

Streamlined Quantitative Imaging Biomarker Development

Generalization of radiomics through automated machine learning



Martijn P. A. Starmans

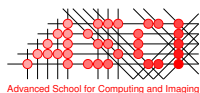
Streamlined Quantitative Imaging Biomarker Development

*Generalization of radiomics through
automated machine learning*

Martijn Pieter Anton Starmans

Acknowledgements:

This work is part of the research programme STRaTeGy with project numbers 14929, 14930, and 14932, which is (partly) financed by the Dutch Research Council (NWO).



This work was carried out in the ASCI graduate school. ASCI dissertation series number 431.

For financial support for the publication of this thesis the following organizations are gratefully acknowledged: NWO, the ASCI graduate school, Quantib BV, and the department of Radiology and Nuclear Medicine of Erasmus MC.

ISBN: 978-94-6416-970-6
Cover: Susan Starre & Martijn Starmans
Layout: Martijn Starmans
Printing: Ridderprint | www.ridderprint.nl

© **Martijn Pieter Anton Starmans, 2022**

Except for the following chapters:

Chapter 2: © Elsevier Inc., 2020

Chapter 5: © Wiley, 2019

Chapter 6: © Elsevier Inc., 2020

Chapter 10: © Society for Endocrinology, 2021

All rights reserved. No part of this thesis may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without written permission from the author or, when appropriate, from the publisher.

Streamlined Quantitative Imaging Biomarker Development

Generalization of radiomics through automated machine learning

Gestroomlijnde ontwikkeling van kwantitatieve
biomarkers op basis van beeldvorming

Generalisatie van radiomics door automatische machine learning

THESIS

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof. dr. A.L. Bredenoord

and in accordance with the decision of the Doctorate Board.

The public defence shall be held on
Tuesday 01 Februari 2022 at 13.00 hrs

by

Martijn Pieter Anton Starmans
born in Velsen, The Netherlands

Erasmus University Rotterdam



Doctoral Committee

Promotors Prof. dr. W.J. Niessen

Other members Prof. dr. M.W. Vernooij
 Prof. dr. ir. A.L.A.J. Dekker
 Dr. K. Lekadir

Co-promotors Dr. ir. S. Klein
 Dr. J.J. Visser

This thesis is dedicated to the memory of my mother

Contents

1	Introduction	3
1.1	Personalized medicine	4
1.2	Radiomics: biomarkers based on quantitative medical imaging features	4
1.3	Research aim	7
1.4	Outline	8
	Part I Adaptive radiomics framework	11
<hr/>		
2	Radiomics: Data mining using quantitative medical image features	13
3	Reproducible radiomics through automated machine learning validated on twelve clinical applications	19
3.1	Introduction	21
3.2	Methods	23
3.3	Experiments	32
3.4	Results	36
3.5	Discussion	36
3.6	Conclusions	43
4	The WORC* database: MRI and CT scans, segmentations, and clinical labels for 930 patients from six radiomics studies	53
4.1	Value of the data	55
4.2	Data description	55
4.3	Experimental design, materials and methods	61
4.4	Ethics statement	63
4.5	Acknowledgments	63
4.6	CRedit author statement	64
4.7	Declaration of competing interest	64
	Part II Radiomics biomarkers in clinical applications	67
<hr/>		

5 Radiomics approach to distinguish between well differentiated liposarcomas and lipomas on MRI	69
5.1 Introduction	71
5.2 Methods	71
5.3 Results	75
5.4 Discussion	81
5.A Radiomics feature extraction	83
5.B Technical details on decision model creation	84
6 Differential diagnosis and mutation stratification of desmoid-type fibromatosis on MRI using radiomics	93
6.1 Introduction	95
6.2 Material and methods	96
6.3 Results	99
6.4 Discussion	105
6.5 Conclusions	109
6.A Radiomics feature extraction	110
6.B Adaptive workflow optimization for automatic decision model creation	111
7 Differential diagnosis and molecular stratification of gastrointestinal stromal tumors on CT images using a radiomics approach	127
7.1 Introduction	129
7.2 Methods	129
7.3 Results	133
7.4 Evaluation of models for the differential diagnosis	133
7.5 Discussion	138
7.6 Conclusion	141
7.A Radiomics feature extraction	142
7.B Adaptive workflow optimization for automatic decision model creation	143
8 A multi-center, multi-vendor study to evaluate the generalizability of a radiomics model for classifying prostate cancer: high grade vs. low grade	153
8.1 Introduction	155
8.2 Material and methods	156
8.3 Results	161
8.4 Discussion	163
8.5 Conclusions	165
8.A Radiomics features extraction	165
8.B Adaptive workflow optimization for automatic decision model creation	167
9 The <i>BRAF</i> P.V600E mutation status of melanoma lung metastases cannot be discriminated on computed tomography by LIDC criteria nor radiomics using machine learning	173
9.1 Introduction	175
9.2 Material and methods	176
9.3 Results	179
9.4 Discussion	182

9.5	Conclusions	185
9.A	Radiomics feature extraction	186
9.B	Model optimization	186
10	Predicting symptomatic mesenteric mass in small intestinal neuroendocrine tumors using radiomics	197
10.1	Introduction	199
10.2	Materials and methods	200
10.3	Results	205
10.4	Discussion	210
10.5	Conclusion	211
10.A	Radiomics feature extraction	212
10.B	Significant features	213
11	Distinguishing pure histopathological growth patterns of colorectal liver metastases on CT using deep learning and radiomics: a pilot study	221
11.1	Introduction	223
11.2	Methods and materials	224
11.3	Segmentation	224
11.4	Results	229
11.5	Discussion	233
11.6	Conclusions	236
11.A	Feature extraction	236
11.B	Model optimization	237
12	Automated differentiation of malignant and benign primary solid liver lesions on MRI: an externally validated radiomics model	247
12.1	Introduction	249
12.2	Materials and methods	250
12.3	Results	254
12.4	Discussion	259
12.A	Pathological examination	263
12.B	Radiomics feature extraction	264
12.C	Radiomics decision model creation	265
	General discussion and summary	275
13	Discussion	277
13.1	Contributions and impact	278
13.2	Roadmap for future research and vision	285
13.3	Conclusion	295
	Summary	299
	Nederlandse samenvatting	305

Acknowledgements	311
About the author	321
Publications	323
PhD portfolio	333
Acronyms	341
Bibliography	349

1.

Introduction

1.1 Personalized medicine

In the last decades, there has been a paradigm shift in healthcare, moving from a reactive, one-size-fits-all approach, towards a more proactive, personalized approach [1, 2, 3]. In personalized medicine, healthcare takes an individual person's unique characteristics into account, with a focus on the individual's outcomes instead of general population statistics, and a focus on prevention instead of solely on treatment. This requires the integration of data from various sources, such as genetic, anatomic, environmental, metabolomic, clinical, laboratory, and imaging data, see Figure 1.1. Therefore, personalized medicine heavily relies on multidisciplinary health teams to integrate all data in order to gain a comprehensive understanding of a person's health status. As the amount of health data has drastically increased, these teams face the increasingly complicated task of combining all the available data to support screening, diagnosis, prognosis, monitoring, treatment planning (e.g. chemotherapy, radiotherapy, immunotherapy), drug usage, surgery, follow-up, and so on.

To aid in this process, personalized medicine generally involves clinical decision support systems, including technologies leveraging big data to relate specific patient characteristics to clinical variables, so-called *biomarkers* [4]. Biomarkers relate to clinical variables such as a biological state, outcome or condition. Especially in cancer medicine, there is a high need for accurate biomarkers, as cancer is a heterogeneous disease with a wide variety of presentations [1, 5]. Hence, personalized medicine has received steep interest in oncology, with medical imaging, such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Positron Emission Tomography (PET) and Ultrasound (US), gaining an increasingly important role [1, 6, 7]. Medical imaging has several advantages over other data acquisition methods, as it is relatively quick, non-invasive (depending on the type of imaging), rich in information, and can be conducted repeatedly at various stages of healthcare.

Currently, in clinical practice, medical imaging is assessed by radiologists, which is generally qualitative and observer dependent. In the oncology domain, various guidelines have been proposed to overcome these issues in specific applications. Examples include RECIST [8] to evaluate treatment response for tumors, LIRADS [9] to assess liver lesions in patients with chronic liver disease, PI-RADS [10] to assess prostate cancer, and the World Health Organization (WHO) guidelines for classification of tumors of the central nervous system [11], the digestive system [12], or soft tissue tumors [13]. However, quantitative, objective biomarkers are required to leverage the full potential of medical imaging. Moreover, as medical imaging has become more accessible, there is a worldwide shortage of (specialized) radiologists, increasing the need for clinical decision support systems that reduce the working load on radiologists [6, 14].

1.2 Radiomics: biomarkers based on quantitative medical imaging features

Quantitative imaging biomarkers describe specific properties of an image in a quantitative way. They can describe the properties of a complete image, or those of a specific region of interest, e.g. a tumor or an organ. Within the field of

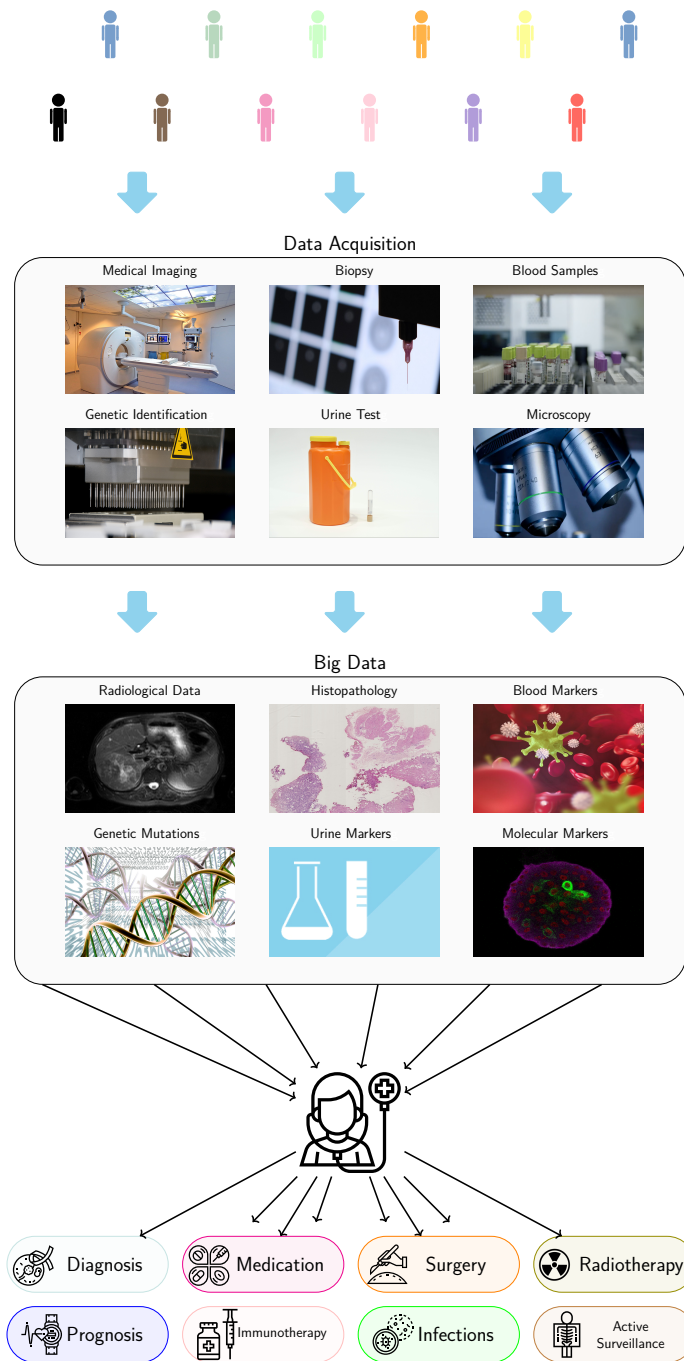


Figure 1.1: Illustration of the challenges of personalized medicine. Each patient has unique characteristics. To uncover these characteristics, a variety of data acquisition methods can be used. This results in a big amount of data being available for each patient. The complicated task of the clinicians is to, for each patient, make decisions on a suitable healthcare plan using the gathered data.

radiology, the term “radiomics” has been coined to describe the use of a large number of quantitative medical imaging features to predict clinical variables [15]. The hypothesis of radiomics is that, since there is a relation between a person’s anatomy, physiology, metabolism, proteins and, genome, there exists a relation between imaging features and these underlying variables (Figure 1.2). Hence, imaging data may be used to create biomarkers to predict these underlying variables. This is especially useful when gathering information on these underlying variables in another way (e.g. chromatography, histopathology, genetic sequencing, based on material from biopsies or resections) is more expensive, time-consuming, invasive, high-risk, or even impossible.

To create radiomics biomarkers, machine learning can be used to discover much more complex features and patterns than humans, and is thus a powerful method to establish relations between imaging features and clinical variables. The use of machine learning in radiomics has led to a rise in popularity, which has resulted in a large number of papers, biomarkers, and radiomics methods being proposed [6, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28].

However, radiomics faces several challenges. In a new clinical application, the main challenge is to find a suitable radiomics method from the wide variety of available options. Most published radiomics methods roughly consist of the same steps: data acquisition and preparation, segmentation, feature extraction, and data mining. The data mining step may itself consist of a combination of various steps: 1) feature imputation; 2) feature scaling; 3) feature selection; 4) dimensionality reduction; 5) resampling; and 6) machine learning algorithms to find relationships between the remaining features and the clinical labels or outcomes. For each of these

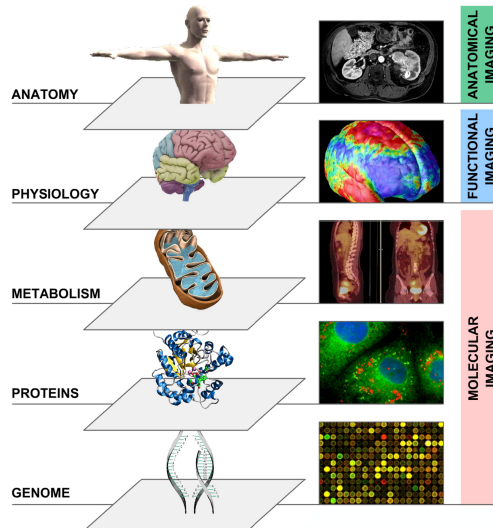


Figure 1.2: Illustration of the hypothesis behind radiomics: since there is a relation between a person’s anatomy, physiology, metabolism, proteins and, genome, there exists a relation between imaging features and these underlying biological variables. Reprinted from [15] with permission from Elsevier (<https://www.sciencedirect.com/science/article/pii/S0959804911009993>).

steps, numerous algorithms have been proposed. Most algorithms have parameters, whose values need to be tuned per application as these influence the performance. As most steps are not independent, and the performance of an algorithm depends on its parameter values, finding the most suitable algorithm and parameter values for each step is not trivial.

Currently, in a new clinical application, finding the optimal radiomics method out of the wide range of available options has to be done manually through a heuristic trial-and-error process. This process has several disadvantages, as it: 1) is time-consuming; 2) requires expert knowledge; 3) does not guarantee that an optimal solution is found; 4) negatively affects the reproducibility; 5) has a high risk of overfitting when not carefully conducted [24, 29]; and 6) limits the translation to clinical practice [20].

Radiomics faces several additional challenges that are vital for the translation to clinical practice.:

1. There is a need for publicly sharing large, multi-center cohorts, to improve the training of radiomics methods, to benchmark radiomics methods, and especially for external validation [1, 17, 20, 24, 25, 27, 30, 31].
2. There is a lack of image acquisition standardization, while radiomics methods are generally sensitive to acquisition variations [16, 18].
3. There is a lack of reproducibility of both radiomics methods and biomarkers [18, 20, 24].

1.3 Research aim

The overall aim of this thesis is to address these challenges, thereby streamlining radiomics research, facilitating the reproducibility of radiomics methods and biomarkers, and ultimately simplifying the use of radiomics in (new) clinical applications. To this end, the following three objectives have been identified.

Our first objective was to propose an adaptive framework to automatically construct and optimize the radiomics method per application. We hypothesized that, instead of manually tuning a radiomics method per application, it should be possible to create one radiomics method that works on multiple applications. Clinically, radiomics applications may be independent and show substantial differences (e.g., prostate cancer versus Alzheimer's disease). Technically, however, the radiomics methods used often show substantial overlap.

Our second objective was to evaluate our adaptive framework on a large number of different clinical applications. In this way, we extensively validated our method and evaluated its generalizability across clinical applications. To maximize the clinical relevance, we focused on oncology applications with a clear need for clinical decision support systems. Moreover, we aimed to facilitate generalization of the resulting biomarkers across image acquisition protocols and thus across clinical centres, increasing the feasibility of applying such a biomarker in routine clinical practice. To this end, we aimed to collect routinely collected, clinically representative, multi-center datasets to train and evaluate our biomarkers.

Our third objective was to make (part of) the collected datasets publicly available and release our code for all methods and experiments open-source. This database would facilitate the reproducibility of our radiomics methods and biomarkers. Additionally, it would enable other researchers to improve the training of radiomics methods and externally validate radiomics biomarkers, and would facilitate public benchmarking.

1.4 Outline

This thesis is divided in two parts. The first part focuses on describing the field of radiomics, our proposed adaptive radiomics method, and our publicly released database. The second part describes in detail the evaluation of our adaptive radiomics method to develop radiomics biomarkers in nine different clinical applications.

Part I [Chapter 2](#) serves as an introduction to radiomics, introduces common terminology, provides an overview of popular approaches, and serves as a guide through the several aspects of designing a radiomics study. Additionally, it describes some of the limitations and future prospects of radiomics.

[Chapter 3](#) describes how recent advances in automated machine learning (AutoML) [32] are exploited to create an adaptive radiomics method. The method is implemented in a Python toolbox, which is coined WORC (Workflow for Optimal Radiomics Classification), and made open-source. We validate our method and evaluate its generalizability in twelve clinical applications.

[Chapter 4](#) describes the publicly released WORC database, consisting of MRI and CT scans, segmentations, and clinical labels for 930 patients from six radiomics studies.

Part II For each of the chapters in this part, an in-depth evaluation of WORC in a different clinical application is provided to answer the following research questions:

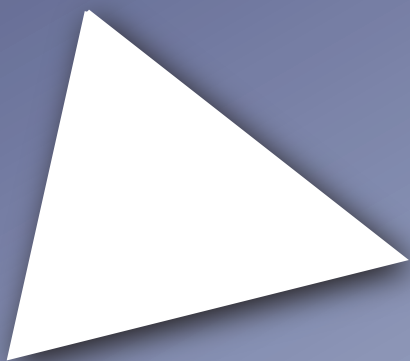
- [Chapter 5](#): can radiomics distinguish between well differentiated liposarcomas and lipomas on MRI?
- [Chapter 6](#): can radiomics distinguish desmoid-type fibromatosis (DTF) from non-DTF tumors in the DTF differential diagnosis on MRI, and predict genetic mutations in DTF?
- [Chapter 7](#): can radiomics distinguish gastrointestinal stromal tumors (GISTs) from non-GIST tumors in the GIST differential diagnosis on CT, and predict genetic mutations in GISTs?
- [Chapter 8](#): can radiomics classify high grade versus low grade prostate cancer on multi-parametric MRI?
- [Chapter 9](#): can radiomics determine the *BRAF* P.V600E mutation status of melanoma lung metastases on CT?

- [Chapter 10](#): can radiomics predict symptomatic mesenteric mass in small intestinal neuroendocrine tumors on CT?
- [Chapter 11](#): can radiomics distinguish pure replacement from pure desmoplastic histopathological growth patterns (HGP) of colorectal liver metastases on CT?
- [Chapter 12](#): can radiomics distinguish malignant from benign primary solid liver lesions on MRI?

Lastly, [Chapter 13](#) discusses the main findings of this thesis, including my methodological, clinical, open science, and education contributions, and provides a roadmap for future research in this field.

Part I

Adaptive radiomics framework



2.

Radiomics: Data mining using quantitative medical image features

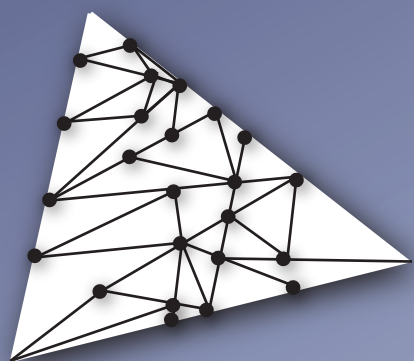
Based on: **M. P. A. Starmans***, S. R. van der Voort*, J. M. Castillo T, J. F. Veenland, S. Klein, and W. J. Niessen, "Radiomics: Data mining using quantitative medical image features," in *Handbook of Medical Image Computing and Computer Assisted Intervention*, S. K. Zhou, D. Rueckert, and G. Fichtinger, Eds. Academic Press, 2020, ch. 18, pp. 429–456. DOI: [10.1016/B978-0-12-816176-0.00023-5](https://doi.org/10.1016/B978-0-12-816176-0.00023-5)

* indicates equal contributions

Abstract

Radiomics uses multiple image features from medical imaging data to predict clinical variables. Various features can be constructed to describe the properties of the full image, or those of a specific region of interest such as a tumor. These features may be related to a wide variety of clinical variables, such as disease characteristics, genetics and therapy response. This can be done through the use of machine learning, which enables the training of a model on these features using data of patients for which the relevant clinical variables are already known. The resulting models may be used as a diagnostic aid for the prediction of labels such as tumor phenotype and therapy response in new patients. Thereby, radiomics can provide a non-invasive alternative for invasive procedures, such as biopsies, to uncover disease characteristics or clinical outcomes. Radiomics therefore has a high potential to be a valuable tool for clinical practice. This may explain the rise in popularity of radiomics in the medical imaging research field in recent years, resulting in many methods and applications. This chapter provides a guide through the several aspects of designing a radiomics study.

Due to copyright restrictions the full text of this chapter cannot be shared publicly.
It is available at <https://doi.org/10.1016/B978-0-12-816176-0.00023-5>.



3.

Reproducible radiomics through automated machine learning validated on twelve clinical applications

Based on: **M. P. A. Starmans**, S. R. van der Voort, T. Phil, M. J. M. Timbergen, M. Vos, G. A. Padmos, W. Kessels, D. Hanff, D. J. Grünhagen, C. Verhoef, S. Sleijfer, M. J. van den Bent, M. Smits, R. S. Dwarkasing, C. J. Els, F. Fiduzi, G. J. L. H. van Leenders, A. Blazevic, J. Hofland, T. Brabander, R. van Gils, G. J. H. Franssen, R. A. Feelders, W. W. de Herder, F. E. Buisman, F. E. J. A. Willemssen, B. Groot Koerkamp, L. Angus, A. A. M. van der Veldt, A. Rajcic, A. E. Odink, M. Deen, J. M. Castillo T, J. F. Veenland, I. Schoots, M. Renckens, M. Doukas, R. A. de Man, J. N. M. Ijzermans, R. L. Miclea, P. B. Vermeulen, E. E. Bron, M. G. Thomeer, J. J. Visser, W. J. Niessen, and S. Klein, “Reproducible radiomics through automated machine learning validated on twelve clinical applications,” *Submitted*, 2021. arXiv: [2108.08618](https://arxiv.org/abs/2108.08618) [eess.IV]

Abstract

Radiomics uses quantitative medical imaging features to predict clinical outcomes. While many radiomics methods have been described in the literature, these are generally designed for a single application. The aim of this study is to generalize radiomics across applications by proposing a framework to automatically construct and optimize the radiomics workflow per application. To this end, we formulate radiomics as a modular workflow, consisting of several components: image and segmentation preprocessing, feature extraction, feature and sample preprocessing, and machine learning. For each component, a collection of common algorithms is included. To optimize the workflow per application, we employ automated machine learning using a random search and ensembling. We evaluate our method in twelve different clinical applications, resulting in the following area under the curves: 1) liposarcoma (0.83); 2) desmoid-type fibromatosis (0.82); 3) primary liver tumors (0.81); 4) gastrointestinal stromal tumors (0.77); 5) colorectal liver metastases (0.68); 6) melanoma metastases (0.51); 7) hepatocellular carcinoma (0.75); 8) mesenteric fibrosis (0.81); 9) prostate cancer (0.72); 10) glioma (0.70); 11) Alzheimer's disease (0.87); and 12) head and neck cancer (0.84). Concluding, our method fully automatically constructs and optimizes the radiomics workflow, thereby streamlining the search for radiomics biomarkers in new applications. To facilitate reproducibility and future research, we publicly release six datasets, the software implementation of our framework (open-source), and the code to reproduce this study.

3.1 Introduction

In the last decades, there has been a paradigm shift in health care, moving from a reactive, one-size-fits-all approach, towards a more proactive, personalized approach [1, 2, 3]. To aid in this process, personalized medicine generally involves clinical decision support systems such as *biomarkers*, which relate specific patient characteristics to some biological state, outcome or condition. To develop such biomarkers, medical imaging has gained an increasingly important role [1, 7]. Currently, in clinical practice, medical imaging is assessed by radiologists, which is generally qualitative and observer dependent. Therefore, there is a need for quantitative, objective biomarkers to leverage the full potential of medical imaging for personalized medicine to improve patient care.

To this end, machine learning, both using conventional and deep learning methods, has shown to be highly successful for medical image classification and has thus become the *de facto* standard. Within the field of radiology, the term “radiomics” has been coined to describe the use of a large number of quantitative medical imaging features in combination with (typically conventional) machine learning to create biomarkers [15]. Predictions for example relate to diagnosis, prognosis, histology, treatment planning (e.g. chemotherapy, radiotherapy, immunotherapy), treatment response, drug usage, surgery, and genetic mutations. The rise in popularity of radiomics has resulted in a large number of papers and a wide variety of methods [6, 16, 17, 18, 20, 21, 22, 23, 24, 26, 27]. In a new radiomics application, finding the optimal method out of the wide range of available options has to be done manually through a heuristic trial-and-error process. This process has several disadvantages, as it: 1) is time-consuming; 2) requires expert knowledge; 3) does not guarantee that an optimal solution is found; 4) negatively affects the reproducibility; 5) has a high risk of overfitting when not carefully conducted [24, 29]; and 6) limits the translation to clinical practice [20].

The aim of this study is to streamline radiomics research, facilitate radiomics’ reproducibility, and simplify its application by proposing a framework to fully automatically construct and optimize the radiomics workflow per application. Most published radiomics methods roughly consist of the same steps: image segmentation, preprocessing, feature extraction, and classification. Hence, as radiomics methods show substantial overlap, we hypothesize that it should be possible to automatically find the optimal radiomics model in a new clinical application by collecting numerous methods in one single framework and systematically comparing and combining all included components.

To optimize the radiomics workflow, we exploit recent advances in automated machine learning (AutoML) [33]. We define a radiomics workflow as a specific combination of algorithms and their associated hyperparameters, i.e., parameters that need to be set before the actual learning step. To create a modular design, we standardize the components of radiomics workflows, i.e., separating the workflows in components with fixed inputs, functionality, and outputs. For each component, we include a large number of algorithms and their associated hyperparameters. We focus on conventional radiomics pipelines, i.e., using conventional machine learning, for the following reasons: 1) radiomics methods are quick to train, hence

AutoML is feasible to apply; 2) the radiomics search space is relatively clear, as radiomics workflows typically follow the same steps, further enhancing the feasibility of AutoML; 3) as there is a large number of radiomics papers, the impact of such a method is potentially large; and 4) radiomics is also suitable for small datasets, which is relevant for (rare) oncological applications [23, 24]. We describe the construction of a radiomics workflow per application as a Combined Algorithm Selection and Hyperparameter (CASH) optimization problem [34], in which we include both the choice of algorithms and their associated hyperparameters. The CASH problem is solved through a brute-force randomized search, identifying the most promising workflows. To boost performance and stability, an ensemble is taken over the most promising workflows to combine them in a single model. Through this use of adaptive workflow optimization, our framework automatically constructs and optimizes the radiomics workflow for each application.

To validate our approach and evaluate its generalizability, we evaluate our framework on twelve different clinical applications using three publicly available datasets and nine in-house datasets. To facilitate reproducibility, six of the in-house datasets with data of in total 930 patients are publicly released with this paper [35] (i.e., Chapter 4 of this thesis). To further facilitate reproducibility, we have made the software implementation of our method, and the code to perform our experiments on all datasets open-source [36, 37].

3.1.1 Background: Radiomics

To outline the context of this study, we here present some background on typical radiomics studies. Generally, a radiomics study can be seen as a collection of various steps: data acquisition and preparation, segmentation, feature extraction, and data mining [19] (i.e., Chapter 2 of this thesis). In this study, we consider the data, i.e., the images, ground truth labels, and segmentations, to be given; data acquisition and segmentation algorithms are therefore outside of the scope of this study.

First, radiomics workflows commonly start with preprocessing of the images and the segmentations to compensate for undesired variations in the data. For example, as radiomics features may be sensitive to image acquisition variations, harmonizing the images may improve the repeatability, reproducibility, and overall performance [18]. Examples of preprocessing steps are normalization of the image intensities to a similar scale, or resampling all images (and segmentations) to the same voxel spacing.

Second, quantitative image features are computationally extracted. As most radiomics applications are in oncology, feature extraction algorithms generally focus on describing properties of a specific region of interest, e.g., a tumor, and require a segmentation. Features are typically split in three groups [38, 39]: 1) first-order or histogram, quantifying intensity distributions; 2) morphology, quantifying shape; and 3) higher-order or texture, quantifying spatial distributions of intensities or specific patterns. Typically, radiomics studies extract hundreds or thousands of features, but eliminate a large part through feature selection in the data mining step. Many open-source toolboxes for radiomics feature extraction exist, such as MaZda [40], CGITA [41], CERR [42], IBEX [43], PyRadiomics [44], CaPTk [45], LIFEx [46],

and RaCat [47]. A comprehensive overview of radiomics toolboxes can be found in Song *et al.* [24].

Lastly, the data mining component may itself consist of a combination of various components: 1) feature imputation; 2) feature scaling; 3) feature selection; 4) dimensionality reduction; 5) resampling; 6) (machine learning) algorithms to find relationships between the remaining features and the clinical labels or outcomes. While these methods are often seen as one component, i.e., the data mining component, we split the data mining step into separate components (Subsection 3.2.2).

3.2 Methods

This study focuses on binary classification problems, as these are most common in radiomics [24].

3.2.1 Adaptive workflow optimization

The aim of our framework is to automatically construct and optimize the radiomics workflow out of a large number of algorithms and their associated hyperparameters. To this end, we have identified three key requirements. First, as the optimal combination of algorithms may vary per application, our optimization strategy should adapt the workflow per application. Second, while model selection is typically performed before hyperparameter tuning, it has been shown that these two problems are not independent [34]. Thus, combined optimization is required. Third, to prevent over-fitting, all optimization should be performed on a training dataset and thereby independent from the test dataset [24, 29, 33]. As manual model selection and hyperparameter tuning is not feasible in a large solution space and not reproducible, all optimization should be automatic.

The Combined Algorithm Selection and Hyperparameter (CASH) optimization problem

To address the three identified key requirements, we propose to formulate the complete radiomics workflow as a Combined Algorithm Selection and Hyperparameter (CASH) optimization problem, which previously has been defined in AutoML for machine learning model optimization [34]. For a single algorithm, the associated hyperparameter space consists of all possible values of all the associated hyperparameters. In machine learning, given a dataset $\mathcal{D} = \{(\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n)\}$ consisting of features \vec{x} and ground truth labels y for n objects or samples, and a set of algorithms $\mathcal{A} = \{A^{(1)}, \dots, A^{(m)}\}$ with associated hyperparameter spaces $\Delta^{(1)}, \dots, \Delta^{(m)}$, the CASH problem is to find the algorithm A^* and associated hyperparameter set λ^* that minimize the loss \mathcal{L} :

$$A^*, \lambda^* \in \underset{A^{(j)} \in \mathcal{A}, \lambda \in \Delta^{(j)}}{\operatorname{argmin}} \frac{1}{k_{\text{training}}} \sum_{i=1}^{k_{\text{training}}} \mathcal{L} \left(A_{\lambda}^{(j)}, \mathcal{D}_{\text{train}}^{(i)}, \mathcal{D}_{\text{valid}}^{(i)} \right), \quad (3.1)$$

where a cross-validation with k_{training} iterations is used to define subsets of the full dataset for training ($\mathcal{D}_{\text{train}}^{(i)}$) and validation ($\mathcal{D}_{\text{valid}}^{(i)}$). In order to combine model selection and hyperparameter optimization, the problem can be reformulated as a pure hyperparameter optimization problem by introducing a new hyperparameter λ_r that selects between algorithms: $\Delta = \Delta^{(1)} \cup \dots \cup \Delta^{(m)} \cup \{\lambda_r\}$ [34]. Thus, λ_r defines which algorithm from \mathcal{A} and which associated hyperparameter space Δ are used. This results in:

$$\lambda^* \in \operatorname{argmin}_{\lambda \in \Delta} \frac{1}{k_{\text{training}}} \sum_{i=1}^{k_{\text{training}}} \mathcal{L} \left(\lambda, \mathcal{D}_{\text{train}}^{(i)}, \mathcal{D}_{\text{valid}}^{(i)} \right). \quad (3.2)$$

We extend the CASH problem to the complete radiomics workflow, consisting of various components. The parameters of all algorithms are treated as hyperparameters. Furthermore, instead of introducing a single hyperparameter to select between algorithms, we define multiple algorithm selection hyperparameters. Two categories are distinguished: 1) for optional components, an *activator* hyperparameter is introduced to determine whether the component is actually used or not; and 2) for mandatory components, an integer *selector* hyperparameter is introduced to select one of the available algorithms. Optional components that contain multiple algorithms have both an *activator* and *selector* hyperparameter. We thus reformulate CASH for a collection of t algorithm sets $\mathcal{A}_C = \mathcal{A}_1 \cup \dots \cup \mathcal{A}_t$ and the collection of associated hyperparameter spaces $\Delta_C = \Delta_1 \cup \dots \cup \Delta_t$. Including the *activator* and *selector* model selection parameters within the hyperparameter collections, similar to Equation 3.2, this results in:

$$\lambda^* \in \operatorname{argmin}_{\lambda_C \in \Delta_C} \frac{1}{k_{\text{training}}} \sum_{i=1}^{k_{\text{training}}} \mathcal{L} \left(\lambda_C, \mathcal{D}_{\text{train}}^{(i)}, \mathcal{D}_{\text{valid}}^{(i)} \right). \quad (3.3)$$

A schematic overview of the algorithm and hyperparameter search space is shown in Figure 3.1. The resulting framework is coined WORC (Workflow for Optimal Radiomics Classification). Including new algorithms and hyperparameters in this reformulation is straight-forward, as these can simply be added to \mathcal{A}_C and Δ_C , respectively.

As a loss function \mathcal{L} , we use the weighted F1-score, which is the harmonic mean of precision and recall, and thus a class-balanced performance metric:

$$F_{1,w} = 2 \sum_{c=1}^{n_{\text{classes}}} \frac{N_c}{N_{\text{total}}} \frac{\text{PREC}_c \times \text{REC}_c}{\text{PREC}_c + \text{REC}_c}, \quad (3.4)$$

where the number of classes $n_{\text{classes}} = 2$ for binary classification, N_c the number of samples of class c , N_{total} the total number of samples, and PREC_c and REC_c the precision and recall of class c , respectively.

As optimization strategy, we use a straightforward random search algorithm, as it is efficient and often performs well [48]. In this random search, N_{RS} workflows are randomly sampled from the search space Δ_C , and their $F_{1,w}$ scores are calculated.

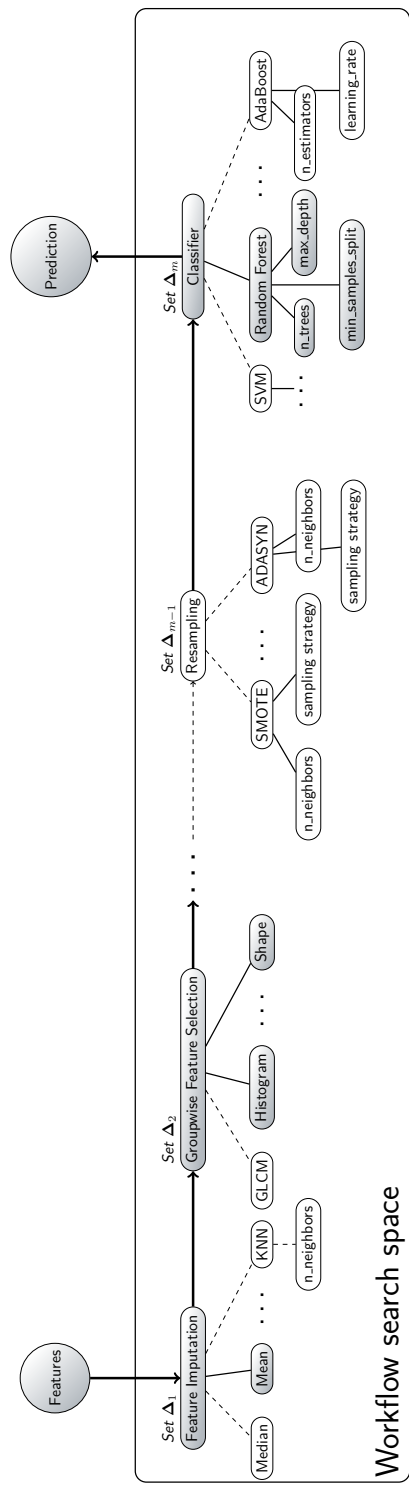


Figure 3.1: Schematic overview of the workflow search space in our framework. The search space consists of various sequential sets of algorithms, where each algorithm may include various hyperparameters, as indicated by the leaves in the trees. An example of a workflow, i.e., a specific combination of algorithms and parameters, is indicated by the gray nodes. Abbreviations: AdaBoost: adaptive boosting; ADASYN: adaptive synthetic sampling; KNN: k-nearest neighbor; GLCM: gray level co-occurrence matrix; SMOTE: synthetic minority oversampling technique; SVM: support vector machine.

Ensembling

In radiomics studies showing the performance of multiple approaches there is often not a clear winner: many workflows generally have similar predictive accuracy. However, despite having similar overall accuracies, the actual prediction for an individual sample may considerably vary per workflow. Moreover, due to the CASH optimization, the best performing solution is likely to overfit. Hence, by combining different workflows in an ensemble, the performance and generalizability of radiomics models may be improved [49].

Furthermore, ensembling may serve as a form of regularization, as local minima in the optimization are balanced by the other solutions in the ensemble. When repeating the optimization, due to the randomness of the search, which single workflow performs best and thus the predictions per sample may vary. This especially occurs when using a small number of random searches. An ensemble may therefore lead to a more stable solution of the random search.

Therefore, instead of selecting the single best workflow, we propose to use an ensemble \mathcal{E} . Various ensembling algorithms have been proposed in literature [50]. Optimizing the ensemble construction on the training dataset may in itself lead to overfitting. Thus, we propose to use a simple approach of combining a fixed number N_{ens} of the best performing workflows by averaging their predictions (i.e., the posterior probabilities for binary classification). The workflows are ranked based on their mean $F_{1,w}$ on the validation datasets.

The WORC optimization algorithm

The optimization algorithm of our WORC framework is depicted in [Algorithm 1](#). All optimization is performed on the training dataset by using a random-split cross-validation with $k_{\text{training}} = 5$, using 80% for training and 20% for validation in a stratified manner, to make sure the distribution of the classes in all sets is similar to the original. A random-split cross-validation is used as this allows a fixed ratio between the training and validation datasets independent of k_{training} , and is consistent with our evaluation setup ([Subsection 3.2.3](#)). The algorithm returns an ensemble \mathcal{E} .

3.2.2 Radiomics components

In order to formulate radiomics as a CASH problem, the workflow needs to be modular and consist of standardized components. In this way, for each component, a set of algorithms and hyperparameters can be defined. We therefore split the radiomics workflow into the following components: image and segmentation preprocessing ([3.2.2](#)), feature extraction ([3.2.2](#)), feature and sample preprocessing ([3.2.2](#)), and machine learning ([3.2.2](#)). For each component, we have included a collection of commonly used algorithms. An overview of the default included components, algorithms, and associated hyperparameters in the WORC framework is provided in [Table 3.1](#).

Table 3.1: Overview of the algorithms and associated hyperparameter search spaces in the random search as used in the WORC framework for binary classification problems. Definitions: $\mathcal{B}(p)$: Bernoulli distribution, equaling value *True* with probability p ; $\mathcal{C}(c)$ a categorical distribution over c categories; $\mathcal{U}(\min, \max)$: uniform distribution; $\mathcal{U}^d(\min, \max)$: uniform distribution with only discrete values; $\mathcal{U}^l(\min, \max)$: uniform distribution on a logarithmic scale. Abbreviations: AdaBoost: adaptive boosting; ADASYN; adaptive synthetic sampling; KNN: k-nearest neighbors; LDA: linear discriminant analysis; LR: logistic regression; PCA: principal component analysis; RBF: radial basis function; QDA: quadratic discriminant analysis; RF: random forest; SMOTE: synthetic minority oversampling technique; SVM: support vector machine; XGBoost: extreme gradient boosting.

Step	Component	Algorithm	Hyperparameter	Distribution
1	Feature Selection	Group-wise selection	Activator	$\mathcal{B}(1.0)$
2	Feature Imputation		Activator per group	$17 \times \mathcal{B}(0.5)$
			Selector	$\mathcal{C}(5)$
		Mean	-	-
		Median	-	-
		Mode	-	-
		Constant (zero)	-	-
		KNN	Nr. Neighbors	$\mathcal{U}^d(5, 10)$
3	Feature Selection	Variance Threshold	Activator	$\mathcal{B}(1.0)$
4	Feature Scaling	Robust z-scoring	-	-
5	Feature Selection	RELIEF	Activator	$\mathcal{B}(0.2)$
			Nr. Neighbors	$\mathcal{U}^d(2, 6)$
			Sample size	$\mathcal{U}(0.75, 0.95)$
			Distance P	$\mathcal{U}^d(1, 4)$
			Nr. Features	$\mathcal{U}^d(10, 50)$
6	Feature Selection	SelectFromModel	Activator	$\mathcal{B}(0.2)$
			Type	$\mathcal{C}(3)$
			LASSO alpha	$\mathcal{U}(0.1, 1.5)$
			RF Nr. Trees	$\mathcal{U}^d(10, 100)$
7	Dimensionality Reduction	PCA	Activator	$\mathcal{B}(0.2)$
			Type	$\mathcal{C}(4)$
8	Feature Selection	Univariate testing	Activator	$\mathcal{B}(0.2)$
			Threshold	$\mathcal{U}^l(10^{-3}, 10^{-2.5})$
9	Resampling		Activator	$\mathcal{B}(0.2)$
			Selector	$\mathcal{U}^d(1, 6)$
		RandomUnderSampling	Strategy	$\mathcal{C}(4)$
		RandomOverSampling	Strategy	$\mathcal{C}(4)$
		NearMiss	Strategy	$\mathcal{C}(4)$
		NeighborhoodCleaningRule	Strategy	$\mathcal{C}(4)$
			Nr. Neighbors	$\mathcal{U}^d(3, 15)$
			Cleaning threshold	$\mathcal{U}(0.25, 75)$
		SMOTE	Type	$\mathcal{C}(4)$
			Strategy	$\mathcal{C}(4)$
			Nr. Neighbors	$\mathcal{U}^d(3, 15)$
		ADASYN	Strategy	$\mathcal{C}(4)$
			Nr. Neighbors	$\mathcal{U}^d(3, 15)$
10	Classification		Selector	$\mathcal{U}^d(1, 8)$
		SVM	Kernel	$\mathcal{C}(3)$
			Regularization	$\mathcal{U}^l(10^0, 10^6)$
			Polynomial degree	$\mathcal{U}^d(1, 7)$
			Homogeneity	$\mathcal{U}(0, 1)$
			RBF γ	$\mathcal{U}^l(10^{-5}, 10^5)$
		RF	Nr. Trees	$\mathcal{U}^d(10, 100)$
			Min. samples / split	$\mathcal{U}^d(2, 5)$
			Max. depth	$\mathcal{U}^d(5, 10)$
		LR	Regularization	$\mathcal{U}(0.01, 1)$
			Solver	$\mathcal{C}(2)$
			Penalty	$\mathcal{C}(3)$
			L_1 -ratio	$\mathcal{U}(0, 1)$
		LDA	Solver	$\mathcal{C}(3)$
			Shrinkage	$\mathcal{U}^l(10^{-5}, 10^5)$
		QDA	Regularization	$\mathcal{U}^l(10^{-5}, 10^5)$
		Gaussian Naive Bayes	Regularization	$\mathcal{U}(0, 1)$
		AdaBoost	Nr. Estimators	$\mathcal{U}^d(10, 100)$
			Learning rate	$\mathcal{U}^l(0.01, 1)$
		XGBoost	Nr. Rounds	$\mathcal{U}^d(10, 100)$
			Max. depth	$\mathcal{U}^d(3, 15)$
			Learning rate	$\mathcal{U}^l(0.01, 1)$
			γ	$\mathcal{U}(0.01, 10)$
			Min. child weight	$\mathcal{U}^d(1, 7)$
			% Random samples	$\mathcal{U}(0.3, 1.0)$

Algorithm 1 The WORC optimization algorithm

```

1: procedure WORC( $\Delta_C, N_{RS}, k_{\text{training}}, N_{\text{ens}}$ )
2:   for  $n \in \{1, \dots, N_{RS}\}$  do
3:      $\lambda_n \leftarrow \text{Random}(\Delta_C)$ 
4:      $\mathcal{L}_n = \frac{1}{k_{\text{training}}} \sum_{i=1}^{k_{\text{training}}} \mathcal{L}(\lambda_n, \mathcal{D}_{\text{train}}^{(i)}, \mathcal{D}_{\text{valid}}^{(i)})$ 
5:   end for
6:    $\Delta_{\text{ranked}} \leftarrow \text{Rank}(\{\lambda_1, \dots, \lambda_{N_{RS}}\} \propto \{\mathcal{L}_1, \dots, \mathcal{L}_{N_{RS}}\})$ 
7:    $\Delta_{\text{ens}} \leftarrow \Delta_{\text{ranked}}[1 : N_{\text{ens}}]$ 
8:   Retrain  $\Delta_{\text{ens}}$  on full training set
9:   Combine  $\Delta_{\text{ens}}$  into ensemble  $\mathcal{E}$ 
10:  return  $\mathcal{E}$ 
11: end procedure

```

Image and segmentation preprocessing

Before feature extraction, image preprocessing such as image quantization, normalization, resampling or noise filtering may be applied [16, 38, 44]. By default no preprocessing is applied. The only exception is image normalization (using z-scoring), which we apply in modalities that do not have a fixed unit and scale (e.g. qualitative MRI, ultrasound), but not in modalities that have a fixed unit and scale (e.g. Computed Tomography (CT), quantitative MRI such as T1 mapping).

Feature extraction

For each segmentation, 564 radiomics features quantifying intensity, shape, orientation and texture are extracted through the open-source feature toolboxes PyRadiomics [44] and PREDICT [51]. A comprehensive overview is provided in Table 3.A.1. Thirteen intensity features describe various first-order statistics of the raw intensity distributions within the segmentation, such as the mean, standard deviation, and kurtosis. Thirty-five shape features describe the morphological properties of the segmentation, and are extracted based only on the segmentation, i.e., not using the image. These include shape descriptions such as the volume, compactness, and circular variance. Nine orientation features describe the orientation and positioning of the segmentation, i.e., not using the image. These include the major axis orientations of a 3D ellipse fitted to the segmentation, the center of mass coordinates and indices. Lastly, 507 texture features are extracted, which include commonly used algorithms such as the Gray Level Co-occurrence Matrix (GLCM) (144 features) [39], Gray Level Size Zone Matrix (GLSZM) (16 features) [39], Gray Level Run Length Matrix (GLRLM) (16 features) [39], Gray Level Dependence Matrix (GLDM) (14 features) [39], Neighborhood Grey Tone Difference Matrix (NGTDM) (5 features) [39], Gabor filters (156 features) [39], Laplacian of Gaussian (LoG) filters (39 features) [39], and Local Binary Patterns (LBP) (39 features) [52]. Additionally, two less common feature groups are defined: based on local phase [53] (39 features) and vesselness filters [54] (39 features).

Many radiomics studies include datasets with variations in the slice thickness due to heterogeneity in the acquisition protocols. This may cause feature values to be dependent on the acquisition protocol. Moreover, the slice thickness is often substantially larger than the pixel spacing. Hence, extracting robust 3D features may be hampered by these variations, especially for low resolutions. To overcome this issue, a 2.5D approach is used: all features except the histogram features are extracted per 2D axial slice and aggregated over all slices. Afterwards, several first-order statistics over the feature distributions are evaluated and used as actual features, see also [Table 3.A.1](#).

In addition to these features, depending on the application, clinical characteristics, e.g. age and sex, and manually scored features, e.g. based on visual inspection by a radiologist, can be added.

Some of the features have parameters themselves, such as the scale on which a derivative is taken. As some features are rather computationally expensive to extract, we do not include these parameters directly as hyperparameters in the CASH problem. Instead, the features are extracted for a predefined range of parameter values. In the next components, feature selection algorithms are employed to select the most relevant features and thus parameters. The used parameter ranges are reported in [Table 3.A.1](#).

Radiomics studies may involve multiple scans per sample, e.g. in multimodal (MRI + CT) or multi-contrast (T1-weighted MRI + T2-weighted MRI) studies. Commonly, radiomics features are defined on a single image, which also holds for the features described in this study. Hence, when multiple scans per sample are included, the 564 radiomics features are extracted per scan and concatenated.

Feature and sample preprocessing

We define feature and sample preprocessing as all algorithms that can be used between the feature extraction and machine learning components. The order of these algorithms in the WORC framework is fixed and given in [Table 3.1](#).

Feature imputation is employed to replace missing feature values. Values may be missing when a feature could not be defined and computed, e.g. a lesion may be too small for a specific feature to be extracted. Algorithms for imputation include: 1) mean; 2) median; 3) mode; 4) constant value (default: zero); and 5) nearest neighbor approach.

Feature scaling is employed to ensure that all features have a similar scale. As this generally benefits machine learning algorithms, this is always performed through z-scoring. A robust version is used, where outliers, defined as feature values outside the 5th – 95th percentile range are excluded before computation of the mean and standard deviation.

Feature selection or dimensionality reduction algorithms may be employed to select the most relevant features and eliminate irrelevant or redundant features. As multiple algorithms may be combined, instead of defining feature selection or dimensionality reduction as a single step, each algorithm is included as a single step in the workflow with an *activator* hyperparameter to determine whether the algorithm is used or not.

Algorithms included are:

1. A group-wise feature selection, in which groups of features (i.e., intensity, shape, and texture feature subgroups) can be selected or eliminated. To this end, each feature group has an *activator* hyperparameter. This algorithm serves as regularization, as it randomly reduces the feature set, and is therefore always used. The group-wise feature selection is the first step in the workflows, as it reduces the computation time of the other steps by reducing the feature space.
2. A variance threshold, in which features with a low variance (< 0.01) are removed. This algorithm is always used, as this serves as a feature sanity check with almost zero risk of removing relevant features. The variance threshold is applied before the feature scaling, as this results in all features having unit variance.
3. Optionally, the RELIEF algorithm [55], which ranks the features according to the differences between neighboring samples. Features with more differences between neighbors of different classes are considered higher in rank.
4. Optionally, feature selection using a machine learning model [56]. Features are selected based on their importance as given by a machine learning model trained on the dataset. Hence, the used algorithm should be able to give the features an importance weight. Algorithms included are LASSO, logistic regression, and random forest.
5. Optionally, principal component analysis (PCA), in which either only those linear combinations of features are kept which explained 95% of the variance in the features, or a fixed number of components (10, 50, or 100) is selected.
6. Optionally, individual feature selection through univariate testing. To this end, for each feature, a Mann-Whitney U test is performed to test for significant differences in distribution between the classes. Afterwards, only features with p-values below a certain threshold are selected. The (non-parametric) Mann-Whitney U test was chosen as it makes no assumptions about the distribution of the features.

RELIEF, selection using a model, PCA, and univariate testing have a 27.5% chance to be included in a workflow in the random search, as this gives an equal chance of applying any of these or no feature selection algorithm. The feature selection algorithms may only be combined in the mentioned order in the WORC framework.

Resampling algorithms may be used, primarily to deal with class imbalances. These include various algorithms from the `imbalanced-learn` toolbox [57]: 1) random under-sampling; 2) random over-sampling; 3) near-miss resampling; 4) the neighborhood cleaning rule; 5) SMOTE [58] (regular, borderline, Tomek, and the edited nearest neighbors variant); and 6) ADASYN [59]. All algorithms can apply four out of five different resampling strategies, resampling: 1) the minority class (not for undersampling algorithms); 2) all but the minority class; 3) the majority class (not for oversampling algorithms); 4) all but the majority class; and 5) all classes.

Machine learning

For machine learning, we mostly use methods from the `scikit-learn` toolbox [60]. The following classification algorithms are included: 1) logistic regression; 2) support vector machines (with a linear, polynomial, or radial basis function kernel); 3) random forests; 4) naive Bayes; 5) linear discriminant analysis; 6) quadratic discriminant analysis (QDA); 7) AdaBoost [61]; and 8) extreme gradient boosting (XGBoost) [62]. The associated hyperparameters for each algorithm are depicted in Table 3.1.

3.2.3 Statistics

Evaluation using a single dataset is performed through a random-split cross-validation with $k_{\text{test}} = 100$, see Figure 3.A.1(a) for a schematic overview. A random-split cross-validation was chosen, as it has a relatively low computational complexity while facilitating estimation of the generalization error [63, 64]. In each iteration, the data is randomly split in 80% for training and 20% for testing in a stratified manner. In each random-split iteration, all CASH optimization is performed within the training set according to Algorithm 1 to eliminate any risk of overfitting on the test set. When a fixed, independent training and test set are used, only the second, internal random-split cross-validation with $k_{\text{training}} = 5$ on the training set for the CASH optimization is used, see Figure 3.A.1(b).

Performance metrics used for evaluation of the test set include the Area Under the Curve (AUC), calculated using the Receiver Operating Characteristic (ROC) curve, $F_{1,w}$, sensitivity, specificity, precision, recall, accuracy, and Balanced Classification Rate (BCR) [65]. When a single dataset is used, and thus a $k_{\text{test}} = 100$ random-split cross-validation, 95% confidence intervals of the performance metrics are constructed using the corrected resampled t-test, thereby taking into account that the samples in the cross-validation splits are not statistically independent [64]. When a fixed training and test set are used, 95% confidence intervals are constructed using 1000x bootstrap resampling of the test dataset and the standard method for normal distributions ([66], table 6, method 1). ROC confidence bands are constructed using fixed-width bands [67].

3.2.4 Software implementation

The WORC toolbox is implemented in Python3 and available open-source [36] under the Apache License, Version 2.0. The WORC toolbox supports Unix and Windows operating systems. Documentation on the WORC toolbox can be found online [68], and several tutorials are available¹. Basic usage only requires the user to specify the locations of the used data (i.e., images, segmentations, ground truth). A minimal working example of the WORC toolbox interface in Python3 is shown in Algorithm 3.A.1.

The WORC toolbox makes use of the `fastr` package [69], an automated workflow engine. `fastr` does not provide any actual implementation of the required

¹<https://github.com/MStarmans91/WORCTutorial>

(radiomics) algorithms, but serves as a computational workflow engine, which has several advantages. Firstly, *fastr* requires workflows to be modular and split into standardized components or *tools*, with standardized inputs and outputs. This nicely connects to the modular design of WORC, for which we therefore wrapped each component as a tool in *fastr*. Alternating between feature extraction toolboxes can be easily done by changing a single field in the WORC toolbox configuration. Second, provenance is automatically tracked by *fastr* to facilitate repeatability and reproducibility. Third, *fastr* offers support for multiple execution plugins in order to be able to execute the same workflow on different computational resources or clusters. Examples include linear execution, local threading on multiple CPUs, and SLURM [70]. Fourth, *fastr* is agnostic to software language. Hence, instead of restricting the user to a single programming language, algorithms (e.g. feature toolboxes) can be supplied in a variety of languages such as Python, Matlab, R and command line executables. Fifth, *fastr* provides a variety of import and export plugins for loading and saving data. Besides using the local file storage, these include use of XNAT [71].

The computation time of a WORC experiment roughly scales with k_{training} , k_{test} , and N_{RS} . A high degree of parallelization for all these parameters is possible, as all workflows can be executed independent of each other. We choose to run the iterations of k_{test} sequential instead of in parallel to maintain a sustainable computational load. For the k_{training} iterations and N_{RS} samples, all workflows are run in parallel. The default experiments in this study consist of executing 500000 workflows ($k_{\text{training}} = 5$, $k_{\text{test}} = 100$, and $N_{RS} = 1000$). On average, experiments in our study had a computation time of approximately 18 hours on a machine with 24 Intel E5-2695 v2 CPU cores, hence roughly 10 minutes per train-test cross-validation iteration. The contribution of the feature extraction to the computation time is negligible.

3.3 Experiments

3.3.1 Evaluation of default configuration on twelve different clinical applications

In order to evaluate our WORC framework, experiments were performed on twelve different clinical applications: see Table 3.2 for an overview of the twelve datasets, and Figure 3.2 for example images from each dataset. All datasets are multi-center with heterogeneity in the image acquisition protocols. For each experiment, per patient, one or more scan(s) and segmentation(s), and a ground truth label are provided. All scans were made at “baseline”, i.e., before any form of treatment or surgery. One dataset (the Glioma dataset) consists of a fixed, independent training and test set and is thus evaluated using 1000x bootstrap resampling. In the other eleven datasets, the performance is evaluated using the $k_{\text{test}} = 100$ random-split cross-validation.

The first six datasets (Lipo, Desmoid, Liver, GIST, CRLM, and Melanoma) are publicly released as part of this study, see [35] (i.e., Chapter 4 of this thesis) for more details. Three datasets (HCC, MesFib, and Prostate) cannot be made publicly

Table 3.2: Overview of the twelve datasets used in this study to evaluate our WORC framework. Abbreviations: ADC: Apparent Diffusion Coefficient; CT: Computed Tomography; DWI: Diffusion Weighted Imaging; MRI: Magnetic Resonance Imaging; T1w: T1 weighted; T2w: T2 weighted.

#	Dataset	Patients	Modality	Segmentation	Description
1.	Lipo ^O	115	T1w MRI	Tumor	Distinguishing well-differentiated liposarcoma from lipoma in 116 lesions from 115 patients [72] (i.e., Chapter 5 of this thesis).
2.	Desmoid ^O	203	T1w MRI	Tumor	Differentiating desmoid-type fibromatosis from soft-tissue sarcoma [73] (i.e., Chapter 6 of this thesis).
3.	Liver ^O	186	T2w MRI	Tumor	Distinguishing malignant from benign primary solid liver lesions [74] (i.e., Chapter 12 of this thesis).
4.	GIST ^O	246	CT	Tumor	Differentiating gastrointestinal stromal tumors (GIST) from other intra-abdominal tumors in 247 lesions from 246 patients [75] (i.e., Chapter 7 of this thesis).
5.	CRLM ^O	77	CT	Tumor	Distinguishing replacement from desmoplastic histopathological growth patterns in colorectal liver metastases (CRLM) in 93 lesions from 77 patients [76] (i.e., Chapter 11 of this thesis).
6.	Melanoma ^O	103	CT	Tumor	Predicting the <i>BRAF</i> mutation status in melanoma lung metastases in 169 lesions from 103 patients [77] (i.e., Chapter 9 of this thesis).
7.	HCC	154	T2w MRI	Liver	Distinguishing livers in which no hepatocellular carcinoma (HCC) developed from livers with HCC at first detection during screening [78].
8.	MesFib	68	CT	Surrounding mesentery	Identifying patients with mesenteric fibrosis at risk of developing intestinal complications [79] (i.e., Chapter 10 of this thesis).
9.	Prostate	40	T2w MRI, DWI, ADC	Lesion	Classifying suspected prostate cancer lesions in high-grade (Gleason > 6) versus low-grade (Gleason ≤ 6) in 72 lesions from 40 patients [80].
10.	Glioma	413	T1w & T2w MRI	Tumor	Predicting the 1p/19q co-deletion in patients with presumed low-grade glioma with a training set of 284 patients and an external validation set of 129 patients [81].
11.	Alzheimer	848	T1w MRI	Hippocampus	Distinguishing patients with Alzheimer's disease from cognitively normal individuals in 848 subjects based on baseline T1w MRIs [82].
12.	H&N	137	CT	Gross tumor volume	Predicting the T-stage (high (≥ 3) or low (< 3)) in patients with head-and-neck cancer [83].

^ODataset publicly released as part of this study [35] (i.e., Chapter 4 of this thesis).

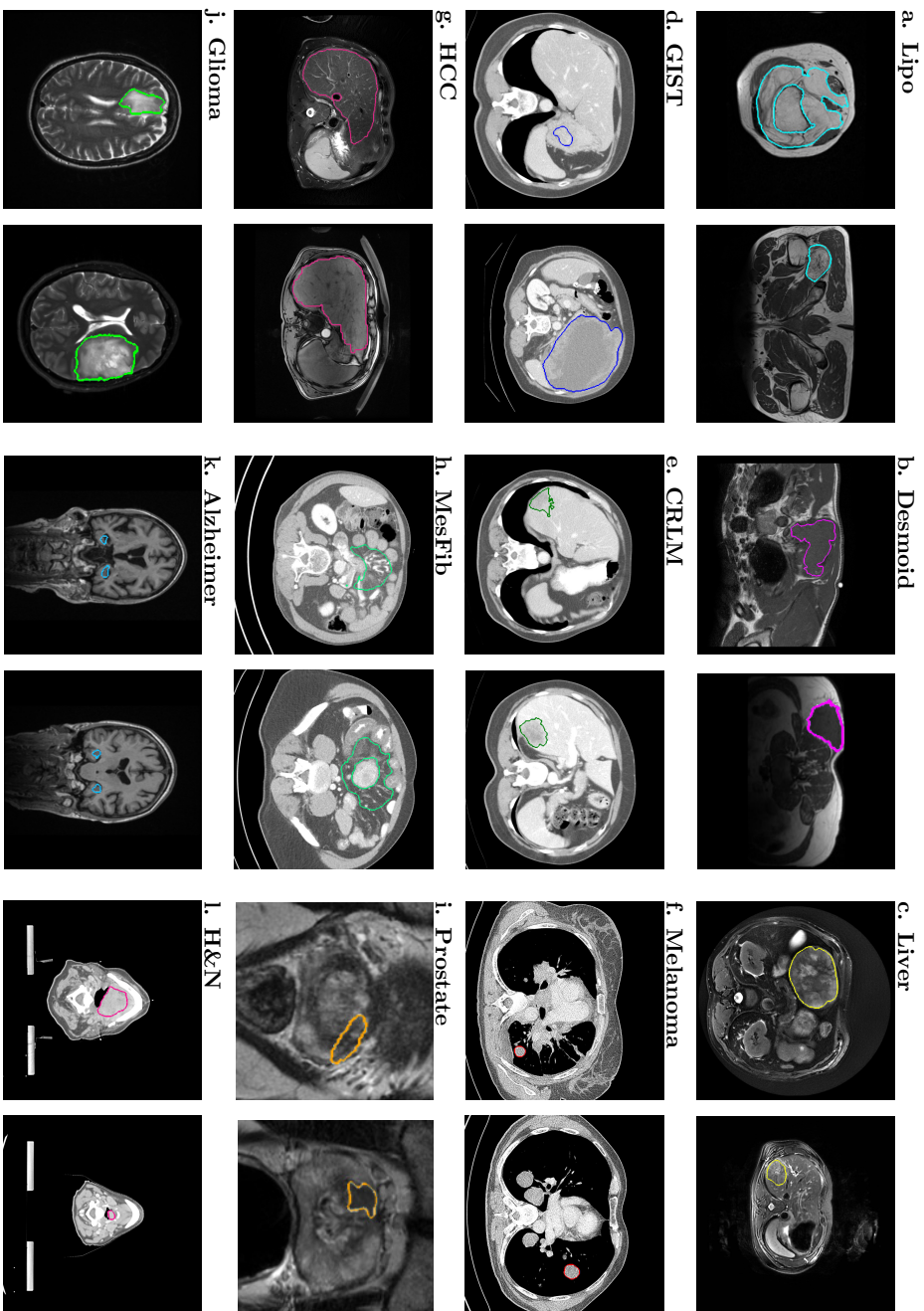


Figure 3.2: Examples of the 2D slices from the 3D imaging data from the twelve datasets used in this study to evaluate our WORC framework. For each dataset, for one patient of each of the two classes, the 2D slice in the primary scan direction (e-g., axial) with the largest area of the segmentation is depicted; the boundary of the segmentation is projected in color on the image. The datasets included were from different clinical applications: a. lipomatous tumors [72] (i.e., Chapter 5 of this thesis); b. desmoid-type fibromatosis [73] (i.e., Chapter 6 of this thesis); c. primary solid liver tumors [74] (i.e., Chapter 12 of this thesis); d. gastrointestinal stromal tumors [75] (i.e., Chapter 7 of this thesis); e. colorectal liver metastases [76] (i.e., Chapter 11 of this thesis); f. melanoma [77] (i.e., Chapter 9 of this thesis); g. hepatocellular carcinoma [78]; h. mesenteric fibrosis [79] (i.e., Chapter 10 of this thesis); i. prostate cancer [80]; j. low grade glioma [81]; k. Alzheimer's disease [82]; and l. head and neck cancer [83].

available. The final three datasets (Glioma, Alzheimer, and H&N) are already publicly available, and were described in previous studies [81, 83, 84].

For the Glioma dataset, the raw imaging data was not available. Instead, pre-computed radiomics features are available [85], which were directly fed into WORC.

The Alzheimer dataset was obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI was to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD). For up-to-date information, see www.adni-info.org. This dataset will be referred to as the “Alzheimer” dataset. Here, radiomics was used to distinguish patients with AD from cognitively normal (CN) individuals. The cohort as described by Bron *et al.* [82] was used, which includes 334 patients with AD and 520 CN individuals with approximately the same mean age in both groups (AD: 74.9 years, CN: 74.2 years). The hippocampus was chosen as region of interest for the radiomics analysis, as this region is known to suffer from atrophy early in the disease process of AD. Automatic hippocampus segmentations were obtained for each patient using the algorithm described by Bron *et al.* [86].

The H&N dataset [83] was obtained from a public database² and directly fed into WORC. For each lesion, the first gross tumor volume (GTV-1) segmentation was used as region of interest for the radiomics analysis. Patients without a CT scan or a GTV-1 segmentation were excluded.

3.3.2 Influence of the number of random search iterations and ensemble size

An additional experiment was conducted to investigate the influence of the number of random search iterations N_{RS} and ensemble size N_{ens} on the performance. For reproducibility, this experiment was performed using the six datasets publicly released in this study (Lipo, Desmoid, Liver, GIST, CRLM, and Melanoma). We hypothesize that increasing N_{ens} at first will improve the performance and stability, and after some point, when the ratio N_{ens}/N_{RS} becomes too high, will reduce the performance and stability as bad solutions are added to the ensemble.

We varied the number of random search iterations ($N_{RS} \in [10; 50; 100; 1000; 10000; 25000]$) and the ensemble size ($N_{ens} \in [1(\text{i.e., no ensembling}); 10; 50; 100]$) and repeated each experiment ten times with different seeds for the random number generator. To limit the computational burden, $k_{test} = 20$ was used instead of the default $k_{test} = 100$, and the $N_{RS} = 25000$ experiment was only performed once instead of ten times. For each configuration, both the average performance and the stability were assessed in terms of the mean and standard deviation of $F_{1,w}$. Based on these experiments, the default number of random search iterations and ensemble size for the WORC optimization algorithm were determined and used in all other experiments.

²<https://xnat.bmia.nl/data/projects/stwstrategyhn1>

3.4 Results

3.4.1 Application of the WORC framework to twelve datasets

Error plots of the AUCs from the application of our WORC framework with the same default configuration on the twelve different datasets are shown in [Figure 3.3](#); detailed performances, including other metrics, are shown in [Table 3.3](#); the ROC curves are shown in [Figure 3.A.2](#). In eleven of the twelve datasets, we successfully found a prediction model, with mean AUCs of 0.83 (Lipo), 0.82 (Desmoid), 0.81 (Liver), 0.77 (GIST), 0.68 (CRLM), 0.75 (HCC), 0.81 (MesFib), 0.72 (Prostate), 0.70 (Glioma), 0.87 (Alzheimer), and 0.84 (H&N). In the Melanoma dataset, the mean AUC (0.51) was similar to that of guessing (0.50).

3.4.2 Influence of the number of random search iterations and ensemble size

The performance of varying the number of random search iterations N_{RS} and ensemble size N_{ens} in the first six datasets is reported in [Table 3.4](#).

For five out of six datasets in this experiment (Lipo, Desmoid, Liver, GIST, and CRLM), the mean performance generally improved when increasing both N_{RS} and N_{ens} . The sixth dataset (Melanoma) is an exception, as the performances for varying N_{RS} and N_{ens} was similar. This can be attributed to the fact that it is the only dataset in this study where we could not successfully construct a productive model.

In the first five datasets, the mean $F_{1,w}$ for the lowest values, $N_{RS} = 1$ (i.e., only trying one random workflow) and $N_{ens} = 1$ (i.e., no ensembling), was 0.75 (Lipo), 0.61 (Desmoid), 0.66 (Liver), 0.67 (GIST), and 0.54 (CRLM). The mean performance for the highest values, $N_{RS} = 25000$ and $N_{ens} = 100$, was substantially higher for all five datasets (Lipo: 0.84; Desmoid: 0.72; Liver: 0.80; GIST: 0.76; and CRLM: 0.63). The mean $F_{1,w}$ of $N_{RS} = 1000$ was very similar to that of $N_{RS} = 25000$, while $N_{RS} = 25000$ took 25 times longer to execute than $N_{RS} = 1000$. This indicates that at some point, here $N_{RS} = 1000$, increasing the computation time by trying out more workflows does not result in an increase in performance on the test set anymore.

At $N_{RS} = 10$ and $N_{ens} = 1$, the standard deviation of the $F_{1,w}$ (Lipo: 0.026; Desmoid: 0.023; Liver: 0.022; GIST: 0.038; and CRLM: 0.027) was substantially higher than at $N_{RS} = 10000$, $N_{ens} = 100$ (Lipo: 0.001; Desmoid: 0.004; Liver: 0.002; GIST: 0.002; and CRLM: 0.005). This indicates that increasing N_{RS} and N_{ens} improves the stability of the model. The standard deviations of $N_{RS} = 10000$ were similar to $N_{RS} = 1000$, illustrating that, similar to the mean performance, the stability at some point converges. For each N_{RS} value, the standard deviation at first decreased when increasing N_{ens} , but increased when N_{ens} became similar or equal to N_{RS} .

