four other benign Bosniak IIF CCRL that were misclassified had predicted average risk of 0.59. The only Bosniak IIF malignant lesion in the testing dataset was correctly predicted as malignant with a risk of 0.72. Two of the benign Bosniak III lesions (4 of 9 Bosniak III lesions) were correctly classified with an average risk of 0.32. The two other benign Bosniak III CCRLs that were misclassified had predicted average risk of 0.61. All five malignant Bosniak III lesions were correctly classified with an average risk of 0.72. Among Bosniak IV lesions (all malignant), only one lesion was predicted as benign with a risk of 0.41 (the same than for the radiomics model). The average risk for other Bosniak IV malignant lesions was 0.89.

Also, removing lesions with radius smaller than 10 mm slightly increased the diagnostic performance of radiomics for predicting malignancy (testing AUC of 0.92 vs. 0.91). Thus, radiomics-based random forest modeling was more accurate for characterizing malignancy than SPHARM descriptors (testing AUC of 0.92 vs. 0.83; testing balanced accuracy of 0.90 vs. 0.81).

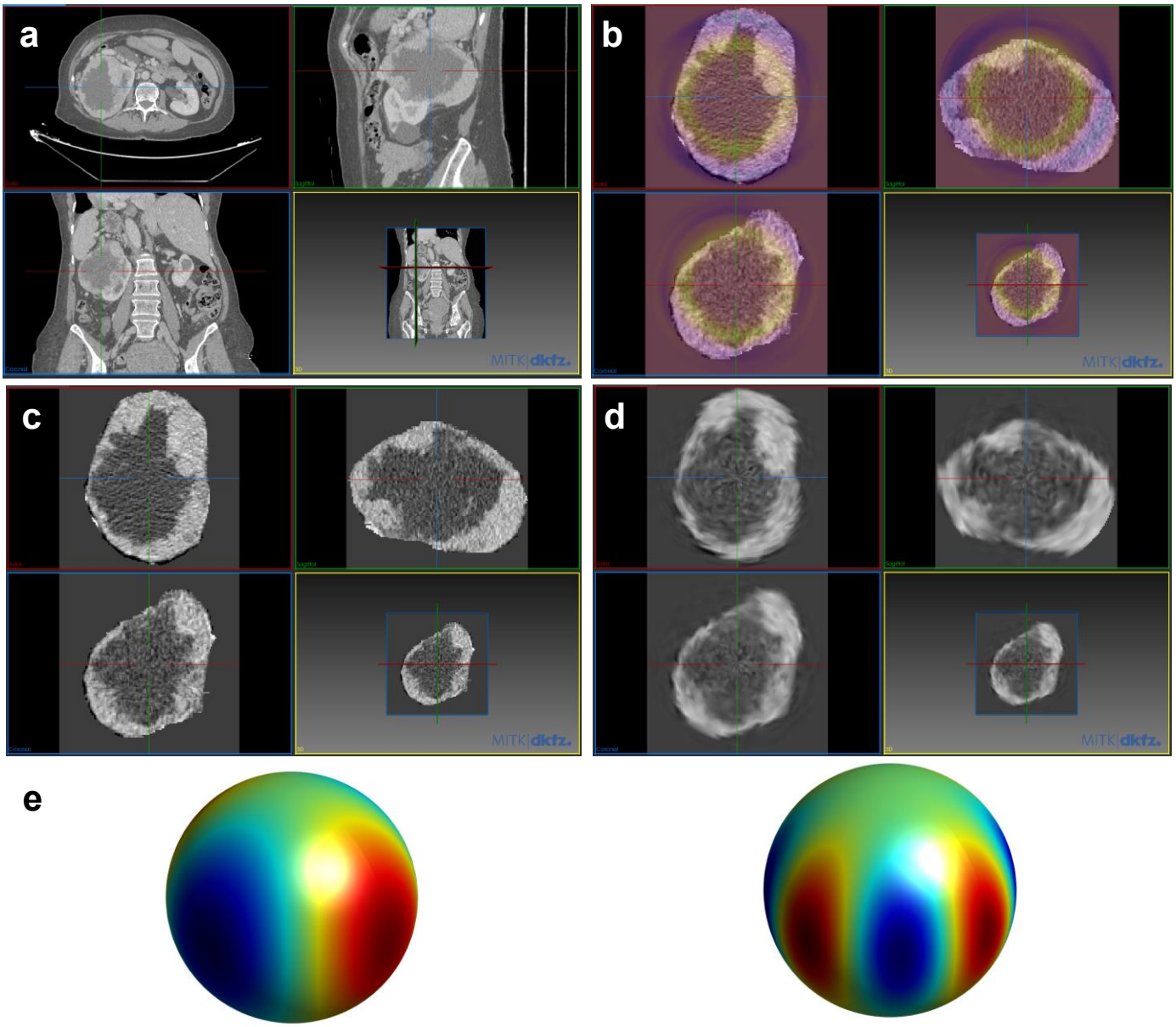


Figure 4.11 (a) Contrast-enhanced CT of a malignant renal cyst (Bosniak IV), (b) segmented tumor with overlayed full reconstructed classification matrix of SPHARM descriptor, (c) segmented CT image and its (d) reconstruction with SPHARM coefficients (L_{max}, R_{max} = 25, 25), and (e) harmonic filters (l, m, r = 2, 2, 7) and (l, m, r = 4, 4, 9), expressed within the spherical yellow overlay in panel (b) as predictors of malignancy. Predicted risk from the tensor logistic regression model was 0.89.

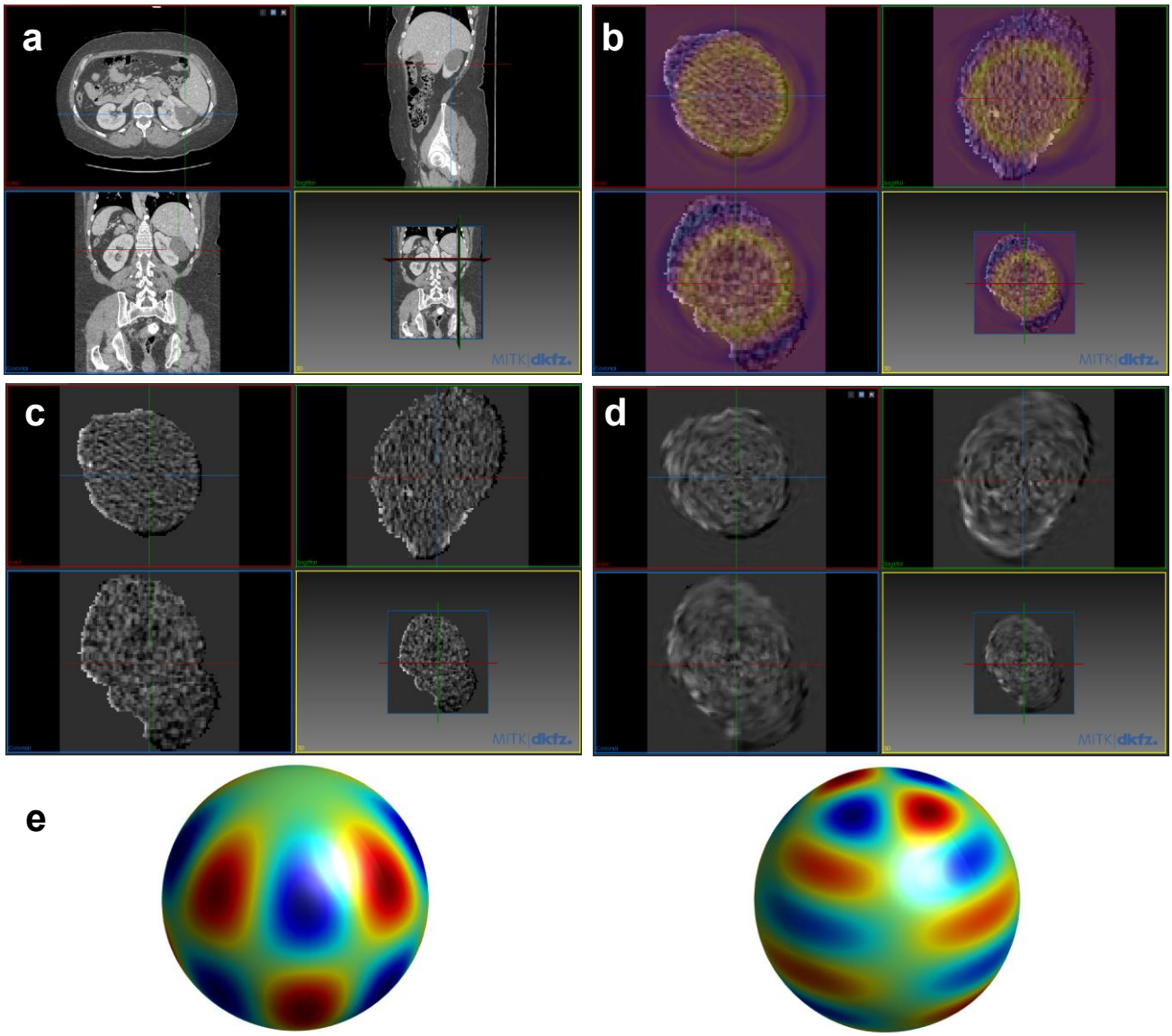


Figure 4.12 (a) Contrast-enhanced CT of a benign renal cyst (Bosniak IIF), (b) segmented tumor with overlayed full reconstructed classification matrix of SPHARM descriptor, (c) segmented CT image and its (d) reconstruction with SPHARM coefficients (L_{max}, R_{max} = 25, 25), and (e) harmonic filters (l, m, r = 5, 4, 16) and (l, m, r = 7, 2, 18), expressed within the blue overlay in panel (b) as predictors of benign category. Predicted risk from the tensor logistic regression model was 0.22.

4.3.2. Analyses in Spherical Volumes of Interest

Radiomics and SPHARM Analyses

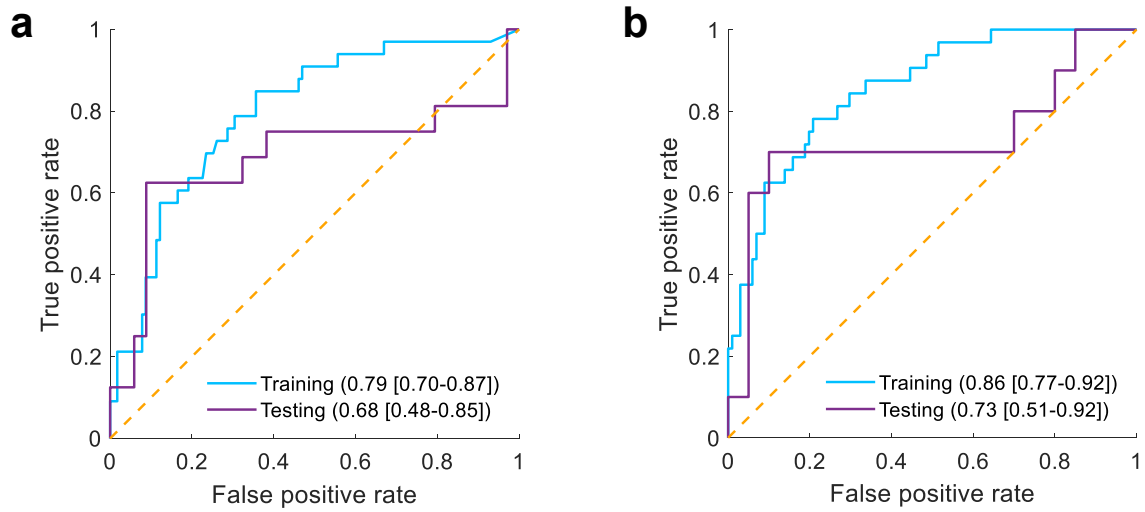


Figure 4.13 Training and testing ROC curves for differentiating benign from malignant renal cystic lesions in spherical volumes of interest with (a) the 5 most important radiomics features and (b) SPHARM descriptors.

When applying our pipelines in spherical VOIs, a significant decrease in diagnostic performance was observed for both methods. The average Dice coefficient between segmented VOIs and spherical VOIs was 0.76 ± 0.12 (range: 0.54-0.91) across the population, indicating significant changes in VOI. A bigger decrease in AUC was observed for radiomics features extracted from these spherical VOIs, going from 0.92 to 0.68 in AUC on the testing dataset, while that of SPHARM coefficients decreased to a lesser extent from 0.83 to 0.73 (**Figure 4.13**). This decrease in performance observed for radiomics features meant that the 3 first-order statistics features – median, 10^{th} percentile, and 90^{th} percentile – and the 2 texture features – dependence entropy (GLDM) and informational measure of correlation 1 (GLCM) – changed nonlinearly from segmented VOIs to spherical VOIs. In fact, including surrounding tissue such as bowels could induce significant changes in histogram statistics, especially if air is found in bowel areas (-1000 HU). Similarly, for texture features, the interrelationships of voxel intensities when including more image content from larger volumes will change, even though included features were selected for their stability when subject to such VOIs changes. A similar decrease in performance was seen for SPHARM descriptors. This could be direct results that discriminative information captured in cysts walls must have changed significantly on spherical VOIs, now including surrounding tissue.

4.4. Summary of Findings and Discussion

Summary of Findings

The radiomics-based random forest and the SPHARM decomposition-based TensorReg models achieved high diagnostic performance in distinguishing benign from malignant CCRLs. The radiomics model was more accurate for characterizing malignancy than the SPHARM model. Simple descriptive statistics (median, 10^{th} and 90^{th} percentiles) combined with one co-occurrence and one dependence of gray-level matrices texture features provided the high and robust diagnostic accuracy observed for radiomics features in this study. In fact, overall higher enhancement values and more complex and chaotic textures were directly associated to cysts' malignancy in the developed radiomics signature. The SPHARM decomposition matrices seemed to capture the presence of enhancing nodules and microcalcifications in the inner part of the cysts as predictors of malignancy, and enhancement of outer wall of the cysts as a predictor for benign cysts. This localization property of SPHARM descriptors enabled us to analyze the underlying information captured by the decomposition coefficients of the SPHARM method. While both SPHARM and spherical VOIs explored in this study underperformed compared to radiomics in semi-automated segmentations, experiments in the next chapter (**Chapter 5**) show cases where there are advantages to using SPHARM and also spherical VOIs.

Reproducibility

Stringent inclusion criteria on CE-CT protocols were used to limit the inter-scan variability of acquired CE-CT images. Our inclusion of CE-CT scans with standard-of-care variable tube current (mA) automatically adapting detector exposure, but with fixated tube kilovoltage (kVp) might explain the good radiomics features reproducibility observed in this study. In fact, varying mA was previously found to have little impact on radiomics features' stability while having fixated kVp was associated with good features reproducibility (79). Only 17 radiomics features were excluded from analyses. Almost all excluded features were textural features (16 of 17, 94%), as reported in previous studies (30, 73). Selection of optimized preprocessing steps did increase features reproducibility extracted from CE-CT images at the nephrographic phase. For instance, up-sampling CE-CT images to a isotropic size of 0.5 x 0.5 x 0.5 mm^3 with bin width of 15 HU

would have resulted in a smaller number of included features, going from average ICC (± standard deviation) of 0.896 ± 0.134 to 0.876 ± 0.178 (**Figure 4.3**). The significance of this difference also lies in the standard deviations of these ICCs since the increase in standard deviation indicates that more features were below the inclusion threshold, going from 17 to 27 excluded features. This shows that comprehensive analyses of preprocessing parameters are required and should be reported for all radiomics studies, since selecting default settings of a given platform will not lead to optimal feature reproducibility. Unreproducible features should always be reported and the inclusion of all implemented radiomics features of the employed platform as a default should be avoided. It also supports that care should be taken when designing radiomics studies and defining inclusion criteria of CT acquisition parameters. Including scans at different kVp or with different reconstruction kernels might not provide in the end comparable radiomics features (30, 76-78).

For SPHARM decompositions, the pipeline was developed according to previously reported methods on volumetric shapes (117, 127) and according to our preliminary results on volumetric texture benchmark datasets. The absence of need for image preprocessing makes SPHARM an interesting approach in the context of harmonization of imaging data analysis. These invariance properties emerge as a direct result of the modeling of intensities distribution on shells based on spherical harmonics functions. By evaluating the L2-norm of SPHARM coefficients along order *m* for each degree *l*, SPHARM descriptors represent energies at each frequency band and radius, hence become comparable across varying image intensities from which they are extracted. SPHARM descriptors were also optimized to capture relevant image information without modeling noise in the image, as shown by reconstructions in preliminary analyses on 3D textures and by the high MSSI obtained for SPHARM descriptors-based reconstructions of renal cysts. However, further studies including different CT acquisition parameters should be performed to assess if SPHARM descriptors are significantly confounded by the use of different kVp or reconstruction kernels, since image contrast and physical signal acquired change when such parameters vary (30, 76-78).

Radiomics and SPHARM Analyses

We did not expect any added clinical value from these analyses in Bosniak categories I, II, or IV since Bosniak classification is highly associated with benignity for Bosniak categories I and

II and malignancy for Bosniak category IV. However, it is interesting to note that both models showed high diagnostic performance in these categories. In fact, in the testing dataset, all benign CCRL were correctly predicted in Bosniak categories I and II. For the radiomics model, 12 of the 16 malignant CCRL were correctly predicted, while 15 of the 16 malignant CCRL were correctly predicted by the SPHARM model, although with increased false positive rate. Interestingly, the four incorrectly classified malignant lesions by the random forest were low grade tumors compared to other malignant lesions. The radiomics model was more accurate than the SPHARM model and appeared to be more specific, while SPHARM modeling was more sensitive. Although increased sensitivity is desired in the prediction of malignancy, associated increase in false positives may result in incidental morbidity (153). The current clinical specificity of radiologists using the Bosniak classification system was evaluated at 74% (95% confidence interval: 64%-82%) in a recent meta-analysis including 2578 lesions (153). Thus, since high accuracy and increased specificity compared to that of radiologists (94% [83-100]) were obtained through radiomics-based random forest modeling, implementation of such a computational framework could provide improved decision-making for clinicians in the stratification of patients for surgery. One of the strengths of this radiomics-based model was also that most important features were first-order features (global distribution of pixel values within the tumor) within the core VOI. They also reflect radiological features proposed in the 2019 Bosniak classification update, as low grade benign cysts tend to be hyperattenuating and homogeneous while high grade malignant cysts tend to show enhanced textured appearance due to the development of cysts' nodules and microcalcifications (**Figure 4.1**; **Table 4.2**) (39). Moreover, the consistent diagnostic performance from training to testing on an external dataset showed the high robustness of the radiomics pipeline and of the random forest model developed in this study. Robustness is also achieved through the choice of feature since 10^{th} and 90^{th} percentiles are more stable under VOI changes for instance than maximum or minimum values which could pick up random fluctuations of high or low HU intensities. After optimized preprocessing of images, z-score normalization of features before random forest modeling also enabled us to provide comparable features across institutions (102, 103).

Our interpretation of SPHARM descriptors was highly associated to clinical features of malignancy as seen on CE-CT images, *i.e.* textured enhancement patterns in cysts' nodules and microcalcifications, and of features observed in benign renal cysts, *i.e.* wall enhancement with

minimal texture and no enhancement in cysts' center. Thus, it even related to radiological signatures proposed in the 2019 Bosniak classification update (**Table 4.2**), similar to what was reported for the proposed radiomics signature.

Analyses in Spherical Volumes of Interest

The significant drop in AUC observed for radiomics features in the diagnosis of malignancy in spherical VOIs supported the need for expert precise manual or semi-automated segmentations to capture tumoral characteristics (142, 143). However, VOI variations going from original segmentations to spherical VOIs were significantly greater than that made in VOI variations during the radiomics feature selection process, as indicated by poor Dice coefficients. Thus, most features might not have been deemed reproducible under such changes in VOIs. Because of the spherical topology observed in renal cysts, we expected that the added material outside of accurately segmented tumors, but within spherical VOIs, would not have altered SPHARM descriptors in a way that would have drastically affected its discriminatory power. However, the significant decrease in performance also observed for the SPHARM decomposition method does support the need for accurate delineation of tumor volumes for adapted characterization of cysts' malignancy. However, as detailed in **Chapter 5**, there are situations in which this hypothesis holds, indicating that accurate predictive models do not always need precise time-consuming segmentations. On the other hand, since we simply repeated the pipelines developed for segmented tumors to features extracted from spherical VOIs, it was expected that the classification performance would change. Conceptually, the underlying pathophysiological information captured in the original signal from the core of segmented cysts have changed by including signal unknown to the developed models coming from surrounding tissues included in spherical VOIs. Thus, using the same classification models initially fitted to signal coming solely from within the cysts might not reflect the radiological signature associated with the change in segmentation. However, we also wanted our methods to capture information that was solely associated with tumoral information and not with surrounding organs. Therefore, pipelines and random forest models for both methods were not built again from the beginning after extracting descriptors from these novel spherical VOIs.

Limitations

Our study had the following limitations. First, we did not analyze the diagnostic performance of both pipelines applied on CE-CT images at the available non-enhanced, corticomedullary, or excretory delayed phases. Radiomics features or SPHARM descriptors extracted from these images or dynamic analyses of how extracted features evolved across phases might provide relevant underlying information on pathological processes (161). On the other hand, since it is at the nephrographic phase of CE-CT that relevant enhancement patterns are seen and used for visual classification by radiologists, using features extracted from nephrographic phase was more consistent with clinical standard of care. Second, in the current version of our work, diagnostic performances of radiologists' classification using the Bosniak system 2008 and its 2019 Silverman's update were not compared to that of histopathological analyses or against one another. To fully assess if our models can provide support to radiologists, and improve specificity and performance, we will need these comparisons. This was a secondary aim of the study in preparation which will be included in the manuscript for publication. In fact, our team aimed at comparing the diagnostic performance of the 2008 Bosniak diagnostic classification system with the proposed 2019 Silverman's updated Bosniak classification guidelines and to evaluate the interobserver variability of these two radiological classification systems. We also aimed to define the most accurate and reproducible visually assessable qualitative set of CE-CT features to predict malignancy of Bosniak cysts using the 2019 Bosniak classification and to propose a new classification for renal cysts in the widely used American College of Radiology Data System. These goals were separate from my contributions but will be included in the final version of the manuscript to be submitted for journal publication. Third, SPHARM descriptors analyses were not performed on the whole study population. In fact, small lesions made of few voxels cannot be sampled by up to 25 shells. This was a limitation of the SPHARM method, which is also observed in radiomics texture analyses. For texture features extraction, a voxel neighborhood is required to assess mathematical inter-relationships of voxel intensities in a VOI. A recent study suggested that the minimal VOI size for texture features to be evaluated was 1000 mm^3 (*e.g.* 10 x 10 x 10 mm^3) (162). In our study, all cysts had at least a diameter of 10 mm. Thus, all cysts respected this limit and this issue did not arise in radiomics analyses. However, slightly increased performance was still observed in the subpopulation with cysts of minimal radius of 10 mm for radiomics features. Future studies will need to assess the minimal VOI size needed for SPHARM decomposition, given

preselected radial and frequency expansions. Finally, we did not report comparisons of SPHARM descriptors for different radial and frequency decompositions. Instead, the frequency decomposition selection was guided by preliminary analyses on 3D textures and by previous work on 3D shapes (17, 127), which all consistently indicated that L_{max} = 25 resulted in optimized VOI encoding for reconstruction and classification. Similarly, radial sampling every voxel on a radius extending from tumor center was found to be accurate for reconstruction and classification in preliminary analyses. Thus, given that renal cysts of category IIF and increasing have minimal radii of 15 mm (156), 30 shells would perfectly sample cysts imaged with 0.5 x 0.5 mm^2 in-plane resolution on CE-CT, with one shell every voxel on the radius. To provide an intermediate radial sampling to account for the distribution of cysts' sizes, a radial sampling of 25 shells was found to be optimal for SPHARM-based reconstruction and classification. This was shown by high MSSI between inverse SPHARM-reconstructed VOIs compared to original VOIs and by good classification performance of SPHARM descriptors in the training and testing datasets.

Discussion

We anticipate that quantitative image analysis of renal cysts based on CE-CT images will help nephrologists to optimize patient management. Proposed methods could lead to better malignancy risk stratification of CCRLs in order to reduce unnecessary surgery or follow-up (false positive Bosniak IIF to IV cysts). In addition, quantitative image analysis could help with the detection of missed malignancies (false negative Bosniak I and II cysts). Since CE-CT is commonly used for the detection of enhancement within renal lesions, this work is clinically translatable and could help with clinical decision-making and avoiding unnecessary surgeries due to cysts misclassification.

This work could form the basis of automated renal cyst risk stratification and potentially become a physician-assisting tool that can be deployed at imaging workstations. Bosniak classification is well correlated with the increase of malignancy risk but is limited for grading lesion complexity using the IIF and III categories, resulting in unnecessary surgeries or follow-up (39). Our models may help avoiding these procedures which add to the burden on the healthcare system. Combining SPHARM decomposition with radiomics in future work could harness the

discriminatory ability of radiomics found in this study to the localization ability of SPHARM descriptors providing robust interpretable modeling of malignancy in CCRLs.

In conclusion, proposed models have achieved high diagnostic performance in distinguishing benign from malignant complex cystic renal lesions. Future studies should reassess patient management based on Bosniak classification and on the risk of malignancy predicted by our models, especially the radiomics-based classification which led to the highest diagnostic performance with increased specificity compared to that of radiologists as reported in the literature. Extracting descriptors from spherical VOIs decreased significantly this performance which suggest that computational methods – at least in the context of CE-CT of CCRL – should extract information from signal coming from the tumor itself, and not roughly including surrounding tissues. Finally, this study relies on its robustness through the development of a state-of-the-art radiomics pipeline using the well-established open-source Pyradiomics platform, and of a SPHARM pipeline with interesting invariance properties, which could both help to produce more reliable and reproductible analyses in the future. Including genomic profiles of lesions in the modeling could further allow to predict malignancy and differentiate disease subtypes, extending this work to so-called radiogenomics (163, 164).

5. MULTI-PARAMETRIC MRI OF ENDOMETRIAL TUMORS

5.1. Background

In developed countries, endometrial cancer is currently the most common gynecological malignancy and the fourth most common cancer in women (165). After diagnosis, the mainstay of treatment is total hysterectomy with bilateral salpingo-oophorectomy (166-170). Treatment may include surgical removal of pelvic and para-aortic lymph nodes, adjuvant chemotherapy, or radiotherapy (171, 172). Over the years, the International Federation of Gynecology and Obstetrics (FIGO) staging system has shown high prognostic value for determining risk groups in endometrial cancer (169, 173, 174). Prognostic risk factors include histologic grade, intra-tumoral heterogeneity, presence of myometrial invasion (MI), and lymphovascular space invasion (LVSI), visually assessed by trained radiologists on medical images or by pathologists on histology specimens (166-169, 175). To provide a comprehensive staging of these factors, state-of-the-art histopathological analyses are currently performed in clinical settings as the reference standard either on preoperative biopsies or on the surgical specimen (170, 176-179). For preliminary screening and risk stratification, preoperative biopsies tend to underestimate of tumor grade (176, 177). Given the high sampling variability of biopsy and the importance of capturing endometrial intra-tumoral heterogeneity (12), noninvasive quantitative whole tumor assessment could provide comprehensive staging of the depth of MI, the presence of LVSI, and of tumor grade (27, 28). Hence, there is a need for noninvasive assessment of histopathological prognostic risk factors in endometrial cancer which could help with patients' stratification for surgeries.