3.5 Discussion

In this study, we proposed a framework to automatically construct and optimize radiomics workflows to generalize radiomics across applications. To evaluate the performance and generalization, we applied our framework to twelve different,

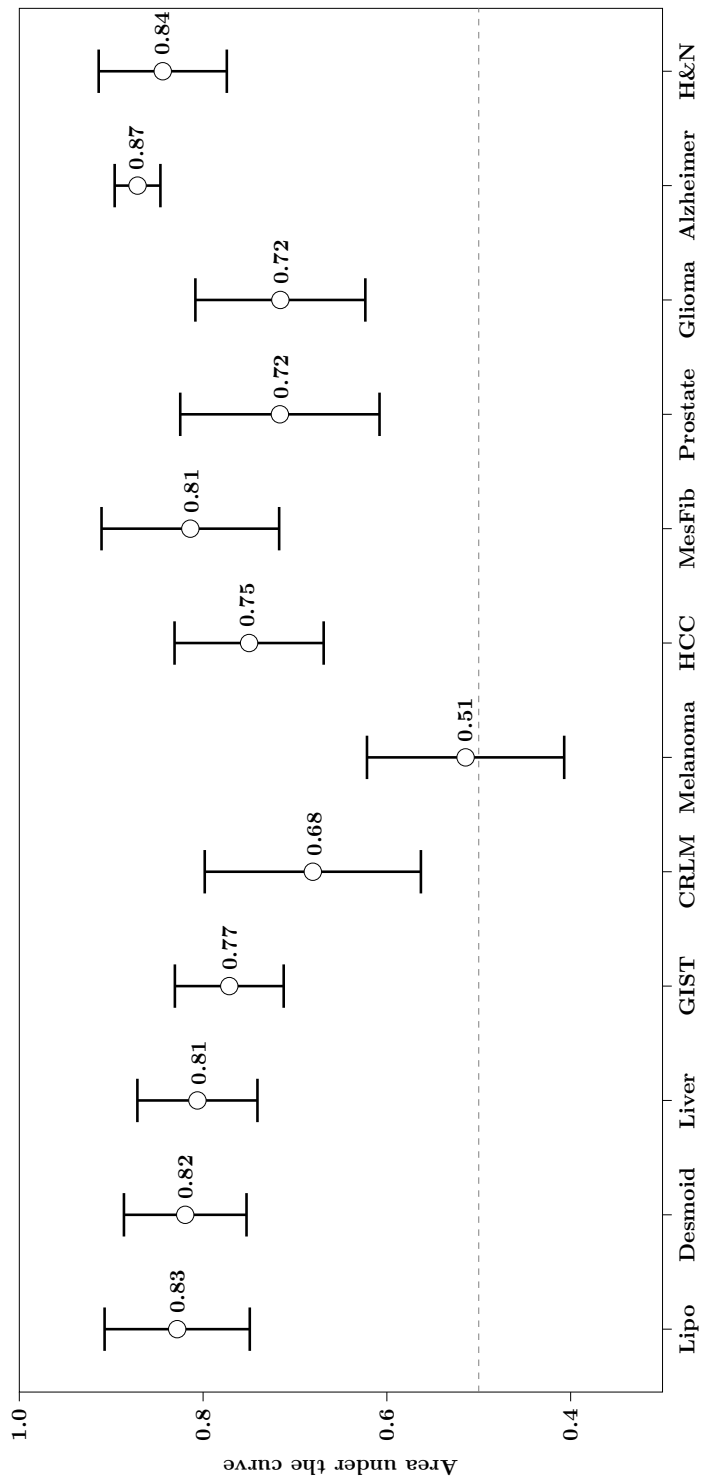


Figure 3.3: Error plots of the area under the receiver operating characteristic curve (AUC) of the radiomics models on twelve datasets. The error plots represent the 95% confidence intervals, estimated through $k_{\text{est}} = 100$ random-split cross-validation on the entire dataset (all except Glioma) or through 1000x bootstrap resampling of the independent test set (Glioma). The circle represents the mean (all except Glioma) or point estimate (Glioma), which is also stated to the right of each circle. The dashed line corresponds to the AUC of random guessing (0.50).

Table 3.3: Classification results of our WORC framework on the twelve datasets. For all metrics, the mean and 95% confidence intervals (CIs) are reported. Abbreviations: AUC: area under the receiver operating characteristic curve; BCR: balanced classification rate [65]; $F_{1,w}$: weighted F1-score. ^x: 95% CI constructed through a $k_{test} = 100$ random-split cross-validation; ^b: 95% CI constructed through a 1000x bootstrap resampling of the test set.

Dataset	Lipo ^x	Desmoid ^x	Liver ^x	GIST ^x	CRLM ^x	Melanoma ^x
AUC	0.83 [0.75, 0.91]	0.82 [0.75, 0.87]	0.81 [0.74, 0.87]	0.77 [0.71, 0.83]	0.68 [0.56, 0.80]	0.51 [0.41, 0.62]
BCR	0.74 [0.66, 0.82]	0.73 [0.66, 0.79]	0.72 [0.65, 0.80]	0.70 [0.65, 0.76]	0.62 [0.51, 0.72]	0.51 [0.42, 0.60]
$F_{1,w}$	0.73 [0.65, 0.82]	0.76 [0.69, 0.82]	0.72 [0.65, 0.79]	0.70 [0.65, 0.75]	0.61 [0.51, 0.72]	0.50 [0.41, 0.59]
Sensitivity	0.72 [0.60, 0.84]	0.59 [0.46, 0.72]	0.74 [0.64, 0.84]	0.67 [0.58, 0.76]	0.60 [0.44, 0.75]	0.53 [0.38, 0.67]
Specificity	0.75 [0.63, 0.88]	0.87 [0.80, 0.93]	0.71 [0.60, 0.83]	0.73 [0.65, 0.81]	0.64 [0.47, 0.80]	0.49 [0.36, 0.62]
Dataset	HCC ^x	MesFib ^x	Prostate ^x	Glioma ^b	Alzheimer ^x	H&N ^x
AUC	0.75 [0.67, 0.83]	0.81 [0.72, 0.91]	0.72 [0.61, 0.82]	0.70 [0.62, 0.81]	0.87 [0.85, 0.90]	0.84 [0.77, 0.91]
BCR	0.69 [0.61, 0.78]	0.72 [0.62, 0.82]	0.67 [0.57, 0.78]	0.62 [0.55, 0.69]	0.78 [0.75, 0.80]	0.75 [0.67, 0.82]
$F_{1,w}$	0.69 [0.60, 0.78]	0.72 [0.61, 0.83]	0.67 [0.56, 0.78]	0.53 [0.43, 0.63]	0.79 [0.77, 0.82]	0.75 [0.67, 0.83]
Sensitivity	0.72 [0.59, 0.85]	0.78 [0.61, 0.94]	0.67 [0.49, 0.85]	0.36 [0.25, 0.47]	0.69 [0.64, 0.75]	0.79 [0.68, 0.90]
Specificity	0.67 [0.54, 0.80]	0.67 [0.49, 0.85]	0.68 [0.53, 0.82]	0.89 [0.79, 0.98]	0.86 [0.83, 0.89]	0.71 [0.58, 0.83]

Table 3.4: Mean and standard deviation (Std) for the weighted F1-score when ten times repeating experiments with varying number of random search iterations (N_{RS}) and ensemble size (N_{ens}) on six different datasets (Lipo, Desmoid, Liver, GIST, CRLM, and Melanoma). The color coding of the mean indicates the relative performance on each dataset (green: high; red: low); the color coding of the standard deviation indicates the relative variation on each dataset (dark: high; light: low).

Lipo	$N_{RS} = 10$		$N_{RS} = 50$		$N_{RS} = 100$		$N_{RS} = 1000$		$N_{RS} = 10000$		$N_{RS} = 25000$
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean
$N_{ens} = 1$	0.754	0.026	0.772	0.021	0.790	0.026	0.784	0.016	0.790	0.012	0.800
$N_{ens} = 10$	0.771	0.007	0.819	0.015	0.833	0.005	0.841	0.004	0.830	0.004	0.830
$N_{ens} = 50$	-	-	0.801	0.008	0.815	0.004	0.855	0.002	0.843	0.002	0.836
$N_{ens} = 100$	-	-	-	-	0.808	0.006	0.853	0.002	0.848	0.001	0.842
Desmoid	$N_{RS} = 10$		$N_{RS} = 50$		$N_{RS} = 100$		$N_{RS} = 1000$		$N_{RS} = 10000$		$N_{RS} = 25000$
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean
$N_{ens} = 1$	0.607	0.023	0.621	0.020	0.612	0.024	0.634	0.018	0.679	0.012	0.689
$N_{ens} = 10$	0.660	0.020	0.701	0.012	0.690	0.013	0.697	0.015	0.706	0.010	0.719
$N_{ens} = 50$	-	-	0.696	0.012	0.709	0.008	0.712	0.008	0.715	0.004	0.717
$N_{ens} = 100$	-	-	-	-	0.699	0.005	0.717	0.005	0.719	0.004	0.717
Liver	$N_{RS} = 10$		$N_{RS} = 50$		$N_{RS} = 100$		$N_{RS} = 1000$		$N_{RS} = 10000$		$N_{RS} = 25000$
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean
$N_{ens} = 1$	0.661	0.022	0.703	0.027	0.713	0.019	0.743	0.023	0.773	0.008	0.778
$N_{ens} = 10$	0.709	0.016	0.755	0.015	0.762	0.011	0.792	0.007	0.805	0.005	0.806
$N_{ens} = 50$	-	-	0.753	0.013	0.767	0.005	0.797	0.004	0.801	0.003	0.807
$N_{ens} = 100$	-	-	-	-	0.766	0.006	0.793	0.003	0.798	0.002	0.803
GIST	$N_{RS} = 10$		$N_{RS} = 50$		$N_{RS} = 100$		$N_{RS} = 1000$		$N_{RS} = 10000$		$N_{RS} = 25000$
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean
$N_{ens} = 1$	0.668	0.038	0.709	0.023	0.712	0.023	0.725	0.019	0.735	0.010	0.733
$N_{ens} = 10$	0.683	0.018	0.744	0.017	0.749	0.009	0.758	0.008	0.756	0.003	0.763
$N_{ens} = 50$	-	-	0.717	0.019	0.738	0.008	0.764	0.002	0.762	0.002	0.762
$N_{ens} = 100$	-	-	-	-	0.725	0.009	0.766	0.002	0.761	0.002	0.761
CRLM	$N_{RS} = 10$		$N_{RS} = 50$		$N_{RS} = 100$		$N_{RS} = 1000$		$N_{RS} = 10000$		$N_{RS} = 25000$
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean
$N_{ens} = 1$	0.545	0.027	0.572	0.038	0.555	0.025	0.589	0.026	0.583	0.015	0.591
$N_{ens} = 10$	0.586	0.025	0.611	0.014	0.619	0.011	0.621	0.010	0.625	0.008	0.615
$N_{ens} = 50$	-	-	0.620	0.014	0.635	0.013	0.633	0.008	0.626	0.005	0.635
$N_{ens} = 100$	-	-	-	-	0.633	0.013	0.639	0.007	0.621	0.005	0.633
Melanoma	$N_{RS} = 10$		$N_{RS} = 50$		$N_{RS} = 100$		$N_{RS} = 1000$		$N_{RS} = 10000$		$N_{RS} = 25000$
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean
$N_{ens} = 1$	0.500	0.018	0.509	0.020	0.506	0.015	0.528	0.021	0.546	0.020	0.552
$N_{ens} = 10$	0.489	0.018	0.506	0.016	0.508	0.011	0.522	0.015	0.539	0.011	0.553
$N_{ens} = 50$	-	-	0.488	0.011	0.495	0.008	0.520	0.009	0.534	0.005	0.537
$N_{ens} = 100$	-	-	-	-	0.490	0.010	0.513	0.004	0.529	0.004	0.536

independent clinical applications, while using the exact same configuration. We were able to find a classification model in eleven applications, indicating that our WORC framework can be used to automatically find radiomics signatures in various clinical applications.

The increase in radiomics studies in recent years has led to a wide variety of radiomics algorithms and related software implementations [17, 24]. For a new clinical application, finding a suitable radiomics workflow has to be done manually, which is a tedious and time consuming process lacking reproducibility. We exploited advances in automated machine learning in order to fully automatically construct complete radiomics workflows from a large search space of radiomics components, including image preprocessing, feature calculation, feature and sample preprocessing, and machine learning algorithms. Hence, our WORC framework streamlines the construction and optimization of radiomics workflows in new applications, and thus facilitates probing datasets for radiomics signatures.

In the field of radiomics, there is a lack of reproducibility, while this is vital for the transition of radiomics models to clinical practice [18, 24]. A recent study [28] even warned that radiomics research must achieve “higher evidence levels” to avoid a reproducibility crisis such as the recent one in psychology [87]. Hence, to facilitate reproducibility, besides automating the radiomics workflows construction, we have publicly released six datasets with a total of 930 patients [35] (i.e., Chapter 4 of this thesis), and made the WORC toolbox and the code to perform our experiments on all datasets open-source [36, 37]. Besides a lack of reproducibility, there is a positive publication bias in radiomics, with as few as 6% of the studies between 2015 and 2018 showing negative results as reported by Buvat *et al.* [88]. They indicate that, to overcome this bias, sound methodology, robustness, reproducibility, and standardization are key. By addressing these factors in our study, including extensive validation of our framework on twelve different clinical applications, we hope to contribute to overcoming the challenges for publishing negative results.

From the twelve datasets included in this study, the melanoma dataset is the only dataset for which we were not able to find a biomarker, which is studied in detail in Angus *et al.* [77] (i.e., Chapter 9 of this thesis). Additionally, Angus *et al.* [77] showed that scoring by a radiologist also led to a negative result. This validates our framework, showing that it does not invent a relation when one does not exist.

Several initiatives towards standardization of radiomics have been formed. The Radiomics Quality Score (RQS) was defined to assess the quality of radiomics studies [30]. While the RQS provides guidelines for the overall study evaluation and reporting, it does not provide standardization of the radiomics workflows or algorithms themselves. The Imaging Biomarker Standardization Initiative (IBSI) [39, 89] provides guidelines for the radiomics feature extraction component and standardization for a set of 174 features (we use 564 features by default, of which a part is included in IBSI). In this study, we complement these important initiatives by addressing the standardization of the radiomics workflow itself.

Related to this work, AutoML has previously been used in radiomics using Tree Based Optimization Tool (TPOT) [90] by Su *et al.* [91] to predict the H3 K27M mutation in midline glioma and Sun *et al.* [92] to predict invasive placentation. These studies are examples of using AutoML to optimize the machine learning

component of radiomics in two specific applications. In this study, we streamlined the construction and optimization of the complete radiomics workflow, included a large collection of commonly used radiomics algorithms and algorithms in the search space, and extensively validated our approach and evaluated its generalizability in twelve different applications. Additionally, our work shows similarities with the Medical Segmentation Decathlon (MSD) [93], in which algorithms were compared on a multitude of segmentation tasks. To this end, the MSD provided data representative of various challenges in medical imaging and created a framework for benchmarking segmentation algorithms and evaluating their generalizability. Although not in a challenge design, our contributions are similar, but on a different task, as we focus on radiomics, i.e., classification of clinical outcomes, instead of segmentation. Moreover, besides comparing a large collection of algorithms, we optimized combining them in a radiomics prediction model using AutoML and ensembling.

The field of medical deep learning faces several similar challenges to conventional radiomics [21, 22, 23, 26]: a lack of standardization, a wide variety of available algorithms, and the need for tuning of model selection and hyperparameters per application. The same problem thus persists: on a given application, from all available deep learning algorithms, how to find the optimal (combination of) workflows? Here, we showed that automated machine learning may be used to streamline this process for conventional radiomics algorithms. Hence, future research may include a similar framework to WORC to facilitate construction and optimization of deep learning workflows, including the full workflow from image to prediction, or a hybrid approach combining deep learning and conventional radiomics. In the field of computer science, the automatic deep learning model selection is addressed in Neural Architecture Search (NAS) [94], which is currently a hot topic in the field of AutoML [95]. In deep learning for medical imaging, NAS is still at an early stage, and the available algorithms mostly focus on segmentation [96]. While the main concept of our framework, i.e., the CASH optimization, could be applied in a similar fashion for deep learning, this poses several challenges. First, deep learning models generally take a lot longer to train, in the order of hours or even days, compared to less than a second for conventional machine learning methods. Our extensive optimization and cross-validation setup is therefore not feasible. Second, the deep learning search space is less clear due to the wide variety of design options, while conventional radiomics workflows typically follow the same steps. Lastly, while current NAS approaches mostly focus on architectural design hyperparameters, pre- and post-processing choices may be equally important to include in the search space [97]. Most NAS methods jointly optimize the network hyperparameters and weights through gradient based optimizations. As the pre- and post-processing are performed outside of the network and require *selector* type hyperparameters, combined optimization with the architectural design options is not trivial.

The two main components of the WORC optimization algorithm are the random search and the ensemble. Our results show that, in line with our hypothesis, increasing N_{ens} at first improves both the performance and the stability of the resulting models. However, as we also hypothesized, when the ratio $N_{\text{ens}}/N_{\text{RS}}$ becomes too large, the performance and stability decrease. On the six datasets in this experiment, the performance and stability at $N_{\text{RS}} = 1000$ was similar to that at

$N_{RS} = 25000$, while the computation time does increase. Therefore, $N_{RS} = 1000$ was chosen as the default in the WORC optimization algorithm, together with $N_{ens} = 100$ to have an optimal N_{ens}/N_{RS} ratio.

For the three previously publicly released datasets from other studies, we compared the performance of our WORC framework to that of the original studies. In the Glioma dataset, our performance (AUC of 0.70) was similar to the original study (van der Voort *et al.* [81]: AUC of 0.72). We thus showed that that our framework was able to successfully construct a signature using an external set of features. Moreover, as the Glioma dataset consists of a separate training and external validation set, we also verified the external-validation setup (Figure 3.A.1 b). In the Alzheimer dataset, our performance (AUC of 0.87) was also similar to the original study (Bron *et al.* [82]: AUC range of 0.80 - 0.94, depending on the level of preprocessing). However, Bron *et al.* [82] used whole-brain voxel-wise features, while we used radiomics features extracted from the hippocampus only. We may therefore have missed information from other brain regions, having a negative effect on the performance in our study. On the H&N dataset, Aerts *et al.* [83] did not evaluate the prognostic value of radiomics for predicting the T-stage, but rather the association through the concordance index (0.69). Moreover, Aerts *et al.* [83] trained a model on a separate dataset of patients with a different clinical application (lung tumors) and externally validated the signature on the H&N dataset, while we performed an internal cross-validation on the H&N dataset. As the lung dataset is not publicly available (anymore), the original experimental setup could not be replicated. Hence the results cannot be directly compared. Concluding, to the extent possible when comparing the results, our WORC framework showed a similar performance as the original studies.

In principle, in any radiomics application, our WORC framework can be used to construct and optimize the radiomics workflow. However, there is a trade-off between the brute-force optimization of our WORC algorithm versus using prior (domain) knowledge to develop a “logical” algorithm. Nonetheless, even in a small search space, deciding purely based on prior knowledge which algorithm will be optimal is complex and generally not feasible. Therefore, we suggest to use domain knowledge to reduce the search space, as it may be possible on certain applications to determine which algorithms *a priori* have a (near) zero chance of succeeding. The WORC optimization algorithm can be used to construct and optimize the radiomics workflow within the remaining search space. Moreover, when the optimal solution is expected to not be included in the default WORC search space and thus a new radiomics method is proposed, this can be added to our framework in a straightforward manner. This facilitates systematic comparison of the new method with the existing, already included methods, and combining the new method with (parts of) the existing methods to optimize the radiomics workflow and increase the overall performance.

In this study, we have not directly compared the performance of WORC to the current standard practice in radiomics. Implicitly, one could argue that this is already done by WORC as various workflows are compared. The comparison is complicated by the lack of standardized methods in radiomics, resulting in variation in “standard” practice. Based on the literature, standard practice can be defined as *a priori* selecting specific methods, commonly one feature selection and one machine learning method,

only tuning a small set of related parameters [6, 16, 17, 18, 20, 21, 22, 23, 24, 26, 27]. Effectively, this corresponds with substantially limiting the WORC search space and not using an ensemble, thus resulting in a similar computation time unless N_{RS} is changed. Future research therefore includes the comparison of the WORC optimization algorithm with this standard practice in radiomics.

Future research could include, firstly, the use of more advanced optimization strategies to improve the performance. Generally, random search, as we use in the WORC optimization algorithm, serves as a solid baseline for optimization problems [48]. However, there is no guarantee that the optimum has been found, and the result may differ when repeating an experiment. The original study introducing CASH used Bayesian optimization, which may overcome these issues [34]. Other strategies include multi-fidelity optimization (e.g. bandits), genetic or evolutionary algorithms [91], or gradient based optimization [33]. However, the hyperparameter space in the WORC framework is relatively large due to the inclusion of multiple (optional) algorithm collections instead of just one, making optimization more complex and computationally expensive. Moreover, optimizing the performance further on the validation set may result in overfitting [33], therefore actually resulting in worse generalization. Secondly, as we evaluated our framework on twelve different datasets, when applying WORC on a new dataset, meta-learning could be used to learn from the results on these previous twelve datasets [33]. Especially on smaller datasets, taking into account which solutions worked best on previous datasets may improve the performance and lower the computation time. Thirdly, future research into the use of more advanced ensembling strategies may also improve the performance and stability [98]. Lastly, our framework may be used on other clinical applications to automatically optimize radiomics workflows. While we only showed the use of our framework on CT and MRI, the used features have also been shown to be successful in other modalities such as PET [99] and ultrasound [100], and thus the WORC framework could also be useful in these modalities.

3.6 Conclusions

In this study, we proposed a framework for the fully automatic construction and optimization of radiomics workflows to generalize radiomics across applications. The framework was validated on twelve different, independent clinical applications, on eleven of which our framework automatically constructed a successful radiomics model. On the three datasets of these that were previously publicly released and analyzed with different methods, we achieved a similar performance as that of the original studies. Hence, our framework may be used to streamline the construction and optimization of radiomics workflows on new applications, and thus for probing datasets for radiomics signatures. By releasing six datasets publicly, and the WORC toolbox implementing our framework and the code to reproduce the experiments of this study open-source, we aim to facilitate reproducibility and validation of radiomics algorithms.

Data statement

Six of the datasets used in this study (Lipo, Desmoid, Liver, GIST, CRLM, and Melanoma), comprising a total of 930 patients, are publicly released as part of this study and hosted via a public XNAT³ as published in Starmans *et al.* [35] (i.e., Chapter 4 of this thesis). By storing all data on XNAT in a structured and standardized manner, experiments using these datasets can be easily executed at various computational resources with the same code.

Three datasets were already publicly available as described in Section 3.3. The other three datasets could not be made publicly available. The code for the experiments on the nine publicly available datasets is available on GitHub [37].

Acknowledgments

The authors thank Laurens Groenendijk for his assistance in processing the data and in the anonymization procedures, and Hakim Achterberg for his assistance in the development of the software. This study is supported by EuCanImage (European Union's Horizon 2020 research and innovation programme under grant agreement Nr. 952103). This work was partially carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

Data collection and sharing for this project was partially funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

³<https://xnat.bmia.nl/data/projects/worc>

Funding

Martijn P. A. Starmans and Jose M. Castillo T. acknowledge funding from the research program STRaTeGy with project numbers 14929, 14930, and 14932, which is (partly) financed by the Netherlands Organization for Scientific Research (NWO). Sebastian R. van der Voort acknowledges funding from the Dutch Cancer Society (KWF project number EMCR 2015-7859). Part of this study was financed by the Stichting Coolingsingel (reference number 567), a Dutch non-profit foundation.

Competing interests statement

Wiro J. Niessen is founder, scientific lead, and shareholder of Quantib BV. Jacob J. Visser is a medical advisor at Contextflow. Astrid A. M. van der Veldt is a consultant (fees paid to the institute) at BMS, Merck, MSD, Sanofi, Eisai, Pfizer, Roche, Novartis, Pierre Fabre and Ipsen. The other authors do not declare any conflicts of interest.

CRedit author statement

M.P.A.S., W.J.N., and S.K. provided the conception and design of the study. M.P.A.S., M.J.M.T., M.V., G.A.P., W.K., D.H., D.J.G., C.V., S.S., R.S.D., C.J.E., F.F., G.J.L.H.v.L., A.B., J.H., T.B., R.v.G., G.J.H.F., R.A.F., W.W.d.H., F.E.B., F.E.J.A.W., B.G.K., L.A., A.A.M.v.d.V., A.R., A.E.O., J.M.C.T., J.V., I.S., M.R., Mic.D., R.d.M., J.IJ., R.L.M., P.B.V., E.E.B., M.G.T., and J.J.V. acquired the data. M.P.A.S., S.R.v.d.V., M.J.M.T., M.V., A.B., F.E.B., L.A., Mit.D., J.M.C.T., R.L.M., E.B., M.G.T. and S.K. analyzed and interpreted the data. M.P.A.S., S.R.v.d.V., T.P., and Mit.D. created the software. M.P.A.S. and S.K. drafted the article. All authors read and approved the final manuscript.

Ethics statement

The protocol of this study conformed to the ethical guidelines of the 1975 Declaration of Helsinki. Approval by the local institutional review board of the Erasmus MC (Rotterdam, the Netherlands) was obtained for collection of the WORC database (MEC-2020-0961), and separately for eight of the studies using in-house data (Lipo: MEC-2016-339, Desmond: MEC-2016-339, Liver: MEC-2017-1035, GIST: MEC-2017-1187, CRLM: MEC-2017-479, Melanoma: MEC-2019-0693, HCC: MEC-2018-1621, Prostate: NL32105.078.10). The need for informed consent was waived due to the use of anonymized, retrospective data. For the last study involving in-house data, the Mesfib study, as the study was retrospectively performed with anonymized data, no approval from the ethical committee or informed consent was required.

Appendix

Algorithm 3.A.1 Minimal working example of the WORC toolbox interface in Python

```
from WORC import SimpleWORC

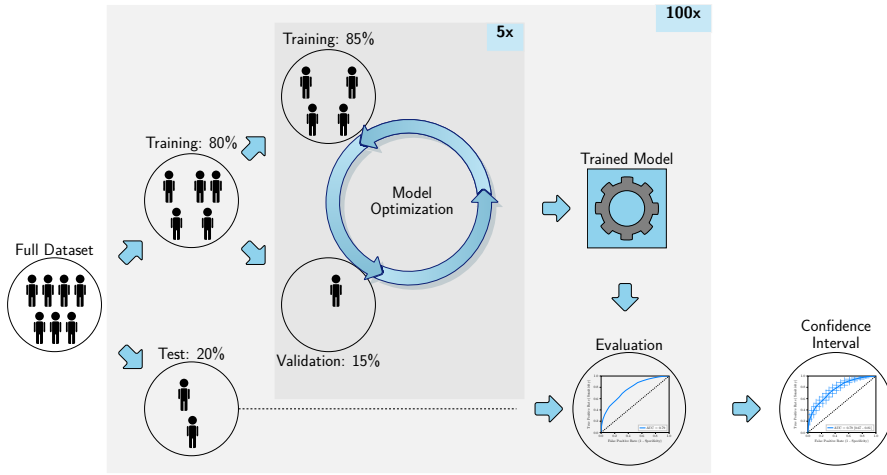
# Create a Simple WORC object
experiment = SimpleWORC(experiment_name)

# Set the input data according to the variables we defined earlier
experiment.images_from_this_directory(imagedatadir)
experiment.segmentations_from_this_directory(imagedatadir)
experiment.labels_from_this_file(label_file)
experiment.predict_labels(label_name)

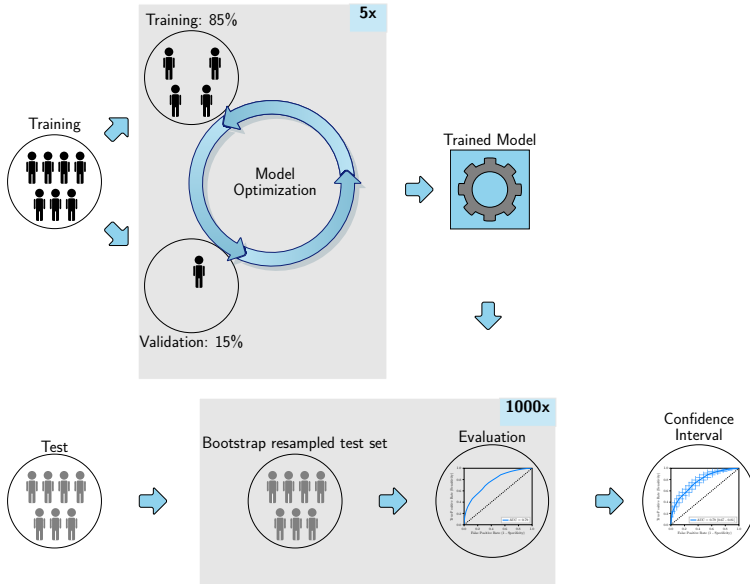
# Use the standard workflow for binary classification
experiment.binary_classification()

# Change a configuration field to only use an SVM
experiment.add_config_overrides({'Classification': {'classifiers': 'SVM'}})

# Run the experiment!
experiment.execute()
```



A. Internal validation



B. External validation

Figure 3.A.1: Cross-validation setups used by our WORC framework for optimization and evaluation. When a single dataset is used, internal validation is performed through a $k_{\text{test}} = 100$ random-split cross-validation (a). When fixed, separate training and test datasets are used, external validation is performed by developing the model on the training set and evaluating the performance on the test set through 1000x bootstrap resampling (b). Both include an internal $k_{\text{training}} = 5$ random-split cross-validation on the training set to split the training set into parts for actual training and validation, in which the model optimization is performed. The final selected model, trained on the full training dataset, is used for independent testing on the test dataset.

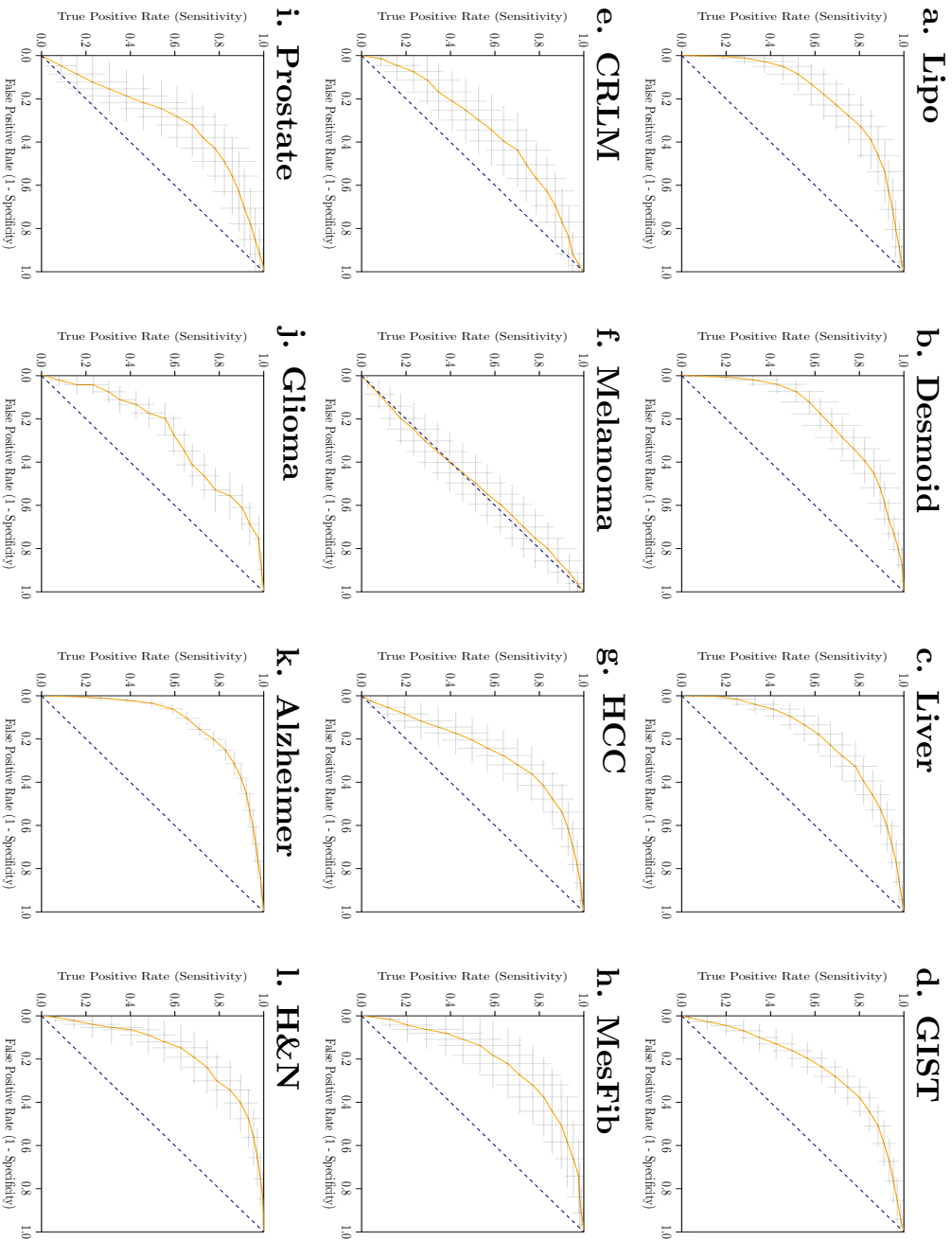
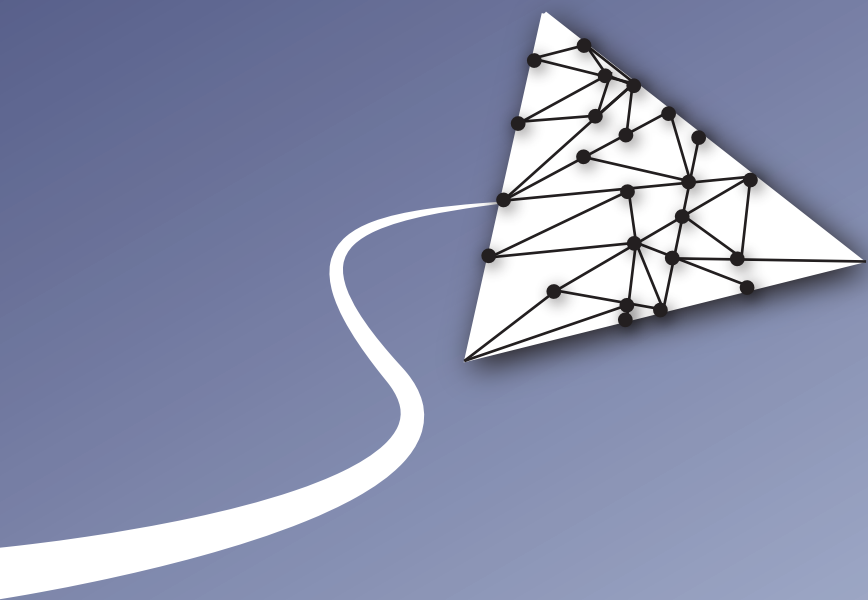


Figure 3.A.2: ROC curves. The datasets include: a. lipomatous tumors [72] (i.e., Chapter 5 of this thesis); b. desmoid-type fibromatosis [73] (i.e., Chapter 6 of this thesis); c. primary solid liver tumors [74] (i.e., Chapter 12 of this thesis); d. gastrointestinal stromal tumors [75] (i.e., Chapter 7 of this thesis); e. colorectal liver metastases [76] (i.e., Chapter 11 of this thesis); f. melanoma [77] (i.e., Chapter 9 of this thesis); g. hepatocellular carcinoma [78]; h. mesenteric fibrosis [79] (i.e., Chapter 10 of this thesis); i. prostate cancer [80]; j. low grade glioma [81]; k. Alzheimer's disease [82]; and l. head and neck cancer [83].

Table 3.A.1: Overview of the 564 features used by default in the `WORC` framework. GLCM features were calculated in four different directions (0, 45, 90, 135 degrees) using 16 gray levels and pixel distances of 1 and 3. LBP features were calculated using the following three parameter combinations: 1 pixel radius and 8 neighbors, 2 pixel radius and 12 neighbors, and 3 pixel radius and 16 neighbors. Gabor features were calculated using three different frequencies (0.05, 0.2, 0.5) and four different angles (0, 45, 90, 135 degrees). LoG features were calculated using three different widths of the Gaussian (1, 5 and 10 pixels). Vessel features were calculated using the full mask, the edge, and the inner region. Local phase features were calculated on the monogenic phase, phase congruency and phase symmetry.

Histogram (13 features)	LoG (13*3=39 features)	Vessel (12*3=39 features)	GLCM (MS) (6*3*4*2=144 features)	Gabor (13*3*4=156 features)	NGTDM (5 features)	LBP (13*3=39 features)
min	min	min	contrast (normal, MS mean + std)	min	busyness	min
max	max	max	dissimilarity (normal, MS mean + std)	max	coarseness	max
mean	mean	mean	homogeneity (normal, MS mean + std)	mean	complexity	mean
median	median	median	angular second moment (ASM) (normal, MS mean + std)	median	contrast	median
std	std	std	energy (normal, MS mean + std)	std	strength	std
skewness	skewness	skewness	correlation (normal, MS mean + std)	skewness		skewness
kurtosis	kurtosis	kurtosis		kurtosis		kurtosis
peak	peak	peak		peak		peak
peak position	peak position	peak position		peak position		peak position
range	range	range		range		range
energy	energy	energy		energy		energy
quartile range	quartile range	quartile range		quartile range		quartile range
entropy	entropy	entropy		entropy		entropy
GLSZM (16 features)	GLRM (16 features)	GLDM (14 features)	Shape (35 features)	Orientation (9 features)	Local phase (13*3=39 features)	
Gray Level Non Uniformity	Gray Level Non Uniformity	Dependence Entropy	compactness (mean + std)	theta_x	min	
Gray Level Non Uniformity Normalized	Gray Level Non Uniformity Normalized	Dependence Non-Uniformity	radial distance (mean + std)	theta_y	max	
Gray Level Variance	Gray Level Variance	Dependence Non-Uniformity Normalized	roughness (mean + std)	theta_z	mean	
High Gray Level Zone Emphasis	High Gray Level Run Emphasis	Dependence Variance	convexity (mean + std)	COM index x	median	
Large Area Emphasis	Long Run Emphasis	Gray Level Non-Uniformity	circular variance (mean + std)	COM index y	std	
Large Area High Gray Level Emphasis	Long Run High Gray Level Emphasis	Gray Level Variance	principal axes ratio (mean + std)	COM index z	skewness	
Large Area Low Gray Level Emphasis	Long Run Low Gray Level Emphasis	High Gray Level Emphasis	elliptic variance (mean + std)	COM x	kurtosis	
Low Gray Level Zone Emphasis	Low Gray Level Run Emphasis	Large Dependence Emphasis	solidity (mean + std)	COM y	peak	
SizeZoneNonUniformity	RunEntropy	Large Dependence High Gray Level Emphasis	area (mean, std, min + max)	COM z	peak position	
SizeZoneNonUniformityNormalized	RunLengthNonUniformity	Large Dependence Low Gray Level Emphasis	volume (total, mesh, volume)		range	
SmallAreaEmphasis	RunLengthNonUniformityNormalized	Low Gray Level Emphasis	elongation		energy	
SmallAreaHighGrayLevelEmphasis	RunPercentage	Small Dependence Emphasis	flatness		quartile	
SmallAreaLowGrayLevelEmphasis	RunVariance	Small Dependence High Gray Level Emphasis	least axis length		entropy	
ZoneEntropy	ShortRunEmphasis	Small Dependence Low Gray Level Emphasis	major axis length			
ZonePercentage	ShortRunHighGrayLevelEmphasis		minor axis length			
ZoneVariance	ShortRunLowGrayLevelEmphasis		maximum diameter 3D (rows, columns, slices)			
			maximum diameter 2D			
			sphericity			
			surface area			
			surface volume ratio			

*Abbreviations: COM: center of mass; GLCM: gray level co-occurrence matrix; MS: multi slice; NGTDM: neighborhood gray tone difference matrix; GLSZM: gray level size zone matrix; GLRLM: gray level run length matrix; LBP: local binary patterns; LoG: Laplacian of Gaussian; std: standard deviation.



4.

The WORC* database: MRI and CT scans, segmentations, and clinical labels for 930 patients from six radiomics studies

Based on: **M. P. A. Starmans**, M. J. M. Timbergen, M. Vos, G. A. Padmos, D. J. Grünhagen, C. Verhoef, S. Sleijfer, G. J. L. H. van Leenders, F. E. Buisman, F. E. J. A. Willemssen, B. G. Koerkamp, L. Angus, A. A. M. van der Veldt, A. Rajicic, A. E. Odink, M. Renckens, M. Doukas, R. A. de Man, J. N. M. IJzermans, R. L. Miclea, P. B. Vermeulen, M. G. Thomeer, J. J. Visser, W. J. Niessen, and S. Klein, "The WORC* database: MRI and CT scans, segmentations, and clinical labels for 930 patients from six radiomics studies," *Submitted*, 2021. medRxiv: [2021.08.19.21262238](https://doi.org/10.1101/2021.08.19.21262238) (eess.IV)

Abstract

The WORC database consists in total of 930 patients composed of six datasets gathered at the Erasmus MC, consisting of patients with: 1) well-differentiated liposarcoma or lipoma (115 patients); 2) desmoid-type fibromatosis or extremity soft-tissue sarcomas (203 patients); 3) primary solid liver tumors, either malignant (hepatocellular carcinoma or intrahepatic cholangiocarcinoma) or benign (hepatocellular adenoma or focal nodular hyperplasia) (186 patients); 4) gastrointestinal stromal tumors (GISTs) and intra-abdominal gastrointestinal tumors radiologically resembling GISTs (246 patients); 5) colorectal liver metastases (77 patients); and 6) lung metastases of metastatic melanoma (103 patients). For each patient, either a magnetic resonance imaging (MRI) or computed tomography (CT) scan, collected from routine clinical care, one or multiple (semi-)automatic lesion segmentations, and ground truth labels from a gold standard (e.g., pathologically proven) are available. All datasets are multi-center imaging datasets, as patients referred to our institute often received imaging at their referring hospital. The dataset can be used to validate or develop radiomics methods, i.e., using machine or deep learning to relate the visual appearance to the ground truth labels, and automatic segmentation methods. See also the research article related to this dataset: Starmans et al., *Reproducible radiomics through automated machine learning validated on twelve clinical applications*, Submitted (i.e., [Chapter 3](#) in this thesis).

4.1 Value of the data

- This dataset provides imaging data, outlined lesions, age, sex, and ground truth labels (e.g., diagnosis, genetic mutations, biological characteristics), mostly obtained from pathology, for a large number of patients from six different cancer studies. Publicly sharing imaging data with ground truth labels and segmentations benefits reproducibility, enables external validation, and hence accelerates transition to clinical practice [18, 24, 31]. This dataset has been collected in routine clinical care at multiple centers, thus representing the real-life variability and heterogeneity of the data. For these reasons, this dataset is a valuable resource.
- This dataset will be beneficial for researchers working on computer aided diagnosis for cancer based on imaging, specifically in the areas of liposarcoma, desmoid type-fibromatosis, gastrointestinal stromal lesions, sarcoma, primary liver cancer, (colorectal) liver metastases, and (melanoma) lung metastases.
- This data can be used to validate or develop radiomics methods (i.e., using conventional machine learning or deep learning to relate the visual appearance to the ground truth labels) and automated segmentation methods. For example, the data can be used as a large, heterogeneous independent test set, or to increase the size and heterogeneity of train sets for developing new methods.

4.2 Data description

The WORC dataset contains 930 patients and is composed of six radiomics studies, coined the Lipo ([Subsection 4.2.1](#)), Desmoid ([Subsection 4.2.2](#)), Liver ([Subsection 4.2.3](#)), GIST ([Subsection 4.2.4](#)), CRLM ([Subsection 4.2.5](#)), and Melanoma ([Subsection 4.2.6](#)) dataset. All datasets were collected at the Erasmus MC, Rotterdam, the Netherlands, but are multi-center imaging datasets, as patients referred to our institute often received imaging at their referring hospital. Example images of each dataset are shown in [Figure 4.1](#).

For each study, five different sources of data are provided:

1. Routine clinical MRI (Lipo, Desmoid, Liver) or CT (GIST, CRLM, Melanoma) scans
2. Details on the acquisition protocols ([Subsection 4.2.7](#))
3. Lesion segmentations
4. Age and sex
5. Pathological ground truth labels

The data is available on an XNAT server; an online platform to store (medical) imaging data in a standardized way, allowing access through both a Graphical User Interface (GUI) and an Application Programming Interface (API) [71]. The datasets

Table 4.1: Specifications of the data.

Subject	Medical Imaging
Specific subject area	Routine MRI and CT scans, lesion segmentations, clinical labels of six radiomics studies
Type of data	Medical Imaging data (NIfTI files): MRI data: T1-weighted T2-weighted CT data Medical Imaging metadata (JSON files) Segmentations (NIfTI files) Patient data (Excel files): Age Sex Pathological ground truth (Excel files, subject level variables)
How data were acquired	MRI and CT scans were acquired on 177 different scanners. Age and sex were obtained from patient records. Ground truth data were obtained from a gold standard, mainly by pathological analysis of tumor tissue obtained from either biopsy or resection. An exception was made for “typical” focal nodular hyperplasia (FNH) [12], which was confirmed radiologically. Whole-tumor segmentations were semi-automatically annotated by various observers.
Data format	Raw
Parameters for data collection	MRI and CT scans were acquired with a variety of image acquisition protocols.
Description of data collection	Pre-treatment imaging data and ground truth data were retrospectively included at the Erasmus MC from patients with: Well-differentiated liposarcoma or lipoma between 2009 - 2018 Desmoid-type fibromatosis and extremity soft-tissue-sarcoma between 1990 - 2018 Primary solid liver tumors between 2002 - 2018 Gastrointestinal stromal tumors or similar intra-abdominal tumors between 2004 - 2017 Colorectal liver metastases between 2003 - 2015 Lung metastases of melanoma between 2012 - 2018

Data source location	Erasmus MC (University Medical Center), Rotterdam, The Netherlands
Data accessibility	Repository name: Health-RI XNAT Data identification number: WORC Direct URL to data: https://xnat.bmia.nl/data/projects/worc Data usage agreement: https://xnat.bmia.nl/data/projects/worc/resources/License/files/WORC_data_license.pdf Data downloader: https://doi.org/10.5281/zenodo.5119040
Related research article	Starmans <i>et al.</i> [101] (i.e., Chapter 3 of this thesis)

for this study are publicly hosted on the Health-RI XNAT ¹. Code to download the data locally, and code to reproduce the experiments from Starmans *et al.* [102] on these datasets, have been released open-source [37].

For each study, details on the ground truth labels and the data collection are given in the respective subsections. The acquisition protocol details for all studies are described in [Subsection 4.2.7](#). The scans have been converted from DICOM to NIfTI using the `dcm2nii` toolbox version v1.0.20180518 [103]. For each patient, a single scan is included and provided as NIfTI files named “*image.nii.gz*”. The associated details on the scan acquisition protocol are given in a JSON file named “*metadata.json*”. The corresponding segmentation is given in the NIfTI file “*segmentation.nii.gz*”, where a label of 1 indicates a lesion and a label of 0 indicates background. For the CRLM dataset, multiple segmentations of various lesions made by multiple observers are given, see [Subsection 4.2.5](#). The ground truth pathological labels for all studies are combined in the Excel sheet “*Clinical_data.xlsx*” and as labels on subject level in the XNAT project to allow for easier automatic processing.

4.2.1 The Lipo dataset

This dataset consists of 115 patients with either a well-differentiated liposarcoma (WDLPS) ($N = 58$) or lipoma ($N = 58$), as described in Vos *et al.* [72] (i.e., [Chapter 5](#) of this thesis). One patient has both a WDLPS and a lipoma, thus the dataset in total contains 116 lesions. For each patient, a T1-weighted MRI scan is provided. The ground truth label, i.e., whether a lesion was a WDLPS or lipoma, is represented by the *MDM2* amplification. The *MDM2* amplification status for each patient is provided, where patients have label 1 if the lesion was a WDLPS, and label 0 if the lesion was a lipoma.

For the patient with both a WDLPS and a lipoma, a segmentation is provided for each lesion: “*segmentation_WDLPS.nii.gz*” and “*segmentation_lipoma.nii.gz*”

¹<https://xnat.bmia.nl/data/projects/worc>

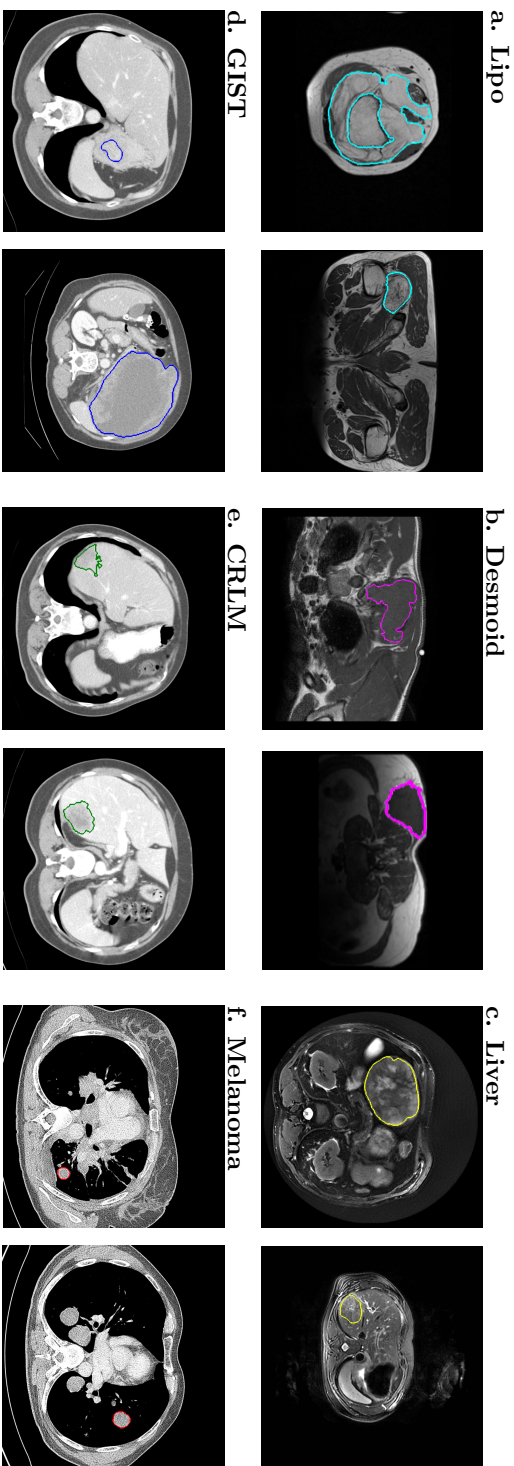


Figure 4.1: Examples of the 2D slices from the 3D imaging data from the six datasets included in the WORC dataset. For each dataset, for one patient of each of the two classes, the 2D slice in the primary scan direction (e.g., axial) with the largest area of the segmentation is depicted; the boundary of the segmentation is projected in color on the image. The datasets included were from different clinical applications: a. lipomatous tumors [72] (i.e., Chapter 5 of this thesis); b. desmoid-type fibromatosis [73] (i.e., Chapter 6 of this thesis); c. primary solid liver tumors [74] (i.e., Chapter 12 of this thesis); d. gastrointestinal stromal tumors [75] (i.e., Chapter 7 of this thesis); e. colorectal liver metastases [76] (i.e., Chapter 11 of this thesis); and f. melanoma [77] (i.e., Chapter 9 of this thesis).

4.2.2 The Desmoid dataset

This dataset consists of 203 patients with either desmoid-type fibromatosis (DTF) ($N = 72$) or extremity soft-tissue sarcomas (STS), i.e., the non-DTF group ($N = 131$), as described in Timbergen *et al.* [73] (i.e., Chapter 6 of this thesis). The non-DTF group consists of 64 myxofibrosarcomas, 31 leiomyosarcomas, and 36 myxoid liposarcomas. For each patient, a T1-weighted MRI scan is provided. The ground truth label, i.e., whether a lesion was a DTF or one of the non-DTF phenotypes, was confirmed by histology. The differential diagnosis for each patient is provided, where patients have label 1 if the lesion was a DTF, and label 0 if the lesion was a non-DTF. The subtype of the non-DTF lesions is also provided.

4.2.3 The Liver dataset

This dataset consists of 186 patients with either a malignant ($N = 94$) or benign ($N = 93$) primary solid liver tumor, as described in Starmans *et al.* [74] (i.e., Chapter 12 of this thesis). For each patient, a T2-weighted MRI scan is provided. The malignant group includes 81 hepatocellular carcinoma (HCC) and 13 intrahepatic cholangiocarcinoma (iCCA); the benign group includes 48 hepatocellular adenoma (HCA) and 44 FNH. The ground truth label, i.e., the phenotype of a lesion, was based on pathology. An exception are “typical” FNH [12], for which the ground truth was established radiologically. The differential diagnosis for each patient is provided, where patients have label 1 if the lesion was malignant, and label 0 if the lesion was benign. The phenotype of the lesions is also provided.

4.2.4 The GIST dataset

This dataset consists of 246 patients with either gastrointestinal stromal lesions (GISTs) ($N = 125$) or intra-abdominal tumors radiologically resembling GIST (non-GIST) ($N = 122$), as described in Starmans *et al.* [75] (i.e., Chapter 7 of this thesis). One patient has two GISTs, thus the dataset in total contains 247 lesions. The non-GIST group consists of 22 schwannoma, 25 leiomyosarcoma, 25 leiomyoma, 25 esophageal or gastri junctional adenocarcinoma, and 25 lymphoma. For each patient, a contrast-enhanced venous phase CT scan is provided. The ground truth label, i.e., whether a lesion was a GIST or one of the non-GIST phenotypes, was confirmed by histology. The differential diagnosis for each patient is provided, where patients have label 1 if the lesion was a GIST, and label 0 if the lesion was a non-GIST. The subtype of the non-GIST lesions is also provided.

4.2.5 The CRLM dataset

This dataset consists of 77 patients with a total of 93 colorectal liver metastases (CRLM) with either a 100% desmoplastic histopathological growth patterns (HGP) [104] ($N = 46$) or 100% replacement HGP ($N = 47$), as described in Starmans *et al.* [76] (i.e., Chapter 11 of this thesis)². For each patient, a portal venous phase CT

²Starmans *et al.* [76] reported a total of 76 patients, but the dataset did actually contain 77 patients.

scan is provided. The ground truth label, i.e., whether a lesion had a desmoplastic or replacement HGP, was determined on hematoxylin and eosin stained tissue sections. The HGP type for each patient is provided, where patients have label 1 if the lesions had replacement HGP, and label 0 if the lesions had a desmoplastic HGP. As the HGP is assumed to be the same for all lesions of a subject, the ground truth is provided on subject level.

For each patient, for each lesion, segmentations by three clinicians (STUD1, PhD, RAD) and a Convolutional Neural Network (CNN) are available: e.g. “*segmentation_lesion1_STUD1.nii.gz*”, “*segmentation_lesion1_PhD.nii.gz*”, “*segmentation_lesion1_RAD.nii.gz*”, and “*segmentation_lesion1_CNN.nii.gz*”. Additionally, each lesion was segmented a second time by the first observer (STUD2), and is named e.g. “*segmentation_lesion1_STUD2.nii.gz*”. Note that 8 out of the 93 lesions (9%) were missed by the CNN, and thus do not include a CNN segmentation

4.2.6 The Melanoma dataset

This dataset consists of 169 lung metastases of 103 patients with *BRAF* mutated ($N = 51$) or *BRAF* wild type ($N = 52$) metastatic melanoma, as described in Angus *et al.* [77] (i.e., Chapter 9 of this thesis). For each patient, a contrast-enhanced thoracic CT scan is provided. When multiple lesions were included, the corresponding segmentations are named “*segmentation_lesion1.nii.gz*”, “*segmentation_lesion2.nii.gz*”, and so on. The ground truth label, i.e., whether lesions from a patient were *BRAF* mutated or *BRAF* wild type, is provided, where patients have label 1 if the lesions were *BRAF* mutated, and label 0 if the lesions were *BRAF* wild type. As the *BRAF* mutation is assumed to be the same for all lesions of a subject, the ground truth is provided on subject level.

4.2.7 Acquisition protocol details

From the original DICOM files from the MRI and CT scans, the values of several tags were extracted to provide information on the used acquisition protocols, which for each scan are included in a *metadata.json* file.

For both MRI and CT scans, the following general acquisition protocol details from the following DICOM tags are included:

(0008, 0060) Modality	(0018, 0083) Number of averages
(0008, 0070) Manufacturer	(0018, 0084) Spacing between slices
(0008, 1090) Model name	(0018, 0093) Percent sampling
(0018, 0020) Scanning sequence	(0018, 1030) Protocol name
(0018, 0022) Scan options	(0018, 5100) Patient position
(0018, 0023) Acquisition type	(0020, 0037) Orientation
(0018, 0024) Sequence name	(0028, 0030) Pixel spacing
(0018, 0050) Slice thickness	

For each MRI scan, the following specific acquisition protocol details from the following DICOM tags are additionally included:

(0018, 0080) Repetition time	(0018, 0091) Echo train length
(0018, 0081) Echo time	(0018, 1250) Coil
(0018, 0082) Inversion time	(0018, 1310) Acquisition matrix
(0018, 0084) Imaging frequency	(0018, 1312) Encoding direction
(0018, 0087) Tesla	(0018, 1314) Flip angle

For each CT scan, the following specific acquisition protocol details from the following DICOM tags are additionally included:

(0018, 0060) KVP (kilovoltage peak)	(0018, 1210) Convolution kernel
-------------------------------------	---------------------------------

4.3 Experimental design, materials and methods

4.3.1 The Lipo dataset

Patients that were either referred to/discussed at, or diagnosed/treated at the Erasmus MC Cancer Institute, Rotterdam, the Netherlands, between December 2009 and August 2018 with a pathologically confirmed diagnosis of lipoma or WDLPS were retrospectively included. Inclusion criteria were: a known *MDM2* amplification status tested by fluorescence *in situ* hybridization (FISH); and at least a T1-weighted MRI sequence available before treatment (if applicable).

The lipoma and WDLPS lesions were segmented semi-automatically on the T1-weighted MRI [105]. All images were segmented independently by either a medical masters student or a PhD candidate with an MD degree. Both were blinded to the type of lipomatous lesion. To validate segmentation accuracy, a sample set was verified by a musculoskeletal radiologist, specialized in soft-tissue sarcomas (4 years of experience). Semi-automatic results were always reviewed and manually corrected when necessary, to assure the result resembled manual segmentation.

4.3.2 The Desmoid dataset

Patients that were either referred to/discussed at, or diagnosed/treated at the Erasmus MC Cancer Institute, Rotterdam, the Netherlands, between 1990 and 2018 with histologically proven primary or recurrent DTF, or a malignant extremity STS, were retrospectively included. Inclusion criteria were: at least a T1-weighted MRI sequence available before treatment (if applicable); for the STS, a histologically proven primary myxofibrosarcoma, myxoid liposarcoma or leiomyosarcoma of the extremities.

The DTF and STS were all segmented semi-automatically on the T1-weighted MRI [105]. All images were segmented independently by either a medical masters student or a PhD candidate with an MD degree under supervision of a musculoskeletal radiologist (4 years of experience). Both were blinded to the type of lesion. Semi-automatic results were always reviewed and manually corrected when necessary, to assure the result resembled manual segmentation.

4.3.3 The Liver dataset

Patients that were either referred to/discussed at, or diagnosed/treated at the Erasmus MC Cancer Institute, Rotterdam, the Netherlands, between 2002 and 2018 with a primary solid liver lesion were retrospectively included. Inclusion criteria were: HCC, iCCA, HCA or FNH; pathologically proven phenotype; and availability of a T2-weighted MRI scan. An exception to the pathologically proven phenotype was made for typical FNH, which are routinely not biopsied and diagnosed radiologically [106], as typical FNH imaging characteristics are 100% specific [107]. Exclusion criteria were: maximum diameter equal to or smaller than 3 cm; underlying liver disease; and significant imaging artefacts.

The lesions were all segmented semi-automatically on the T2-weighted MRI [105]. All images were segmented independently by one of two experienced abdominal radiologists (21 and 8 years of experience). Both were blinded to the type of lesion. Semi-automatic results were always reviewed and manually corrected when necessary, to assure the result resembled manual segmentation.

4.3.4 The GIST dataset

Patients that were either referred to/discussed at, or diagnosed/treated at the Erasmus MC Cancer Institute, Rotterdam, the Netherlands, between 2004 and 2017 with a histopathologically proven primary GIST or intra-abdominal tumors radiologically resembling GIST were retrospectively included. The inclusion criterion was availability of at least a contrast-enhanced venous-phase CT prior to treatment. The sample sizes of the non-GIST and the GIST cohort were matched. The non-GIST subtypes were balanced, i.e. a similar number of patients per subtype was randomly included.

The lesions were all segmented semi-automatically on the CT scan [105]. All images were segmented independently by either a medical masters student or a PhD candidate with an MD degree under supervision of a musculoskeletal radiologist (5 years of experience). Both were blinded to the type of lesion. Semi-automatic results were always reviewed and manually corrected when necessary, to assure the result resembled manual segmentation.

4.3.5 The CRLM dataset

Patients that were surgically treated at the Erasmus MC Cancer Institute, Rotterdam, the Netherlands, between 2003 and 2015 with CRLM were included. Inclusion criteria were: availability of at least a contrast-enhanced venous-phase CT prior to treatment; available hematoxylin and eosin stained tissue sections; either a 100% desmoplastic HGP or a 100% replacement HGP. Exclusion criteria were: recurrent CRLM or CRLM requiring two-staged resections; and treatment with preoperative chemotherapy, since chemotherapy may alter HGPs [104]. HGPs were scored on resection specimens according to the consensus guidelines by an expert pathologist (PV) [108].

The lesions were all segmented semi-automatically on the CT scan [105]. Lesion segmentation was performed by four observers: a medicine student with no relevant

experience (STUD1), a PhD student (PhD) with limited experience, an expert abdominal radiologist (RAD), and an automatic CNN. The student segmented all lesions a second time (STUD2). All observers were blinded to the type of lesion. Semi-automatic results were always reviewed and manually corrected when necessary, to assure the result resembled manual segmentation.

The CNN used for the automatic segmentations was the Hybrid-Dense-UNet, which achieved state-of-the-art performance on the LITS liver tumor segmentation challenge and is open-source [109, 110]. The original CNN as trained on the LITS data was used. From the CNN lesion segmentations, only lesions that had histology were extracted, and the segmentations were saved per lesion.

4.3.6 The Melanoma dataset

Patients that were diagnosed with metastatic melanoma at the Erasmus MC Cancer Institute, Rotterdam, the Netherlands, between January 2012 and February 2018 were retrospectively included. Inclusion criteria were: known tumor *BRAF* mutation, diagnostic contrast-enhanced thoracic CT scan prior to commencement of any systemic therapy, and at least one lung metastasis of ≥ 10 mm evaluable according to Response Evaluation Criteria In Solid Tumors (RECIST) v1.1 [8]. Patients with *BRAF* mutations other than p.V600E were excluded. Formalin-fixed paraffin embedded material of the primary tumor and/ or metastasis was tested for *BRAF* (exon 15) using a polymerase chain reaction based assay or next generation sequencing as part of standard care.

Per patient, up to two lung lesions ≥ 10 mm were selected by a clinician supervised by an experienced chest radiologist and segmented semi-automatically on the CT scan [105]. In patients with >2 lung metastases of ≥ 10 mm, either the two largest or the two most easily distinguishable lesions were segmented (i.e., two separate lesions were preferred over two adjacent lesions). The clinician was blinded to the type of lesion. Semi-automatic results were always reviewed and manually corrected when necessary, to assure the result resembled manual segmentation.

4.4 Ethics statement

The study protocol for the collection of the WORC database conformed to the ethical guidelines of the 1975 Declaration of Helsinki. Approval by the local institutional review board of the Erasmus MC (Rotterdam, the Netherlands) was obtained for collection of the WORC database (MEC-2020-0961), and separately for the six included studies (Lipo: MEC-2016-339, Desmoid: MEC-2016-339, Liver: MEC-2017-1035, GIST: MEC-2017-1187, CRLM: MEC-2017-479, Melanoma: MEC-2019-0693). The need for informed consent was waived due to the use of anonymized, retrospective data.

4.5 Acknowledgments

The authors thank Laurens Groenendijk for his assistance in processing the data and in the anonymization procedures. Martijn P. A. Starmans acknowledges funding from the research program STRaTeGy with project numbers 14929 and 14930, which

is (partly) financed by the Netherlands Organization for Scientific Research (NWO). Part of this study was financed by the Stichting Coolsingel (reference number 567), a Dutch non-profit foundation. This study is supported by EuCanImage (European Union's Horizon 2020 research and innovation programme under grant agreement Nr. 952103).

4.6 CRediT author statement

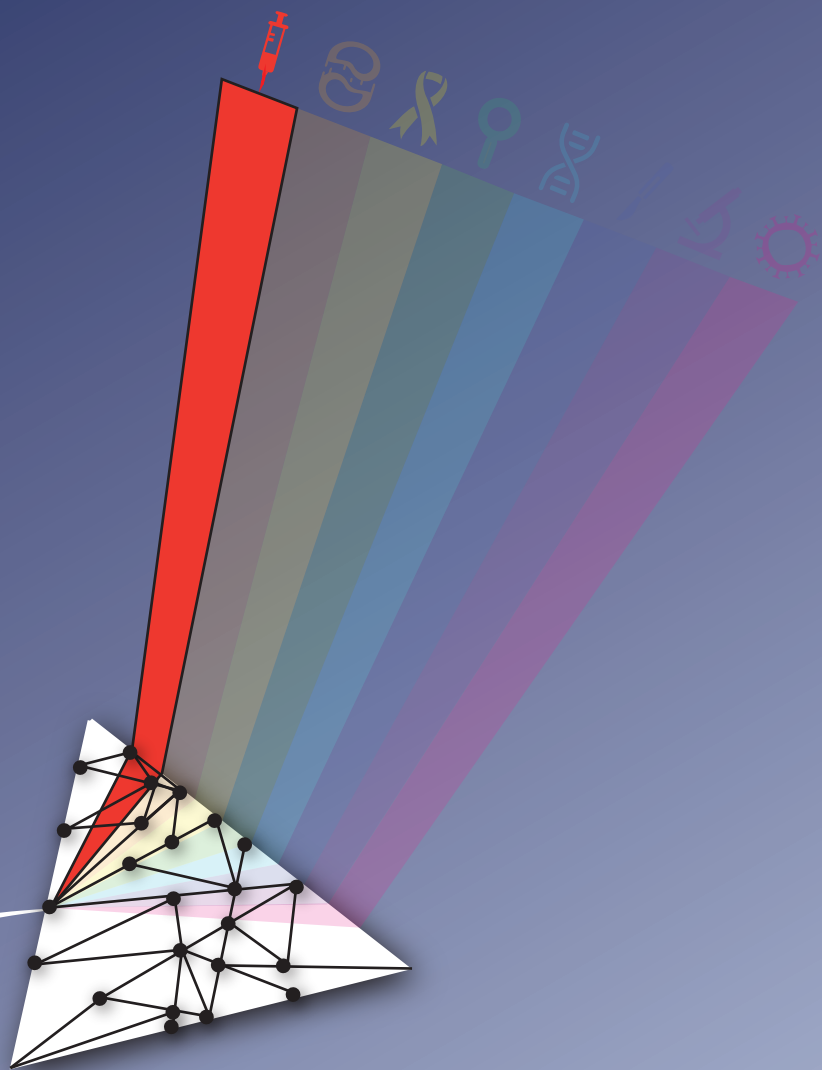
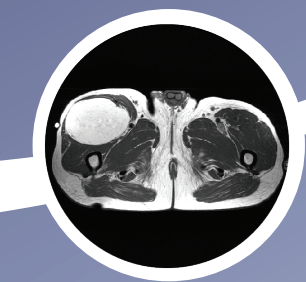
M.P.A.S., M.J.M.T., M.V., D.J.G., C.V., S.S., F.E.B., L.A., A.A.M.v.d.V., R.L.M., M.G., J.J.V., W.J.N., and S.K. provided the conception and design of the study. M.P.A.S., M.J.M.T., M.V., G.A.P., D.J.G., C.V., S.S., G.J.L.H.v.L., F.E.B., F.E.J.A.W., B.G.K., L.A., A.A.M.v.d.V., A.R., A.E.O., M.R., M.D., R.d.M., J.IJ., R.L.M., P.B.V., M.G.T., and J.J.V. acquired the data. M.P.A.S., M.J.M.T., M.V., F.E.B., L.A., R.L.M., M.G.T. and S.K. analyzed and interpreted the data. M.P.A.S. created the software. M.P.A.S. and S.K. drafted the article. All authors read and approved the final manuscript.

4.7 Declaration of competing interest

Wiro J. Niessen is founder, scientific lead, and shareholder of Quantib BV. Jacob J. Visser is a medical advisor at Contextflow. Astrid A. M. van der Veldt is a consultant (fees paid to the institute) at BMS, Merck, MSD, Sanofi, Eisai, Pfizer, Roche, Novartis, Pierre Fabre and Ipsen. The other authors do not declare any conflicts of interest.

Part II

Radiomics biomarkers in clinical applications



5.

Radiomics approach to distinguish between well differentiated liposarcomas and lipomas on MRI

Based on: M. Vos*, **M. P. A. Starmans***, M. J. M. Timbergen, S. R. van der Voort, G. A. Padmos, W. Kessels, W. J. Niessen, G. J. L. H. van Leenders, D. J. Grünhagen, S. Sleijfer, C. Verhoef, S. Klein, and J. J. Visser, "Radiomics approach to distinguish between well differentiated liposarcomas and lipomas on MRI," *British Journal of Surgery*, vol. 106, no. 13, pp. 1800–1809, Dec. 2019. doi: [10.1002/bjs.11410](https://doi.org/10.1002/bjs.11410)

* indicates equal contributions

Abstract

Background: Well differentiated liposarcoma (WDLPS) can be difficult to distinguish from lipoma. Currently, this distinction is made by testing for *MDM2* amplification, which requires a biopsy. The aim of this study was to develop a noninvasive method to predict *MDM2* amplification status using radiomics features derived from MRI.

Methods: Patients with an *MDM2*-negative lipoma or *MDM2*-positive WDLPS and a pretreatment T1-weighted MRI scan who were referred to Erasmus MC between 2009 and 2018 were included. When available, other MRI sequences were included in the radiomics analysis. Features describing intensity, shape and texture were extracted from the tumour region. Classification was performed using various machine learning approaches. Evaluation was performed through a 100 times random-split cross-validation. The performance of the models was compared with the performance of three expert radiologists.

Results: The data set included 116 tumours (58 patients with lipoma, 58 with WDLPS) and originated from 41 different MRI scanners, resulting in wide heterogeneity in imaging hardware and acquisition protocols. The radiomics model based on T1 imaging features alone resulted in a mean area under the curve (AUC) of 0.83, sensitivity of 0.68 and specificity of 0.84. Adding the T2-weighted imaging features in an explorative analysis improved the model to a mean AUC of 0.89, sensitivity of 0.74 and specificity of 0.88. The three radiologists scored an AUC of 0.74 and 0.72 and 0.61 respectively; a sensitivity of 0.74, 0.91 and 0.64; and a specificity of 0.55, 0.36 and 0.59.