Many meta-analyses have shown the diagnostic value of magnetic resonance imaging (MRI) for assessing histopathological features of endometrial cancer, especially deep MI (a representation of an invasion of the myometrium is seen on **Figure 5.1**) (42, 180-183). Since the depth of MI is known to be the most important morphologic prognostic factor for endometrial cancer and MRI's high soft tissue contrast enables to differentiate the extent of MI, MR examinations have became essential in preoperative staging (41, 184). Current recommendations for staging endometrial carcinoma on imaging are mostly qualitative and do not provide comprehensive guidelines for stratifying patients based on all histopathological features.

Therefore, the development of comprehensive noninvasive diagnostic tools based on MRI for preoperative risk stratification is required to provide reliable and reproducible methods for determining tumors' stage and aggressiveness.

To address this urgent need, many studies in the last decade have assessed the capacity of radiomics features extracted from single sequence or multi-parametric MRI (mpMRI) to diagnose three main outcomes *i.e.* high grade, deep myometrial invasion, and lymphovascular space invasion (18, 50-57). In these studies, expert segmentations were required to delimit the extent of endometrial tumors on MR images from many sequences which correspond to hours of work burdening the already overloaded radiologists' workload (14, 15).

Since high intra-tumoral heterogeneity is generally associated with tumor's inherent aggressiveness (185), radiomics texture features used in various cross-sectional imaging modalities have been studied as a detection, diagnosis, prognosis, characterization, and response assessment method for various tumor types (185-192). Our group previously proposed a mathematical model using 2D MRI-based texture features for risk assessment in endometrial cancer (18). In this previous study, the proposed model was highly associated with deep MI and could distinguish low grade from high grade tumors. Interestingly, the model achieved similar performance to that of subspecialty radiologists in the assessment of deep MI. One could argue that the absence of an external testing dataset to validate these results cannot show the robustness of 2D radiomics analysis method in this previous study. Therefore, the main goal of our study was to build upon

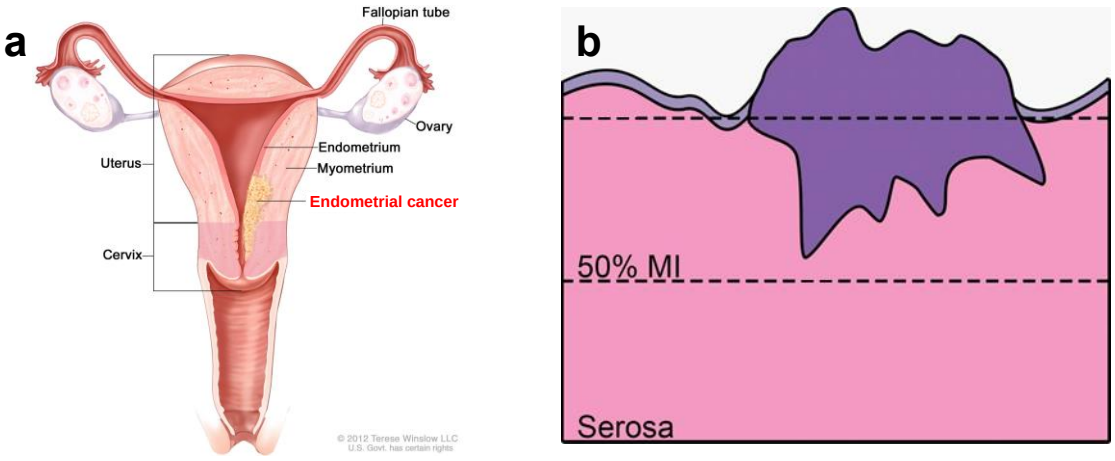


Figure 5.1 (a) Anatomical representation of endometrial cancer and (b) schema of myometrial invasion, where pale purple represents the endometrium, dark purple is the lesion, and pink is the myometrium. Adapted from National Cancer Institute 2012 and van der Putten LJM 2017 (179).

previous work and retrospectively evaluate the utility of 3D mpMRI-based radiomics features and of spherical harmonics (SPHARM) descriptors in endometrial cancer to predict deep MI and high grade endometrial tumors using histology as the reference standard. Our secondary aim was to assess the ability of both methods to perform in spherical VOIs extended from expert manual segmentations.

Table 5.1 2009 revised International Federation of Gynecology and Obstetrics (FIGO) staging for endometrial cancer. Adapted from Creasman W 2009 (193, 194).

FIGO	Pathological changes	MRI findings
Stage I	Tumor confined to the uterus	-Normal or thickened endometrial stripe with diffuse or focal abnormal signal intensity -Intact junctional zone with smooth endometrial-myometrial interface
IA	No or less than half myometrial invasion.	-Signal intensity of tumor extends into myometrium <50% -Partial thickness disruption of junctional zone with irregular endometrial-myometrial interface
IB	Invasion equal to or more than half of the myometrium.	-Signal intensity of tumor extends into myometrium ≥50% -Partial or full thickness disruption of junctional zone with irregular endometrial-myometrial interface
Stage II	Tumor invades cervix but does not extend beyond uterus	-Internal and endocervical canal are widened -Disruption of enhancing endocervical mucosa
Stage III	Tumor extends beyond uterus but not outside the pelvis	
IIIA	Invasion of serosa, adnexa, or positive peritoneal cytology	-Disruption of continuity of outer myometrium -Irregular uterine configuration
IIIB	Vaginal and/or parametrial involvement	-Segmental loss of hypointense vaginal wall
IIIC	Metastases to pelvic and/or para-aortic lymph nodes	-Regional lymph nodes larger than 1cm in diameter
Stage IV	Tumor extends outside of the pelvis or invades bladder or rectal mucosa	
IVA	Invasion of bladder or rectal mucosa	-Tumor signal disrupts normal tissue planes with loss of low signal intensity of bladder or rectal wall
IVB	Distant metastases (includes intra-abdominal or inguinal lymph nodes)	-Tumor masses in distant organs or anatomic sites

5.2. Materials and Methods

Study Population

This international dual-center retrospective study included patients over 18 years old with endometrial lesions. An institutional review board approval and waiver for informed consent were obtained at both participating institutions, McGill University Health Centre (Montreal, Canada) and Hôpital Lariboisière, Assistance Publique-Hôpitaux (Paris, France). MRI and clinical data of

patients who underwent an MR examination with pelvic protocol before surgery between January 2011 and July 2015 were included. All included female patients had histology-proven endometrial cancer as assessed by surgical staging. Patients were excluded if the lesion measured less than 1cm in diameter or if it was nonvisible on MRI; if the MRI examination or the pathologic report was incomplete; if they had any history of neoadjuvant therapy prior to surgery; or if patients had other malignancies. Data from included patients were deidentified and processed according to European and Canadian Laws. Patients from the first institution from which originated this study were included in the training dataset while patients from the second institutions formed the external testing dataset.

Magnetic Resonance Imaging Examinations

All MRI studies were performed on 1.5 T MRI scanner (training dataset: Signa Excite; General Electric [GE] Healthcare, Waukesha, WI, USA; testing dataset: Magnetom Avanto, Siemens Healthcare, Erlangen, Germany) using the vendor specific phased-array pelvic surface coils. Fasted patients were scanned in supine position and administered intramuscularly 40mg of hyoscine butyl bromide (Buscopan, Boehringer, Ingelheim, Germany) prior to acquisition to decrease peristalsis. MR examinations included the following standard-of-care diagnostic sequences for pelvic protocols: fast spin echo T2-weighted imaging, echo planar imaging diffusion-weighted imaging (DWI) at b = 0 and 1000 s/mm^2, and 3D gradient echo T1-weighted dynamic contrast-enhanced (DCE)-MRI (**Figure 5.2**; **Table 5.2**). Administration of 0.1 mmol of a gadolinium contrast agent (Gadovist, Bayer, Leverkusen, Germany) per kg of body weight was performed intravenously prior to DCE-MRI. DCE-MRI acquisitions were performed at pre-contrast in sagittal and axial oblique planes, at post-contrast at 25 seconds, 60 seconds, and 120 seconds in the sagittal plane, and at delayed phase 240 seconds in the axial oblique plane). The axial oblique plane was positioned perpendicularly to the endometrial cavity (*i.e.* short axis view).

Histopathological Analysis

Histopathological sampling of surgical specimen was obtained for all patients. Analyses of specimens were used to determine tumor subtype, histopathological grade, and the presence of deep (≥ 50%) MI. The low grade category included FIGO grades 1 and 2, while high grade category

included FIGO grade 3 and non-endometrioid subtypes. The percentage of MI depth was assessed as a function of the penetration depth of the tumor between the endometrium and the serosa, where 50% was considered at the middle of the myometrium (**Figure 5.1**). Pathologists were blinded to imaging results and radiologists were blinded to pathology results.

Table 5.2 MR examinations acquisition parameters.

	MR Sequence	Acquisition plane	TR/TE (msec)	Acquisition matrix	Field of view (cm)	Slice thickness (mm)
T2w MRI	FSE	axial, sagittal, coronal, oblique axial*	4000-4575/100	512×256	24	4
DWI	EPI	sagittal, oblique axial*	5000/69	128×256	32	6
DCE-MRI	3D GRE	sagittal, oblique axial*	3.6/1.75	320×192	26	4

MRI Segmentation

Segmentations of 3D endometrial tumor volume of interest (VOI) were manually drawn in consensus by two experienced radiologists in pelvic MRI on image volumes from 6 different MR sequences: T2-weighted MRI, DWI (at b = 1000 s/mm^2), apparent diffusion coefficient (ADC) maps generated from the voxel-wise combination of b = 0 s/mm^2 and 1000 s/mm^2 (as ADC = $-\frac{1}{b}\ln(\mathrm{DWI}_{b=1000\mathrm{s/mm^2}}/\mathrm{DWI}_{b=0\mathrm{s/mm^2}})$, and second phase (at 60s [DCE2]), third phase (120s [DCE3]), and delayed phase (240s after injection, often referred to as post-gadolinium [PostGado]) of DCE-MRI. Pre-contrast and first phase (25s) DCE-MRI were not included analyses due to insufficient contrast between the endometrial lesion and the surrounding myometrium. VOIs were delineated avoiding peripheral borders of lesions to avoid including adjacent myometrium or healthy endometrium. For DWI and ADC, the plane between sagittal and oblique axial with best quality was used. Spherical VOIs were extended from these original manual segmentations based on the largest radius of each tumor.

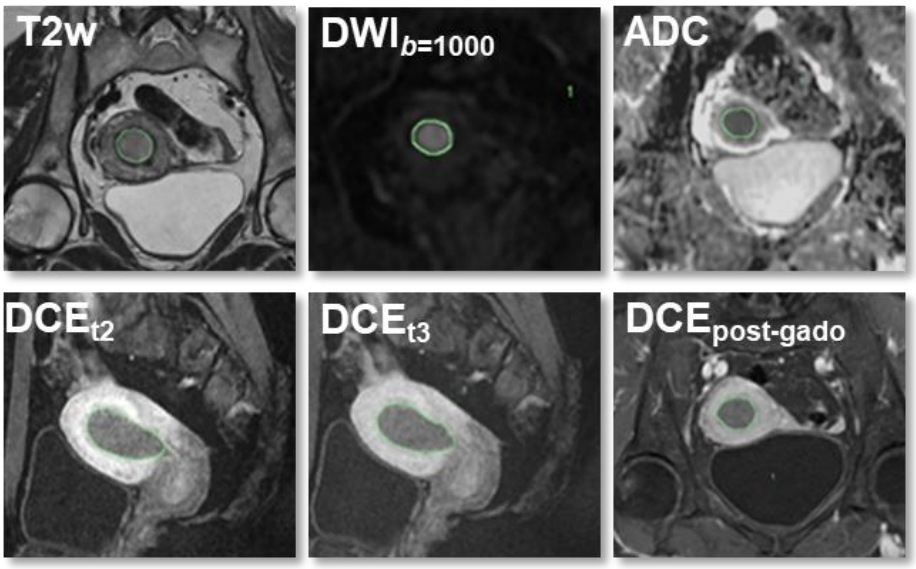


Figure 5.2 Segmented endometrial tumor on the 6 MR contrasts included.

Computational Methods

Both computational pipelines described in **Chapter 2** and **Chapter 3** were applied in this clinical study to extract radiomics features and SPHARM descriptors as surrogate biomarkers of histopathological features of endometrial cancer. Prior to radiomics features extraction, images and VOIs obtained in each MR sequence were resampled to achieve isotropic voxels and gray levels were discretized. Each feature was calculated with 16 different sets of extraction parameters (4 gray-level bin width sizes x 4 isotropic voxel sizes). Gray-level bin width sizes were 15, 20, 25, and 30; and resampled voxel sizes were 0.5, 1, 2, and 3 mm^3. Fixed bin size (absolute gray-level discretization) was used to increase reproducibility of radiomics features which are known to be confounded by gray-level discretization on MR images (88). Thus, as discussed in **Chapter 2**, we had to perform image intensity normalization prior to absolute gray-level discretization. Pyradiomics-integrated normalization was used to rescale the mean of voxel intensities across the whole image at 300 and their standard deviation at 100, similar to what recent studies recommended (46, 86, 88-94). B-spline interpolation was used for image resampling to isotropic voxel size in Pyradiomics (160). The set of preprocessing steps leading to the highest radiomics features reproducibility was then assessed based on intraclass correlation coefficients (ICC) of radiomics features under VOI variations. ICC were obtained by comparing features across three VOIs: original, dilated, and eroded VOIs. Dilated and eroded VOIs were obtained from the original

manual segmentation by adding or removing a one voxel-thick surface, respectively. Unstable (ICC < 0.80) and multicollinear correlated radiomics features (Spearman's $\rho > 0.95$) were removed prior to random forest feature selection. All modeling was conducted in Python 3.7.4 using the Scikit-learn package (109). The 5 most important radiomics features across all MR sequences for the diagnosis of each histopathological feature were identified with a random forest model after minimizing trees' depth, number of trees, and maximal features possibly splitting at each node. Training and validation were performed using bootstrapped out-of-bag samples and 95% confidence intervals were reported.

SPHARM decomposition was performed by sampling 25 shells at different radii which were decomposed up to a maximal SPHARM degree $L_{max} = 25$, resulting in 25 x 25 matrices for each MR contrast on native images. The ability of SPHARM descriptors to encode and reconstruct endometrial tumors' volumes was assessed by comparing these original volumes to inverse SPHARM-reconstructed volumes with the multi-scale structural similarity index (MSSI). A tensor logistic regressor (TensorReg) was used to fit a classification matrix of SPHARM coefficients for classifying histopathological features of endometrial cancer. All SPHARM and TensorReg analyses were performed in MATLAB (2020a, Mathworks, Natick, MA, USA). Using SPHARM decomposition of each MR contrast's images, training and validation were performed using stratified 5-fold cross-validation and using bootstrapping of training folds to produce predictions while adjusting the regularization parameter to minimize both training and validation error. The obtained SPHARM predictions of each contrast were subsequently combined by fitting a simple regularized logistic regression model using bootstrapping to produce 95% confidence intervals. Both models were independently tested with an external testing dataset from the Hôpital Lariboisière using descriptors extracted i) from expert segmentations and ii) spherical VOIs. Expert segmentations were compared with spherical VOIs with Dice coefficients across MR sequences.

Receiver operating characteristics (ROC) curves were reported based on models' predictions on the training and testing datasets. Thresholds maximizing Youden's index (sensitivity+specificity-1) were identified and associated diagnostic performance metrics of models were reported, *i.e.* sensitivity, specificity, balanced accuracy, positive predictive value, and negative predictive value.

5.3. Results

Population characteristics

Between January 2011 and July 2015, 94 patients who underwent 1.5 T MRI examination before surgery for endometrial carcinoma at McGill University Health Center were included (mean age: 65.5 years [range: 43-90 years]). In this training dataset, 43 patients (43 of 94, 45.7%) had deep MI, and 33 (33 of 94, 35.1%) had high FIGO grade as assessed by histology (**Table 5.3**). At Hôpital Lariboisière, a total of 63 patients were included (mean age: 67.1 years [range: 44-88 years]). In this external testing dataset, 36 patients (36 of 63, 57.1%) had deep MI, and 15 (15 of 63, 23.8%) had high FIGO grade as assessed by histology (**Table 5.3**).

Table 5.3 Patient surgical histopathological findings.

	Training dataset (*n* = 94)	**Testing dataset (*n* = 63)**
Deep myometrial invasion		
< 50% myometrial invasion	51 (54.3%)	27 (42.9%)
Deep (> 50%) myometrial invasion	43 (45.7%)	36 (57.1%)
Histopathological grade		
Low (grade 1and 2)	61 (64.9%)	48 (76.2%)
High (grade 3 and non-endometriod)	33 (35.1%)	15 (23.8%)

5.3.1. Multi-parametric MRI Results

Reproducibility

The reproducibility analysis revealed that the preprocessing parameters which led to the highest overall ICC across all MR sequences were resampled isotropic voxels size of 1 mm^3 and normalized absolute discretized MR images with fixed bin size of 25 (**Figure 5.3** and **Figure 5.4**). Using this set of preprocessing steps, 56 radiomics features were excluded from further analysis since their associated ICC were below the predefined 0.80 reproducibility threshold as shown on **Figure 5.5**, 19 extracted from ADC maps, 20 from DWI, 1 from second phase DCE-MRI, 1 from third phase DCE-MRI, 15 from delayed phase DCE-MRI, and none from T2-weighted MRI. There were 11 first-order features excluded (2 extracted from ADC maps, 4 from DWI, and 5 from delayed phase DCE-MRI), no shape-based features excluded, and 45 textural features excluded (17 extracted from ADC maps, 16 from DWI, 1 from second-phase and 1 from third-phase DCE-MRI (both large area high gray-level emphasis [GLSZM]), and 10 from delayed-phase DCE-MRI). Thus, there were 580 features left from the 636 (91%) originally extracted features (*i.e* 106 features

per sequence) after this first feature selection step. Secondly, highly correlated multicollinear features were removed (Spearman's rho > 0.95) resulting in keeping only 361 features of the 580 selected reproducible features (62%, or 57% of originally extracted features).

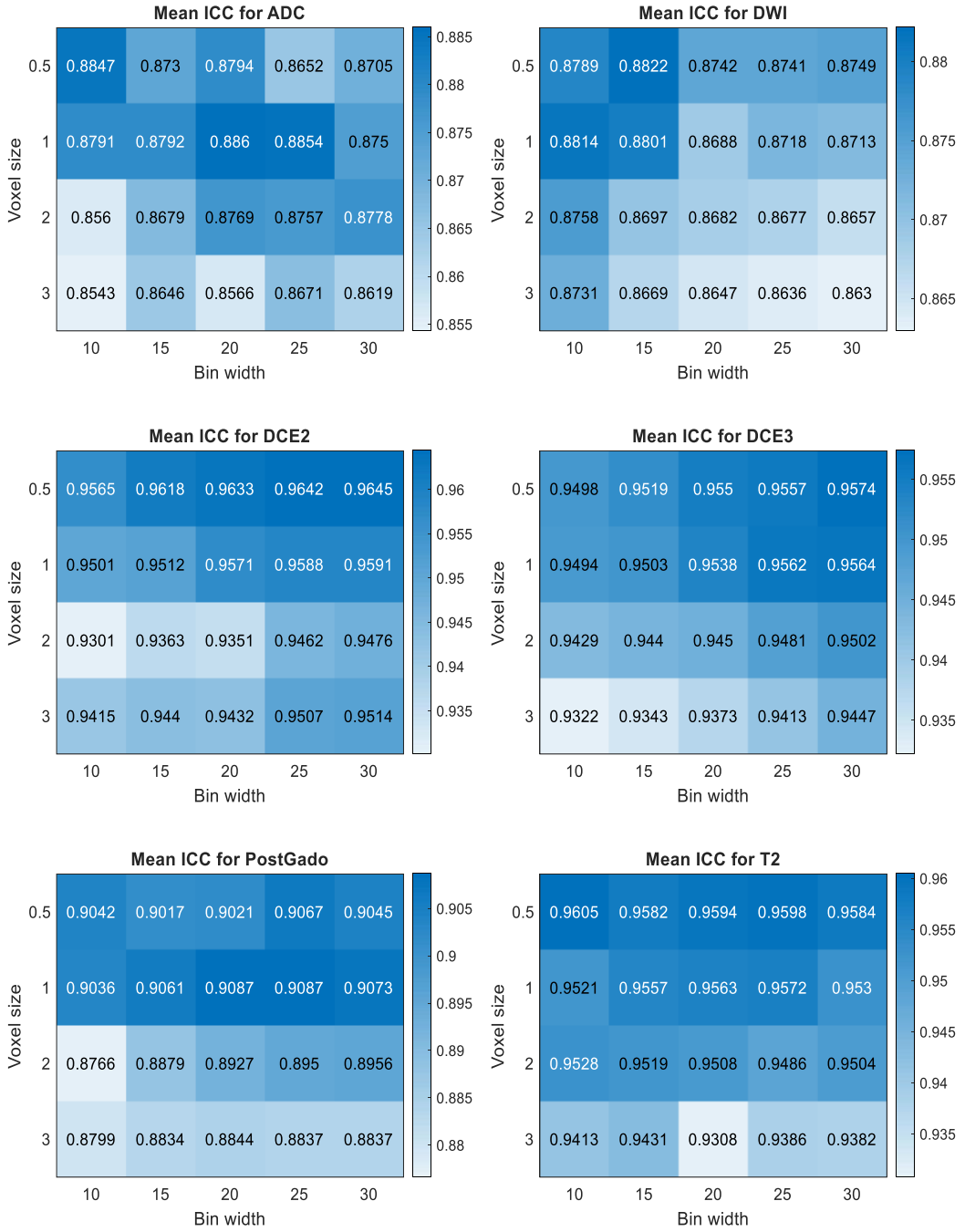


Figure 5.3 Mean intra-class correlation coefficients for assessing reproducibility of radiomics features evaluated in mpMRI of endometrial tumours.

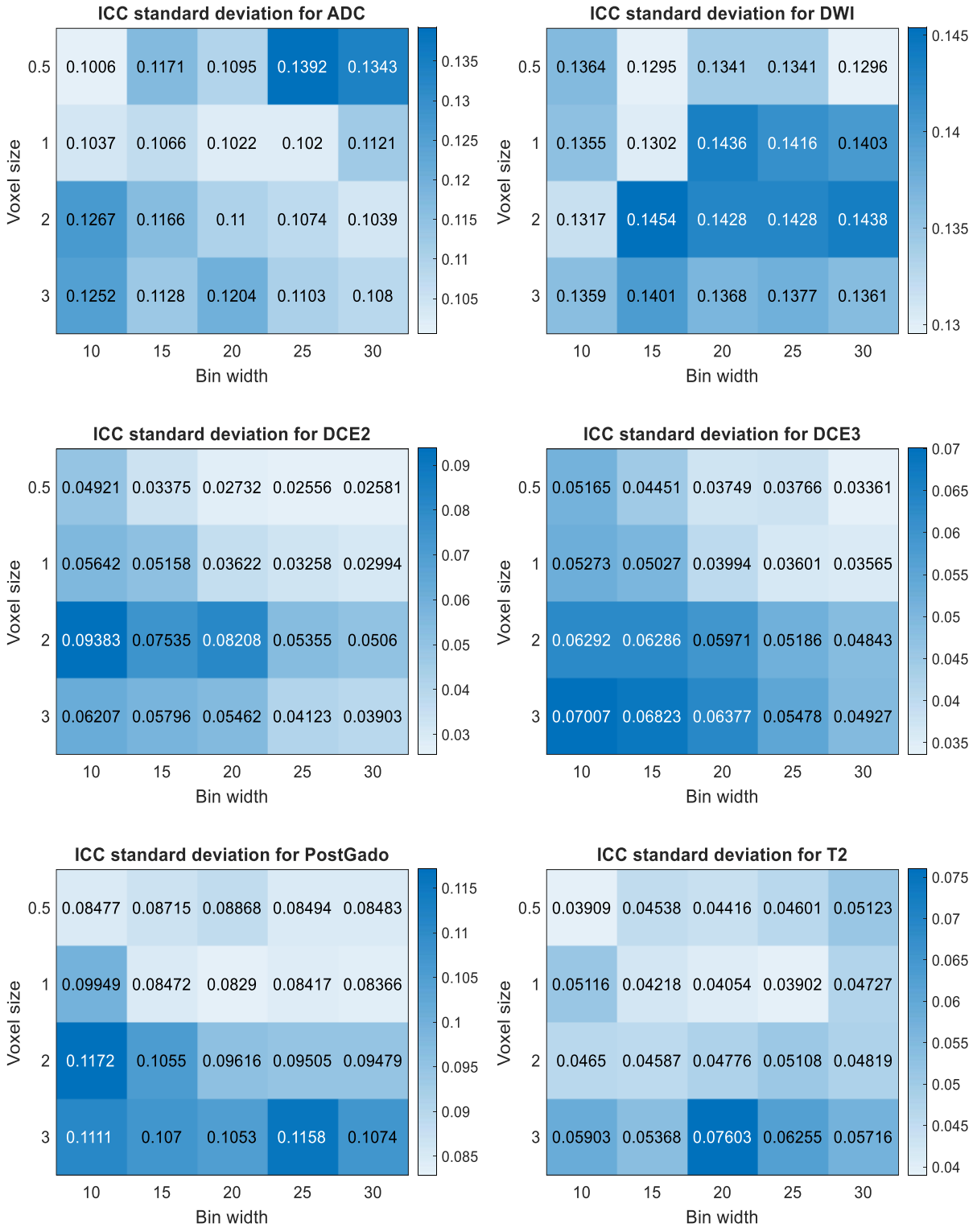


Figure 5.4 Standard deviation of intra-class correlation coefficients for assessing reproducibility of radiomics features evaluated in mpMRI of endometrial tumors.