Conclusion: Radiomics is a promising, non-invasive method for differentiating between WDLPS and lipoma, outperforming the scores of the radiologists. Further optimization and validation is needed before introduction into clinical practice.

5.1 Introduction

Lipomatous tumours are the most commonly observed soft tissue tumours, mostly owing to the high incidence of benign lipomas. Also within the malignant spectrum of soft tissue tumours (soft tissue sarcomas), liposarcoma is among the most frequently observed subtype [111]. Well differentiated liposarcoma (WDLPS) represents the largest subgroup of liposarcomas; these low-grade, locally aggressive tumours are characterized by amplification of the *MDM2* gene [111]. In rare cases, WDLPS can progress into a more aggressive subtype: dedifferentiated liposarcoma (DDLPS), which has a poorer prognosis [111].

Several differences between lipoma and WDLPS on MRI have been described in the literature: size, location, tumour depth and intratumour heterogeneity. However, as there can be considerable overlap between these features, distinguishing between the two tumour types remains difficult, even for trained radiologists [112, 113, 114, 115, 116]. As the differences between lipoma/WDLPS and DDLPS are more obvious, this distinction can accurately be made solely by eye, [115, 117, 118, 119, 120].

An accurate diagnosis is needed to provide patients with the correct treatment and follow-up. Whereas lipomas do not necessarily need to be excised, patients with WDLPS are generally considered candidates for surgery [121]. Currently, the standard way to differentiate lipoma from WDLPS is through a biopsy, which is tested for *MDM2* amplification using fluorescence in situ hybridization (FISH). Amplification of the *MDM2* gene is present in WDLPS, but absent in lipoma [111, 122, 123]. Taking a biopsy is an invasive and painful procedure for the patient, and is associated with risks, depending on tumour location, and potential sampling error.

The field of radiomics is based on the hypothesis that there is a relationship between medical imaging features and the underlying biological information, such as genetic aberrations [15]. Radiomics approaches have already been used in soft tissue sarcomas to predict other outcomes, such as differentiating between benign and malignant soft tissue tumours in general (not specifically lipomatous tumours) [124], between intermediate- and high-grade soft tissue sarcomas [125], and predicting the risk of lung metastases from soft tissue sarcoma of the extremities [126]. Based on these results, it was hypothesized that radiomics might also be able to differentiate WDLPS from lipoma.

The aim of this study was to develop a model that predicts *MDM2* amplification status using a radiomics approach, thereby differentiating WDLPS from lipoma. MRI scans obtained during routine diagnostic evaluation were used. Additionally, the performance of this model was compared with that of three trained radiologists reading the images. Finally, patients with DDLPS were included and classified by the radiologists to confirm that these tumours have distinct imaging features and can be identified without the help of additional models or tests.

5.2 Methods

Patients with a pathologically confirmed diagnosis of lipoma, WDLPS or DDLPS, a known *MDM2* amplification status tested by FISH, and with at least a T1-weighted MRI sequence available before treatment (if applicable) were included. All patients

were either referred to/discussed at, or diagnosed/treated at the Erasmus MC Cancer Institute, Rotterdam, the Netherlands, between December 2009 and August 2018. As a result, some of the MRI scans were made in the referring hospitals. The study was reviewed and approved by the local medical ethics review committee (MEC-2016-339), and performed in accordance with national and international legislation. Need for informed consent was waived owing to the retrospective and anonymized nature of the study.

To explore the potential predictive value of different MRI sequences, several additional sequences were included, when available. Based on their use in clinical practice, the sequences were grouped into: plain T1 (T1); T1 with fat saturation (T1-FS) including T1 inversion recovery (IR) approaches (T1-IR; a combination of Spectral Presaturation with Inversion Recovery (SPIR), Short τ Inversion Recovery (STIR), Spectral Attenuated Inversion Recovery (SPAIR) and Turbo Inversion Recovery Magnitude (TIRM)); T1 with gadolinium contrast (T1-GD); T1 with fat saturation and gadolinium contrast (T1-FS-GD) including T1-IR with GD; T2 imaging (T2) including T2-Fast Field Echo (T2FFE) and T2*; and T2-FS including T2-IR.

5.2.1 Segmentation

The lipoma and WDLPS lesions were segmented semiautomatically on the T1 images to indicate the regions of interest (ROIs) [105]. All images were segmented independently by either a medical masters student or a PhD candidate with an MD degree. Both were blinded to the type of lipomatous tumour. To validate segmentation accuracy, a sample set was verified by a musculoskeletal radiologist, specialized in soft tissue sarcomas. Median tumour size, defined as the maximum diameter in centimetres, and tumour volume, with corresponding i.q.r. values, were extracted from the segmentations. TheDDLPS images were used only for visual classification by the radiologists, and therefore not segmented.

To transfer the segmentations to the other sequences, all sequences were spatially aligned to the T1 sequence using automated image registration (elastix software [127]), thereby compensating for patient movement between scans. Quality assurance was done by visual inspection.

5.2.2 Radiomics feature extraction

Quantitative imaging features related to intensity, shape and texture were extracted from the ROIs using PyRadiomics software [44, 51]. More details can be found in Section 5.A. The shape features quantified were morphological properties such as volume and similarity to a circle. Intensity features were quantified using first-order statistics such as the mean and standard deviation. Texture features quantified more complex properties, such as the presence of heterogeneity and speckle patterns. When a scan type was missing for a patient, the feature values for the missing image type were imputed.

5.2.3 Additional features

Several additional features were selected based on the available literature and clinical relevance, including patient characteristics (age, sex and tumour location (extremity, trunk, head and neck or pelvis)) and manually scored features (tumour depth (superficial or deep), unilobular or multilobular tumour, atypical appearance on T1 image (yes or no)). These are referred to as patient and manually scored features respectively. Tumours were considered superficial when entirely located above the fascia, or as deep-seated when located beneath the fascia, or with invasion of the fascia.

5.2.4 Decision model creation

To create a decision model from the features, the Workflow for Optimal Radiomics Classification (WORC) toolbox [36] was used. A schematic overview of the radiomics methodology is shown in [Figure 5.1](#).

In WORC, decision model creation is divided into several steps. These steps include, for example, selection of features that offer the highest predictive value and machine learning to discover the patterns in these features that distinguish between WDLPS and lipoma. For each of these steps, numerous algorithms have been proposed in the literature. WORC performs an exhaustive search amongst these algorithms, in a fully automated way, and establishes the combination of algorithms that maximizes the prediction accuracy. As the single best solution may be a coincidental finding, the 50 best performing solutions were combined into a single model, with the purpose of creating a more robust model and boosting performance. More details can be found in [Section 5.B](#).

5.2.5 Experimental set-up

To assess the predictive value of the T1 imaging features, and the additional patient and manually scored features, five models were trained and tested based on: imaging features only (model 1); patient features only (model 2); manually scored features only (model 3); a combination of imaging features and manually scored features (model 4); and volume only (model 5). The fifth model was included because WDLPS is generally larger than lipoma [113]. Additionally, to investigate the potential of the features independent of volume, these five models were evaluated on a volume-matched cohort, that is a subset of the data in which the distribution of tumour volume was similar among WDLPS and lipoma. These models were trained on the full data set, but tested only on patients from the volume-matched cohort.

Next, the potential value of other MRI sequences was explored by training and testing multiple imaging-based radiomics models using combinations of the various MRI sequences. When a model showed more potential than the T1 imaging-only model, it was evaluated on the volume-matched cohort as well.

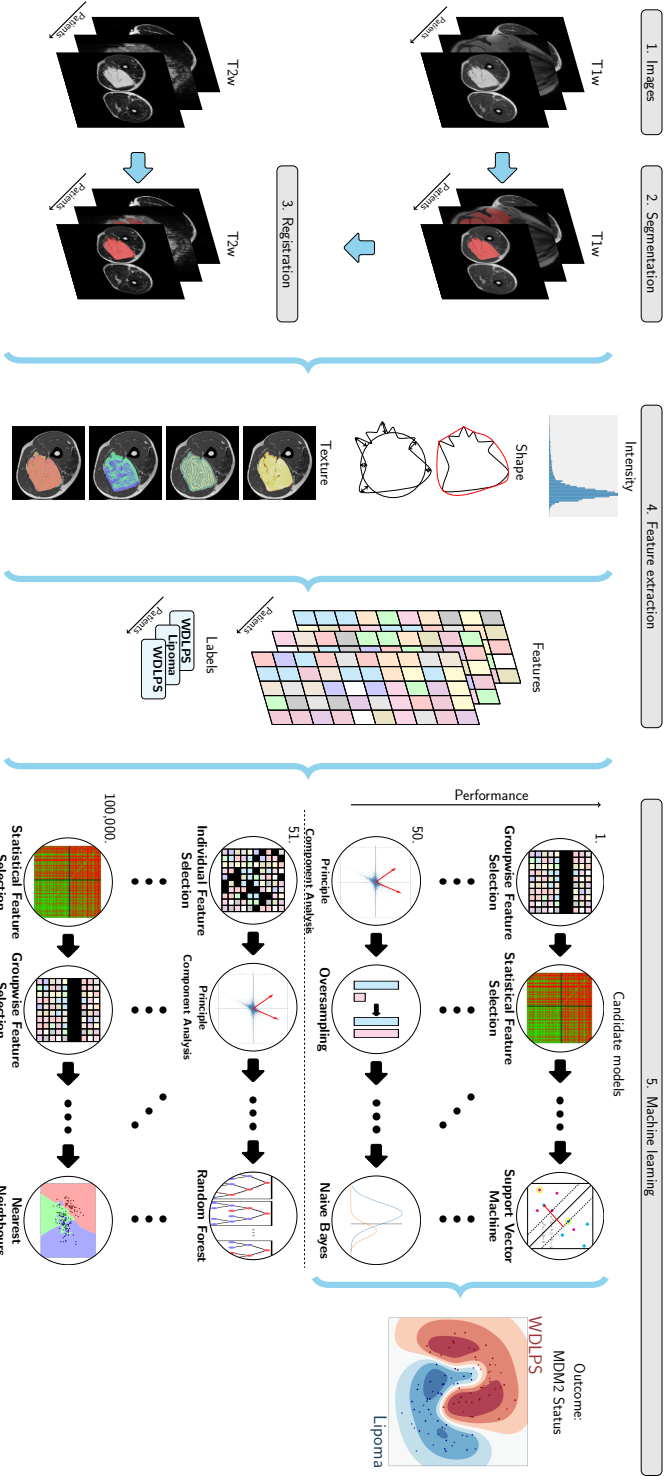


Figure 5.1: Schematic overview of the radiomics approach. Inputs to the algorithm are T1- and T2-weighted magnetic resonance images of well differentiated liposarcoma (WDLPS) and lipoma (1). Processing steps include segmentation of the tumour on the T1 image (2), registration of the T1 to the T2 image to transform this segmentation to the T2 image (3), feature extraction from both the T1 and T2 images (4) and the creation of a decision model from the features (5), using an ensemble of the best 50 workflows from 100 000 candidate workflows, workflows are different combinations of the different processing and analysis steps (for example the classifier used).

5.2.6 Evaluation

Model evaluation was performed through cross-validation. The data were randomly split for 100 iterations, using 80 per cent for training and 20 per cent for testing. In each iteration, automatic workflow optimization was performed on the training set in an internal ten times random split cross-validation (Figure 5.A.1). Thus, the models were optimized solely on the training set; the test set was used only for evaluation of the final model. All splitting was done in a stratified manner to keep the balance between WDLPS and lipoma similar in all data sets.

Performance was evaluated using the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, accuracy, sensitivity, specificity, negative predictive value and positive predictive value, averaged over the 100 cross-validation iterations. Positive *MDM2* amplification status (WDLPS) was defined as the positive class. Ninety-five per cent confidence intervals for the mean performance measures were constructed using the corrected resampled t test based on all 100 cross-validation iterations, thereby taking into account that the samples in the cross-validation splits were not statistically independent [64].

5.2.7 Model insights

Insight into the model was gained by ranking the patients from typical to atypical for both lipoma and WDLPS, based on the consistency of the model predictions. This was determined by the number of times (percentage) that a patient was classified correctly when included in the test set. Typical examples were patients who were always classified correctly; and atypical vice versa. In addition, to identify the individual imaging features included in the radiomics model and to assess their respective contribution to the model, univariable statistical testing of the imaging features was undertaken using the Mann–Whitney U test. P values were corrected for multiple testing using the Bonferroni correction.

5.2.8 Classification by radiologists

Three radiologists with expertise in soft tissue tumours classified the lipomatous tumours; radiologists 1, 2 and 3 had 3, 10 and 5 years of experience respectively. First, the radiologists had to classify the tumours as either DDLPS or WDLPS/lipoma (non-DDLPS), to confirm that DDLPS can be recognized visually. Regardless of whether a tumour was classified as DDLPS or not, the tumours subsequently had to be classified as *MDM2*-negative (lipoma) or *MDM2*-positive (WDLPS/DDLPS). The classification was done using a ten-point scale to indicate the certainty of the radiologists. The radiologists had access to all sequences that were available for each patient, as well as the age and sex.

5.3 Results

In total, 138 tumours were included: 58 patients had an *MDM2*-negative lipoma, 58 had an *MDM2*-positive WDLPS and 22 had an *MDM2*-positive DDLPS. Most

patients were men (60.1 per cent) and had a deep-seated tumour located in a leg. Median WDLPS size was 20.4 cm and median volume was 36.3 cl, compared with 12.3 cm and 12.9 cl for lipoma (Table 5.1).

Most of the patients underwent surgery: 32 with a lipoma, 50 with a WDLPS and 19 of those with a DDLPS. The eight patients with a WDLPS who did not have surgery were treated conservatively with an active surveillance approach, whereas the three with a DDLPS who did not have surgery had an inoperable tumour.

The 116 lipoma and WDLPS scans came from 41 different MRI scanners; there was wide heterogeneity in imaging hardware and acquisition protocols used, reflected in differences in magnetic field strength (1.5T, 98 scans; 1T, 10 scans; 3T, 8 scans), manufacturer (Siemens, Munich, Germany, 45 scans; Philips, Amsterdam, the Netherlands, 45 scans; GE, Chicago, Illinois, USA, 26 scans), scanner model (19 different ones), slice thickness, repetition time and echo time. Additional sequences besides T1 were available in subsets of patients: T1-FS in 55 patients (47.4 per cent), T1-GD in 42 patients (36.2 per cent), T1-FS-GD in 80 patients (69.0 per cent), T2 in 76 patients (65.5 per cent) and T2-FS in 92 patients (79.3 per cent) (Table 5.A.1).

5.3.1 Evaluation of radiomics models based on T1 imaging and additional features

The performances of models 1–5 are shown in Figure 5.2 and Table 5.A.2. Model 1, based on the T1 imaging features, resulted in an AUC of 0.83, sensitivity of 0.68 and specificity of 0.84. Model 2, based on patient features, had a lower AUC (0.75), higher sensitivity (0.77), but lower specificity (0.59). Similarly, model 3, based on manually scored features, also had a lower AUC (0.72), higher sensitivity (0.76) and lower specificity (0.57). Model 4, combining the imaging and manually scored features, performed worse than model 1, implying that imaging features are sufficient as input. Finally, model 5, based on volume alone, performed similarly to model 1 with an AUC of 0.83, sensitivity of 0.67 and specificity of 0.84. Although the performance metrics were similar for models 1 and 5, the ROC curves in Figure 5.2 show some differences. The ROC curve for the volume model (Figure 5.2e) has some sharp bends, while that for the T1 imaging model is smoother (Figure 5.2a).

5.3.2 Evaluation of the radiomics models with additional MRI sequences

Most models with an additional MRI sequence had a similar performance to the T1 imaging model (Table 5.A.3). However, the model combining the T1 and T2 imaging features showed a clear improvement in performance, with an AUC of 0.89, sensitivity of 0.74 and specificity of 0.88. The distribution of patient characteristics and the distribution of WDLPS and lipoma were similar across patients who had a T2 scan, indicating that the added value is within the T2 imaging features and not a result of incidental correlation with these characteristics, for example owing to selection bias.

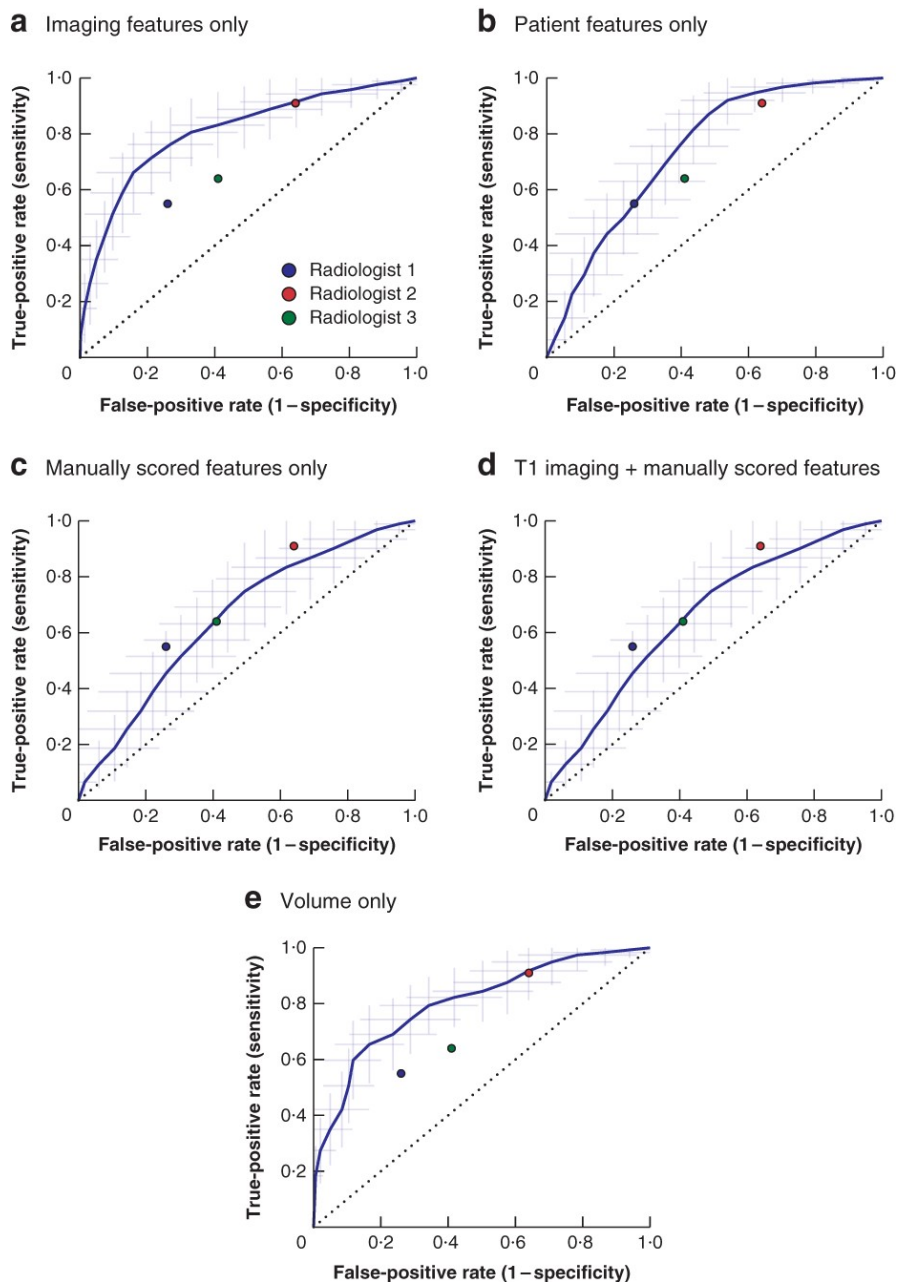


Figure 5.2: Receiver operating characteristic (ROC) curves for the radiomics models based on the T1-weighted MRI sequence. **a** Using imaging features only, **b** using patient features only, **c** using manually scored features only, **d** using T1 imaging features combined with manually scored features, and **e** using volume only. The shaded area indicates the 95 per cent confidence intervals of the 100 times random-split cross-validation; the curve is fit through their means. The performance of the three radiologists is shown.

Table 5.1: Characteristics of the patients with lipomatous tumours. *With percentages in parentheses unless indicated otherwise; †values are median (i.q.r.). WDLPS, well differentiated liposarcoma; DDLPS, dedifferentiated liposarcoma.

	No. of patients* (n = 138)
Age (years)†	64 (54–71)
Sex ratio (M: F)	83: 55
Diagnosis	
Lipoma	58 (42.0)
WDLPS	58 (42.0)
DDLPS	22 (15.9)
Tumour location	
Upper extremity	14 (10.1)
Lower extremity	71 (51.4)
Trunk	37 (26.8)
Head and neck	6 (4.3)
Retroperitoneum and pelvis	6 (4.3)
Paratesticular	4 (2.9)
Tumour depth	
Superficial	20 (14.5)
Deep	118 (85.5)
Tumour size (cm)†	
Lipoma	12.3 (9.3–15.5)
WDLPS	20.4 (15.9–26.3)
Tumour volume (cl)†	
Lipoma	12.9 (4.6–25.0)
WDLPS	36.3 (22.9–85.5)

*With percentages in parentheses unless indicated otherwise; †values are median (i.q.r.). WDLPS, well differentiated liposarcoma; DDLPS, dedifferentiated liposarcoma.

5.3.3 Evaluation of models on volume-matched cohort

Model 5, based on volume alone, illustrated that volume is indeed a strong predictive factor. The 17 tumours with a volume above 70 cl were all WDLPS, whereas the 21 tumours with a volume below 7 cl were all lipoma. In the volume-matched cohort, consisting of the other 78 tumours with a volume between 7 and 70 cl, the volume distributions for WDLPS and lipoma were more similar. As only the T2 scans provided additional value over the T1 imaging features, the T1+T2 imaging model was evaluated for the volume-matched cohort as well.

The performance of both imaging-based models (T1 and T1+T2) was worse on the volume-matched cohort (T1: AUC 0.69; T1+T2: AUC 0.81) (Table 5.2) than on the entire cohort (AUC 0.83 and 0.89 respectively) (Table 5.A.3). The models based on the patient and manually scored features performed similarly to the models tested on the full cohort. The model based on volume alone still performed above

chance (mean AUC 0.64), but considerably worse than on the entire data set. In this volume-matched data set, both the T1 imaging model (AUC 0.69, sensitivity 0.60, specificity 0.74) and the T1+T2 imaging model (AUC 0.81, sensitivity 0.66, specificity 0.84) performed considerably better than volume alone (Table 5.2). This showed that these models were not based solely on volume, and that other features provided additional predictive value over volume.

5.3.4 Model insights

Of the 116 lipomatous tumours, 69 (26 WDLPS, 43 lipoma) were always classified correctly by model 1 in all 100 cross-validation iterations. In contrast, 13 tumours (9 WDLPS, 4 lipoma) were always classified incorrectly. Figure 5.3 shows four MRI slices of such typical and atypical examples of lipoma and WDLPS. The lesions that were always classified incorrectly were checked for possible sampling error of the biopsy. The *MDM2* amplification status of eight of the 13 tumours always classified incorrectly was already determined on the resection specimen (6 WDLPS, 2 lipoma). For the other five patients, in whom it was tested on the biopsy (3 WDLPS, 2 lipoma), pathological examination of the resection specimen confirmed the diagnosis, except for one patient with a lipoma who did not undergo surgery. In the other patient with a lipoma, the resection specimen again tested negative for *MDM2* amplification. The three WDLPS resection specimens were not retested.

Analysis of feature importance was done for the volume-matched cohort, as the results on the full data set were dominated by volume-related measures. In total, 16 individual features were found to be significant after Bonferroni correction on the volume-matched cohort (Figure 5.A.2, supporting information). These included 11 shape features (including several volume-related statistics), four texture features and one intensity feature.

5.3.5 Radiomics models compared with radiologists

On the entire cohort, the AUCs of all three radiologists (0.74, 0.72 and 0.61 for radiologist 1, 2 and 3 respectively) (Table 5.A.4) were below the lower limit of the 95 per cent c.i. of the T1 imaging model (0.75 to 0.90) (Figure 5.2 and Table 5.A.2), as well as of the 95 per cent c.i. of the T1+T2 imaging model (0.83 to 0.95) (Table 5.A.3).

Table 5.2: Performance of radiomics models trained on the full cohort, but evaluated in the volume-matched cohort.

	T1 imaging features	T1 + T2 imaging features	Patient features	Manually scored features	Volume
AUC	0.69 [0.58, 0.80]	0.81 [0.72, 0.90]	0.74 [0.64, 0.84]	0.67 [0.56, 0.77]	0.64 [0.53, 0.74]
Accuracy	0.67 [0.57, 0.76]	0.75 [0.66, 0.83]	0.66 [0.56, 0.75]	0.60 [0.51, 0.69]	0.66 [0.57, 0.74]
Sensitivity	0.60 [0.45, 0.75]	0.66 [0.52, 0.79]	0.69 [0.55, 0.83]	0.70 [0.53, 0.87]	0.50 [0.36, 0.64]
Specificity	0.74 [0.60, 0.87]	0.84 [0.71, 0.96]	0.62 [0.48, 0.76]	0.51 [0.36, 0.65]	0.82 [0.71, 0.92]
NPV	0.66 [0.54, 0.77]	0.72 [0.60, 0.83]	0.68 [0.56, 0.79]	0.65 [0.49, 0.80]	0.62 [0.53, 0.71]
PPV	0.72 [0.58, 0.85]	0.81 [0.69, 0.93]	0.65 [0.54, 0.76]	0.59 [0.49, 0.69]	0.74 [0.61, 0.87]

Values are mean (95 per cent c.i.) over the cross-validation iterations. AUC: area under the curve; NPV: negative predictive value; PPV: positive predictive value.

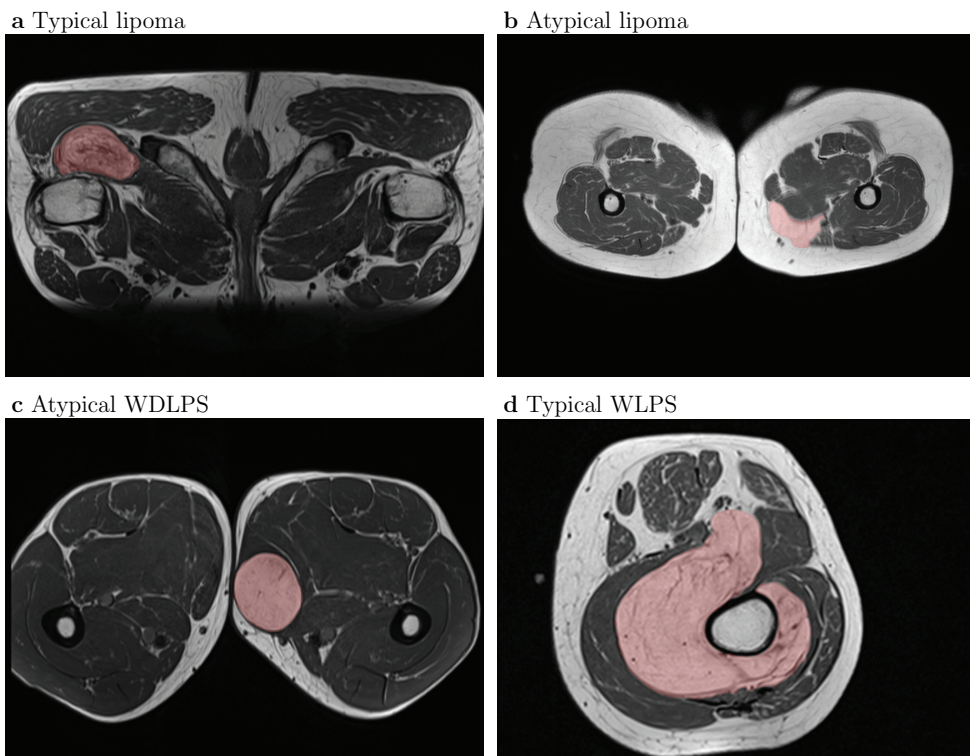


Figure 5.3: Examples of typical and atypical lipomas and well differentiated liposarcomas. **a** Typical lipoma, **b** atypical lipoma, **c** atypical well differentiated liposarcoma (WDLPS) and **d** typical WDLPS. The typical examples are from two patients always classified correctly by the T1 imaging model; the atypical examples are from two patients always classified incorrectly by the T1 imaging model.

The radiologists achieved sensitivity values similar to (0.64 and 0.74) or higher (0.91) than those of the radiomics models (T1: 0.68; T1+T2: 0.74), but their specificity was much lower (radiomics: 0.84 and 0.88 respectively; radiologists 1–3: 0.55, 0.36 and 0.59 respectively). The Cohen's κ value was 0.24, 0.04 and 0.40 for all pairs of radiologists, with a mean of 0.23, indicating poor interobserver agreement.

On the volume-matched cohort, the radiologists had a performance (AUC 0.68, 0.74 and 0.55) (Table 5.A.4) more similar to that of the T1 imaging model (AUC 0.69) (Table 5.2). On average, the T1 imaging model still performed better in terms of specificity (radiomics: 0.74; radiologists 1–3: 0.58, 0.37 and 0.50), whereas the radiologists again performed better on sensitivity (radiomics: 0.60; radiologists 1–3: 0.65, 0.88 and 0.60). However, the T1+T2 imaging model performed much better (AUC 0.81, sensitivity 0.66, specificity 0.84) than both the T1 imaging model and the radiologists. On this cohort, the Cohen's κ values were 0.18, -0.04 and 0.34 for all pairs of radiologists, with a mean of 0.16, again indicating poor interobserver agreement.

5.3.6 Distinction between dedifferentiated liposarcoma and well differentiated liposarcoma/lipoma

Besides classifying lipoma and WDLPS, the radiologists also classified the scans from 22 patients with DDLPS to evaluate whether DDLPS can indeed be identified by imaging only, without the help of additional models. Radiologists 1–3 had an AUC of 0.97, 0.91 and 0.90 respectively; a sensitivity of 0.95, 0.95 and 0.91; and a specificity of 0.95, 0.56 and 0.89 in distinguishing DDLPS from non-DDLPS (WDLPS/lipoma) (Table 5.A.4).

5.4 Discussion

This study shows that there is a relationship between quantitative MRI features and *MDM2* amplification status, and that radiomics is a promising non-invasive method for differentiating lipoma from WDLPS. Although the radiologists were able to distinguish between DDLPS and non-DDLPS, they were outperformed by the T1 and T1+T2 imaging models in differentiating WDLPS from lipoma. Moreover, the agreement between radiologists was very poor, whereas the radiomics-based predictions were objective and reproducible (given a tumour segmentation).

Remarkably, the model trained on volume alone had a similar performance to the T1 imaging model, which included many additional features. However, in the volume-matched data set, the T1 imaging model performed considerably better than the volume-only model, indicating that other features do provide additional predictive value. It is already known that WDLPS is on average larger than lipoma [113], and the relationship with volume (or size) in our data set was also strong; the database did not contain lipoma larger than 70 cl or WDLPS smaller than 7 cl although these do exist [128, 129]. However, all WDLPS lesions start as small tumours and grow over time, so the measured tumour volume depends on the moment of presentation, and a small or intermediate tumour volume is therefore not a reliable biomarker. Future research should include expansion of the data set to make the volume distributions more representative (including lipoma larger than 70 cl and WDLPS smaller than 7 cl), thereby making the radiomics model less volume-dependent.

The models trained solely on either the patient or manually scored features performed slightly worse than the model trained on the T1 imaging features only. As the combined model did not outperform the T1 imaging model, the manually scored features did not add much in the search for the best radiomics model. Additionally, the manually scored features may be observer-dependent, and thus prone to subjectivity. Although patient features (age, sex and tumour location) are objective, the distribution in the present data set may not be representative of clinical practice. For example, none of the patients with WDLPS were younger than 35 years, there were no lipomas among patients older than 82 years, no lipomas in the head and neck region, and no WDLPS in the pelvis or shoulder/trunk; all these might occur in daily clinical practice. Therefore, the imaging-only models have more potential as an objective tool in clinical practice.

The results of present study are similar to those of Thornhill *et al.* [130], who used a comparable approach and showed that lipomas can be distinguished from liposarcomas by texture and shape analysis. Strong points of the present study include the larger sample size (116 versus 44 in Thornhill *et al.* [130]). Thornhill *et al.* [130] also included other liposarcoma subtypes in their model, such asDDLPS and myxoid liposarcoma (8 of 20 included liposarcomas). These other liposarcoma subtypes have distinct radiological features [115, 119], which in general can be easily discriminated from lipomas by experienced radiologists. By solely including the two tumour types that are the most difficult to distinguish (WDLPS and lipoma) in the radiomics model, the present data set is more challenging and more clinically relevant. In contrast to the cases described by Thornhill *et al.* [130], the diagnosis of all patients in the present data set was confirmed by verifying the *MDM2* amplification status using FISH, the current standard for diagnosing and differentiating between lipoma and WDLPS [111, 122, 123]. The present radiomics model only requires routine MRI scans (T1, and optionally T2) without contrast injection; the other sequences did not add any predictive value to the model. As almost all standard MRI protocols include a T1 and T2 sequence, the present radiomics method is generalizable, feasible and applicable for use in daily practice. Finally, these radiomics models were developed and evaluated on a heterogeneous data set, thereby increasing the chance that the reported performance can be reproduced in a routine clinical setting when using other MRI scanners.

Advantages of using a radiomics approach over pathological assessment to differentiate between lipoma and WDLPS include sparing patients an invasive and painful biopsy, and saving the substantial costs of a radiologist performing the imaging-guided biopsy and of the pathologist assessing it, including the costs of molecular testing by FISH. Radiomics makes use of MRI images obtained during routine diagnostic evaluation and patients do not need to become a widely available tool, patients with WDLPS can be identified and referred to a soft tissue sarcoma expert centre at an earlier stage, with potential beneficial effects on further diagnostics, treatment and follow-up.

Several limitations of this study should be noted, besides the volume bias already mentioned. First, segmentation of ROIs of the tumours was done manually, which inherently leads to both interobserver and intraobserver variability, as has been quantified for other cancer types [131, 132, 133]. Variability in segmenting the ROIs might lead to variability in the extracted imaging features and subsequently influence the classification of tumours. Additionally, manual segmentation is rather time-consuming. This could be addressed by use of automated segmentation tools that might be available in the future. Second, variation in imaging protocols might have influenced the imaging statistics. No restrictions were put on the T1 MRI sequences regarding field strength, slice thickness, or other MRI acquisition settings, as selecting a single protocol is an unrealistic reflection of daily clinical practice and would have made the results non-generalizable. Instead, this study shows that the present radiomics approach is robust to these variations by both training and testing the model on heterogeneous data. Third, the model is based on retrospectively collected data, which might have led to selection and information bias. This potential selection bias might have occurred particularly in the lipoma subgroup, as usually

only large and atypical lipomas are referred to a sarcoma centre. However, this probably made the data set even more challenging and relevant, as these can be seen as the complex cases. Addition of the ‘small and typical’ lipomas would have made the classification easier, and radiomics is not needed to make the distinction for such lipomas.

The present radiomics model could serve as a non-invasive, quick and low-cost alternative to a biopsy. Although the model needs optimization to match the accuracy of a biopsy, there could be a certain patient group for whom the model may already be useful. For example, patients at high risk of complications of biopsy, or those in whom the radiomics model can predict the *MDM2* amplification status with a high degree of certainty, could already be treated according to the prediction of the radiomics model. Although further research is required to identify which patients could benefit most from the present model, initial misclassification of a WDLPS as a lipoma would not harm the patient, considering that active surveillance seems a safe option in patients without (invalidating) symptoms and/or tumour growth, at least in the short term³⁰. In addition, the performance of the radiomics model improved substantially when T2 images were added. However, only 65.5 per cent of the patients had a T2 scan available, so for a follow-up study it is proposed to use MRI with at least both T1 and T2 sequences.

5.4.1 Acknowledgements

M.V. and M.P.A.S. contributed equally to this study. The authors thank E. H. G. Oei and D. F. Hanff for classifying the lipomatous tumours. This study was financed by the Stichting Coolensingel (reference no. 567), a Dutch non-profit foundation. M.P.A.S. acknowledges funding from the research programme STRaTeGy (project no. 14929-14930), which is partly financed by the Netherlands Organization for Scientific Research (NWO). W.J.N. is founder, scientific lead and stock holder of Quantib. Disclosure: The authors declare no other conflicts of interest.

Appendix

Appendix 5.A Radiomics feature extraction

In this study, radiomics features quantifying intensity, shape and texture were extracted. Intensity features were extracted using the histogram of all intensity values within the Regions of Interest (ROIs) and included several first order statistics such as the mean, standard deviation and kurtosis. Shape features were extracted by solely using the ROI and included shape descriptions such as the compactness, roundness and circular variance. Additionally, the volume and orientation of the ROI were used. Texture features were extracted using the Gray Level Co-occurrence Matrix, Gray Level Size Zone Matrix Gray Level Run Length Matrix and Neighborhood Grey Tone Difference Matrix. All features were extracted using the defaults for MR images from PyRadiomics.

The used dataset is highly heterogeneous in terms of acquisition protocols. Especially the variations in slice thickness and contrast may cause feature values to

be highly dependent on the acquisition protocol. The slice thickness varies between 2.5mm and 10mm. Hence, extracting robust 3D features may be hampered by these variations, especially for the low resolutions. To overcome this issue, all features are extracted per 2D axial slice and aggregated over all slices. Due to the slice thickness and pixel spacing heterogeneity, the images were not resampled. Due to variations in especially the magnetic field strength, echo time, and repetition time, the image contrast highly varies, which will affect the feature values. To overcome this, each 3D MRI is normalized using z-scoring before feature extraction.

The code to extract the features has been published open-source [134].

Appendix 5.B Technical details on decision model creation

The Workflow for Optimal Radiomics Classification (WORC) toolbox[36] makes us of adaptive algorithm optimization to create the optimal performing workflow from a variety of methods. We define a workflow as a sequential combination of algorithms and their respective parameters.

WORC includes algorithms to perform feature imputation, feature selection, feature scaling, oversampling, and machine learning. Feature selection was performed to eliminate features which are not useful to distinguish between WDLPS and lipoma. These included; 1) a group-wise search, in which specific groups of features (i.e. intensity, shape, and the several subgroups of texture features as defined in Supplementary Materials 1) are selected or deleted; 2) a variance threshold, in which features with a low variance are removed; and 3) principal component analysis (PCA), in which only those linear combinations of features were kept which explained a large part of the variance in the features.

Feature scaling was performed to make all features have the same scale, as otherwise the machine learning methods may focus only on those features with large values. This was done through z-scoring, i.e. subtracting the mean value followed by division by the standard deviation. In this way, all features had a mean of zero and a variance of one.

Oversampling was used to make sure the classes (i.e. WDLPS and lipoma) were balanced in the training dataset. These include 1) random oversampling, which randomly repeats patients of the minority class; and 2) SMOTE [58], which creates new synthetic patients using a combination of the patients in the minority class.

Lastly, machine learning methods were used to determine a decision rule to distinguish between WDLPS and lipoma. These included 1) logistic regression; 2) support vector machines; 3) random forests; 4) naive Bayes; and 5) linear and quadratic discriminant analysis.

Most of the included methods require specific settings or parameters to be set, which may have a large impact on the performance. As these parameters have to be determined before executing the workflow, these are so-called “hyperparameters”. In WORC, we treat all parameters of all methods as hyperparameters, since they may all influence the decision model creation. Hence, we simultaneously determine which combination of algorithms and hyperparameters performs best.

In the training phase, a total of 100,000 pseudo-randomly generated workflows is created and executed. The workflows are ranked from best to worst based on the

F1-score, which is the harmonic average of precision and recall. Due to the large number of workflows executed, there is a chance that the best performing workflow is overfitting, i.e. looking at too much detail or even noise in the training dataset. Hence, to create a more robust model and boost performance, WORC combines the 50 best performing methods into a single decision model, which is known as ensembling. The ensemble is created through averaging of the probabilities, i.e. the chance of a patient being WDLPS or lipoma, of these 50 workflows.

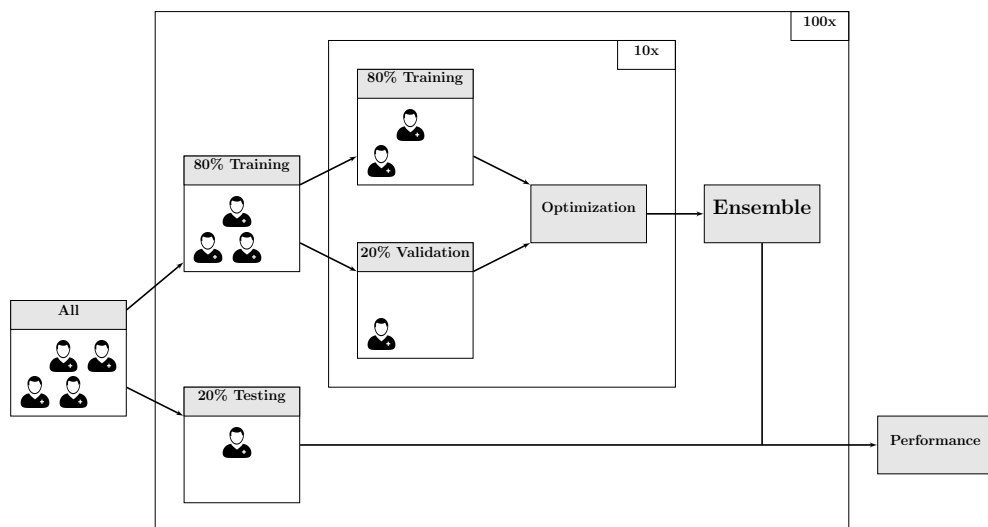


Figure 5.A.1: Visualization of the 100x stratified random-split cross-validation, including a second cross-validation within the training set to perform the automatic workflows optimization. Optimization was done solely on the training set in order to prevent overfitting on the test set. The ensemble averages the predictions of the best 50 performing workflows to create a more robust model.

Table 5.A.1: Several properties of the acquisition protocols of the 116 T1-weighted MRI sequences of patients with lipoma or well-differentiated liposarcoma (WDLPS) that were used to build the radiomics model.

Property	N	%		
Magnetic field strength				
1T	10	9.6		
1.5T	98	84.5		
3T	8	6.9		
Manufacturer				
Siemens	45	38.8		
Philips	45	38.8		
GE	26	22.4		
Setting (Unit)	Mean	Std.	Min.	Max.
Slice Thickness (mm)	4.77	1.14	2.5	10.0
Repetition time (ms)	555	108	280	831
Echo time (ms)	13.2	4.3	5.7	37
Available MRI sequences		N	%	
T1	116	100		
T1-FS	55	47.4		
T1-GD	42	36.2		
T1-FS-GD	80	69.0		
T2	76	65.5		
T2-FS	92	79.3		

Std.: standard deviation, min.: minimum value, max.: maximum value, mm: millimeters, ms: milliseconds, FS: Fat Saturation, GD: gadolinium contrast.

Table 5.A.2: Performance of the radiomics models based on T1 imaging features only; patient features only; manually scored features only; the combination of T1 imaging and manually scored features; and of volume only on the full dataset. Performance for the radiomics models is reported for each experiment as mean [95% confidence interval] over the cross-validation iterations.

	Model 1 T1 imaging features	Model 2 Patient features	Model 3 Manually scored features	Model 4 T1 imaging + manually scored features	Model 5 Volume
AUC	0.83 [0.75, 0.90]	0.75 [0.64, 0.85]	0.72 [0.62, 0.81]	0.69 [0.58, 0.79]	0.83 [0.75, 0.91]
Accuracy	0.68 [0.67, 0.84]	0.68 [0.59, 0.76]	0.67 [0.57, 0.76]	0.61 [0.51, 0.70]	0.76 [0.67, 0.84]
Sensitivity	0.68 [0.53, 0.82]	0.77 [0.63, 0.90]	0.76 [0.58, 0.94]	0.53 [0.37, 0.68]	0.67 [0.52, 0.81]
Specificity	0.84 [0.72, 0.95]	0.59 [0.45, 0.72]	0.57 [0.43, 0.71]	0.69 [0.54, 0.84]	0.84 [0.71, 0.97]
NPV	0.73 [0.63, 0.82]	0.73 [0.61, 0.85]	0.73 [0.59, 0.86]	0.60 [0.50, 0.69]	0.75 [0.66, 0.83]
PPV	0.82 [0.70, 0.93]	0.66 [0.58, 0.73]	0.64 [0.54, 0.74]	0.64 [0.51, 0.76]	0.81 [0.69, 0.93]

AUC: area under the curve, NPV: negative predictive value, PPV: positive predictive value

Table 5.A.3: Performance of radiomics models trained on features extracted from various MRI sequences on the full dataset. Performance is reported as mean [95% confidence interval] over the cross-validation iterations.

	T1	T1 + T1-FS	T1 + T1-GD	T1 + T1-FS-GD	T1 + T2	T1 + T2-FS
AUC	0.83 [0.75, 0.90]	0.84 [0.75, 0.92]	0.81 [0.72, 0.90]	0.81 [0.73, 0.89]	0.89 [0.83, 0.95]	0.81 [0.73, 0.88]
Accuracy	0.68 [0.67, 0.84]	0.77 [0.69, 0.85]	0.76 [0.67, 0.84]	0.75 [0.66, 0.83]	0.81 [0.74, 0.87]	0.74 [0.66, 0.81]
Sensitivity	0.68 [0.53, 0.82]	0.69 [0.56, 0.82]	0.69 [0.56, 0.82]	0.66 [0.51, 0.81]	0.74 [0.61, 0.86]	0.66 [0.53, 0.79]
Specificity	0.84 [0.72, 0.95]	0.84 [0.73, 0.95]	0.77 [0.71, 0.83]	0.84 [0.72, 0.95]	0.88 [0.78, 0.98]	0.82 [0.70, 0.93]
NPV	0.73 [0.63, 0.82]	0.74 [0.65, 0.82]	0.73 [0.64, 0.82]	0.72 [0.63, 0.81]	0.78 [0.69, 0.86]	0.72 [0.63, 0.80]
PPV	0.82 [0.70, 0.93]	0.83 [0.72, 0.93]	0.80 [0.69, 0.91]	0.81 [0.69, 0.93]	0.88 [0.78, 0.97]	0.79 [0.68, 0.90]

AUC: area under the curve, NPV: negative predictive value, PPV: positive predictive value, FS: Fat Saturation, GD: gadolinium contrast

Table 5.A.4: Performance of the three radiologists in differentiating between well-differentiated liposarcomas and lipomas on both the full and volume-matched cohort, and in differentiating dedifferentiated liposarcoma (DDLPS) and non-DDLPS (well-differentiated liposarcoma (WDLPS) / lipomas).

	Full cohort			Volume-matched cohort			DDLPS vs. non-DDLPS		
	Rad. 1	Rad. 2	Rad.3	Rad. 1	Rad. 2	Rad. 3	Rad. 1	Rad. 2	Rad. 3
AUC	0.74	0.72	0.61	0.68	0.74	0.55	0.97	0.91	0.90
Accuracy	0.64	0.64	0.61	0.62	0.63	0.55	0.95	0.62	0.89
Sensitivity	0.74	0.91	0.64	0.65	0.88	0.60	0.95	0.95	0.91
Specificity	0.55	0.36	0.59	0.58	0.37	0.50	0.95	0.56	0.89
NPV	0.68	0.81	0.62	0.61	0.74	0.54	0.99	0.98	0.98
PPV	0.62	0.59	0.61	0.62	0.59	0.56	0.78	0.29	0.61

AUC: area under the curve; NPV: negative predictive value; PPV: positive predictive value; Rad.: radiologist

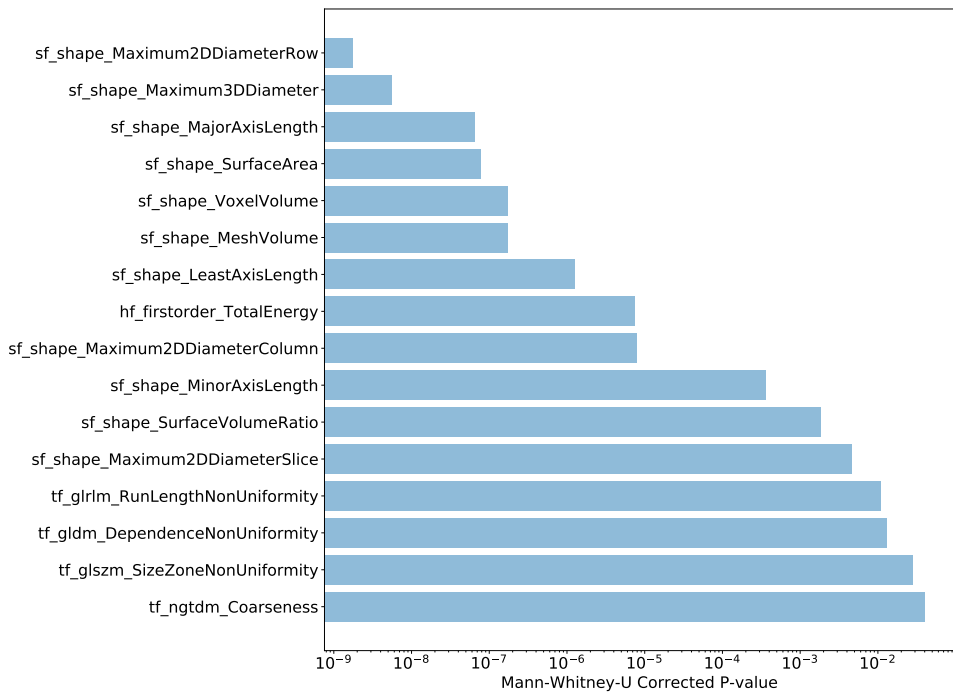
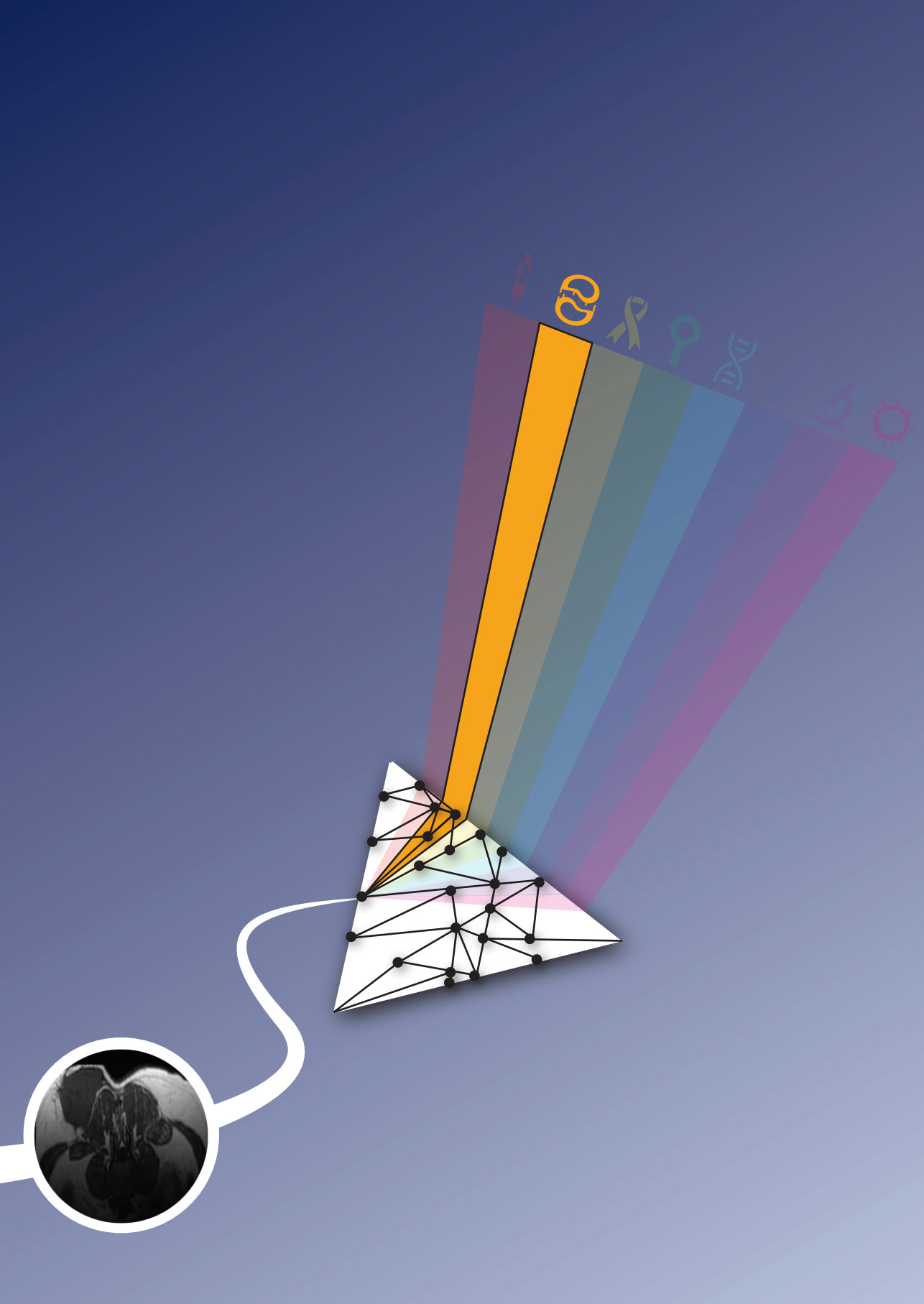


Figure 5.A.2: P-values of Mann-Whitney U tests of feature values for WDLPS and lipomas. Only the features that had a corrected P-value < 0.05 were included in the graph. The labels on the y-axis correspond to the feature names: see [Section 5.A](#) for more details.



6.

Differential diagnosis and mutation stratification of desmoid-type fibromatosis on MRI using radiomics

Based on: M. J. M. Timbergen*, **M. P. A. Starmans***, G. A. Padmos, D. J. Grünhagen, G. J. L. H. van Leenders, D. F. Hanff, C. Verhoef, W. J. Niessen, S. Sleijfer, S. Klein, and J. J. Visser, "Differential diagnosis and mutation stratification of desmoid-type fibromatosis on MRI using radiomics," *European Journal of Radiology*, vol. 131, p. 109266, Oct. 2020. DOI: [10.1016/j.ejrad.2020.109266](https://doi.org/10.1016/j.ejrad.2020.109266)

* indicates equal contributions

Abstract

Purpose: Diagnosing desmoid-type fibromatosis (DTF) requires an invasive tissue biopsy with β -catenin staining and *CTNNB1* mutational analysis, and is challenging due to its rarity. The aim of this study was to evaluate radiomics for distinguishing DTF from soft tissue sarcomas (STS), and in DTF, for predicting the *CTNNB1* mutation types.

Methods: Patients with histologically confirmed extremity STS (non-DTF) or DTF and at least a pretreatment T1-weighted (T1w) MRI scan were retrospectively included. Tumors were semi-automatically annotated on the T1w scans, from which 411 features were extracted. Prediction models were created using a combination of various machine learning approaches. Evaluation was performed through a 100x random-split cross-validation. The model for DTF vs. non-DTF was compared to classification by two radiologists on a location matched subset.

Results: The data included 203 patients (72 DTF, 131 STS). The T1w radiomics model showed a mean AUC of 0.79 on the full dataset. Addition of T2w or T1w post-contrast scans did not improve the performance. On the location matched cohort, the T1w model had a mean AUC of 0.88 while the radiologists had an AUC of 0.80 and 0.88, respectively. For the prediction of the *CTNNB1* mutation types (S45 F, T41A and wild-type), the T1w model showed an AUC of 0.61, 0.56, and 0.74.

Conclusions: Our radiomics model was able to distinguish DTF from STS with high accuracy similar to two radiologists, but was not able to predict the *CTNNB1* mutation status.

6.1 Introduction

Sporadic desmoid-type fibromatosis (DTF) is a rare borderline, soft tissue tumor arising in musculoaponeurotic structures [111]. Worldwide epidemiological data is lacking, but population studies in Scandinavia and the Netherlands show a low incidence of 2.4–5.4 cases per million per year [135, 136]. Early recognition and diagnosis of DTF is therefore challenging.

On MRI, DTF can display a wide variety of enhancement patterns [137]. DTF has imaging characteristics that are often associated with soft tissue sarcomas (STS), such as crossing fascial boundaries, an invasive growth pattern, little central necrosis, mildly hyperintense on T1-weighted (T1w) MRI, and hyperintense and heterogeneous on T2-weighted (T2w) MRI with hypointense bands [138]. Hence, the distinction between DTF and STS, i.e. non-DTF, can be difficult. An invasive tissue biopsy, with additional immunohistochemical staining for β -catenin and mutation analysis of the *CTNNB1* (β -catenin) gene, is therefore currently required to differentiate DTF from non-DTF [139].

As DTF is a borderline tumor who is unable to metastasize, and requires a different treatment regimen than malignant STS, this distinction is highly relevant. Differentiation between DTF and STS based on imaging would be beneficial because of the rarity of DTF, making clinical and pathological recognition challenging. Furthermore, DTF exhibits an aggressive growth pattern and growth might be stimulated after (surgical) trauma, including biopsies [140]. Avoiding (multiple) harmful biopsies which potentially cause tumor growth is therefore of great importance.

Several studies have addressed the prognostic role of the *CTNNB1* mutation in DTF [141, 142, 143], as serine 45 (S45F) tumors appear to have a higher risk of recurrence after surgery compared to threonine 41 (T41A) and wild type (WT) (i.e. no *CTNNB1* mutation [144]) tumors [145]. Obtaining the *CTNNB1* mutation status is for diagnostic purposes and to guide the clinical work-up, but, for now, the *CTNNB1* mutation status has no therapeutic consequences [146]. The majority of DTF harbors a *CTNNB1* mutation at either T41A or S45 F [141]. Assessment of the mutation status is currently done by Sanger Sequencing or Next Generation Sequencing, which are time consuming and expensive.

In radiomics, large amounts of quantitative imaging features are related to clinical outcome [19] (i.e., [Chapter 2](#) of this thesis). Radiomics may serve as a non-invasive surrogate to contribute to diagnosis, prognosis and treatment planning [147, 148]. Based on the results of previous studies in cancer [23], we hypothesized that radiomics may also be useful in DTF.

This study investigated whether a radiomics model based on MRI is able to 1) distinguish DTF from non-DTF in the extremities, and 2) to predict the *CTNNB1* mutation status of DTF. Additionally, in the DTF vs. non-DTF distinction, we evaluated which of the included MRI sequences has the highest predictive value.

6.2 Material and methods

6.2.1 Data collection

Approval by the Erasmus Medical Center (MC) institutional review board (MEC-2016-339) was obtained. Patients diagnosed or referred to the Erasmus MC between 1990-2018 with a histologically proven primary or recurrent DTF were included. This resulted in a multicenter imaging dataset as patients referred to our sarcoma expert institute often received imaging at their referring hospital. The most frequently used imaging modality prior to treatment was T1w-MRI, and its availability was used as an inclusion criterion [23]. When available, other sequences such as T2w, T1w post-contrast, dynamical contrast enhanced (DCE), proton density (PD) and diffusion weighted imaging (DWI) MRI were collected.

For the differential diagnosis (DTF vs. non-DTF), histologically confirmed malignant extremity STS were included. Benign STS were excluded, because this distinction is clinically less relevant. Nonextremity STS were excluded because of the infrequent use of MRI. Although DTF tumors commonly occur in the abdominal wall, their differential diagnosis is broad and includes pseudo-tumors such as myositis, nodular fasciitis and hematomas, and tumors such as lipomas, STS, endometriosis, carcinomas, lymphomas and metastasis [149]. Hence, we decided to focus on the distinction between DTF and STS, and included patients with a histologically proven primary fibromyxosarcoma, myxoid liposarcoma or leiomyosarcoma of the extremities. Similar to the DTF, patients with at least a pre-treatment T1w-MRI were retrospectively included.

Sex, age at diagnosis, and tumor location were collected. For the DTF, in case of a missing *CTNNB1* mutation status, Sanger Sequencing was performed after review of formalin-fixed paraffin-embedded tumor sections by a pathologist. Cases with a known *CTNNB1* mutation did not undergo additional review by a pathologist. Poor scan quality (e.g. artifacts), poor DTF DNA quality with failure of sequencing, and *CTNNB1* mutation other than S45F, T41A or WT led to exclusion.

6.2.2 Radiomics feature extraction

The tumors were all manually segmented once on the T1w-MRI by one of two clinicians under supervision of a musculoskeletal radiologist (4 years of experience). A subset of 30 DTF was segmented by both clinicians, in which intra-observer variability was evaluated through the pairwise Dice Similarity Coefficient (DSC), with DSC > 0.70 indicating good agreement [150]. To transfer the segmentations to the other sequences, all sequences were automatically aligned to the T1w-MRI using image registration with the Elastix software [127]. For each lesion, per MRI sequence, 411 features quantifying intensity, shape and texture were extracted. Details can be found in [Section 6.A](#) and [Table 6.A.2](#).

6.2.3 Decision model creation

To create a decision model from the features, the WORC toolbox was used, see [Figure 6.1](#) [36, 72, 151]. In WORC, the decision model creation consists of several

steps, e.g. feature selection, resampling, and machine learning. WORC performs an automated search amongst a variety of algorithms for each step and determines which combination of algorithms maximizes the prediction performance on the training set. More details can be found in [Section 6.B](#).

For the differential diagnosis cohort, a binary classification model was created using a variety of machine learning models. For the DTF cohort (predicting the *CTNNB1* mutation), a multiclass classification model was created using random forests.

6.2.4 Evaluation

Evaluation of all models was done through a 100x random-split cross-validation. In each iteration, the data was randomly split in 80 % for training and 20 % for testing in a stratified manner, to make sure the distribution of the classes in all sets was similar to the original ([Figure 6.A.1](#)). Within the training set, model optimization was performed using an internal cross-validation (5x). Hence, all optimization was done on the training set to eliminate any risk of overfitting on the test set.

Performance was evaluated using the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve, balanced classification accuracy (BCA), sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV). For the multiclass models, we reported the multiclass AUC [152] and overall BCA [65]. The positive classes included: DTF in the differential diagnosis, and the presence of the mutation in the mutation analysis. The 95 % confidence intervals were constructed using the corrected resampled t-test, thereby taking into account that the samples in the cross-validation splits are not statistically independent [64]. Both the mean and the confidence intervals are reported. ROC confidence bands were constructed using fixed-width bands [67].

To assess the predictive value of the various features, models were trained based on: 1) volume; 2) age and sex; 3) T1w-MRI imaging; 4) T1w-MRI imaging, age and sex. Model 1 was created to verify that the imaging models were not solely based on volume. Model 2 was created to evaluate potential age and gender biases. In model 4, the imaging and clinical characteristics are combined by using both the imaging features and age and sex as features for a total of 413 features. This allows WORC to combine the imaging and clinical characteristics in the most optimal way. Additionally, a model was made for each combination of T1w-MRI and one of the other included MRI sequences (e.g. based on T1w-MRI and T2w-MRI) to evaluate the added value of these other sequences. When a sequence was missing for a patient, feature imputation was used to estimate the missing values.

The code for the feature extraction, model creation and evaluation has been published open-source [153].

6.2.5 Model insight

To explore the predictive value of individual features, the Mann-Whitney U univariate statistical test was used. P-values were corrected for multiple testing using the Bonferroni correction, and were considered statistically significant at a p-value <0.05.

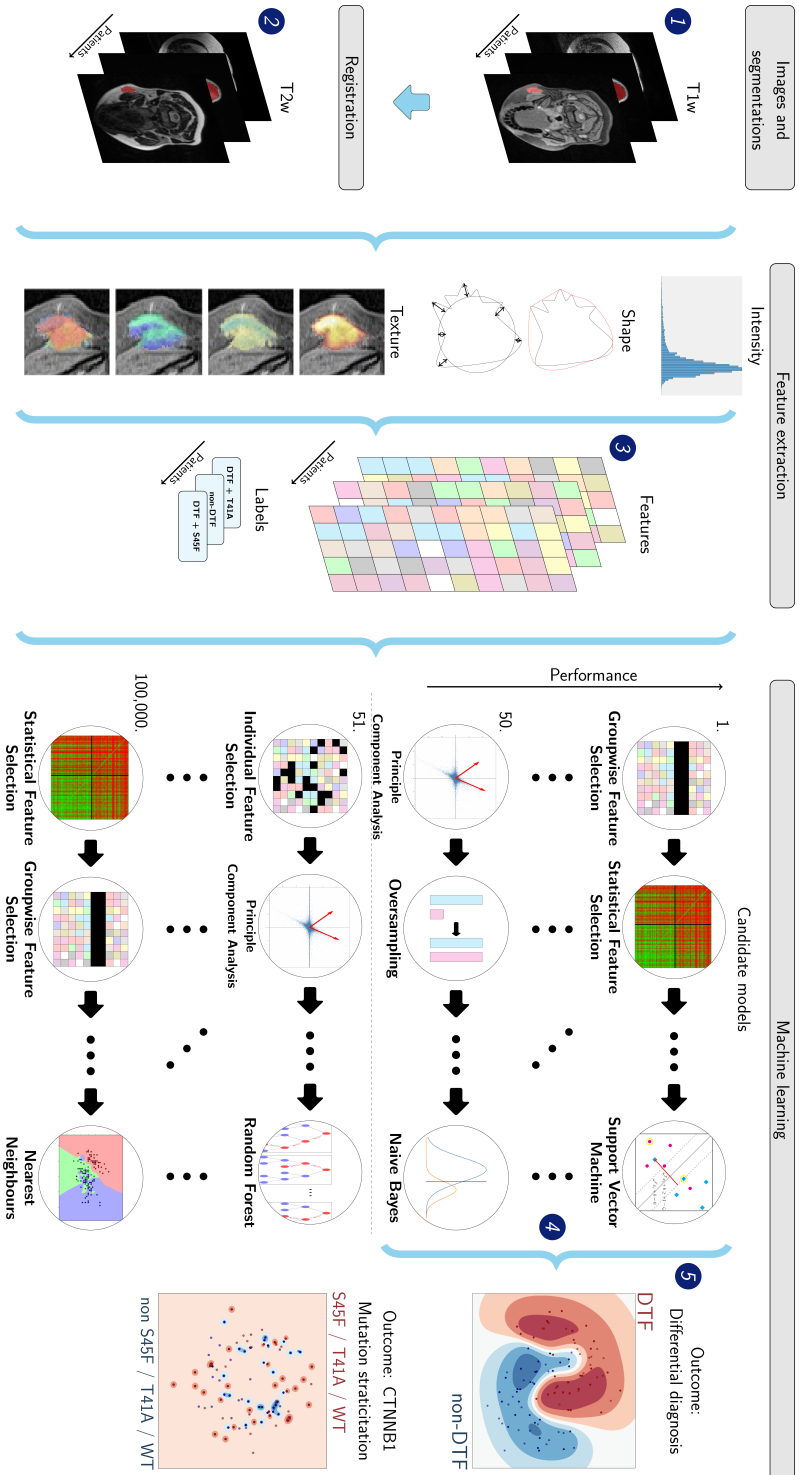


Figure 6.1: Schematic overview of the radiomics approach, adapted from Vos *et al.* [72] (i.e., Chapter 5 of this thesis). Processing steps include segmentation of the tumor on the T1-weighted (T1w) MRI (1), registration of the T1w to the T2-weighted (T2w) MRI to transform this segmentation to the T2w-MRI (2), feature extraction from both the T1w-MRI and the T2w-MRI (3) and the creation of machine learning decision models (5), using an ensemble of the best 50 workflows from 100,000 candidates (4), where the workflows are different combinations of the processing and analysis steps. DTF, desmoid-type fibromatosis.

Feature robustness to variations in the segmentations was assessed on the subset of 30 DTF segmented by two observers using the intra-class correlation coefficient (ICC), where an ICC > 0.75 indicated good reliability [154]. To evaluate model reliability, a separate model was trained using only these features with a good reliability. To gain insight into the models, the patients were ranked based on the consistency of the model predictions. Typical examples for each class consisted of the patients that were correctly classified in all cross-validation iterations; atypical vice versa.

6.2.6 Classification by radiologists

To compare the models with clinical practice, the tumors were classified by two musculoskeletal radiologists (5 and 4 years of experience), which had access to all available MRI sequences, age, and sex. They were specifically instructed to distinguish between STS and DTF. Classification was made on a ten-point scale to indicate the radiologists' certainty. As only extremity STS were selected for the non-DTF group, a location-matched database was used. This included all extremity DTF and the same number of non-DTF. Agreement between the radiologists was evaluated using Cohen's kappa. The radiomics models were evaluated as well in this cohort. In each cross-validation iteration, these models were trained on 80 % of the full dataset, but tested only on patients from the location-matched cohort in the other 20 % of the dataset. The DeLong test was used to compare the AUCs [155].

6.3 Results

6.3.1 Study selection and population

The dataset included 203 patients; see [Table 6.1](#) for the clinical characteristics. The differential diagnosis cohort consisted of 64 fibromyxosarcomas, 31 leiomyosarcomas, 36 myxoid liposarcomas, and 72 DTFs (65 primary, 7 recurrent), of which 61 were suitable for the mutation analysis.

The dataset originated from 68 scanners, resulting in a large heterogeneity in the acquisition protocols, see [Table 6.2](#). From the 72 patients in the DTF cohort, there were 30 T1w post-contrast (42 %), 49 T1w postcontrast FatSat (68 %), 34 T2w (47 %), 33 T2w FatSat (46 %), 3 proton density (PD) (4 %), 18 DCE (25 %) and 3 DWI (4 %) MRI scans. Due to the limited availability of the PD, DCE, and DWI sequences, besides the T1w-MRI, only the T1w post-contrast and T2w (with/without FatSat) sequences were analyzed.

On the subset of 30 DTF that was segmented by both observers, the mean DSC was 0.77 (standard deviation of 0.20), indicating good agreement. An example of the image registration results is depicted in [Figure 6.2](#).

6.3.2 Differential diagnosis

The performance of models 1–6 for the differential diagnosis is shown in [Table 6.3](#). Model 1, based on volume, showed little predictive value (mean AUC of 0.69). Model 2, based on age and sex, performed better (mean AUC of 0.86). Model 3, based

Table 6.1: Clinical characteristics of both cohorts.