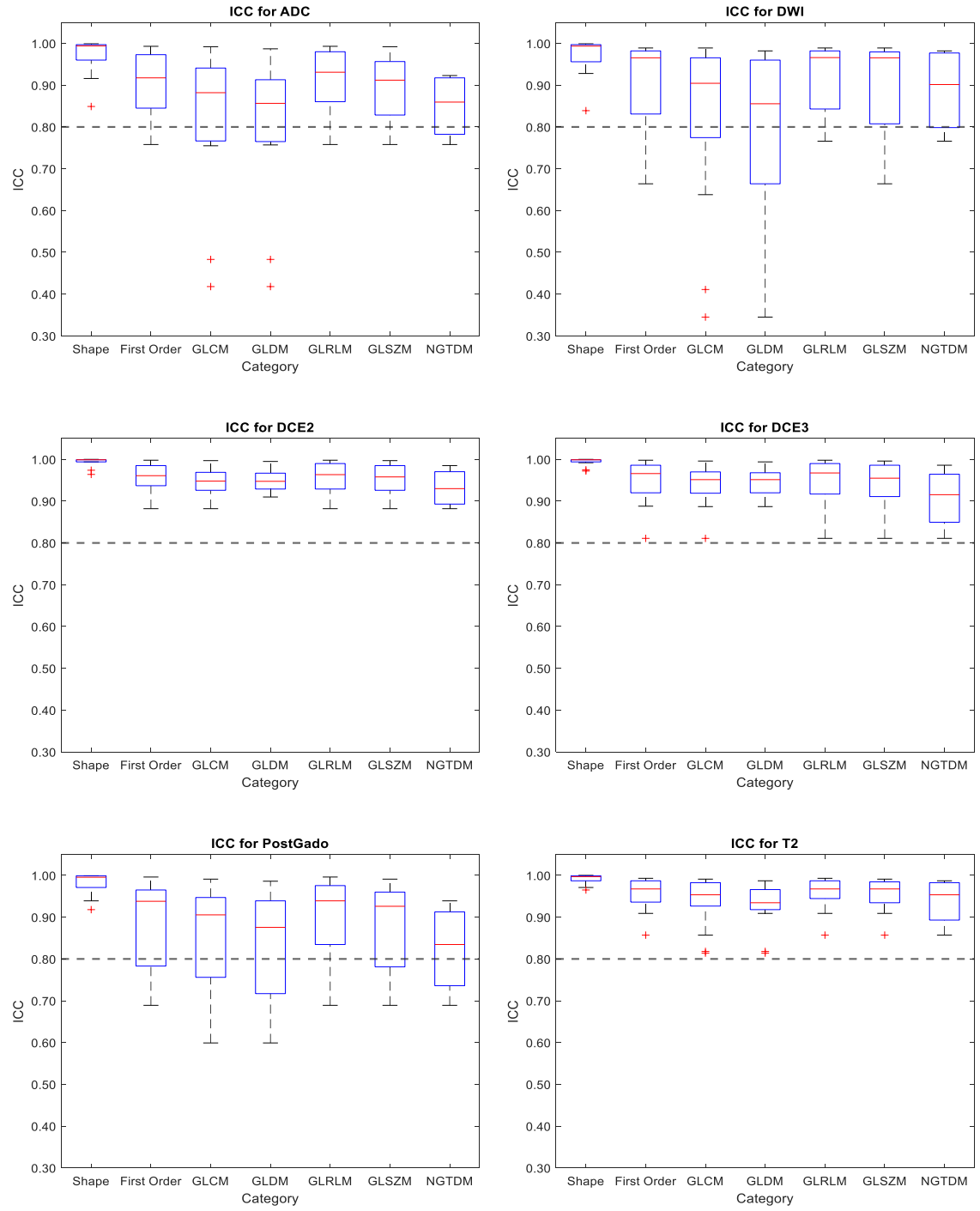


Figure 5.5 Intra-class correlation coefficients for assessing reproducibility of radiomics features in each class extracted from preprocessed MR images with fixed bin width of 25 and resampled isotropic voxels of 1 mm^3.

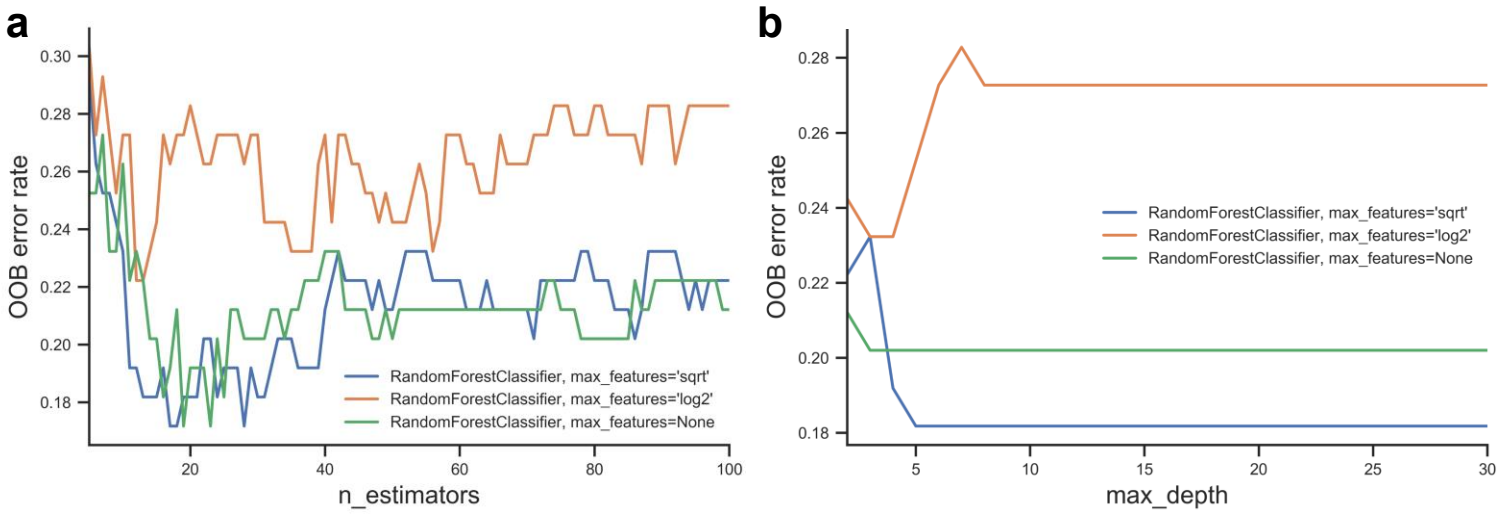


Figure 5.6 Out-of-bag error as a function of (a) the number of trees and (b) the maximal depth of each tree in the random forest. Minimizing these hyperparameters decreases over-fitting.

Radiomics in Deep Myometrial Invasion

Prior to feature selection, each radiomics feature used alone for differentiating <50% MI from deep MI with an area under the ROC curve (AUC) greater than 0.75 in the training dataset were reported in **Table 5.4**. Fifty features of the set of reproducible features were identified (6 extracted from ADC maps, 6 from DWI, 7 from second phase DCE-MRI, 9 from third phase DCE-MRI, 17 from delayed phase DCE-MRI, and 5 from T2-weighted MRI). Four were first-order statistics, 24 were shape features, and 22 were texture features. Interestingly, simple and easily assessed morphological descriptors extracted from different MR contrasts such as lesions' least axis length, minor axis length, or maximum 2D diameter slice led to AUCs greater than 0.77. These highly discriminating shape-based features suggested that basic analysis the segmented VOI shape might lead to improved classification of the depth of MI.

Table 5.4 Most discriminative single radiomics features for predicting deep myometrial invasion (AUC ≥ 0.75).

#	**Radiomics features**	**AUC**
1	PostGado_shape_LeastAxisLength	0.812
2	T2_shape_LeastAxisLength	0.800
3	PostGado_gldm_DependenceEntropy	0.795
4	DCE3_shape_LeastAxisLength	0.785
5	DWI_shape_LeastAxisLength	0.785
6	DCE3_glszm_ZoneEntropy	0.782
7	PostGado_glcm_Idn	0.777
8	ADC_shape_MinorAxisLength	0.777
9	PostGado_shape_MeshVolume	0.775
10	PostGado_shape_VoxelVolume	0.775
11	PostGado_shape_MinorAxisLength	0.772
12	PostGado_glcm_Correlation	0.772
13	T2_ngtdm_Busyness	0.771
14	ADC_shape_Maximum2DDiameterSlice	0.771
15	DCE2_shape_LeastAxisLength	0.770
16	DCE3_glcm_Correlation	0.769
17	PostGado_glcm_Idmn	0.768
18	ADC_shape_LeastAxisLength	0.768
19	PostGado_gldm_LargeDependenceHighGrayLevelEmphasis	0.766
20	DWI_shape_MinorAxisLength	0.766
21	DWI_glszm_LargeAreaHighGrayLevelEmphasis	0.765
22	T2_glrlm_GrayLevelNonUniformity	0.764
23	DCE3_glcm_Idmn	0.764
24	T2_gldm_GrayLevelNonUniformity	0.763
25	PostGado_glrlm_RunLengthNonUniformity	0.760
26	PostGado_glszm_ZoneEntropy	0.759
27	DCE2_glcm_Idmn	0.759
28	DCE3_shape_Maximum2DDiameterColumn	0.758
29	DWI_gldm_GrayLevelNonUniformity	0.758
30	PostGado_shape_SurfaceArea	0.757
31	DCE3_glcm_MCC	0.757
32	DWI_shape_Maximum2DDiameterSlice	0.757
33	ADC_shape_SurfaceArea	0.756
34	PostGado_glszm_LargeAreaHighGrayLevelEmphasis	0.755
35	ADC_shape_VoxelVolume	0.755
36	ADC_shape_MeshVolume	0.755
37	DWI_shape_VoxelVolume	0.755
38	DCE2_glcm_Idn	0.754
39	DCE3_glcm_Idn	0.753
40	DCE2_shape_Maximum2DDiameterColumn	0.753
41	PostGado_gldm_DependenceNonUniformity	0.753
42	DWI_shape_MeshVolume	0.753
43	DCE3_gldm_DependenceNonUniformity	0.752
44	T2_shape_MinorAxisLength	0.752
45	DCE2_firstorder_Energy	0.752
46	DCE2_firstorder_TotalEnergy	0.751
47	PostGado_firstorder_TotalEnergy	0.751
48	PostGado_shape_Maximum2DDiameterSlice	0.751
49	DCE2_gldm_DependenceNonUniformity	0.750
50	DCE3_firstorder_TotalEnergy	0.750

After the exclusion of unreproducible and multicollinear correlated features, the random forest features selection process based on Gini impurity led to the inclusion of a total of 5 radiomics features, over which the diagnostic performance was not significantly increased by the inclusion of more features. To limit overfitting and to minimize out-of-bag error in the model, the number of trees included in random forest modeling was limited to 20, the depth of each decision tree was limited to 5, and the number of features to possibly split at each node was limited to the squared root of the number of features (**Figure 5.6**). The selected discriminative radiomics features were 2 shape-based features – least axis length extracted from delayed phase DCE-MRI and maximum 2D diameter slice extracted from ADC maps – and 3 second-order texture statistics – zone entropy from gray-level size zone matrices (GLSZM), dependence entropy from gray-level dependencies matrices (GLDM), and dependence nonuniformity from GLDM all extracted from delayed phase DCE-MRI. Included features were all among the single most discriminating (AUC > 0.75) when simply using one feature in ROC analysis in the training dataset as reported in **Table 5.4**. Their distributions against one another are represented in pair plots in **Figure 5.7** where blue dots represent subjects with deep MI and orange dots represent subjects without deep MI (<50% MI).

The final random forest model was built including only these 5 features (ADC maps-based maximum 2D diameter slice and delayed phase DCE-MRI-based least axis length, zone entropy [GLSZM], dependence entropy [GLDM], and dependence nonuniformity [GLDM]). Cross-validated training with bootstrapping on balanced samples of the training dataset with the random forest classifier resulted in high diagnostic performance for predicting deep MI as shown by ROC analysis (**Figure 5.8**; **Table 5.5**). This model was then applied to the external testing dataset on which the random forest classifier also provided high performance. When using the 5 most reproducible and discriminative features, the random forest model resulted in diagnostic performance on the testing dataset with an AUC of 0.81. Sensitivity, specificity and balanced accuracy were respectively 86%, 75% and 81% in the testing dataset.

2D planes of the random forest classification decision boundaries are represented in **Figure 5.9**, where orange and red are associated again with <50% MI and blue with deep MI. The overall smooth classification heatmaps confirm that overfitting was limited in our random forest modeling. It also shows that using combination of only 2 radiomics features might not fully allow to assess the depth of MI, as seen from significant overlap between classes.

Table 5.5 Diagnostic performance of random forest classifier on radiomics features to predict deep myometrial invasion.

		AUC	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)
Deep MI	Training	0.86 (0.75–0.93)	84 (71–93)	88 (77–95)	86 (78–92)	84 (71–93)	88 (75–94)
	Testing	0.81 (0.68–0.88)	86 (68–94)	75 (60–93)	81 (70–91)	82 (67–91)	80 (60–90)

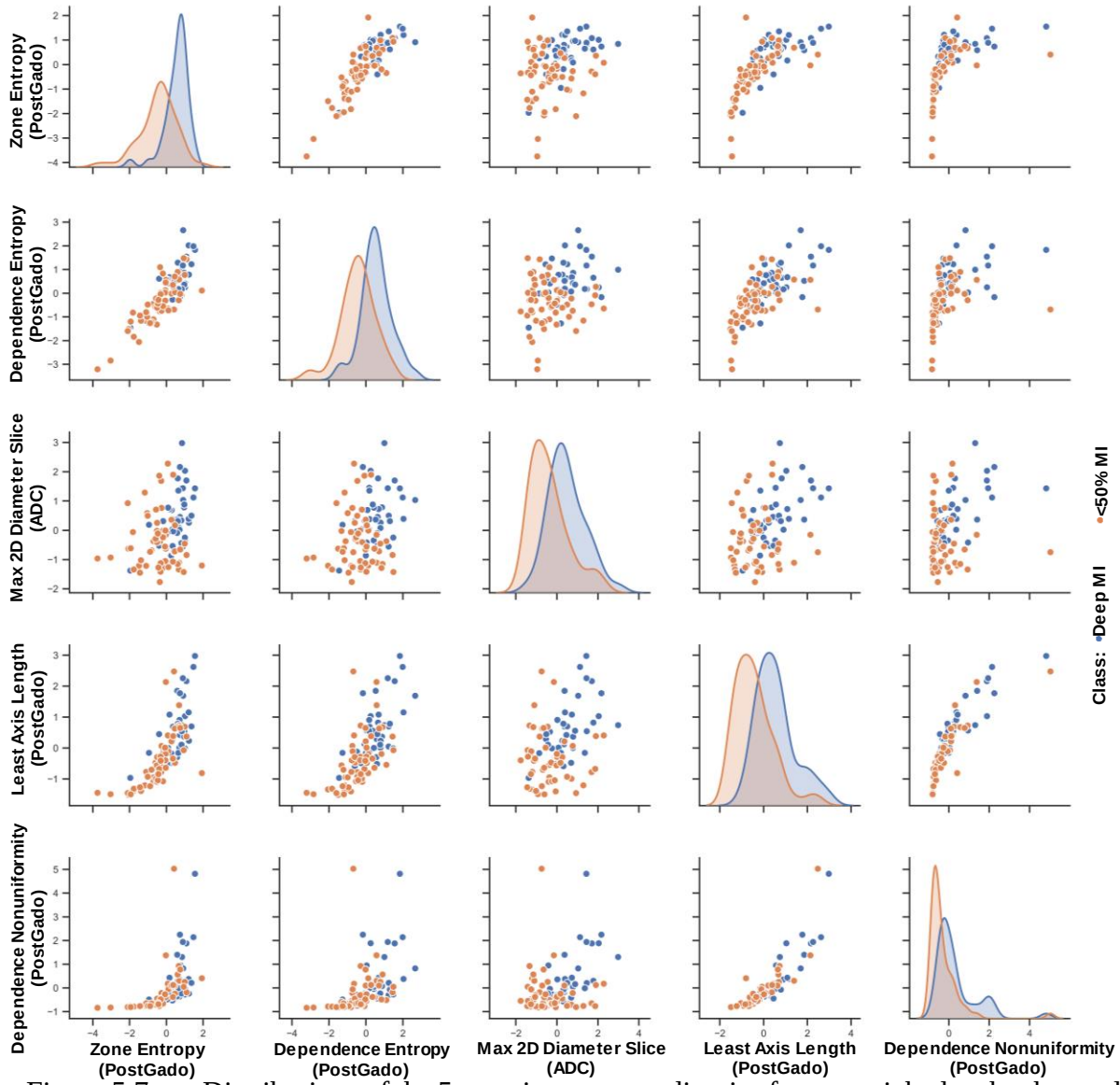


Figure 5.7 Distributions of the 5 most important radiomics features picked up by the random forest model for classifying deep myometrial invasion.

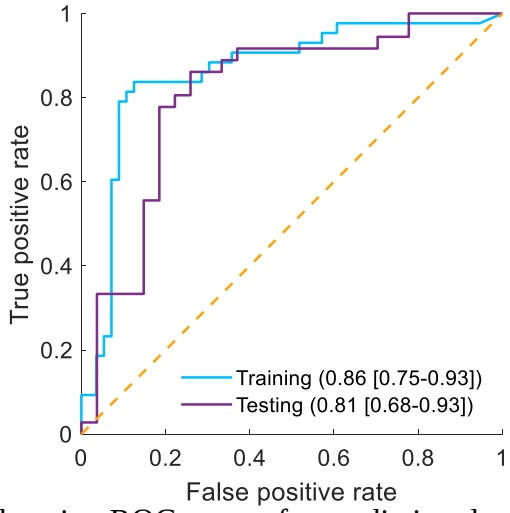


Figure 5.8 Training and testing ROC curves for predicting deep myometrial invasion with a radiomics-based random forest model.

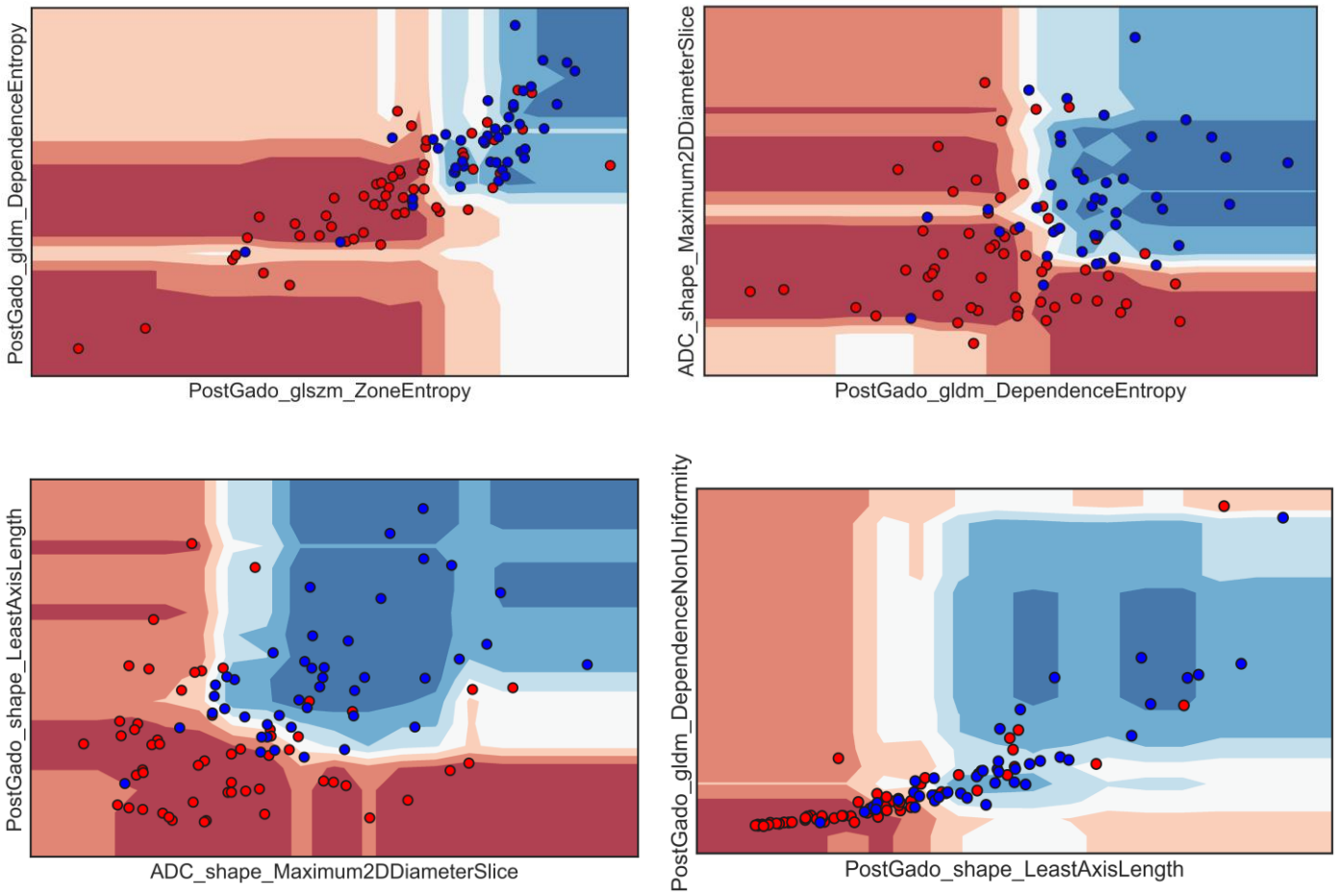
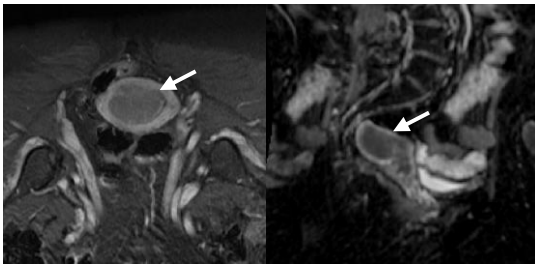
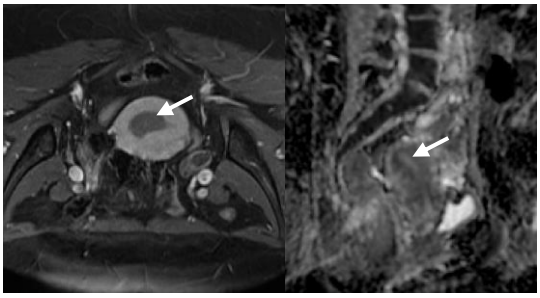


Figure 5.9 Random forest classifier visualization of 2D classification planes of most important features for the classification of deep myometrial invasion on mpMRI.



Deep MI (FIGO IIIA), Predicted Risk = 0.92

Zone Entropy (GLSZM, PG) =	7.16
Dependence Entropy (GLDM, PG) =	7.47
Maximum 2D Diameter Slice (ADC) =	70.9mm
Least Axis Length (PG) =	26.0mm
Dependence NonUniformity (GLDM, PG) =	2625



<50% MI (FIGO IA), Predicted Risk = 0.11

Zone Entropy (GLSZM, PG) =	6.73
Dependence Entropy (GLDM, PG) =	7.62
Maximum 2D Diameter Slice (ADC) =	38.9mm
Least Axis Length (PG) =	14.5mm
Dependence NonUniformity (GLDM, PG) =	790

Figure 5.10 Radiomics features extracted on delayed phase DCE-MRI and ADC maps in two subjects with and without deep myometrial invasion. PG = PostGado.

Examples of well classified depth of MI on endometrial lesions with radiomics features are shown in **Figure 5.10**. Notable quantitative features of these lesions on ADC maps and delayed phase DCE-MRI were that the lesion with deep MI was larger than that without deep MI (maximum 2D diameter slice, 70.9 mm vs. 38.9 mm, respectively; least axis length, 26.0 mm vs. 14.5 mm, respectively). For texture features, entropy was used as a measure of randomness and uncertainty in two matrix representations, GLSZM and GLDM. Dependence nonuniformity measured the similarity of GLDMs, with higher values associated to more heterogeneity among dependencies. Thus, tumors with deep MI appeared to show more textured appearance on delayed phase DCE-MRI, with more heterogeneous zones and chaotic voxel intensities dependencies compared to that of the lesion without deep MI.

Radiomics in High Grade

For differentiating low from high grade endometrial cancer, no radiomics features resulted in AUC > 0.75 when simply using one feature in ROC analysis. Only 14 resulted in AUC > 0.72 (**Table 5.6**). The selected 5 most discriminative radiomics features for predicting high grade incorporated radiomics features only from second-order statistics but from 5 different MR sequences: contrast from neighbouring gray-tone difference matrix (NGDTM) extracted from T2-weighted MRI, gray-level variance from GLCM extracted from ADC maps, and large run high gray-level emphasis from gray-level run-length matrix (GLRLM) extracted from third phase DCE-MRI, inverse difference moment normalized from GLCM extracted from DWI, and gray-level nonuniformity from GLDM extracted from second phase DCE-MRI. Their distributions against one another are represented in pair plots in **Figure 5.11** where blue dots represent subjects with high grade and orange dots represent subjects with low grade.

Bootstrapped training on balanced samples of the training dataset with our random forest classifier on these 5 features resulted in good diagnostic performance for differentiating low grade from high grade endometrial cancer (**Figure 5.12**; **Table 5.7**). In fact, the classifier robustly performed from training to testing with AUC of 0.74. Sensitivity, specificity and balanced accuracy were respectively 65%, 84% and 75% in the testing dataset.

Table 5.6 Most discriminative radiomics features extracted from mpMRI for predicting high grade endometrial cancer (AUC ≥ 0.72).

#	**Radiomics Features**	**AUC**
1	DWI_glcm_ClusterTendency	0.734
2	DCE2_firstorder_Median	0.732
3	DCE3_firstorder_Median	0.732
4	DCE2_firstorder_RootMeanSquared	0.731
5	DCE2_firstorder_Mean	0.729
6	DCE3_firstorder_RootMeanSquared	0.727
7	DWI_glcm_SumEntropy	0.727
8	DCE3_firstorder_Mean	0.726
9	DCE2_firstorder_90Percentile	0.723
10	DWI_glcm_ClusterProminence	0.722
11	DWI_glcm_SumSquares	0.722
12	DWI_glcm_JointEntropy	0.721
13	DWI_glrlm_HighGrayLevelRunEmphasis	0.720
14	DWI_gldm_HighGrayLevelEmphasis	0.720

However, as seen on features distribution on **Figure 5.11**, we cannot simply discriminate linearly lesions based on these selected features. Complex discriminating decision trees had to be combined to produce accurate classification. However, one could argue that an AUC of 0.74 is not good, especially in the staging of cancerous lesions. Examples of endometrial tumors with low and with high grade are represented on **Figure 5.13** for visualization.

Table 5.7 Diagnostic performance of random forest classifier on radiomics features to predict high grade endometrial cancer.