	Differential diagnosis cohort				Mutation analysis cohort	
	DTF n=72	Fibro- myxosarcoma n=64	Leiomyosarcoma n=31	Myxoid liposarcoma n=36	DTF n=61	
Sex						
Male	16 (22%)	41 (64%)	19 (61 %)	22 (61%)	15 (25%)	
Female	56 (78%)	23 (36%)	12 (39%)	14 (39%)	46 (75%)	
Age median (IQR)	36 (23-47)	67 (54-77)	66 (55-73)	42 (35-56)	36 (22-47)	
Tumor location						
Head / neck	12 (17%)	-	-	-	11 (18%)	
Chest aperture	4 (6%)	-	-	-	3 (5%)	
Abdominal wall	24 (33%)	-	-	-	16 (26%)	
Back	11 (15%)	-	-	-	10 (16 %)	
Intra-abdominal	1 (1%)	-	-	-	1 (2%)	
Upper extremity	5 (7%)	6 (9%)	7 (23%)	1 (3%)	5 (8%)	
Lower extremity	15 (21%)	58 (91%)	24 (77%)	35 (97 %)	15 (25%)	
Tumor size in cm/ ^a mut ^b /fa	6.3 (4.1-9.8)	7.0 (4.9-12.9)	8.3 (5.2-9.4)	12.8 (8.5-15.3)	6.3 (4.1-9.5)	
Volume in cl median (IQR)	2.0 (0.5-9.8)	5.6 (1.1-34.1)	8.2 (1.7-11.4)	16.8 (5.2-37.4)	2.2 (0.7-9.6)	
Mutation type						
T41A	NA	NA	NA	NA	24 (39%)	
S45F	NA	NA	NA	NA	16 (26%)	
Wild-type	NA	NA	NA	NA	21 (34%)	
MRI sequences						
T2w FS	33 (46%)	37 (58%)	15 (48%)	16 (44%)	26 (43%)	
T2w non-FS	32 (70%)	37 (64%)	19 (39%)	19 (43%)	26 (61%)	
T1w PC FS	49 (70%)	32 (50%)	19 (48%)	22 (51%)	43 (70%)	
T1w PC non-FS	30 (43%)	24 (48%)	11 (23%)	17 (33%)	25 (35%)	

*Abbreviations: DTF: desmoid-type fibromatosis; IQR: interquartile range; cm: centimeter; d: centiliter; MRI: magnetic resonance imaging; FS: FatSat; PC: postcontrast.
Percentages might not add up to 100 % in total because of rounding.
^a Maximum diameter automatically measured in three planes.

Table 6.2: Properties of the acquisition protocols of the 203 T1-weighted MRI sequences in the dataset.

Property	Number	%		
Magnetic field strength				
1T	20	10		
1.5T	167	82		
3T	16	8		
Manufacturer				
Siemens	93	46		
Philips	79	39		
General Electrics	27	13		
Toshiba	4	2		
Setting (Unit)	Mean	Std.	Min	Max
Slice Thickness (mm)	4.66	1.45	1.0	11.0
Repetition time (ms)	619	533	0.0	4620
Echo time (ms)	14	7	2.0	94.0

*Abbreviations: T: tesla; Std: standard deviation; mm: millimeter; ms: milliseconds.

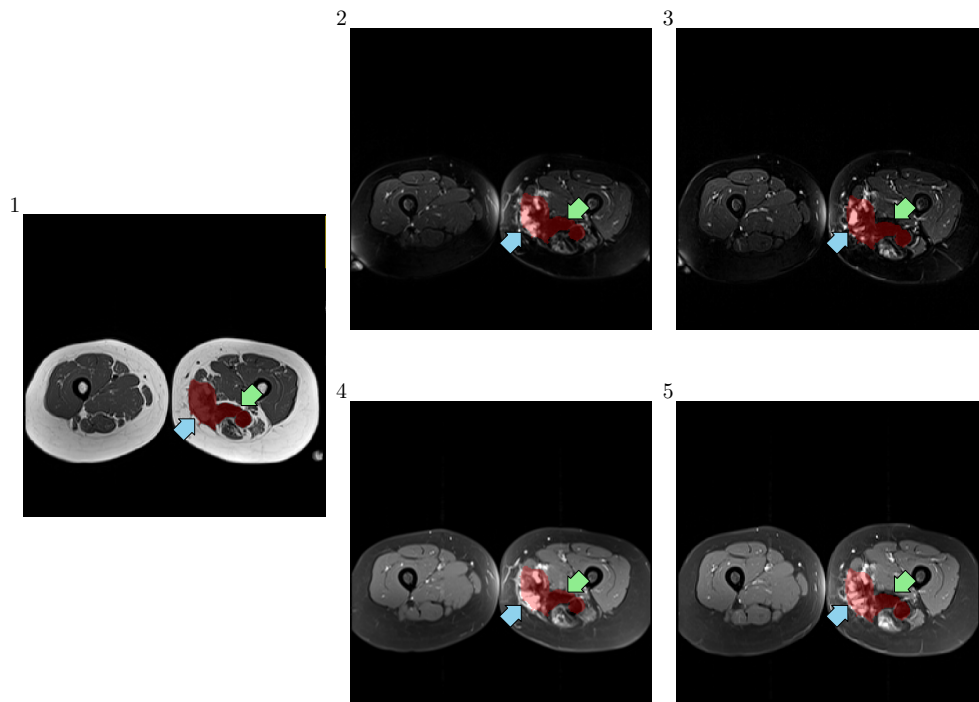


Figure 6.2: Segmentations on various MRI sequences before and after applying image registration in a desmoid-type fibromatosis case. The arrows are at the same position in each image and point at two details where the (mis)alignment is evident. (1) Original T1-weighted (T1w) MRI; (2) Original T2w-MRI; (3) Registered T2w-MRI; (4) Original T1w post-contrast MRI ; (5) Registered T1w post-contrast MRI.

on T1w-MRI, had a mean AUC of 0.79, thus performing worse than age and sex, but better than volume alone. Model 4, combining the T1w-MRI, age, and sex, showed little improvement in terms of mean AUC (0.88) over model 2. Addition of a T2w-MRI, i.e. model 5, or T1 post-contrast MRI, i.e. model 6, both with or without FatSat, both yielded a minor overall improvement over model 3 (mean AUC of 0.84 and 0.84, respectively). These observations were confirmed by the ROC curves in [Figure 6.3](#). The models using either only non-FatSat or FatSat scans, both for the T2w and T1w post-contrast MRI, faired similar, see [Table 6.A.1](#).

6.3.3 Comparison with radiologists

As described in the methods, for the comparison with radiologists, a location-matched cohort consisting of all extremity DTFs and an equal amount of extremity non-DTF was used. To this end, all 20 extremity DTFs and 20 randomly selected extremity non-DTFs were included in the location-matched cohort. The performance of radiomics and the radiologists in this cohort is shown in [Table 6.4](#): model 1 and 5–6 were omitted from the results for brevity. The AUCs of the radiomics models (model 2: 0.93; model 3: 0.88; model 4: 0.98) were generally higher than both radiologists 1 (0.80) and 2 (0.88). This is confirmed by the ROC curves in [Figure 6.4](#). Cohen's kappa between the two radiologists was 0.40, indicating intermediate observer agreement. A DeLong power analysis of the AUCs resulted in a power of only 0.1. Due to the limited power, the p-values of the DeLong test were omitted.

6.3.4 CTNNB1 mutation status stratification

[Table 6.5](#) depicts the performance of the radiomics models for the *CTNNB1* mutation stratification. Model 4, using T1w-MRI, age, and sex, had a high specificity (S45 F: 0.83, T41A: 0.59 and WT: 0.72), but a sensitivity similar to guessing (S45 F: 0.15, T41A: 0.49 and WT: 0.56). This indicates a strong bias in the models towards the negative classes, i. e. not-S45 F, not-T41A and not-WT. As model 4 did not perform well, models 1, 2, and 3 were omitted from the results, as these contain a subset of these features. Adding the T2w or T1w post-contrast imaging, i. e. models 5 and 6,

Table 6.3: Performance of the radiomics models for the DTF differential diagnosis based on: model 1: volume only; model 2: age and sex only; model 3: T1w imaging features, including volume; model 4: the combination of T1w imaging features and age and sex; model 5: the combination of T1w and T2w imaging features; and model 6: the combination of T1w and T1w post-contrast imaging features. Outcomes are presented with the 95% confidence interval.

	Model 1 Volume	Model 2 Age + Sex	Model 3 T1w	Model 4 T1w + Age + Sex	Model 5 T1w + T2w	Model 6 T1w + T1w post-contrast
AUC	0.69 [0.61, 0.76]	0.86 [0.79, 0.92]	0.79 [0.73, 0.85]	0.88 [0.82, 0.93]	0.84 [0.78, 0.89]	0.84 [0.78, 0.90]
BCA	0.59 [0.53, 0.65]	0.78 [0.71, 0.86]	0.71 [0.65, 0.77]	0.79 [0.72, 0.86]	0.68 [0.62, 0.75]	0.75 [0.69, 0.81]
Sensitivity	0.80 [0.70, 0.91]	0.78 [0.66, 0.90]	0.61 [0.49, 0.72]	0.70 [0.57, 0.83]	0.43 [0.31, 0.55]	0.62 [0.52, 0.73]
Specificity	0.39 [0.28, 0.49]	0.79 [0.71, 0.87]	0.81 [0.73, 0.89]	0.88 [0.82, 0.94]	0.94 [0.88, 0.99]	0.88 [0.82, 0.95]
NPV	0.50 [0.71, 0.89]	0.88 [0.81, 0.94]	0.80 [0.76, 0.75]	0.85 [0.80, 0.91]	0.76 [0.72, 0.80]	0.81 [0.76, 0.85]
PPV	0.41 [0.36, 0.46]	0.72 [0.57, 0.76]	0.64 [0.53, 0.75]	0.76 [0.67, 0.86]	0.80 [0.66, 0.94]	0.76 [0.65, 0.88]

[†]Abbreviations: T1w: T1-weighted; T2w: T2-weighted; AUC: area under the receiver operator characteristic curve; BCA: balanced classification accuracy; NPV: negative predictive value; PPV: positive predictive value.

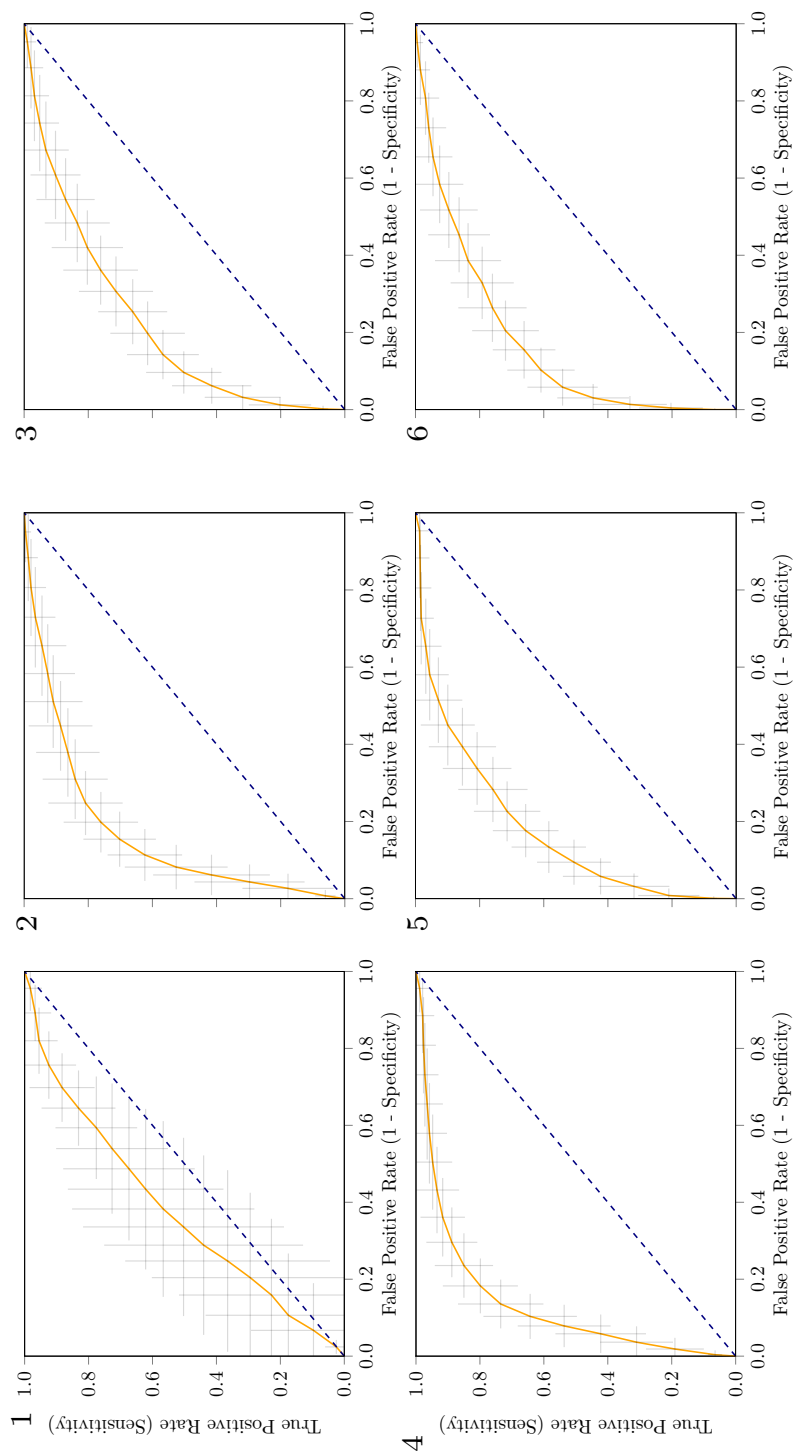


Figure 6.3: Receiver operating characteristic curves of the radiomics models based on volume (1); age and sex (2); T1-weighted (T1w) features (3); T1w features, age, and sex (4); T1w + T2 weighted imaging features (5); and T1w + T1w post-contrast imaging features (6). The grey crosses identify the 95 % confidence intervals of the 100x random-split cross-validation; the orange curve depicts the mean.

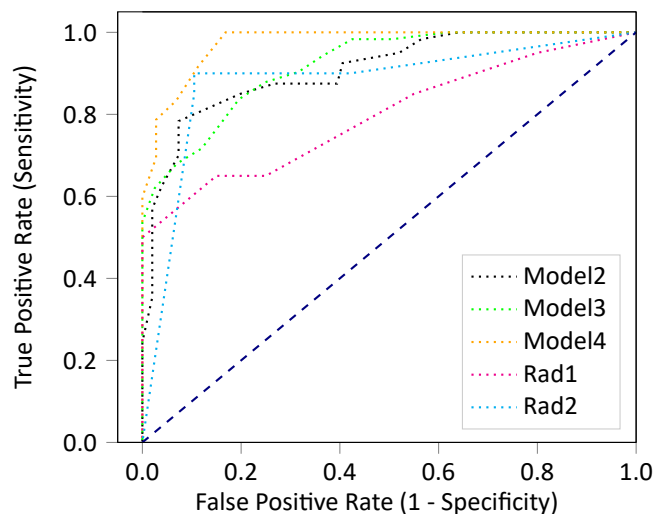


Figure 6.4: Receiver operating characteristic curves of the radiomics models based on age and sex (model 2); imaging (model 3); and imaging, age and sex (model 4); and those of the radiologists (Rad1 and Rad2), in the locationmatched cohort.

Table 6.4: Performance of the two radiologists and the radiomics models in differentiating between DTF (n = 20) and non-DTF (n = 20) in the location-matched cohort. Outcomes are presented with the 95% confidence interval.

	Model 2 Age + Sex	Model 3 T1w	Model 4 T1w + Age + Sex	Rad 1	Rad 2
AUC	0.93 [0.84, >1]	0.87 [0.73, >1]	0.98 [0.92, >1]	0.80	0.88
BCA	0.85 [0.71, 1.00]	0.71 [0.56, 0.87]	0.88 [0.77, 0.99]	0.75	0.90
Sensitivity	0.79 [0.57, >1]	0.49 [0.21, 0.77]	0.78 [0.57, 1.00]	0.65	0.90
Specificity	0.90 [0.71, >1]	0.93 [0.78, >1]	0.98 [0.91, >1]	0.85	0.89
NPV	0.82 [0.61, >1]	0.65 [0.43, 0.76]	0.82 [0.64, >1]	0.71	0.89
PPV	0.91 [0.72, >1]	0.81 [0.47, >1]	0.98 [0.91, >1]	0.81	0.90

*Abbreviations: T1w: T1-weighted; AUC: area under the receiver operator characteristic curve; BCA: balanced classification accuracy; PPV: positive predictive value; NPV: negative predictive value.

did not improve the performance. Hence, the models using either only non-FatSat or FatSat scans were omitted, as these contain subsets of the scans from models 5 and 6.

6.3.5 Model insight

As the *CTNNB1* mutation status stratification models did not perform well, the model insight analysis was only conducted for the differential diagnosis. The p-values from the Mann-Whitney U test between the DTF and non-DTF patients of all features are shown in Table 6.A.3. In the feature importance analysis, 76 T1w-MRI features had significant p-values (5.4×10^{-8} to 4.8×10^{-2}). These included two intensity features (entropy and peak), two shape features (radial distance and volume), and 72 texture features. The p-value of age (1×10^{-11}) was lower than that of all imaging features. The ICC values of all T1w-MRI features are shown in Table 6.A.4. Of the 411 features, 270 (66 %) had an ICC > 0.75 and thus good reliability. Only using these features with a good reliability in model 3 did not alter the performance.

As we are mostly interested in which imaging features define typical DTF, and not age and sex, the patient ranking was conducted for model 3. Of the 203 patients, 104 tumors (24 DTFs, 80 non-DTFs) were always classified correctly by model 3, i.e. in all 100 cross-validation iterations. Nineteen tumors (17 DTFs, 2 non-DTFs) were always classified incorrectly. In Figure 6.5, MRI slices of such typical and atypical examples of DTFs are shown.

6.4 Discussion

This study showed that radiomics based on T1w-MRI can distinguish from STS. Adding T2w or T1w post-contrast MRI did not substantially improve the model. The DTF *CTNNB1* mutation status could not be predicted through radiomics. To our knowledge, this is the first study to evaluate the DTF differential diagnosis and mutation status through an automated radiomics approach.

Age and sex appeared to be strong predictors for the diagnosis of DTF, performing better than T1w-MRI. The combination of imaging, age and sex did not improve the model. This implies that age and sex are sufficient for distinguishing DTF from STS. In line with previous nationwide DTF cohort studies, females represented the majority of our cohort, with a lower median age compared to the median age of the patients from the non-DTF group [135, 156]. The relation in our database may however be too strong, and thereby not representative of clinical practice. For example, above 63 years of age, our database included 60 non-DTF and only a single DTF. While the peak incidence of DTF is between 20–40 years, DTF can affect patients of all ages with reported ranges from 2 to 90 years [156]. Simply classifying all tumors in patients above 63 years as non-DTF, regardless of any tumor (imaging) information, is unfeasible. Such a model cannot be applied in the general population, while the model purely based on T1w-MRI imaging, as it does not use any population-based information. Our cohort might be biased due to the focus on MRI and the extremity as a location, while other modalities (e.g. CT or ultrasound) may be used for certain locations or for certain types of patients. Further research

Table 6.5: Performance of the random forest multilabel radiomics models for the DTF CTNNB1 mutation stratification based on: model 4: T1w imaging features, age and sex; model 5: T1w + T2w imaging features; and model 6: T1w + T1w post-contrast imaging features. Model 4 was evaluated for a single class (S45F, T41A, and WT) or the overall performance (All). Outcomes are presented with the 95% confidence interval.

	Model 4 - S45F	Model 4 - T41A	Model 4 - WT	Model 4 - All	Model 5 - All	Model 5 - All
	T1w + Age + Sex	T1w + Age + Sex	T1w + Age + Sex	T1w + Age + Sex	T1w + T2w	T1w + T1w post-contrast
AUC	0.61 [0.44, 0.77]	0.56 [0.43, 0.68]	0.74 [0.60, 0.87]	0.63 [0.54, 0.72]	0.63 [0.53, 0.72]	0.60 [0.50, 0.69]
BCA	0.48 [0.35, 0.61]	0.53 [0.42, 0.64]	0.65 [0.54, 0.75]	0.56 [0.47, 0.64]	0.57 [0.48, 0.66]	0.53 [0.44, 0.61]
Sensitivity	0.15 [<0, 0.37]	0.49 [0.27, 0.71]	0.56 [0.35, 0.77]	N/A	N/A	N/A
Specificity	0.83 [0.67, 0.98]	0.59 [0.41, 0.76]	0.72 [0.55, 0.89]	N/A	N/A	N/A
NPV	0.76 [0.70, 0.82]	0.65 [0.53, 0.77]	0.73 [0.64, 0.82]	N/A	N/A	N/A
PPV	0.17 [<0, 0.45]	0.42 [0.28, 0.56]	0.59 [0.40, 0.77]	N/A	N/A	N/A

* Abbreviations: T1w: T1-weighted MRI; T2w: T2-weighted MRI; AUC: area under the receiver operator characteristic curve; BCA: balanced classification accuracy; PPV: positive predictive value; NPV: negative predictive value; WT: wild-type; NA: not applicable

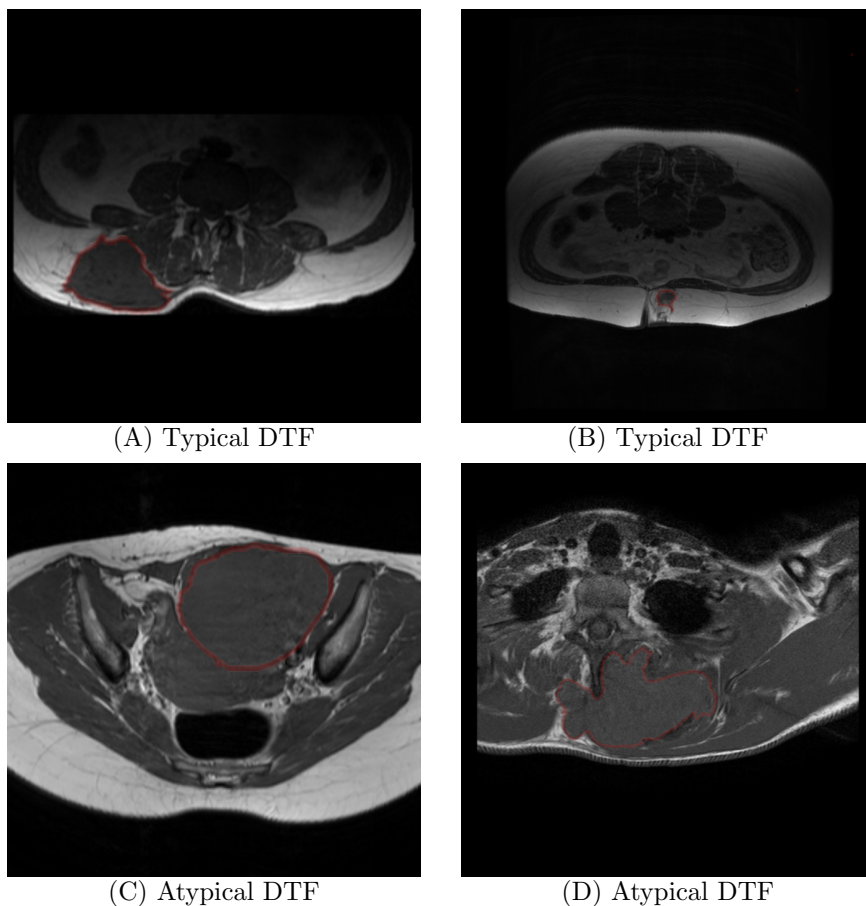


Figure 6.5: The typical examples (A and B) are two cases always classified correctly by the T1-weighted (T1w) imaging model; the atypical examples (C and D) are two cases always classified incorrectly by the T1w imaging model.

should include the expansion of our dataset to make especially the age distribution more representative.

To estimate the clinical value of our model, we compared the performance with the assessment of two radiologists. The model based on imaging performed similar to the radiologists. The model combining age, sex and imaging features, using the same dataset as the radiologist, had a higher AUC than the musculoskeletal radiologists. However this model may suffer from the selection bias as mentioned in the previous section. The agreement between the radiologists was intermediate, indicating observer dependence in the prediction. The radiomics model is observer independent, assuming the segmentation is reproducible as indicated by the high DSC and ICC, and will always give the same prediction on the same image.

The DTF differential diagnosis is highly important for treatment decisions, but

difficult on imaging due to its rarity, while using invasive biopsies brings risks such as tumor growth. The use of our T1w-MRI radiomics model may therefore aid early recognition and diagnosis of DTF, thus shortening the diagnostic delay by enabling direct referral to an STS expertise center. Since all routine MRI protocols include a T1w- MRI, our radiomics method is generalizable, feasible and applicable for use in daily clinical practice. After further model optimization, it may serve as a quick, non-invasive, and low-cost alternative for a biopsy, currently limited to extremities due to the used dataset.

Additionally, we investigated the predictive value of sequences other due to the multicenter imaging dataset. Although T2w-MRI is often used to correlate DTF signal intensity with prognosis or response to therapy [157, 158, 159, 160], in the current study T2w-MRI added little predictive value to the T1w-MRI, similar to the T1w post-contrast MRI. This may however be attributed to the fact that these sequences were only available for a subset of the patients. Our cohort contained too few patients with PD, DCE, or DWI sequences to be analyzed. However, there is little to no indication of the added value of these sequences in DTF [161, 162, 163].

The second aim of this study was to predict the DTF *CTNNB1* mutation status. Our radiomics model was not able to stratify the *CTNNB1* mutation type, which is in line with the absence of literature linking DTF MRI appearance to the *CTNNB1* mutation.

The current study enclosed several limitations. First, due to the rarity of DTF, the DTF sample size was limited and possibly too small for the mutation stratification model to learn from. This also resulted in little statistical power for the mutation analysis, as shown by the large width of our confidence intervals, and for the comparison with the radiologists in the differential diagnosis. Besides primary tumors, the DTF cohort contained also recurrent tumors. As this number was low, and to our knowledge, there are no indications that recurrent DTF appear different on MRI than primary DTF, the expected influence is small. Within the DTF cohort, the WT group was relatively large and might have been subjected to incorrect allocation, as Sanger Sequencing is not always sensitive enough to detect all mutations [144]. The results of the *CTNNB1* mutation status stratification showed a strong bias towards the majority classes, which may be attributed to the class imbalance. Although we exploited commonly used imbalanced learning strategies such as than T1w-MRI. The number of available sequences was however limited resampling and ensembling. other strategies may improve the performance. Second, only extremity DTFs were included for comparison with STS. This was due to the limited availability of MRI in non-extremity soft tissue tumors. However, this is not representative for the entire DTF population, which also occurs frequently in the abdominal wall and trunk [136]. Third, the current radiomics approach requires manual annotations. While accurate, this process is also time consuming and subject to some observer variability as indicated by our DSC, and thus limits the transition to clinical practice. Automatic segmentation methods, for example deep learning, may help to overcome these limitations [164]. Lastly, the dataset originated from 68 different scanners, which resulted in substantial heterogeneity in the acquisition protocols. The lack of standard imaging parameters can be problematic as these can affect the appearance of the tumor and thus the radiomics performance. However, our

method was successfully able to create diagnostic models despite these differences. As these models were trained on a variety of imaging protocols, there is an increased chance that the reported performance can be reproduced in a routine clinical setting when using other MRI scanners. Using a single-scanner with dedicated tumor protocols may improve the model performance, but will limit the generalizability.

Future work should firstly focus on the prospective validation of our findings. Although we did use a multicenter imaging dataset and performed a rigorous cross-validation experiment strictly separating training from testing data, we did not validate our model on an independent, external dataset. Afterwards, the radiomics model could be used to predict clinical outcomes of DTF receiving active surveillance or systemic treatment.

6.5 Conclusions

Our radiomics approach is capable of distinguishing DTF from non- DTF tumors on T1w-MRI, and can potentially aid diagnosis and shorten diagnostic delay. The performance of the model was similar to that of two experienced musculoskeletal radiologists. The model was not able to predict *CTNNB1* mutation status of DTF tumors. Further optimization and external validation of the model is needed to incorporate radiomics in clinical practice.

Sources of funding for research This study was financed by the Stichting Coolsingel (reference number 567), a Dutch non-profit foundation.

CRedit authorship contribution statement Milea J.M. Timbergen: Conceptualization, Resources, Investigation, Funding acquisition, Writing - original draft, Data curation. Martijn P.A. Starmans: Conceptualization, Methodology, Data curation, Resources, Software, Validation, Visualization, Formal analysis, Writing - original draft. Guillaume A. Padmos: Resources, Investigation, Writing - review & editing. Dirk J. Grünhagen: Conceptualization, Supervision, Writing - review & editing. Geert J.L.H. van Leenders: Resources, Investigation, Writing - review & editing. D.F. Hanff: Resources, Writing - review & editing. Cornelis Verhoef: Conceptualization, Supervision, Funding acquisition, Writing - review & editing. Wiro J. Niessen: Methodology, Software, Writing - review & editing. Stefan Sleijfer: Conceptualization, Supervision, Writing - review & editing. Stefan Klein: Conceptualization, Methodology, Formal analysis, Software, Supervision, Writing - review & editing. Jacob J. Visser: Conceptualization, Data curation, Investigation, Methodology, Project administration, Supervision, Validation, Writing - review & editing.

Declaration of competing interest Wiro J. Niessen is founder, scientific lead and stock holder of Quantib BV. The other authors do not declare any conflicts of interest.

Acknowledgements Martijn P.A. Starmans acknowledges funding from the research program STRaTeGy (project number 14929-14930), which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO). This work was

partially carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

Appendix

Appendix 6.A Radiomics feature extraction

This appendix is similar to Vos *et al.* [72] (i.e., [Chapter 5](#) of this thesis), but details relevant for the current study are highlighted.

A total of 411 radiomics features were used in this study. All features were extracted using the defaults for MRI scans from the Workflow for Optimal Radiomics Classification (WORC) toolbox [36], which internally uses the PREDICT [51] and PyRadiomics [44] feature extraction toolboxes. The code to extract the features for this specific study has been published open-source [153]. An overview of all features is depicted in [Table 6.A.2](#). For details on the mathematical formulation of the features, we refer the reader to Zwanenburg *et al.* [39]. More details on the extracted features can be found in the documentation of the respective toolboxes, mainly the WORC documentation [68].

The features can be divided in several groups. Twelve histogram features were extracted using the histogram of all intensity values within the Regions of Interest (ROIs), i.e. the tumors, and included several first-order statistics such as the mean, standard deviation and kurtosis. To create the histogram, the images were binned using a fixed number of 50 bins. Seventeen shape features were extracted based only on the ROI, i.e. not using the image, and included shape descriptions such as the volume, compactness, roundness and circular variance. The orientation of the ROI was described by three features, which represent the three major axis angles of a 3-D ellipse fitted to the ROI. Lastly, 379 texture features were extracted using the Gray Level Co-occurrence Matrix (144 features), Gray Level Size Zone Matrix (16 features), Gray Level Run Length Matrix (16 features), Gabor filters (72 features), Laplacian of Gaussian filters (36 features), vessel (i.e. tubular structure) filters (36 features) [54], local phase filters (36 features) [53], Local Binary Patterns (18 features), and the Neighborhood Grey Tone Difference Matrix (5 features).

Most of the texture features include parameters to be set for the extraction. Beforehand the values of the parameters which will result in features with the highest discriminative power for the classification at hand (e.g. DTF vs non-DTF) is not known. Including these parameters in the workflow optimization, see [Section 6.B](#), would lead to repeated computation of the features, resulting in a redundant decrease in computation time. Therefore, alternatively, these features are extracted at a range of parameters as is default in WORC. The hypothesis is that the features with high discriminative power will be selected by the feature selection methods and/or the machine learning methods as described in [Section 6.B](#). The parameters used are described in [Table 6.A.2](#).

The dataset used in this study is highly heterogeneous in terms of acquisition protocols. Especially the variations in slice thickness may cause feature values to be highly dependent on the acquisition protocol as this varied between 1.0 mm and 11 mm,. Hence, extracting robust 3D features may be hampered by these variations,

especially for low resolutions. To overcome this issue, all features were extracted per 2D axial slice and aggregated over all slices. Afterwards, several first-order statistics over the feature distributions were evaluated and used in the machine learning approach. The images were not resampled, as this would result in interpolation errors. Due to variations in especially the magnetic field strength, echo time, and repetition time, the image contrast highly varies, which would affect the feature values. To partially overcome this, each 3D MRI was normalized using z-scoring before feature extraction. These settings are also the default in WORC.

Appendix 6.B Adaptive workflow optimization for automatic decision model creation

This appendix is similar to Vos *et al.* [72] (i.e., [Chapter 5](#) of this thesis), but details relevant for the current study are highlighted.

The Workflow for Optimal Radiomics Classification (WORC) toolbox [36] makes use of adaptive algorithm optimization to create the optimal performing workflow from a variety of methods. WORC defines a workflow as a sequential combination of algorithms and their respective parameters. To create a workflow, WORC includes algorithms to perform feature scaling, feature imputation, feature selection, over-sampling, and machine learning. If used, as some of these steps are optional as described below, these methods are performed in the same order as described in this appendix. More details can be found in the WORC documentation [68].

Feature scaling was performed to make all features have the same scale, as otherwise the machine learning methods may focus only on those features with large values. This was done through z-scoring, i.e. subtracting the mean value followed by division by the standard deviation, for each individual feature. In this way, all features had a mean of zero and a variance of one.

In the analysis including the T2w or T1w post contrast sequences, in case of a missing sequence, feature imputation was used to estimate replacement values for the missing sequence. Strategies for imputation included 1) the mean; 2) the median; 3) the most frequent value; and 4) a nearest neighbor approach.

Feature selection was performed to eliminate features which were not useful to distinguish between the classes, e.g. DTF vs. non-DTF. These included; 1) a variance threshold, in which features with a low variance (<0.01) are removed. This method was always used, as this serves as a feature sanity check with almost zero risk of removing relevant features; 2) optionally, a group-wise search, in which specific groups of features (i.e. intensity, shape, and the subgroups of texture features as defined in [Section 6.A](#)) are selected or deleted. To this end, each feature group had an on/off variable which is randomly activated or deactivated, which were all included as hyperparameters in the optimization; 3) optionally, individual feature selection through univariate testing. To this end, for each feature, a Mann-Whitney U test is performed to test for significant differences in distribution between the labels (e.g. DTF vs non-DTF). Afterwards, only features with a p-value above a certain threshold are selected. A Mann-Whitney U test was chosen as features may not be normally distributed and the samples (i.e. patients) were independent; and 4) optionally,

principal component analysis (PCA), in which either only those linear combinations of features were kept which explained 95% of the variance in the features or a limited amount of components (between 10 – 50). These feature selection methods may be combined by WORC, but only in the mentioned order.

Oversampling was used to make sure the classes were balanced in the training dataset. These included; 1) random oversampling, which randomly repeats patients of the minority class; and 2) the synthetic minority oversampling technique (SMOTE) [58], which creates new synthetic “patients” using a combination of the features in the minority class. Randomly, either one of these methods or no oversampling method was used.

Lastly, machine learning methods were used to determine a decision rule to distinguish the classes. These included; 1) logistic regression; 2) support vector machines; 3) random forests; 4) naive Bayes; and 5) linear and quadratic discriminant analysis.

Most of the included methods require specific settings or parameters to be set, which may have a large impact on the performance. As these parameters have to be determined before executing the workflow, these are so-called “hyperparameters”. In WORC, all parameters of all mentioned methods are treated as hyperparameters, since they may all influence the decision model creation. WORC simultaneously estimates which combination of algorithms and hyperparameters performs best. A comprehensive overview of all parameters is provided in the WORC documentation [68].

By default in WORC, the performance is evaluated in a 100x random-split train-test cross-validation. In the training phase, a total of 100,000 pseudo-randomly generated workflows is created. These workflows are evaluated in a 5x random-split cross-validation on the training dataset, using 85% of the data for actual training and 15% for validation of the performance. All described methods were fit on the training datasets, and only tested on the validation datasets. The workflows are ranked from best to worst based on their mean performance on the validation sets using the F1-score, which is the harmonic average of precision and recall. Due to the large number of workflows executed, there is a chance that the best performing workflow is overfitting, i.e. looking at too much detail or even noise in the training dataset. Hence, to create a more robust model and boost performance, WORC combines the 50 best performing workflows into a single decision model, which is known as ensembling. These 50 best performing workflows are re-trained using the entire training dataset, and only tested on the test datasets. The ensemble is created through averaging of the probabilities, i.e. the chance of a patient being DTF or non-DTF, of these 50 workflows.

A full experiment consists of executing 50 million workflows (100,000 pseudo-randomly generated workflows times a 5x train-validation cross-validation times 100x train-test cross-validation), which can be parallelized. The computation time of training or testing a single workflow is on average less than a second, depending on the size of the dataset both in terms of samples (i.e. patients) and features. The largest experiment in this study, i.e. the differential diagnoses including 203 patients with both a T1w and T2w MRI had a computation time of approximately 32 hours on a 32 CPU core machine. The contribution of the feature extraction to the computation

time is negligible.

The code for the model creation, including more details, has been published open-source as well [[153](#)].

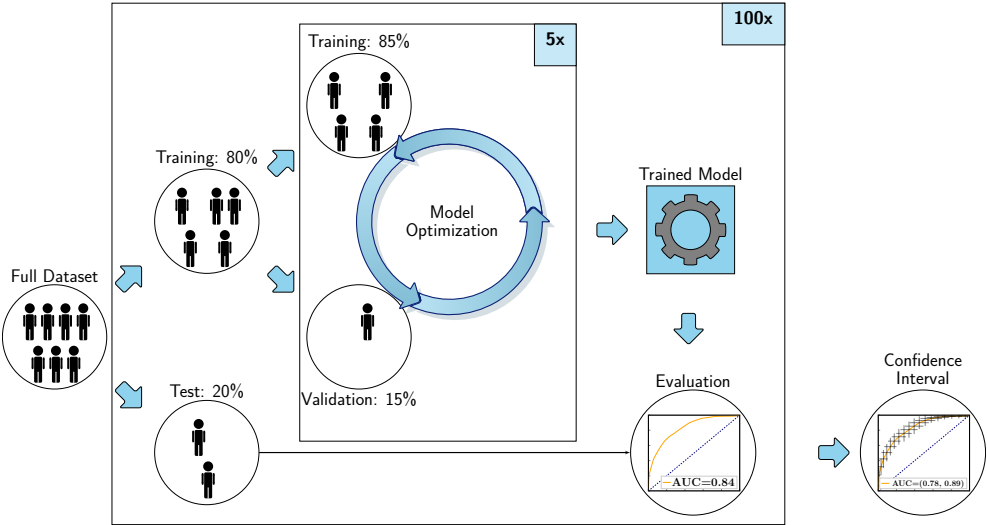


Figure 6.A.1: Visualization of the 100x random split-cross validation, including a second cross validation within the training set.

Table 6.A.1: Performance of the radiomics models for the DTF differential diagnosis based on T1w and T2w non-FatSat imaging features; T1w and T2w FatSat imaging features; T1w and T1w post-contrast non-FatSat imaging features; and T1w and T1w post-contrast FatSat imaging features. Outcomes are presented with the 95% confidence interval.

	T1w + T2w non-FatSat	T1w + T2w FatSat	T1w + T1w post-contrast non-FatSat	T1w + T1w post-contrast FatSat
AUC	0.83 [0.76, 0.89]	0.83 [0.77, 0.89]	0.80 [0.74, 0.85]	0.82 [0.75, 0.88]
BCA	0.64 [0.58, 0.71]	0.66 [0.59, 0.72]	0.73 [0.67, 0.79]	0.72 [0.66, 0.79]
Sensitivity	0.32 [0.19, 0.44]	0.34 [0.20, 0.47]	0.60 [0.49, 0.72]	0.59 [0.48, 0.70]
Specificity	0.97 [0.92, >1]	0.97 [0.94, 1.00]	0.85 [0.79, 0.92]	0.86 [0.79, 0.94]
NPV	0.74 [0.70, 0.77]	0.74 [0.70, 0.78]	0.79 [0.74, 0.84]	0.79 [0.74, 0.83]
PPV	0.87 [0.68, >1]	0.88 [0.71, >1]	0.71 [0.60, 0.82]	0.72 [0.61, 0.84]

*Abbreviations: T1w: T1-weighted images, T2w: T2-weighted images; AUC: area under the receiver operator characteristic curve; BCA: balanced classification accuracy; PPV: positive predictive value; NPV: negative predictive value

Table 6.A.2: Overview of the 411 features used in this study. GLCM and GLCMMS features were calculated in four different directions (0, 45, 90, 135 degrees) using 16 gray levels and pixel distances of 1 and 3. LBP features were calculated using the following three parameter combinations: 1 pixel radius and 8 neighbours, 2 pixel radius and 12 neighbours, and 3 pixel radius and 16 neighbours. Gabor features were calculated using three different frequencies (0.05, 0.2, 0.5) and four different angles (0, 45, 90, 135 degrees). LoG features were calculated using three different widths of the Gaussian (1, 5 and 10 pixels). Vessel features were calculated using the full mask, the edge, and the inner region. Local phase features were calculated on the monogenic phase, phase congruency and phase symmetry.

Histogram (12 features):	Shape (17 features):	Orientation (3 features):	GLCM (6*4*2=48 features):	GLCMMS (12*4*2=96 features):	NGTDM (5 features):	LBP (6*3=18 features):
min max mean median std skewness kurtosis peak range energy quartile range entropy	compactness (mean + std) radius (mean + std) roughness (mean + std) convexity (mean + std) circular variance (mean + std) principal axes ratio (mean + std) elliptic variance (mean + std) solidity (mean + std) volume	theta_x theta_y theta_z	contrast dissimilarity homogeneity angular second moment (ASM) energy correlation	contrast (mean + std) disimilarity (mean + std) homogeneity (mean + std) ASM (mean + std) energy (mean + std) correlation (mean + std)	busyness coarseness complexity contrast strength	mean std median kurtosis skewness peak
GLSZM (16 features):	GLRM (16 features):	Gabor (6*4*3=72 features):	LoG (12*3=36 features)	Vessel (12*3=36 features)	Local phase (12*3=36 features)	
GrayLevelNonUniformity GrayLevelNonUniformityNormalized GrayLevelVariance HighGrayLevelZoneEmphasis LargeAreaEmphasis LargeAreaHighGrayLevelEmphasis LargeAreaLowGrayLevelEmphasis LowGrayLevelZoneEmphasis SizeZoneNonUniformity SizeZoneNonUniformityNormalized SmallAreaEmphasis SmallAreaHighGrayLevelEmphasis SmallAreaLowGrayLevelEmphasis ZoneEntropy ZonePercentage ZoneVariance	GrayLevelNonUniformity GrayLevelNonUniformityNormalized GrayLevelVariance HighGrayLevelRunEmphasis LongRunEmphasis LongRunHighGrayLevelEmphasis LongRunLowGrayLevelEmphasis LowGrayLevelRunEmphasis RunEntropy RunLengthNonUniformity RunLengthNonUniformityNormalized RunPercentage RunVariance ShortRunEmphasis ShortRunHighGrayLevelEmphasis ShortRunLowGrayLevelEmphasis	mean std min max skewness kurtosis	min max mean median std skewness kurtosis peak range energy quartile entropy	min max mean median std skewness kurtosis peak range energy quartile entropy	min max mean median std skewness kurtosis peak range energy quartile entropy	

* Abbreviations: GLCM: gray level co-occurrence matrix; GLCMMS: GLCM multislice; NGTDM: neighborhood gray tone difference matrix; GLSZM: gray level size zone matrix; GLRLM: gray level run length matrix; LBP: local binary patterns; LoG: Laplacian of Gaussian; std: standard deviation.

Table 6.A.3: P-values after Bonferonni correction of features in a Mann-Whitney U test between desmoid type fibromatosis (DTF) and non-DTF patients. Only the features with significant p-values ($p < 0.05$) are depicted. Besides the feature names, several of the labels also include the parameters used. More details on the features can be found in [Section 6.A](#).

Feature label	P-value
tf_Gabor_0.5A1.57mean	5.39E-08
logf_energy_sigma10	7.98E-07
tf_GLSZM_LargeAreaHighGrayLevelEmphasis	1.18E-06
logf_energy_sigma1	1.64E-06
logf_peak_sigma5	1.69E-06
tf_Gabor_0.5A1.57max	1.87E-06
logf_peak_sigma10	1.93E-06
logf_energy_sigma5	2.30E-06
tf_GLRLM_LongRunHighGrayLevelEmphasis	3.08E-06
tf_GLRLM_LongRunEmphasis	4.97E-06
tf_GLSZM_LargeAreaEmphasis	5.66E-06
hf_peak	7.41E-06
logf_peak_sigma1	7.47E-06
phasef_phasesym_peak_WL3_N5	7.74E-06
tf_Gabor_0.5A1.57std	1.15E-05
vf_Frangi_inner_peak_SR(1.0, 10.0)_SS2.0	1.54E-05
semf_Gender	1.58E-05
phasef_phasecong_peak_WL3_N5	1.60E-05
tf_Gabor_0.5A1.57min	1.89E-05
tf_Gabor_0.5A2.36mean	1.92E-05
tf_GLRLM_RunPercentage	2.36E-05
tf_Gabor_0.5A1.57skew	3.69E-05
sf_rad_dist_avg_2D	4.06E-05
vf_Frangi_full_peak_SR(1.0, 10.0)_SS2.0	4.89E-05
vf_Frangi_edge_peak_SR(1.0, 10.0)_SS2.0	4.89E-05
tf_GLRLM_RunVariance	5.09E-05
phasef_monogenic_peak_WL3_N5	5.23E-05
phasef_monogenic_energy_WL3_N5	5.74E-05
tf_GLSZM_ZoneVariance	5.82E-05
tf_GLSZM_LargeAreaLowGrayLevelEmphasis	6.73E-05
tf_Gabor_0.5A1.57kurt	6.82E-05
hf_entropy	2.29E-04
tf_GLRLM_RunEntropy	2.41E-04
tf_GLRLM_GrayLevelNonUniformity	2.99E-04
tf_GLCMMS_correlationd1.0A1.0std	3.10E-04
tf_GLCMMS_correlationd1.0A1.0mean	3.97E-04
tf_Gabor_0.5A2.36std	4.02E-04
tf_GLCMMS_dissimilarityd1.0A1.0mean	5.21E-04
logf_entropy_sigma1	6.88E-04
tf_GLCMMS_dissimilarityd1.0A1.0std	7.95E-04

tf_Gabor_0.2A0.0skew	1.12E-03
tf_GLRLM_LongRunLowGrayLevelEmphasis	1.19E-03
tf_GLRLM_RunLengthNonUniformityNormalized	1.73E-03
tf_GLCMMS_homogeneityd1.0A1.0mean	2.18E-03
vf_Frangi_edge_min_SR(1.0, 10.0)_SS2.0	2.59E-03
vf_Frangi_full_min_SR(1.0, 10.0)_SS2.0	2.59E-03
sf_volume_2D	2.71E-03
vf_Frangi_edge_mean_SR(1.0, 10.0)_SS2.0	3.55E-03
vf_Frangi_full_mean_SR(1.0, 10.0)_SS2.0	3.55E-03
tf_GLCMMS_contrastd1.0A1.0std	3.63E-03
tf_GLCMMS_contrastd1.0A1.0mean	4.25E-03
vf_Frangi_full_median_SR(1.0, 10.0)_SS2.0	4.83E-03
vf_Frangi_edge_median_SR(1.0, 10.0)_SS2.0	4.83E-03
phasef_phasecong_energy_WL3_N5	5.85E-03
logf_entropy_sigma5	7.59E-03
vf_Frangi_edge_quartile_range_SR(1.0, 10.0)_SS2.0_Features_0	7.84E-03
vf_Frangi_full_quartile_range_SR(1.0, 10.0)_SS2.0_Features_0	7.84E-03
phasef_phasesym_median_WL3_N5	7.99E-03
tf_GLCMMS_homogeneityd1.0A1.0std	9.21E-03
tf_Gabor_0.5A2.36min	9.93E-03
tf_Gabor_0.05A0.0mean	1.02E-02
tf_Gabor_0.5A0.79min	1.21E-02
tf_Gabor_0.2A0.79min	1.39E-02
tf_GLCM_correlationd1.0A1.0	1.41E-02
tf_GLRLM_ShortRunEmphasis	1.44E-02
tf_GLCM_homogeneityd3.0A3.0	1.59E-02
tf_Gabor_0.2A0.0mean	1.61E-02
tf_GLSZM_ZonePercentage	1.96E-02
vf_Frangi_inner_min_SR(1.0, 10.0)_SS2.0	2.15E-02
tf_Gabor_0.05A0.0kurt	2.40E-02
tf_Gabor_0.05A0.0max	3.38E-02
logf_entropy_sigma10	3.41E-02
tf_Gabor_0.05A0.0skew	3.76E-02
tf_GLRLM_ShortRunLowGrayLevelEmphasis	3.88E-02
tf_Gabor_0.05A0.0std	4.53E-02
tf_GLCMMS_correlationd3.0A3.0std	4.60E-02
tf_GLCM_homogeneityd1.0A1.0	4.81E-02

*Abbreviations: GLCM: gray level co-occurrence matrix; GLCMMS: GLCM multislice; NGTDM: neighborhood gray tone difference matrix; GLSZM: gray level size zone matrix; GLRLM: gray level run length matrix; LBP: local binary patterns; LoG: Laplacian of Gaussian; std: standard deviation.

Table 6.A.4: Intra-class correlation coefficient (ICC) values of all features among segmentations of two clinicians in a set of 30 desmoid type fibromatosis patients. Only the features with an ICC > 0.75, which are considered as reliable, are included. Besides the feature names, several of the labels also include the parameters used. More details on the features can be found in [Section 6.A](#).

Feature label	ICC
tf_Gabor_0.05A0.79std	0.75
tf_Gabor_0.5A2.36kurt	0.75
tf_GLSZM_SizeZoneNonUniformityNormalized	0.75
tf_GLCM_ASMd1.0A2.36	0.76
logf_mean_sigma1	0.76
tf_Gabor_0.5A2.36skew	0.76
tf_GLRLM_GrayLevelVariance	0.76
tf_GLRLM_GrayLevelNonUniformityNormalized	0.76
tf_GLCMMS_contrastd1.0A0.0std	0.76
tf_Gabor_0.5A0.79mean	0.76
tf_GLCM_ASMd1.0A0.0	0.76
phasef_phasesym_median_WL3_N5	0.77
tf_GLCM_ASMd1.0A1.57	0.77
tf_LBP_skew_R8_P24	0.77
tf_Gabor_0.5A0.79max	0.77
tf_GLSZM_SmallAreaHighGrayLevelEmphasis	0.77
tf_Gabor_0.5A0.79skew	0.77
tf_GLCMMS_ASMd1.0A1.57mean	0.78
tf_GLCMMS_contrastd1.0A0.79std	0.78
tf_GLCMMS_contrastd1.0A0.79mean	0.78
tf_Gabor_0.5A0.79std	0.78
tf_GLSZM_SmallAreaLowGrayLevelEmphasis	0.79
tf_GLCMMS_dissimilarityd1.0A2.36mean	0.79
tf_GLCMMS_ASMd1.0A0.0mean	0.79
tf_Gabor_0.05A0.79min	0.79
hf_entropy	0.79
tf_Gabor_0.5A0.0kurt	0.79
tf_GLCMMS_ASMd1.0A2.36std	0.80
tf_GLCMMS_homogeneityd1.0A1.57std	0.80
tf_GLCMMS_dissimilarityd1.0A2.36std	0.80
tf_GLSZM_SmallAreaEmphasis	0.80
tf_GLCMMS_ASMd1.0A0.79std	0.80
tf_GLCMMS_dissimilarityd1.0A1.57std	0.80
tf_GLCMMS_ASMd1.0A0.0std	0.80
tf_Gabor_0.5A0.0std	0.80
tf_GLCMMS_contrastd1.0A0.0mean	0.80
vf_Frangi_edge_std_SR(1.0, 10.0)_SS2.0	0.80
vf_Frangi_full_std_SR(1.0, 10.0)_SS2.0	0.80
tf_GLCMMS_ASMd1.0A1.57std	0.81
tf_LBP_mean_R8_P24	0.81
tf_GLCMMS_ASMd1.0A0.79mean	0.81

tf_NGTD_M_Contrast	0.81
tf_GLCMMS_dissimilarityd1.0A1.57mean	0.81
tf_GLCMMS_dissimilarityd1.0A0.79mean	0.81
tf_GLCMMS_energyd1.0A1.57mean	0.81
tf_GLCMMS_ASMD1.0A2.36mean	0.81
vf_Frangi_inner_quartile_range_SR(1.0, 10.0)_SS2.0	0.81
tf_LBP_std_R8_P24	0.81
tf_GLCMMS_homogeneityd1.0A0.0std	0.82
tf_GLCMMS_dissimilarityd1.0A0.0std	0.82
tf_GLCMMS_homogeneityd1.0A0.79mean	0.82
tf_Gabor_0.2A2.36std	0.82
tf_GLCMMS_homogeneityd1.0A2.36mean	0.83
tf_GLCMMS_energyd1.0A0.0mean	0.83
tf_GLCMMS_dissimilarityd1.0A0.79std	0.83
tf_GLCMMS_energyd1.0A0.0std	0.83
tf_GLCMMS_correlationd1.0A0.0std	0.83
tf_GLCMMS_energyd1.0A1.57std	0.83
tf_Gabor_0.5A2.36std	0.84
tf_GLCMMS_energyd1.0A0.79mean	0.84
hf_median	0.84
vf_Frangi_inner_max_SR(1.0, 10.0)_SS2.0	0.84
tf_GLCMMS_energyd1.0A0.79std	0.84
tf_Gabor_0.5A2.36mean	0.84
tf_GLCMMS_energyd1.0A2.36std	0.84
tf_GLCMMS_homogeneityd1.0A0.79std	0.84
tf_GLCMMS_homogeneityd1.0A2.36std	0.84
tf_GLCMMS_homogeneityd1.0A1.57mean	0.84
tf_GLSZM_ZoneEntropy	0.84
vf_Frangi_inner_range_SR(1.0, 10.0)_SS2.0	0.85
tf_GLCMMS_energyd1.0A2.36mean	0.85
phasef_phasecong_kurtosis_WL3_N5	0.85
tf_GLCMMS_correlationd1.0A0.79mean	0.85
logf_median_sigma1	0.85
tf_Gabor_0.2A1.57min	0.85
tf_Gabor_0.2A1.57kurt	0.86
tf_GLCMMS_correlationd1.0A0.0mean	0.86
tf_GLCMMS_dissimilarityd1.0A0.0mean	0.86
tf_GLCMMS_homogeneityd1.0A0.0mean	0.86
tf_GLCMMS_correlationd1.0A0.79std	0.86
tf_Gabor_0.5A0.0max	0.86
tf_GLCMMS_correlationd1.0A1.57std	0.87
tf_LBP_mean_R15_P36	0.87
tf_Gabor_0.05A1.57kurt	0.87
sf_cvar_avg_2D	0.87
vf_Frangi_inner_std_SR(1.0, 10.0)_SS2.0	0.87
phasef_phasecong_mean_WL3_N5	0.87
tf_LBP_std_R3_P12	0.87

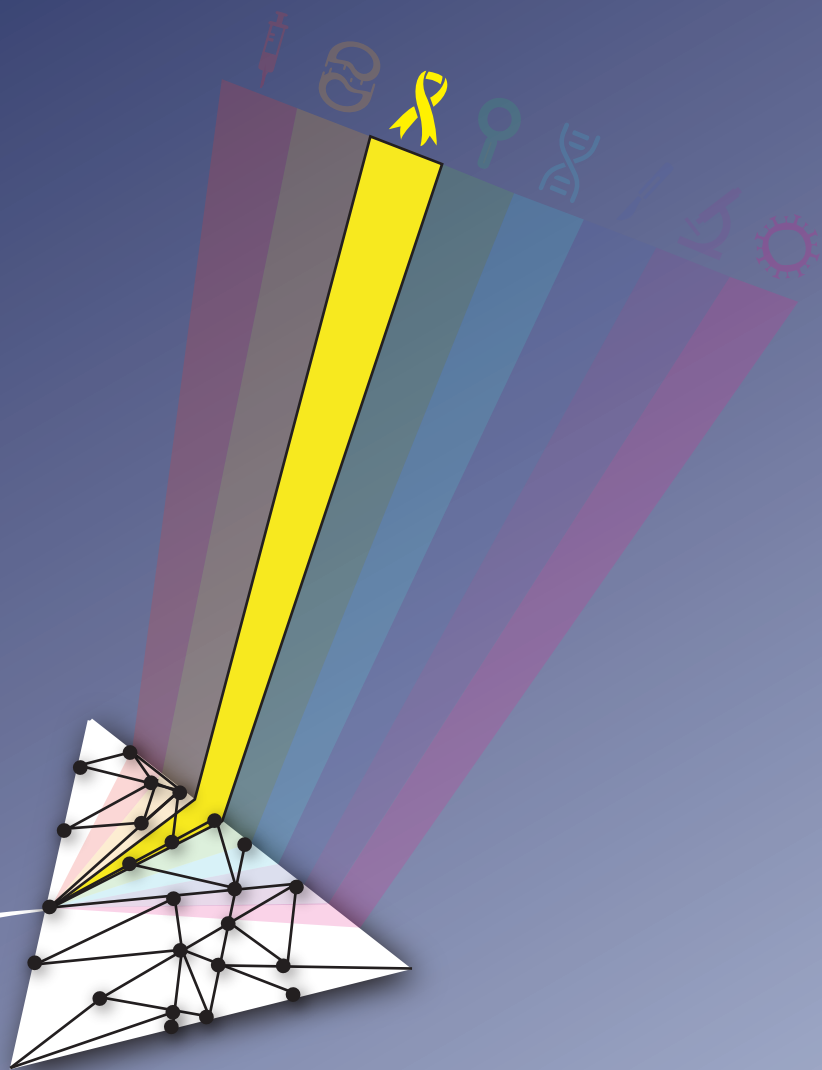
tf_LBP_std_R15_P36	0.88
tf_GLCM_energyd3.0A0.79	0.88
tf_GLCM_energyd3.0A1.57	0.88
tf_Gabor_0.2A2.36skew	0.88
tf_Gabor_0.5A1.57std	0.88
tf_GLCMMS_correlationd1.0A2.36std	0.88
tf_GLCM_energyd3.0A0.0	0.88
phasef_monogenic_mean_WL3_N5	0.88
tf_GLCM_energyd1.0A0.79	0.88
tf_GLRLM_LongRunEmphasis	0.88
tf_GLCMMS_correlationd1.0A2.36mean	0.88
tf_Gabor_0.05A0.79max	0.89
tf_GLCM_energyd3.0A2.36	0.89
tf_GLSZM_HighGrayLevelZoneEmphasis	0.89
tf_GLSZM_LowGrayLevelZoneEmphasis	0.89
tf_GLCM_energyd1.0A2.36	0.89
tf_Gabor_0.2A2.36max	0.89
tf_GLCM_energyd1.0A0.0	0.89
tf_GLRLM_LongRunHighGrayLevelEmphasis	0.89
vf_Frangi_inner_entropy_SR(1.0, 10.0)_SS2.0	0.89
tf_Gabor_0.5A1.57max	0.89
tf_GLCM_energyd1.0A1.57	0.89
tf_Gabor_0.2A0.79skew	0.89
tf_Gabor_0.05A0.79skew	0.90
phasef_monogenic_median_WL3_N5	0.90
logf_mean_sigma5	0.90
tf_Gabor_0.05A2.36min	0.90
hf_mean	0.90
tf_Gabor_0.5A0.0min	0.90
tf_Gabor_0.2A0.0kurt	0.90
tf_GLRLM_LongRunLowGrayLevelEmphasis	0.90
tf_Gabor_0.2A0.79min	0.90
tf_Gabor_0.2A1.57std	0.91
phasef_phasecong_skewness_WL3_N5	0.91
hf_min	0.91
tf_Gabor_0.2A0.79kurt	0.91
phasef_phasesym_max_WL3_N5	0.91
phasef_phasesym_range_WL3_N5	0.91
tf_Gabor_0.2A0.79std	0.91
phasef_phasesym_mean_WL3_N5	0.91
hf_quartile_range	0.91
tf_Gabor_0.2A2.36min	0.91
vf_Frangi_edge_max_SR(1.0, 10.0)_SS2.0	0.92
vf_Frangi_full_max_SR(1.0, 10.0)_SS2.0	0.92
vf_Frangi_edge_range_SR(1.0, 10.0)_SS2.0	0.92
vf_Frangi_full_range_SR(1.0, 10.0)_SS2.0	0.92
logf_skewness_sigma1	0.92

sf_prax_avg_2D	0.92
tf_Gabor_0.05A2.36max	0.92
tf_GLCMMS_correlationd1.0A1.57mean	0.92
tf_Gabor_0.2A0.0skew	0.92
hf_max	0.92
tf_Gabor_0.2A0.79max	0.93
phasef_phasecong_std_WL3_N5	0.93
phasef_phasesym_std_WL3_N5	0.93
tf_GLRLM_RunEntropy	0.93
hf_range	0.93
tf_Gabor_0.05A2.36skew	0.93
tf_Gabor_0.05A0.79kurt	0.93
tf_Gabor_0.05A0.0std	0.93
hf_std	0.93
tf_Gabor_0.2A1.57skew	0.94
sf_rad_dist_std_2D	0.94
tf_Gabor_0.05A1.57min	0.94
tf_Gabor_0.05A2.36std	0.94
tf_Gabor_0.2A1.57mean	0.94
tf_Gabor_0.05A2.36kurt	0.94
tf_Gabor_0.2A0.0std	0.94
tf_Gabor_0.05A0.0max	0.94
tf_Gabor_0.2A2.36mean	0.94
phasef_phasesym_quartile_range_WL3_N5	0.95
logf_min_sigma1	0.95
tf_Gabor_0.2A2.36kurt	0.95
tf_GLCM_contrastd3.0A2.36	0.95
phasef_monogenic_min_WL3_N5	0.95
logf_range_sigma1	0.95
tf_GLCM_homogeneityd3.0A0.79	0.95
hf_peak	0.95
tf_GLCM_homogeneityd3.0A0.0	0.95
tf_Gabor_0.2A0.79mean	0.96
tf_GLCM_homogeneityd3.0A1.57	0.96
vf_Frangi_edge_entropy_SR(1.0, 10.0)_SS2.0	0.96
vf_Frangi_full_entropy_SR(1.0, 10.0)_SS2.0	0.96
tf_GLCM_contrastd3.0A0.0	0.96
logf_max_sigma1	0.96
phasef_monogenic_range_WL3_N5	0.96
tf_Gabor_0.05A1.57std	0.96
tf_GLCM_homogeneityd3.0A2.36	0.96
tf_GLCM_contrastd3.0A1.57	0.96
tf_Gabor_0.05A2.36mean	0.96
tf_GLRLM_RunVariance	0.96
tf_GLCM_contrastd1.0A0.0	0.96
phasef_phasecong_entropy_WL3_N5	0.96
logf_min_sigma5	0.96

logf_max_sigma10	0.96
tf_GLCM_contrastd3.0A0.79	0.96
tf_GLCM_homogeneityd1.0A0.79	0.96
phasef_monogenic_max_WL3_N5	0.96
tf_GLCM_contrastd1.0A2.36	0.96
tf_Gabor_0.05A1.57skew	0.96
tf_GLCM_contrastd1.0A0.79	0.96
phasef_phasecong_max_WL3_N5	0.96
phasef_phasecong_range_WL3_N5	0.96
logf_range_sigma10	0.96
tf_GLCM_homogeneityd1.0A2.36	0.97
logf_min_sigma10	0.97
phasef_monogenic_std_WL3_N5	0.97
phasef_monogenic_quartile_range_WL3_N5	0.97
tf_GLCM_homogeneityd1.0A0.0	0.97
vf_Frangi_inner_kurtosis_SR(1.0, 10.0)_SS2.0	0.97
tf_GLCM_homogeneityd1.0A1.57	0.97
tf_GLCM_contrastd1.0A1.57	0.97
tf_Gabor_0.2A0.0min	0.97
tf_Gabor_0.05A0.0mean	0.97
tf_GLCM_dissimilarityd1.0A0.0	0.97
tf_Gabor_0.5A0.0mean	0.97
tf_GLCM_dissimilarityd3.0A2.36	0.97
phasef_phasesym_entropy_WL3_N5	0.97
phasef_monogenic_entropy_WL3_N5	0.97
logf_kurtosis_sigma1	0.98
tf_GLCM_dissimilarityd1.0A2.36	0.98
tf_GLCM_dissimilarityd1.0A1.57	0.98
tf_GLCM_dissimilarityd3.0A1.57	0.98
tf_Gabor_0.2A1.57max	0.98
logf_range_sigma5	0.98
vf_Frangi_inner_skewness_SR(1.0, 10.0)_SS2.0	0.98
tf_GLCM_dissimilarityd1.0A0.79	0.98
tf_GLCM_dissimilarityd3.0A0.79	0.98
logf_std_sigma5	0.98
tf_GLCM_dissimilarityd3.0A0.0	0.98
tf_Gabor_0.05A1.57mean	0.98
logf_std_sigma1	0.98
logf_mean_sigma10	0.98
tf_Gabor_0.5A1.57mean	0.98
hf_energy	0.98
logf_entropy_sigma1	0.98
tf_Gabor_0.05A0.0min	0.98
logf_std_sigma10	0.98
logf_entropy_sigma10	0.98
logf_median_sigma10	0.98

tf_Gabor_0.2A0.0mean	0.99
tf_Gabor_0.2A0.0max	0.99
logf_entropy_sigma5	0.99
phasef_phasecong_energy_WL3_N5	0.99
sf_rad_dist_avg_2D	0.99
phasef_phasesym_energy_WL3_N5	0.99
phasef_monogenic_kurtosis_WL3_N5	0.99
logf_max_sigma5	0.99
tf_Gabor_0.05A1.57max	0.99
vf_Frangi_inner_energy_SR(1.0, 10.0)_SS2.0	0.99
logf_skewness_sigma5	0.99
logf_quartile_range_sigma10	0.99
tf_GLSZM_GrayLevelNonUniformity	0.99
logf_energy_sigma1	0.99
logf_quartile_range_sigma5	1.00
logf_kurtosis_sigma5	1.00
logf_kurtosis_sigma10	1.00
logf_skewness_sigma10	1.00
logf_peak_sigma10	1.00
tf_GLSZM_SizeZoneNonUniformity	1.00
logf_energy_sigma10	1.00
logf_quartile_range_sigma1	1.00
sf_volume_2D	1.00
tf_GLRLM_RunLengthNonUniformity	1.00
logf_peak_sigma1	1.00
logf_peak_sigma5	1.00
logf_energy_sigma5	1.00
phasef_monogenic_peak_WL3_N5	1.00
vf_Frangi_inner_peak_SR(1.0, 10.0)_SS2.0	1.00
tf_Gabor_0.5A1.57min	1.00
phasef_phasesym_peak_WL3_N5	1.00
phasef_monogenic_energy_WL3_N5	1.00
phasef_phasecong_peak_WL3_N5	1.00
vf_Frangi_edge_peak_SR(1.0, 10.0)_SS2.0	1.00
vf_Frangi_full_peak_SR(1.0, 10.0)_SS2.0	1.00
tf_GLRLM_GrayLevelNonUniformity	1.00
phasef_phasecong_min_WL3_N5	1.00
phasef_phasesym_min_WL3_N5	1.00
tf_LBP_median_R8_P24	1.00
tf_LBP_peak_R15_P36	1.00
tf_LBP_peak_R3_P12	1.00
tf_LBP_peak_R8_P24	1.00

*Abbreviations: GLCM: gray level co-occurrence matrix; GLCMMS: GLCM multislice; NGTDM: neighborhood gray tone difference matrix; GLSZM: gray level size zone matrix; GLRLM: gray level run length matrix; LBP: local binary patterns; LoG: Laplacian of Gaussian; std: standard deviation.



7.

Differential diagnosis and molecular stratification of gastrointestinal stromal tumors on CT images using a radiomics approach

Based on: **M. P. A. Starmans***, M. J. M. Timbergen*, M. Vos, M. Renckens, D. J. Grünhagen, G. J. L. H. van Leenders, R. S. Dwarkasing, F. E. J. A. Willemsen, W. J. Niessen, C. Verhoef, S. Sleijfer, J. J. Visser, and S. Klein, "Differential diagnosis and molecular stratification of gastrointestinal stromal tumors on CT images using a radiomics approach," *Accepted*. arXiv: [2010.06824](https://arxiv.org/abs/2010.06824) [[eess.IV](#)]

* indicates equal contributions

Abstract

Distinguishing gastrointestinal stromal tumors (GISTs) from other intra-abdominal tumors and GISTs molecular analysis is necessary for treatment planning, but challenging due to its rarity. The aim of this study was to evaluate radiomics for distinguishing GISTs from other intra-abdominal tumors, and in GISTs, predict the *c-KIT*, *PDGFRA*, BRAF mutational status and mitotic index (MI). All 247 included patients (125 GISTs, 122 non-GISTs) underwent a contrast-enhanced venous phase CT. The GIST vs. non-GIST radiomics model, including imaging, age, sex and location, had a mean area under the curve (AUC) of 0.82. Three radiologists had an AUC of 0.69, 0.76, and 0.84, respectively. The radiomics model had an AUC of 0.52 for *c-KIT*, 0.56 for *c-KIT* exon 11, and 0.52 for the MI. Hence, our radiomics model was able to distinguish GIST from non-GISTs with a performance similar to three radiologists, but was not able to predict the *c-KIT* mutation or MI.

7.1 Introduction

Gastrointestinal stromal tumors (GISTs) are rare mesenchymal tumors of the gastrointestinal tract, with an estimated incidence between 10-15 cases per million inhabitants per year [165, 166]. The most common tumor locations are the stomach (56%) and the small intestine (32%); less common locations are the esophagus (<1%) and the colorectal region (6%) [165]. Differentiating GISTs from other intra-abdominal tumors (non-GISTs), such as schwannomas, leiomyosarcomas, leiomyomas, esophageal/gastric junctional adenocarcinomas, and lymphomas is highly important for treatment planning [167]. Computed tomography (CT) is the imaging modality of choice in GIST diagnosis [168], but as the differential diagnosis remains challenging, assessment through an invasive tissue biopsy is generally required [169]. A non-invasive and quicker alternative may aid in the early assessment of GISTs.

Treatment planning of GISTs is also based on their molecular profile. The mitotic index (MI) reflects the proliferative rate of GISTs, correlates with survival and risk of metastatic spread [170], and as such determines whether or not a patient with localized disease should get adjuvant systemic treatment. Treatment decisions are also based on the mutational status of GISTs. *PDGFRA* exon 18 mutated (Asp842Val) GISTs are resistant to imatinib [171] and alternative treatments are being explored in this specific subgroup. GISTs with a *c-KIT* exon 11 mutation also have shown a greater sensitivity for imatinib than those with a *c-KIT* exon 9 mutations [167], hence the latter are often treated with a higher imatinib dose. The MI and these genetic mutations are currently assessed through an invasive tissue biopsy.

The field of radiomics relates imaging features to molecular characteristics in order to non-invasively contribute to diagnosis, prognosis and treatment decisions. Several radiomics studies have shown promising results in risk stratification of GISTs [172, 173, 174, 175, 176, 177, 178, 179, 180, 181]. However, radiomics has not been previously used to distinguish GISTs from non-GISTs, nor to predict the mutational status or the MI.

The aim of this study was to evaluate whether an automatically optimized radiomics model based on CT is capable of 1) differentiating GISTs from other intra-abdominal tumors resembling GIST prior to treatment, i.e. the differential diagnosis; and 2) predicting the presence and type of mutation (*BRAF*, *PDGFRA* and *c-KIT*) and the MI of GISTs, i.e. the molecular analysis.