		AUC	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)
High Grade	Training	0.74 (0.63–0.83)	90 (76–98)	60 (52–72)	74 (57–86)	55 (41–69)	92 (80–98)
	Testing	0.74 (0.61–0.86)	65 (46–84)	84 (70–92)	75 (65–86)	56 (41–68)	89 (75–95)

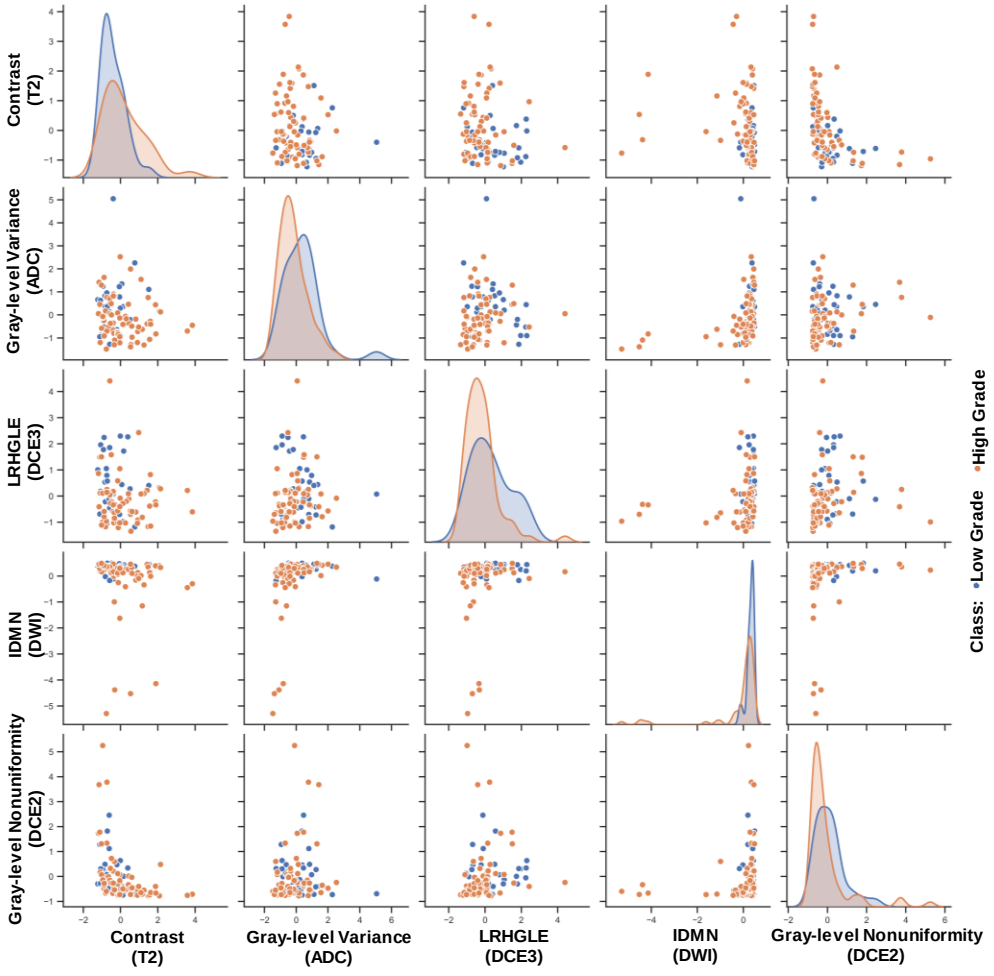


Figure 5.11 Distributions of the 5 most important radiomics features picked up by the random forest model for classifying low grade from high grade endometrial tumors.

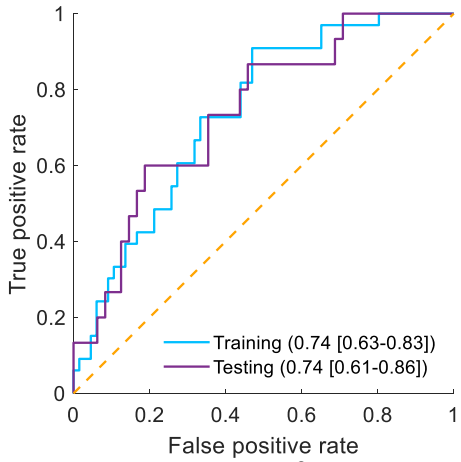
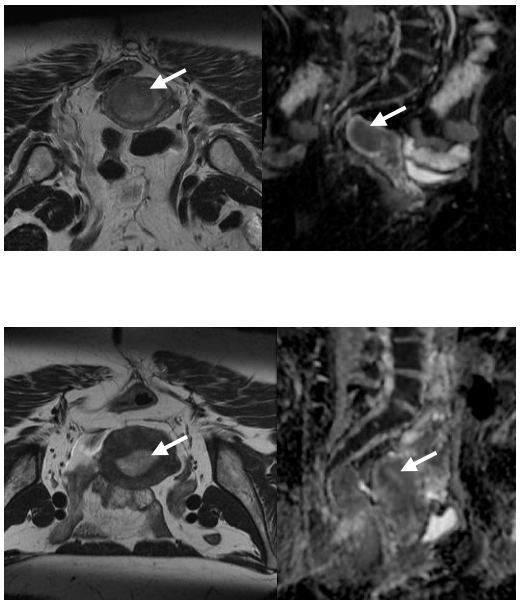


Figure 5.12 Training and testing ROC curves for predicting high grade endometrial cancer with the 5 most important radiomics features picked up by the random forest classifier.



High Grade (FIGO IIIA), Predicted Risk = 0.76

Contrast (NGTDM, T2w) = 0.023

GL Variance (GLSZM, ADC) = 28.3

LRHGLE (GLRLM, DCE_{t3}) = 564.6

IDMN (GLCM, DWI_{b1000}) = 0.998

GL Nonuniformity (GLDM, DCE_{t2})= 1491.6

Low Grade (FIGO IA), Predicted Risk = 0.22

Contrast (NGTDM, T2w) = 0.058

GL Variance (GLSZM, ADC) = 17.1

LRHGLE (GLRLM, DCE_{t3}) = 426.0

IDMN (GLCM, DWI_{b1000}) = 0.996

GL Nonuniformity (GLDM, DCE_{t2})= 400.0

Figure 5.13 Radiomics features extracted on mpMRI (T2-weighted MRI and ADC maps on the left panels) in two subjects with low grade and high grade endometrial tumors.

SPHARM in Deep Myometrial Invasion

Table 5.8 Patient surgical histologic findings in subgroup with endometrial tumors with minimal radius of 10 mm.

	Training dataset (*n* = 89 [94.7%])	**Testing dataset (*n* = 60 [95.2%])**
Deep myometrial invasion		
< 50% myometrial invasion	48 (53.9%)	25 (41.7%)
Deep (> 50%) myometrial invasion	41 (46.1%)	35 (58.3%)
Histopathological grade		
Low (grade 1 and 2)	60 (67.4%)	45 (75.0%)
High (grade 3 and non-endometriod)	29 (32.6%)	15 (25.0%)

For this section of analyses, a subset of the population with a minimal endometrial lesions' radius of 10 mm was selected to assess the diagnostic performance of SPHARM descriptors. Using SPHARM decomposition with 25 shells and up to $L_{max} = 25$, the reconstruction ability of SPHARM descriptors was assessed with MSSI. Mean MSSI across the population was 0.92 ± 0.05 (range: 0.73-0.99) for ADC maps, 0.96 ± 0.04 (0.80-0.99) for DWI, 0.90 ± 0.06 (0.67-0.98) for T2-weighted MRI, 0.89 ± 0.04 (0.79-0.97) for second phase DCE-MRI, 0.89 ± 0.04 (0.79-0.96) for third phase DCE-MRI, and 0.90 ± 0.05 (0.76-0.98) for delayed phase DCE-MRI. Thus, SPHARM ability for encoding volumes and their content from MR images was high across all MR sequences and could adequately reconstruct these volumetric images. SPHARM decomposition was performed for each MR contrast and predictions from trained TensorReg model on these extracted SPHARM descriptors were combined with a simple logistic regressor. The set of previously identified 5 most discriminative radiomics features were used in the same random forest model to enable comparisons on this subset of the population.

Table 5.9 Detailed diagnostic performance of radiomics features and SPHARM descriptors for predicting deep myometrial invasion combining all MR sequences.

		AUC	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	**PPV (%)**	**NPV (%)**
DMI	SPHARM, Training	0.94 (0.85–0.98)	82 (68–93)	93 (73–100)	88 (77–94)	93 (78–100)	82 (64–92)
	SPHARM, Testing	0.94 (0.85–1.00)	100 (100–100)	74 (51–92)	90 (78–98)	88 (71–98)	100 (100–100)
	Radiomics, Training	0.92 (0.82–0.95)	93 (82–98)	75 (62–85)	82 (75–90)	74 (60–85)	94 (82–98)
	Radiomics, Testing	0.92 (0.81–0.98)	82 (65–93)	80 (51–94)	81 (70–91)	89 (70–97)	72 (55–91)

Combined predictions from mpMRI-SPHARM decompositions resulted in high diagnostic performance for differentiating <50% MI from deep MI as shown by ROC analysis on training and testing datasets (**Figure 5.14**; **Table 5.9**). In fact, the proposed model robustly performed from training to testing with a consistent AUC of 0.94. These results and their robustness arise from the combination of SPHARM decompositions from multiple MR contrasts. The logistic regression model combining predictions from each SPHARM descriptor for predicting deep MI was:

$$\log\left(\frac{\hat{y}_{MI}}{1-\hat{y}_{MI}}\right) = -2.66 + 1.92x_{ADC} + 0.01x_{DWI} + 0.47x_{T2} + 2.12x_{DCE2} + 0.75x_{DCE3} + 0.47x_{PostGado}$$

Thus, the most relevant MR contrasts for the prediction of the depth of MI with SPHARM decomposition were ADC maps ($\beta_{ADC} = 2.12$) and the second phase DCE-MRI ($\beta_{DCE2} = 2.12$), based on regularized regression coefficients β. Interestingly, most important radiomics features were also extracted from ADC maps, but not from second phase DCE-MRI. Also, removing lesions with radius smaller than 10 mm also increased the diagnostic performance of radiomics for predicting deep MI (testing AUC of 0.92 vs. 0.81; testing balanced accuracy of 0.81 vs. 0.81). SPHARM was more accurate in this subset of the population for characterizing the depth of MI than radiomics (testing AUC of 0.94 vs. 0.92; testing balanced accuracy of 0.90 vs. 0.81).

Further analyses in important single MR contrasts – ADC maps and second phase DCE-MRI – are provided in the next sections to investigate the high diagnostic performance observed when using SPHARM for predicting deep MI.

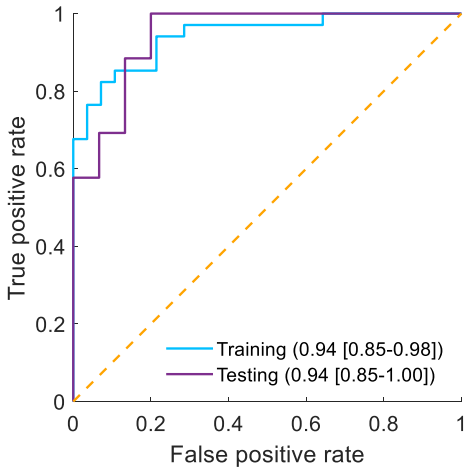


Figure 5.14 Training and testing ROC curves for predicting deep myometrial invasion with SPHARM descriptors combining all MRI sequences. Predictions for each sequence were combined with a linear logistic regressor.

SPHARM in High Grade

Table 5.10 Detailed diagnostic performance of radiomics features and SPHARM descriptors for predicting high grade endometrial cancer combining all MR sequences.

		AUC	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	**PPV (%)**	**NPV (%)**
High grade	SPHARM, Training	0.89 (0.76–0.96)	92 (74–100)	82 (67–92)	86 (75–92)	76 (59–90)	82 (64–92)
	SPHARM, Testing	0.81 (0.64–0.90)	93 (67–100)	63 (45–79)	78 (64–86)	58 (37–86)	95 (68–100)
	Radiomics, Training	0.79 (0.72–0.88)	90 (76–97)	68 (56–79)	76 (68–86)	60 (48–75)	94 (83–98)
	Radiomics, Testing	0.72 (0.58–0.83)	93 (65–100)	55 (41–69)	66 (52–78)	41 (25–58)	96 (80–100)

Combined predictions from mpMRI SPHARM decomposition resulted in high diagnostic performance for differentiating low grade from high grade as shown by ROC analysis on training and testing datasets (**Figure 5.15**; **Table 5.10**). The proposed model resulted in high AUC at training (0.89) and testing (0.81). The logistic regression model combining predictions for predicting high grade from each SPHARM descriptor was:

$$\log\left(\frac{\hat{y}_{HG}}{1-\hat{y}_{HG}}\right) = -3.63 + 1.25\beta_{ADC} + 0.01\beta_{DWI} + 0.51\beta_{T2} + 0.71\beta_{DCE2} + 1.22\beta_{DCE3} + 2.03\beta_{PostGado}$$

Thus, the most relevant MR contrast for the prediction of the tumor grade with SPHARM decomposition was delayed phase DCE-MRI ($\beta_{PostGado} = 2.31$), based on regularized regression coefficients β. Interestingly, no radiomics features extracted from delayed phase DCE-MRI were included in the final random forest model reported above including the 5 most discriminative radiomics features. Removing lesions with radius smaller than 10 mm decreased the diagnostic performance of the previously selected 5 most discriminating radiomics features for predicting high grade (testing AUC of 0.74 vs. 0.72; testing balanced accuracy of 0.75 vs. 0.66). In fact, smaller lesions are often related to lower grade, which might explain this decrease in performance (41). Thus, SPHARM was more accurate in this subset of the population for characterizing the depth of MI than radiomics (testing AUC of 0.81 vs. 0.72; testing balanced accuracy of 0.78 vs. 0.66). Further analyses only on delayed phase DCE-MRI are provided in next sections to address more in-depth reasons for this high diagnostic performance observed when using SPHARM for predicting high grade endometrial cancer.

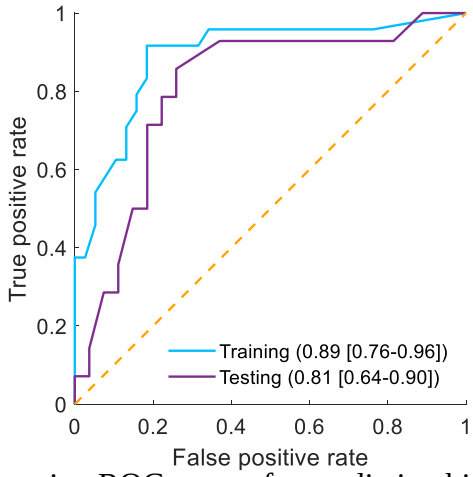


Figure 5.15 Training and testing ROC curves for predicting high grade endometrial cancer with SPHARM descriptors combining all MRI sequences. Predictions for each sequence were combined with a linear logistic regressor.

5.3.2. Single Contrast MRI Results

After combining predictions from all available MR contrasts, we explored single contrasts based on current literature and on important MR contrasts identified by regression models. More specifically, we found that ADC maps extracted from DWI and also DCE-MRI could be highly predictive of the depth of MI quantitatively, but also visually to delineate the extent of endometrial tumors (41, 42, 180, 181, 183, 184, 195-199). Similarly, DCE-MRI is widely used clinically for staging endometrial tumors (41, 184). Thus, we went further and propose here three single MR sequence analyses for predicting deep MI and high grade in endometrial cancer. This was also performed to further analyze SPHARM results presented in previous sections on mpMRI.

Table 5.11 Detailed diagnostic performance of radiomics features and SPHARM descriptors for predicting deep myometrial invasion on ADC maps.

ADC		**AUC**	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	**PPV (%)**	**NPV (%)**
DMI	SPHARM, Training	0.91 (0.81–0.97)	75 (56–87)	93 (77–100)	82 (71–90)	93 (74–100)	74 (56–87)
	SPHARM, Testing	0.85 (0.71–0.94)	62 (43–80)	93 (70–100)	78 (62–88)	94 (65–100)	60 (41–76)
	Radiomics, Training	0.77 (0.63–0.88)	74 (61–87)	72 (59–86)	74 (61–86)	67 (63–71)	79 (70–87)
	Radiomics, Testing	0.76 (0.65–0.84)	92 (78–98)	52 (35–72)	74 (61–84)	72 (58–84)	82 (58–100)

5.3.2.1. Analyses in ADC Maps for Predicting Deep MI

The SPHARM pipeline was repeated using solely ADC maps for predicting deep MI. For comparison, radiomics analyses' diagnostic performance was also assessed by evaluating the set of 5 most important features on ADC maps with random forest modeling. Briefly, these radiomics features consisted in two first-order statistics – skewness and kurtosis – and three shape-based features – minor axis length, surface volume ratio, and maximum 2D diameter slice.

SPHARM decomposition on ADC maps resulted in high diagnostic performance for predicting deep MI as shown by ROC analysis on training and testing datasets (**Figure 5.16**; **Table 5.11**). The proposed model resulted in high AUC at training (0.91) and testing (0.85). Using SPHARM descriptors extracted solely from ADC maps still decreased diagnostic performance for predicting deep MI compared to combining all MR contrasts (testing AUC of 0.94 vs. 0.85; testing balanced accuracy of 0.90 vs. 0.78). Using features extracted solely from ADC maps also decreased the diagnostic performance of radiomics for predicting deep MI (testing AUC of 0.92 vs. 0.76; testing balanced accuracy of 0.81 vs. 0.74). The drop in AUC was higher for radiomics than for SPHARM. Thus, SPHARM was more accurate on ADC maps for characterizing the depth of MI than radiomics (testing AUC of 0.85 vs. 0.76; testing balanced accuracy of 0.78 vs. 0.74).

On **Figure 5.16** is also represented the 2D classification matrix of TensorReg regression coefficients. We see that lower frequency content (blue negative coefficients) was predictor of <50% MI across all radii of ADC maps. Higher frequency content (red positive coefficients) was

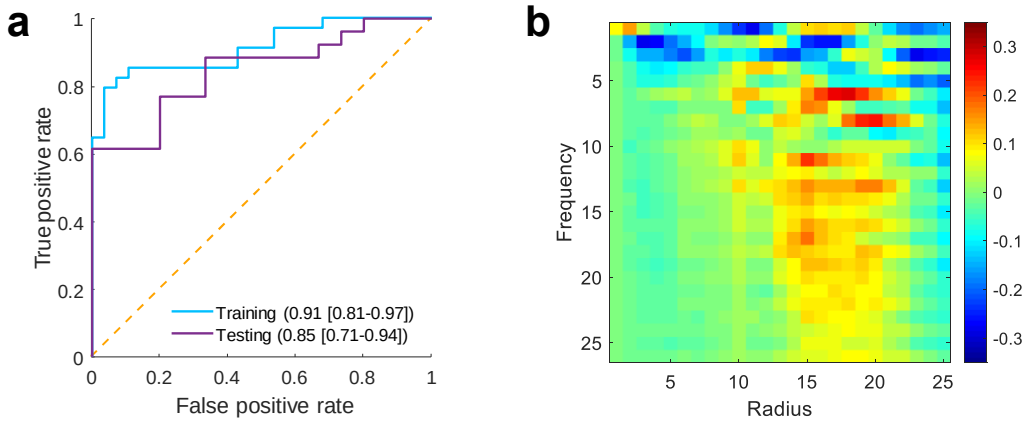


Figure 5.16 (a) Training and testing ROC curves for predicting deep myometrial invasion with SPHARM descriptors on ADC maps, and (b) classification matrix picked up by the logistic tensor regressor trained on the SPHARM descriptors.

predictor of deep MI, mostly around middle-outer radii. Examples of reconstruction and important harmonic filters in lesions showing deep MI or <50% MI are shown on **Figure 5.17** and **Figure 5.18**, respectively. Harmonic filters predictors of deep MI show higher frequency repeated patterns compared to those seen for the lesion without deep MI. In fact, harmonics of lower order and especially at lower radius might have captured the more uniform and overall smaller lesions seen in endometrial tumors with shallow MI or without MI. This is also consistent with shape-based features selected through random forest features selection which were highly correlated with the presence of deep MI. With both techniques, it seemed that rougher textured patterns in larger VOI were associated with deep MI while shallower uniform appearances in smaller VOI were associated with < 50% MI. SPHARM decomposition was the most accurate technique for differentiating absence or shallow MI (<50%) from deep MI using all MR contrasts and only ADC maps.

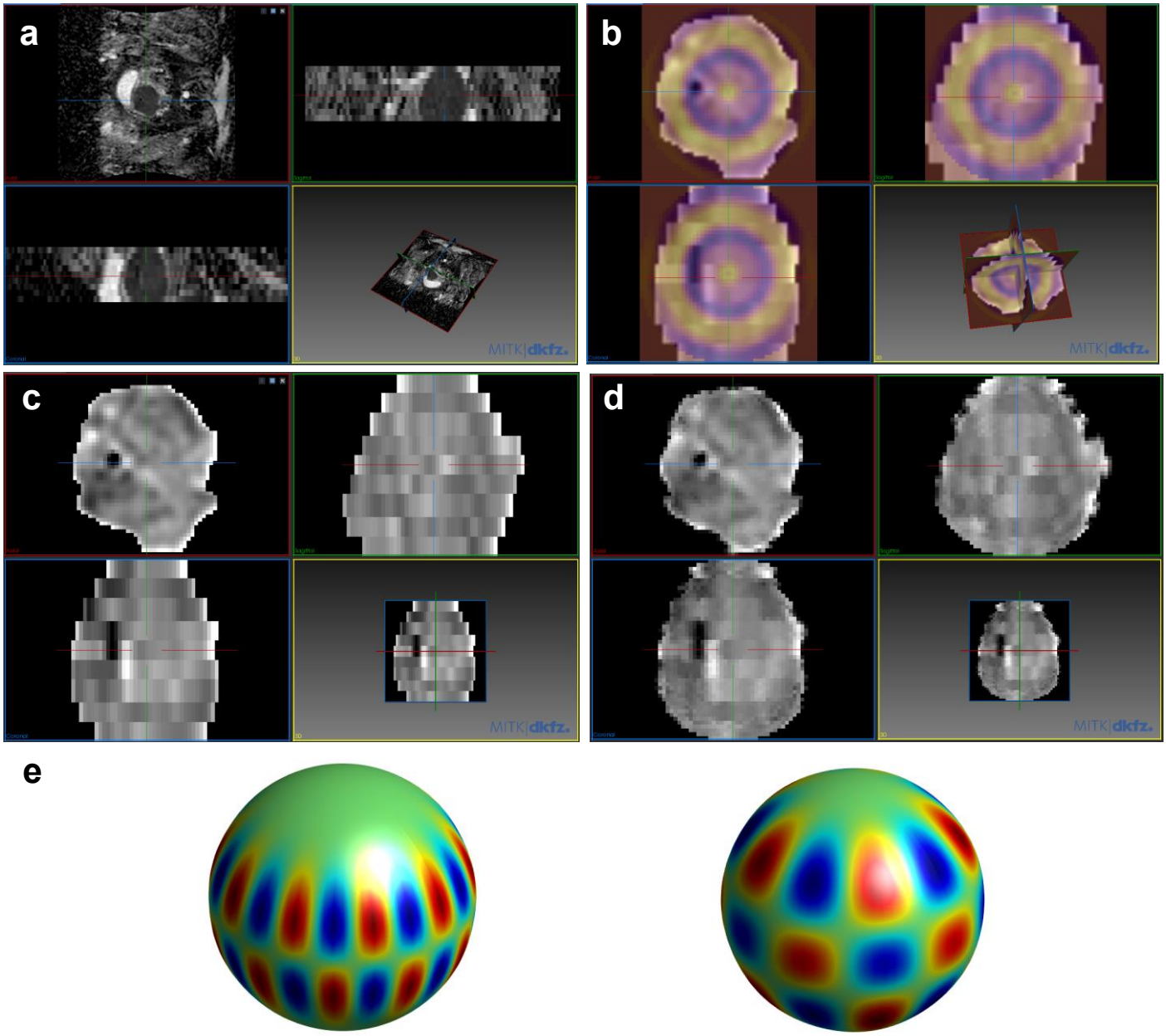


Figure 5.17 (a) ADC map of endometrial tumor with deep MI, (b) segmented tumor with overlayed full reconstructed classification matrix of SPHARM descriptor, (c) segmented ADC maps and its (d) reconstruction with SPHARM coefficients (L_{max}, R_{max} = 25, 25), and (e) harmonic filters with (l, m, r = 11, 10, 15) and (l, m, r = 7, 5, 20), expressed within the spherical yellow overlay in panel (b) as predictors of deep myometrial invasion.

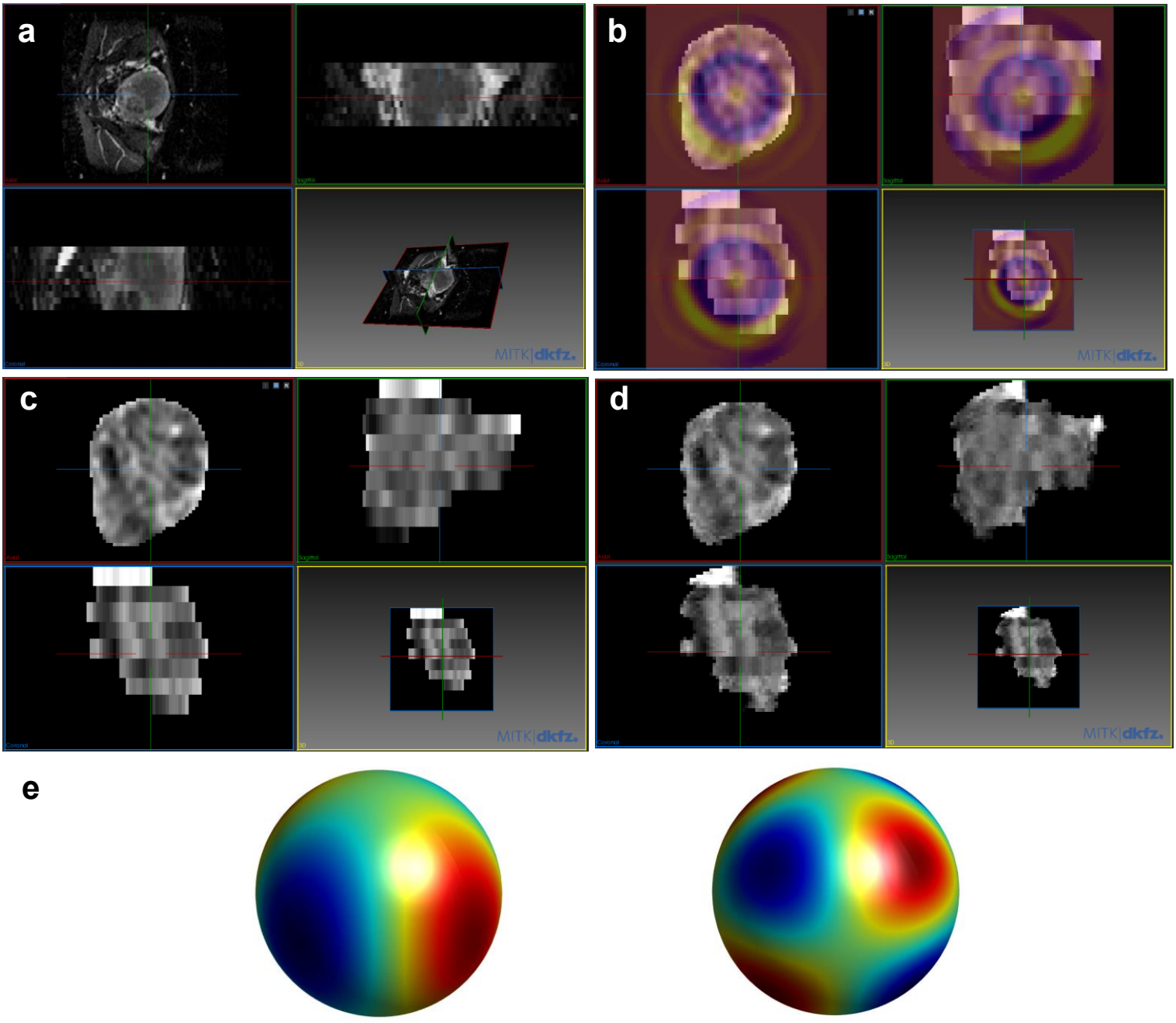


Figure 5.18 (a) ADC map of endometrial tumor without deep MI, (b) segmented tumor with overlayed full reconstructed classification matrix of SPHARM descriptor, (c) segmented ADC maps and its (d) reconstruction with SPHARM coefficients (L_{max}, R_{max} = 25, 25), and (e) harmonic filters with (l, m, r = 2, 2, 3) and (l, m, r = 3, 2, 12), expressed within the spherical blue overlay in panel (b) as predictors of <50% myometrial invasion.

5.3.2.2. Analyses in DCE-MRI for Predicting Deep MI

Table 5.12 Detailed diagnostic performance of radiomics features and SPHARM descriptors for predicting deep myometrial invasion on second enhancement phase of DCE-MRI.

DCE2		AUC	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)
DMI	SPHARM, Training	0.87 (0.77–0.94)	67 (51–83)	90 (72–97)	79 (64–86)	87 (65–96)	71 (54–84)
	SPHARM, Testing	0.86 (0.63–0.96)	95 (73–100)	58 (30–86)	82 (65–94)	81 (63–96)	88 (40–100)
	Radiomics, Training	0.70 (0.54–0.82)	86 (68–100)	55 (40–72)	71 (60–81)	60 (54–64)	85 (67–100)
	Radiomics, Testing	0.68 (0.53–0.81)	70 (50–82)	60 (40–76)	65 (53–75)	70 (53–82)	59 (41–78)

The SPHARM pipeline was repeated using solely second phase DCE-MRI for predicting deep MI. For comparison, radiomics analyses' diagnostic performance was also reassessed by evaluating the set of 5 most important features on second phase DCE-MRI with random forest modeling. These radiomics features consisted in three shape-based features – least axis length, flatness, and mesh volume – and 2 second-order statistics – inverse difference moment normalized from GLCM and dependence nonuniformity from GLDM.

SPHARM decomposition on second phase DCE-MRI resulted in high diagnostic performance for predicting deep MI as shown by ROC analysis on training and testing datasets (**Figure 5.19**; **Table 5.12**). The proposed model robustly performed from training to testing as shown by high consistent AUCs of 0.87 and 0.86, respectively. Using SPHARM descriptors extracted solely from second phase DCE-MRI still decreased in diagnostic performance for predicting deep MI compared to when combining all MR contrasts (testing AUC of 0.94 vs. 0.86; testing balanced accuracy of 0.90 vs. 0.82). Using features extracted solely from second phase DCE-MRI also decreased the diagnostic performance of radiomics for predicting deep MI but to a higher extent (testing AUC of 0.92 vs. 0.68; testing balanced accuracy of 0.81 vs. 0.65). In fact, this drop in AUC was higher for radiomics than for SPHARM. Overall, SPHARM was more accurate on second phase DCE-MRI for characterizing the depth of MI than radiomics (testing AUC of 0.86 vs. 0.68; testing balanced accuracy of 0.82 vs. 0.65).

On **Figure 5.19** is also represented the resulting 2D classification matrix of TensorReg regression coefficients. Interestingly, low frequency content narrowly located around radius 15 was

predictor of <50% MI on second phase DCE-MRI. For predicting deep MI, both high and lower frequencies were used for classification across different radii. Examples of lesions' reconstructions on DCE-MRI and important harmonic filters with deep MI or with <50% MI are shown in **Figure 5.20** and **5.21**, respectively.

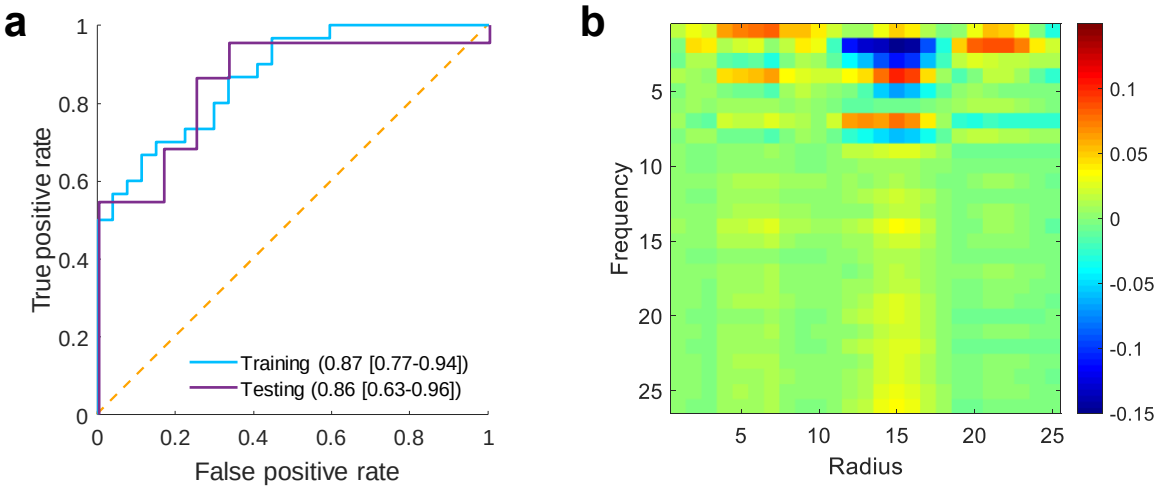


Figure 5.19 (a) Training and testing ROC curves for predicting deep myometrial invasion with SPHARM descriptors on second phase DCE-MRI, and (b) classification matrix picked up by the logistic tensor regressor.

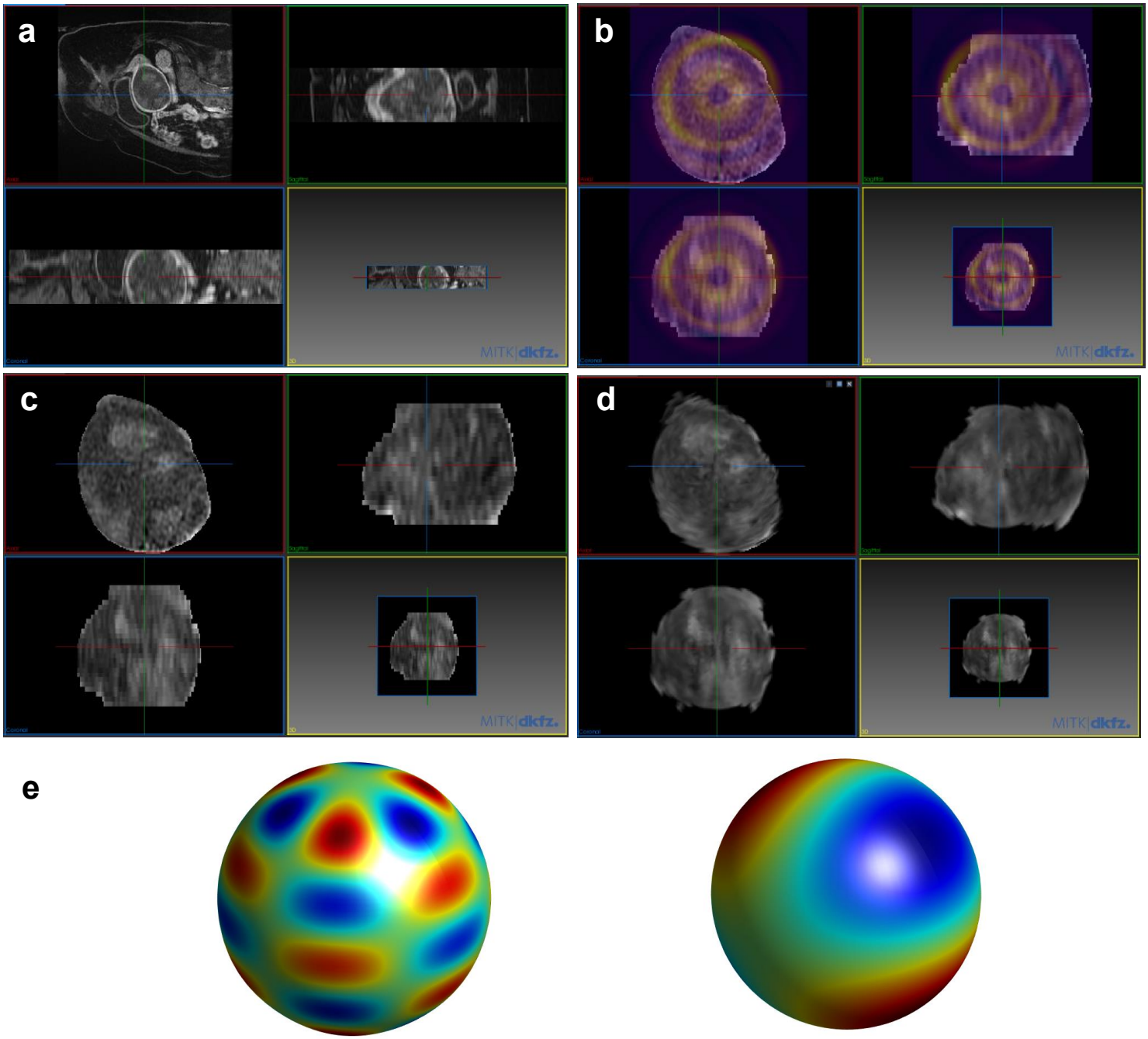


Figure 5.20 (a) Second DCE-MRI phase images of endometrial tumor with deep MI, (b) segmented tumor with overlayed full reconstructed classification matrix of SPHARM descriptor, (c) segmented second phase DCE-MRI and its (d) reconstruction with SPHARM coefficients (L_{max}, R_{max} = 25, 25), and (e) harmonic filters (l, m, r = 7, 3, 15) and (l, m, r = 2, 1, 22), expressed within the spherical yellow overlay in panel (b) as predictors of deep MI.

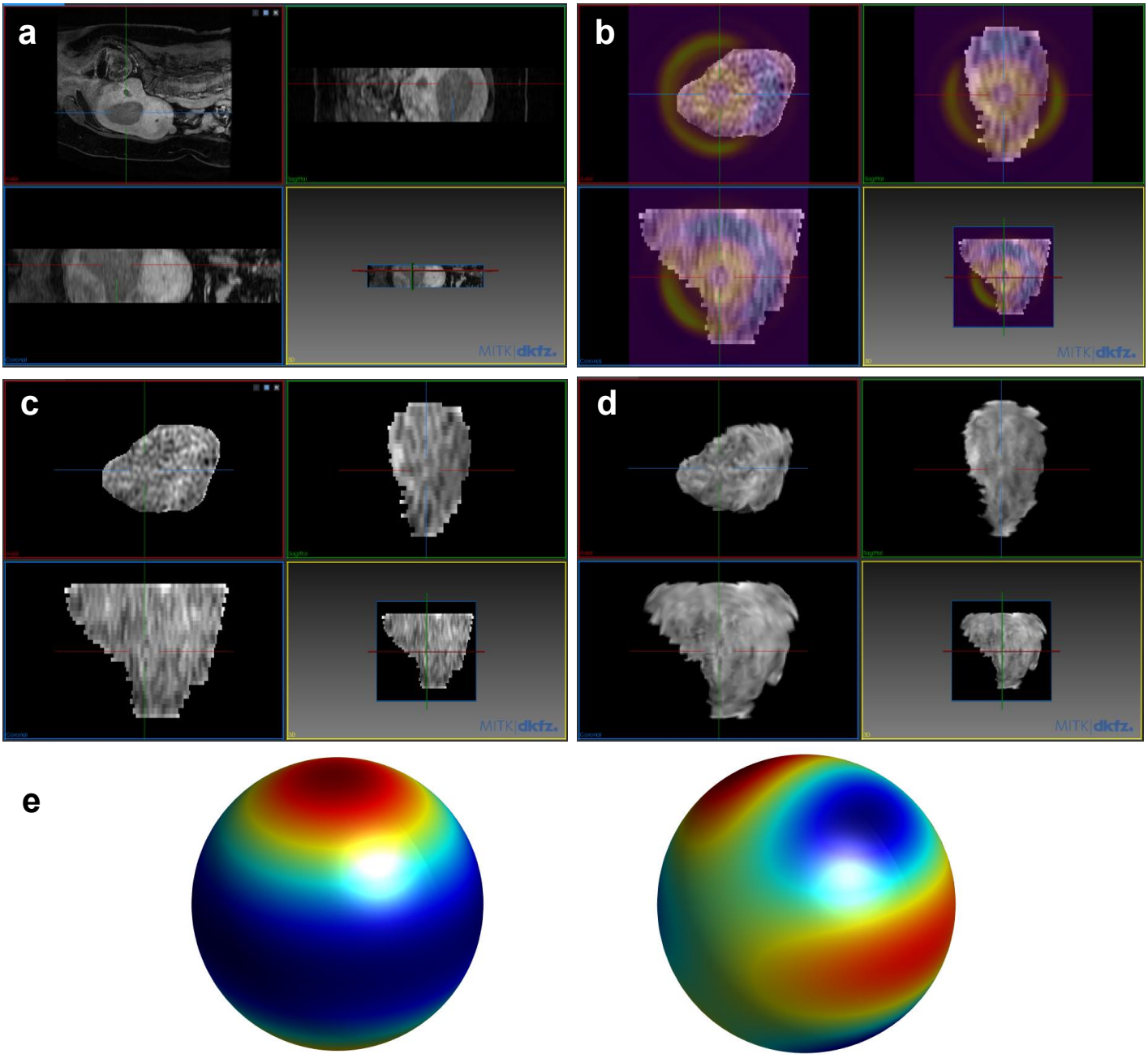


Figure 5.21 (a) Second DCE-MRI phase images of endometrial tumor with deep MI, (b) segmented tumor with overlayed full reconstructed classification matrix of SPHARM descriptor, (c) segmented second phase DCE-MRI and its (d) reconstruction with SPHARM coefficients (L_{max}, R_{max} = 25, 25), and (e) harmonic filters (l, m, r = 2, 0, 15) and (l, m, r = 3, 1, 16), expressed within the spherical yellow overlay in panel (b) as predictors of <50% MI.

5.3.2.3. Analyses in DCE-MRI for Predicting High Grade

Table 5.13 Detailed diagnostic performance of radiomics features and SPHARM descriptors for predicting high grade on delayed enhancement phase of DCE-MRI.

PostGado		AUC	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)
High Grade	SPHARM, Training	0.77 (0.62–0.88)	91 (68–100)	58 (45–76)	74 (64–84)	58 (44–76)	91 (70–100)
	SPHARM, Testing	0.79 (0.62–0.94)	82 (55–100)	68 (54–84)	75 (65–86)	59 (45–77)	89 (66–98)
	Radiomics, Training	0.61 (0.54–0.82)	58 (40–74)	69 (58–81)	62 (60–81)	47 (33–64)	74 (65–85)
	Radiomics, Testing	0.51 (0.53–0.81)	31 (10–56)	79 (78–97)	53 (53–75)	43 (14–80)	72 (58–93)

The SPHARM pipeline was repeated using solely delayed phase DCE-MRI for differentiating low from high grade endometrial cancer. For comparison, radiomics analyses' diagnostic performance was also assessed by evaluating the set of 5 most important features on delayed phase DCE-MRI with random forest modeling. Briefly, these radiomics features consisted in one first-order statistic – total energy – and four texture features – size zone nonuniformity normalized from GLSZM, long run high gray-level emphasis from GLRLM, dependence variance from GLDM, and small dependence low gray-level emphasis from GLDM. SPHARM decomposition on delayed phase DCE-MRI maps resulted in good diagnostic performance for predicting high grade with AUC of 0.77 at training and of 0.79 at testing (**Figure 5.22**; **Table 5.13**). The estimated performance from training to testing increased, suggesting that the developed SPHARM signature was highly robust and accurate across centers. Using SPHARM descriptors extracted only from delayed phase DCE-MRI decreased diagnostic performance for predicting high grade compared to when combining all MR contrasts (testing AUC of 0.81 vs. 0.79; testing balanced accuracy of 0.78 vs. 0.75). The same was observed for radiomics but to a higher extent as modeling with only delayed phase DCE-MRI features was similar to that of random classification (testing AUC of 0.72 vs. 0.51; testing balanced accuracy of 0.66 vs. 0.53). Indeed, the drop in AUC was higher for radiomics than for SPHARM. Therefore, SPHARM decomposition was significantly more accurate on delayed phase DCE-MRI for predicting high grade than radiomics (testing AUC of 0.79 vs. 0.51; testing balanced accuracy of 0.75 vs. 0.53).

On **Figure 5.22** is also represented the 2D classification matrix of TensorReg regression coefficients. Interestingly, the SPHARM signature developed for predicting high grade endometrial cancer was different than that for deep MI. We see that lower frequency content (red positive coefficients) was predictor of high grade on inner radii of delayed DCE-MRI. Higher frequency content, especially at $L_{max} = 5$ (blue negative coefficients), was predictor of low grade, mostly around middle-outer radii. Indeed, homogeneous low frequency inner content seemed to be associated to high grade tumors on delayed phase DCE-MRI. This suggests that an overall clearance of enhancement of the inner compartment of the tumor leading to more homogeneous and dark appearance might have been caught as predictive by the SPHARM decomposition. On the opposite, the higher frequency textured patterns which was associated with low grade lesions might reflect residual contrast material, especially at tumor edges, revealing tumor substructures. Since, healthier tissue will enhance more than highly cellular cancerous tissue, this might explain textured patterns caught on outer rims of endometrial tumors as predictors of benignity.

Examples of reconstruction and important harmonic filters in high grade and low grade lesions MI are shown on **Figures 5.23** and **5.24**, respectively. Harmonic filters predictors of high grade show smooth uniform appearance, while those predictors of low grade show textured repeated patterns.

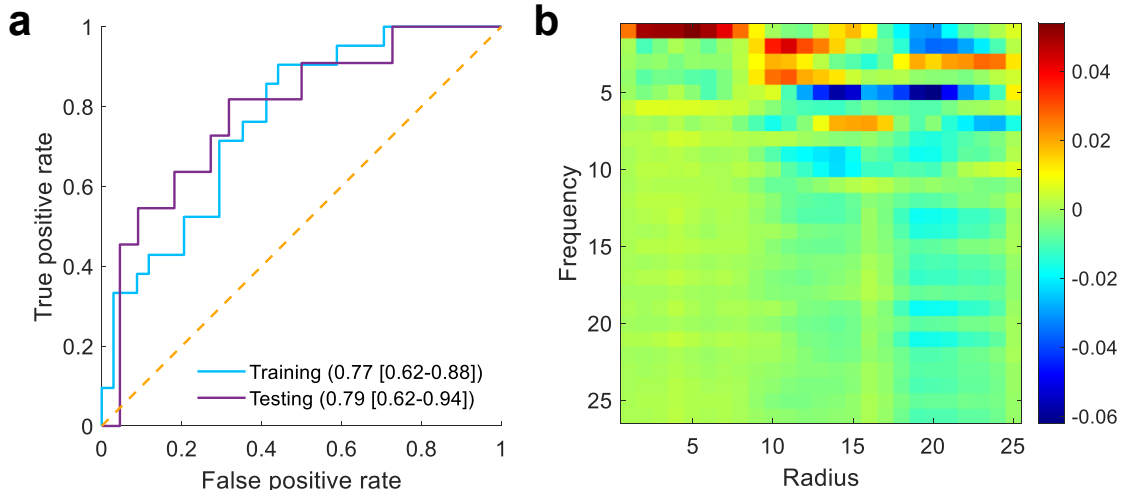


Figure 5.22 (a) Training and testing ROC curves for differentiating low from high grade endometrial cancer with SPHARM descriptors on delayed phase DCE-MRI, and (b) classification matrix picked up by TensorReg modeling.

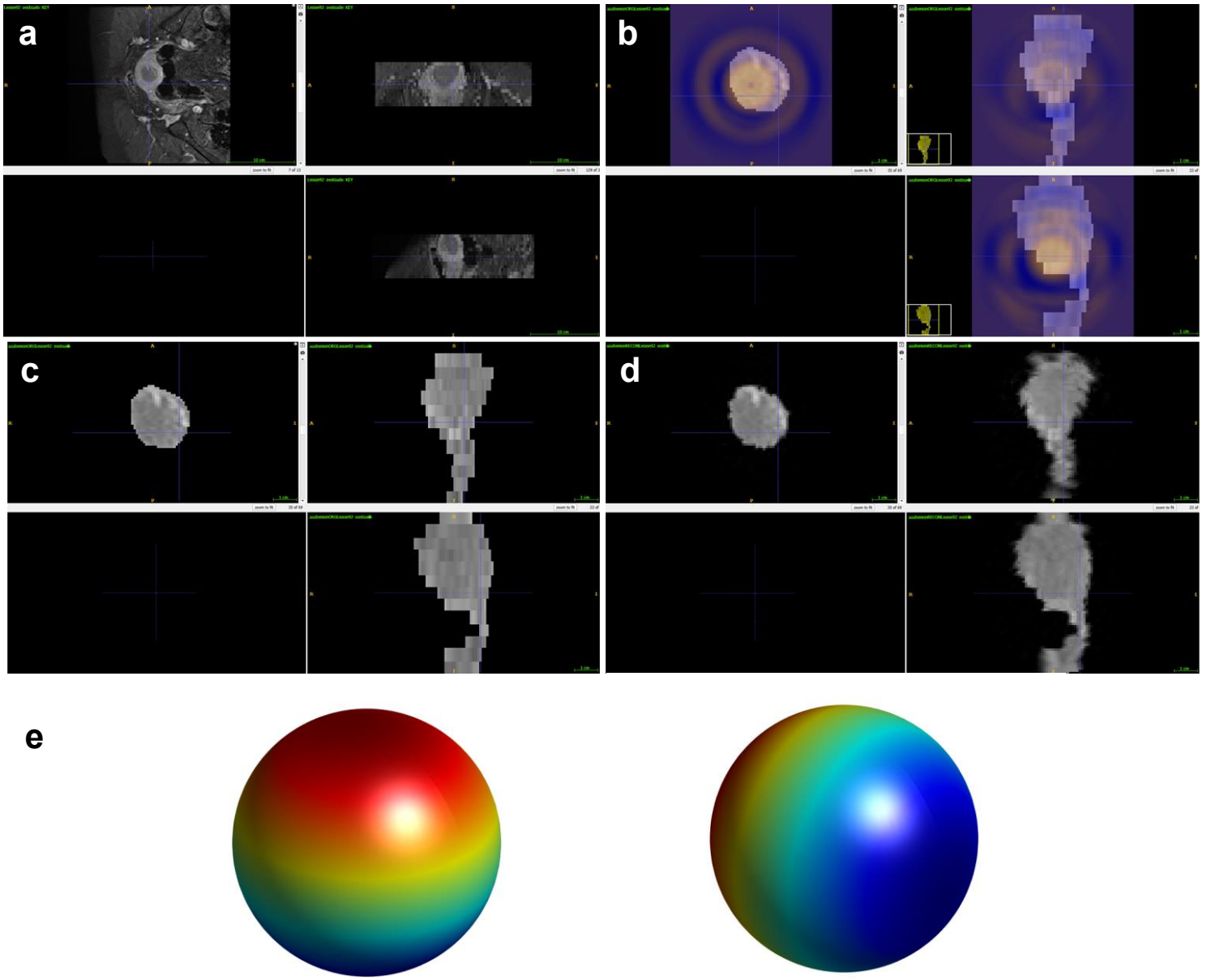


Figure 5.23 (a) Delayed phase DCE-MRI of endometrial tumor with high grade, (b) segmented tumor with overlayed full reconstructed classification matrix of SPHARM descriptor, (c) segmented delayed phase DCE-MRI and (d) its reconstruction with SPHARM coefficients (L_{max}, R_{max} = 25, 25), and (e) harmonic filters (l, m, r = 1, 0, 6) and (l, m, r = 1, 1, 8), expressed within the spherical yellow overlay in panel (b) as predictors of high grade.

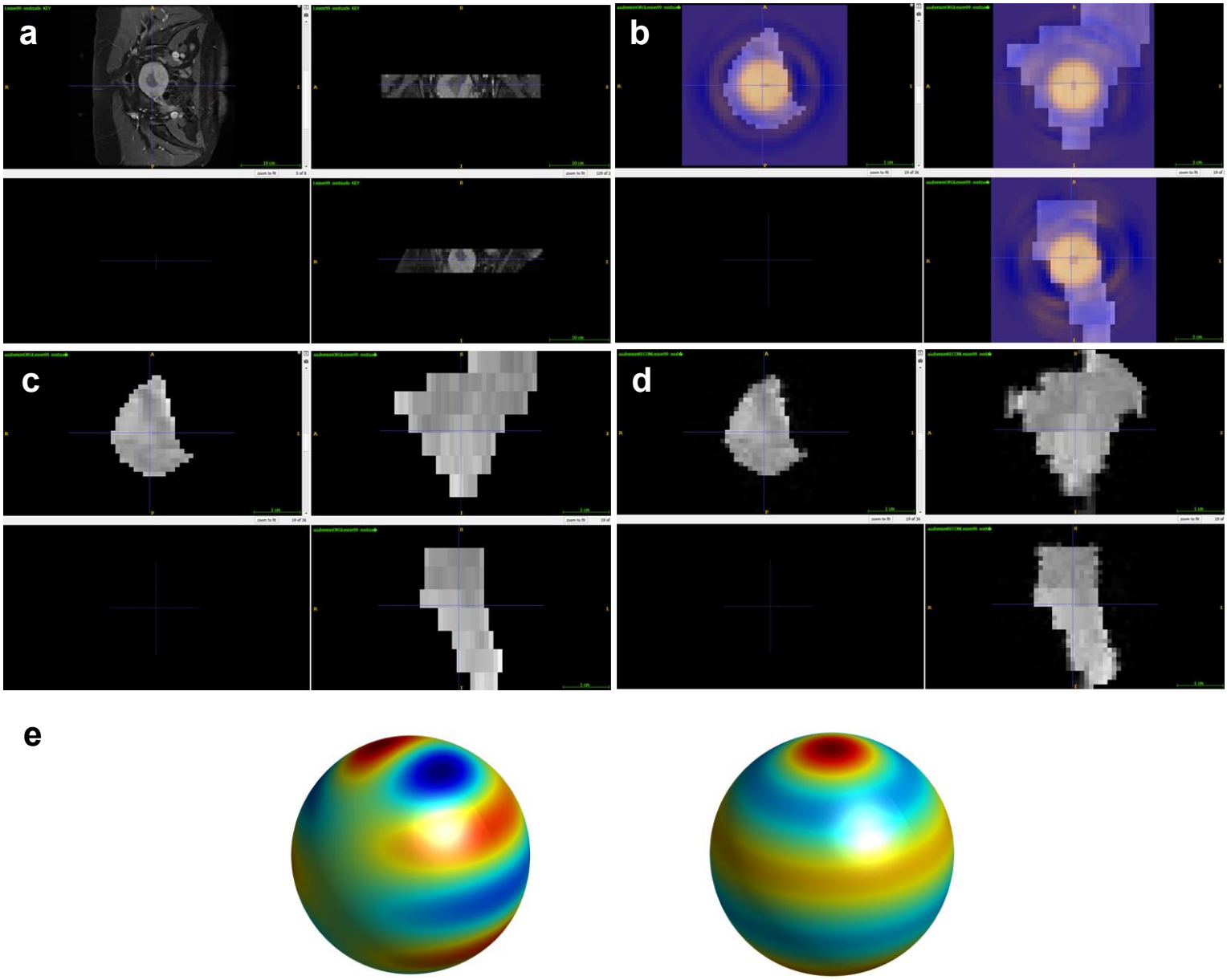


Figure 5.24 (a) Delayed phase DCE-MRI of endometrial tumor with low grade, (b) segmented tumor with overlayed full reconstructed classification matrix of SPHARM descriptor, (c) segmented delayed phase DCE-MRI and (d) its reconstruction with SPHARM coefficients (L_{max}, R_{max} = 25, 25), and (e) harmonic filters (l, m, r = 5, 0, 15) and (l, m, r = 5, 1, 20), expressed within the spherical blue overlay in panel (b) as predictors of low grade.