7.2 Methods

7.2.1 Data collection

Approval by the Erasmus MC institutional review board was obtained (MEC-2017-1187). Patients from our institute between 2004-2017 with a histopathologically proven primary GIST or intra-abdominal tumors resembling GIST with at least a contrast-enhanced venous phase CT prior to treatment [167, 182], were retrospectively included. Several GISTs may have been included in the Dutch GIST registry. Exact numbers on potential overlap with previous studies using the registry cannot be determined. As no radiomics studies on this registry have been published,

potential overlap has little relevance. Age at diagnosis, sex, and tumor location were collected. Tumor location was based on radiology reports and categorized into: (distal) esophagus, stomach, small intestine, colon, rectum, pelvis, mesentery, uterus, and other. The sample sizes of the non-GIST and the GIST cohort were matched. The non-GIST subtypes were balanced, i.e. a similar number of patients per subtype was randomly included. GISTs with a known mutation status and/or MI, prior to therapy were included for the molecular analysis. Both were obtained from pathology reports and analyzed on either the primary lesion or, in case of metastatic disease at first presentation, on secondary lesions. The mutation was categorized as 'absent' or 'present' for each type (e.g. *c-KIT*) and subtype (e.g. *c-KIT* exon 11). The MI (expressed in high power fields (HPF), magnification 40x, totaling 5mm²), determined on biopsy or excision material, was split into low ($\leq 5/50$ HPF) and high ($> 5/50$ HPF) [183]. An adjusted MI was calculated per 50 HPF when the MI was not counted per 50 HPF. In case of unknown mutation status or MI, the case was excluded from the particular analysis.

7.2.2 Radiomics

The radiomics workflow is depicted in [Figure 7.1](#), adapted from Vos *et al.* [72] (i.e., [Chapter 5](#) of this thesis). The tumors were all manually segmented once by one of two clinicians under supervision of a musculoskeletal radiologist (5 years of experience) using in-house developed software [105]. A subset of 30 GISTs was segmented by both clinicians, in which intra-observer variability was evaluated through the pairwise Dice Similarity Coefficient (DSC), with DSC > 0.70 indicating good agreement [150]. For each lesion, 564 features quantifying intensity, shape, and texture were extracted. For details, see [Section 7.A](#). To create a decision model from the features, the WORC toolbox was used [36, 72, 151]. In WORC, the decision model creation consists of several steps, e.g. feature selection, resampling, and machine learning. WORC performs an automated search amongst a variety of algorithms for each step and determines which combination maximizes the prediction performance on the training set. For details, see [Section 7.B](#). The code for the feature extraction and model creation has been published open-source [184].

7.2.3 Robustness to segmentation and image acquisition variations

Radiomics' robustness to segmentation variations was assessed using the intra-class correlation coefficient (ICC) of the features on the subset of 30 GISTs which were segmented by two observers. "Good" and "excellent" reliability were defined by ICC > 0.75 and ICC > 0.90 , respectively [154]. Moreover, the impact of ICC-based feature selection on model performance was assessed by creating models using only features with good or excellent reliability.

Robustness to variations in the acquisition parameters was assessed by using ComBat harmonization [185, 186]. In ComBat, feature distributions are harmonized for variations in the imaging acquisition, e.g. due to differences in hospitals, manufacturers, or acquisition parameters. When dividing the dataset into groups based on these variations, the groups have to remain sufficiently large to estimate the har-

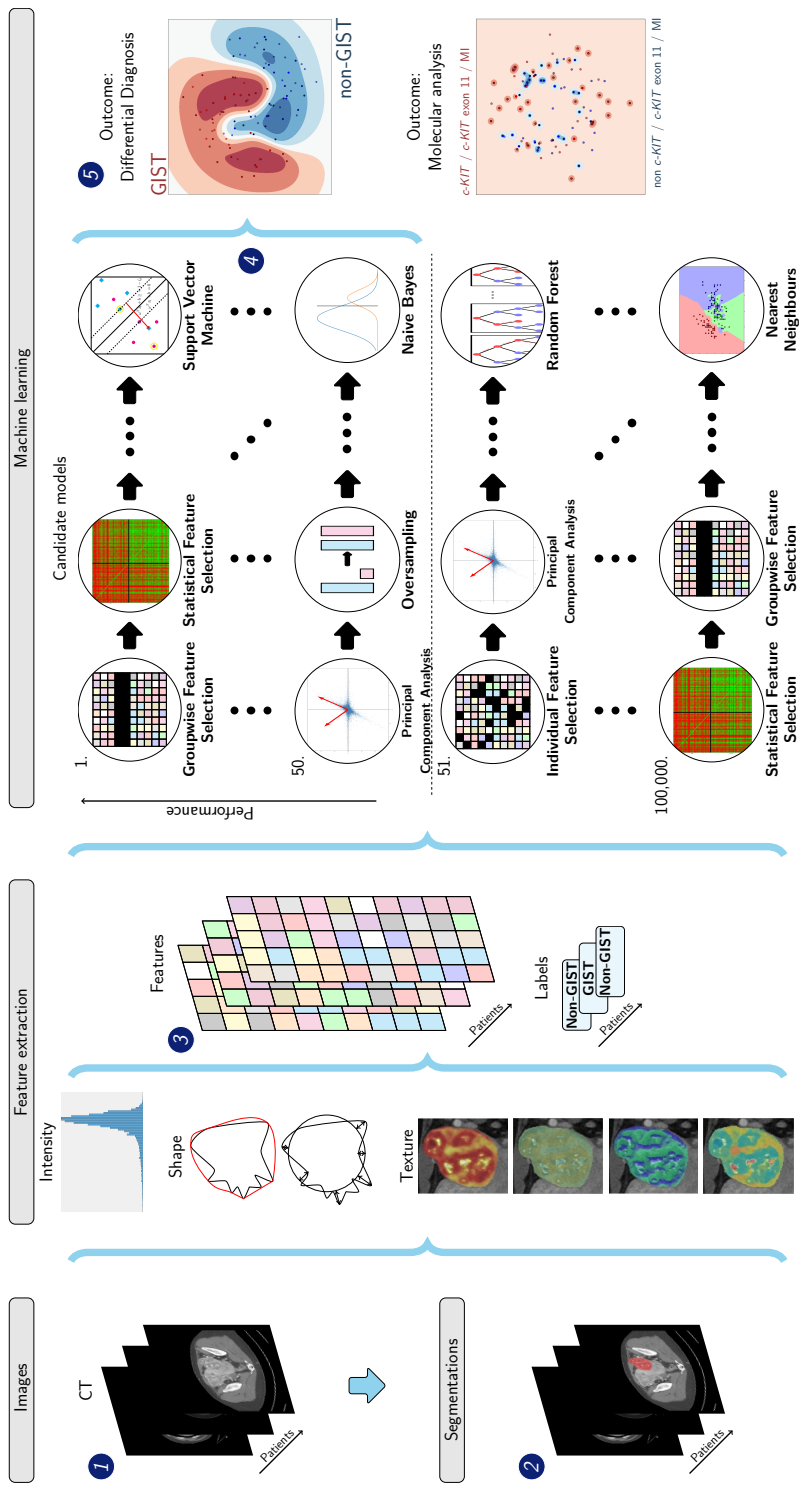


Figure 7.1: Schematic overview of the radiomics approach. Input to the algorithm are the CT images (1). Processing steps then include segmentation of the tumor (2), feature extraction (3) and the creation of machine learning decision models (5), using an ensemble of the best 50 workflows from 100,000 candidate workflows (4), which are different combinations of the different processing and analysis steps (e.g. the classifier used). * Abbreviations: GIST, gastrointestinal stromal tumor.

monization parameters. In our study, groups were defined based on manufacturer alone, or based on protocol, defined as the combination of manufacturer and slice thickness (above or below the median). No moderation variable was used.

7.2.4 Experimental setup

Evaluation of all models was done through a 100x random-split cross-validation. In each iteration, the data was randomly split in 80% for training and 20% for testing in a stratified manner, to make sure the distribution of the classes in all sets was similar to the original, see [Figure 7.A.1](#). Within the training set, model optimization was performed using an internal cross-validation (5x). Hence, all optimization was done on the training set to eliminate any risk of overfitting on the test set.

Performance was evaluated using the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve, balanced classification accuracy (BCA) [65], sensitivity, and specificity. The positive classes were defined as: GIST, the presence of the mutations, and a high MI in the respective analyses. The 95% confidence intervals (CIs) were constructed using the corrected resampled t-test based on the results from all 100 cross-validation iterations [64]. Both the mean and the confidence intervals are reported. ROC confidence bands were constructed using fixed-width bands [67].

To assess the predictive value of the various features, models were trained based on: 1) volume; 2) location; 3) age and sex; 4) imaging; 5) age, sex, and imaging; and 6) age, sex, imaging, and tumor location. Models 2 and 6, assessing the predictive value of location, were included in the differential diagnosis because the radiologists also used tumor location.

In the mutation stratification, only the subset of patients with a known mutation (sub)type was taken into account for each analysis.

7.2.5 Model insight

To explore the predictive value of individual features, the Mann-Whitney U univariate statistical test was used for continuous variable, and a Chi-square test for categorical variables. P-values were corrected for multiple testing using the Bonferroni correction, and were considered statistically significant at a p-value <0.05. To gain insight into the models, the patients were ranked based on the consistency of the model predictions. Typical examples for each class consisted of the patients that were correctly classified in all cross-validation iterations; atypical vice versa. To estimate model robustness to segmentation and acquisition protocol variations, for the differential diagnosis, additional imaging-only models (i.e. model 4) were created using only reliable features through ICC-based feature selection and ComBat harmonization, respectively.

7.2.6 Performance of the radiologists

To compare the models with clinical practice, three radiologists (5, 15 and 12 years of experience) independently evaluated the tumors. Evaluation was done on a ten-point

scale to indicate the scoring certainty, i.e. 1 = strongly disagree GIST, 5 = mildly disagree GIST, 6 = mildly agree GIST, 10 = strongly agree GIST. The radiologists were blinded for the diagnosis but had access to the CT scan, patient age and sex. Only the differential diagnosis was scored, as the mutation and MI are based on pathology in clinical practice. The radiologists' agreement was evaluated using Cohen's Kappa. To enable direct statistical comparison between the radiologists' performances on the one hand and the best radiomics model on the other hand, the radiomics model was evaluated in an additional leave-one-out cross-validation, after which the DeLong test was used to compare the AUCs [155].

7.3 Results

7.3.1 Study population and dataset

The dataset included 247 patients (125 GISTs, 122 non-GISTs), which were all included in the differential diagnosis analysis. Sclerosing mesenteritis (N=16) and inflammatory fibroid polyp (N=4) were excluded due to their small numbers. Clinical characteristics of the dataset are summarized in Table 7.1. The dataset of 247 CT scans originated from 66 different scanners, resulting in variation in the acquisition protocols, see Table 7.1. The scans originated from four different manufacturers (Siemens, Berlin, Germany: 126, Philips, Eindhoven, the Netherlands: 63, General Electric, Boston, United States: 10, Toshiba, Tokyo, Japan: 48). On the subset of 30 GISTs that was segmented by both observers, the mean DSC was 0.84 (standard deviation of 0.20), indicating good agreement.

Two patients were excluded for the molecular radiomics analysis as the molecular characteristics were obtained after receiving systemic treatment. A total of 123 GISTs were included in the cohort for the molecular analysis. The mutation analysis was performed on tissue obtained from the primary lesion, except for three patients for which this was performed on a metastatic hepatic lesion. A *c-KIT* mutational analysis was performed in 98/123 (80%) GIST patients. One patient had a *c-KIT* mutation which was not further specified. Twenty-six (27%) patients had no *c-KIT* mutation. The majority of patients had a *c-KIT* exon 11 mutation (N=59, 60%). Due to the low numbers of *c-KIT* exon 9 (N=10), *c-KIT* exon 13 (N=2), *PDGFRA* (N=14), and *BRAF* (N=0), these mutations were excluded from further analysis.

The MI was available in 90/123 (73%) GISTs (55 low, 35 high). The MI of 33 (37%) GISTs was converted to the adjusted MI. The MI was determined on excision material in 54 (60%) patients, and on biopsy material in 36 (40%) patients, including one patient in which the MI was based on the hepatic GIST metastasis.

7.4 Evaluation of models for the differential diagnosis

The performances of the models distinguishing GISTs from non-GISTs are shown in Table 7.2 and Figure 7.2. On average, model 1, based solely on volume, did not perform well (AUC of 0.56). Model 2, based on location, performed better (AUC of 0.82), but showed a sharp cutoff in the ROC curve (Figure 7.2b). Model 3, based on age and sex, did not perform well (AUC of 0.61). Model 4, based on CT imaging

Table 7.1: Clinical and CT scan characteristics of the dataset.

	GISTs	Schwannoma	Leiomyo-sarcoma	Leiomyoma	Esophageal/ gastric junctional adenocarcinoma	Lymphoma
Number	125	22	25	25	25	25
Sex						
Male	66 (53%)	11 (50%)	7 (28%)	6 (24%)	16 (64%)	18 (72%)
Female	59 (47%)	11 (50%)	18 (72%)	19 (76%)	9 (36%)	7 (28%)
Age at diagnosis^d	64 (56-72)	59 (45-67)	60 (53-71)	49 (41-59)	65 (56-74)	62 (52-67)
Tumor location^b						
(Distal) esophagus	-	-	-	6 (24%)	5 (20%)	-
Stomach	80 (64%)	2 (9.1%)	1 (4%)	3 (12%)	20 (80%)	2 (8%)
Small intestine	29 (23%)	-	1 (4%)	-	-	4 (16%)
Colon	1 (1%)	-	2 (8%)	-	-	1 (4%)
Rectum	7 (6%)	-	-	-	-	-
Pelvis	1 (1%)	7 (31.8%)	5 (0%)	2 (8%)	-	1 (4%)
Mesentery	-	-	-	-	-	7 (28%)
Uterus	-	-	2 (8%)	13 (52%)	-	-
Other	7 (6%)	13 (59.1%)	14 (56%)	1 (4%)	-	10 (40%)
Tumor volume (cl)^e	15.7 (4.3-52.6)	13.9 (1.6-29.7)	12.9 (6.7-99.6)	8.2 (1.6-25.5)	1.6 (0.7-3.1)	9.4 (4.6-29.4)
Acquisition protocol						
Slice thickness (mm) ^{a,c}	5.0 (3.0-5.0)	5.0 (2.0-6.0)	5.0 (3.0-5.0)	3.0 (3.0-5.0)	4.0 (3.0-5.0)	3.0 (3.0-3.0)
Pixel spacing (mm) ^{a,c}	0.72 (0.68-0.78)	0.74 (0.68-0.79)	0.72 (0.68-0.78)	0.75 (0.68-0.84)	0.74 (0.66-0.78)	0.77 (0.69-0.85)
Tube current (mA) ^{a,c}	189 (129-283)	162 (115-206)	221 (160-349)	210 (147-395)	210 (142-312)	207 (145-301)
Peak kilovoltage ^{a,c}	120 (100-120)	120 (120-120)	120 (100-120)	120 (100-120)	120 (100-120)	100 (100-100)

^aMedian (inter quartile range)

^b Percentages may not add up to 100% because of rounding

^c Other values than those given in the median and inter quartile range do occur

^d Abbreviations: GIST: gastrointestinal stromal tumor; cl: centiliter; mm: millimeter; mA: milli Ampère

features, performed better with a mean AUC of 0.74. Model 5, combining imaging with age and sex, did not yield an improvement (AUC of 0.70). Model 6, adding tumor location, did yield an improvement (AUC of 0.82).

7.4.1 Comparison with radiologists

The performance of the radiologists is shown in [Table 7.2](#) and [Figure 7.2](#). Compared to model 6, which had the same inputs, i.e. based on imaging, age, sex, and tumor location, the AUCs of the first two radiologists (0.69 and 0.76) were lower, while the AUC of the third radiologists was similar (0.84). All radiologists had a relatively high sensitivity (0.74, 0.90, and 0.78), but a low specificity (0.60, 0.44, and 0.74). Cohen's kappa measures between the pairs of radiologists were 0.20, 0.31 and 0.33, all indicating poor inter-observer agreement. The Delong test between the pairs of radiologists indicated a significant difference in performance for radiologists 1 versus 3 ($p=6 \times 10^{-5}$) and 2 versus 3 ($p=0.01$); for radiologist 1 versus 2, the power was too low to claim insignificance. Radiomics model 6 evaluated in a leave-one-out cross-validation (AUC of 0.82) also performed statistically significantly better than the first radiologist ($p=0.0018$); for comparison with the other radiologists, the power was too low to claim insignificance.

7.4.2 Evaluation of models for the molecular analysis

For the *c-KIT* mutation stratification and MI predictions, the performance of the radiomics model based on age, sex and imaging features (model 5) is depicted in [Table 7.3](#).

In the mutation stratification, the radiomics models had a mean AUC of 0.52, a low specificity (0.01), and a high sensitivity (0.97) for predicting the presence of a *c-KIT* mutation in general (model 5A). Predicting the presence of a *c-KIT* exon 11 mutation (model 5B) performed similar (AUC of 0.56). The MI prediction (model 5C) had a mean AUC of 0.52, a high specificity (0.71) and a low sensitivity (0.30). All models thus focus on the majority class and perform close to guessing, as is confirmed by the ROC curves in [Figure 7.A.2](#). As models 1, 3 and 4 include a subset of the features from model 5, which already did not perform well, these results are omitted. Model 2 and 6 were only used in the differential diagnosis.

7.4.3 Model insight

As the molecular analysis models did not perform well, the model insight analysis was only conducted for the differential diagnosis. The p-values of the feature importance analysis are shown in [Table 7.A.1](#). In total, 43 features had significant p-values after Bonferroni correction (1.1×10^{-17} to 4.6×10^{-2}). These included the tumor location (1.1×10^{-17}), two intensity features, three orientation features, four shape features of which three related to the tumor area, and 33 texture features. A list of these features and their p-values has been added to the mentioned published code [184]. Volume was not found to be significant.

Results on ranking patients from typical to atypical are only shown for the model based on imaging, i.e. model 4, as we were interested in the imaging features that

Table 7.2: Performance of the radiomics models for the differential diagnosis based on 1) volume; 2) location; 3) age and sex; 4) imaging features; 5) imaging features, age and sex; and 6) imaging features, age, sex and tumor location, and of the three radiologists (R1, R2 and R3). Values are presented with their 95% confidence intervals.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	R1	R2	R3
	<i>Volume</i>	<i>Location</i>	<i>Age + sex</i>	<i>Imaging</i>	<i>Imaging + age</i> + sex	<i>Imaging + age</i> + sex + <i>location</i>			
AUC	0.56 [0.48, 0.64]	0.82 [0.76, 0.88]	0.61 [0.55, 0.68]	0.74 [0.67, 0.81]	0.70 [0.64, 0.77]	0.82 [0.76, 0.87]	0.69	0.76	0.84
BCA	0.55 [0.49, 0.60]	0.82 [0.76, 0.88]	0.56 [0.51, 0.62]	0.67 [0.60, 0.74]	0.66 [0.59, 0.73]	0.74 [0.68, 0.80]	0.67	0.67	0.76
Sensitivity	0.28 [0.15, 0.41]	0.93 [0.88, 0.98]	0.62 [0.54, 0.71]	0.58 [0.46, 0.72]	0.58 [0.46, 0.70]	0.71 [0.61, 0.82]	0.74	0.90	0.78
Specificity	0.82 [0.70, 0.93]	0.71 [0.60, 0.82]	0.50 [0.41, 0.60]	0.75 [0.67, 0.84]	0.75 [0.64, 0.85]	0.77 [0.68, 0.85]	0.60	0.44	0.74

* Abbreviations: AUC: area under the receiver operating characteristic curve; BCA: balanced classification accuracy; R1, R2 and R3: radiologists 1, 2 and 3

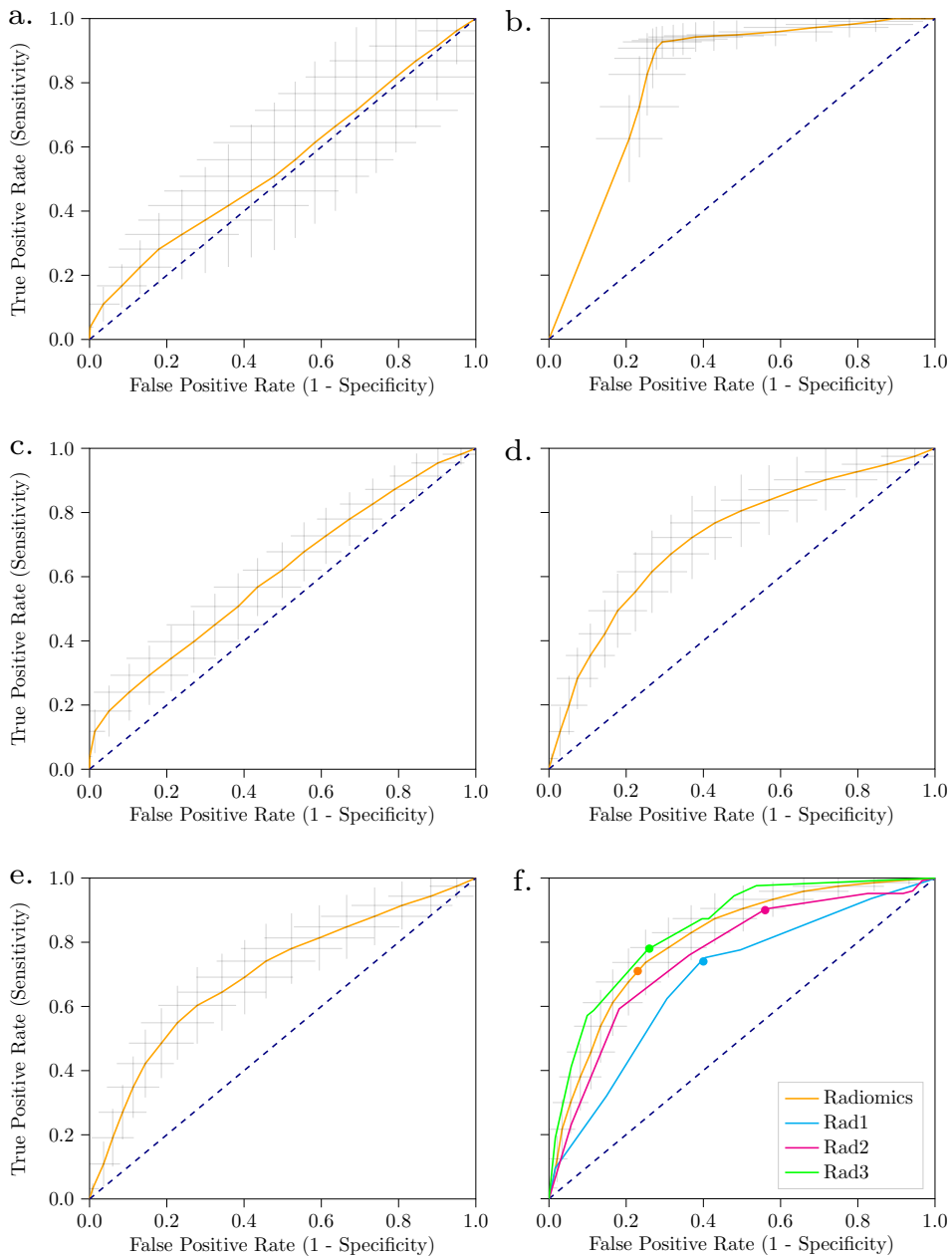


Figure 7.2: Receiver operating characteristic curves of the radiomics models for the differential diagnosis based on volume (a); location (b); age and sex (c); imaging (d); imaging, age, and sex (e); imaging, age, sex, and tumor location (f). Additionally in figure (f), the curves for scoring by three radiologists are shown, and the cut-off points for both the radiomics model and the radiologists. For the radiomics models, the grey crosses identify the 95% confidence intervals of the 100x random-split cross-validation; the orange curve is fit through their means.

Table 7.3: Performance of radiomics model 5, based on imaging, age, and sex, for the GIST mutation stratification and the mitotic index for A) *c-KIT* presence vs. absence B) *c-KIT* exon 11 presence vs. absence; and C) mitotic index ($\leq 5/50$ HPF vs. $>5/50$ HPF). The number of patients included in each analysis (N) is mentioned in the heading. Values are presented with their 95% confidence intervals.

	Model 5A <i>c-KIT</i> (N=98)	Model 5B <i>c-KIT</i> exon 11 (N=96)	Model 5C Mitotic index (N=90)
AUC	0.52 [0.38, 0.66]	0.56 [0.44, 0.67]	0.52 [0.38, 0.65]
BCA	0.49 [0.46, 0.52]	0.52 [0.44, 0.61]	0.51 [0.41, 0.60]
Sensitivity	0.97 [0.91, >1.00]	0.78 [0.64, 0.91]	0.30 [0.12, 0.47]
Specificity	0.01 [<0.00, 0.07]	0.27 [0.11, 0.43]	0.71 [0.56, 0.87]

* Abbreviations: AUC: area under the receiver operating characteristic curve; BCA: balanced classification accuracy; PPV: positive predictive value; NPV: negative predictive value

defined typical GISTs. Of the 247 patients, 104 tumors (44 GISTs, 60 non-GISTs, 42%) were always classified correctly, and were thus considered typical. Twenty-nine tumors (18 GISTs, 11 non-GISTs, 12%) were always classified incorrectly and thus atypical. In [Figure 7.3](#), four CT slices of such typical and atypical examples of GISTs are shown. Visual inspection of the tumors defined as typical or atypical by the radiomics model showed a relation with necrosis (more present in typical GIST, typically a necrotic core) and shape (more compact, circular and non-lobulated for typical GIST). The patients which were equally often classified as GIST and non-GIST in the cross-validation iterations were mostly small tumors. These typical characteristics and the difficulty with small tumors correspond to the literature for GIST risk stratification [[180](#), [187](#)]. Smaller tumors were also more often misclassified by the radiologists in our study.

A list of the ICC values of all imaging features has been added to the mentioned published code [[184](#)]. Of the 564 imaging features, 327 (58%) had an ICC > 0.75 and thus good reliability, 197 (34%) had an ICC > 0.90 and thus excellent reliability. Only using features with a good or excellent reliability in model 4 did not substantially alter the performance (AUC of 0.76 and 0.75, respectively), see [Table 7.A.2](#). Similarly, using ComBat to harmonize the features for manufacturer or protocol differences did not substantially alter the performance either (AUC of 0.76 and 0.73, respectively), see [Table 7.A.2](#).

7.5 Discussion

Radiomics can distinguish GISTs from other intra-abdominal tumors with a performance similar to three radiologists. Radiomics could not predict the presence and subtype of *c-KIT* mutations or the MI.

Diagnosing GISTs is currently done through a biopsy, aided by manually scored imaging features [[168](#), [188](#), [189](#)]. The ability to distinguish GISTs from non-GISTs on routine CT scans through radiomics could be a non-invasive and quick alternative for the initial assessment of intra-abdominal tumors. The use of our model would

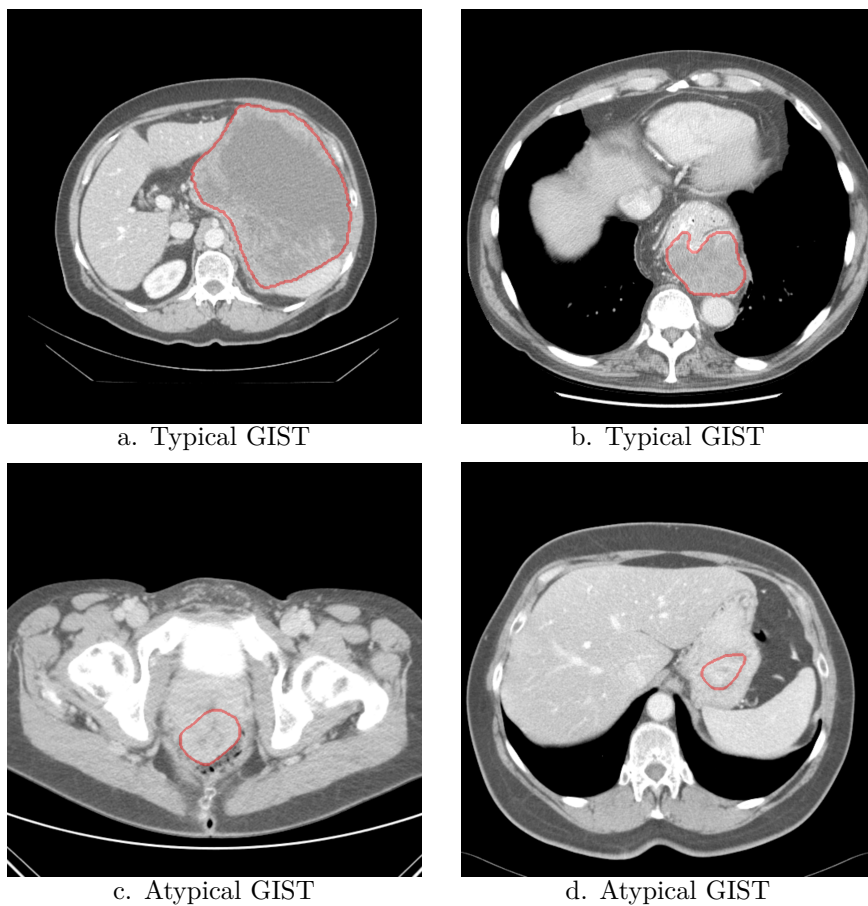


Figure 7.3: Examples of GISTs always correctly or always incorrectly predicted by the radiomics CT imaging model, i.e. model 4. The typical examples (a and b) are two of the GISTs always classified correctly by the model; the atypical examples (c and d) are two of the GISTs always classified incorrectly by the model.

aid quick referral of GIST patients from a peripheral hospital to a center of expertise without the need to wait for an invasive biopsy and time-consuming pathology analysis, and it would prevent GIST patients being missed (i.e. false negatives), unnecessary referral or even treatment for non-GIST (i.e. false positives). Additionally, for non-GIST benign abnormalities, our differential diagnosis model prevents further dissemination investigation and pathologic examinations. To our knowledge, this is the first study to evaluate the GIST differential diagnosis on many locations through an automated radiomics approach on a large, multi-scanner dataset and compare the performance with radiologists.

The performance of the differential diagnosis imaging-only model was similar to two radiologists, and significantly better than one. The agreement between the radiologists was poor, indicating observer dependence in the prediction, and there were significant performance differences. The advantage of the radiomics model is that it is automatic and observer independent, assuming the segmentation is reproducible as indicated by the high DSC, and that it will always give the same prediction on the same image, thereby improving over manual scoring.

Tumor location is highly relevant for distinguishing GISTs from non-GISTs as GISTs grow typically in the stomach or small intestines [165]. In our study, tumor location was based on radiology reports, which is subjective and occasionally fails to report the true tumor primary origin [183]. Moreover, the tumor location distribution in our dataset may not be a correct representation of the overall population, e.g. only non-GISTs were located in the uterus. Despite the subjectivity of potential bias in tumor location, we added location to the imaging model for a fair comparison with the radiologists. Further research on location-matched datasets is required to investigate the value of location in the GIST differential diagnosis model.

In the literature, risk classification of outcomes such as recurrence or aggressive behavior for GISTs has mostly been based on criteria such as the Armed Forces Institute of Pathology criteria, modified National Institutes of Health consensus criteria of 2008, and the modified Fletcher classification system [167, 190, 191, 192, 193]. Several studies to evaluate radiomics for this risk stratification have been conducted over the last years [172, 173, 174, 175, 176, 177, 178, 179, 180, 181]. These studies illustrate the clinical need for new methods to stratify GISTs for guiding treatment decisions and show the potential of applying radiomics in the setting of GIST.

Our first contribution with respect to the existing literature is the focus on the diagnostic trajectory of GISTs by aiming to predict the differential diagnosis, whereas existing studies mainly focus on risk classification [167, 190, 191, 192, 193]. Second, our method determines the optimal radiomics pipeline from a large number of radiomics algorithms and parameters, automatically evaluating a large number of radiomics methods, whereas existing studies typically report the results of a “hand-crafted”, manually optimized radiomics pipeline [172, 173, 174, 175, 176, 177, 178, 179, 180, 181]. Moreover, through an extensive cross-validation scheme, all model optimization was performed on the training dataset, eliminating the risk of overfitting of the model on the test set. This increases the chances of generalizability of our performance estimates. Lastly, we evaluated the model’s robustness to segmentation and scanner variations. In our results, neither ICC-based features

selection nor ComBat harmonization substantially altered the performance. As the model performance did not alter when using these measures to increase radiomics robustness, this may indicate that the model is already robust to these variations. Evaluating ComBat with the differential diagnosis (e.g. GIST or non-GIST) as moderation variable did lead to a near perfect performance (AUC of 0.99), but similar results were obtained with randomly labeling patients as GIST or non-GIST, indicating that this near perfect performance was a result of overfitting by ComBat.

Our study has several limitations. First, there was substantial heterogeneity in the acquisition protocols. This heterogeneity may have (negatively) affected the performance. Nevertheless, even on this heterogeneous dataset, the radiomics model achieved promising performance, similar to three experienced radiologists, suggesting high generalizability of the model. Second, the dataset for the mutation analysis was small (N=98 GISTs with known *c-KIT* mutation), which may have been too small for radiomics to learn from. Third, the use of different gene panels for the GIST mutational analysis over the years might have led to a potential underestimation of mutation prevalence in the current cohort, as newer sequencing techniques use larger gene panels and have a higher sensitivity. Additionally, only for a subset of the patients (e.g. 90 of the 125 (72%) in the MI analysis) complete histologic data was available. No data regarding the clinical outcome such as survival or recurrence was available for the GISTs. Fourth, the current radiomics approach requires manual segmentation. While accurate, this process is also time consuming and potentially subject to observer variability, although the DSC indicated good agreement and our ICC-based feature selection shows that only using reliable features resulted in a similar performance as using all features. Automatic segmentation methods, such as using deep learning, may help to overcome this limitation. Lastly, the current study has a retrospective study design. A prospective study confirming our results is needed.

Future work should focus on the external validation of our findings on an independent, external dataset. Additionally, extension of the dataset will lead to more statistical power, may improve the performance as the model has more cases to learn from, and may facilitate more data driven approaches such as deep learning. Also, this may result in sufficient samples to study prediction of *PDGFRA*, *BRAF*, and other rare *c-KIT* mutations. Importantly, an alternative to non-invasively determine the mutational status of a GIST is by ctDNA [194]. With better performance of both methods, the combination of radiomics and ctDNA assessment would allow to assess in patients with metastatic disease the most important determinants rendering an invasive biopsy redundant. Eventually, this may be followed by a prospective clinical trial with harmonized acquisition protocols in which the performance, as well as the cost-effectiveness, are assessed.

7.6 Conclusion

Our radiomics model was able to distinguish GIST from non-GIST intra-abdominal tumors based on pre-treatment CT imaging with a performance similar to three experienced radiologists. Our model may therefore aid clinicians early on in the

diagnostic chain. The model was not able to predict the *c-KIT* mutational status and the MI.

Acknowledgements Martijn P. A. Starmans acknowledges funding from the research program STRaTeGy (project number 14929-14930), which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO). This work was partially carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

Competing interests Wiro J. Niessen is founder, scientific lead and stock holder of Quantib BV. The other authors do not declare any conflicts of interest.

Appendix

Appendix 7.A Radiomics feature extraction

This supplemental material is similar to Vos *et al.*, Timbergen *et al.* [72, 73] (i.e., Chapter 5 and Chapter 6 of this thesis), but details relevant for the current study are highlighted.

A total of 564 radiomics features were used in this study. All features were extracted using the defaults for CT scans from the Workflow for Optimal Radiomics Classification (WORC) [36], which internally uses the PREDICT [51] and PyRadiomics [44] feature extraction toolboxes. An overview of all features is depicted in Table 7.A.3. For details on the mathematical formulation of the features, we refer the reader to Zwanenburg *et al.* [39]. More details on the extracted features can be found in the documentation of the respective toolboxes, mainly the WORC documentation [68].

For CT scans, the images are by default not normalized as the scans already have a fixed unit and scale (i.e. Hounsfield), contrary to MRI. The images were not resampled, as this would result in interpolation errors. The code to extract the features has been published open-source [184].

The features can be divided in several groups. Thirteen intensity features were extracted using the histogram of all intensity values within the ROIs and included several first-order statistics such as the mean, standard deviation and kurtosis. These describe the distribution of Hounsfield units within the lesion. Thirty-five shape features were extracted based only on the ROI, i.e. not using the image, and included shape descriptions such as the volume, compactness and circular variance. These describe the morphological properties of the lesion. Nine orientation features were used, describing the orientation of the ROI, i.e. not using the image. Lastly, 483 texture features were extracted using Gabor filters (144 features), Laplacian of Gaussian filters (36 features), vessel (i.e. tubular structures) filters (36 features) [54], the Gray Level Co-occurrence Matrix (144 features) [39], the Gray Level Size Zone Matrix (16 features) [39], the Gray Level Run Length Matrix (16 features) [39], the Gray Level Dependence Matrix (14 features) [39], the Neighbourhood Grey Tone Difference Matrix (5 features) [39], Local Binary Patterns (18 features) [52], and local

phase filters (36 features) [53, 195]. These features describe more complex patterns within the lesion, such as heterogeneity, occurrence of blob-like structures, and presence of line patterns.

Most of the texture features include parameters to be set for the extraction. Beforehand the values of the parameters that will result in features with the highest discriminative power for the classification at hand (e.g. GIST vs non-GIST) are not known. Including these parameters in the workflow optimization, see [Section 7.B](#), would lead to repeated computation of the features, resulting in a redundant decrease in computation time. Therefore, alternatively, these features are extracted at a range of parameters as is default in WORC. The hypothesis is that the features with high discriminative power will be selected by the feature selection methods and/or the machine learning methods as described in [Section 7.B](#). The parameters used are described in [Table 7.A.3](#).

The dataset used in this study is heterogeneous in terms of acquisition protocols. Especially the variations in slice may cause feature values to be dependent on the acquisition protocol. Hence, extracting robust 3D features may be hampered by these variations, especially for low resolutions. To overcome this issue, all features were extracted per 2D axial slice and aggregated over all slices, which is default in WORC. Afterwards, several first-order statistics over the feature distributions were evaluated and used in the machine learning approach.

Appendix 7.B Adaptive workflow optimization for automatic decision model creation

This appendix is similar to Vos *et al.*, Timbergen *et al.* [72, 73] (i.e., [Chapter 5](#) and [Chapter 6](#) of this thesis), but details relevant for the current study are highlighted.

The Workflow for Optimal Radiomics Classification (WORC) toolbox [36] makes use of adaptive algorithm optimization to create the optimal performing workflow from a variety of methods. WORC defines a workflow as a sequential combination of algorithms and their respective parameters. To create a workflow, WORC includes algorithms to perform feature scaling, feature imputation, feature selection, over-sampling, and machine learning. If used, as some of these steps are optional as described below, these methods are performed in the same order as described in this appendix. More details can be found in the WORC documentation [68]. The code to use WORC for creating the differential diagnosis and molecular analysis decision models in this specific study has been published open-source [184].

Feature scaling was performed to make all features have the same scale, as otherwise the machine learning methods may focus only on those features with large values. This was done through z-scoring, i.e. subtracting the mean value followed by division by the standard deviation, for each individual feature. In this way, all features had a mean of zero and a variance of one. A robust version of z-scoring was used, in which outliers, i.e. values below the 5th percentile or above the 95th percentile, were excluded from computing the mean and variance.

When a feature could be computed, e.g. a lesion is too small for specific feature to be extracted or a division by zero occurs, feature imputation was used to estimate

replacement values for the missing values. Strategies for imputation included 1) the mean; 2) the median; 3) the most frequent value; and 4) a nearest neighbor approach. Feature selection was performed to eliminate features which were not useful to distinguish between the classes, e.g. GIST vs. non-GIST. These included; 1) a variance threshold, in which features with a low variance (<0.01) are removed. This method was always used, as this serves as a feature sanity check with almost zero risk of removing relevant features; 2) optionally, a group-wise search, in which specific groups of features (i.e. intensity, shape, and the subgroups of texture features as defined in [Section 7.A](#)) are selected or deleted. To this end, each feature group had an on/off variable which is randomly activated or deactivated, which were all included as hyperparameters in the optimization; 3) optionally, individual feature selection through univariate testing. To this end, for each feature, a Mann-Whitney U test was performed to test for significant differences in distribution between the labels (e.g. GIST vs non-GIST). Afterwards, only features with a p-value above a certain threshold were selected. A Mann-Whitney U test was chosen as features may not be normally distributed and the samples (i.e. patients) were independent; and 4) optionally, principal component analysis (PCA), in which either only those linear combinations of features were kept which explained 95% of the variance in the features or a limited number of components (between 10 – 50). These feature selection methods may be combined by WORC, but only in the mentioned order.

Various resampling strategies can optionally be used, which can be used to overcome class imbalances and reduce overfitting on specific training samples. These included various methods from the imbalanced-learn toolbox [57]; random over-sampling, random under-sampling, near-miss resampling, the neighborhood cleaning rule, ADASYN, and SMOTE (regular, borderline, Tomek and the edited nearest neighbors).

Lastly, machine learning methods were used to determine a decision rule to distinguish the classes. These included; 1) logistic regression; 2) support vector machines; 3) random forests; 4) naive Bayes; and 5) linear and quadratic discriminant analysis.

Most of the included methods require specific settings or parameters to be set, which may have a large impact on the performance. As these parameters have to be determined before executing the workflow, these are so-called "hyperparameters". In WORC, all parameters of all mentioned methods are treated as hyperparameters, since they may all influence the decision model creation. WORC simultaneously estimates which combination of algorithms and hyperparameters performs best. A comprehensive overview of all parameters is provided in the WORC documentation [68].

By default in WORC, the performance is evaluated in a 100x random-split train-test cross-validation. In the training phase, a total of 25,000 pseudo-randomly generated workflows is created. These workflows are evaluated in a 5x random-split cross-validation on the training dataset, using 85% of the data for actual training and 15% for validation of the performance. All described methods are fit on the training datasets, and only tested on the validation datasets. The workflows are ranked from best to worst based on their mean performance on the validation sets using the F1-score, which is the harmonic average of precision and recall. Due

to the large number of workflows that is executed, there is a chance that the best performing workflow is overfitting, i.e. looking at too much detail or even noise in the training dataset. Hence, to create a more robust model and boost performance, WORC combines the 50 best performing workflows into a single decision model, which is known as ensembling. These 50 best performing workflows are re-trained using the entire training dataset, and only tested on the test datasets. The ensemble is created through averaging of the probabilities, i.e. the chance of a patient being GIST or non-GIST, of these 50 workflows.

A full experiment consists of executing 12.5 million workflows (25,000 pseudo-randomly generated workflows, times a 5x train-validation cross-validation times 100x train-test cross-validation), which can be parallelized. The computation time of training or testing a single workflow is on average less than a second, depending on the size of the dataset both in terms of samples (i.e. patients) and features. The largest experiment in this study, i.e. the differential diagnosis including 247 patients had a computation time of approximately 32 hours on a 32 CPU core machine. The contribution of the feature extraction to the computation time was negligible.

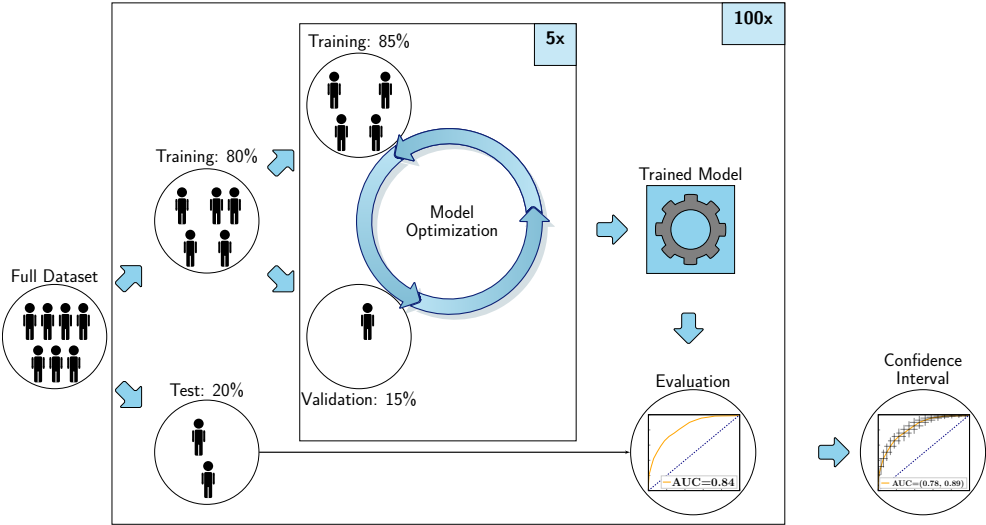


Figure 7.A.1: Visualization of the 100x random split-cross validation, including a second cross validation within the training set.

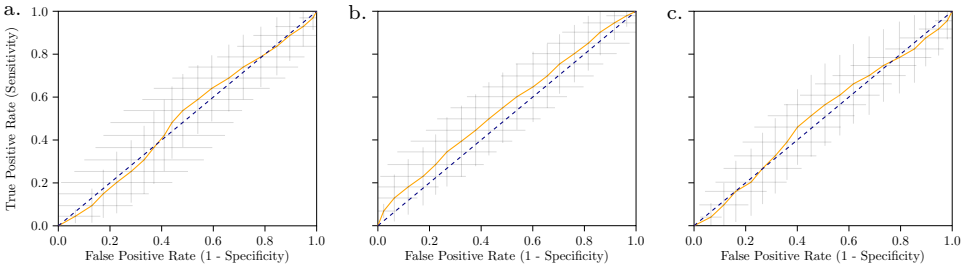


Figure 7.A.2: Receiver operating characteristic curves of the radiomics models based on the CT imaging features, age at diagnosis and sex for (a) *c-KIT* presence vs. absence; (b) *c-KIT* exon 11 presence vs. absence; and (c) mitotic index ($\leq 5/50$ HPF vs. $> 5/50$ HPF). The grey crosses identify the 95% confidence intervals of the 100x random-split cross-validation; the orange curve is fit through their means.

Table 7.A.1: P-values of features from univariate tests between GIST and non-GIST patients after Bonferonni correction. A Mann-Whitney U test was used for continuous variables, a Chi-square test for categorical variables. Only features with a p-value < 0.05, which are considered statistically significant, are shown. Besides the feature names, several of the feature labels also include the parameters used. More details on the features can be found in Supplemental Materials 1.

Label	Mann-Whitney U p-value	Chi2 p-value
semf_location		1.14×10^{-17}
of_COM_x	7.46×10^{-8}	
hf_energy	1.59×10^{-5}	
of_COM_Index_x	3.06×10^{-5}	
tf_Gabor_mean_F0.2_A0.79	3.37×10^{-4}	
tf_GLRLM_LongRunEmphasis	1.10×10^{-3}	
tf_GLRLM_RunVariance	1.15×10^{-3}	
tf_GLSZM_ZonePercentage	1.31×10^{-3}	
tf_GLRLM_ShortRunEmphasis	1.31×10^{-3}	
tf_GLRLM_RunPercentage	1.37×10^{-3}	
tf_GLRLM_RunLengthNonUniformityNormalized	1.37×10^{-3}	
tf_GLDM_DependenceVariance	1.49×10^{-3}	
tf_Gabor_mean_F0.2_A2.36	1.50×10^{-3}	
tf_GLDM_LargeDependenceEmphasis	1.57×10^{-3}	
tf_GLDM_SmallDependenceLowGrayLevelEmphasis	2.97×10^{-3}	
tf_GLCMMS_homogeneityd3.0A0.79mean	4.12×10^{-3}	
tf_Gabor_energy_F0.2_A0.79	4.91×10^{-3}	
tf_GLRLM_LongRunHighGrayLevelEmphasis	5.86×10^{-3}	
tf_GLDM_SmallDependenceEmphasis	6.45×10^{-3}	
tf_Gabor_energy_F0.2_A2.36	7.20×10^{-3}	
sf_area_std_2D	7.21×10^{-3}	
tf_GLDM_DependenceNonUniformityNormalized	8.36×10^{-3}	
tf_GLDM_LargeDependenceLowGrayLevelEmphasis	8.70×10^{-3}	
tf_Gabor_energy_F0.2_A1.57	8.97×10^{-3}	
tf_GLDM_LargeDependenceHighGrayLevelEmphasis	0.010	
tf_Gabor_mean_F0.2_A0.0	0.011	
tf_GLCMMS_homogeneityd3.0A0.0mean	0.013	
sf_area_max_2D	0.015	
tf_GLSZM_LargeAreaHighGrayLevelEmphasis	0.016	
tf_GLSZM_LargeAreaEmphasis	0.016	
tf_GLSZM_ZoneVariance	0.016	
tf_Gabor_energy_F0.2_A0.0	0.017	
hf_min	0.017	
sf_area_avg_2D	0.022	
tf_GLRLM_LongRunLowGrayLevelEmphasis	0.024	
of_COM_y	0.025	
tf_GLSZM_LargeAreaLowGrayLevelEmphasis	0.027	
tf_Gabor_median_F0.05_A2.36	0.027	
sf_shape_Maximum2DDiameterSlice	0.031	
tf_GLRLM_GrayLevelNonUniformityNormalized	0.039	
vf_Frangi_inner_energy_SR(1.0_10.0)_SS2.0	0.039	
tf_Gabor_mean_F0.5_A2.36	0.045	
tf_Gabor_kurtosis_F0.05_A0.79	0.046	

* Abbreviations: GLCM: gray level co-occurrence matrix; GLCMMS: GLCM multislice; NGTDM: neighborhood gray tone difference matrix; GLSZM: gray level size zone matrix; GLRLM: gray level run length matrix; LBP: local binary patterns; LoG: Laplacian of Gaussian; std: standard deviation.

Table 7.A.2: Performance of the radiomics models for the differential diagnosis based on imaging using only features with good ($ICC > 0.75$) or excellent ($ICC > 0.90$) reliability; and using ComBat harmonization per manufacturer or per protocol (manufacturer and high/low slice thickness). For each metric, the mean and 95% confidence interval over the 100x random-split cross-validation iterations are given.

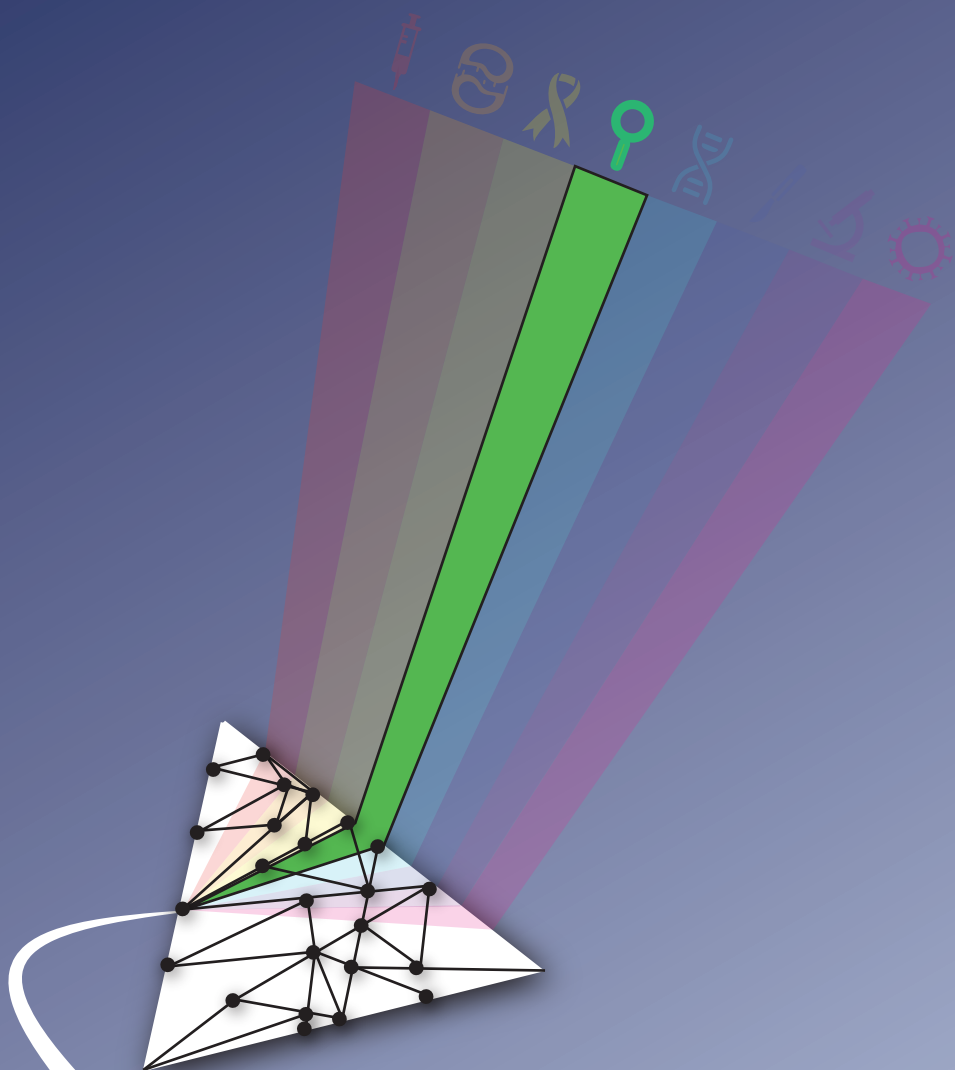
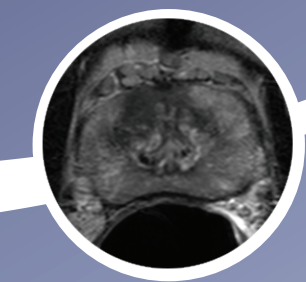
	ICC >0.75	ICC >0.90	ComBat - Manufacturer	ComBat - Protocol
AUC	0.76 [0.70, 0.82]	0.75 [0.69, 0.82]	0.76 [0.69, 0.82]	0.73 [0.66, 0.80]
BCA	0.70 [0.64, 0.76]	0.69 [0.63, 0.75]	0.70 [0.65, 0.75]	0.67 [0.61, 0.73]
Sensitivity	0.64 [0.52, 0.76]	0.59 [0.49, 0.70]	0.65 [0.56, 0.74]	0.60 [0.49, 0.71]
Specificity	0.75 [0.66, 0.84]	0.79 [0.71, 0.88]	0.75 [0.65, 0.85]	0.75 [0.66, 0.84]

*Abbreviations: AUC: area under the receiver operating characteristic curve; BCA: balanced classification accuracy.

Table 7.A.3: Overview of the 564 features used in this study. GLCM features were calculated in four different directions (0, 45, 90, 135 degrees) using 16 gray levels and pixel distances of 1 and 3. LBP features were calculated using the following three parameter combinations: 1 pixel radius and 8 neighbours, 2 pixel radius and 12 neighbours, and 3 pixel radius and 16 neighbours. Gabor features were calculated using three different frequencies (0.05, 0.2, 0.5) and four different angles (0, 45, 90, 135 degrees). LoG features were calculated using three different widths of the Gaussian (1, 5 and 10 pixels). Vessel features were calculated using the full mask, the edge, and the inner region. Local phase features were calculated on the monogenic phase, phase congruency and phase symmetry.

Histogram (13 features)	LoG (13*3=39 features)	Vessel (12*3=39 features)	GLCM (MS) (6*3*4*2=144 features)	Gabor (13*3*4=156 features)	NGTDM (5 features)	LBP (13*3=39 features)
min	min	min	contrast (normal, MS mean + std)	min	busyness	min
max	max	max	dissimilarity (normal, MS mean + std)	max	coarseness	max
mean	mean	mean	homogeneity (normal, MS mean + std)	mean	complexity	mean
median	median	median	angular second moment (ASM) (normal, MS mean + std)	median	contrast	median
std	std	std	energy (normal, MS mean + std)	std	strength	std
skewness	skewness	skewness	correlation (normal, MS mean + std)	skewness		skewness
kurtosis	kurtosis	kurtosis		kurtosis		kurtosis
peak	peak	peak		peak		peak
peak position	peak position	peak position		peak position		peak position
range	range	range		range		range
energy	energy	energy		energy		energy
quartile range	quartile range	quartile range		quartile range		quartile range
entropy	entropy	entropy		entropy		entropy
GLSZM (16 features)	GLRM (16 features)	GLDM (14 features)	Shape (35 features)	Orientation (9 features)	Local phase (13*3=39 features)	
Gray Level Non Uniformity	Gray Level Non Uniformity	Dependence Entropy	compactness (mean + std)	theta_x	min	
Gray Level Non Uniformity Normalized	Gray Level Non Uniformity Normalized	Dependence Non-Uniformity	radial distance (mean + std)	theta_y	max	
Gray Level Variance	Gray Level Variance	Dependence Non-Uniformity Normalized	roughness (mean + std)	theta_z	mean	
High Gray Level Zone Emphasis	High Gray Level Run Emphasis	Dependence Variance	convexity (mean + std)	COM index x	median	
Large Area Emphasis	Long Run Emphasis	Gray Level Non-Uniformity	circular variance (mean + std)	COM index y	std	
Large Area High Gray Level Emphasis	Long Run High Gray Level Emphasis	Gray Level Variance	principal axes ratio (mean + std)	COM index z	skewness	
Large Area Low Gray Level Emphasis	Long Run Low Gray Level Emphasis	High Gray Level Emphasis	elliptic variance (mean + std)	COM x	kurtosis	
Low Gray Level Zone Emphasis	Low Gray Level Run Emphasis	Large Dependence Emphasis	solidity (mean + std)	COM y	peak	
SizeZoneNonUniformity	RunEntropy	Large Dependence High Gray Level Emphasis	area (mean, std, min + max)	COM z	peak position	
SizeZoneNonUniformityNormalized	RunLengthNonUniformity	Large Dependence Low Gray Level Emphasis	volume (total, mesh, volume)		range	
SmallAreaEmphasis	RunLengthNonUniformityNormalized	Low Gray Level Emphasis	elongation		energy	
SmallAreaHighGrayLevelEmphasis	RunPercentage	Small Dependence Emphasis	flatness		quartile	
SmallAreaLowGrayLevelEmphasis	RunVariance	Small Dependence High Gray Level Emphasis	least axis length		entropy	
ZoneEntropy	ShortRunEmphasis	Small Dependence Low Gray Level Emphasis	major axis length			
ZonePercentage	ShortRunHighGrayLevelEmphasis		minor axis length			
ZoneVariance	ShortRunLowGrayLevelEmphasis		maximum diameter 3D			
			maximum diameter 2D (rows, columns, slices)			
			sphericity			
			surface area			
			surface volume ratio			

* Abbreviations: COM: center of mass; GLCM: gray level co-occurrence matrix; MS: multi slice; NGTDM: neighborhood gray tone difference matrix; GLSZM: gray level size zone matrix; GLRLM: gray level run length matrix; LBP: local binary patterns; LoG: Laplacian of Gaussian; std: standard deviation.



8.

A multi-center, multi-vendor study to evaluate the generalizability of a radiomics model for classifying prostate cancer: high grade vs. low grade

Based on: J. M. Castillo T, **M. P. A. Starmans**, M. Arif, W. J. Niessen, S. Klein, C. H. Bangma, I. G. Schoots, and J. F. Veenland, "A multi-center, multi-vendor study to evaluate the generalizability of a radiomics model for classifying prostate cancer: High grade vs. low grade," *Diagnostics*, vol. 11, no. 2, p. 369, 2 Feb. 2021. doi: [10.3390/diagnostics11020369](https://doi.org/10.3390/diagnostics11020369)

Abstract

Radiomics applied in MRI has shown promising results in classifying prostate cancer lesions. However, many papers describe single-center studies without external validation. The issues of using radiomics models on unseen data have not yet been sufficiently addressed. The aim of this study is to evaluate the generalizability of radiomics models for prostate cancer classification and to compare the performance of these models to the performance of radiologists. Multiparametric MRI, photographs and histology of radical prostatectomy specimens, and pathology reports of 107 patients were obtained from three healthcare centers in the Netherlands. By spatially correlating the MRI with histology, 204 lesions were identified. For each lesion, radiomics features were extracted from the MRI data. Radiomics models for discriminating high-grade (Gleason score ≥ 7) versus low-grade lesions were automatically generated using open-source machine learning software. The performance was tested both in a single-center setting through cross-validation and in a multi-center setting using the two unseen datasets as external validation. For comparison with clinical practice, a multi-center classifier was tested and compared with the Prostate Imaging Reporting and Data System version 2 (PIRADS v2) scoring performed by two expert radiologists. The three single-center models obtained a mean AUC of 0.75, which decreased to 0.54 when the model was applied to the external data, the radiologists obtained a mean AUC of 0.46. In the multi-center setting, the radiomics model obtained a mean AUC of 0.75 while the radiologists obtained a mean AUC of 0.47 on the same subset. While radiomics models have a decent performance when tested on data from the same center(s), they may show a significant drop in performance when applied to external data. On a multi-center dataset our radiomics model outperformed the radiologists, and thus, may represent a more accurate alternative for malignancy prediction.

8.1 Introduction

Prostate cancer (PCa) is the most common malignancy and second leading cause of cancer-related death in men [196]. From all patients diagnosed with PCa, those with low-grade lesions might be candidates for active surveillance, whereas patients with high-grade PCa require treatment [197]. The gold standard for PCa assessment in current clinical practice is histopathological verification of biopsy cores [197]. These cores are evaluated by a pathologist and assigned a grade using the Gleason score (GS). However, this procedure has shown to be susceptible to under-diagnosis of high-grade PCa and over-diagnosis of low grade PCa [198].

Multi-parametric magnetic resonance imaging (mpMRI) has received increasing interest for diagnosing, monitoring and treatment follow up for PCa. MpMRI allows noninvasive visualization of the whole prostatic tissue and extraction of quantitative parameters such as tissue density and permeability. To evaluate mpMRI, radiologists use the Prostate Imaging Reporting and Data System (PIRADS) v2, with a grading scale from one (highly unlikely to be clinically significant prostate cancer) to five (highly likely to be clinically significant prostate cancer) [10]. Nevertheless, mpMRI interpretation is challenging and prone to inter- and intra-reader variability among expert radiologists [198].

By extracting multiple imaging features, radiomics has the potential to evaluate the mpMRI data in a more objective way. In the context of PCa, the literature has shown evidence of the potential of radiomics in classifying PCa lesions [80, 199, 200, 201], with promising performances in terms of sensitivity and specificity [202]. Nevertheless, current studies on prostate MRI radiomics still lack the quality required to allow their introduction in clinical practice [202, 203]. This is due to the fact that most of the radiomics studies validated their approach by splitting their original dataset in training and validation subsets, while only a few studies performed a validation using an external set [204, 205, 206]. The latter evaluation is more relevant for a clinical context, where new data can present variations that were not taken into account when the original model was created. Three sources of variations can be identified: at the patient level, at the level of the MRI scanner, and at the level of the clinician. At the patient level: a model created with patient data collected in a specialized treatment centre, will differ from a model based on data collected in a hospital with a surveillance function. Magnetic resonance (MR) images vary between vendors and between scanner types from the same vendor, even if the same acquisition parameters are used. Current evidence shows that is possible to overcome these differences by applying feature harmonization techniques [207]. These techniques aim to estimate the statistical differences between imaging features computed from different data sets and apply a correction for it. To our knowledge there is no scientific evidence reporting the usage of feature harmonization in the context of PCa classification. At the clinician level: the pathologist reports, which are used as ground truth for the model, are based on the visual Gleason grading of pathologists, who are prone to considerable inter-observer variation [208, 209]. Therefore, the question arises what performance can be expected when testing radiomics models on unseen multi-center multi-vendor data: how generalizable are radiomics model in the context of PCa? The number of studies addressing

generalizability is limited. To our knowledge, few studies tested their model's generalizability for PCa detection regarding tumor aggressiveness using multiple scanners [210, 211, 212]. Only a few studies have validated their methods using external datasets for PCa tumor grade prediction [202]. When radiomics models are being considered as decision support tools for clinical practice, the generalizability issue should be addressed.

The main contribution of this study is two-fold. First, we assessed the generalizability of a radiomics approach for classifying PCa in a multi-center, multi-vendor setting. Second, in the same setting we compared the classification performance of radiologists to the performance of our radiomics model.

8.2 Material and methods

Our patient cohort was obtained from three healthcare centers in the Netherlands in the context of the Prostate Cancer Molecular Medicine project (PCMM), in Table 1 some of the clinical variables of this set are summarized. A Kruskal–Wallis test was performed to check whether the median of the GS distribution, volume, and prostatic specific antigen (PSA) of the included data sets were comparable.

The data usage of this study was approved by the medical ethics review committee of Erasmus MC under the number NL32105.078.10. In this PCMM-project, the mpMRI and pathology data of men with localized PCa who were scheduled for prostatectomy were prospectively collected from 2011 to 2014. In this study, we will refer to the data from the respective centers as data set A, B and C. The data of each center were visually graded by a radiologist and a pathologist working at that center. In total we included 107 patients for whom MRI, pathology images and reports were available. The distribution was as follows: A = 29, B = 38 and C = 40, the details regarding the MRI scanners and acquisition parameters of each set are described in Section 8.A. The dataset shows considerable variability, with images acquired with scanners from three different vendors, using various voxel sizes and b values for the diffusion weighted sequences. In deriving our radiomics models we included the T2-weighted (T2w) and the diffusion weighted imaging (DWI) sequences and the apparent diffusion coefficient maps (ADC) derived from the DWI images.

All 107 patients had their prostate surgically removed. After the prostatectomy, the prostate was cut into 3 mm thick slices. Of the top of each slice, a photograph was taken, and 4μm coupes were cut and stained with H&E. Based on the H&E, the pathologist marked the areas with cancerous tissue on the photographs and assigned a GS to each tumor region. In Figure 1 the number of lesions per GS found in each set is summarized. We grouped lesions with a $GS \leq 6$ as low-grade tumors and lesions with a $GS \geq 7$ as high-grade tumors. Out of the 107 patients, 204 lesions in total were processed, 92 (45%) low-grade and 112 (55%) high-grade. The methods used to correlate the lesions found in the pathology with MRI are explained in the following section.

Table 8.1: Prostate Cancer Molecular Medicine (PCMM) data set clinical variables and lesions characteristics. PIRADS grading performed by radiologist 1 (R1) and 2 (R2). Age of patients for data sets B and C was not available (NA). .

Center	A	B	C
Number of Patients	29	38	40
Age at Diagnosis (mean ± std years)	64 ± 7	NA	NA
PSA before treatment (mean ± std ng/mL)	12 ± 10	9 ± 5	10 ± 8
Lesions Characteristics			
Number of lesions	204		
Lesion location			
PZ	33	59	45
TZ	15	23	26
AFS	NA	2	1
Lesion volume (median and IQR mL)	1.6 (0.2–1.8)	1.4 (0.1–1.5)	0.8 (0.2–1.1)
Radiologist PIRADS grading		R1	R2
I		0	4
II		16	9
III		21	36
IV		33	34
V		43	61
Total		113	144

*Abbreviations: PZ: Peripheral zone. TZ: transition zone. AFS: anterior fibromuscular stroma. IQR: interquartile range.

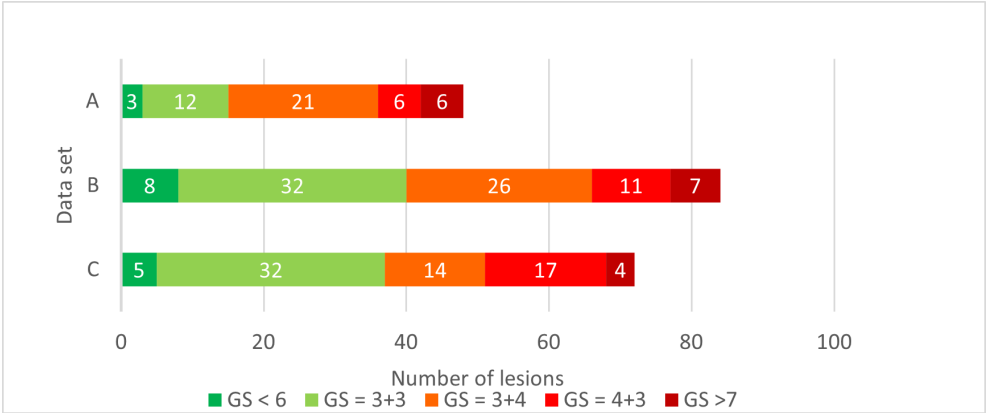


Figure 8.1: Distribution of Gleason grading of identified lesions at radical prostatectomy specimen of three different centers. The number of lesions per group is shown in white.

8.2.1 Ground truth construction: Pathology-MRI correlation

A mask of identified lesions based on microscopy analysis (H&E staining) was manually drawn by a pathologist on the prostatectomy specimens' photos. Using in house software implemented in Mevislab (v-2.2.1, Germany) [213], the macroscopy images of the prostatectomy specimen were manually registered and stacked to generate a prostate volume to enable the registration with MRI. Then, based on the prostate borders, prostate masks were manually drawn on the MR and macroscopy images. Afterwards, these two masks were manually aligned in 3D by rotation, translation, and scaling of the pathology volume. Subsequently, the translation in slice-direction was fine-tuned while inspecting the pathology and the corresponding T2w slices. As the last step, the lesion segmentation from the pathology volume was overlaid on the T2w volume.

8.2.2 Image pre-processing

In order to address the variation in image resolution between and within data sets, the MR images were resampled to a voxel grid of $0.27 \text{ mm} \times 0.27 \text{ mm} \times 3 \text{ mm}$, which was the spacing used in the largest proportion (36%) of the T2w images.

8.2.3 Radiomics generalizability evaluation

To assess the generalizability of our radiomics models, we used the experimental setup as shown in Figure 2. Image data from a single center was used to train a radiomics classifier for each center. On this training set, an $100\times$ internal random-split cross-validation was used to assess the single center performance. Finally, the model was evaluated using the other two sets to assess the generalizability; this procedure was repeated with each set. The details regarding the development of the radiomics classifiers are explained in the following section.

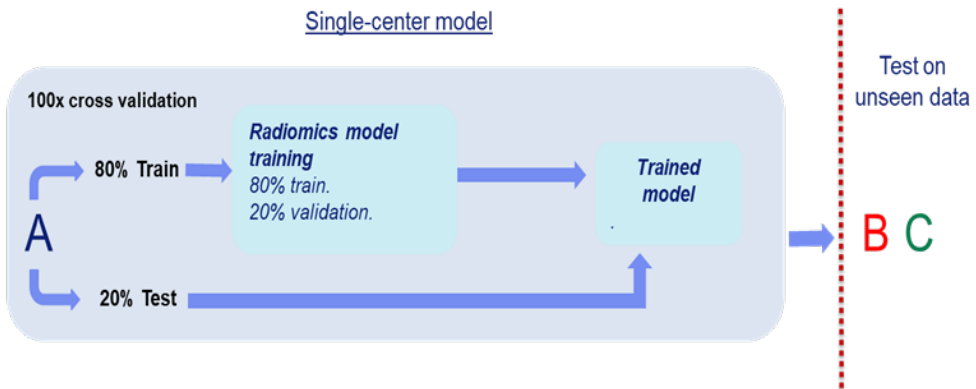


Figure 8.2: Scheme of the generalization experiment setting. In this example dataset A is used to develop a model. The model is tested on the other two sets (B and C)..

To generate the radiomics classifiers for each data set, we used the open-source Workflow for Optimal Radiomics Classification (v-3.3.2, Rotterdam, The Netherlands) platform (WORC) with the default settings [36] and another setting including feature harmonization with ComBat [185]. WORC performs an automatic search amongst a wide variety of algorithms and their corresponding parameters to determine the optimal combination that maximizes the prediction performance on the training set, a schematic overview of the method is shown in Figure 3. The workflow starts with the user defining a region of interest (ROI) from the image, which in our case was the delineation obtained by the pathology–MRI correlation. Within these tumor masks, features quantifying intensity, shape, texture and orientation were extracted from the T2w, ADC and the highest b-value image available from the DWI images. Following feature extraction, a decision model was created, which in WORC consist of several steps, such as feature selection, oversampling and machine learning methods. WORC automatically optimizes the radiomics pipeline: during each iteration WORC generates 100,000 workflows by using different combinations of methods and parameters. At the end of each cross validation, the 50 best performing solutions were combined in an ensemble as a single classification model. The final ensemble of 50 classifiers is the resulting radiomics model, the performance of which is evaluated on the independent test set (external evaluation). Feature selection was done to select the most predictive features through enabling/disabling entire families of features (e.g., shape, local binary patterns, texture based on grey-level co-occurrence matrices). The code utilized for these experiments is available online in a GitHub repository [214].

8.2.4 Radiomics classifier evaluation

The internal evaluation of the model was performed by using a $100\times$ random-split cross validation: First, the data set was split into 80% for training and 20% for testing. After this, 20% of the training set was used as validation set. This validation set was used in each training iteration to select the best parameters in order to optimize the prediction accuracy. The remaining 20% was used for performance evaluation: area under the curve (AUC), receiver operating characteristic (ROC) curve, sensitivity, and specificity. The high-grade tumors were considered the positive class. To compute the 95% confidence intervals (CI) in the cross-validation experiment, we used the corrected resampled t-test [64]. ROC confidence bands were constructed using fixed-width bands [67].

To analyze the impact of having multiple lesions from the same patient, we performed the external evaluation both at the lesion and patient level. At the patient level, for each patient only the highest grade lesion was taken into account.

8.2.5 Comparison of our radiomics model with the clinical assessment using PIRADS v2

To compare the classification performance of a multi-center radiomics model with the clinical assessment using the PIRADS v2 score, a test set was evaluated by both radiomics and the radiologist, see Figure 4. The PIRADS scoring of the lesions was

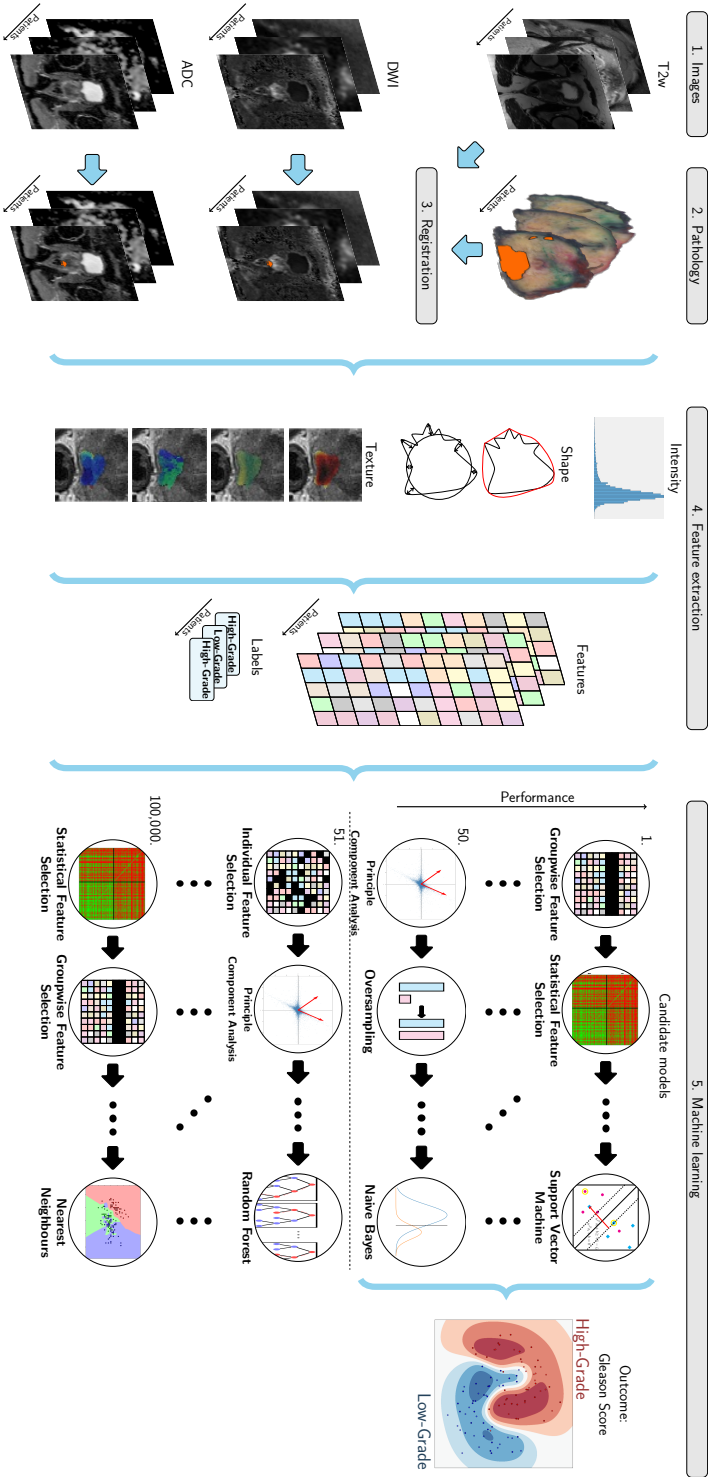


Figure 8.3: (1) The magnetic resonance sequences to be used in the model are defined. (2) The lesions from the pathology are copied and registered to the T2w sequence. (3) The diffusion weighted imaging (DWI) and apparent diffusion coefficient (ADC) are resampled and registered to the T2w. (4) Features are extracted from the T2w, DWI and ADC. (5) A radiomics model is created from the features, using an ensemble of the best 50 workflows from 100,000 candidate workflows, where the workflows are different combinations of the different classifiers.

done by two radiologists with 4 years and 10 years of experience, respectively, from of the partaking centers A and B, fully blinded from histopathology results. The lesions graded as having a PIRADS ≥ 3 were considered positive for high-grade PCa and the lesions with a score ≤ 2 as negative for high-grade PCa.

For this experiment, in order to avoid a bias towards a single center, we created a test set (D) by randomly selecting 20% of the data from each of the three centers. From this set, the lesions that were not detected by one of the two radiologists were removed from the study since our goal was to compare the classification performance, not the detection rate. Subsequently, the remaining patient data (ABC*) was used to train a radiomics model to classify the patients in set D. The end performance for either radiologist and the radiomics model was computed on patient level classification.

8.3 Results

Statistical analysis of clinical variables The median of the Gleason Score (H = 4.63, $p = 0.09$), the lesion volume (H = 5.85, $p = 0.06$) and PSA (H = 1.99, $p = 0.36$) were similar for the three data sets.

Radiomics model generalizability Table 2 shows the results for the generalizability test. Overall, it can be seen that even though reasonable performances in terms of AUC (mean = 0.75) were obtained from the internal cross-validations, when the models were tested on the other data sets, the performances dropped considerably (mean AUC = 0.54). The inclusion of feature harmonization with ComBat did not improve the performance of the radiomics models. The performance metrics on the external validation sets were comparable when evaluated lesion and patient wise. Meanwhile, radiologists' performance (mean AUC = 0.47) shows high sensitivity with a low specificity.

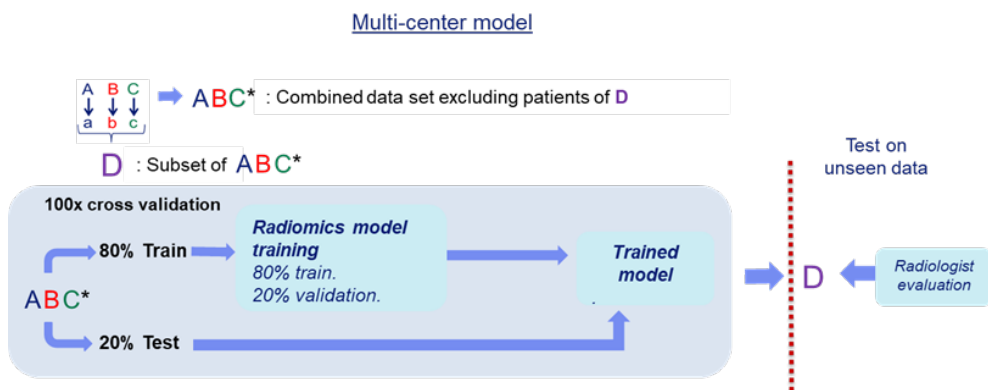


Figure 8.4: Scheme of the comparison experiment of our multi-center radiomics model with the evaluation by the radiologist. A randomly selected of patients in ABC was set apart as test set (D), the rest of the data (ABC*) was used to develop the multi-center radiomics model.

Table 8.2: Generalization study results. Internal: internal evaluation was performed using a 100× random-split cross-validation, reported with confidence interval. External: by training in one dataset, testing on the two remaining datasets.

Model	Internal	External LC	External CH	External PC	R1 and R2
Trained on A	A	B and C			
AUC	0.75 (0.58–0.92)	0.43	0.49	0.55	0.44
Sensitivity	0.91 (0.82–1.00)	0.80	0.78	0.81	0.80
Specificity	0.30 (0.03–0.55)	0.22	0.27	0.21	0.06
Trained on B	B	A and C			
AUC	0.69 (0.57–0.81)	0.60	0.57	0.55	0.50
Sensitivity	0.64 (0.47–0.80)	0.43	0.74	0.86	0.88
Specificity	0.67 (0.50–0.83)	0.62	0.38	0.25	0.13
Trained on C	C	A and B			
AUC	0.80 (0.68–0.92)	0.60	0.62	0.65	0.44
Sensitivity	0.74 (0.66–0.86)	0.52	0.51	0.48	0.69
Specificity	0.66 (0.50–0.82)	0.63	0.69	0.63	0.19

*Abbreviations: LC: lesion level classification. PC: patient level classification. AUC: area under the curve. CH: Test result using ComBat feature harmonization. R1 and R2: radiologist 1 and 2.

8.3.1 Comparison of Our Radiomics Model with the Clinical Assessment using PIRADS v2

The resulting test set was composed of 16 patients with high-grade lesions and eight patients with low-grade lesions. Table 3 presents the results of the classification performance for the internal cross-validation and the performance on the test set (ABC*) for the model and the two radiologists. It can be seen that the radiomics model outperformed (AUC = 0.75) the radiologist classification with the PIRADS score (AUC of 0.50 and 0.44). Radiologists achieved a decent sensitivity (0.76 and 0.88), but near-zero specificity (0.25 and 0.0), whereas the radiomics model achieved a sensitivity of 0.88 and a specificity of 0.63.

Table 8.3: Performance comparison of the multi-center radiomics model with the PIRADS score performed by two radiologists. Internal: internal cross validation results reported with confidence intervals. Model: results from the multi-center model for the unseen data.

Metrics	Internal	Model	R1	R2
AUC	0.72 (0.64–0.79)	0.75	0.50	0.44
Sensitivity	0.76 (0.66–0.89)	0.88	0.76	0.88
Specificity	0.55 (0.44–0.66)	0.63	0.25	0.00

*Abbreviations: AUC: area under the curve; R1 and R2: radiologist 1 and 2, respectively.

8.4 Discussion

The expanding usage of prostate MRI for PCa diagnosis has brought an increased interest in radiomics research for tumor classification. As a result, many approaches have been proposed, and promising results have been presented, thus raising the opportunity of using these models in daily clinical workflow. However, there is limited evidence regarding the performance of these models with unseen data in a new clinical contexts, for instance with MR scanners from different vendors and/or grading by different pathologists and/or different patient profiles. Investigating how these changes affect radiomics performance is required prior to applying these models in a clinical setting.

In this study we developed radiomics classifiers starting from three independent sets and evaluated the performance on the unseen data of the other centers. To compensate for the differences between data sets and reduce the negative effects on performance that these differences might have, resampled all the images in our experiments to the same voxel size, and used the same method to correlate the pathology data to the MR data. Furthermore, we applied techniques such as normalization and class unbalance correction. While obtaining a decent performance working with data from a single center, our results showed a substantial decline in performance when evaluating the radiomics models on external data. Thus, since an internal validation on a single-center dataset is not representative of external performance, it is advisable to carry out external validations to have a realistic estimation of predictive power.

The decline in performance is most probably related to several factors. One important factor affecting the feature computation is the dependency of the radiomics features on MR scanning parameters [215]. It has been shown that image normalization applied with variety of approaches or pre-filtering cannot overcome the scan-feature dependency problem [216]. Recent literature shows evidence that it is possible to overcome the scanner-feature dependency issue by applying feature harmonization techniques such as ComBat [185]. In our experiments, we applied feature harmonization using ComBat, however the inclusion of this technique did not improve our results while testing on the external sets.

Another factor is that the delineations on the pathology data were carried out by different pathologists working at the different centers. These delineations were transferred to the MRI, but the delineation is a factor that influences the feature computation [19] (i.e., Chapter 2 of this thesis), compromising the likeness of the features computed from different datasets. In clinical practice, the delineation of lesions in MRI is mostly performed by a single clinician, which makes it unfeasible to test feature robustness for several delineations. Furthermore, manual delineation by specialists is time consuming and potentially subject to observer variability. Utilizing either assisted or fully automatic segmentation methods available [217, 218] for the prostate and PCa lesions could improve feature computation consistency, important for radiomics approaches, and positively impact the model generalizability.

Various studies have assessed the use of radiomics in PCa classification on mpMRI [202]. To our knowledge, this is the first study to specifically address the generalizability of radiomics models in the context of PCa classification. Our study

consisted of multi-centric data sets: image data from multiple vendors and multiple scanners from the same vendor, two different radiologists diagnosing the patients, three different pathology departments grading histology slices of prostatectomies as ground truth. There are studies in which one factor is varied, e.g., the study published by Dinh et al. [219]. In their study they developed a model specifically for peripheral zone PCa detection, maintaining the model's performance between two MR scanners belonging to different vendors. However, in their experiments the data were acquired from the same center, evaluated, and processed by the same radiologists and pathologists. This might have affected positively the performance of their method.

When comparing our radiomics model to the PIRADS v2 scoring by radiologists, our results show that the radiologists achieved high sensitivity at the cost of a low specificity, while our model increased specificity substantially. This high sensitivity with PIRADS v2 may translate in clinical practice in overdiagnosis and overtreatment. A radiomics model may not only provide a more objective quantitative support tool to recommend surveillance for those cases where treatment may not instantly be required, but should also maintain a high sensitivity for those cases with aggressive PCa. However, it is important to take into account the data that the radiomics model was developed on, and the setting the model will be applied in. In other words, the safe utilization of a radiomics model in the clinic is feasible, as long as the population on which it is applied, holds similar characteristics to the population used to develop the model.

Our study has some limitations. First, our ground truth tumor grading is based on one pathologist per center, which can cause discrepancies in lesion delineations and grading. Having a consensus ground truth could have positively impacted our performance. However, this limitation represents current clinical practice, where the reader agreement between pathologists is between 70–80% [208, 209]. Secondly, the number of patients included per medical center is limited. However, the total number of patients in our study is higher than the average value of 80 patients found in similar radiomics studies [185]. Thirdly, the clinical assessment was performed using the PIRADS classification v2.0 because v2.1 was not available at the moment of the readings. Finally, we did not include clinical variables or epidemiological factors in our model. This information plays a role in clinical decision making, therefore, including this information may have a positive impact on the end performance in a multi-center and multi-vendor setting. Although, clinical patient information such as the level of PSA, the patient risk group and the outcome of the digital rectal examination were not available for a substantial number of patients which represented an obstacle to include these variables. Despite the previous limitations, our study contributes to the field of PCa classification using radiomics by: (1) being the first study with the generalizability of PCa classification radiomics models as main focus; (2) making our scientific code available in a public repository. As regards this last point, we would like to invite the scientific community to test this code on their own data sets and so promote discussions and future collaborations. Additionally, we would like to make some recommendations for future work: when developing a generalizable radiomics model for PCa classification the data should represent the variation present in the clinical practice with data of several centers

with various pathologists and radiologists, and multiple MRI scanners from multiple vendors. The validation of the model should be performed in a prospective cohort.

8.5 Conclusions

In this paper we assessed the generalizability of radiomics models in the context of PCa grading. When limited to a specific center or, e.g., to a specific scanner or specific setting, these models perform well and may represent a valuable tool to differentiate lowgrade from high grade tumors. However, when applying radiomics on data from different centers and/or scanners, a considerable drop in performance can be expected, making these models less reliable in this context.

To become clinical viable and support clinical decision making, training and validation of radiomics models should be performed in multi-center scenarios with data representative of the population on which the model will be applied.

Author contributions: Conceptualization, J.M.C.T., W.J.N., I.G.S. and J.F.V.; data curation, J.M.C.T., C.H.B., J.F.V.; formal analysis, J.M.C.T., M.P.A.S., M.A., W.J.N., S.K. and J.F.V.; funding acquisition, W.J.N. and J.F.V.; investigation, J.M.C.T., M.P.A.S., M.A., W.J.N., S.K., I.G.S. and J.F.V.; methodology, J.M.C.T., M.P.A.S., M.A., W.J.N., S.K., I.G.S. and J.F.V.; project administration, W.J.N. and J.F.V.; resources, J.M.C.T., W.J.N., C.H.B., I.G.S. and J.F.V.; software, J.M.C.T. and M.P.A.S.; supervision, W.J.N., I.G.S. and J.F.V.; validation, J.M.C.T., M.P.A.S., S.K., I.G.S. and J.F.V.; visualization, J.M.C.T., M.P.A.S. and J.F.V.; writing—original draft, J.M.C.T. and J.F.V.; writing—review and editing, M.P.A.S., M.A., W.J.N., S.K., C.H.B., I.G.S. and J.F.V. All authors have read and agreed to the published version of the manuscript.

Funding: This work is part of the research program strategy with project numbers 14929, 14930, and 14932, which is (partly) financed by the Netherlands organization for scientific research (NWO).

Data availability statement: Please refer to suggested Data Availability Statements in section “MDPI Research Data Policies” at <https://www.mdpi.com/ethics.data> managing tasks related to the PCMM data set.

Acknowledgments: We would like to specially acknowledge the support given by Tim Hulsen with the management and collection of the data.

Conflicts of interest: The authors declare no conflict of interest.

Appendix

Appendix 8.A Radiomics features extraction

This supplemental material is similar to Vos *et al.*, Timbergen *et al.* [72, 73] (i.e., Chapter 5 and Chapter 6 of this thesis), but details relevant for the current study are

highlighted.

A total of 540 radiomics features were used in this study. All features were extracted using Workflow for Optimal Radiomics Classification (WORC) [36], which internally uses the PREDICT [51] and PyRadiomics [44] feature extraction toolboxes. For details on the mathematical formulation of the features, we refer the reader to Zwanenburg *et al.* [39]. More details on the extracted features can be found in the documentation of the respective toolboxes, mainly the WORC documentation [68].

For CT scans, the images are by default not normalized as the scans already have a fixed unit and scale (i.e., Hounsfield), contrary to MRI. The images were not resampled, as this would result in interpolation errors. The code to extract the features has been published open-source [184].

The features can be divided in several groups. Thirteen intensity features were extracted using the histogram of all intensity values within the ROIs and included several first-order statistics such as the mean, standard deviation and kurtosis. Thirty-five shape features were extracted based only on the ROI, i.e., not using the image, and these included shape descriptions, such as the volume, compactness and circular variance. These describe the morphological properties of the lesion. Nine orientation features were used, describing the orientation of the ROI, i.e., not using the image. Lastly, 483 texture features were extracted using Gabor filters (144 features), Laplacian of Gaussian filters (36 features), vessel (i.e., tubular structures) filters (36 features) [54], the Gray Level Co-occurrence Matrix (144 features) [39], the Gray Level Size Zone Matrix (16 features) [39], the Gray Level Run Length Matrix (16 features) [39], the Gray Level Dependence Matrix (14 features) [39], the Neighbourhood Grey Tone Difference Matrix (five features) [39], Local Binary Patterns (18 features) [52], and local phase filters (36 features) [53, 195]. These features describe more complex patterns within the lesion, such as heterogeneity, occurrence of blob-like structures, and presence of line patterns.

8

Most of the texture features include parameters to be set for the extraction. Beforehand the values of the parameters that will result in features with the highest discriminative power for the classification at hand (e.g., high grade vs. low grade) are not known. Including these parameters in the workflow optimization, see [Section 8.B](#), would lead to repeated computation of the features, resulting in a redundant decrease in computation time. Therefore, alternatively, these features are extracted at a range of parameters as is default in WORC. The hypothesis is that the features with high discriminative power will be selected by the feature selection methods and/or the machine learning methods, as described in [Section 8.B](#).

The dataset used in this study is heterogeneous in terms of acquisition protocols. Especially the variations in slice may cause feature values to be dependent on the acquisition protocol. Hence, extracting robust 3D features may be hampered by these variations, especially for low resolutions. To overcome this issue, all features were extracted per 2D axial slice and aggregated over all slices, which is default in WORC. Afterwards, several first-order statistics over the feature distributions were evaluated and used in the machine learning approach.

Appendix 8.B Adaptive workflow optimization for automatic decision model creation

This appendix is similar to Vos *et al.*, Timbergen *et al.* [72, 73] (i.e., [Chapter 5](#) and [Chapter 6](#) of this thesis), but details relevant for the current study are highlighted. The Workflow for Optimal Radiomics Classification (WORC) toolbox [36] makes use of adaptive algorithm optimization to create the optimal performing workflow from a variety of methods. WORC defines a workflow as a sequential combination of algorithms and their respective parameters. To create a workflow, WORC includes algorithms to perform feature scaling, feature imputation, feature selection, over-sampling, and machine learning. If used, as some of these steps are optional as described below, these methods are performed in the same order as described in this appendix. More details can be found in the WORC documentation [68]. The code to use WORC for creating the differential diagnosis and molecular analysis decision models in this specific study has been published open-source [184].

Feature scaling was performed to make all features have the same scale, as otherwise the machine learning methods may focus only on those features with large values. This was done through z-scoring, i.e., subtracting the mean value followed by division by the standard deviation, for each individual feature. In this way, all features had a mean of zero and a variance of one. A robust version of z-scoring was used, in which outliers, i.e., values below the fifth percentile or above the 95th percentile, were excluded from computing the mean and variance.

When a feature could be computed, e.g., a lesion is too small for a specific feature to be extracted or a division by zero occurs, feature imputation was used to estimate replacement values for the missing values. Strategies for imputation included: (1) the mean; (2) the median; (3) the most frequent value; and (4) a nearest neighbor approach.

Feature selection was performed to eliminate features which were not useful to distinguish between the classes. These included: (1) a variance threshold, in which features with a low variance (<0.01) are removed. This method was always used, as this serves as a feature sanity check with almost zero risk of removing relevant features; (2) optionally, a group-wise search, in which specific groups of features (i.e., intensity, shape, and the subgroups of texture features, as defined in [Section 8.A](#), are selected or deleted. To this end, each feature group had an on/off variable which is randomly activated or deactivated, which were all included as hyperparameters in the optimization; (3) optionally, individual feature selection through univariate testing. To this end, for each feature, a Mann–Whitney U test was performed to test for significant differences in distribution between the labels. Afterwards, only features with a p-value above a certain threshold were selected. A Mann–Whitney U test was chosen as features may not be normally distributed and the samples (i.e., patients) were independent; and (4) optionally, principal component analysis (PCA), in which either only those linear combinations of features were kept which explained 95% of the variance in the features or a limited number of components (between 10 – 50). These feature selection methods may be combined by WORC, but only in the mentioned order.

Various resampling strategies can optionally be used, which can be used to

overcome class imbalances and reduce overfitting on specific training samples. These included various methods from the imbalanced-learn toolbox [57]; random over-sampling, random under-sampling, near-miss resampling, the neighborhood cleaning rule, ADASYN, and SMOTE (regular, borderline, Tomek and the edited nearest neighbors).

Lastly, machine learning methods were used to determine a decision rule to distinguish the classes. These included: (1) logistic regression; (2) support vector machines; (3) random forests; (4) naive Bayes; and (5) linear and quadratic discriminant analysis.

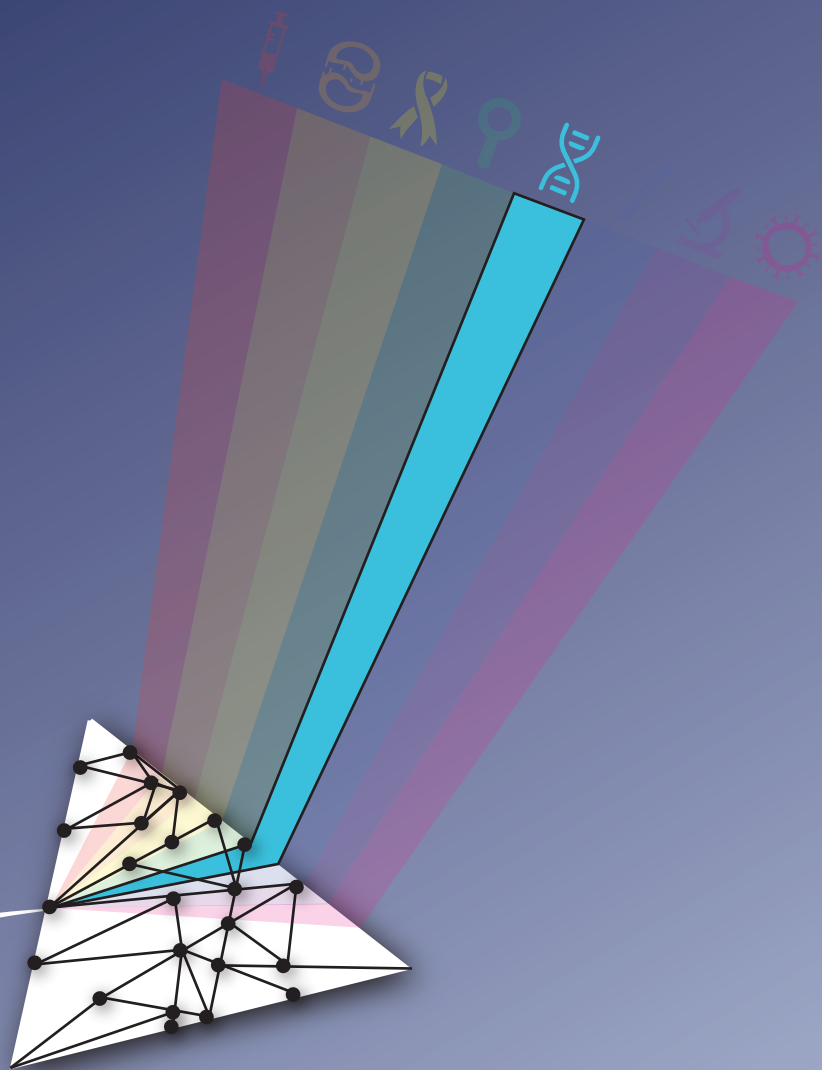
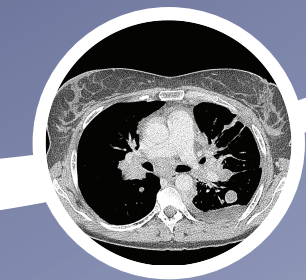
Most of the included methods require specific settings or parameters to be set, which may have a large impact on the performance. As these parameters have to be determined before executing the workflow, these are so-called “hyperparameters”. In WORC, all parameters of all mentioned methods are treated as hyperparameters, since they may all influence the decision model creation. WORC simultaneously estimates which combination of algorithms and hyperparameters performs best. A comprehensive overview of all parameters is provided in the WORC documentation [68].

By default, in WORC, the performance is evaluated in a $100\times$ random-split train-test cross-validation. In the training phase, a total of 100,000 pseudo-randomly generated workflows is created. These workflows are evaluated in a $5\times$ random-split cross-validation on the training dataset, using 80% of the data for actual training and 20% for validation of the performance. All described methods are fit on the training datasets, and only tested on the validation datasets. The workflows are ranked from best to worst based on their mean performance on the validation sets using the F1-score, which is the harmonic average of precision and recall. Due to the large number of workflows that is executed, there is a chance that the best performing workflow is overfitting, i.e., looking at too much detail or even noise in the training dataset. Hence, to create a more robust model and boost performance, WORC combines the 50 best performing workflows into a single decision model, which is known as ensembling. These 50 best performing workflows are re-trained using the entire training dataset, and only tested on the test datasets. The ensemble is created through averaging of the probabilities, i.e., the chance of lesion with high grade or low grade, of these 50 workflows. A full experiment consists of executing 50 million workflows (100,000 pseudo-randomly generated workflows, times a $5\times$ train-validation cross-validation times $100\times$ train-test cross-validation), which can be parallelized.

Table 8.A.1: Table describing scanners characteristics at the three clinical sites.

Center	Vendor	Model	Magnetic field (Tesla)	#Patients	Sequence	Voxel Size (mm)	B-values	Endorectal Coil
A	GE Medical Systems	MR750	3T	21	T2	0.37 x 0.37 x 3.00	50/400/800	No
					DWI	1.09 x 1.09 x 4.00		
	GE Medical Systems	MR450	1.5T	3	T2 DWI	0.47 x 0.47 x 3.00 1.25 x 1.25 x 4.00	100/500/1000	No
B	SIEMENS	Avanto	1.5T	5	T2	0.70 x 0.70 x 3.00	50/400/600	No
					DWI	1.85 x 1.85 x 6.00		
C	Philips Healthcare	Achieva	3T	38	T2	0.27 x 0.27 x 3.00	150/300/450/600/750	Yes
					DWI	1.03 x 1.03 x 3.00		
	SIEMENS	TrioTim Skyra	3T	17 23	T2 DWI	0.63 x 0.63 x 3.00 2.00 x 2.00 x 4.00	50/500/800	No
					T2 DWI	0.60 x 0.60 x 3.00 2.00 x 2.00 x 4.00		

*Abbreviations: GE: General Electric; T: Tesla; T2: T2-weighted sequence; DWI: Diffusion weighted imaging.



9.

The *BRAF* P.V600E mutation status of melanoma lung metastases cannot be discriminated on computed tomography by LIDC criteria nor radiomics using machine learning

Based on: L. Angus*, **M. P. A. Starmans***, A. Rajcic, A. E. Odink, M. Jalving, W. J. Niessen, J. J. Visser, S. Sleijfer, S. Klein, and A. A. M. van der Veldt, "The BRAF P.V600E mutation status of melanoma lung metastases cannot be discriminated on computed tomography by LIDC criteria nor radiomics using machine learning," *Journal of Personalized Medicine*, vol. 11, no. 4, p. 257, 4 Apr. 2021. doi: [10.3390/jpm11040257](https://doi.org/10.3390/jpm11040257)

* indicates equal contributions

Abstract

Patients with *BRAF* mutated (*BRAF*-mt) metastatic melanoma benefit significantly from treatment with *BRAF* inhibitors. Currently, the *BRAF* status is determined on archival tumor tissue or on fresh tumor tissue from an invasive biopsy. The aim of this study was to evaluate whether radiomics can predict the *BRAF* status in a non-invasive manner. Patients with melanoma lung metastases, known *BRAF* status, and a pretreatment computed tomography scan were included. After semi-automatic annotation of the lung lesions (maximum two per patient), 540 radiomics features were extracted. A chest radiologist scored all segmented lung lesions according to the Lung Image Database Consortium (LIDC) criteria. Univariate analysis was performed to assess the predictive value of each feature for *BRAF* mutation status. A combination of various machine learning methods was used to develop *BRAF* decision models based on the radiomics features and LIDC criteria. A total of 169 lung lesions from 103 patients (51 *BRAF*-mt; 52 *BRAF* wild type) were included. There were no features with a significant discriminative value in the univariate analysis. Models based on radiomics features and LIDC criteria both performed as poorly as guessing. Hence, the *BRAF* mutation status in melanoma lung metastases cannot be predicted using radiomics features or visually scored LIDC criteria.

9.1 Introduction

Cutaneous melanoma is an aggressive skin cancer most commonly occurring on the ultra-violet light exposed skin of Caucasians [220, 221]. In Europe, it is the 8th most common malignancy in men and the 5th most common in women, with an annual incidence of 144,200 new cases and 27,100 deaths [222]. In the coming years, the incidence of melanoma is expected to increase rapidly, resulting in an increased melanoma-associated mortality [223].

The introduction of new systemic treatment modalities, including immunotherapy and *BRAF* inhibitors, has significantly improved the prognosis of patients with metastatic melanoma [224]. Approximately 50% of melanomas harbor a mutation in the *BRAF* gene, with p.V600E being the most common variant [225, 226, 227]. Patients with *BRAF*-mutant (*BRAF*mt) melanoma benefit significantly from treatment with *BRAF* inhibitors and onset of response is often rapid [228]. To enhance response rates and duration of response, patients are usually treated with a combination of a *BRAF* and a MEK inhibitor [229, 230, 231, 232]. Due to the therapeutic consequences, determination of the *BRAF* mutation status in patients with metastatic melanoma is mandatory according to the European Society of Medical Oncology guidelines [233].

Currently, the *BRAF* mutation status is usually determined by molecular analysis of a metastatic lesion [222]. However, tissue biopsies are invasive, thereby exposing patients to potential risks including bleeding, infection and in case a lung biopsy is taken the risk of pneumothorax. In addition, molecular analyses can be time-consuming, especially when the tumor specimen has been archived at another hospital. Since patients with metastatic melanoma can experience rapidly progressive disease with life-threatening symptoms and an urgent medical need for systemic therapy, faster and less invasive diagnostics to determine the *BRAF* mutation status may significantly improve patient management.

Recently, various tumor characteristics have been predicted non-invasively using quantitative imaging features, also referred to as “radiomics”. In non-small cell lung cancer, radiomics on computed tomography (CT) can predict tumor stage and epidermal growth factor receptor (EGFR) mutation status [83, 234, 235, 236, 237, 238, 239, 240, 241]. In patients with primary colorectal cancer, a CT radiomics signature that was associated with *BRAF* mutation status [179]. CT-based radiomics has been applied to predict response to immunotherapy in melanoma lymph node metastases [242], but with little success (area under the curve (AUC) of 0.64). The value of radiomics for predicting *BRAF* mutation status has not been investigated. If CT-based radiomics could predict *BRAF* mutation status with a high positive predictive value, this may provide a faster and more patient-friendly alternative to determine the *BRAF* mutation status in metastatic melanoma.

The aim of this study was to evaluate the utility of CT-based radiomics to predict *BRAF* mutation status (mutant versus wild type) in metastatic melanoma. In metastatic melanoma, lung metastases are relatively easy to annotate on CT as compared to other metastases since they can be clearly distinguished from healthy lung tissue. Therefore, the aim of this study was to evaluate the utility of CT-based radiomics to predict *BRAF* mutation status (mutant versus wild type) in melanoma lung metastases.

9.2 Material and methods

9.2.1 Data collection

This study was approved by the Erasmus MC institutional research board (MEC-2019-0693). Anonymized patient data was used and therefore need for written informed consent was waived by the Institutional Review Board. All patients diagnosed with metastatic melanoma at the Erasmus MC between January 2012 and February 2018 were included retrospectively if they met the following pre-specified criteria: known tumor *BRAF* mutation, diagnostic contrast-enhanced thoracic CT scan prior to commencement of any systemic therapy, and at least one lung metastasis of ≥ 10 mm evaluable according to Response Evaluation Criteria In Solid Tumors (RECIST) v1.1 [8]. Patients with *BRAF* mutations other than p.V600E were excluded from the analysis, since *BRAF* inhibitors may be less effective in patients with other *BRAF* mutations [243]. Formalin-fixed paraffin embedded material of the primary tumor and/ or metastasis is tested for *BRAF* (exon 15) using a polymerase chain reaction based assay or next generation sequencing as part of standard care.

9.2.2 Radiomics

Lung metastases were measured according to RECIST v1.1 [8]. For 3D segmentation, up to two lung lesions ≥ 10 mm were selected by a clinician supervised by an experienced chest radiologist. In patients with >2 lung metastases of ≥ 10 mm, either the two largest or the two most easily distinguishable lesions were segmented (i.e., two separate lesions were preferred over two adjacent lesions). Using in-house developed software [105], selected lung metastases were segmented semi-automatically using a lung window for visualization. The result was visually inspected and manually corrected when necessary by an experienced chest radiologist to ensure that the semi-automatic segmentation resembled the manual segmentation. The clinician and chest radiologist were both blinded for *BRAF* mutation status. From each segmented lesion, 540 radiomics features were extracted to quantify intensity, shape and texture. Details are described in Section 9.A. To create a decision model using these features, the Workflow for Optimal Radiomics Classification (WORC) toolbox was used (Figure 9.1) [36, 72, 151]. Details are described in Section 9.A. In brief, the creation of a decision model in WORC consists of several steps, including selection of relevant features, resampling and machine learning techniques to identify patterns to distinguish *BRAF*-mt from *BRAF* wild type (*BRAF*-wt) lesions. WORC performs an automated search including a variety of algorithms for each step and determines which combination of algorithms maximizes the predictive performance on the training set. The open-source code for the feature extraction and model optimization has been published [244].

9.2.3 Scoring by radiologist

An experienced chest radiologist (certified for 8 years) scored the segmented lung lesions. There are no guidelines to differentiate histologic subtypes in lung metastases; therefore, the Lung Image Database Consortium (LIDC) criteria were used.

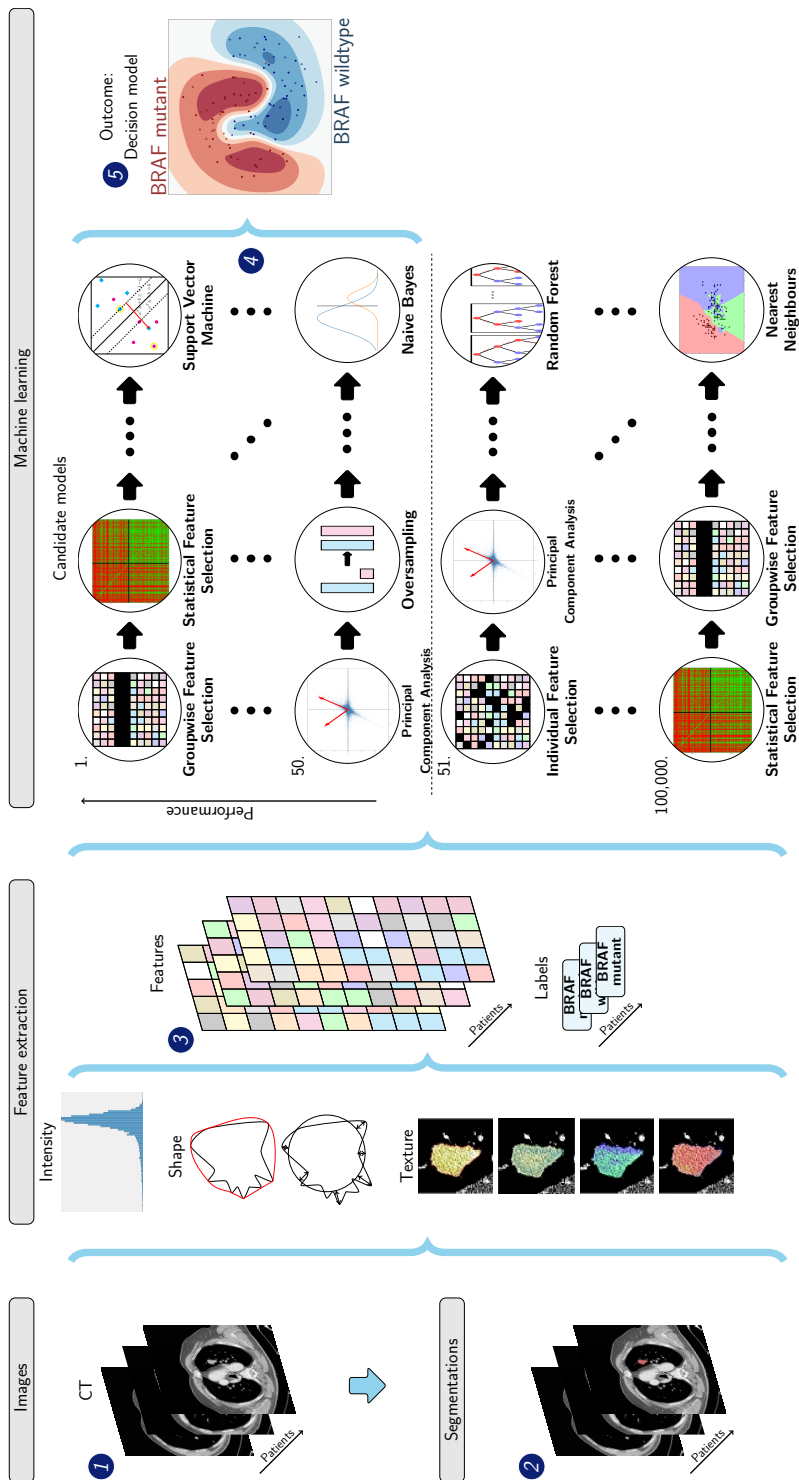


Figure 9.1: Schematic overview of the radiomics approach: adapted from Vos *et al.* [72] (i.e., Chapter 5 of this thesis). Inputs to the algorithm are (1) contrast-enhanced thoracic CT images of patients with *BRAF* mutated or *BRAF* wild type metastatic melanoma and (2) a segmentation of the lung metastasis. Processing steps include (3) feature extraction and (5) the creation of a machine learning decision model, using (4) an ensemble of the best 50 workflows from 100,000 candidate workflows, which are different combinations of the different processing and analysis steps (e.g., the classifier used).

These criteria were developed to standardize the description of radiological features of lung abnormalities in clinical practice [245]. The following LIDC features were rated: subtlety, calcification, internal structure, lobulation, likelihood of malignancy, margin, sphericity, spiculation and texture (see Table 9.A.1 for the rating system). The radiologist was blinded for the *BRAF* status, but not to the diagnosis of metastatic melanoma and had access to the CT scan, age and sex of the patient.

9.2.4 Experimental setup

To assess the predictive value of quantitative imaging features (i.e., radiomics features) and LIDC features, five models were trained and tested using WORC based on: (1) automatically extracted radiomics features only (2) similar to model 1, but only including the largest lesion per patient; (3) similar to model 1, but only including patients with *NRAS* and *BRAF* wild type melanoma for the comparison with *BRAF*-mt; (4) manually scored LIDC features only; and (5) a simple benchmark model. Model 2 was applied to assess a potential bias for patients with multiple lesions. Model 3 was included because activating *NRAS* mutations could potentially result in a similar phenotype as *BRAF*-mt, since mutations in both genes lead to activation of the mitogen-activated protein kinase (MAPK) pathway. The simple benchmark model was evaluated in a similar way as model 1, i.e., using all lesions and automatically extracted radiomics features. Model 5 was applied to compare the performance of WORC to a simple benchmark machine learning model, which uses binary logistic regression with LASSO (least absolute shrinkage and selection operator) feature selection (i.e., ElasticNet).

9.2.5 Statistics

To assess the predictive value of the individual features, the Mann-Whitney U test was performed for univariate analyses of continuous variables and Pearson's chi-squared test was used for categorical variables. For radiomics, p-values were corrected for multiple testing using the Bonferroni correction according to the default in WORC. A p-value of <0.05 was considered to be statistically significant.

Evaluation of the radiomics models was performed using a 100x random-split cross-validation. In each iteration, the data was randomly split into 80% for training and 20% for testing in a stratified manner to guarantee a similar distribution of the classes in the training and test set as compared to the original set. Metastases from the same patients were always grouped together in either the training or test set. To eliminate the risk of overfitting, in each iteration, all model optimization was performed strictly within the training set by using a second internal 5x random-split cross-validation (see Figure 9.A.1). The final model consists of an ensemble of the 50 best workflows, i.e., combination of methods and parameters, each defined by a specific set of hyperparameters. This final model may be different in each of the 100x random-split cross-validation iterations. For each of the five models described in the experimental setup, these sets hyperparameters are included with the code [244]. Details are described in Section 9.B.

The performance of all four models was described by the AUC of the receiver operating characteristic (ROC) curve, accuracy, sensitivity, specificity, negative predictive value (NPV) and positive predictive value (PPV). The positive class was defined as *BRAF*-mt. For each metric, the average over the 100 cross-validation iterations and a 95% confidence interval (CI) were reported. The 95% CIs were constructed using the corrected resampled t-test based on the results from all 100 cross-validation iterations, thereby taking into account that the samples in the cross-validation splits are not statistically independent [64]. ROC confidence bands were constructed using fixed-width bands [67].

9.3 Results

9.3.1 Study population

In total, 103 patients were included, see [Figure 9.A.2](#) for a flowchart of patient inclusion. Characteristics of these patients and their CT scans are summarized in [Table 9.1](#). The median age was 65 years (interquartile range (IQR) 52–72) and 50.5% of the patients were men. *BRAF* mutation status was either determined on the primary tumor (N = 20), local recurrence (N = 3), or metastasis (N = 79). In these lesions, *BRAF* p.V600E was detected in 51 patients, whereas 52 patients had *BRAF*-wt melanomas. In total, 103 CT scans were acquired from 10 different CT scanners, resulting in the inclusion of data acquired with different acquisition protocols ([Table 9.1](#)). Although for all acquisition parameters the difference between *BRAF*-mt and *BRAF*-wt was not statistically significant, the difference in tube current reached almost statistical significance ($p = 0.05$).

9.3.2 Radiomics and LIDC features and models

In total, 169 lung metastases in 103 patients were segmented. [Figure 9.2](#) illustrates randomly selected segmentations of lung metastases from patients with *BRAF*-mt and *BRAF*-wt metastatic melanoma. Median volume of segmented lung lesions was 18.3 mL (IQR: 7.3–48.6 mL). None of the radiomics or LIDC features were significantly different between *BRAF*-mt and *BRAF*-wt lung metastases, as none of the features had a p -value < 0.05 after Bonferroni correction. LIDC criteria scores are shown in [Table 9.2](#). Using all 169 lung metastases, the radiomics model (model 1) resulted in a mean AUC of 0.49, sensitivity of 0.61 and specificity of 0.37 ([Figure 3A](#), [Table 9.2](#)). Model 2, i.e., only inclusion of the largest lesion per patient, slightly improved the performance (AUC of 0.65), whereas model 3, i.e., only inclusion of *BRAF*-wt melanoma who were also *NRAS* wild type, still had a poor performance (AUC of 0.49) ([Figure 3B, C](#), [Table 9.2](#)). In addition, model 4, i.e., based on the LIDC features scored by a radiologist, resulted in an AUC of 0.46 ([Figure 3D](#)). The simple benchmark (model 5) resulted in a similar performance (AUC of 0.50).

Table 9.1: Patient and imaging characteristics. Values in parentheses are percentages unless indicated otherwise.

Patient	BRAF-mt (N=51)	BRAF-wt (N=52)	P-value
Age (years) [§]	59 (50-69)	66 (57-74)	0.048
Sex			0.768
Male	25 (49)	27 (52)	
Female	26 (51)	25 (48)	
Primary tumor localization			0.027
Skin	49 (96)	42 (81)	
Mucosal	0 (0)	6 (11)	
Unknown	2 (4)	4 (8)	
Determination of <i>BRAF</i> -mutation status			0.851
Primary tumor	9 (18)	11 (21)	
Local recurrence	1 (2)	2 (4)	
Metastasis	40 (78)	39 (75)	
Unknown	1 (2)	0 (0)	
<i>NRAS</i> mutation status [§]			Not determined
Mutant	-	22 (42)	
Wild type	-	23 (44)	
Unknown	-	7 (2)	
Imaging			
Acquisition protocol			
Slice thickness (mm) ^{§,1}	1.5 (1.5, 1.5)	1.5 (1.5, 1.5)	0.23
Pixel spacing (mm) [§]	0.68 (0.64, 0.74)	0.67 (0.61, 0.73)	0.16
Tube current (mA) [§]	405 (278, 553)	333 (210, 490)	0.05
Peak kilovoltage ^{§,1}	120 (120, 120)	120 (118, 120)	0.44
Contrast Agent			0.84
Visipaque 320	35	37	
Ultravist	1	0	
Omnipaque	1	1	
Optiray	0	1	
Unknown	14	13	
Number of segmented lesions per patient			0.54
One	20 (39)	17 (33)	
Two	31 (61)	35 (67)	

Values in parentheses are percentages unless stated otherwise. [†]Values are median (Inter quartile range). [§]*NRAS* and *BRAF* mutations are mutually exclusively occurring; hence, we did not test for significance between *BRAF* wild type versus mutant cases. ¹Other values than those given in the median and inter quartile range do occur.

Table 9.2: Performance of the models for *BRAF* mutation prediction based on different sets of features and lesions.

	Model 1 Radiomics All Lesions - WORC	Model 2 Radiomics Largest Lesion	Model 3 Radiomics <i>NRAS</i> Wild Type	Model 4 LIDC All Lesions	Model 5 Radiomics All Lesions - Benchmark
AUC	0.49 [0.38, 0.59]	0.65 [0.51, 0.79]	0.49 [0.37, 0.61]	0.46 [0.38, 0.55]	0.50 [0.42, 0.58]
Accuracy	0.48 [0.39, 0.57]	0.61 [0.50, 0.72]	0.65 [0.58, 0.71]	0.49 [0.42, 0.56]	0.50 [0.43, 0.57]
Sensitivity	0.61 [0.44, 0.77]	0.61 [0.42, 0.80]	0.94 [0.87, 1.00]	0.29 [0.11, 0.48]	0.56 [0.32, 0.80]
Specificity	0.37 [0.22, 0.52]	0.60 [0.38, 0.82]	0.08 [0.00, 0.17]	0.66 [0.46, 0.86]	0.44 [0.20, 0.69]
NPV	0.53 [0.39, 0.66]	0.61 [0.46, 0.76]	0.35 [0.00, 0.75]	0.52 [0.42, 0.61]	0.43 [0.21, 0.66]
PPV	0.45 [0.37, 0.53]	0.63 [0.48, 0.77]	0.67 [0.62, 0.72]	0.44 [0.30, 0.58]	0.47 [0.37, 0.56]

*Abbreviations: AUC: area under the receiver operating characteristic curve; NPV: negative predictive value; PPV: positive predictive value.

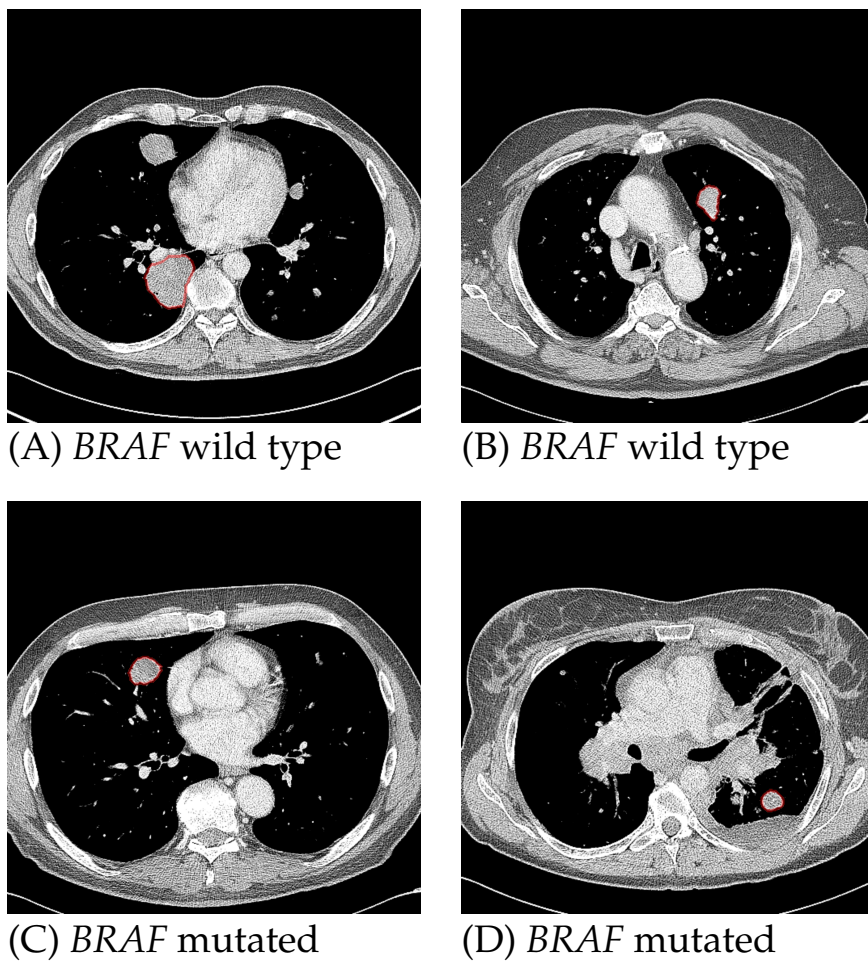


Figure 9.2: Examples of *BRAF* wild type (A,B) and *BRAF* mutant (C,D) lung metastases of four patients with metastatic melanoma. Contours of the segmentations of the selected metastases are shown in red.

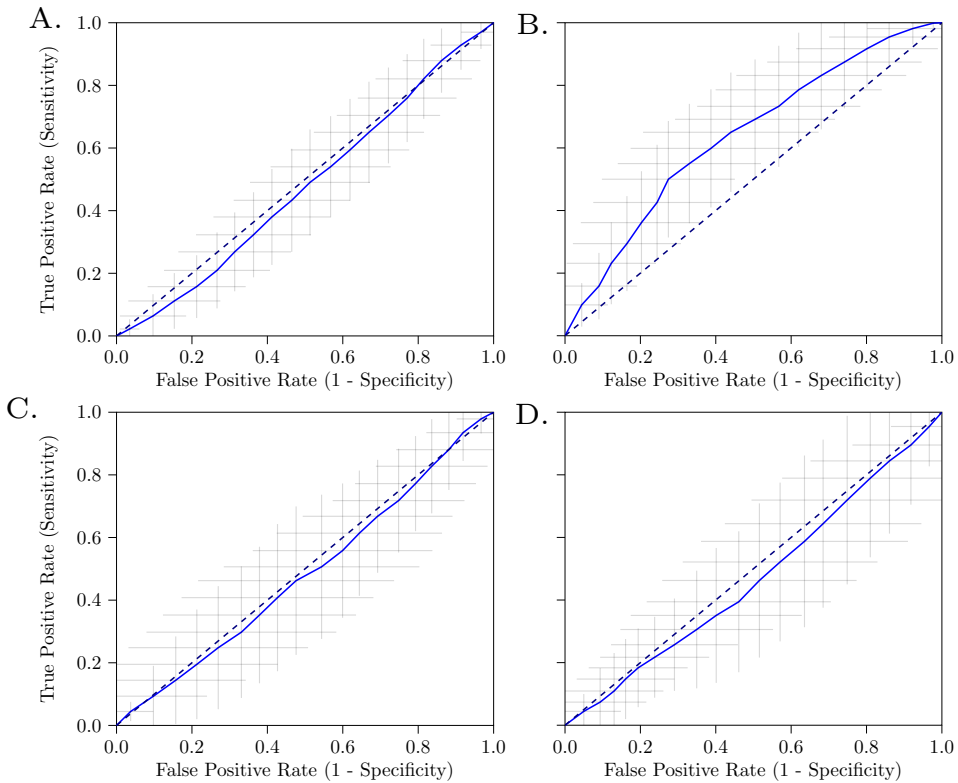


Figure 9.3: Receiver operating characteristic (ROC) curve of the radiomics model of all lesions (A), only the largest lesion (B), only *BRAF* wild type lesions with *NRAS* wild type (C) and LIDC features (D). The crosses identify the 95% confidence intervals of the 100x random-split cross-validation; the blue curve is fit through their means.

9.4 Discussion

9

The results of this study show that there is no association between radiomics features of lung metastases and the *BRAF* mutation status in patients with metastatic melanoma. Our model using only the largest lesion per patient performed best with a moderate mean AUC, but still none of the features had any individual discriminative value. In addition, the performance confidence intervals (e.g., the sensitivity and specificity) still included many values below the performance of guessing. The LIDC criteria as scored by a thorax radiologist also failed to discriminate the *BRAF* mutation status in melanoma lung metastases.

Despite the remarkable success of *BRAF* inhibitors and immunotherapy in patients with metastatic melanoma, only a subset of patients benefits from these therapies [230, 246]. Tools to select the patients most likely to benefit are of great interest and this has resulted in several radiomics studies aiming to predict tumor response. Similar to our study, previous radiomics models, either to predict therapy

response or survival, had a low to moderate performance in metastatic melanoma [242, 247, 248]. In the largest radiomics study in melanoma thus far, 483 lesions from 80 melanoma patients were included and a greater morphological heterogeneity of lymph nodes determined by CT was associated with immunotherapy response, resulting in a moderate AUC of 0.64 [242]. However, the model performed poorly in lung and liver lesions (AUC of 0.55). Comparable to our CT-based findings, a recent study showed that radiomics features derived from 18F-FDG positron emission tomography (PET) to determine the *BRAF* p.V600E mutation status also had a moderate performance (AUC of 0.62). They studied 176 lesions, including 18 lung lesions from 70 patients with melanoma (35 *BRAF*-mt and 35 *BRAF*-wt) [249]. To the best of our knowledge, this PET study [249] and our CT study are the first melanoma studies aiming to predict *BRAF* p.V600E mutation status, showing that neither PET nor CT radiomics features can discriminate between patients with *BRAF*-mt and *BRAF*-wt melanomas. We therefore believe that our comprehensive study provides insight into the potential of radiomics in this area, which can guide future research [88].

The lack of discrimination between *BRAF*-mt and *BRAF*-wt melanoma could potentially be explained by activating mutations in the *NRAS* gene in *BRAF*-wt melanoma. Since *NRAS* and *BRAF* are involved in the same pathway, i.e., the MAPK pathway, activating *NRAS* and *BRAF* mutations could result in a similar phenotype. Therefore, we evaluated an additional model which only included *NRAS* wild type lesions in patients with *BRAF*-wt melanoma (model 3). In our cohort of patients with *BRAF*-wt melanoma, 22 out of 45 (49%) patients—with known *NRAS* mutation status—had a *NRAS* mutation. Exclusion of all patients with *NRAS* mutation or unknown *NRAS* mutation status resulted in an AUC of 0.54 (95% CI 0.44–0.64). Based on these findings, it is very unlikely that inclusion of *NRAS* mutant melanomas negatively impacted our results. In addition, our findings are supported by the low predictive value of PET radiomics in the same setting in which patients with *NRAS* mutations were also excluded [249].

Our study was designed for a comprehensive evaluation of the relationship between CT imaging features and the *BRAF* mutation status in melanoma lung metastases. To our knowledge this is currently the largest CT-based radiomics study on the *BRAF* mutation status in patients with metastatic melanoma and with 103 subjects even large for a radiomics study [24]. It is unlikely that treatment-related resistance mechanisms influenced the outcome, since the study population was treatment-naïve, thereby reflecting the appearance of untreated melanoma lung metastases. The investigated patient population only included melanoma patients for whom correct determination of the *BRAF* status is of utmost importance for rapid treatment stratification. The WORC radiomics method applied has been previously validated to predict mutation status of several genes in other tumor types, such as lipoma and liposarcoma [72] (i.e., Chapter 5 of this thesis), desmoids [73] (i.e., Chapter 6 of this thesis), gastrointestinal stromal tumors [75] (i.e., Chapter 7 of this thesis), liver cancer [105, 250], prostate cancer [80] (i.e., Chapter 8 of this thesis) and mesenteric fibrosis [79] (i.e., Chapter 10 of this thesis). In these previous studies, the radiomics models had a much better performance (mean AUCs between 0.71–0.89) and multiple features were statistically significant in univariate statistical testing.

In the current study, none of the radiomics features had any discriminative value; therefore, it can be concluded that radiomics features of melanoma lung metastases are not related to the *BRAF* mutation status. WORC includes a wide variety of radiomics approaches and automatically optimizes the combination, thereby evaluating many different approaches. Moreover, a different normalization method, combining z-scoring with a logarithmic transform and a correction term to better cope with outliers and non-normally distributed features [251], yielded similar negative results (model 1: AUC of 0.49). Hence, it is unlikely that a different radiomics approach will lead to a positive result.

In addition to the radiomics analysis, a radiologist visually evaluated the lesions. Similar to radiomics results, the radiologist could not discriminate between *BRAF*-wt and *BRAF*-mt lesions by applying the LIDC criteria. Although radiomics can potentially correlate imaging features with clinical outcome even in cases where a radiologist cannot, the relation between quantitative imaging features and clinical outcome is considered stronger when clinical outcomes can be discriminated visually by a radiologist. This was not evident in the current study and it can be considered additional evidence that a CT-based radiomics signature probably does not exist for the *BRAF* mutation status in melanoma lung metastases. Although radiomics is promising in other fields of research, it is not expected that all cytogenetic changes are associated with morphological changes. Consequently, it is unlikely that every DNA alteration can be detected by radiomics.

Our study has several limitations. Firstly, the *BRAF* mutation status was often determined on other tumor tissue than the segmented lung metastases. The *BRAF* status was determined on biopsy material from a lung metastasis, which did not necessarily match the segmented lung lesion, in only 12 patients. Although the concordance rate of the *BRAF* mutation status between primary melanoma and metastases is quite high [227, 252, 253], a recent meta-analysis showed a pooled discrepancy rate of 13.4% between primary melanomas and metastases and a 7.3% discrepancy rate between metastatic sites [254]. Hence, tumor heterogeneity might have caused misclassification of *BRAF* mutation status, thereby negatively affecting the results. Ideally, in prospective radiomics studies, genomic and radiomics analyses are performed on the same tumor site. Secondly, the segmentation of regions of interest (ROI) was performed semi-automatically. Automatic segmentation methods may improve the consistency of the segmentations and thus affect the radiomics model. However, due to the clear distinction of lung lesions and their surroundings, it is not expected that automatic segmentation will substantially alter the results. Thirdly, the heterogeneity in the acquisition protocols may have negatively affected the performance of our radiomics model. These variations may have led to variations in the imaging features, which complicate the recognition of patterns. Using a single acquisition protocol would give an estimate of the performance unaffected by such variations. However, the variations in the acquisition protocols were small, making it unlikely this significantly affected the results of the current study. Feature selection methods based on feature test-retest reproducibility could be investigated in future work [18, 255]. The difference in tube current between *BRAF*-mt and *BRAF*-wt almost reached statistical significance and could have been implicitly used by the model to distinguish these lesions. However, our results show that, despite this difference, the

performance of the model was similar to guessing. Lastly, although training data were strictly separated from test data in cross-validation, we did not validate our findings on an independent, external dataset.

9.5 Conclusions

In summary, our study demonstrates that neither CT-based radiomics features, nor CT-derived LIDC features scored by a radiologist can discriminate between *BRAF* mutant and *BRAF* wild type lung metastases in patients with metastatic melanoma. Therefore, CT based parameters cannot replace determination of *BRAF* mutation status on tumor tissue.

Author contributions Conceptualization, L.A., M.P.A.S., J.J.V., S.S., S.K. and A.A.M.v.d.V.; methodology, L.A., M.P.A.S., J.J.V., S.S., S.K. and A.A.M.v.d.V.; software, M.P.A.S., S.K. and W.J.N.; formal analysis, L.A. and M.P.A.S.; investigation, L.A., A.R., M.P.A.S. and A.E.O.; resources, A.A.M.v.d.V. and S.K.; data curation, L.A. and M.P.A.S.; writing—original draft preparation, L.A., A.R. and M.P.A.S.; writing—review and editing, L.A., M.P.A.S., A.R., A.E.O., M.J., W.J.N., J.J.V., S.S., S.K., A.A.M.v.d.V.; visualization, M.P.A.S.; supervision, A.A.M.v.d.V. and S.K.; project administration, L.A.; funding acquisition, not applicable. All authors have read and agreed to the published version of the manuscript.

Funding This research received no external funding. M.P.A.S. acknowledges funding from the research program STRaTeGy (project number 14929-14930), which is (partly) financed by the Netherlands Organization for Scientific Research (NWO). This work was partially carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

Institutional review board statement The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of the Erasmus MC (MEC-2019-0693, date of approval 29 November 2019). Informed Consent Statement: Since we used anonymized patient data, written informed consent was waived by the Institutional Review Board. Data Availability Statement: Imaging and clinical research data are not available at this time. Programming code is openly available on Zenodo at <https://doi.org/10.5281/zenodo.4644067>.

Conflicts of interest L.A. reports receiving a consulting honorarium from Merck and a speaking honorarium from Pfizer; and A.A.M.v.d.V., receiving consulting honoraria from Sanofi, Roche, Merck Sharp & Dohme, Pfizer, Eisai, Ipsen, Novartis, Pierre Fabre and Bristol-Myers Squibb. W.J.N. is founder, scientific lead and stock holder of Quantib BV. M.J. reports consulting honoraria to institution from Merck, BMS, Novartis, Pierre Fabre, Tesaro, AstraZeneca and performs clinical studies with BMS, AbbVie, Merck, Cristal Therapeutics. No other potential conflict of interest relevant to this publication was reported.

Appendix

Appendix 9.A Radiomics feature extraction

This supplementary material is similar to Vos *et al.* [72] (i.e., [Chapter 5](#) of this thesis), but details relevant for the current study are highlighted.

A total of 540 radiomics features were used in this study. All features were extracted using the defaults for CT scans from the Workflow for Optimal Radiomics Classification (WORC) toolbox [36], which internally uses the PREDICT [51] and PyRadiomics [44] feature extraction toolboxes. For CT scans, the images are not normalized as the scans already have a fixed unit and scale (i.e. Hounsfield), contrary to MRI. The code to extract the features for this specific study has been published open-source [244]. An overview of all features is depicted in [Table 9.A.2](#). For details on the mathematical formulation of the features, we refer the reader to Zwanenburg *et al.* [39]. More details on the extracted features can be found in the documentation of the PREDICT, PyRadiomics, and mainly the WORC documentation [68].

The features can be divided in several groups. Thirteen intensity features were extracted using the histogram of all intensity values within the ROIs and included several first-order statistics such as the mean, standard deviation and kurtosis. These describe the distribution of Hounsfield units within the lesion. Thirty-five shape features were extracted based only on the ROI, i.e. not using the image, and included shape descriptions such as the volume, compactness and circular variance. These describe the morphological properties of the lesion. Nine orientation features were used, describing the orientation of the ROI, i.e. not using the image. Lastly, 483 texture features were extracted using Gabor filters (144 features), Laplacian of Gaussian filters (36 features), vessel (i.e. tubular structures) filters (36 features) [54], the Gray Level Co-occurrence Matrix (144 features) [39], the Gray Level Size Zone Matrix (16 features) [39], the Gray Level Run Length Matrix (16 features) [39], the Gray Level Dependence Matrix (14 features) [39], the Neighbourhood Grey Tone Difference Matrix (5 features) [39], Local Binary Patterns (18 features) [52], and local phase filters (36 features) [53]. These features describe more complex patterns within the lesion, such as heterogeneity, occurrence of blob-like structures, and presence of line patterns.

Appendix 9.B Model optimization

This appendix is similar to Vos *et al.* [72] (i.e., [Chapter 5](#) of this thesis), but details relevant for the current study are highlighted. The Workflow for Optimal Radiomics Classification (WORC) toolbox [36] makes use of adaptive algorithm optimization to create the optimal performing workflow from a variety of methods. WORC defines a workflow as a sequential combination of algorithms and their respective parameters. To create a workflow, WORC includes algorithms to perform feature scaling, feature imputation, feature selection, oversampling, and machine learning. If used, as some of these steps are optional as described below, these methods are performed in the same order as described in this appendix. More details can be found in the WORC

documentation [68]. The code to use WORC for creating the *BRAF* decision models in this specific study has been published open-source [244].

When a feature could not be computed, e.g. the lesion is too small or a division by zero occurs, feature imputation was used to estimate replacement values for the missing values. Strategies for imputation included 1) the mean; 2) the median; 3) the most frequent value; and 4) a nearest neighbor approach.

Feature scaling was performed to make all features have the same scale, as otherwise the machine learning methods may focus only on those features with large values. This was done through z-scoring, i.e. subtracting the mean value followed by division by the standard deviation, for each individual feature. In this way, all features had a mean of zero and a variance of one. A robust version of z-scoring was used, in which outliers, i.e. values below the 5th percentile or above the 95th percentile, are excluded from computing the mean and variance.

Feature selection was performed to eliminate features which were not useful to distinguish between the classes, i.e. *BRAF* mutant vs. *BRAF* wild-type. These included; 1) a variance threshold, in which features with a low variance (<0.01) are removed. This method was always used, as this serves as a feature sanity check with almost zero risk of removing relevant features; 2) optionally, a group-wise search, in which specific groups of features (i.e. intensity, shape, and the subgroups of texture features as defined in Section 9.A) are selected or deleted. To this end, each feature group has an on/off variable which is randomly activated or deactivated, which were all included as hyperparameters in the optimization; 3) optionally, individual feature selection through univariate testing. To this end, for each feature, a Mann-Whitney U test is performed to test for significant differences in distribution between the labels (e.g. *BRAF* mutant vs *BRAF* wild-type). Afterwards, only features with a p-value above a certain threshold are selected. A Mann-Whitney U test was chosen as features may not be normally distributed and the samples (i.e. lesions) were independent; and 4) optionally, principal component analysis (PCA), in which either only those linear combinations of features were kept which explained 95% of the variance in the features or a limited amount of components (between 10 – 50). These feature selection methods may be combined by WORC, but only in the mentioned order.

Oversampling was used to make sure the classes were balanced in the training dataset. These included; 1) random oversampling, which randomly repeats patients of the minority class; and 2) the synthetic minority oversampling technique (SMOTE) [58], which creates new synthetic “lesions” using a combination of the features in the minority class. Randomly, either one of these methods or no oversampling method was used.

Lastly, machine learning methods were used to determine a decision rule to distinguish the classes. These included; 1) logistic regression; 2) support vector machines; 3) random forests; 4) naive Bayes; and 5) linear and quadratic discriminant analysis.

Most of the included methods require specific settings or parameters to be set, which may have a large impact on the performance. As these parameters have to be determined before executing the workflow, these are so-called “hyperparameters”. In WORC, all parameters of all mentioned methods are treated as hyperparameters,

since they may all influence the decision model creation. WORC simultaneously estimates which combination of algorithms and hyperparameters performs best. A comprehensive overview of all parameters is provided in the WORC documentation [68].

By default in WORC, the performance is evaluated in a 100x random-split train-test cross-validation. In the training phase, a total of 100,000 pseudo-randomly generated workflows is created. These workflows are evaluated in a 5x random-split cross-validation on the training dataset, using 85% of the data for actual training and 15% for validation of the performance. All described methods were fit on the training datasets, and only tested on the validation datasets. The workflows are ranked from best to worst based on their mean performance on the validation sets using the F1-score, which is the harmonic average of precision and recall. Due to the large number of workflows executed, there is a chance that the best performing workflow is overfitting, i.e. looking at too much detail or even noise in the training dataset. Hence, to create a more robust model and boost performance, WORC combines the 50 best performing workflows into a single decision model, which is known as ensembling. These 50 best performing workflows are re-trained using the entire training dataset, and only tested on the test dataset. The ensemble is created through averaging of the probabilities, i.e. the chance of a lesion being *BRAF* mutant or *BRAF* wild-type, of these 50 workflows.