5.3.3. Analyses in Spherical Volumes of Interest

Radiomics and SPHARM Analyses for Predicting Deep MI

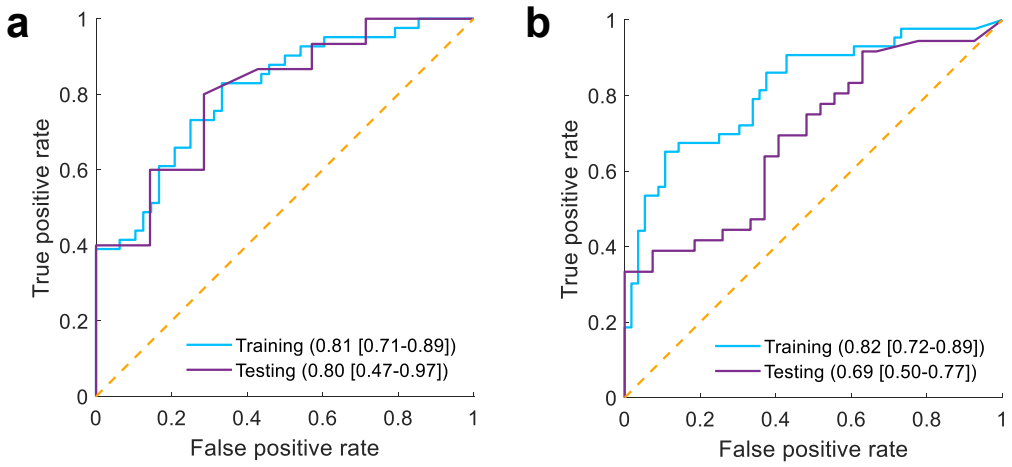


Figure 5.25 Training and testing ROC curves for predicting deep myometrial invasion with (a) SPHARM descriptors and (b) radiomics features combining all MRI sequences in spherical VOIs extended from manual segmentations.

After performing analyses in single contrasts, we assessed the ability of both methods to perform in spherical VOIs extended from expert manual segmentations using all MR sequences. Previously developed models for both computational methods were applied again without any changes, using the same 5 most important radiomics features with the same random forest classifier and the same logistic regression model based on TensorReg predictions from SPHARM decompositions of MR contrasts.

The average Dice coefficient across the population between segmented VOIs and extended spherical VOIs was 0.71 ± 0.12 (range: 0.63-0.93) for ADC maps, 0.70 ± 0.13 (0.60-0.92) for DWI, 0.67 ± 0.15 (0.54-0.91) for T2-weighted MRI, 0.61 ± 0.12 (0.40-0.92) for second phase DCE-MRI, 0.61 ± 0.16 (0.38-0.88) for third phase DCE-MRI, and 0.71 ± 0.12 (0.52-0.94) for delayed phase DCE-MRI. These poor Dice coefficients indicated significant region changes from manual to spherical VOI.

For predicting deep MI, there was a significant decrease in performance going from expert segmentations to spherical VOIs for SPHARM-based modeling (testing AUC = 0.94 vs. 0.80) and for radiomics-based modeling (testing AUC = 0.81 vs. 0.69) (**Figure 5.25**; **Table 5.14**).

Interestingly, even when using these rougher segmentations including surrounding tissue around endometrial lesions, the SPHARM decomposition method still robustly performed with good AUCs from training to testing. This was not observed for radiomics-based modeling which showed poor performance on the testing dataset. This meant that the 5 most important features selected, ADC maps-based maximum 2D diameter slice, and delayed phase DCE-MRI-based least axis length, zone entropy (GLSZM), dependence entropy (GLDM), and dependence nonuniformity (GLDM), changed nonlinearly from segmented VOIs to spherical VOIs. This can be understood especially for texture features which are computed from interrelationships of voxel intensities. Hence, when including signal from neighboring voxels outside the tumor which have a completely different appearance on delayed phase DCE-MRI (*i.e.* healthy endometrium and myometrium compared to tumoral tissue), the textural representation matrices from which texture features are extracted must have changed accordingly. However, we could have expected that shape-based feature would not have been altered as much, even though maximum 2D diameter slice and least axis length values evaluated on a sphere become redundant information in the model.

Table 5.14 Detailed diagnostic performance of SPHARM descriptors and radiomics features for predicting deep MI on mpMRI in spherical VOIs.

		AUC	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	**PPV (%)**	**NPV (%)**
DMI	SPHARM, Training	0.81 (0.71–0.89)	77 (52–93)	79 (62–97)	79 (64–86)	82 (64–98)	69 (51–84)
	SPHARM, Testing	0.80 (0.47–0.97)	81 (55–94)	73 (45–100)	78 (62–91)	86 (58–100)	65 (38–90)
	Radiomics, Training	0.82 (0.72–0.89)	84 (72–93)	63 (48–78)	75 (64–81)	71 (58–80)	80 (66–91)
	Radiomics, Testing	0.69 (0.50–0.77)	77 (58–92)	45 (33–65)	63 (50–77)	42 (25–58)	84 (63–94)

Radiomics and SPHARM Analyses for Predicting High Grade

For differentiating low grade from high grade tumors, there was an increase in diagnostic performance going from expert segmentations to spherical VOIs for the SPHARM-based model (testing AUC = 0.81 vs. 0.87) (**Figure 5.26**; **Table 5.15**). This appeared very curious compared to what was observed for predicting deep MI, and in the previous study on CE-CT of renal cysts. The performance of the radiomics-based model decreased as seen for deep MI but to a greater extent, turning the random forest classifier into a random classifier with testing AUC approaching 0.50

(testing AUC = 0.74 vs. 0.52). Since only texture features were extracted for predicting high grade, signal changes associated with the inclusion of healthy myometrium and endometrium in the spherical VOIs changed nonlinearly extracted features, even though they were deemed reproducible under similar VOI changes initially. However, these selected texture features were not highly discriminating in the first place, as shown in pair plots on **Figure 5.11**.

The increased performance of the SPHARM-based TensorReg model in spherical VOIs for predicting high grade endometrial cancer emerged from the robustness of the regularized logistic regression model combining SPHARM predictions from each sequence. The most important sequence was found to be the delayed phase of DCE-MRI on which good contrast is obtained between lesions and surrounding healthy myometrium and endometrium. Thus, if this contrast was high on most sequences, SPHARM decompositions might have been able to captured the shape and intensity differences related to tissue pertaining to the tumor compared to that of surrounding tissues and enhanced this tumoral information based on the previous modeling on segmented tumors. Moreover, since SPHARM coefficients which were highly predictive of malignancy on delayed phase DCE-MRI were mainly located on inner radii of endometrial lesions as shown on **Figure 5.22**, most of this information was not impacted and still captured during classification of low grade from high grade endometrial tumors using SPHARM descriptors.

Table 5.15 Detailed diagnostic performance of SPHARM descriptors and radiomics features for predicting high grade on mpMRI in spherical VOIs.

		AUC	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	**PPV (%)**	**NPV (%)**
High grade	SPHARM, Training	0.91 (0.86–0.96)	82 (66–97)	92 (81–100)	88 (76–96)	94 (74–100)	72 (56–86)
	SPHARM, Testing	0.87 (0.76–0.97)	72 (56–94)	99 (82–100)	86 (70–99)	81 (55–100)	78 (58–89)
	Radiomics, Training	0.62 (0.49–0.73)	88 (72–97)	41 (31–53)	58 (35–68)	44 (32–56)	87 (70–97)
	Radiomics, Testing	0.52 (0.34–0.67)	80 (55–100)	38 (26–53)	48 (35–61)	29 (15–46)	85 (66–96)

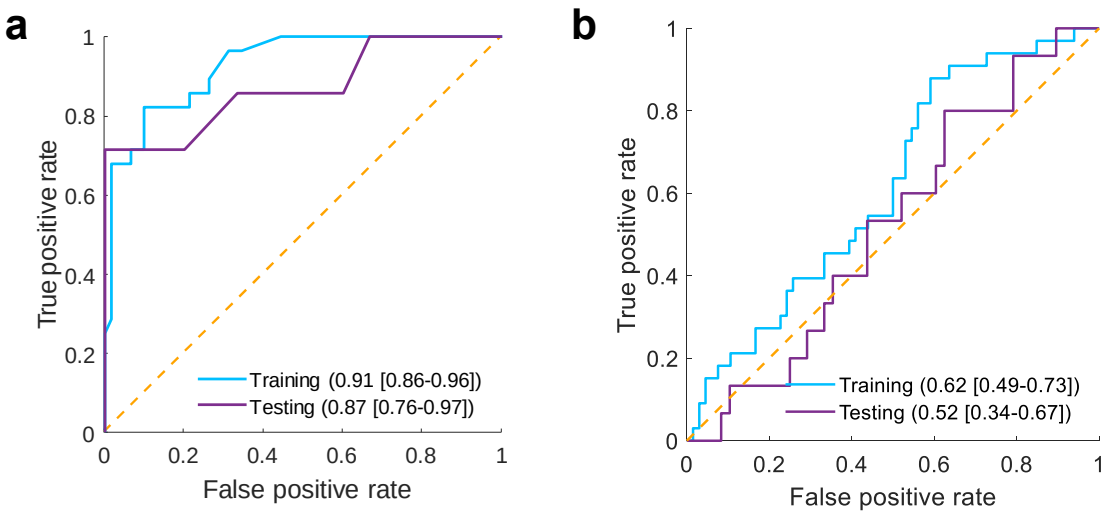


Figure 5.26 Training and testing ROC curves for predicting high grade endometrial cancer with (a) SPHARM descriptors and (b) radiomics features combining all MR sequences in spherical segmentations extended from manual segmentations.

5.4. Summary of Findings and Discussion

Summary

Radiomics- and SPHARM-based diagnostic models have been developed in this thesis for assessing histopathological features of endometrial cancer on mpMRI. We focused on standardized validation and reproducibility of these two computational methods, notably by using an independent testing dataset, confirming the robustness of radiomics and SPHARM pipelines. Indeed, models applied on the testing dataset from another center demonstrated consistently clinically acceptable performance. For the prediction of deep MI, combined shape and texture features extracted from ADC maps and delayed phase DCE-MRI resulted in high diagnostic performance, similar to that of SPHARM decomposition. For predicting high grade, combination of texture features extracted from almost all available MR contrasts resulted in good diagnostic performance, while SPHARM decomposition coefficients provided higher performance on a single contrast, the delayed phase DCE-MRI. The SPHARM-based modeling provided good diagnostic performance for predicting deep MI or high grade endometrial cancer in spherical VOIs, while the diagnostic performance of the radiomics-based modeling was significantly decreased when tested in these rougher and approximative VOIs.

Reproducibility

There were no radiomics features extracted from T2-weighted images excluded from analyses. In fact, T2-weighted MRI-extracted features showed high reproducibility given the preprocessing steps selected. This is consistent with a study by Bianchini *et al* which reported excellent reproducibility of radiomics features when extracted from T2-weighted images acquired with small differences in echo time (TE) (< 5 ms) and repetition time (TR) (< 600 ms) (86). In fact, our study's inclusion criteria reflected this, which might explain the high overall reproducibility of T2-weighted MRI-based radiomics features (**Table 5.2**). TE and TR changes alter the weighting between T1 and T2 relaxation times and modify image contrast, which lead to non-linear signal differences that might not be removable even through image normalization (86). Thus, choosing narrow ranges of TE and TR in the inclusion criteria of radiomics studies might lead to better reproducibility. The choice of optimized preprocessing steps increased the reproducibility of radiomics features across MR sequences, similar to what was previously reported (88). For example, down-sampling T2-weighted images to a isotropic size of 3 x 3 x 3 mm^3 with bin width of 20 would have resulted in features exclusion, going from average ICC (± standard deviation) of 0.957 ± 0.039 to 0.931 ± 0.076 (**Figure 5.3**; **Figure 5.4**). For features extracted from DCE-MRI phases, only one radiomics feature was excluded for second phase (60 s) and third phase (120 s) DCE-MRI acquired in sagittal view, showing overall excellent reproducibility (average ICC = 0.959 ± 0.033 and 0.956 ± 0.036, respectively). However, this was not observed for the delayed phase (240 s) DCE-MRI acquired in axial oblique view, for which a total of 15 features were excluded (average ICC = 0.909 ± 0.084). The same T1-weighting than for second and third phase DCE-MRI was used for acquiring images at the delayed phase. This suggests that the change in acquisition plane might have impacted features reproducibility (sagittal vs. axial oblique). Since the axial oblique plane is manually selected to be perpendicular to the endometrial cavity on the scout scan during acquisitions, there is an operator dependence introduced in this step of the MR examination. This would be consistent with the lower feature reproducibility also observed for ADC and DWI (average ICC = 0.885 ± 0.102 and 0.872 ± 0.142; and excluded features = 19 and 20; respectively) which were also reconstructed with an axial oblique view. Even though some features had to be excluded because of poor reproducibility, the overall ICC for each sequence was at least very good on MR

contrasts of endometrial cancer (mean ICC > 0.85). Preprocessing of images and z-score normalization of features before random forest modeling enabled us to provide comparable features across institutions, as shown by the proposed robustly performing models from training to testing on datasets from different institutions (102, 103). Moreover, controlling hyperparameters of random forest classifiers minimized overfitting on the training dataset, providing good translation from training to testing. In fact, leaving the random forest algorithm building as many trees of uncontrolled depth will necessarily lead to high diagnostic performance during the training phase, with "pure" samples of each binary category at the leaf level of each tree. This hyperparameter control step should be carefully performed whenever using ensemble learners such as random forests.

SPHARM decompositions were performed according to previously reported methods on volumetric shapes characterization (117, 127) and according to our preliminary results on volumetric texture benchmark datasets. SPHARM also avoided several of the preprocessing steps associated with radiomics studies, not just the resampling to isotropic voxel size, but also the signal discretization in bins, as detailed previously in **Chapter 4**. High MSSI were obtained for SPHARM descriptors-based reconstructions of endometrial lesions across all MR sequences. This indicated that SPHARM descriptors could accurately capture and encode structural and textural information in SPHARM coefficients given the selected frequency and radial expansions of decompositions, and even reconstruct these structures. Further standardization of SPHARM descriptors will be required to assess the impact of MR acquisition parameters on the SPHARM decomposition method. If coefficients could robustly encode the structural and textural information from 3D volumetric images acquired with different TE or TR, or with spin echo vs. gradient echo sequences, these quantitative imaging biomarkers could provide significant clinical value.

To address the need for standardization in imaging biomarkers extracted from MRI data in oncology, a task group report from the American Association of Physicists in Medicine (AAPM) is currently being written at the time of the redaction of this thesis, which is expected to be published before 2020-12-31 (AAPM Task Group 294 Magnetic Resonance Biomarkers in Radiation Oncology, aapm.org/org/structure/?committee_code=TG294). Expert

recommendations from groups like these will allow further harmonization of mpMRI-based radiomics or SPHARM biomarkers of different cancer sites, including endometrial cancer.

Radiomics and SPHARM Analyses

With SPHARM decomposition coefficients, the diagnostic performance was high for detecting deep MI using all MR contrasts, but also when using only ADC maps or second phase DCE-MRI. On ADC maps which are often related to tissue cellularity (41, 184), SPHARM coefficients which were the most predictive of deep MI were on higher radii and at higher frequencies, *i.e.* more textured, while coefficients predictors of the absence of deep MI were at more inner radii and were associated with lower frequency content, *i.e.* more uniform. In fact, lower order harmonics at lower radius seemed to capture the overall more uniform appearance of smaller endometrial lesions without deep MI. For detecting deep MI, higher order SPHARM decomposition coefficients were required to fully capture the radiological signature of MI. This is consistent with findings from our radiomics analyses in which textural features were also important, as extracted from delayed phase DCE-MRI, but also morphological features since larger tumors were associated with greater depth of MI also extracted from ADC maps. Even though some prior reports reported good diagnostic performance for mean ADC in tumors' VOI for predicting deep MI (195, 199), mean ADC was not represented in the group of predictive features for this outcome in **Table 5.4**. In fact, the mean ADC obtained in our population was not predictive of MI, similar to what Rechichi G *et al* also reported (200). Again, in radiomics analyses, the set of 5 most discriminative features for predicting deep MI were texture and shape features extracted from either ADC maps or delayed phase DCE-MRI.

SPHARM decompositions on second phase DCE-MRI were even more predictive of the depth of MI. This phase of DCE-MRI is the actual phase used clinically to evaluate the extent of MI, since high contrast is obtained between the myometrium and the lesion (184). Thus, it is conceptually even more clinically relevant to perform our analyses on this specific time point of the DCE-MRI series. The high diagnostic performance revealed by SPHARM coefficients was not observed in radiomics features which led to poor classification of deep MI when using solely features extracted from second phase DCE-MRI. Although high feature reproducibility

was seen for radiomics as extracted on this phase, selected features did not reflect the pathological signature of deep MI even though they were shape-based and textural features, similar to what was found in the mpMRI radiomics model. Thus, SPHARM decomposition was the most accurate method for detecting deep MI on ADC maps (testing AUC = 0.85) and on second phase DCE-MRI (testing AUC = 0.86), and increased diagnostic performance was observed when combining predictions from all MR contrasts (testing AUC = 0.94). Since the depth of MI is known to be the most important morphologic prognostic factor in endometrial cancer, an accurate and robust preoperative mpMRI-based quantitative biomarker could provide clinical benefit in patients stratification for surgery (41, 184). Another advantage of the SPHARM method is that the orientation of the invasion of the lesion in the myometrium will not affect its diagnostic performance since SPHARM coefficients are rotation-invariant by construction. Hence, lesions can be compared with this technique from one patient to another, no matter the lesion's orientation.

Clinically, the FIGO grade of endometrial tumors is assessed based on morphological features seen on T2-weighted MRI and on DCE-MRI where excellent contrast is obtained between hyperenhancing myometrium and nonenhencing endometrial tumors (41, 184). Since morphological features are generally used in the clinical setting for tumor staging, it might explain why the radiomics signature built only from texture features did not provide high diagnostic performance for differentiating low grade from high grade endometrial tumors. Interestingly, there were no shape features included in the set of single most discriminative radiomics features in **Table 5.6**. Some prior studies found good diagnostic performance of radiomics features extracted from ADC maps for differentiating low grade from high grade endometrial cancer (52, 55). In our population, the only feature extracted from ADC maps which was included in the final radiomics model to differentiate high grade from lower grade endometrial tumors was gray-level variance from GLSZM. Interestingly, a recent study by Yamada I *et al* also included ADC-based GLSZM texture features in their final random forest model for differentiating low from high grade endometrial cancer (55), suggesting consistent radiomics-based signature of higher grade tumors. Since ADC maps characterize the extent of tissue cellularity by assessing the impedance of water molecules diffusion, the discriminative capacity of variance of ADC zones decoded the more heterogeneous cellularity seen in high

grade endometrial lesions (41), as assessed by radiomics-based modeling. The mpMRI-based SPHARM decomposition method provided significantly improved diagnostic performance compared to radiomics (testing AUC = 0.81 vs. 0.72 in the subset of the population). In fact, combining predictions of TensorReg models for each MR contrast decomposition resulted in a predictive signature of tumor grade. SPHARM decompositions on single contrast delayed phase DCE-MRI were also very predictive of high grade endometrial lesions, where the homogeneous nonenhancing core of the lesions associated with low frequency content captured in SPHARM descriptors seemed to correlate with high grade tumors while textured enhancing edges with low grade.

Analyses in Spherical Volumes of Interest

Decreases in AUC when changing from expert segmentations to spherical VOIs were observed both for SPHARM and radiomics models for predicting deep MI. This decrease in performance was also observed for radiomics-based modeling for predicting high grade endometrial cancer, but not for SPHARM decompositions for predicting this same outcome which performance increased. During expert segmentations, tumor contour was defined as areas of intermediate signal intensity on T2-weighted images that were different from the normal adjacent low signal intensity myometrium. On DCE-MRI, areas of lower signal intensity in the adjacent myometrium were considered to be the tumor. Restricted diffusion areas represented as high signal intensity on DWI or low signal intensity on ADC maps were also regarded as tumor areas (41). When extending a sphere from these initial contours, no considerations of such signal changes were used since only the longest tumor radius was extended from the segmentation's center of mass. Thus, if the endometrial tumor had an ovoid shape rather than a spherical morphology, then the sphere-shaped segmentation included more surrounding tissue. This explains the low Dice coefficients between manual segmentations and spherical VOIs obtained across MR contrasts. This also explains the observed decreased discriminative power of texture features originally assessed as relevant for each classification task, *i.e.* < 50% MI vs. deep MI and low grade vs. high grade. However, the ability of SPHARM coefficients to accurately differentiate low (1 and 2) from high FIGO grades (3 and non-endometroid type) in spherical VOIs was an interesting finding. In fact, through robust modeling of SPHARM-based

TensorReg predictions from each sequence with a regularized logistic regressor, SPHARM decompositions were able to encode tumoral shape and intensity differently than that of surrounding healthy myometrium and endometrium. As these healthy tissues tend to be more enhanced than the lesion (41, 184), SPHARM seemed to be able to separately describe less enhanced tumor content on DCE-MRI from enhancing healthy tissue, most notably. As predictive SPHARM coefficients for high grade tumors were mainly located in central radii of lesions, this signature was not influenced by including surrounding tissue in the sphere VOI. Textured patterns on outer shells were predictor for low grade endometrial tumors on delayed phase DCE-MRI. Including more tissue in a sphere where high contrast is seen between the lesions and the healthy myometrium might also have increased the performance if these contrast differences were captured as textured patterns by SPHARM decomposition. This contrast between tumoral tissue and healthy tissue also enables the descriptors to use the previously developed signature in segmented tumors as predictive information. In fact, the relevant tumoral signal captured previously was coming from the hypoenhancing region in the novel spherical VOI, while the enhancing regions, now included in the sphere, used to be voxels which values were set to 0. Therefore, the same contrast is seen by the computational method and the radiological signature developed previously by SPHARM coefficients for predicting high grade in segmented lesions is still relevant even in rougher approximative spherical VOIs.

An advantage of spherical VOIs is that they have the potential of drastically reducing inter-observer variability while accelerating analyses. In fact, there are not many ways to position a sphere englobing a tumor based on its maximal diameter. Thus, we can envision that a physician could simply measure the 2D maximal diameter of an endometrial tumor, which is often used as a metric of treatment response (11), and a sphere would automatically be created from this measurement. A SPHARM decomposition of the spherical VOI would then be applied to identify tumoral grade, with the possibility of producing harmonic “heatmaps” of important SPHARM filters on the 3D MR image. Thus, the physician or radiologist could interpret what information was captured by SPHARM descriptors locally on the image, in a similar manner than convolutional filters from convolutional neural networks might be used to interpret important image features characterizing a cancerous lesion picked up during training (136).

Limitations

Our study had the following limitations. First, in the VOIs drawn by two radiologists in consensus, the inter-reader agreement was not evaluated. Further studies are needed to evaluate the interobserver agreement in terms of the VOI placement. In fact, an additional aim of the manuscript in preparation that was not included in this thesis was to compare the interobserver agreement of faster semi-automated segmentations to that of expert time-consuming manual segmentations. These goals were part of my contributions but will only be included in the final version of the manuscript to be submitted. However, analyses in spherical VOIs have shown the importance of extracting information from tumor regions to capture relevant radiomics-based signatures of deep MI or high grade endometrial cancer, which was not observed for the SPHARM method in predicting high grade. Second, the proposed radiomics model did not provide excellent diagnostic accuracy for the prediction of high grade endometrial cancer. Combined textural features identified through random forest modeling could not linearly discriminate low from high grade tumors. Nevertheless, the radiomics-based modeling on mpMRI still demonstrated a stable accuracy in the diagnosis of high grade lesions, from training to testing. However, multi-parametric and single contrast MRI-based SPHARM modeling seemed to be a better diagnostic solution for the staging of endometrial cancer, with increased diagnostic performance compared to that of radiomics. Limitations of the SPHARM decomposition method have been more extensively discussed in the previous chapter. Hence, we refer the reader to this section for further analyses of computational methods' limitations.

Discussion

Deep MI and tumor grade have been widely used to stratify patients according to the risk of local, regional, or distant recurrence, enabling patients' selection for surgeries including pelvic and para-aortic lymphadenectomy (172, 201). Since radiomics features and SPHARM descriptors can provide information regarding these high-risk factors from standard-of-care MRI prior to surgery, it may help with preoperative risk stratification and optimal selection of patients who require more extensive surgery while avoiding overtreatment of low-risk patients (2).

For diagnosis of deep MI and high grade tumors, our study demonstrated comparable diagnostic accuracy with that of the prior study using 2D mpMRI-based radiomics features by our team (81% and 78%, respectively) (18). However, since we did not directly compare 3D radiomics to 2D radiomics, it remains to be evaluated whether models using 3D radiomics features can yield better diagnostic performance than 2D models, even though the former provide more comprehensive characterization of the lesion (10). Moreover, in the previous report by Ueno Y *et al*, the average performance of radiologists for predicting deep MI was estimated at an accuracy of 81% (18). Hence, the radiomics model on the whole population achieved the exact same accuracy (81%) on the whole population and the SPHARM model did even better on the subset of the population (90%). This suggests that such computational methods could provide clinical benefit for preoperative risk stratification in patients with endometrial cancer. To further assess if the inclusion of these methods could really benefit the current clinical pipeline, the agreement of radiologists' staging with that of radiomics and SPHARM will be assessed in the final version of this study to be submitted for journal publication.

While many MR contrasts are generally acquired to provide comprehensive characterization of endometrial cancer (41, 184), using single contrasts has been shown to provide accurate predictions of deep MI and of high grade with the SPHARM decomposition method, especially on single phases from multiphase DCE-MRI. Since MR examinations tend to be time-consuming and expensive, the acquisition of a single MR sequence instead of a full set of clinically available sequences could accelerate clinical workflow while only losing little accuracy as a trade-off. Computational enhancement of underlying pathological information encoded in medical images with proposed methods could be a clinically acceptable and viable option in some settings where scanning time is limited. The same reasoning can be applied on spherical VOIs. In some contexts, not having to spend weeks and months to precisely contour tumors slice-by-slice on numerous MR contrasts (if MRI units are available) may be worth the decrease in performance. Indeed, a simpler model that is clinically efficient might be more translatable than a better performing model that cannot be implemented due to practical time constraints (10), or that can be implemented but only after unreasonable delays which could impact clinical care or delay treatments. This is especially applicable to community or rural

clinics that may lack radiological expertise for accurate tumor delineation or time-consuming scans.

In conclusion, even though more clinical evidence and research efforts for harmonization are required before these computational methods can be used for actual clinical decision-making, this study suggested that radiomics-based and SPHARM-based modeling may become relevant comprehensive tools for preoperative risk stratification in patients with endometrial cancer. Our results showed that volumetric mpMRI-based computational methods could diagnose histopathological features of endometrial cancer, deep MI and high FIGO grade, which are important prognostic factors. Hence, such techniques could offer clinical benefit for preoperative risk stratification in patients with endometrial cancer.

6. CONCLUSION

This thesis presented methodological developments from the field of computer vision applied to the field of diagnostic radiology in clinical retrospective studies. Our contributions are both technical and clinical. The main technical contributions are i) adapting spherical harmonics (SPHARM) decomposition to medical imaging and its comparison to radiomics, and ii) the comparison of fast spherical segmentations to that of expert radiologists' manual segmentations in the diagnosis of histology-defined features of the disease. The main clinical contributions are i) the development of a radiomics pipeline based on state-of-the-art radiomics literature and on the Image biomarkers standardization initiative (IBSI)'s recommendations, and ii) the implementation of SPHARM and radiomics pipelines in two clinical studies, identifying renal cysts malignancy on contrast-enhanced computed tomography, and identifying histopathological features of endometrial cancer on multi-parametric magnetic resonance imaging.

Throughout both studies, we observed the importance of preprocessing of images prior to radiomics features extraction. Radiomics studies should always report these analyses of optimal preprocessing parameters following IBSI's recommendations and literature on reproducibility of these mathematical descriptors. In the SPHARM pipeline, those steps were not necessary given the properties of SPHARM decompositions. More simple and reproducible analyses could make computational methods easier to translate into clinical practice (10). Thus, in the case of radiomics, filtering of images to obtain invariance properties is desirable and might be an interesting recommendation, discussed in IBSI's second chapter (66). However, these local image descriptors, such as filtered images after convolutional filtering used in other radiomics studies, were not studied in this work given the lack of standardization guidelines at the time of redaction of this thesis. Now that the publication of a first version of IBSI's second chapter on these analyses is out in preprint, there should be further explorations of these computational approaches and their properties for radiomics studies, while focusing on harmonization (66). Furthermore, we have shown that it is possible in some cases to reduce the burden associated with long multi-sequence MRI scans and time-consuming manual segmentations, by using single MR contrast or a simple spherical volume of interest with the

proposed SPHARM decomposition method. In clinical practice, since time is money (202), it is of high importance for clinical translation that such computer-aided diagnosis techniques relieve the burdened radiologists' and radio-oncologists' workload (14, 15), especially in settings where radiological expertise may be lacking and/or resources may be limited, such as in developing countries.

The proposed methods also have the potential of addressing a clinical need in risk stratification of low-risk patients. Currently, it is not clinically feasible to have histopathological confirmation for each patient with preoperative image-based diagnosis of low-risk renal cyst or endometrial lesion since it is costly and would require an invasive surgical procedure or a biopsy. Thus, if a quantitative imaging biomarker highly associated with disease's grade or malignancy supports the radiological finding that a given lesion has a lower risk of progression, then this patient will not be undergoing such procedures since both assessments are concordant. On the opposite, a lesion with radiological assessment indicating low risk but with a quantitative biomarker suggesting high risk will be sent for a biopsy to obtain histopathological analyses of this lesion to inform further decisions. The proposed computational methods have been shown to relate with high-risk factors from standard CE-CT and MR images before surgery. Thus, they could provide support with risk stratification early in the clinical decision pipeline for optimal patient selection for surgery and to avoid overtreatment of low-risk lesions with concordant radiological and computational assessments.

Otherwise, the field of radiomics is moving towards deep learning-based features extraction (47, 203). In fact, radiomics features can be implemented as part of a deep learning framework, combining predictions from three-dimensional convolutional neural networks with clinical outcomes to construct more complex and predictive tumor signatures, but also to preprocess or adapt image resolution prior to features extraction (72). Comprehensive predictions integrating features from clinical data and deep learning approaches to characterize tumor regions could also provide insight on which part of the tumor lead to the diagnosis made, using convolutional filters, hence addressing the interpretability issue of radiomics studies by providing information localization in volumes of interest (136), analogous to what was proposed with the SPHARM pipeline. Such methods could also encompass inter-lesions heterogeneity and not only intra-tumoral heterogeneity of the primary tumor – as we did throughout this thesis

– to encompass all important clinical variables and to enable precision oncology (1). Implementing radiogenomics in future computational pipelines could also improve diagnoses and add another dimension based on genomic signatures of pathological features of cancer (204). For SPHARM decompositions, future work should investigate the use of radial basis functions instead of shell sampling at equispaced radii, to provide 3D mathematical descriptors of tumors independent tumor size, thus not limited to larger tumors. Computational approaches like the one proposed by Galinsky VL and Frank LR combining SPHARM decompositions as angular basis functions and 3D spherical Bessel functions as radial basis functions could better characterize 3D volumes as a whole, at different angular and radial frequency expansions, while avoiding the need for sampling of shells (120).

Finally, across the two clinical studies explored in this thesis, we have discussed the need for standardization of radiomics features extraction and analyses. However, there is also an urging need for harmonizing the reporting of results in radiomics studies (205). Following guidelines from consortiums such as the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) initiative is required for the translation of radiomics into the clinic and to enable precision oncology (206). This is also a strong argument for open science since we will need to have access to computational methods developed by different research groups and to more clinical imaging data from multiple institutions to build models and compare their efficacy across centers and modalities. The development and implementation of surrogate imaging biomarkers such as those discussed in this thesis will require multi-centre technical and clinical validation following clear guidelines. Thus, there are exciting opportunities for medical physicists in imaging biomarkers metrology awaiting!

REFERENCES

1. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015;372(9):793-5.
2. Sala E, Rockall AG, Freeman SJ, Mitchell DG, Reinhold C. The added role of MR imaging in treatment stratification of patients with gynecologic malignancies: what the radiologist needs to know. Radiology. 2013;266(3):717-40.
3. DeVita Jr VT, Lawrence T, Rosenberg SA. Cancer: principles & practice of oncology: annual advances in oncology: Lippincott Williams & Wilkins, 2012.
4. Harry VN, Semple SI, Parkin DE, Gilbert FJ. Use of new imaging techniques to predict tumour response to therapy. Lancet Oncol. 2010;11(1):92-102.
5. Kyle SD, Law WP, Miles KA. Predicting tumour response. Cancer Imaging. 2013;13(3):381-90.
6. O'Connor JP, Aboagye EO, Adams JE, et al. Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol. 2017;14(3):169-86.
7. Amin S, Bathe OF. Response biomarkers: re-envisioning the approach to tailoring drug therapy for cancer. BMC Cancer. 2016;16(1):850.
8. Aerts HJ. The Potential of Radiomic-Based Phenotyping in Precision Medicine: A Review. JAMA Oncol. 2016;2(12):1636-42.
9. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.
10. Savadjiev P, Chong J, Dohan A, et al. Image-based biomarkers for solid tumor quantification. Eur Radiol. 2019;29(10):5431-40.
11. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45(2):228-47.
12. Gerlinger M, Rowan AJ, Horswell S, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012;366(10):883-92.
13. Choy G, Khalilzadeh O, Michalski M, et al. Current Applications and Future Impact of Machine Learning in Radiology. Radiology. 2018;288(2):318-28.
14. Bhargavan M, Kaye AH, Forman HP, Sunshine JH. Workload of radiologists in United States in 2006-2007 and trends since 1991-1992. Radiology. 2009;252(2):458-67.
15. Khan SH, Hedges WP. Workload of consultant radiologists in a large DGH and how it compares to international benchmarks. Clin Radiol. 2013;68(5):e239-44.
16. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications. 2014;5:4006.
17. Kazhdan M, Funkhouser T, Rusinkiewicz S. Rotation invariant spherical harmonic representation of 3D shape descriptors. Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on Geometry processing: Eurographics Association, 2003; p. 156-64.
18. Ueno Y, Forghani B, Forghani R, et al. Endometrial Carcinoma: MR Imaging-based Texture Model for Preoperative Risk Stratification-A Preliminary Analysis. Radiology. 2017;284(3):748-57.
19. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.
20. Huang Y, Liu Z, He L, et al. Radiomics Signature: A Potential Biomarker for the Prediction of Disease-Free Survival in Early-Stage (I or II) Non-Small Cell Lung Cancer. Radiology. 2016;281(3):947-57.
21. Larue RT, Defraene G, De Ruysscher D, Lambin P, van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. Br J Radiol. 2017;90(1070):20160665.
22. Savadjiev P, Chong J, Dohan A, et al. Demystification of AI-driven medical image interpretation: past, present and future. Eur Radiol. 2019;29(3):1616-24.
23. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016;278(2):563-77.
24. Fournier LS, Oudard S, Thiam R, et al. Metastatic renal carcinoma: evaluation of antiangiogenic therapy with dynamic contrast-enhanced CT. Radiology. 2010;256(2):511-8.
25. Ronot M, Clift AK, Vilgrain V, Frilling A. Functional imaging in liver tumours. J Hepatol. 2016;65(5):1017-30.

26. Lo Gullo R, Daimiel I, Morris EA, Pinker K. Combining molecular and imaging metrics in cancer: radiogenomics. Insights into Imaging. 2020;11(1):1.
27. Poleszczuk J, Enderling H. Tumor Composition Depends on the Viewing Angle. In: Matthäus F, Matthäus S, Harris S, Hillen T, eds. The Art of Theoretical Biology. Cham: Springer International Publishing, 2020; p. 130-1.
28. Poleszczuk J, Hahnfeldt P, Enderling H. Evolution and phenotypic selection of cancer stem cells. PLoS Comput Biol. 2015;11(3):e1004025.
29. Wakabayashi T, Ouhmich F, Gonzalez-Cabrera C, et al. Radiomics in hepatocellular carcinoma: a quantitative review. Hepatol Int. 2019;13(5):546-59.
30. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int J Radiat Oncol Biol Phys. 2018;102(4):1143-58.
31. Ursprung S, Beer L, Bruining A, et al. Radiomics of computed tomography and magnetic resonance imaging in renal cell carcinoma-a systematic review and meta-analysis. Eur Radiol. 2020.
32. Lubner MG, Smith AD, Sandrasegaran K, Sahani DV, Pickhardt PJ. CT Texture Analysis: Definitions, Applications, Biologic Correlates, and Challenges. Radiographics. 2017;37(5):1483-503.
33. Hu W, Yang H, Xu H, Mao Y. Radiomics based on artificial intelligence in liver diseases: where we are? Gastroenterol Rep (Oxf). 2020;8(2):90-7.
34. Liu YI, Shin LK, Jeffrey RB, Kamaya A. Quantitatively defining washout in hepatocellular carcinoma. AJR Am J Roentgenol. 2013;200(1):84-9.
35. Bae KT. Intravenous contrast medium administration and scan timing at CT: considerations and approaches. Radiology. 2010;256(1):32-61.
36. McEvoy SH, McCarthy CJ, Lavelle LP, et al. Hepatocellular carcinoma: illustrated guide to systematic radiologic diagnosis and staging according to guidelines of the American Association for the Study of Liver Diseases. Radiographics. 2013;33(6):1653-68.
37. Xu X, Zhang HL, Liu QP, et al. Radiomic analysis of contrast-enhanced CT predicts microvascular invasion and outcome in hepatocellular carcinoma. J Hepatol. 2019;70(6):1133-44.
38. Kocak B, Durmaz ES, Erdim C, Ates E, Kaya OK, Kilickesmez O. Radiomics of Renal Masses: Systematic Review of Reproducibility and Validation Strategies. AJR Am J Roentgenol. 2020;214(1):129-36.
39. Silverman SG, Pedrosa I, Ellis JH, et al. Bosniak Classification of Cystic Renal Masses, Version 2019: An Update Proposal and Needs Assessment. Radiology. 2019;292(2):475-88.
40. Silverman SG, Israel GM, Herts BR, Richie JP. Management of the incidental renal mass. Radiology. 2008;249(1):16-31.
41. Beddy P, O'Neill AC, Yamamoto AK, Addley HC, Reinhold C, Sala E. FIGO staging system for endometrial cancer: added benefits of MR imaging. Radiographics. 2012;32(1):241-54.
42. Bi Q, Chen Y, Wu K, et al. The Diagnostic Value of MRI for Preoperative Staging in Patients with Endometrial Cancer: A Meta-Analysis. Acad Radiol. 2019.
43. Gupta RT, Kauffman CR, Polascik TJ, Taneja SS, Rosenkrantz AB. The state of prostate MRI in 2013. Oncology (Williston Park). 2013;27(4):262-70.
44. Lugauer F, Wetzl J. Magnetic Resonance Imaging. In: Maier A, Steidl S, Christlein V, Hornegger J, eds. Medical Imaging Systems: An Introductory Guide. Cham: Springer International Publishing, 2018; p. 91-118.
45. Sun Y, Reynolds HM, Parameswaran B, et al. Multiparametric MRI and radiomics in prostate cancer: a review. Australasian Physical & Engineering Sciences in Medicine. 2019;42(1):3-25.
46. Schwier M, van Griethuysen J, Vangel MG, et al. Repeatability of Multiparametric Prostate MRI Radiomics Features. Sci Rep. 2019;9(1):9441.
47. Khalvati F, Zhang J, Chung AG, Shafiee MJ, Wong A, Haider MA. MPCaD: a multi-scale radiomics-driven framework for automated prostate cancer localization and detection. BMC Medical Imaging. 2018;18(1):16.
48. Schick U, Lucia F, Dissaux G, et al. MRI-derived radiomics: methodology and clinical applications in the field of pelvic oncology. The British Journal of Radiology. 2019;92(1104):20190105.
49. Horvat N, Bates DDB, Petkovska I. Novel imaging techniques of rectal cancer: what do radiomics and radiogenomics have to offer? A literature review. Abdominal Radiology. 2019;44(11):3764-74.
50. Ghosh A, Singh T, Singla V, Bagga R, Srinivasan R, Khandelwal N. DTI histogram parameters correlate with the extent of myoinvasion and tumor type in endometrial carcinoma: a preliminary analysis. Acta Radiol. 2020;61(5):675-84.
51. Luo Y, Mei D, Gong J, Zuo M, Guo X. Multiparametric MRI-Based Radiomics Nomogram for Predicting Lymphovascular Space Invasion in Endometrial Carcinoma. J Magn Reson Imaging. 2020.

52. Nougaret S, Reinhold C, Alsharif SS, et al. Endometrial Cancer: Combined MR Volumetry and Diffusion-weighted Imaging for Assessment of Myometrial and Lymphovascular Invasion and Tumor Grade. Radiology. 2015;276(3):797-808.
53. Stanzione A, Cuocolo R, Del Grosso R, et al. Deep Myometrial Infiltration of Endometrial Cancer on MRI: A Radiomics-Powered Machine Learning Pilot Study. Acad Radiol. 2020.
54. Xu X, Li H, Wang S, et al. Multiplanar MRI-Based Predictive Model for Preoperative Assessment of Lymph Node Metastasis in Endometrial Cancer. Front Oncol. 2019;9:1007.
55. Yamada I, Miyasaka N, Kobayashi D, et al. Endometrial Carcinoma: Texture Analysis of Apparent Diffusion Coefficient Maps and Its Correlation with Histopathologic Findings and Prognosis. Radiology: Imaging Cancer. 2019;1(2):e190054.
56. Yan B, Liang X, Zhao T, Ding C, Zhang M. Is the standard deviation of the apparent diffusion coefficient a potential tool for the preoperative prediction of tumor grade in endometrial cancer? Acta Radiol. 2020:284185120915596.
57. Ytre-Hauge S, Dybvik JA, Lundervold A, et al. Preoperative tumor texture analysis on MRI predicts high-risk disease and reduced survival in endometrial cancer. J Magn Reson Imaging. 2018;48(6):1637-47.
58. Lecler A, Duron L, Balvay D, et al. Combining Multiple Magnetic Resonance Imaging Sequences Provides Independent Reproducible Radiomics Features. Sci Rep. 2019;9(1):2068.
59. Schick U, Lucia F, Dissaux G, et al. MRI-derived radiomics: methodology and clinical applications in the field of pelvic oncology. Br J Radiol. 2019;92(1104):20190105.
60. Vallières M, Zwanenburg A, Badic B, Cheze Le Rest C, Visvikis D, Hatt M. Responsible Radiomics Research for Faster Clinical Translation. J Nucl Med. 2018;59(2):189-93.
61. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017;14(12):749-62.
62. Papanikolaou N, Matos C, Koh DM. How to develop a meaningful radiomic signature for clinical use in oncologic patients. Cancer Imaging. 2020;20(1):33.
63. Yip SS, Aerts HJ. Applications and limitations of radiomics. Phys Med Biol. 2016;61(13):R150-66.
64. Fornacon-Wood I, Mistry H, Ackermann CJ, et al. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. European Radiology. 2020.
65. Zwanenburg A, Vallieres M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020;295(2):328-38.
66. Depeursinge A, Andrearczyk V, Whybra P, et al. Standardised convolutional filtering for radiomics. arXiv preprint arXiv:200605470. 2020.
67. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics. 1973;SMC-3(6):610-21.
68. Galloway MM. Texture analysis using gray level run lengths. Computer Graphics and Image Processing. 1975;4(2):172-9.
69. Thibault G, Fertil B, Navarro C, L., et al. Texture indexes and gray level size zone matrix. Application to cell nuclei classification. 10th International Conference on Pattern Recognition and Information Processing, PRIP 2009. Minsk, Belarus2009; p. 140--5.
70. Sun C, Wee WG. Neighboring gray level dependence matrix for texture classification. Computer Vision, Graphics, and Image Processing. 1983;23(3):341-52.
71. Amadasun M, King R. Textural features corresponding to textural properties. IEEE Transactions on Systems, Man, and Cybernetics. 1989;19(5):1264-74.
72. Marcadent S, Hofmeister J, Preti MG, Martin SP, Ville DVD, Montet X. Generative Adversarial Networks Improve the Reproducibility and Discriminative Power of Radiomic Features. Radiology: Artificial Intelligence. 2020;2(3):e190035.
73. Welch ML, McIntosh C, Haibe-Kains B, et al. Vulnerabilities of radiomic signature development: The need for safeguards. Radiother Oncol. 2019;130:2-9.
74. Bushberg JT, Seibert JA, Leidholdt Jr. EM, Boone JM, Goldschmidt Jr. EJ. The Essential Physics of Medical Imaging. Medical Physics. 2003;30(7):1936-.
75. Radon J. On the Determination of Functions from Their Integral Values along Certain Manifolds. IEEE Trans Med Imaging. 1986;5(4):170-6.
76. Berenguer R, Pastor-Juan MDR, Canales-Vazquez J, et al. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. Radiology. 2018;288(2):407-15.

77. Zhao B, Tan Y, Tsai WY, Schwartz LH, Lu L. Exploring Variability in CT Characterization of Tumors: A Preliminary Phantom Study. Transl Oncol. 2014;7(1):88-93.
78. Lu L, Liang Y, Schwartz LH, Zhao B. Reliability of Radiomic Features Across Multiple Abdominal CT Image Acquisition Settings: A Pilot Study Using ACR CT Phantom. Tomography. 2019;5(1):226-31.
79. Mackin D, Ger R, Dodge C, et al. Effect of tube current on computed tomography radiomic features. Scientific Reports. 2018;8(1):2354.
80. Mackin D, Fave X, Zhang L, et al. Harmonizing the pixel size in retrospective computed tomography radiomics studies. PLoS One. 2017;12(9):e0178524.
81. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Med Phys. 2017;44(3):1050-62.
82. Roa AM, Andersen HK, Martinsen AC. CT image quality over time: comparison of image quality for six different CT scanners over a six-year period. J Appl Clin Med Phys. 2015;16(2):4972.
83. Nishimura DG. Principles of magnetic resonance imaging: Standford Univ., 2010.
84. Paschal CB, Morris HD. K-space in the clinic. J Magn Reson Imaging. 2004;19(2):145-59.
85. Rai R, Holloway LC, Brink C, et al. Multicenter evaluation of MRI-based radiomic features: A phantom study. Med Phys. 2020.
86. Bianchini L, Santinha J, Loução N, et al. A multicenter study on radiomic features from T2-weighted images of a customized MR pelvic phantom setting the basis for robust radiomic models in clinics. arXiv preprint arXiv:200506833. 2020.
87. Sollini M, Antunovic L, Chiti A, Kirienko M. Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. Eur J Nucl Med Mol Imaging. 2019;46(13):2656-72.
88. Duron L, Balvay D, Vande Perre S, et al. Gray-level discretization impacts reproducible MRI radiomics texture features. PLoS One. 2019;14(3):e0213459.
89. Sun C, Tian X, Liu Z, et al. Radiomic analysis for pretreatment prediction of response to neoadjuvant chemotherapy in locally advanced cervical cancer: A multicentre study. EBioMedicine. 2019;46:160-9.
90. Stanzione A, Ricciardi C, Cuocolo R, et al. MRI Radiomics for the Prediction of Fuhrman Grade in Clear Cell Renal Cell Carcinoma: a Machine Learning Exploratory Study. J Digit Imaging. 2020.
91. Scalco E, Belfatto A, Mastropietro A, et al. T2w-MRI signal normalization affects radiomics features reproducibility. Med Phys. 2020;47(4):1680-91.
92. Moradmand H, Aghamiri SMR, Ghaderi R. Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. J Appl Clin Med Phys. 2020;21(1):179-90.
93. Um H, Tixier F, Bermudez D, Deasy JO, Young RJ, Veeraraghavan H. Impact of image preprocessing on the scanner dependence of multi-parametric MRI radiomic features and covariate shift in multi-institutional glioblastoma datasets. Phys Med Biol. 2019;64(16):165011.
94. Lacroix M, Frouin F, Dirand A-S, et al. Correction for Magnetic Field Inhomogeneities and Normalization of Voxel Values Are Needed to Better Reveal the Potential of MR Radiomic Features in Lung Cancer. Frontiers in Oncology. 2020;10(43).
95. Vallières M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Physics in Medicine & Biology. 2015;60(14):5471.
96. Ashrafinia S. Quantitative Nuclear Medicine Imaging using Advanced Image Reconstruction and Radiomics. Baltimore, Maryland, USA: Johns Hopkins University, 2019.
97. Gotz M, Nolden M, Maier-Hein K. MITK Phenotyping: An open-source toolchain for image-based personalized medicine with radiomics. Radiother Oncol. 2019;131:108-11.
98. Pfaehler E, Zwanenburg A, de Jong JR, Boellaard R. RaCaT: An open source and easy to use radiomics calculator tool. PLoS One. 2019;14(2):e0212223.
99. Zwanenburg A, Leger S, Agolli L, et al. Assessing robustness of radiomic features by image perturbation. Sci Rep. 2019;9(1):614.
100. Vallières M, Kay-Rivest E, Perrin LJ, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Scientific Reports. 2017;7(1):10117.
101. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychological reports. 1966;19(1):3-11.