The code for the model creation, including more details, has been published open-source [244].

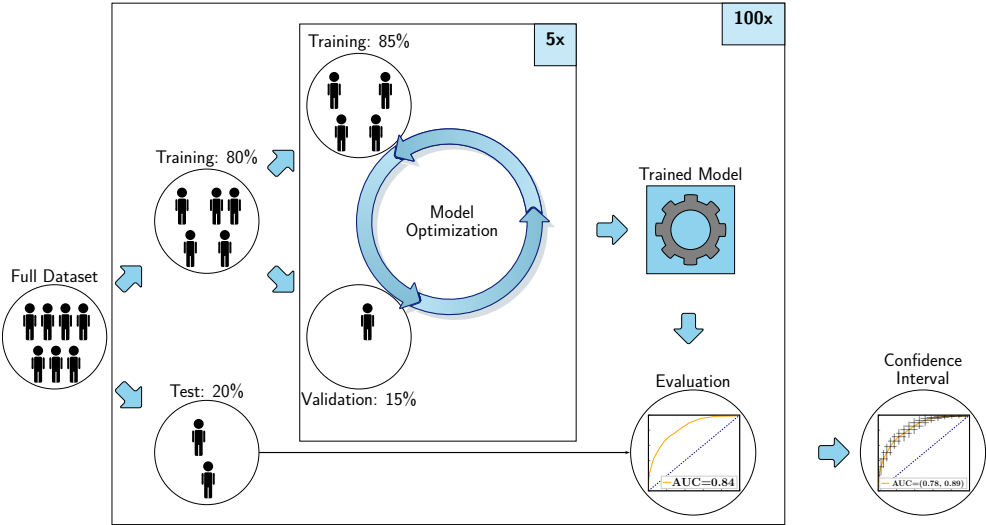


Figure 9.A.1: Visualization of the 100x random-split cross-validation, including a second 5x random-split cross-validation within the training set.

Table 9.A.1: LIDC Nodule Characteristics, Definitions, and Ratings [245].

Characteristic	Ratings	Description
Calcification (categorical)	1 Popcorn 2 Laminated 3 Solid 4 Non-central 5 Central 6 Absent	Calcification appearance in the nodule - the smaller the nodule, the more likely it must contain calcium in order to be visualized. Benignity is highly associated with central, non-central, laminated, and popcorn calcification
Internal structure (categorical)	1 Soft tissue 2 Fluid 3 Fat 4 Air	Expected internal composition of the nodule
Lobulation (ordinal)	1 Marked 2 . 3 . 4 . 5 None	Whether a lobular shape is apparent from the margin or not - lobulated margin is an indication for benignity
Malignancy (ordinal)	1 Highly unlikely 2 Moderately unlikely 3 Indeterminate 4 Moderately suspicious 5 Highly suspicious	Likelihood of malignancy of the nodule - malignancy is associated with large nodule size while small nodules are more likely to be benign. Most malignant nodules are non-calcified and have speculated margins.
Margin (ordinal)	1 Poorly defined 2 . 3 . 4 . 5 Sharp	How well defined the margins of the nodules are
Sphericity (ordinal)	1 Linear 2 . 3 Ovoid 4 . 5 Round	Dimensional shape of nodule in terms of roundness
Spiculation (ordinal)	1 Marked 2 . 3 . 4 . 5 None	Degree to which the nodule exhibits spicules, spike-like structures, along its border - spiculated margin is an indication of malignancy
Subtlety (ordinal)	1 Extremely subtle 2 Moderately subtle 3 . 4 Fairly subtle 5 Obvious	Difficulty in detection - refers to the contrast between the lung and its surroundings
Texture (ordinal)	1 Nonsolid 2 . 3 Part-solid/mixed 4 . 5 Solid	Internal density of a nodule - texture plays an important role when attempting to segment a nodule, since part-solid and nonsolid texture can increase the difficulty of defining the nodule boundary

Table 9.A.2: LIDC criteria scored by a thorax radiologist.

	<i>BRAF</i> Mutant (N=82 lesions)	<i>BRAF</i> wild type (N=87 lesions)
Calcification		
Popcorn		
Yes	0	0
No	82	87
Laminated		
Yes	0	0
No	82	87
Solid		
Yes	0	1
No	82	86
Non-central		
Yes	0	0
No	82	87
Central		
Yes	1	0
No	82	87
Absent		
Yes	75	80
No	7	7
Internal structure		
Soft tissue		
Yes	75	81
No	7	6
Fluid		
Yes	0	0
No	82	87
Fat		
Yes	0	0
No	82	87
Air		
Yes	1	1
No	81	86
Lobulation (ordinal)		
1 Marked	10	7
2	1	0
3	4	5
4	20	26
5 None	47	49
Malignancy		
Highly unlikely	8	5
Moderate unlikely	2	0
Indeterminate	0	1
Moderately suspicious	1	1
Highly suspicious	71	80
Margin (ordinal)		
1 Poorly defined	8	5
2	3	1
3	12	11
4	4	12
5 Sharp	55	58
Sphericity (ordinal)		
1 Linear	9	7
2	3	2
3 Ovoid	33	28
4	20	25
5 Round	17	25
Spiculation (ordinal)		
1 Marked	8	6
2	2	1
3	1	2
4	6	6
5 None	65	72
Subtlety		
1 Extremely subtle	7	5
2 Moderately subtle	0	0
3	0	0
4 Fairly subtle	0	1
5 Obvious	75	81
Texture		
1 Nonsolid	10	5
2	0	0
3 Part-solid/mixed	0	0
4	0	0
5 Solid	72	82

Table 9.A.3: Overview of the 540 features used in this study. GLCM features were calculated in four different directions (0, 45, 90, 135 degrees) using 16 gray levels and pixel distances of 1 and 3. LBP features were calculated using the following three parameter combinations: 1 pixel radius and 8 neighbours, 2 pixel radius and 12 neighbours, and 3 pixel radius and 16 neighbours. Gabor features were calculated using three different frequencies (0.05, 0.2, 0.5) and four different angles (0, 45, 90, 135 degrees). LoG features were calculated using three different widths of the Gaussian (1, 5 and 10 pixels). Vessel features were calculated using the full mask, the edge, and the inner region. Local phase features were calculated on the monogenic phase, phase congruency and phase symmetry.

Histogram (13 features)	LoG (12*3=36 features)	Vessel (12*3=36 features)	GLCM (MS) (6*3*4*2=144 features)	Gabor (12*4*3=144 features)	NGTDM (5 features)	LBP (12*3=39 features)
min max mean median std skewness kurtosis peak peak position range energy quartile entropy	min max mean median std skewness kurtosis peak peak position range energy quartile entropy	min max mean median std skewness kurtosis peak peak position range energy quartile entropy	contrast (normal, MS mean + std) dissimilarity (normal, MS mean + std) homogeneity (normal, MS mean + std) angular second moment (ASM) (normal, MS mean + std) energy (normal, MS mean + std) correlation (normal, MS mean + std)	min max mean median std skewness kurtosis peak peak position range energy quartile entropy	busyness coarseness complexity contrast strength	min max mean median std skewness kurtosis peak peak position range energy quartile entropy
GLSZM (16 features)	GLRM (16 features)	GLDM (14 features)	Shape (35 features)	Orientation (9 features)	Local phase (12*3=39 features)	
Gray Level Non Uniformity Gray Level Non Uniformity Normalized Gray Level Variance High Gray Level Zone Emphasis Large Area Emphasis Large Area High Gray Level Emphasis Large Area Low Gray Level Emphasis Low Gray Level Zone Emphasis SizeZoneNonUniformity SizeZoneNonUniformityNormalized Small AreaEmphasis Small AreaHighGrayLevelEmphasis SmallAreaLowGrayLevelEmphasis ZoneEntropy ZonePercentage ZoneVariance	Gray Level Non Uniformity Gray Level Non Uniformity Normalized Gray Level Variance High Gray Level Run Emphasis Long Run Emphasis Long Run High Gray Level Emphasis Long Run Low Gray Level Emphasis Low Gray Level Run Emphasis RunEntropy RunLengthNonUniformity RunLengthNonUniformityNormalized RunPercentage RunVariance ShortRunEmphasis ShortRunHighGrayLevelEmphasis ShortRunLowGrayLevelEmphasis	Dependence Entropy Dependence Non-Uniformity Dependence Non-Uniformity Normalized Dependence Variance Gray Level Non-Uniformity Gray Level Variance High Gray Level Emphasis High Gray Level Emphasis Large Dependence Emphasis Large Dependence High Gray Level Emphasis Large Dependence Low Gray Level Emphasis Low Gray Level Emphasis Small Dependence Emphasis Small Dependence High Gray Level Emphasis Small Dependence Low Gray Level Emphasis	compactness (mean + std) radial distance (mean + std) roughness (mean + std) convexity (mean + std) circular variance (mean + std) principal axes ratio (mean + std) elliptic variance (mean + std) solidity (mean + std) area (mean, std, min + max volume (total, mesh, volume) elongation flatness least axis length minor axis length major axis length maximum diameter 2D maximum diameter 3D (rows, columns, slices) sphericity surface area surface volume ratio	theta_x theta_y theta_z COM index x COM index y COM index z COM x COM y COM z	min max mean median std skewness kurtosis peak range energy quartile entropy	

*Abbreviations: COM: center of mass; GLCM: gray level co-occurrence matrix; MS: multi slice; NGTDM: neighborhood gray tone difference matrix; GLSZM: gray level size zone matrix; GLRM: gray level run length matrix; LBP: local binary patterns; LoG: Laplacian of Gaussian; std: standard deviation.

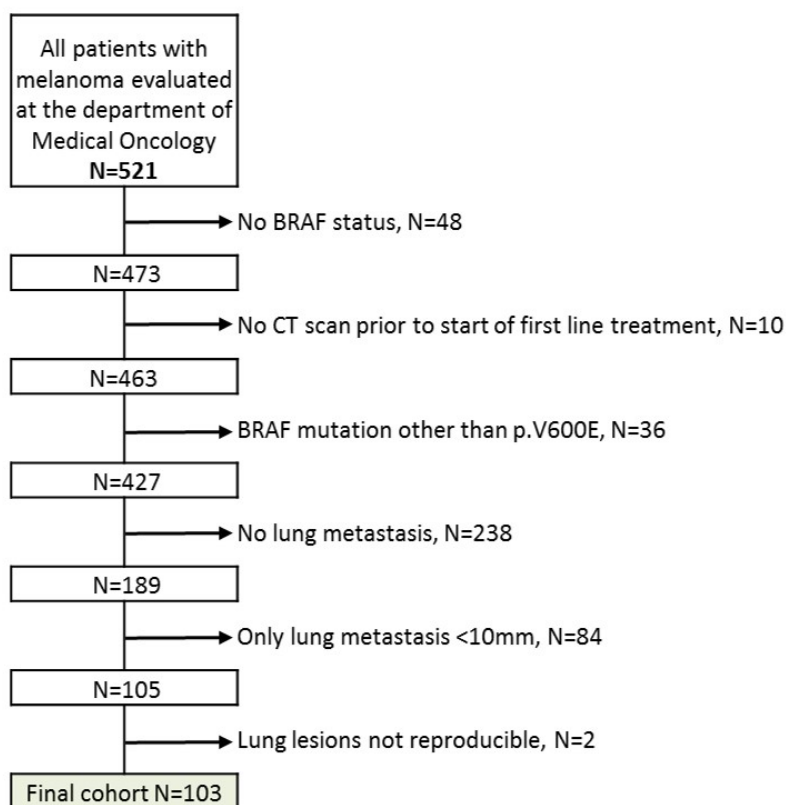
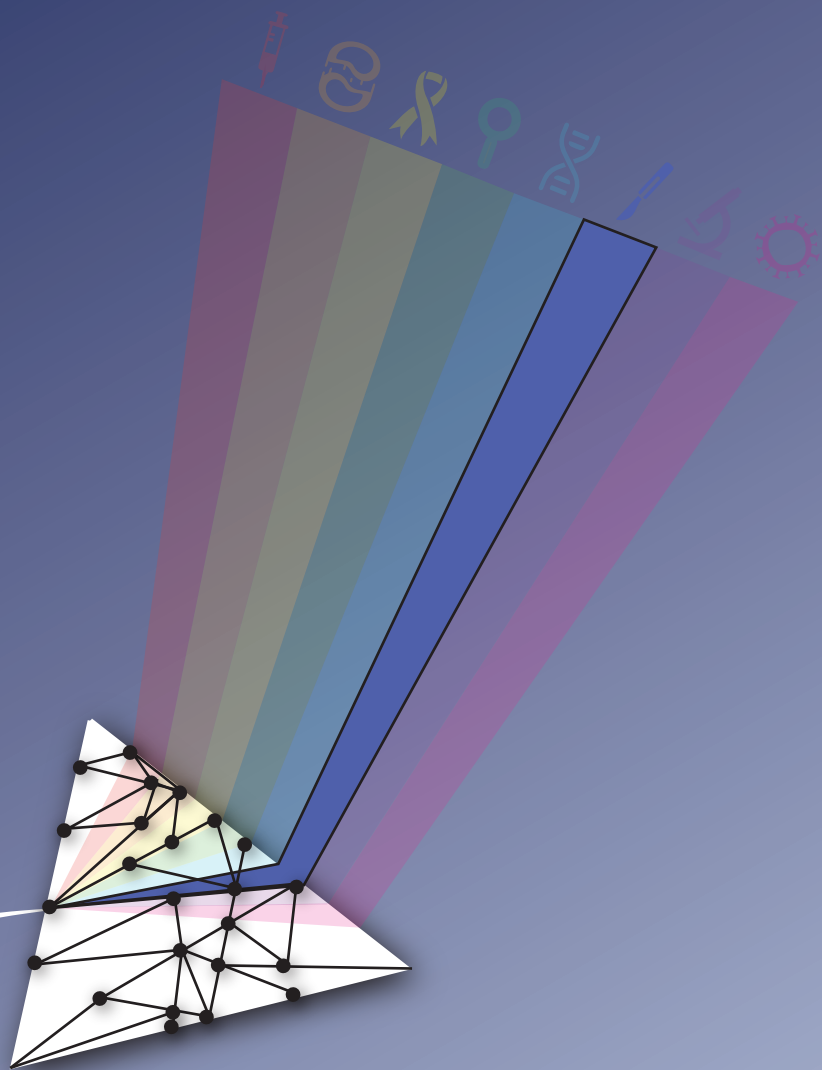


Figure 9.A.2: Flowchart of patient inclusion.



10.

Predicting symptomatic mesenteric mass in small intestinal neuroendocrine tumors using radiomics

Based on: A. Blazevic*, **M. P. A. Starmans***, T. Brabander, R. S. Dwarkasing, R. A. H. van Gils, J. Hofland, G. J. H. Franssen, R. A. Feelders, W. J. Niessen, S. Klein, and W. W. de Herder, "Predicting symptomatic mesenteric mass in small intestinal neuroendocrine tumors using radiomics," *Endocrine-Related Cancer*, vol. 28, no. 8, pp. 529–539, 8 Aug. 1, 2021. DOI: [10.1530/erc-21-0064](https://doi.org/10.1530/erc-21-0064)

* indicates equal contributions

Abstract

Metastatic mesenteric masses of small intestinal neuroendocrine tumors (SI-NETs) are known to often cause intestinal complications. The aim of this study was to identify patients at risk to develop these complications based on routinely acquired CT scans using a standardized set of clinical criteria and radiomics. Retrospectively, CT scans of SI-NET patients with a mesenteric mass were included and systematically evaluated by five clinicians. For the radiomics approach, 1128 features were extracted from segmentations of the mesenteric mass and mesentery, after which radiomics models were created using a combination of machine learning approaches. The performances were compared to a multidisciplinary tumor board (MTB). The dataset included 68 patients (32 asymptomatic, 36 symptomatic). The clinicians had AUCs between 0.62 and 0.85 and showed poor agreement. The best radiomics model had a mean AUC of 0.77. The MTB had a sensitivity of 0.64 and specificity of 0.68. We conclude that systematic clinical evaluation of SI-NETs to predict intestinal complications had a similar performance than an expert MTB, but poor inter-observer agreement. Radiomics showed a similar performance and is objective, and thus is a promising tool to correctly identify these patients. However, further validation is needed before the transition to clinical practice.

10.1 Introduction

Small intestinal neuroendocrine tumors (SI-NETs) are rare neoplasms with a mostly slow, progressive course [256]. Patients frequently present with metastasized disease, the liver and mesentery being the dominant metastatic sites [257]. SI-NETs are known to induce fibrosis, most notably surrounding a metastatic mesenteric mass, via the production of mediators like serotonin. This mesenteric fibrosis causes distortion and traction on the surrounding intestine and can encase mesenteric vessels. In the majority of patients, this leads to severe complications such as intestinal obstruction and ischemia.

In order to prevent future complications, the current European Neuroendocrine Tumor Society (ENETS) guideline advises consideration of prophylactic surgery in these patients [258]. However, not all of these patients may benefit from surgery: approximately 30% of patients with the mesenteric metastasized disease have no abdominal symptoms [259, 260]. In addition, recent studies found no survival or clinical benefit of prophylactic palliative surgery in asymptomatic patients [260, 261]. Nonetheless, it has been suggested that a certain subset of patients might benefit from early surgical intervention [258]. Often the presence of a mesenteric mass and the severity of mesenteric fibrosis are used to determine the necessity of prophylactic palliative surgery. However, there is discordance between the histological and radiological severity of mesenteric fibrosis and the symptomatology [262, 263]. To our knowledge, there is currently no method to reliably identify patients prone to develop intestinal complications due to a SI-NET mesenteric mass.

The currently developed stratification methods for SI-NETs focus solely on overall survival and prognosis and do not include risk factors for intestinal complications due to mesenteric metastasis and fibrosis [264, 265, 266]. Therefore, we propose two novel methods for the identification of complications based on contrast-enhanced abdominal CT scans. First, a visual systematic clinical evaluation of the scan. Second, a data-driven approach to identify predictive features of symptomatic mesenteric masses. To this end, we use radiomics, in which quantitative medical imaging features are related to clinical outcome. Radiomics has been used in combination with CT in various clinical applications, such as liver cancer [105], lung cancer [83], clear cell renal carcinoma [267], and many more [16]. In neuroendocrine tumors, radiomics has been used to predict the grade of pancreatic neuroendocrine tumors [268]. Given the success in these previous studies and the fact that CT scans are routinely acquired for assessing disease progression, we hypothesize that radiomics may be used to quantify the appearance of the SI-NET mesenteric mass and surrounding mesentery. Besides developing a prediction model using radiomics, further analysis of the radiomics features of symptomatic patients may elucidate new insights in the processes involved in the development of symptomatic mesenteric masses.

The aim of this study was to find a method to reliably identify patients at high risk of developing complications from a mesenteric mass and surrounding fibrosis. To this end, routinely acquired CT scans were assessed by five clinicians using systematic clinical evaluation, and a radiomics approach was used in which we assessed the predictive value of (1) the SI-NET mesenteric mass; (2) the surrounding

mesentery; and (3) the mesenteric mass location. To compare the performance with clinical practice, a multidisciplinary tumor board (MTB) evaluated the patients as well.

10.2 Materials and methods

10.2.1 Study population

This study was performed in accordance with the Dutch Code of Conduct for Medical Research of 2004. As the study was retrospectively performed with anonymized data, no approval from the ethical committee or informed consent was required. Patients were retrospectively included from the Rotterdam NET-database, which encompassed all NET patients treated between January 1993 and December 2018 in the Erasmus MC, University Medical Center Rotterdam, the Netherlands. Included cases had a pathologically proven SI-NET and radiological evidence of a metastatic mesenteric mass. A metastatic mesenteric mass was diagnosed if the lesion met the criteria of a malignant mesenteric lymph node on CT scan in accordance with the RECIST 1.1 guidelines, as these are validated criteria to determine disease progression with clear criteria for a malignant lymph node [8, 258].

Patients were included in the symptomatic group in case of palliative abdominal surgery because of intestinal complications, for example, obstruction, ischemia, or perforation. For this group, a venous phase contrast-enhanced abdominal CT scan performed up to 365 days before the surgery was used. Patients were included in the asymptomatic group when none of the mentioned intestinal complications were present, and thus no abdominal surgery was performed, for at least 3 years after the included venous phase contrast-enhanced abdominal CT scan was performed.

Due to the low quality of older scans and to make the outcome more applicable to the current CT technology, only scans between 2008 and 2018 were included. No other restrictions on the acquisition parameters or contrast administration protocol were imposed. It was recorded whether positive enteric contrast was used or not. Baseline characteristics included age, sex, tumor grade according to WHO criteria, ENETS disease stage, plasma chromogranin A (CgA) level, and 24-h urinary 5-hydroxyindoleacetic acid (5-HIAA) excretion [258].

10.2.2 Segmentation

For each patient, three regions of interest (ROIs) were segmented: (1) the mesenteric mass (MM); (2) the surrounding mesentery (SM); and (3) the origin of the superior mesenteric artery (SMA). Segmentation was performed manually per voxel by a clinician (AB) and reviewed by a nuclear physician (TB). For segmentation of the MM, a mesenteric node of ≥ 15 mm on the short axis was selected in accordance with RECIST 1.1 criteria for target lymph nodes [8]. In case of multiple pathological mesenteric nodes, the largest mesenteric node was selected, since the desmoplastic reaction occurs principally around the dominant mesenteric node [269]. The SM was segmented by annotating the mesentery between the MM and the surrounding bowel wall with a maximum distance of 30 mm between the MM and SM contour. This

cutoff was chosen instead of annotating the entire mesentery between the MM and bowel wall to reduce differences in the segmentations due to variations in mesenteric retraction across patients. Determination of the exact middle of the SMA origin, that is, one point on one slice, is difficult due to the variable and often high slice thickness (e.g. 5 mm) of the scans and is potentially observer dependent. Instead, to improve reproducibility, for all scans, the first 10 mm of the SMA branching from the abdominal aorta were manually delineated per voxel. The center of this ROI was used to calculate the location features as described in [Subsection 10.2.4](#).

10.2.3 Systematic clinical evaluation by clinicians

The criteria for the systematic clinical evaluation are shown in [Table 10.A.1](#). Fibrosis was classified as: grade 1 (<10 thin radiating strands), grade 2 (>10 thin and <10 thick radiating strands), grade 3 (>10 thick radiating strands) [269]. Mesenteric mass staging was classified as: stage I when the mesenteric mass is located close to the intestine; stage II involves arterial branches close to the origin in the mesenteric artery; stage III extends along the SMA; and stage IV masses grow around the mesenteric artery with involvement of the first jejunal arteries [270]. As mesenteric metastases are known to compromise mesenteric vasculature, vessel encasement (tumor tissue surrounding the vessel), signs of intestinal edema (thickened mucosal and submucosal layers) or bowel wall ischemia (thickened bowel wall with diminished contrast-enhancement) were also scored. The criteria were scored by five clinicians: two radiologists (Rad1 and Rad2, 15 and 5 years of experience, respectively), a nuclear medicine physician (Nucl, 4 years of experience), a surgeon (Surg, 10 years of experience), and an endocrinologist (End, 30 years of experience).

10.2.4 Radiomics

From both the MM and the SM segmentations, 564 features quantifying intensity, shape, and texture were extracted: these will be referred to as the MM features and the SM features, respectively. The MM and SM features total 1128 imaging features per patient. More details on the extracted features can be found in [Section 10.A](#) and [Table 10.A.2](#). The positions of the MM with respect to the SMA (x, y, and z) were also extracted, which we refer to as location features. These location features were used to approximate the established classification of the lymph node metastases stage [270]. We included these location features since lesions more proximal to the origin of the SMA tend to be more often symptomatic [271], bringing the total number of features to 1131.

To create a decision model from the features, the WORC toolbox was used, see [Figure 10.1](#) [36, 151]. In WORC, the decision model creation is divided in several steps, for example, feature selection, resampling and machine learning. For each step, a number of different methods are included. WORC performs an automated, exhaustive search among a variety of algorithms and their parameters to establish workflows that maximize performance and determine which combination of algorithms maximizes the prediction performance on the training set.

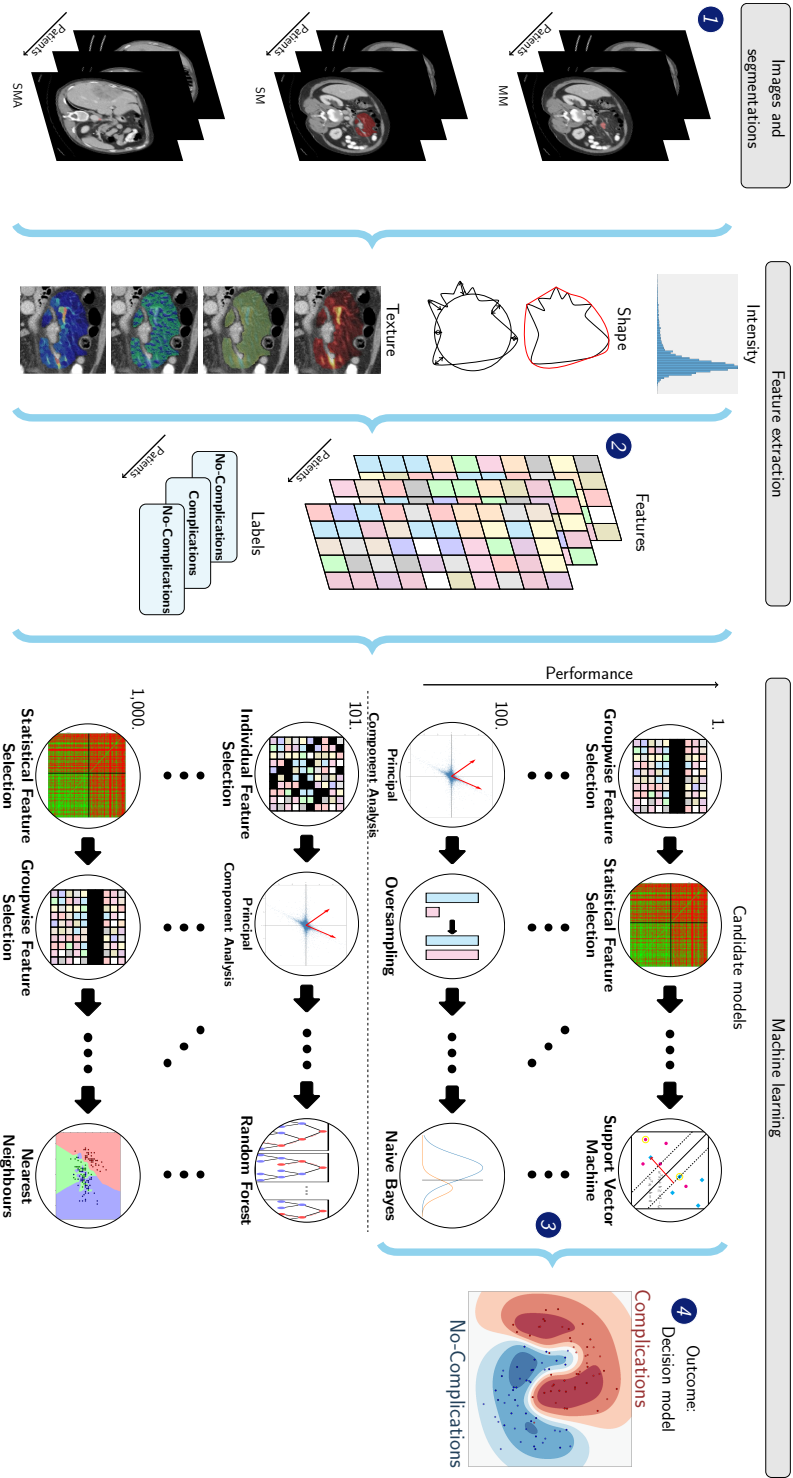


Figure 10.1: Schematic overview of the radiomics approach. Processing steps include segmentation of the mesenteric mass (MM), surrounding mesentery (SM), and superior mesenteric artery (SMA) on the CT scan (1), feature extraction from the CT based on the segmentations (2), and the creation of a decision model (4), using an ensemble of the best 100 workflows from 1000 candidate workflows (3), where the workflows are different combinations of the processing and analysis steps (e.g. the classifier used). Adapted, with permission, from Vos *et al.* [72] (i.e., Chapter 5 of this thesis): the images under (1), texture features, numbers at (3), and output at (4) have been modified with respect to the original figure.

Several models were created using different features to assess the predictive value of the various characteristics in predicting the development of symptomatic mesenteric mass: (1) age and sex; (2) baseline characteristics; (3) MM features; (4) SM features; (5) location features; (6) MM and SM features combined; (7) MM, SM and location features combined; (8) all features combined; and (9) similar to model₈, but excluding patients with positive enteric contrast. Model₁ and model₂ were created to assess whether simple, objective characteristics may provide information on symptom development. Model₉ was created to assess the impact of the usage of enteric contrast in the CT scans on the model performance. Even though the main area of interest is mesentery and not the bowel lumen, which is mostly affected by the contrast, the differences in appearance may influence the feature values and thus potentially bias the models. A schematic overview of the various models is given in Table 10.1. The code for both the feature extraction and creation of the decision models using WORC has been published open-source [272].

10.2.5 Comparison with clinical practice

In order to compare the performance of our model with current clinical practice, the CT scans were evaluated by the MTB from the Erasmus MC, an ENETS center of excellence. The MTB was asked to determine whether the patient was likely to develop intestinal complications due to the mesenteric mass and fibrosis within 1 year (yes/no), based on the same CT scan used for the radiomics analysis. The MTB assessed features such as bowel wall ischemia, edema, and severity of mesenteric fibrosis and vessel encasement. However, as there is no established method to use these features to guide decision-making, the features were simply assessed and expert opinion was used to determine if the patient is likely to develop intestinal complications, which resembles clinical practice.

Table 10.1: Description of the nine models to assess the predictive value of various feature groups in predicting abdominal complications.

Model	Enteric contrast	Radiomics features	Non-imaging features	Number of patients
Model ₁	Yes	None	Age, sex	68
Model ₂	Yes	None	All*	68
Model ₃	Yes	MM	None	68
Model ₄	Yes	SM	None	68
Model ₅	Yes	Location	None	68
Model ₆	Yes	MM, SM	None	68
Model ₇	Yes	MM, SM, Location	None	68
Model ₈	Yes	MM, SM, Location	All*	68
Model ₉	No	MM, SM, Location	All*	16

*Abbreviations: Age, sex, tumor grade, ENETS disease stage, CgA, 5-HIAA. CgA, serum chromogranin A; normal range < 94 µg/L; 5-HIAA, urinary 5-HIAA excretion; normal range < 50 µmol/24 h.

10.2.6 Statistical analysis

Differences between the asymptomatic and symptomatic groups in baseline clinical characteristics were evaluated using SPSS software (version 21 for Windows, SPSS Inc.). Data were presented as the median and inter-quartile range (IQR; 25th–75th percentiles) or percentage with count. Continuous data were compared by using a Mann–Whitney U test. A chi-square test was performed for the comparison of categorical data. A P-value of < 0.05 was considered statistically significant.

Agreement between the different raters in the systematic clinical evaluation was determined using Fleiss Kappa, where a value < 0.40 indicated poor agreement [273].

The statistics for the radiomics models were evaluated using the WORC software [36, 151]. To evaluate the significance of individual features, a Mann–Whitney U test was used. The P-values were corrected for multiple testing using the Bonferroni correction. A P-value of < 0.05 was considered statistically significant.

In all radiomics experiments, evaluation was implemented through a 100× random-split cross-validation, with 80% of the data used for training and 20% for independent testing, see Figure 10.2. On the training set, another random-split cross-validation was performed, splitting the dataset in 85% for training and 15% for validation to be used for the model optimization. Hence, all optimization was done on the training dataset: the test dataset was only used for evaluation to prevent overfitting on the test dataset. In both cross-validations, splitting was done in a stratified manner, to ensure that the balance between the asymptomatic and symptomatic groups was similar in training and test set.

To gain insight into the predictions of the model, patients were ranked from typical to atypical for both the asymptomatic and the symptomatic group, based on the model prediction consistency. This was determined by the number of times

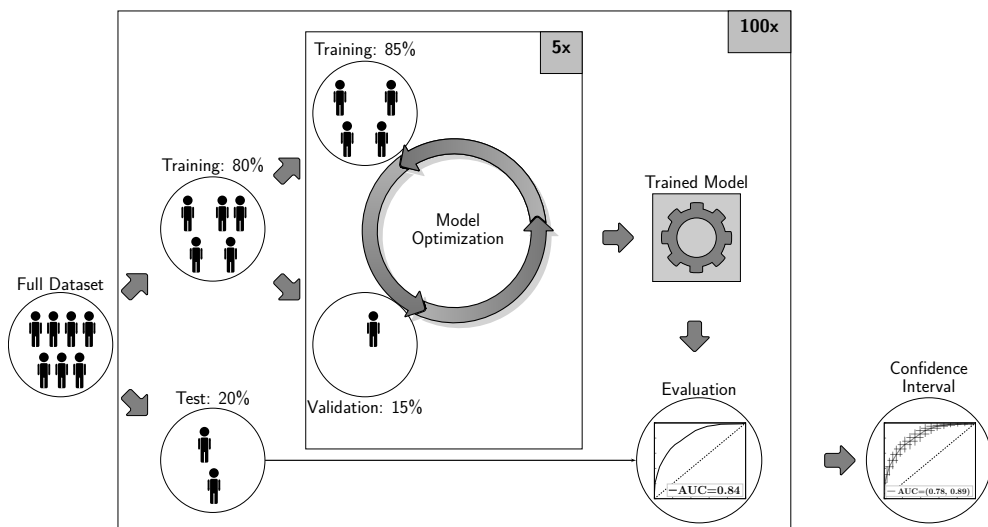


Figure 10.2: Visualization of the 100× random-split cross-validation, including a second 5× random-split cross-validation within the training set in which the model optimization was conducted.

(percentage) that a patient was classified correctly when included in the test set. Typical examples for each class were defined as patients who were always classified correctly; atypical vice versa.

Performance was evaluated using the area under the receiver operating characteristic (ROC) curve (AUC), balanced classification accuracy (BCA), sensitivity, and specificity. For the radiomics models, 95% CIs on the average performance metrics over all 100 cross-validation iterations were constructed using the corrected resampled t-test, thereby taking into account that samples in the crossvalidation splits are not statistically independent [64]. ROC confidence bands were constructed using fixed-width bands [67]. For the MTB, 95% CIs were constructed with Graphpad Software Prism using the modified Wald method. In all analyses, the symptomatic group was defined as the positive class.

10.3 Results

10.3.1 Dataset characteristics

A total of 68 patients was included, with 32 in the asymptomatic group and 36 in the symptomatic group. There were no significant differences between the groups in baseline clinical characteristics (Table 10.2). For the asymptomatic group, the median time between the CT scan and development of abdominal symptoms or endofollow-up was 70.5 months (IQR; 50–86 months). For the symptomatic group, the median time between CT scan and palliative surgery was 97 days (IQR; 49–140 days). In the symptomatic group, indications for surgery were respectively: obstruction ($n = 19$, 53%), pain ($n = 13$, 36%), ischemia ($n = 2$, 6%), and perforation ($n = 2$, 6%). For 32 patients, laparotomy findings revealed macroscopic signs of mesenteric fibrosis and, when acute pain was present preoperatively, signs of ischemia were present in 59% ($n = 19$). In the remaining four operated patients, documentation of findings during surgery was scarce.

The resulting multi-center CT dataset originated from 29 different scanners and thereby showed substantial heterogeneity in the imaging protocols (Table 10.2). Statistically significant differences in the distribution of the parameters between the CT scans of the symptomatic and the asymptomatic group were found for the use of enteric contrast, pixel spacing, tube current, and kilovoltage peak. However, the absolute differences were generally small, for example, 0.73 mm vs 0.75 mm in mean pixel spacing. Additionally, nine different reconstruction kernels were used.

10.3.2 Feature significance

After correcting for multiple testing, from the 1137 features (1128 imaging, 3 location, and 6 patient characteristics), 73 were found to have significant P-values (0.003–0.045), see Figure 10.A.1. These included only features extracted from the SM: a more detailed description of these features is given in Section 10.B. No shape features, thus also not the SM volume, were found to be significant.

Table 10.2: Baseline characteristics of the 68 patients. Numerical data are presented as median with interquartile range (IQR) in brackets. Categorical data are presented as percentages with a count in brackets. P-values are calculated using a Mann–Whitney U test for numerical data, a chi-square test for categorical data.

Characteristic	Symptomatic (N = 36)	Asymptomatic (N = 32)	P-value
Clinical			
Age	66 [55 – 74]	62 [54 – 72]	0.90
Male	56% (20)	78% (25)	0.072
CgA	343 [178 – 1057]	170 [72 – 415]	0.27
5-HIAA	163 [59 – 481]	126 [78 – 288]	0.46
Tumor grade			0.40
Grade I	56 % (20)	56 % (18)	
Grade II	31 % (11)	19 % (6)	
Unknown	14% (5)	25% (8)	
ENETS disease stage			0.15
Stage III	22% (8)	9% (3)	
Stage IV	78% (28)	91% (29)	
CT Imaging			
Enteric contrast	36% (13)	9% (3)	0.009
Pixel spacing (mm)	0.73 [0.70, 0.77]	0.75 [0.73, 0.79]	0.04
Slice thickness (mm)	3.0 [3.0, 3.25]	3.0 [3.0, 5.0]	0.19
Tube current (mA)	158 [99, 312]	271 [144, 346]	0.034
Kilovoltage peak	100 [100, 120]	120 [100, 120]	0.020
Manufacturer			0.55
Siemens	30	30	
Philips	2	1	
Toshiba	3	1	
Unknown	1	0	
Surgery indication			
Obstruction	53% (19)		
Pain	36% (13)		
Ischemia	6% (2)		
Perforation	6% (2)		

*Abbreviations: CgA, serum chromogranin A; normal range < 94 $\mu\text{g/L}$; 5-HIAA, urinary 5-HIAA excretion; normal range < 50 $\mu\text{mol/24 h}$.

10.3.3 Systematic evaluation by clinicians

The performance of the systematic clinical evaluation by the five raters is shown in [Table 10.3](#); their ROC curves are shown in [Figure 10.3](#). While all clinicians performed better than guessing (0.50), their AUCs varied (radiologists: 0.85 and 0.76, nuclear physician: 0.71, surgeon: 0.82, endocrinologist: 0.62). Fleiss Kappa between the five clinicians on evaluating patients as asymptomatic or symptomatic was 0.15, indicating poor agreement. The agreement on the classification of the radiological features was also poor (0.06–0.35) ([Table 10.A.1](#)).

10.3.4 Evaluation of radiomics models

The performance of the various radiomics models is shown in [Table 10.3](#). Model₁, using only age and sex, had a poor performance (AUC of 0.49), indicating that age and sex are not related to the risk of developing intestinal complications. Inclusion of tumor grade according to WHO criteria, ENETS disease stage, CgA level, and urinary 5-HIAA excretion, that is, model₂, performed slightly better (AUC of 0.58). Among the models using radiomics features from a single ROI, model₄, including SM, had the best performance (AUC of 0.81, sensitivity of 0.78, specificity of 0.67). Interestingly, model₅, including only the location features of the MM also had fair predictive power (AUC of 0.72). Combining all imaging and location features, model₇, performed similarly (AUC of 0.74, sensitivity of 0.70, specificity of 0.65) to

Table 10.3: Performances of systematic evaluation by five raters and the radiomics models. The radiomics models are based on: age and sex (model₁); all non-imaging features (model₂); features extracted from the mesenteric mass (MM) (model₃); features extracted from the surrounding mesentery (SM) (model₄); only the location (model₅); both MM and SM features (model₆); MM, SM, and location features (model₇); MM, SM, location, and non-imaging features (model₈); similar to model₈ but excluding patients with positive enteric contrast (model₉). Performance for the radiomics models was given as mean (95% CI).

Model	AUC	BCA	Specificity	Sensitivity
Radiologist 1	0.85	0.80	0.84	0.75
Radiologist 2	0.76	0.73	0.66	0.81
Nuclear physician	0.71	0.68	0.91	0.44
Surgeon	0.82	0.79	0.78	0.81
Endocrinologist	0.60	0.59	0.63	0.56
Model ₁	0.49 (0.34, 0.65)	0.50 (0.39, 0.61)	0.49 (0.23, 0.74)	0.52 (0.30, 0.73)
Model ₂	0.58 (0.44, 0.72)	0.58 (0.46, 0.70)	0.55 (0.34, 0.76)	0.61 (0.41, 0.80)
Model ₃	0.65 (0.52, 0.79)	0.61 (0.49, 0.73)	0.61 (0.43, 0.78)	0.61 (0.42, 0.81)
Model ₄	0.81 (0.72, 0.91)	0.72 (0.62, 0.82)	0.67 (0.49, 0.85)	0.78 (0.61, 0.94)
Model ₅	0.72 (0.60, 0.84)	0.63 (0.51, 0.75)	0.60 (0.41, 0.79)	0.67 (0.47, 0.87)
Model ₆	0.77 (0.64, 0.90)	0.71 (0.59, 0.83)	0.69 (0.50, 0.88)	0.73 (0.55, 0.90)
Model ₇	0.74 (0.62, 0.87)	0.68 (0.55, 0.81)	0.65 (0.45, 0.85)	0.70 (0.52, 0.88)
Model ₈	0.79 (0.66, 0.91)	0.72 (0.61, 0.82)	0.72 (0.54, 0.90)	0.71 (0.52, 0.89)
Model ₉	0.77 (0.63, 0.91)	0.69 (0.55, 0.84)	0.74 (0.54, 0.94)	0.64 (0.40, 0.88)

* Abbreviations: AUC, area under the receiver operating characteristic curve; BCA, balanced classification accuracy.

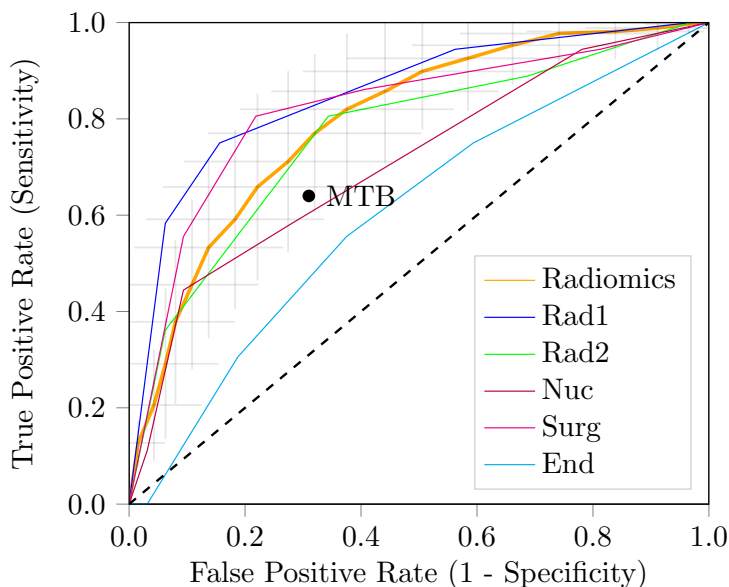


Figure 10.3: Receiver operating characteristic curve of radiomics model₄, based on the surrounding mesentery, and of evaluation by five clinical raters (radiologist 1 (blue), radiologist 2 (green), nuclear physician (purple), surgeon (magenta), and endocrinologist (cyan)). The performance of the multidisciplinary tumor board (MTB) is indicated by a red dot. For the radiomics model, the gray crosses identify the 95% CIs of the performance over the 100× random-split cross-validation iterations; the orange curve is fit through the mean of the CIs.

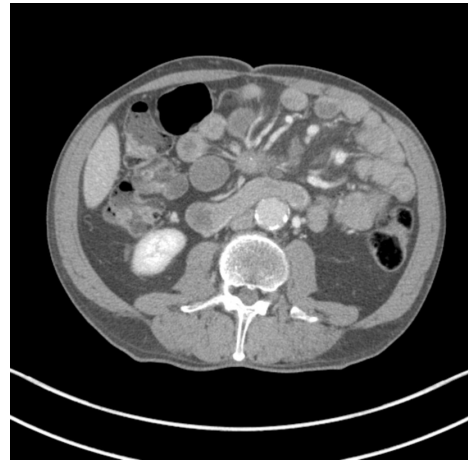
the model based solely on the SM. Inclusion of the patient characteristics (model₈, AUC of 0.79) did not improve the predictive power.

In our dataset, 24% ($n = 16$) of the CT scans were performed with enteric contrast. Of these patients, 18.6% ($n = 3$) were asymptomatic; hence, the distribution of enteric contrast with respect to asymptomatic and symptomatic group was significantly different ($P < 0.05$, Table 10.2). Excluding these patients, that is, model₉, yielded a similar performance (AUC of 0.77).

Of the 68 patients, 35 patients (19 asymptomatic, 16 symptomatic) were always classified correctly, that is, in all 100 cross-validation iterations, by model₄, and are thus considered typical. Of these 32 typical patients, 13 patients (7 asymptomatic, 6 symptomatic) were also correctly classified by all five clinicians. Analogously, 6 patients (3 asymptomatic, 3 symptomatic) were always classified incorrectly and thus considered atypical. In Figure 10.4, four CT slices of such typical and atypical examples of asymptomatic and symptomatic patients are depicted. The patients with enteric contrast were both in the typical ($n = 7$) and atypical ($n = 1$) examples of both classes.



(A) Typical Asymptomatic



(B) Atypical Asymptomatic



(C) Atypical Symptomatic



(D) Typical Symptomatic

Figure 10.4: Examples of typical and atypical surrounding mesentery. The typical examples (A,D) are two of the patients always classified correctly by model₄; the atypical examples (B,C) are two of the patients always classified incorrectly by model₄. (A) Typical asymptomatic, (B) atypical asymptomatic, (C) atypical symptomatic, and (D) typical symptomatic.

10.3.5 Comparison with multidisciplinary tumor board

The MTB prediction of developing intestinal complications had a specificity of 0.69 (95% CI; 0.51, 0.82), a sensitivity of 0.64 (95% CI; 0.48, 0.78), and an accuracy of 0.66 (95% CI; 0.54, 0.77). For the sake of brevity, only the ROC curves of the single-ROI model with the highest AUC, model₄, the five raters, and the MTB performance are depicted in [Figure 10.3](#). The performance of the MTB was slightly below the ROC curve of the mean performance of the radiomics model over all cross-validations, but within the 95% CI.

10.4 Discussion

We evaluated both systematic clinical evaluation and a radiomics approach for reliably identifying patients who are prone to develop complications of the metastatic mesenteric mass and fibrosis and thus may benefit from prophylactic surgery. Our results show that both the systematic clinical evaluation and our best performing radiomics model can identify these patients with a performance similar to a specialized MTB.

To date, there are no clear clinical or radiological predictors for the development of a symptomatic mesenteric mass [262, 263]. Therefore, we evaluated a wide array of clinical characteristics and radiomics features. In contrast to other prognostic models in SI-NETs, we found that clinical characteristics such as age, sex, ENETS disease stage, tumor grade and markers had little to no predictive power for the development of a symptomatic mesenteric mass (model₁ and model₂) [264, 265, 266].

From the radiomics features, only SM features showed statistically significant differences between the asymptomatic and symptomatic patients. No MM or location features showed a statistically significant difference. This highlights the importance of the mesentery surrounding the metastatic mesenteric mass in the development of symptoms. In order to gain insight in the underlying profibrotic mechanisms, we analyzed the predictive features of the SM and found that most (93%) were texture features. Future detailed analysis of the relation between these features and clinical characteristics could elucidate the processes involved in the development of a symptomatic mesenteric mass and fibrosis and guide treatment development. The importance of the SM was also confirmed by the radiomics models, as the model solely using SM features (model₄) was one of the highest ranking models in terms of AUC and the performance was not improved by additional features (i.e. model₆–model₈). Moreover, model₄ is clinically more feasible, as it only requires annotation of the surrounding mesentery. We will, therefore, further refer to model₄ as ‘the radiomics model’.

Systematic evaluation by clinicians resulted in similar discriminative power as the radiomics model. However, evaluation of the separate CT findings demonstrated poor inter-observer agreement, which is in line with findings in the literature [274]. The relatively low degree of the overall agreement further limits the reliability of the prediction by the clinicians. The radiomics model, on the other hand, is independent of the observer and thus any personal training or experience, assuming the segmentation is reproducible. It could therefore be useful in clinics where

there are no NET specialists, to better identify patients that may benefit from prophylactic palliative surgery and refer these patients to a center of expertise. Moreover, reducing the bias in risk evaluation could aid assessment of treatment effectiveness for mesenteric metastases and fibrosis, and the development of clear guidelines for patient selection for prophylactic palliative surgery.

Some limitations to our study should be noted. First, although we used a multi-center imaging dataset and performed a rigorous cross-validation experiment strictly separating training from testing data, we did not validate our model on an independent, external dataset. Moreover, even though our dataset was relatively large considering the rarity of SI-NETs, it was relatively small for a radiomics study [24], which may explain why our CIs are quite wide (e.g. the AUCs span between 20 and 30% of the range). Additionally, testing for statistically significant differences of the AUCs through, for example, a DeLong test was not possible due to limited power. Expanding the size of the dataset may result in an increase in performance and increased statistical power. Second, in line with guidelines from the radiomics field [18], our study included CT scans over a time period of 10 years with variations in acquisition protocols. On one hand, this is a strength of our study, as the radiomics models had predictive value despite substantial acquisition variations. Moreover, as the models were trained on a wide variety of CT scanners and acquisition protocols, we expect the model to be able to accurately make predictions in a wide variety of (routine) settings. On the other hand, heterogeneity may have (negatively) affected our performance. Using a single-scanner study will limit the generalizability but may positively impact the performance. Further research is required to evaluate the influence of acquisition parameters on the model performance. When expanding the dataset to include more patients, feature harmonization techniques such as ComBat may be employed [185]. Third, our model relies on the manual annotation of the ROIs. Manual annotation can be time consuming and may lead to observer dependency of the model. Automation of the segmentation may help overcome these deficits.

To our knowledge, this is the first study that shows the potential of radiomics for the prediction of abdominal complications in SI-NETs. In our study, we used CT, as this was the preferred modality in routine clinical care [275]. Future research may investigate the potential value of other imaging modalities. The usage of MRI might be limited in this context as it holds similar information and is not routinely performed in SI-NETs [275]. On the other hand, use of nuclear imaging in SI-NETs is well-established, especially PET-CT using 68Ga-labeled somatostatin analogs [275]. Moreover, many new molecular imaging probes for the detection of fibrosis and fibrogenesis are being developed (e.g. fibroblast activation protein imaging) [276, 277, 278]. However, further research is required to evaluate the value of these imaging techniques in the context of this study, that is, for the prediction of abdominal complications in SI-NETs, potentially combined with radiomics.

10.5 Conclusion

This study used routinely acquired CT scans to identify SI-NET patients prone to the development of intestinal complications due to a metastatic mesenteric mass

and fibrosis. The CT scans were analyzed by five clinicians with different levels of experience using systematic visual evaluation and a radiomics model. While all clinicians were able to identify patients at risk to some degree, the performance of the clinicians substantially varied and agreement was poor. The radiomics model is based on automatic feature extraction from contrast-enhanced CT scans and mainly driven by the appearance of the surrounding mesentery. The predictive power was similar to that of experienced clinicians and a specialized MTB. It could therefore aid in guiding the clinical decision on which patients should receive prophylactic palliative surgery.

Declaration of interest Wiro Niessen is founder, scientific lead and stock holder of Quantib BV. Tessa Brabander is a member of the joined advisory board of AAA/Novartis, and is a speaker and has received travel fees from AAA/Novartis and Prime Oncology. Wouter de Herder received research support from Ipsen and AAA/Novartis and received speaker fees from AAA/Novartis, Pfizer and Ipsen. The other authors do not declare any conflicts of interest.

Funding Martijn Starmans acknowledges funding from the research program STRaTeGy (project number 14929-14930), which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO). Anela Blazevic received funding from the Ipsen Fund via an unrestricted research fund. This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

Author contribution statement Literature search, study design, data collection, data analysis, data interpretation, writing: Anela Blazevic, Martijn P. A. Starmans; Study design, data collection, data interpretation, revising Tessa Brabander, Stefan Klein; Data interpretation, revising: Johannes Hofland, Gaston J. H. Franssen, Richard A. Feelders, Wiro J. Niessen, Roy Dwarkasing, Renza van Gils, Wouter W. de Herder. In addition to the above, all authors approved of the final version of the manuscript and agree to be accountable for all aspects of the work.

Appendix

Appendix 10.A Radiomics feature extraction

This supplementary material is similar to Vos *et al.*, Timbergen *et al.* [72, 73] (i.e., Chapter 5 and Chapter 6 of this thesis), but details relevant for the current study are highlighted.

A total of 564 radiomics features per region of interest (ROI) were used in this study. An overview of all features is provided in Table 10.A.2. All features were extracted using the defaults for CT scans from the Workflow for Optimal Radiomics Classification (WORC) toolbox [36], which internally uses the PREDICT [51] and PyRadiomics [44] feature extraction toolboxes. The code to extract the features for this specific study has been published open-source [272]. For details on the mathematical formulation of the features, we refer the reader to Zwanenburg *et*

al. [39]. More details on the extracted features can be found in the documentation of the respective toolboxes, mainly the WORC documentation [68].

Intensity features were extracted using the histogram of all intensity values within the ROIs and included several first-order statistics such as the mean, standard deviation and kurtosis. Shape features were extracted based only on the ROI, i.e. not using the image, and included shape descriptions such as the compactness, roundness and circular variance. Additionally, the volume and orientation of the ROIs were used. Texture features were extracted using the Gabor filters, Laplacian of Gaussian filters, Vessel filters [54], local phase filters [53, 195], Local Binary Patterns [52], the Gray Level Co-occurrence Matrix [39], the Gray Level Size Zone Matrix [39], the Gray Level Run Length Matrix [39], the Neighbourhood Grey Tone Difference Matrix [39], and the Gray Level Difference Matrix [39].

Most of the features include parameters to be set for the extraction. Beforehand, the values of the parameters that will result in features with the highest discriminative power for the asymptomatic/symptomatic classification task are not known. Including these parameters in the workflow optimization would lead to repeated computation of the features, resulting in a redundant increase in computation time. Therefore, alternatively, these features are extracted at a range of parameters as is default in WORC. The hypothesis is that the features with high discriminative power will be selected by the feature selection methods and/or the machine learning methods. The parameters used are described in Table 10.A.2.

The imaging data used in this study is multi-center, and therefore heterogeneous in terms of acquisition protocols. Especially the variations in slice thickness may cause feature values to be highly dependent on the acquisition protocol. Hence, extracting robust 3D features may be hampered by these variations, especially for low resolutions. The images were not resampled, as this would result in interpolation errors. To overcome this issue, all features were extracted per 2D axial slice and aggregated over all slices. Afterwards, several first-order statistics over the feature distributions were evaluated and used in the machine learning approach. Additionally, before feature extraction, all images were scaled to Hounsfield Units. As all images had the same unit, no additional normalization was applied.

Appendix 10.B Significant features

After Bonferroni correction, 73 features had a statistically significant distribution ($p < 0.05$ in Mann-Whitney U test) in the asymptomatic and symptomatic group. The p-values and names of these features are depicted in Figure 10.A.1. Several groups of features which quantify similar visual appearances in the images can be identified.

All statistically significant features were extracted from the surrounding mesentery (SM): no features from the mesenteric mass, neither the location or patient characteristics were found to be significant. Out of the 73 statistically significant features, 68 (93%) were texture features, as indicated by the blue bars. Thus, the differences between the symptomatic and asymptomatic patients are mostly explained by texture related characteristics of the surrounding mesentery, and not by characteristics of the CT intensity distribution or the shape and volume of the mesentery.

A total 64 (88%) of the statistically significant features is based on the Gray Level Co-occurrence Matrix (GLCM). In the GLCM, after discretizing the image in a fixed number of values, the co-occurrences of specific values between two pixels are counted. For counting the co-occurrences, different directions (e.g. horizontal, vertical) and spacings (e.g. one pixel, ten pixels) can be used. From the resulting GLCM matrix, several features can be computed, such as the homogeneity (uniform spreading of the counts among the different values), the dissimilarity (two values occur less equal to each other in one configuration (e.g. left low gray value – right high gray value) than the opposite (e.g. right low gray value – left high gray value), and the energy (more instances of intensity value pairs in the image that neighbor each other at higher frequencies). Using different combinations of the angle and the distance, 16 (22%) GLCM homogeneity features were significant, of which nine had the lowest p-values of all features. Hence, for the classification it seemed important whether only specific gray level values occurred often next to each other (low GLCM homogeneity), e.g. homogeneous ROI or very distinct patterns, or whether a wide variety of gray levels occurred often next to each other (high GLCM homogeneity), e.g. heterogeneous ROI or random patterns. Inspection of the distributions of these features showed that; 1) the average of the GLCM homogeneity was generally lower for the symptomatic group, indicating that generally these SMs are more homogeneous; and 2) the outliers of the GLCM homogeneity generally consisted of the asymptomatic group, indicating that symptomatic SMs generally were not extremely heterogeneous or homogeneous but rather in between.

It should be noted that the p-values presented here are not necessarily representative of which feature contribute most to the predictions made by the radiomics models. The combination of methods in the WORC toolbox allows for high order, non-linear combinations of multiple features. Hence, while a feature may have a low value in univariate testing, a multivariate combination of features (with lower univariate predictive value) may result in a better performance. Additionally, the combination of 50 workflows in the final model in WORC serves as a form of regularization to prevent the focus on a single feature (group). In this final model when using the SM features (model₄), all feature groups as defined in [Section 10.A](#) were approximately equally often used.

Thus, while the p-values of univariate statistical testing may give us information about the differences between asymptomatic and symptomatic patients in terms of appearance, a different combination of features may result in a better predictive performance than simply selecting the univariate most significant features.



Figure 10.A.1: P-values of Mann-Whitney U tests of feature values for the asymptomatic and symptomatic group. Purple bars correspond to texture features, violet bars to histogram features.

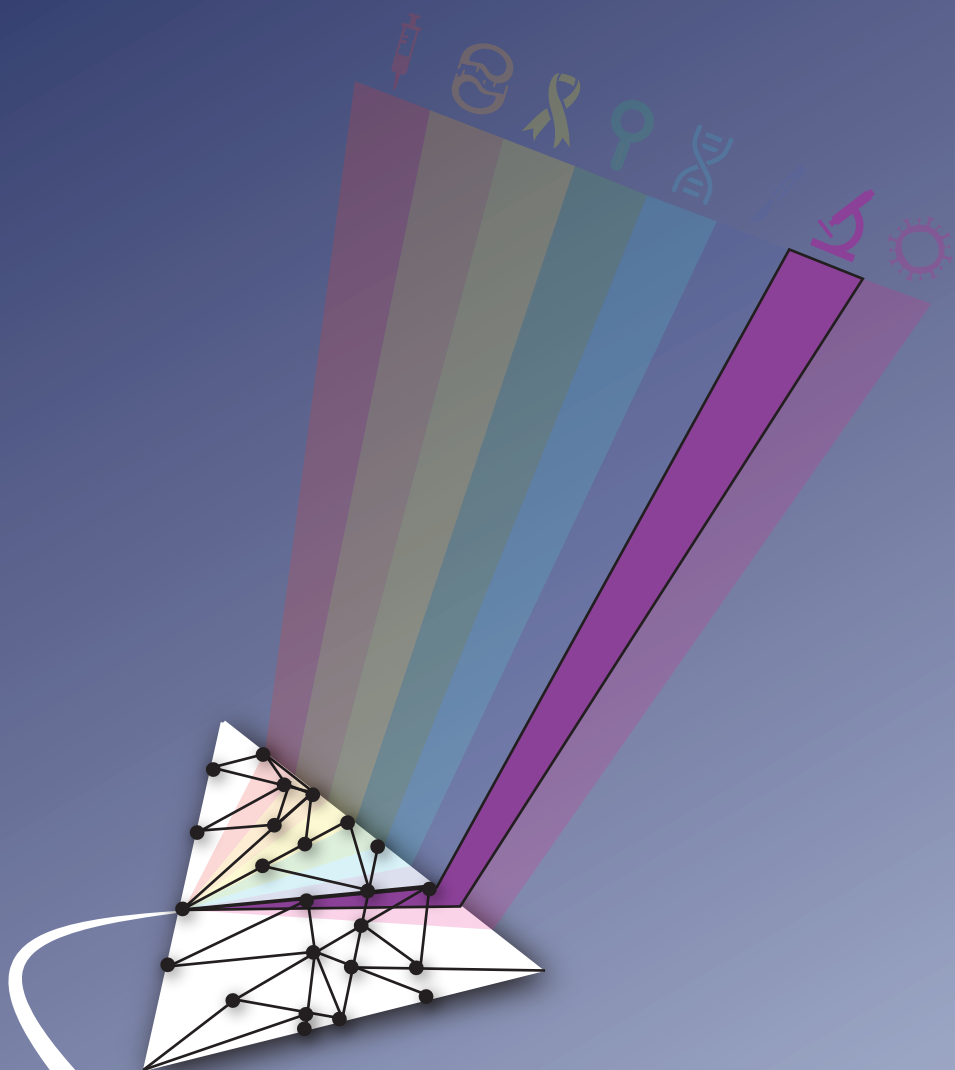
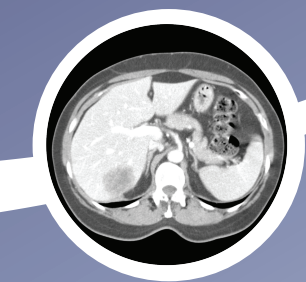
Table 10.A.1: Criteria for systematic evaluation whether patients with SI-NETs are symptomatic or asymptomatic.

Characteristic	Ratings		Fleiss Kappa
Fibrosis (ordinal)	1	Grade 1	0.31
	2	Grade 2	
	3	Grade 3	
Encasement of vessels (ordinal)	1	Yes	0.06
	2	Unsure	
	3	No	
Lymph node location (categorical)	1	Stage I	0.02
	2	Stage II	
	3	Stage III	
	4	Stage IV	
Bowel wall edema (ordinal)	1	Yes	0.35
	2	Unsure	
	3	No	
Bowel wall ischemia (ordinal)	1	Yes	0.17
	2	Unsure	
	3	No	
Asymptomatic (ordinal)	1	Strongly disagree	0.15
	2	Disagree	
	3	Neither agree or disagree	
	4	Agree	
	5	Strongly agree	

Table 10.A.2: Overview of the 564 features used in this study. GLCM features were calculated in four different directions (0, 45, 90, 135 degrees) using 16 gray levels and pixel distances of 1 and 3. LBP features were calculated using the following three parameter combinations: 1 pixel radius and 8 neighbours, 2 pixel radius and 12 neighbours, and 3 pixel radius and 16 neighbours. Gabor features were calculated using three different frequencies (0.05, 0.2, 0.5) and four different angles (0, 45, 90, 135 degrees). LoG features were calculated using three different widths of the Gaussian (1, 5 and 10 pixels). Vessel features were calculated using the full mask, the edge, and the inner region. Local phase features were calculated on the monogenic phase, phase congruency and phase symmetry.

Histogram (13 features)	LoG (13*3=39 features)	Vessel (12*3=39 features)	GLCM (MS) (6*3*4*2=144 features)	Gabor (13*3*3=156 features)	NGTDM (5 features)	LBP (13*3=39 features)
min	min	min	contrast (normal, MS mean + std)	min	busyness	min
max	max	max	dissimilarity (normal, MS mean + std)	max	coarseness	max
mean	mean	mean	homogeneity (normal, MS mean + std)	mean	complexity	mean
median	median	median	angular second moment (ASM) (normal, MS mean + std)	median	contrast	median
std	std	std	energy (normal, MS mean + std)	std	strength	std
skewness	skewness	skewness	correlation (normal, MS mean + std)	skewness		skewness
kurtosis	kurtosis	kurtosis		kurtosis		kurtosis
peak	peak	peak		peak		peak
peak position	peak position	peak position		peak position		peak position
range	range	range		range		range
energy	energy	energy		energy		energy
quartile range	quartile range	quartile range		quartile range		quartile range
entropy	entropy	entropy		entropy		entropy
GLSZM (16 features)	GLRM (16 features)	GLDM (14 features)	Shape (35 features)	Orientation (9 features)	Local phase (13*3=39 features)	
Gray Level Non Uniformity	Gray Level Non Uniformity	Dependence Entropy	compactness (mean + std)	theta_x	min	
Gray Level Non Uniformity Normalized	Gray Level Non Uniformity Normalized	Dependence Non-Uniformity	radial distance (mean + std)	theta_y	max	
Gray Level Variance	Gray Level Variance	Dependence Non-Uniformity Normalized	roughness (mean + std)	theta_z	mean	
High Gray Level Zone Emphasis	High Gray Level Run Emphasis	Dependence Variance	convexity (mean + std)	COM index x	median	
Large Area Emphasis	Long Run Emphasis	Gray Level Non-Uniformity	circular variance (mean + std)	COM index y	std	
Large Area High Gray Level Emphasis	Long Run High Gray Level Emphasis	Gray Level Variance	principal axes ratio (mean + std)	COM index z	skewness	
Large Area Low Gray Level Emphasis	Long Run Low Gray Level Emphasis	High Gray Level Emphasis	elliptic variance (mean + std)	COM x	kurtosis	
Low Gray Level Zone Emphasis	Low Gray Level Run Emphasis	Large Dependence Emphasis	solidity (mean + std)	COM y	peak	
SizeZoneNonUniformity	RunEntropy	Large Dependence High Gray Level Emphasis	area (mean, std, min + max)	COM z	peak position	
SmallAreaNonUniformityNormalized	RunLengthNonUniformity	Large Dependence Low Gray Level Emphasis	volume (total, mesh, volume)		range	
SmallAreaEmphasis	RunLengthNonUniformityNormalized	Low Gray Level Emphasis	elongation		energy	
SmallAreaHighGrayLevelEmphasis	RunPercentage	Small Dependence Emphasis	flatness		quartile	
SmallAreaLowGrayLevelEmphasis	RunVariance	Small Dependence High Gray Level Emphasis	least axis length		entropy	
ZoneEntropy	ShortRunEmphasis	Small Dependence Low Gray Level Emphasis	major axis length			
ZonePercentage	ShortRunHighGrayLevelEmphasis		minor axis length			
ZoneVariance	ShortRunLowGrayLevelEmphasis		maximum diameter 3D			
			maximum diameter 2D (rows, columns, slices)			
			sphericity			
			surface area			
			surface volume ratio			

* Abbreviations: COM: center of mass; GLCM: gray level co-occurrence matrix; MS: multi slice; NGTDM: neighborhood gray tone difference matrix; GLSZM: gray level size zone matrix; GLRLM: gray level run length matrix; LBP: local binary patterns; LoG: Laplacian of Gaussian; std: standard deviation.



11.

Distinguishing pure histopathological growth patterns of colorectal liver metastases on CT using deep learning and radiomics: a pilot study

Based on: **M. P. A. Starmans***, F. E. Buisman*, M. Renckens, F. E. J. A. Willemsen, S. R. van der Voort, B. Groot Koerkamp, D. J. Grünhagen, W. J. Niessen, P. B. Vermeulen, C. Verhoef, J. J. Visser, and S. Klein, "Distinguishing pure histopathological growth patterns of colorectal liver metastases on CT using deep learning and radiomics: A pilot study," *Clinical & Experimental Metastasis*, 2021. doi: [10.1007/s10585-021-10119-6](https://doi.org/10.1007/s10585-021-10119-6)

* indicates equal contributions

Abstract

Histopathological growth patterns (HGPs) are independent prognosticators for colorectal liver metastases (CRLM). Currently, HGPs are determined postoperatively. In this study, we evaluated radiomics for preoperative prediction of HGPs on computed tomography (CT), and its robustness to segmentation and acquisition variations. Patients with pure HGPs [i.e. 100% desmoplastic (dHGP) or 100% replacement (rHGP)] and a CT-scan who were surgically treated at the Erasmus MC between 2003–2015 were included retrospectively. Each lesion was segmented by three clinicians and a convolutional neural network (CNN). A prediction model was created using 564 radiomics features and a combination of machine learning approaches by training on the clinician's and testing on the unseen CNN segmentations. The intra-class correlation coefficient (ICC) was used to select features robust to segmentation variations; ComBat was used to harmonize for acquisition variations. Evaluation was performed through a $100 \times$ random-split cross-validation. The study included 93 CRLM in 76 patients (48% dHGP; 52% rHGP). Despite substantial differences between the segmentations of the three clinicians and the CNN, the radiomics model had a mean area under the curve of 0.69. ICC-based feature selection or ComBat yielded no improvement. Concluding, the combination of a CNN for segmentation and radiomics for classification has potential for automatically distinguishing dHGPs from rHGP, and is robust to segmentation and acquisition variations. Pending further optimization, including extension to mixed HGPs, our model may serve as a preoperative addition to postoperative HGP assessment, enabling further exploitation of HGPs as a biomarker.

11.1 Introduction

Colorectal liver metastases (CRLM) represent approximately 30% of all metastases in patients with colorectal carcinoma [279]. Ten-year survival after CRLM resection is 20%, primarily limited due to recurrent disease [280]. Prognosis estimation is challenging since powerful prognosticators are lacking.

Histopathological growth patterns (HGP) have recently been identified as independent prognosticators in patients after CRLM resection [104]. The interface between tumor cells and normal liver parenchyma (NLP) is characterized by three distinct HGPs: two frequent (desmoplastic HGP (dHGP) and replacement HGP (rHGP), see [Figure 11.A.1](#)) and one rare (pushing HGP) type [108, 281]. A previous study found that dHGP patients have superior survival compared to mixed, replacement or pushing HGP patients [104]. Moreover, recent studies have suggested that HGPs could predict systemic chemotherapy effectiveness [282, 283]. Previous guidelines suggested a cut-off of 50% of a single HGP to determine the dominant HGP [281]. More recent studies have shown that pure HGPs (i.e., 100% of the interface expresses the HGP) appear clinically more relevant [284].

Preoperative HGP assessment is currently not possible, as assessment requires pathology slices of resection specimens to be reviewed with a light microscope. Biopsy material is not suitable due to lesion heterogeneity. Preoperative assessment, however, could provide valuable information on prognosis, could help identifying patients who benefit from perioperative systemic treatment, and could be used to evaluate response treatment by monitoring changes in the HGP [282, 283]. As there is currently no method to assess HGPs preoperatively, investigating these potential improvements is not possible. Hence there is a need to identify HGPs based on medical imaging to exploit the full potential of HGPs as a biomarker, as concluded by a recent review [285].

The field of radiomics has emerged as a non-invasive way to establish relations between quantitative image features and tumor biology or clinical outcomes [19] (i.e., [Chapter 2](#) of this thesis). Several radiomics studies have shown promising results in a wide variety of applications [24]. In CRLM, radiomics has been used to assess chemotherapy response, survival, detect CRLM, and predict mixed HGPs [286, 287, 288, 289, 290]. A major drawback of many radiomics approaches is the dependence on manual segmentations, which may introduce observer variability in the predictions [291, 292, 293]. Additionally, image acquisition variations may affect the predictions [293].