102. Chatterjee A, Vallières M, Dohan A, et al. Creating Robust Predictive Radiomic Models for Data From Independent Institutions Using Normalization. IEEE Transactions on Radiation and Plasma Medical Sciences. 2019;3(2):210-5.
103. Upadhaya T, Vallières M, Chatterjee A, et al. Comparison of Radiomics Models Built Through Machine Learning in a Multicentric Context With Independent Testing: Identical Data, Similar Algorithms, Different Methodologies. IEEE Transactions on Radiation and Plasma Medical Sciences. 2019;3(2):192-200.
104. Efron B. Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics. 1979;7(1):1-26.
105. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. The Journal of Machine Learning Research. 2017;18(1):559-63.
106. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27(8):861-74.
107. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507-17.
108. Menze BH, Kelm BM, Masuch R, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics. 2009;10(1):213.
109. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.
110. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci. 2003;43(6):1947-58.
111. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.
112. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees: CRC press, 1984.
113. Laplace PS. Traité de mécanique céleste. Paris: Imprimerie de Crapelet. Chez J.B.M. Duprat, Libraire pour les Mathématiques, 1798.
114. Schrödinger E. An Undulatory Theory of the Mechanics of Atoms and Molecules. Physical Review. 1926;28(6):1049-70.
115. Berezin FA, Shubin M. The Schrödinger Equation: Springer Science & Business Media, 2012.
116. Zotter F, Frank M. Ambisonic Amplitude Panning and Decoding in Higher Orders. Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality. Cham: Springer International Publishing, 2019; p. 53-98.
117. Funkhouser T, Min P, Kazhdan M, et al. A search engine for 3D models. ACM Trans Graph. 2003;22(1):83–105.
118. Yoshii N, Nimura Y, Fujimoto K, Okazaki S. Spherical harmonics analysis of surface density fluctuations of spherical ionic SDS and nonionic C12E8 micelles: A molecular dynamics study. The Journal of Chemical Physics. 2017;147(3):034906.
119. Shen L, Farid H, McPeek MA. Modeling three-dimensional morphological structures using spherical harmonics. Evolution. 2009;63(4):1003-16.
120. Galinsky VL, Frank LR. Automated segmentation and shape characterization of volumetric data. Neuroimage. 2014;92:156-68.
121. Kainz B, Keraudren K, Kyriakopoulou V, Rutherford M, Hajnal JV, Rueckert D. Fast fully automatic brain detection in fetal MRI using dense rotation invariant image descriptors. 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)2014; p. 1230-3.
122. Styner M, Oguz I, Xu S, et al. Framework for the Statistical Shape Analysis of Brain Structures using SPHARM-PDM. Insight J. 2006(1071):242-50.
123. Chung MK, Dalton KM, Davidson RJ. Encoding Neuroanatomical Information using Weighted Spherical Harmonic Representation. 2007 IEEE/SP 14th Workshop on Statistical Signal Processing2007; p. 146-50.
124. Keiner J, Kunis S, Potts D. Using NFFT 3---A Software Library for Various Nonequispaced Fast Fourier Transforms. ACM Trans Math Softw. 2009;36(4):Article 19.
125. Banerjee J, Moelker A, Niessen WJ, van Walsum T. 3D LBP-Based Rotationally Invariant Region Description. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013; p. 26-37.
126. Skibbe H, Reisert M, Ronneberger O, Burkhardt H. Increasing the Dimension of Creativity in Rotation Invariant Feature Design Using 3D Tensorial Harmonics. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009; p. 141-50.
127. Skibbe H, Wang Q, Ronneberger O, Burkhardt H, Reisert M. Fast computation of 3D spherical Fourier harmonic descriptors - a complete orthonormal basis for a rotational invariant representation of three-dimensional

objects. 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops2009; p. 1863-9.
128. Müller M. Fourier Analysis of Signals. Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications. Cham: Springer International Publishing, 2015; p. 39-114.
129. Brechbühler C, Gerig G, Kübler O. Parametrization of Closed Surfaces for 3-D Shape Description. Computer Vision and Image Understanding. 1995;61(2):154-70.
130. Weisstein EW. Binomial coefficient. mathworld.wolfram.com/BinomialCoefficient. Accessed on 2020-08-16. Wolfram Mathworld.
131. Li S, Makedon F. Spherical parameterization for 3D surface analysis in volumetric images. International Conference on Information Technology: Coding and Computing, 2004 Proceedings ITCC 20042004; p. 643-9 Vol.1.
132. Dutt A, Rokhlin V. Fast Fourier transforms for nonequispaced data. SIAM Journal on Scientific computing. 1993;14(6):1368-93.
133. Shilane P, Min P, Kazhdan M, Funkhouser T. The Princeton Shape Benchmark. Proceedings Shape Modeling Applications, 20042004; p. 167-78.
134. Zhou H, Li L. Regularized matrix regression. J R Stat Soc Series B Stat Methodol. 2014;76(2):463-83.
135. Amir-Moéz AR, Davis C. Generalized Frobenius inner products. Mathematische Annalen. 1960;141(2):107-12.
136. Li Q, Cai W, Wang X, Zhou Y, Feng DD, Chen M. Medical image classification with convolutional neural network. 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV)2014; p. 844-8.
137. Wang Z, Simoncelli EP, Bovik AC. Multiscale structural similarity for image quality assessment. The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 20032003; p. 1398-402 Vol.2.
138. Zhou W, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing. 2004;13(4):600-12.
139. Zhou H, Li L, Zhu H. Tensor Regression with Applications in Neuroimaging Data Analysis. Journal of the American Statistical Association. 2013;108(502):540-52.
140. Jurafsky D, Martin JH. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition: Prentice Hall PTR, 2000.
141. Tsuruoka Y, Tsujii Ji, Ananiadou S. Stochastic gradient descent training for L1-regularized log-linear models with cumulative penalty. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1. Suntec, Singapore: Association for Computational Linguistics, 2009; p. 477–85.
142. Zhang X, Zhong L, Zhang B, et al. The effects of volume of interest delineation on MRI-based radiomics analysis: evaluation with two disease groups. Cancer Imaging. 2019;19(1):89.
143. Qiu Q, Duan J, Duan Z, et al. Reproducibility and non-redundancy of radiomic features extracted from arterial phase CT scans in hepatocellular carcinoma patients: impact of tumor segmentation variability. Quant Imaging Med Surg. 2019;9(3):453-64.
144. Dice LR. Measures of the amount of ecologic association between species. Ecology. 1945;26(3):297-302.
145. Paulhac L, Makris P, Ramel J-Y. A solid texture database for segmentation and classification experiments. VISAPP (2)2009; p. 135-41.
146. Sun M, Bianchi M, Hansen J, et al. Chronic kidney disease after nephrectomy in patients with small renal masses: a retrospective observational analysis. Eur Urol. 2012;62(4):696-703.
147. Sun M, Trinh QD, Bianchi M, et al. A non-cancer-related survival benefit is associated with partial nephrectomy. Eur Urol. 2012;61(4):725-31.
148. Van Poppel H, Da Pozzo L, Albrecht W, et al. A prospective randomized EORTC intergroup phase 3 study comparing the complications of elective nephron-sparing surgery and radical nephrectomy for low-stage renal cell carcinoma. Eur Urol. 2007;51(6):1606-15.
149. Tan HJ, Norton EC, Ye Z, Hafez KS, Gore JL, Miller DC. Long-term survival following partial vs radical nephrectomy among older patients with early-stage kidney cancer. JAMA. 2012;307(15):1629-35.
150. Go AS, Chertow GM, Fan D, McCulloch CE, Hsu CY. Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization. N Engl J Med. 2004;351(13):1296-305.
151. Bosniak MA. The current radiological approach to renal cysts. Radiology. 1986;158(1):1-10.
152. Ascenti G, Mazziotti S, Zimbaro G, et al. Complex cystic renal masses: characterization with contrast-enhanced US. Radiology. 2007;243(1):158-65.

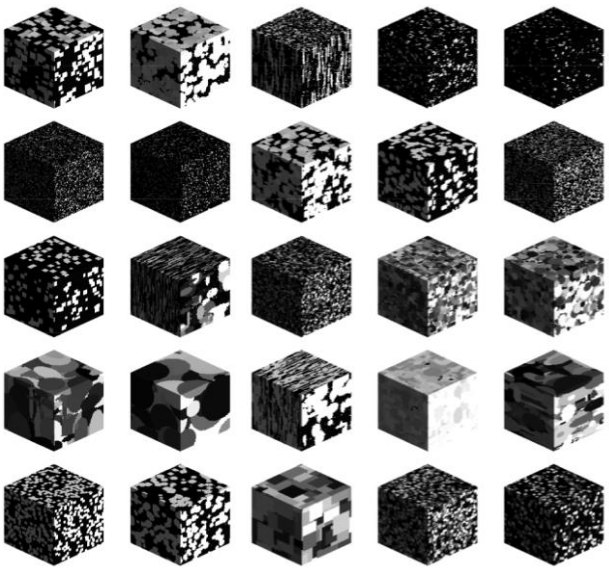
153. Sevcenco S, Spick C, Helbich TH, et al. Malignancy rates and diagnostic performance of the Bosniak classification for the diagnosis of cystic renal lesions in computed tomography - a systematic review and meta-analysis. Eur Radiol. 2017;27(6):2239-47.
154. Wood CG, Stromberg LJ, Harmath CB, et al. CT and MR imaging for evaluation of cystic renal lesions and diseases. Radiographics. 2015;35(1):125-41.
155. Bosniak MA. The use of the Bosniak classification system for renal cysts and cystic tumors. J Urol. 1997;157(5):1852-3.
156. Smith AD, Remer EM, Cox KL, et al. Bosniak category IIF and III cystic renal lesions: outcomes and associations. Radiology. 2012;262(1):152-60.
157. Weibl P, Klatte T, Kollarik B, et al. Interpersonal variability and present diagnostic dilemmas in Bosniak classification system. Scand J Urol Nephrol. 2011;45(4):239-44.
158. Graumann O, Osther SS, Karstoft J, Hørlyck A, Osther PJ. Bosniak classification system: inter-observer and intra-observer agreement among experienced uroradiologists. Acta Radiol. 2015;56(3):374-83.
159. General Electric (GE) Medical Systems GE. LightSpeed™ VCT, Technical Reference Manual, 5340596-1EN, Revision 5. 2011.
160. Lehmann TM, Gonner C, Spitzer K. Addendum: B-spline interpolation in medical image processing. IEEE Trans Med Imaging. 2001;20(7):660-5.
161. Mazzei FG, Mazzei MA, Cioffi Squitieri N, et al. CT perfusion in the characterisation of renal lesions: an added value to multiphasic CT. Biomed Res Int. 2014;2014:135013.
162. Lee S-H, Cho H-h, Lee HY, Park H. Clinical impact of variability on CT radiomics and suggestions for suitable feature selection: a focus on lung cancer. Cancer Imaging. 2019;19(1):54.
163. Karlo CA, Di Paolo PL, Chaim J, et al. Radiogenomics of clear cell renal cell carcinoma: associations between CT imaging features and mutations. Radiology. 2014;270(2):464-71.
164. Alessandrino F, Shinagare AB, Bosse D, Choueiri TK, Krajewski KM. Radiogenomics in renal cell carcinoma. Abdom Radiol (NY). 2019;44(6):1990-8.
165. Parkin D, Whelan S, Ferlay J, Storm H. Cancer Incidence in Five Continents, vol. 1–8, no. 7 of. IARC Cancerbase. 2005.
166. Briet JM, Hollema H, Reesink N, et al. Lymphvascular space involvement: an independent prognostic factor in endometrial cancer. Gynecol Oncol. 2005;96(3):799-804.
167. Huijgens AN, Mertens HJ. Factors predicting recurrent endometrial cancer. Facts Views Vis Obgyn. 2013;5(3):179-86.
168. Uharcek P, Mlyncek M, Ravinger J, Matejka M. Prognostic factors in women 45 years of age or younger with endometrial cancer. Int J Gynecol Cancer. 2008;18(2):324-8.
169. Sorbe B. Predictive and prognostic factors in definition of risk groups in endometrial carcinoma. ISRN Obstet Gynecol. 2012;2012:325790.
170. Amant F, Mirza MR, Koskas M, Creutzberg CL. Cancer of the corpus uteri. Int J Gynaecol Obstet. 2015;131 Suppl 2:S96-104.
171. Braun MM, Overbeek-Wager EA, Grumbo RJ. Diagnosis and Management of Endometrial Cancer. Am Fam Physician. 2016;93(6):468-74.
172. Bogani G, Dowdy SC, Cliby WA, Ghezzi F, Rossetti D, Mariani A. Role of pelvic and para-aortic lymphadenectomy in endometrial cancer: current evidence. J Obstet Gynaecol Res. 2014;40(2):301-11.
173. Lewin SN. Revised FIGO staging system for endometrial cancer. Clin Obstet Gynecol. 2011;54(2):215-8.
174. Endometrial Cancer Treatment (PDQ(R)): Patient Version. PDQ Cancer Information Summaries. Bethesda (MD)2002.
175. Ruz-Caracuel I, Ramon-Patino JL, Lopez-Janeiro A, et al. Myoinvasive Pattern as a Prognostic Marker in Low-Grade, Early-Stage Endometrioid Endometrial Carcinoma. Cancers (Basel). 2019;11(12).
176. Helpman L, Kupets R, Covens A, et al. Assessment of endometrial sampling as a predictor of final surgical pathology in endometrial cancer. Br J Cancer. 2014;110(3):609-15.
177. Leitao MM, Jr., Kehoe S, Barakat RR, et al. Comparison of D&C and office endometrial biopsy accuracy in patients with FIGO grade 1 endometrial adenocarcinoma. Gynecol Oncol. 2009;113(1):105-8.
178. Dimitraki M, Tsikouras P, Bouchlariotou S, et al. Clinical evaluation of women with PMB. Is it always necessary an endometrial biopsy to be performed? A review of the literature. Arch Gynecol Obstet. 2011;283(2):261-6.

179. van der Putten LJ, van de Vijver K, Bartosch C, et al. Reproducibility of measurement of myometrial invasion in endometrial carcinoma. Virchows Arch. 2017;470(1):63-8.
180. Frei KA, Kinkel K, Bonel HM, Lu Y, Zaloudek C, Hricak H. Prediction of deep myometrial invasion in patients with endometrial cancer: clinical utility of contrast-enhanced MR imaging-a meta-analysis and Bayesian analysis. Radiology. 2000;216(2):444-9.
181. Das SK, Niu XK, Wang JL, et al. Usefulness of DWI in preoperative assessment of deep myometrial invasion in patients with endometrial carcinoma: a systematic review and meta-analysis. Cancer Imaging. 2014;14:32.
182. Wu LM, Xu JR, Gu HY, Hua J, Haacke EM, Hu J. Predictive value of T2-weighted imaging and contrast-enhanced MR imaging in assessing myometrial invasion in endometrial cancer: a pooled analysis of prospective studies. Eur Radiol. 2013;23(2):435-49.
183. Andreano A, Rechichi G, Rebora P, Sironi S, Valsecchi MG, Galimberti S. MR diffusion imaging for preoperative staging of myometrial invasion in patients with endometrial cancer: a systematic review and meta-analysis. Eur Radiol. 2014;24(6):1327-38.
184. Sala E, Rockall A, Rangarajan D, Kubik-Huch RA. The role of dynamic contrast-enhanced and diffusion weighted magnetic resonance imaging in the female pelvis. Eur J Radiol. 2010;76(3):367-85.
185. Davnall F, Yip CS, Ljungqvist G, et al. Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? Insights Imaging. 2012;3(6):573-89.
186. Kassner A, Thornhill RE. Texture analysis: a review of neurologic MR imaging applications. AJNR Am J Neuroradiol. 2010;31(5):809-16.
187. O'Connor JP, Rose CJ, Jackson A, et al. DCE-MRI biomarkers of tumour heterogeneity predict CRC liver metastasis shrinkage following bevacizumab and FOLFOX-6. Br J Cancer. 2011;105(1):139-45.
188. Parikh J, Selmi M, Charles-Edwards G, et al. Changes in primary breast cancer heterogeneity may augment midtreatment MR imaging assessment of response to neoadjuvant chemotherapy. Radiology. 2014;272(1):100-12.
189. Bayanati H, R ET, Souza CA, et al. Quantitative CT texture and shape analysis: can it differentiate benign and malignant mediastinal lymph nodes in patients with primary lung cancer? Eur Radiol. 2015;25(2):480-7.
190. Waugh SA, Purdie CA, Jordan LB, et al. Magnetic resonance imaging texture analysis classification of primary breast cancer. Eur Radiol. 2016;26(2):322-30.
191. Hodgdon T, McInnes MD, Schieda N, Flood TA, Lamb L, Thornhill RE. Can Quantitative CT Texture Analysis be Used to Differentiate Fat-poor Renal Angiomyolipoma from Renal Cell Carcinoma on Unenhanced CT Images? Radiology. 2015;276(3):787-96.
192. De Cecco CN, Ganeshan B, Ciolina M, et al. Texture analysis as imaging biomarker of tumoral response to neoadjuvant chemoradiotherapy in rectal cancer patients studied with 3-T magnetic resonance. Invest Radiol. 2015;50(4):239-45.
193. Chaudhry S, Reinhold C, Guermazi A, Khalili I, Maheshwari S. Benign and malignant diseases of the endometrium. Top Magn Reson Imaging. 2003;14(4):339-57.
194. Creasman W. Revised FIGO staging for carcinoma of the endometrium. Int J Gynaecol Obstet. 2009;105(2):109.
195. Deng L, Wang QP, Yan R, et al. The utility of measuring the apparent diffusion coefficient for peritumoral zone in assessing infiltration depth of endometrial cancer. Cancer Imaging. 2018;18(1):23.
196. Guo Y, Wang P, Wang P, et al. Myometrial invasion and overall staging of endometrial carcinoma: assessment using fusion of T2-weighted magnetic resonance imaging and diffusion-weighted magnetic resonance imaging. Onco Targets Ther. 2017;10:5937-43.
197. Meissnitzer M, Forstner R. MRI of endometrium cancer - how we do it. Cancer Imaging. 2016;16:11.
198. Otero-Garcia MM, Mesa-Alvarez A, Nikolic O, et al. Role of MRI in staging and follow-up of endometrial and cervical cancer: pitfalls and mimickers. Insights Imaging. 2019;10(1):19.
199. Takeuchi M, Matsuzaki K, Nishitani H. Diffusion-weighted magnetic resonance imaging of endometrial cancer: differentiation from benign endometrial lesions and preoperative assessment of myometrial invasion. Acta Radiol. 2009;50(8):947-53.
200. Rechichi G, Galimberti S, Signorelli M, et al. Endometrial cancer: correlation of apparent diffusion coefficient with tumor grade, depth of myometrial invasion, and presence of lymph node metastases. AJR Am J Roentgenol. 2011;197(1):256-62.
201. Fotopoulou C, El-Balat A, du Bois A, et al. Systematic pelvic and paraaortic lymphadenectomy in early high-risk or advanced endometrial cancer. Arch Gynecol Obstet. 2015;292(6):1321-7.

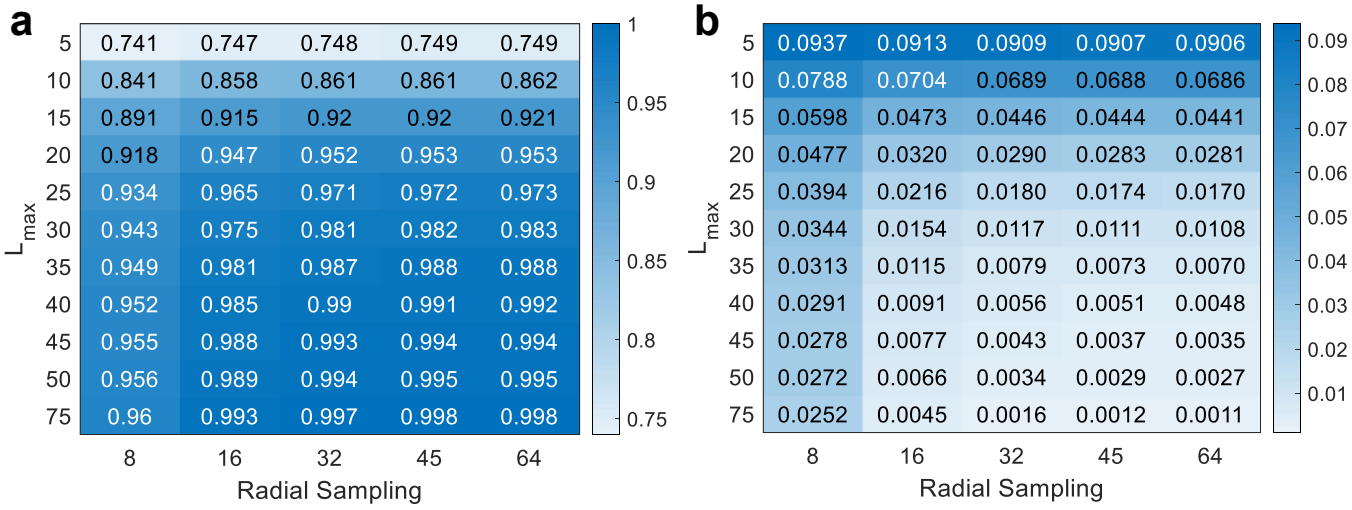
202. Franklin B. Advice to a young tradesman (1748). George Fisher: The American Instructor: or Young Man's Best Companion The Ninth Edition Revised and Corrected Philadelphia: Printed by B Franklin and D Hall, at the New-Printing-Office, in Market-Street. 1748:375-7.
203. Afshar P, Mohammadi A, Plataniotis KN, Oikonomou A, Benali H. From Handcrafted to Deep-Learning-Based Cancer Radiomics: Challenges and Opportunities. IEEE Signal Processing Magazine. 2019;36(4):132-60.
204. Herold CJ, Lewin JS, Wibmer AG, et al. Imaging in the Age of Precision Medicine: Summary of the Proceedings of the 10th Biannual Symposium of the International Society for Strategic Studies in Radiology. Radiology. 2016;279(1):226-38.
205. Park JE, Kim D, Kim HS, et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol. 2020;30(1):523-36.
206. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-73.

APPENDIX

Preliminary Experiments on Volumetric Geometric Synthetic Texture Datasets



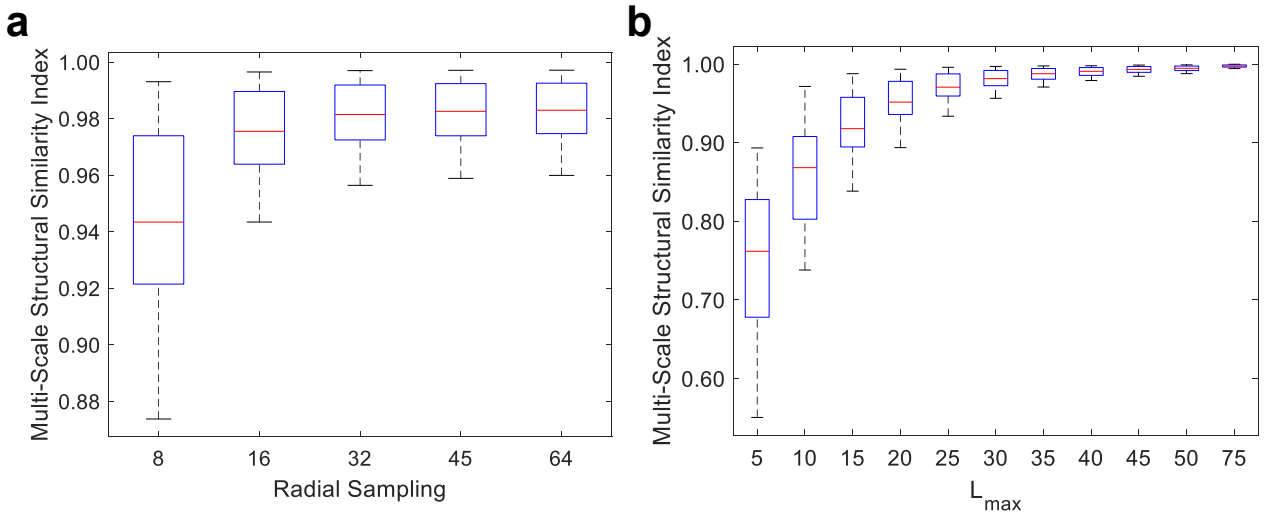
Supplemental Figure 1 Geometric synthetic textures classes from the RFAI dataset from Paulhac L *et al* 2009 (145).



Supplemental Figure 2 (a) Mean and (b) standard deviation of multi-scale structural similarity indexes across different SPHARM expansion and radial sampling in the 3D geometric texture dataset.

Since previous reports mainly assessed the performance of SPHARM descriptors for characterizing 3D shapes (117, 127), we explored and adjusted the SPHARM pipeline on a volumetric synthetic texture benchmark dataset, the RFAI (*Reconnaissance de Formes, Analyse d'Images*) database (145). The second set of textures used consisted in 25 classes of 3D patterns of 64^3 voxels in size with 256 gray-levels built from random insertions of shapes, with 10 examples per class, as shown in **Supplemental Figure 1**. To characterize the ability of SPHARM descriptors to encode these complex textures, the same reconstruction tasks detailed in **Chapter 3** were repeated on decompositions at different radial and frequency levels

For geometric textures, mean MSSI across all classes increased with radial sampling, being at the lowest with 1 sampled shell every 4 voxels (or 8 shells total), and at the highest with 2 sampled shells per voxel (or 64 shells total, which correspond to oversampling). Similarly, MSSI increased with maximal frequency of SPHARM decomposition (L_{max}), being at the lowest with L_{max} = 5 and at the highest with L_{max} = 75 (**Supplemental Figure 2**; **Supplemental Figure 3**). The optimal combination of radial and frequency sampling was found with a radial sampling of 1 shell every voxel (32 shells total) and with frequency decomposition up to L_{max} = 25, similar to what was found on the volumetric Fourier texture dataset in **Chapter 3**. It can be noted that there was no significant gain in MSSI with radial sampling over 32 shells



Supplemental Figure 3 Boxplots of multi-scale structural similarity indexes across (a) radial sampling for fixed L_{max} = 25 and (b) SPHARM expansions for fixed radial sampling of 32 in the 3D geometric texture dataset.

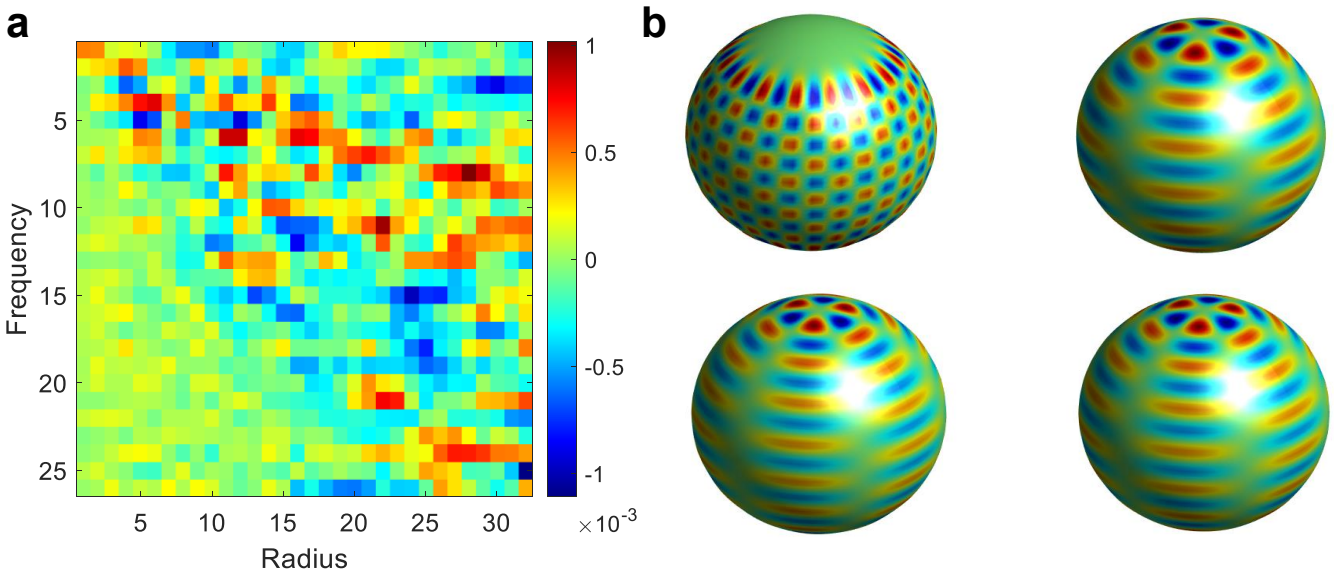
total and over L_{max} = 25. The mean MSSI (± standard deviation) for reconstructed volumetric geometric textures was 0.971 ± 0.018 with 32 sampled shells and with frequency decomposition up to L_{max} = 25. With the same SPHARM decomposition parameters but now reconstructing noisy textures (Gaussian noise) provided in the RFAI database (145), the mean MSSI was 0.942 ± 0.025. Thus, the difference was not significant between reconstructing noisy and normal geometric textures with the optimal set of SPHARM decomposition parameters. Therefore, this expansion was deemed to catch just enough textural details through each SPHARM frequency component to model important textural information, without modeling noise, enabling accurate reconstructions.

For classification, cross-validation with 3 out of 10 examples was used for measuring validation accuracy after fitting models on the bootstrapped samples of the 7 out of 10 examples from each geometric texture class (with a total of 25 classes to classify). The validation classification accuracy was high for SPHARM descriptors using multi-class TensorReg (98% [93%-100%]; **Supplemental Table 1**).

Supplemental Table 1 Accuracy of SPHARM descriptors for classifying the 3D geometric texture dataset.

		Accuracy (%)
SPHARM	Training	96 (90–100)
	Validation	98 (93–100)

Predictive harmonics filters of SPHARM descriptors picked up by TensorReg all modeled SPHARM frequencies across all the expansion. These individual SPHARM descriptors correspond to a relative amount of harmonic filters importance at a given radius (**Supplemental Figure 4**). These preliminary results suggested that SPHARM descriptors could provide accurate encoding of complex volumetric textures.



Supplemental Figure 4 (a) TensorReg 2D classification matrix of regularized regression coefficients on SPHARM descriptors of the volumetric geometric texture dataset. (b) Predictive SPHARM filters on the 3D geometric dataset included orders l, m, r = (24,15,27), (15,3,24), (18,3,25), and (17,3,15), respectively.