The primary aim of this study was to evaluate if radiomics can preoperatively distinguish pure HGPs on computed tomography (CT) scans as a non-invasive addition to postoperative histological assessment, enabling pre-operative treatment response prediction and evaluation. The secondary aim was to evaluate and improve the robustness of the radiomics models to variations in segmentation and acquisition protocol.

11.2 Methods and materials

11.2.1 Patients

This study was performed in accordance with the Dutch Code of Conduct for Medical Research of 2004 and approved by the local institutional review board (“Medische Ethische Toetsings Commissie” (METC), MEC-2017-479). As the study was retrospectively performed with anonymized data, the need for informed consent was waived. Patients surgically treated at the Erasmus MC between 2003–2015 with a preoperative CT-scan in the portal venous phase (PVP) and available hematoxylin and eosin stained tissue sections were included retrospectively. Patients with recurrent CRLM or CRLM requiring two-staged resections were not included. Both synchronous and metachronous resections were allowed. Pre-contrast and arterial phase CT were available in a minority of patients and therefore excluded. Patients treated with preoperative chemotherapy were excluded, since chemotherapy may alter HGPs [104]. HGPs were scored on resection specimens according to the consensus guidelines by an expert pathologist (PV) [108]. In this pilot, we focused on pure HGPs as these appear clinically more relevant than mixed HGPs, as a previous study showed that pure dHGP is an unmatched predictor for improved survival in chemo-naïve patients with CRLM [284]. Furthermore, we hypothesized that the use of radiomics has a higher chance of success in distinguishing pure HGPs, as their morphology is less heterogeneous than mixed HGPs. Patients with pure pushing HGPs were excluded, as this is rare ($< 1\%$) [108, 281, 282, 284]. The pure dHGPs and rHGPs both make up about 20% of the total population of chemo-naïve patients, resulting in inclusion of 40% of all available patients [284].

Various clinical characteristics were collected: age, sex, primary tumor location and nodal status, disease free interval between resection of the colorectal carcinoma and CRLM detection, and the preoperative carcinoembryonic antigen level. Size and number of CRLMs, including ablations without histology, were derived from the CT-scans.

11.3 Segmentation

Lesion segmentation was independently performed by four observers: a medicine student with no relevant experience (STUD1), a PhD student (PhD) with limited experience, an expert abdominal radiologist (RAD), and an automatic CNN. The student segmented all lesions within a week, and immediately afterwards, segmented all lesions a second time (STUD2) to evaluate the intra-observer variability. As the order of segmentation was not the same in the first and second time, but randomized, the time between the first and second segmentation varied between two and seven days. Segmentation agreement between all observer pairs was determined through the pairwise dice similarity coefficient (DSC).

Segmentation by the clinicians was performed with in-house Python-based software [105]. For the lesions, the clinicians could segment manually or semi-automatically using region-growing or slice-to-slice contour propagation. Segmentation was performed per slice in the 2D transverse plane, resulting in a 3D volume.

Semi-automatic results were always reviewed by the individual clinicians and manually corrected when necessary to assure the result resembled manual segmentation.

The Hybrid-Dense-UNet, which achieved state-of-the-art performance on the liver tumor segmentation (LITS) challenge and is open-source, was used to automatically segment the NLP and lesions [109, 110]. The original CNN as trained on the LITS data that was published open-source was used. Lesions which were segmented by the CNN but had no histology were excluded. For lesions that were not segmented by the CNN, but for which histology was available, the segmentation of the radiologist (RAD) was used, resembling implementation in clinical practice. As the Hybrid-Dense-UNet was trained to simultaneously segment the NLP and lesions, this CNN was also used to segment the NLP [109].

11.3.1 Radiomics

From each region of interest (ROI) on the CT, 564 radiomics features were extracted. Features were extracted per segmentation, e.g. for each 3D ROI by each observer. Details can be found in [Section 11.A](#). Based on these features, decision models were created using the workflow for optimal radiomics classification (WORC) toolbox, see [Figure 11.1](#) [36, 72, 73] (i.e., [Chapter 3](#), [Chapter 5](#) and [Chapter 6](#) of this thesis). WORC performs an automated search among a variety of algorithms for each step and determines which combination of algorithms maximizes the prediction performance on the training set. For example, in the machine learning step, one of the eight following algorithms may be used: (1) logistic regression; (2) support vector machines; (3) random forests; (4) naive Bayes; (5) linear discriminant analysis; (6) quadratic discriminant analysis; (7) AdaBoost [61]; and (8) extreme gradient boosting [294]. Details can be found in [Section 11.B](#). The code including all parameters for our experiments has been published open-source [295].

11.3.2 Robustness to segmentation and image acquisition variations

Robustness to segmentation variations was assessed using the intra-class correlation coefficient (ICC) of the features, defining good ($ICC > 0.75$) and excellent ($ICC > 0.90$) reliability [154]. Moreover, the impact of ICC-based feature selection on model performance was assessed by creating models using only these features.

Robustness to variations in the acquisition parameters was assessed by using ComBat [185, 186]. In ComBat, feature distributions are harmonized for variations in the imaging acquisition, e.g. due to differences in hospitals, manufacturers, or acquisition parameters. When dividing the dataset into groups based on these variations, the groups have to remain sufficiently large to estimate the harmonization parameters. In our study, groups were defined based on manufacturer alone or combined with slice thickness (above or below the median) without a moderation variable.

11.3.3 Experimental setup

For each experiment, a 100x random-split cross-validation [63, 64] was performed, randomly splitting the data in each iteration in 80% for training and 20% for test-

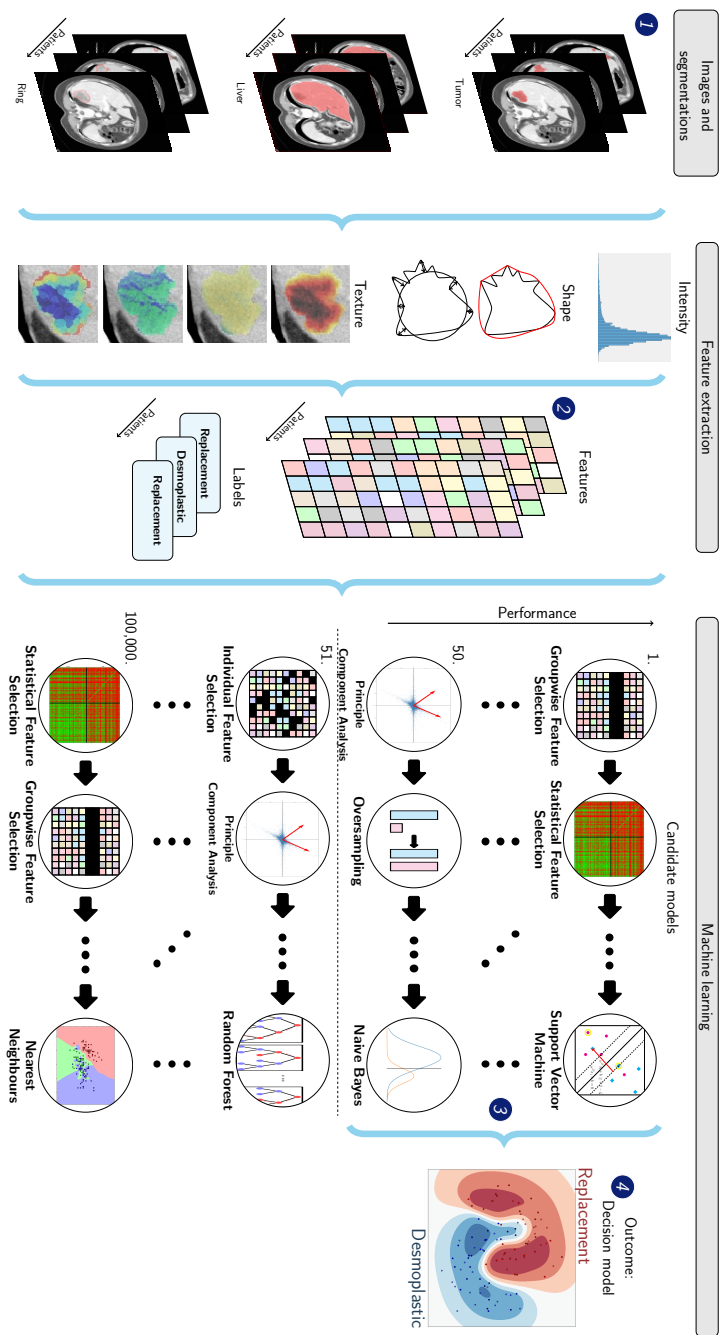


Figure 11.1: Schematic overview of the radiomics approach: adapted from Vos *et al.* [72] (i.e., [Chapter 5](#) of this thesis). Processing steps include segmentation of the lesion and liver, and extraction of the lesion ring (1), feature extraction from the CT based on these regions (2), and the creation of a decision model from the features (4), using an ensemble of the best 50 workflows from 100,000 candidate workflows (3), where the workflows are different combinations of the different processing and analysis steps (e.g. the classifier used).

ing, see [Figure 11.A.2](#). In each iteration, a second, internal 5x random-split cross-validation was performed on the training set, using 85% for training and 15% for validation, where the validation sets were used to optimize the model hyperparameters. Hence, in each iteration, we enforced a strict separation into training, validation and test sets: model construction was performed automatically within the training and validation sets, leaving the test set untouched to minimize the chance of overfitting. The splitting was stratified to maintain a similar dHGP/rHGP ratio in all datasets. Lesions of a patient belonged either all to the training or all to the test dataset.

First, four single-observer radiomics models were created, each using the segmentations of a different observer (STUD2, PhD, RAD, and CNN), but keeping the same observer for training and testing.

Second, a multi-observer radiomics model was trained with segmentations of three observers (STUD2, PhD, and RAD) and tested with segmentations of the fourth, unseen observer (CNN). We hypothesized that a model trained on segmentations from multiple observers may yield a higher performance, and a higher robustness to segmentation variations, as the model is forced to find characteristics shared by all segmentations. For the multi-observer model, the data was split per patient into training and test sets in the same way as in the single-observer model, see [Figure 11.2](#). However, each lesion included in the training set appeared three times, each time with a different segmentation from one of the three observers. The number of training samples was therefore increased to a threefold of the number of training samples used for the single-observer model. This can be seen as a form of data augmentation [296], as compared to the single-observer model, the number of training samples is increased by adding slightly modified copies of the original training samples. Each lesion included in the test set appeared only once, using the segmentation of the CNN.

Third, to estimate model robustness to segmentation and acquisition protocol variations, additional multi-observer models were created using only reliable features (good or excellent) through ICC-based feature selection and ComBat, respectively.

Lastly, features extracted from three other ROIs were evaluated: NLP, and based on the multi-observer setup, NLP plus the lesion, and the lesion border [104, 284], see [Figure 11.A.3](#). Also, to evaluate the predictive value of the clinical characteristics (i.e., 1: age; 2: sex; 3: primary tumor location; 4: primary tumor nodal status; 5: disease free interval; 6: preoperative carcinoembryonic antigen level; 7: CRLM size; and 8: number of CRLMs), two additional HGP prediction models were evaluated using: (1) clinical characteristics (“single-observer”); and (2) imaging and clinical characteristics.

11.3.4 Statistics

The individual predictive values of the radiomics features and the clinical characteristics, and the differences in CT acquisition parameters, were assessed using a Mann–Whitney U test for continuous variables, and a Chi-square test for categorical variables. To this end, the radiomics features extracted from the CNN segmentations were used, as these segmentations were used in the test set in the multi-observer

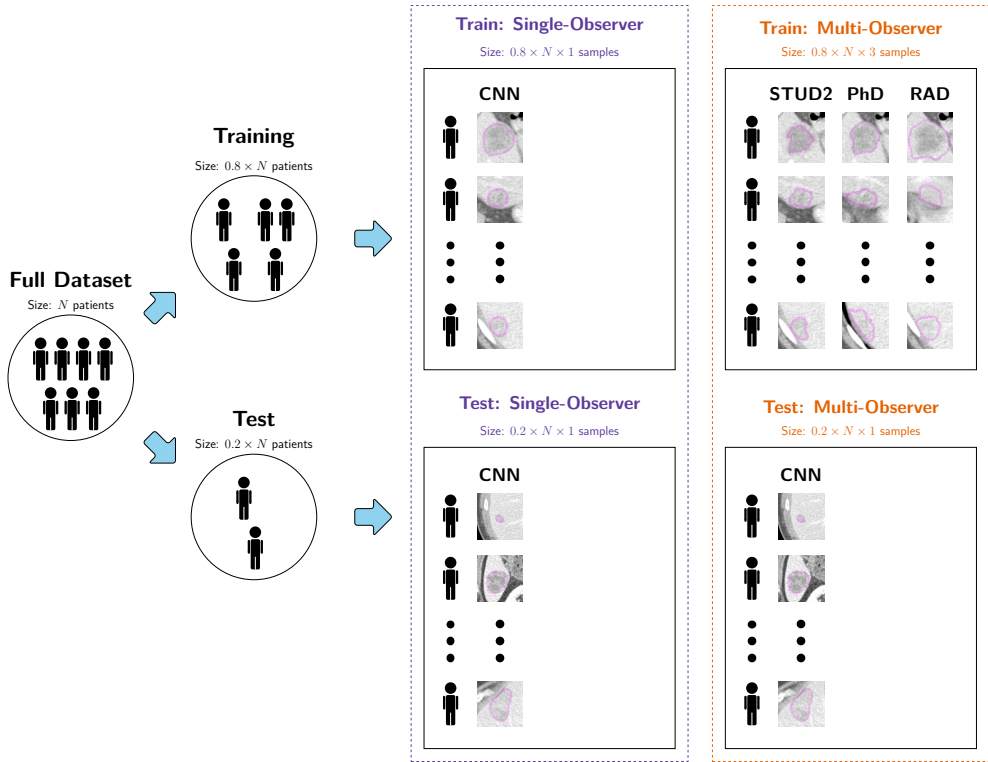


Figure 11.2: Schematic overview of the evaluation setup in a single random-split cross-validation iteration for the single-observer and multi-observer models. For the single-observer models, here illustrated for observer CNN, for both the patients included in the training and in the testing set, each patient appears one time with the segmentation of that single observer. For the multi-observer model, the test set is exactly the same as the single-observer model. However, in the training set, each patient appears three times, each time with a different segmentation from one of the three other observers (STUD2, PhD, and RAD). Hence, in the multi-observer model, the training set size is effectively tripled compared to the single-observer model, while the test set remains unchanged.

models. The p-values of the radiomics features were corrected for multiple testing using the Bonferroni correction (i.e., multiplying the p-values by the number of tests). All p-values were considered statistically significant at a p-value ≤ 0.05 .

Performance was evaluated in the test dataset using accuracy, area under the curve (AUC) of the receiver operating characteristic (ROC) curve, sensitivity, and specificity, averaged over the 100x cross-validation iterations. The corrected resampled t-test was used to construct 95% confidence intervals (CIs), taking into account that the samples in the cross-validation splits are not statistically independent [64]. ROC confidence bands were constructed using fixed-width bands [67]. The positive class was defined as dHGP. The performance estimates in the training dataset are not reported, as these would be too optimistic, since the used methods tend to over-fit on the training dataset [297].

11.4 Results

11.4.1 Dataset

The dataset included 93 lesions (46 dHGP; 47 rHGP) of 76 patients (Table 11.1). The median age was 68 years (interquartile range 60–76 years). No statistically significant differences in clinical characteristics between dHGP and rHGP CRLM were found.

Since the Erasmus MC serves as a tertiary referral, the CT-scans originated from 37 different scanners, resulting in considerable acquisition protocol variations (Table 11.1). The differences in acquisition parameters were not statistically significant, except for pixel spacing ($p = 0.007$, median of 0.78 vs. 0.71 mm). Additionally, nineteen different reconstruction kernels were used, and four manufacturers were present (Siemens: 43, Philips: 16, Toshiba: 16, General Electric: 1).

11.4.2 Segmentation

Lesion segmentation examples are presented in Figure 11.3. The CNN failed to detect 8 of the 93 included lesions (9%), for which the radiologist's segmentation was used. The pairwise DSC to assess the observer segmentation agreement is shown in Table 11.A.1. The intra-observer agreement (DSC of 0.80 for STUD1 and STUD2) was higher than the inter-observer agreement (mean DSC of 0.69 for all other human observers).

11.4.3 Radiomics

In Table 11.2, the performance of the four single-observer models is shown. The mean AUC of all models was above random guessing (0.50), but varied per observer [STUD2: 0.69 (95% CI 0.56–0.82), PhD: 0.66 (95% CI 0.53–0.79), RAD: 0.72 (95% CI 0.59–0.83), and CNN: 0.66 (95% CI 0.54–0.79)]. As the 95% confidence intervals showed substantial overlap, the differences were not statistically significant. Hence, in terms of AUC, the models performed similarly.

In Table 11.3 and Figure 11.4, the multi-observer model performance is shown. Performance was similar [mean AUC of 0.69 (95% CI 0.57–0.81)] to the single-observer models Figure 11.4a. Using only features with good ($N = 263$) [mean AUC of 0.70 (95% CI 0.59–0.81)] or excellent reliability ($N = 166$) [mean AUC of 0.65 (95% CI 0.53–0.77)] across the human observers did not improve the performance (Figure 11.4b). Using ComBat to harmonize the features for manufacturer [mean AUC of 0.64 (95% CI 0.40–0.88)] or protocol [mean AUC of 0.63 (95% CI 0.38–0.87)] differences yielded a minor performance decrease (Figure 11.4c). As there was only one General Electric scan, this scan was omitted from harmonization.

Table 11.4 contains the performances of the models trained on other features, including NLP [mean AUC of 0.65 (95% CI 0.51–0.78)], and based on the multi-observer setup, NLP plus the lesion [mean AUC of 0.63 (95% CI 0.52–0.75)] and the lesion border [mean AUC of 0.67 (95% CI 0.56–0.78)]. Hence, the performance was (slightly) worse than using only lesion features. The model based on clinical characteristics performed similarly to random guessing [mean AUC of 0.56 (95%

Table 11.1: Patient and imaging characteristics of the 76 patients included in this study. P-values are calculated using a Mann–Whitney U test for continuous variables, a chi-square test for continuous variables. P-values in **bold** are deemed significant (< 0.05).

Patient	All	Desmoplastic	Replacement	P-value
Total	76	37 (48.0%)	39 (52.0%)	0.82
Age [†]	68.0 (59.5-75.5)	68.0 (60.0-75.5)	68.0 (59.0-77.0)	
Sex				
Male	44 (57.9%)	24 (64.9%)	20 (51.3%)	0.23
Female	32 (42.1%)	13 (35.1%)	19 (48.7%)	
Primary tumor location				
Right-sided	6 (8.3%)	2 (5.7%)	4 (10.8%)	0.56
Left-sided	29 (54.2%)	21 (60.0%)	18 (48.6%)	
Rectum	27 (37.5%)	12 (34.3%)	15 (40.5%)	
Missing	4			
Nodal status primary tumor				
N0	35 (46.1%)	18 (48.6%)	17 (43.6%)	0.66
N+	41 (53.9%)	19 (51.4%)	22 (56.4%)	
Disease free interval				
≤ 12 months	37 (48.7%)	17 (45.9%)	20 (51.3%)	0.64
≥ 12 months	39 (51.3%)	20 (51.4%)	19 (48.7%)	
Number CRLM				
≤ 1	54 (71.1%)	25 (67.6%)	29 (74.4%)	0.51
≥ 1	22 (28.9%)	12 (34.4%)	10 (25.6%)	
Size largest CRLM				
≤ 5cm	60 (81.1%)	30 (83.3%)	30 (78.9%)	0.63
≥ 5cm	14 (18.9%)	6 (16.7%)	8 (21.1%)	
Missing	2			
CEA*				
≤ 200 µg/L	65 (92.9%)	32 (97.0%)	33 (89.2%)	0.21
≥ 200 µg/L	5 (7.1%)	1 (3.0%)	4 (10.8%)	
Missing	6			
Imaging				
Slice thickness (mm) [†]	5.0 (3.0-5.0)	4.0 (3.0-5.0)	5.0 (3.0-5.0)	0.40
Pixel spacing (mm) [†]	0.74 (0.68-0.78)	0.78 (0.71-0.78)	0.71 (0.67-0.75)	0.007
Tube current (mA) [†]	239 (143-325)	239 (151-305)	232 (135-332)	0.38
Peak kilovoltage [†]	120 (120-120)	120 (120-120)	120 (120-120)	0.09

*Abbreviations: CEA: Carcinoembryonic antigen, CRLM: colorectal liver metastases IQR: interquartile range.

[†] Values are median (Inter quartile range). Other values than those given in the median and inter quartile range may occur.

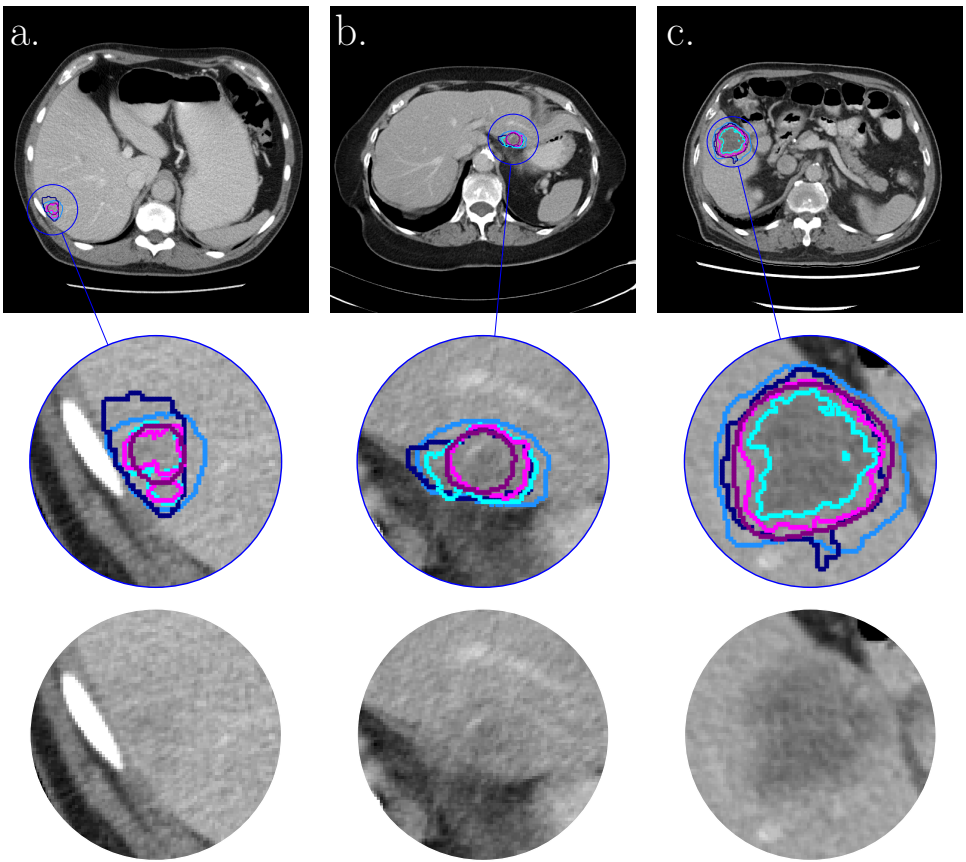


Figure 11.3: Examples of segmentations of three colorectal liver metastases (CRLMs) by the human observers and by the convolutional neural network (CNN) [PhD (dark blue); RAD (light blue); STUD first try (STUD1) (cyan) and second try (STUD2) (magenta); convolutional neural networkCNN (purple)] on a single axial slice of CT-scans. The bottom row depicts the zoomed in region without the segmentation overlays. The three CRLMs displayed are those with a volume at the 25% percentile (a), 50% percentile (b) and 75% percentile (c) of all metastases in the database.

Table 11.2: Performance of the radiomics models using segmentations from single observers (STUD2, PhD, RAD, and CNN) both for the patients in the training sets and the other patients in the test sets. For each metric, the mean and 95% confidence interval over the 100x random-split cross-validation iterations on the test sets are given.

	STUD2	PhD	RAD	CNN
AUC	0.69 [0.56, 0.82]	0.66 [0.53, 0.79]	0.72 [0.59, 0.83]	0.66 [0.54, 0.79]
Accuracy	0.65 [0.55, 0.75]	0.61 [0.50, 0.71]	0.65 [0.55, 0.76]	0.62 [0.52, 0.72]
Sensitivity	0.64 [0.49, 0.80]	0.57 [0.41, 0.72]	0.62 [0.49, 0.76]	0.61 [0.45, 0.76]
Specificity	0.65 [0.48, 0.82]	0.65 [0.49, 0.81]	0.68 [0.52, 0.85]	0.63 [0.47, 0.78]

*Abbreviations: AUC: area under the receiver operator characteristic curve.

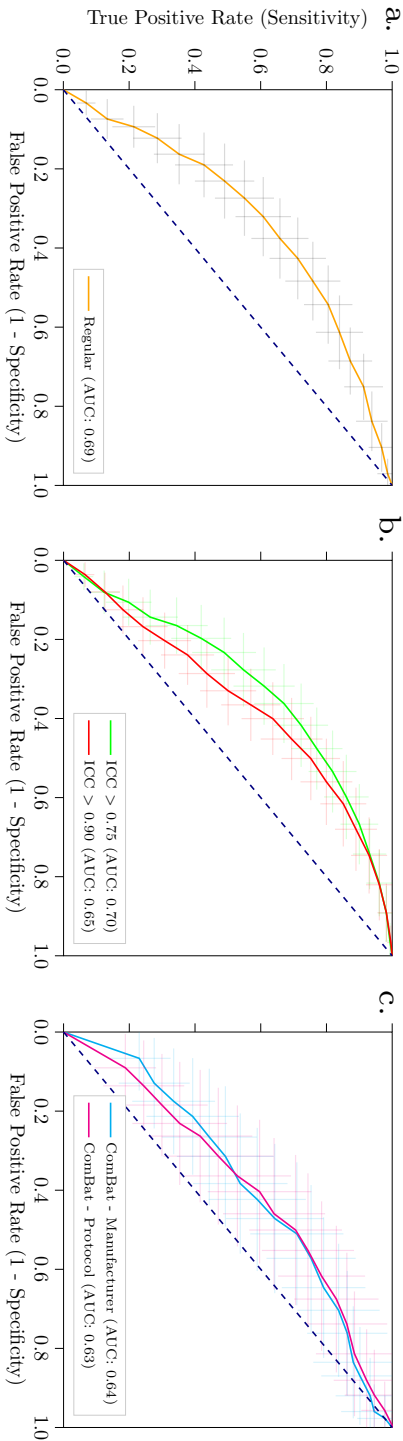


Figure 11.4: Receiver operating characteristic (ROC) curves of the radionics models using segmentations from multiple observers (STUD2, PhD, and RAD) for the patients in the training sets and the segmentations from another observer (CNN) in the other patients in the test sets. These include a the regular model (a); and b using only features with an intra-class correlation coefficient (ICC) larger than 0.75 or 0.90 (b); and c using ComBat to harmonize differences in manufacturer or protocol (c). The crosses indicate the 95% confidence intervals; the curves the means. The dashed lines indicate the performance of random guessing.

Table 11.3: Performance of the radiomics models using segmentations from multiple observers (STUD2, PhD, and RAD) for the patients in the training sets and the segmentations from another observer (CNN) in the other patients in the test sets. The performance is reported for: the regular model; using only features with good (ICC > 0.75) or excellent (ICC > 0.90) reliability; and using ComBat harmonization per manufacturer (Man) or per acquisition protocol (Prot) without a moderation variable. For each metric, the mean and 95% confidence interval over the 100x random-split cross-validation iterations are given.

	Regular	ICC ≥ 0.75	ICC ≥ 0.90	ComBat Man	ComBat Prot
AUC	0.69 [0.57, 0.81]	0.70 [0.59, 0.81]	0.65 [0.53, 0.77]	0.64 [0.40, 0.88]	0.63 [0.38, 0.87]
Accuracy	0.65 [0.54, 0.76]	0.65 [0.55, 0.75]	0.61 [0.50, 0.72]	0.60 [0.41, 0.79]	0.58 [0.39, 0.76]
Sensitivity	0.71 [0.57, 0.86]	0.63 [0.48, 0.78]	0.61 [0.44, 0.77]	0.56 [0.30, 0.82]	0.55 [0.29, 0.81]
Specificity	0.58 [0.41, 0.74]	0.67 [0.51, 0.83]	0.61 [0.45, 0.78]	0.63 [0.33, 0.93]	0.60 [0.29, 0.90]

*Abbreviations: AUC: area under the receiver operator characteristic curve; ICC: intra-class correlation coefficient; Man: Manufacturer; Prot; Protocol.

CI 0.43–0.70]): the model trained on clinical characteristics plus lesion features performed worse than lesion-only [mean AUC of 0.65 (95% CI 0.53–0.77)].

After Bonferroni correction for multiple testing, from the 564 features extracted using the CNN segmentations, only four texture features derived from Gabor filters were found to have statistically significant p-values (0.035–0.010).

11.5 Discussion

The aim of this pilot was to evaluate whether radiomics can distinguish pure dHGP from pure rHGP based on CT-scans and to evaluate its robustness to segmentation and acquisition protocol variations. Despite these variations, our results suggest that radiomics features have predictive value in distinguishing pure HGP on CT-scans, but that caution is warranted when drawing conclusions about the clinical applicability at this stage.

Currently, HGPs can only be determined after surgery using resection specimens. Our radiomics approach may overcome this gap. Preoperative HGP assessment may give an earlier estimate of disease aggressiveness and prognosis, thus improving patient care [285]. A previous study found a 5-year overall survival of 78% in dHGP patients compared to 37% ($p < 0.001$) in patients with other HGPs [284]. Preoperative assessment of HGPs may even imply a practice change, as HGPs may be associated

Table 11.4: Performance of models using features other than only lesion features. These features were extracted from a segmentation of the normal liver parenchyma (NLP); the NLP and the lesion (NLP + Lesion); a ring at the border of the segmentation (Ring); using the clinical characteristics (Clinical); and the clinical characteristics combined with lesion features (Clinical + Lesion).

Metric	NLP	NLP+Lesion	Ring	Clinical	Clinical+Lesion
AUC	0.65 [0.51, 0.78]	0.63 [0.52, 0.75]	0.67 [0.56, 0.78]	0.56 [0.43, 0.70]	0.65 [0.53, 0.77]
Accuracy	0.59 [0.49, 0.70]	0.60 [0.50, 0.71]	0.63 [0.54, 0.73]	0.53 [0.41, 0.64]	0.62 [0.51, 0.72]
Sensitivity	0.52 [0.33, 0.70]	0.60 [0.43, 0.76]	0.67 [0.51, 0.83]	0.56 [0.37, 0.75]	0.62 [0.45, 0.79]
Specificity	0.67 [0.50, 0.85]	0.61 [0.46, 0.75]	0.59 [0.45, 0.74]	0.49 [0.31, 0.67]	0.61 [0.45, 0.77]

*Abbreviations: AUC: area under the receiver operating characteristic curve; NLP: normal liver parenchyma; CNN: convolutional neural network.

with efficacy of systemic chemotherapy [104, 282, 283, 284]. Hence, preoperative HGP assessment through radiomics may also be used predictively to select patients which may benefit from chemotherapy. Moreover, preoperative HGP assessment may enable others to study the full potential of HGP as a biomarker [285]. Although it is difficult at this stage to decide on the accuracy of radiomics-based HGP prediction required for clinical practice, the current performance is likely not sufficient yet and further improvements are warranted.

Our secondary aim was to evaluate and improve the robustness of radiomics to segmentation and acquisition protocol variations. Our results indicate substantial differences between the segmentations. In spite of these differences, our multi-observer model generalized well to segmentations of an unseen “observer”, i.e., the automated CNN. Generally, improving model robustness to segmentation variations is done by selecting only reliable features, i.e., high ICC across multi-observer segmentations [291, 292, 293]. However, in our results, this did not alter the performance, indicating that training on multiple observers already enforced model robustness to segmentation variations. As the unseen observer was a CNN, our combined approach (CNN for segmentation, radiomics for classification) is fully automatic and observer independent. It must be pointed out that, although we used a state-of-the-art CNN ranking second in the renowned LITS challenge [110], 8 lesions (9%) were missed by the CNN. These required manual correction, making the method actually semi-automatic in this minority of cases. The radiologist however initially also missed 19 lesions (20%), which were later corrected based on the pathology outcome, indicating that human observers also miss lesions. Of these 19 lesions, 16 were detected by the CNN. This indicates that the CNN may aid identifying false negatives from the radiologists. However, the CNN detected 257 abnormalities in total, likely including a large number of false positives, which would require correction by the radiologist. Future studies should systematically compare the hit and miss ratios of radiologists and the CNN. Nonetheless, we believe the method’s large degree of automation and its observer independence are highly desirable aspects for use in clinical practice.

Visual inspection of the lesions indicated that the radiologist’s segmentations showed the largest difference with the CNN segmentations. In addition, the radiologist’s segmentations had the lowest overlap (in terms of DSC) with the other observers. Visual inspection indicated that the radiologist generally drew a loose outline around the lesion, and thus ROIs with a relatively large area, while the CNN generally drew conservative outlines, thus ROIs with a relatively small area. Caution should be taken when drawing conclusions, as we only compared ROIs of a single radiologist with the CNN. Moreover, as annotating lesion boundaries is not part of routine clinical practice of radiologists, their segmentations cannot be considered as the ground truth. Additionally, we evaluated models using features extracted from several ROIs to investigate where the most relevant HGP information is. The NLP model performed worse than the lesion-only models. As HGPs are represented at the liver tissue and lesion interface, we expected the combination or usage of the border to be optimal. However, combining these features, or using the border, did not yield an improvement over the lesion-only model. This may be attributed to the fact that determination of the exact border of the lesion is difficult. Our radiomics model uses a more data-driven approach, using 564 features extracted not only from

the lesion boundary but from the full lesion segmentation, and machine learning to determine what information is most relevant. Our results suggest that the lesion itself contains the most informative features. The clinical characteristics did not yield any predictive value on their own, nor added predictive value when combined with the radiomics features. This is in line with the literature, as to our knowledge, no pre-operative biomarkers for HGPs based on clinical characteristics have so far been described [285].

Recently, the value of radiomics to predict HGPs was assessed by Cheng *et al.* [290] using the former consensus guidelines [281]. This study included 126 CRLMs, using for each patient a pre- and post-contrast arterial and PVP CT-scan. An AUC of 0.93 in the training and 0.94 in the validation set was reported, which was much higher than the performance in our study. This difference may be attributed to various factors in the study design. First, we used the more recent clinical guidelines and included only pure HGPs, instead of the previous cut-off of $> 50\%$ of a single HGP [281, 284]. There may be considerable uncertainty in the scoring of pure HGPs, e.g. other HGP types may be missed due to sampling errors [281]. Some cases could be misclassified due to this possible missing information, limiting our performance. Second, Cheng *et al.* [290] used multiple CT-scans per patient: an AUC of 0.79 was obtained in the used validation set when only using the PVP, as we did. Also, we used a multi-center CT dataset with much acquisition protocol heterogeneity, while Cheng *et al.* [290] used a two-center dataset with comparable acquisition protocols. Moreover, our radiomics approach is different, e.g. we used a fully automatic approach optimized on the training set, while the optimization protocol used by Cheng *et al.* [290] is not explicitly mentioned.

There are several limitations to this study. First, our dataset included only pure dHGP or rHGP patients, while mixed and a rare third HGP (pushing) exist as well. The strict selection resulted in a small sample size, which may explain the wide CIs. Due to the large width of the CIs, i.e., the AUCs generally spanned between 15–30% of the range, few claims could be made regarding statistical significance of differences between models. No claims can be made about the performance of the model on mixed HGPs or the pushing HGP. Future studies should include mixed HGPs, which will lead to a larger dataset, and will improve clinical applicability.

Second, we used PVP contrast-enhanced CT-scans, as this was mostly used in clinical routine. Addition of other contrast phases, positron emission tomography or magnetic resonance imaging, may improve the performance [290, 298, 299].

Third, while our CNN produced segmentations similar to the human observers as indicated by the DSC, 8 out of the 93 included lesions were missed. As the CNN segmentations are similar to those of the radiologist and our multi-observer model is robust to segmentation variations, replacing the missed segmentations with the radiologist's is not expected to have substantially influenced our results.

Lastly, our imaging models were trained and evaluated on a multi-center, heterogeneous dataset. On one hand, this is a strength of our study, as the models had predictive value despite substantial acquisition variations. However, heterogeneity may have (negatively) affected our performance. The use of ComBat to compensate for manufacturer variations did not lead to a substantial improvement in prediction accuracy. Additional experiments with ComBat using the HGP as a

“moderation variable” showed a near perfect performance; however, such use of the HGP as a moderation variable in the ComBat algorithm is a form of overfitting, as it uses the ground truth HGP data of the full dataset (including the test set), and it tends to give too optimistic performance estimates. Future research could explore other methods to compensate for manufacturer variations on the one hand while maintaining the distinction between HGPs on the other hand. Alternatively, using a single-scanner study will limit the generalizability, but may positively impact the performance. Additionally, although we used a multi-center dataset, we did not perform an independent, external validation. However, we used a rigorous cross-validation, separating the data 100x in training and testing parts. Hence, as our radiomics approach was optimized on the training set only, the chance of overestimating performance due to “over-engineering” was limited.

Future research could include HGP classification using CNNs. While our current method is largely observer independent, classification without use of any segmentation would be truly observer independent. Also, only four lesion feature showed statistically significant differences between the dHGP and rHGP lesions, suggesting that these features may not be optimal for distinguishing these HGPs. The CNN used for segmentation in our study was not designed for HGP prediction, but rather segmentation of the liver and various liver abnormalities. Features learned by a dedicated classification CNN for HGP prediction may yield more predictive value than the features learned by our segmentation CNN or the generic radiomics features used in our study. This would probably require a larger dataset to learn from.

11.6 Conclusions

Our combination of deep learning for segmentation and radiomics for classification shows potential for automatically distinguishing pure dHGPs from rHGPs of CRLM on CT-scans. The model is observer independent and robust to segmentation variations. However, the current performance is likely not sufficient yet and further improvements are warranted, including extension to mixed HGPs, and external validation. Pending further optimization, radiomics may serve as a non-invasive, preoperative addition to postoperative HGP assessment, enabling pre-operative response prediction, response evaluation, and further studies on HGP as a pre-operative biomarker.

Appendix

Appendix 11.A Feature extraction

This supplementary material is similar to Vos *et al.*, Timbergen *et al.* [72, 73] (i.e., Chapter 5 and Chapter 6 of this thesis), but details relevant for the current study are highlighted.

A total of 564 radiomics features quantifying intensity, shape, orientation and texture were extracted. These features were extracted using the defaults for CT scans from the Workflow for Optimal Radiomics Classification (WORC) toolbox [36],

which internally uses the PREDICT [51] and PyRadiomics [44] feature extraction toolboxes. The code to extract the features for this specific study has been published open-source [295]. An overview of all features is depicted in Table 11.A.2. For details on the mathematical formulation of the features, we refer the reader to Zwanenburg *et al.* [39]. More details on the extracted features can be found in the documentation of the respective toolboxes, mainly the WORC documentation [68].

Before feature extraction, conversion of the CT scan intensities to Hounsfield Units (HU) was performed. The features can be divided in several groups. Intensity features were extracted using the histogram of all intensity values within the ROIs and included several first-order statistics such as the mean, standard deviation and kurtosis. These describe the distribution of Hounsfield units within the lesion. Shape features were extracted based only on the ROI, i.e. not using the image, and included shape descriptions such as the volume, compactness and circular variance. These describe the morphological properties of the lesion. Orientation features were used to describe the orientation of the ROI, i.e. not using the image. Lastly, texture features were extracted using Gabor filters, Laplacian of Gaussian filters, vessel (i.e. tubular structures) filters [54], the Gray Level Co-occurrence Matrix [39], the Gray Level Size Zone Matrix [39], the Gray Level Run Length Matrix [39], the Gray Level Dependence Matrix [39], the Neighbourhood Grey Tone Difference Matrix [39], Local Binary Patterns [52], and local phase filters [53, 300]. These features describe more complex patterns within the lesion, such as heterogeneity, occurrence of blob-like structures, and presence of line patterns.

Most of the texture features include parameters to be set for the extraction. Beforehand the values of the parameters that will result in features with the highest discriminative power for the classification at hand (i.e., dHGP versus rHGP) are not known. Including these parameters in the workflow optimization, see Section 11.B, would lead to repeated computation of the features, resulting in a redundant decrease in computation time. Therefore, alternatively, these features are extracted at a range of parameters as is default in WORC. The hypothesis is that the features with high discriminative power will be selected by the feature selection methods and/or the machine learning methods as described in Section 11.B. The parameters used are described in the caption of Table 11.A.2.

The imaging data used in this study is multi-center, and therefore heterogeneous in terms of acquisition protocols. Especially the variations in slice thickness may cause feature values to be highly dependent on the acquisition protocol. Hence, extracting robust 3D features may be hampered by these variations, especially for low resolutions. The images were not resampled, as this would result in interpolation errors. To overcome this issue, all features were extracted per 2D axial slice and aggregated over all slices. Afterwards, several first-order statistics over the feature distributions were evaluated and used in the machine learning approach. As all images had the same unit (Hounsfield), no additional normalization was applied.

Appendix 11.B Model optimization

This supplementary material is similar to Vos *et al.*, Timbergen *et al.* [72, 73] (i.e., Chapter 5 and Chapter 6 of this thesis), but details relevant for the current study are

highlighted.

The Workflow for Optimal Radiomics Classification (WORC) toolbox [68] makes use automated machine learning to create the optimal performing workflow from a variety of algorithms. Besides deciding whether to use an algorithm, most algorithms require hyperparameters, i.e., parameters that need to be set before the actual learning step, to be tuned to enhance the performance. WORC defines a workflow as a specific sequential combination of algorithms and their respective hyperparameters. In WORC, the radiomics workflow is split into the following components: image and segmentation preprocessing, feature extraction, feature and sample preprocessing, and machine learning. For each component, a collection of algorithms and their associated hyperparameters is included. Given this search space, WORC uses automated machine learning to find the optimal solution. The code to use WORC for creating the decision models in this specific study has been published open-source [295].

The workflows could be constructed from the following default search space in WORC, which components can only be combined in the order listed below:

1. Feature group selection: a group-wise search, in which specific groups of features (i.e., intensity, shape, and the subgroups of texture features as defined in Table 11.A.2) are selected or deleted. To this end, each feature group had an on/off variable which is randomly activated or deactivated, which were all included as hyperparameters in the optimization.
2. Feature imputation: when a feature could not be computed, e.g. a lesion is too small for a specific feature to be extracted, a feature imputation algorithm was used to estimate replacement values for the missing values. Strategies for imputation included 1) the mean; 2) the median; 3) the mode; 4) a constant (default: zero); and 5) a nearest neighbor approach.
3. Feature selection: a variance threshold, in which features with a low variance (<0.01) are removed. This method was always used, as this serves as a feature sanity check with almost zero risk of removing relevant features.
4. Feature scaling was performed to make all features have the same scale, as otherwise the machine learning methods may focus only on those features with large values. This was done through z-scoring, i.e., subtracting the mean value followed by division by the standard deviation, for each individual feature. A robust version of z-scoring was used, in which outliers, i.e., values below the 5th percentile or above the 95th percentile, were excluded from computing the mean and variance.
5. Feature selection: optionally, the RELIEF method [55], which ranks the features according the differences between neighboring samples. Features with more differences between neighbors of different classes (i.e., dHGP versus rHGP) are considered higher in rank.
6. Feature selection: optionally, features are selected by training a machine learning model and selecting features that are regarded important by the model.

Hence the used model should be able to give the features an importance weight. Included model choices are LASSO, logistic regression, and a random forest.

7. Dimensionality reduction: optionally, principal component analysis (PCA) is used, in which either only those linear combinations of features were kept which explained 95% of the variance in the features or a limited number of components (between 10 – 50).
8. Feature selection: optionally, individual features were selected through univariate testing. To this end, for each feature, a Mann-Whitney U test was performed to test for significant differences in distribution between the labels (i.e., dHGP versus rHGP). Afterwards, only features with a p-value above a certain threshold were selected.
9. Resampling: optionally, a resampling strategy could be used, which was used to overcome class imbalances and reduce overfitting on specific training samples. Various methods from the imbalanced-learn toolbox [57] could be used: random over-sampling, random under-sampling, near-miss resampling, the neighborhood cleaning rule, ADASYN, and SMOTE (regular, borderline, Tomek and the edited nearest neighbors variant).
10. Machine learning: lastly, a machine learning methods was used to determine a decision rule to distinguish the classes. Methods included were; 1) logistic regression; 2) support vector machines; 3) random forests; 4) naive Bayes; 5) linear discriminant analysis; 6) quadratic discriminant analysis; 7) AdaBoost [61]; and 8) extreme gradient boosting [62].

The performance of WORC was evaluated in this study through a 100x random-split cross-validation [63, 64], in each iteration splitting the data in 80% for training and 20% for testing. In each cross-validation iteration, all optimization was performed on the training set in order to prevent overfitting on the test set. To prevent overfitting on the training dataset, a 5x random-split stratified cross-validation was performed within the training dataset as well, using 85% for model training and 15% for model validation, see Figure 11.A.2.

WORC states the radiomics workflow as a combined algorithm selection and hyperparameter optimization problem (CASH), as algorithm selection and hyperparameter optimization are often not independent [34]. In each training-test cross-validation iteration, CASH optimization is performed within the training dataset by testing thousand pseudo-randomly generated radiomics workflows from the above search space. These are trained on the five training sets in the 5x random-split training-validation cross-validation, and ranked according to their mean performance on the five validation datasets. As performance metric, the weighted F1-score is used, which is the weighted harmonic average of the precision and recall.

Using only the single workflow that on average performs best on the validation sets may result in poor generalization due to overfitting on the validation sets. Hence, an ensemble was constructed by combining the workflows that perform best on the validation sets [50]. Ensembling was done using the default of WORC by averaging the posteriors of the 100 best workflows.

The following pseudo code illustrates the algorithm of WORC:

- **For** each 100x random-split training-test cross-validation iteration:
 - **Do**: Construct the training dataset by randomly selecting 80% of the patients.
 - **Do**: On this training dataset, define 5x random-split cross-validation splits, selecting in each iteration 85% of the patients for training and 15% for validation.
 - **Do**: Pseudo-randomly sample 1,000 workflows from the search space.
 - **For** each of the 1,000 sampled workflows:
 - * **Do**: Train the workflow on the five training datasets in the 5x random-split cross-validation.
 - * **Do**: Compute the mean weighted F1-score on the corresponding five validation datasets in the 5x random-split cross-validation.
 - **Do**: Rank the 1,000 workflows, retrain the best 100 workflows on the full training set, and combine them into an ensemble model.
 - **Do**: Evaluate the ensemble model on the test dataset, i.e., the remaining 20% of the patients that were not included in the training dataset.

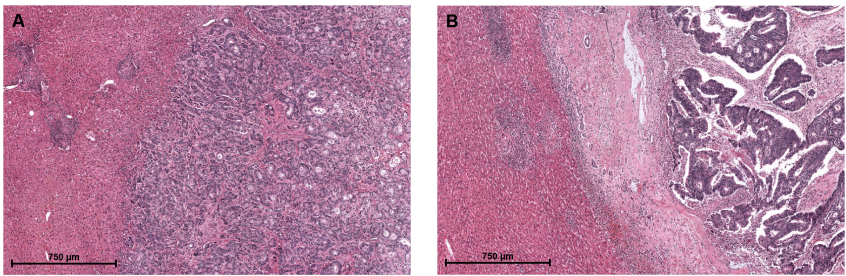


Figure 11.A.1: Replacement type (A) and desmoplastic type (B) HGP on hematoxylin and eosin stained tissue sections.

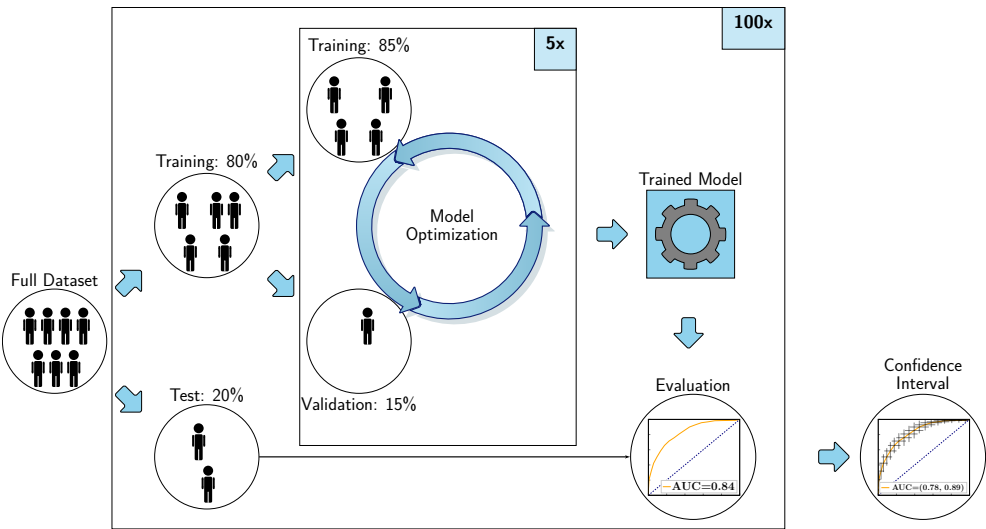


Figure 11.A.2: Visualization of the 100x random split cross-validation, including a second cross-validation within the training set for model optimization. The test dataset is only used for evaluation of the trained model.

Table 11.A.1: Segmentation agreement expressed in Dice Similarity Coefficient (DSC) (mean (standard deviation)) between the observers and the convolutional neural network (CNN) (STUD (1st and 2nd time), PhD, RAD, CNN). The average of the mean and standard deviation of the DSC for each observer are stated in the bottom row.

Observer	STUD1	STUD2	PhD	RAD	CNN
STUD1	-	0.80 (0.15)	0.73 (0.14)	0.60 (0.18)	0.65 (0.26)
STUD2	0.80 (0.15)	-	0.77 (0.13)	0.63 (0.18)	0.66 (0.27)
PhD	0.73 (0.14)	0.77 (0.13)	-	0.69 (0.16)	0.63 (0.25)
RAD	0.60 (0.18)	0.63 (0.18)	0.69 (0.16)	-	0.58 (0.27)
CNN	0.65 (0.26)	0.66 (0.27)	0.63 (0.25)	0.58 (0.27)	-
Average	0.70 (0.18)	0.72 (0.18)	0.71 (0.17)	0.63 (0.20)	0.63 (0.26)

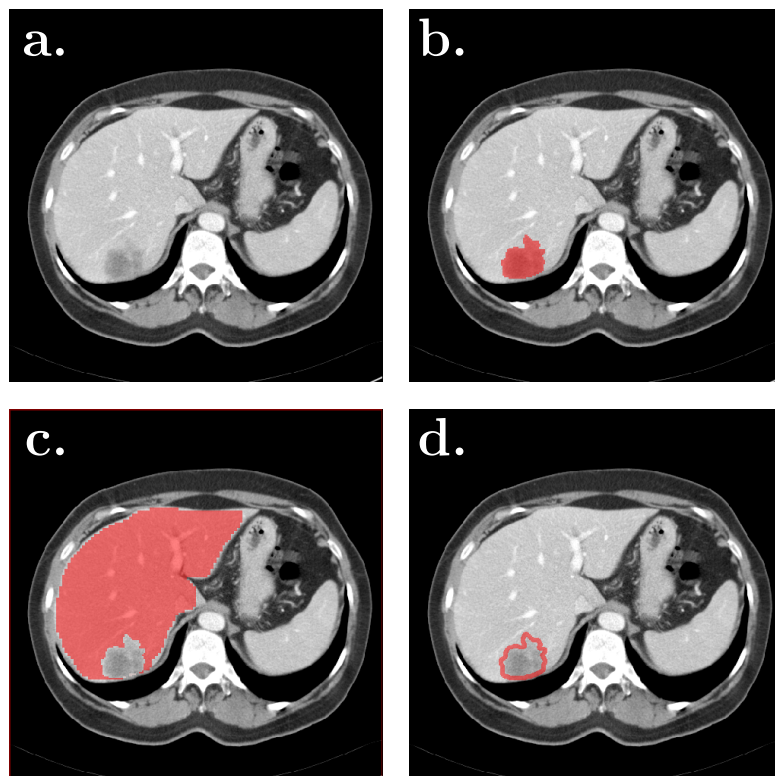
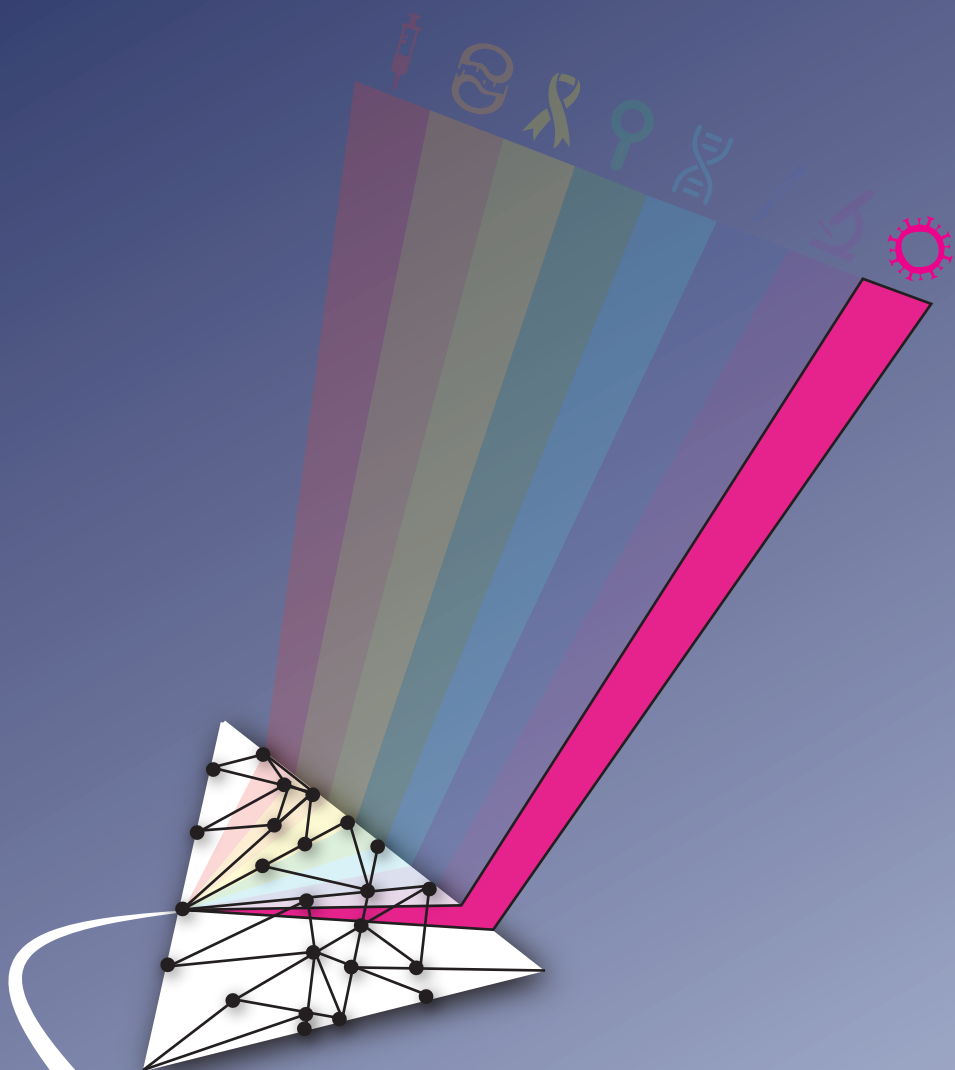
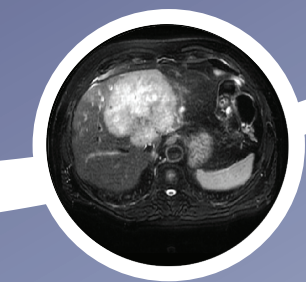


Figure 11.A.3: Examples of segmentations of various regions of interest on a single axial slice of CT-scans. A: CT-scan without segmentation; B: lesion; C: normal liver parenchyma; and D: ring on the border between the lesion and normal liver parenchyma.

Table 11.A.2: Overview of the 564 features used in this study. GLCM features were calculated in four different directions (0, 45, 90, 135 degrees) using 16 gray levels and pixel distances of 1 and 3. LBP features were calculated using the following three parameter combinations: 1 pixel radius and 8 neighbours, 2 pixel radius and 12 neighbours, and 3 pixel radius and 16 neighbours. Gabor features were calculated using three different frequencies (0.05, 0.2, 0.5) and four different angles (0, 45, 90, 135 degrees). LoG features were calculated using three different widths of the Gaussian (1, 5 and 10 pixels). Vessel features were calculated using the full mask, the edge, and the inner region. Local phase features were calculated on the monogenic phase, phase congruency and phase symmetry.

Histogram (13 features)	LoG (13*3=39 features)	Vessel (12*3=39 features)	GLCM (MS) (6*3*4*2=144 features)	Gabor (13*3*3=156 features)	NGTDM (5 features)	LBP (13*3=39 features)
min	min	min	contrast (normal, MS mean + std)	min	busyness	min
max	max	max	dissimilarity (normal, MS mean + std)	max	coarseness	max
mean	mean	mean	homogeneity (normal, MS mean + std)	mean	complexity	mean
median	median	median	angular second moment (ASM) (normal, MS mean + std)	median	contrast	median
std	std	std	energy (normal, MS mean + std)	std	strength	std
skewness	skewness	skewness	correlation (normal, MS mean + std)	skewness		skewness
kurtosis	kurtosis	kurtosis		kurtosis		kurtosis
peak	peak	peak		peak		peak
peak position	peak position	peak position		peak position		peak position
range	range	range		range		range
energy	energy	energy		energy		energy
quartile range	quartile range	quartile range		quartile range		quartile range
entropy	entropy	entropy		entropy		entropy
GLSZM (16 features)	GLRM (16 features)	GLDM (14 features)	Shape (35 features)	Orientation (9 features)	Local phase (13*3=39 features)	
Gray Level Non Uniformity	Gray Level Non Uniformity	Dependence Entropy	compactness (mean + std)	theta_x	min	
Gray Level Non Uniformity Normalized	Gray Level Non Uniformity Normalized	Dependence Non-Uniformity	radial distance (mean + std)	theta_y	max	
Gray Level Variance	Gray Level Variance	Dependence Non-Uniformity Normalized	roughness (mean + std)	theta_z	mean	
High Gray Level Zone Emphasis	High Gray Level Run Emphasis	Dependence Variance	convexity (mean + std)	COM index x	median	
Large Area Emphasis	Long Run Emphasis	Gray Level Non-Uniformity	circular variance (mean + std)	COM index y	std	
Large Area High Gray Level Emphasis	Long Run High Gray Level Emphasis	Gray Level Variance	principal axes ratio (mean + std)	COM index z	skewness	
Large Area Low Gray Level Emphasis	Long Run Low Gray Level Emphasis	High Gray Level Emphasis	elliptic variance (mean + std)	COM x	kurtosis	
Low Gray Level Zone Emphasis	Low Gray Level Run Emphasis	Large Dependence Emphasis	solidity (mean + std)	COM y	peak	
SizeZoneNonUniformity	RunEntropy	Large Dependence High Gray Level Emphasis	area (mean, std, min + max)	COM z	peak position	
SizeZoneNonUniformityNormalized	RunLengthNonUniformity	Large Dependence Low Gray Level Emphasis	volume (total, mesh, volume)		range	
SmallAreaEmphasis	RunLengthNonUniformityNormalized	Low Gray Level Emphasis	elongation		energy	
SmallAreaHighGrayLevelEmphasis	RunPercentage	Small Dependence Emphasis	flatness		quartile	
SmallAreaLowGrayLevelEmphasis	RunVariance	Small Dependence High Gray Level Emphasis	least axis length		entropy	
ZoneEntropy	ShortRunEmphasis	Small Dependence Low Gray Level Emphasis	major axis length			
ZonePercentage	ShortRunHighGrayLevelEmphasis		minor axis length			
ZoneVariance	ShortRunLowGrayLevelEmphasis		maximum diameter 3D			
			maximum diameter 2D (rows, columns, slices)			
			sphericity			
			surface area			
			surface volume ratio			

*Abbreviations: COM: center of mass; GLCM: gray level co-occurrence matrix; MS: multi slice; NGTDM: neighborhood gray tone difference matrix; GLSZM: gray level size zone matrix; GLRLM: gray level run length matrix; LBP: local binary patterns; LoG: Laplacian of Gaussian; std: standard deviation.



12.

Automated differentiation of malignant and benign primary solid liver lesions on MRI: an externally validated radiomics model

Based on: **M. P. A. Starmans**, R. L. Miclea, V. Vilgrain, M. Ronot, Y. Purcell, J. Verbeek, W. J. Niessen, J. N. Ijzermans, R. A. de Man, M. Doukas, S. Klein*, and M. G. Thomeer*, "Automated differentiation of malignant and benign primary solid liver lesions on MRI: An externally validated radiomics model," *Submitted*. medrxiv: [2021.08.10.21261827](https://doi.org/10.1101/2021.08.10.21261827)

* indicates equal contributions

Abstract

Background & Aims: Distinguishing malignant from benign primary solid liver lesions is highly important for treatment planning. However, diagnosis on radiological imaging is challenging. In this study, we developed a radiomics model based on magnetic resonance imaging (MRI) to distinguish the most common malignant and benign primary solid liver lesions, and externally validated the model in two centers.

Approach & Results: Datasets were retrospectively collected from three tertiary referral centers (A, B and C) including data from affiliated hospitals sent for revision. Patients with malignant (hepatocellular carcinoma and intrahepatic cholangiocarcinoma) and benign (hepatocellular adenoma and focal nodular hyperplasia) lesions were included. For each patient, only a T2-weighted MRI was included. A radiomics model was developed on dataset A using a combination of machine learning approaches, and internally evaluated on dataset A through cross-validation. Next, the model was externally validated on datasets B and C, and compared to scoring by two experienced abdominal radiologists on dataset C. In the resulting dataset, in total, 486 patients were included (A: 187, B: 98 and C: 201). Despite substantial MRI acquisition heterogeneity, the radiomics model developed on dataset A had a mean area under the receiver operating characteristic curve (AUC) of 0.78 in the internal validation on dataset A, and a similar AUC in the external validations (B: 0.74, C: 0.76). In dataset C, the two radiologists showed moderate agreement (Cohen's κ : 0.61) and achieved AUCs of 0.86 and 0.82, respectively.

Conclusions: Our radiomics model using T2-weighted MRI only can non-invasively distinguish malignant from benign primary solid liver lesions. External validation indicated that our model is generalizable despite substantial differences in the acquisition protocols.

12.1 Introduction

Liver cancer is the seventh most commonly diagnosed cancer and the third most common cause of cancer deaths worldwide, with approximately 906,000 estimated new cases and 830,000 deaths in 2020 [301]. One of the most important tasks in routine clinical practice is making the distinction between malignant and benign primary solid liver lesions, which substantially influences treatment planning [302, 303]. Commonly, a first assessment is made by the radiologist based on magnetic resonance imaging (MRI). Guidelines such as those from the European Association for the Study of the Liver (EASL) [107, 304] may aid the radiologist. Typically, a mixture of T2-weighted, T1-weighted, dynamic contrast enhanced MRI, diffusion weighted imaging, and the apparent diffusion coefficient (ADC) is used. The diagnosis is often challenging due to the wide variety of liver lesion phenotypes, sizes, and appearances [12], and lack of a clear assessment consensus [305].

Patients from peripheral centers may therefore be referred to tertiary centers for reassessment. This trajectory is time consuming and expensive, while a quick and accurate diagnosis is crucial for the treatment planning. Often, despite imaging, a biopsy may be performed to make the final diagnosis, as indicated by the EASL guidelines. While accurate, biopsies are (minimally) invasive, can be technically challenging, and bring risks such as bleeding and tumor seeding to the patient [306]. Patient treatment may benefit from a non-invasive tool to shorten time to diagnosis by enabling quicker referral, refining patient selection prior to biopsies, and assist diagnosing patients who do not require a biopsy.

In recent years, radiomics, i.e., the use of a large number of quantitative medical imaging features to predict clinical outcomes, has been successfully used in various clinical areas [16, 23], [19] (i.e., Chapter 2 of this thesis). In liver cancer, this has been mostly based on computed tomography to make predictions such as survival, prognosis, and recurrence [286, 288, 307]. For MRI in liver cancer, radiomics has been used to classify focal liver lesions [105, 308, 309, 310], and as LI-RADS [311] surrogate [312]. Radiomics thus shows potential for usage in liver lesion characterization.

However, as concluded in a recent review, the use of radiomics for liver lesion characterization is still at an early stage [313]. First, there is a need for large, multi-center cohorts, especially for external validation [24, 25, 30]. Second, a major challenge is the lack of image acquisition standardization [313], as radiomics methods are generally sensitive to acquisition variations [18], underlining the need for external validation. Rather than requiring a comprehensive, standardized set of multiple MRI sequences, usage of a single sequence would make radiomics models more universally applicable in a routine clinical setting.

The primary aim of this study was therefore to develop a radiomics model based on only T2-weighted MRI to distinguish between the most common malignant and benign primary solid liver lesions, and to externally validate the model in two multi-center cohorts. We used only T2-weighted MRI, as this sequence is widely available, reliable for lesion segmentation, minimally sensitive to motion or breathing artefacts, and informative [107, 304, 311]. Our secondary aim was to compare the performance of radiomics to clinical practice through visual scoring of the lesions by two experienced abdominal radiologists.

12.2 Materials and methods

12.2.1 Data collection

Approval for this study by the institutional review boards of Erasmus MC (Rotterdam, the Netherlands) (MEC-2017-1035), Maastricht UMC+ (Maastricht, the Netherlands) (METC 2018-0742), and Hôpital Beaujon (Paris, France) (N° 2018-002) was obtained. Informed consent was waived due to the use of retrospective, anonymized data. The study protocol conformed to the ethical guidelines of the 1975 Declaration of Helsinki.

Three datasets were collected retrospectively from three tertiary referral centers: all patients diagnosed or referred to A) Erasmus MC between 2002 - 2018; B) Maastricht UMC+ between 2005 - 2018; and C) Hôpital Beaujon, included in reverse chronological order starting at 2018, until in total 201 patients were identified, in accordance with the inclusion and exclusion criteria described below. Imaging data, age, sex, and phenotype were collected for each patient.

Inclusion criteria were: hepatocellular carcinoma (HCC), intrahepatic cholangiocarcinoma (iCCA), hepatocellular adenoma (HCA) or focal nodular hyperplasia (FNH); pathologically proven phenotype, except for “typical” FNH; and availability of a T2-weighted MRI scan. Exclusion criteria were: maximum diameter equal to or smaller than 3 cm; underlying liver disease; and significant imaging artefacts. Details on the pathological examination are given in [Section 12.A](#).

Malignant lesions included HCC (75 - 85% of primary liver cancers), and iCCA (10 - 15% of primary liver cancers) [12]. Benign lesions included HCA (3 - 4 cases per 100,000 person-years in Europe and North America) and FNH (found in 0.8% of all adult autopsies) [12]. The most common benign primary liver lesions, hemangioma, were not included as these are nonsolid and often relatively easy to diagnose on imaging [12, 107]. Only lesions with a pathologically proven phenotype were included to ensure an objective ground truth. Pathological analysis for each patient was performed locally in their admission hospital. An exception was made for typical FNH [12], which are routinely not biopsied and diagnosed radiologically [106], as typical FNH imaging characteristics are 100% specific [12]. Not including these would create a selection bias towards “atypical” FNH: the model performance would then only be evaluated on atypical FNH, and no claims could be made on the performance in typical FNH. In patients with multiple lesions, only the largest one was included.

Patients with underlying liver disease due to alcohol, hepatitis, and vascular liver disease, such as fibrosis or cirrhosis, were excluded, as the a priori chance of a lesion being HCC in these patients is by far the largest [314]. Steatosis was not an exclusion criterium. Diagnosis of liver disease was based on clinical, pathological and/or imaging findings. In case of HCC, cirrhosis was always excluded from biopsy or resection. Lesions with a maximum diameter equal to or smaller than 3 cm were excluded, since in non-cirrhotic livers these have a high probability of being secondary lesions, hemangioma, or cysts [107, 315], which are generally easy to diagnose on imaging [12, 107]. Hence, a radiomics model would have relatively little added value in these patients with underlying liver disease or small lesions. When

T2-weighted MRI with fat saturation was not available, regular T2-weighted MRI was used, similar to clinical practice. Images with significant artefacts (i.e., patient or scanner related) and therefore not suitable for diagnostic purposes, as judged by an experienced radiologist (21 years of experience), were excluded.

12.2.2 Segmentation

Lesion segmentation was done semi-automatically using in-house software [105]. Each lesion was segmented by one of three observers: a radiology resident, and two experienced abdominal radiologists (21 and 8 years of experience). The observers were aware of the inclusion and exclusion criteria, and were asked to segment a primary liver lesion. When the lesions could not be found, e.g. iso-intense lesions, the observers were able to look at the other sequences if available. The observers could segment manually or semi-automatically using region-growing or slice-to-slice contour propagation. Segmentation was performed per slice in the 2D transverse plane, resulting in a 3D volume. Semi-automatic results were always reviewed and manually corrected when necessary, to assure the result resembled manual segmentation. All segmentations were verified by the most experienced radiologist. A subset of 60 lesions (30 from dataset B, 30 from dataset C) was segmented by two observers to assess the intra-observer variability using the pairwise Dice Similarity Coefficient (DSC), with $DSC > 0.70$ indicating good agreement [150].

12.2.3 Radiomics

An overview of the radiomics methodology is depicted in Figure 12.1. As T2-weighted MRI scans do not have a fixed unit and scale, the full images were normalized using z-scoring. No further preprocessing was performed. For each lesion, 564 features quantifying intensity, shape and texture were extracted from the T2-weighted MRI scan. For details, see Section 12.B. To create a decision model from the features, the Workflow for Optimal Radiomics Classification (WORC) toolbox was used [36, 151]. In WORC, decision model creation consists of several steps, e.g. feature selection, resampling, and machine learning. WORC performs an automated search amongst a variety of algorithms for each step and determines which combination maximizes the prediction performance on the training dataset. For details, see Section 12.C. The code for the feature extraction and model creation has been published open-source [316].

12.2.4 Experimental setup

First, to evaluate the predictive value of radiomics within a single center, an internal validation was performed in dataset A through a 100x random-split cross-validation [63, 64], see Figure 12.A.1 A. In each iteration, the data was randomly split into 80% for training and 20% for testing in a stratified manner, to make sure the distribution of classes in all datasets was similar to that in the full dataset.

Second, to evaluate whether a model developed on data from one center generalizes well to unseen data from other centers, two external validations were performed

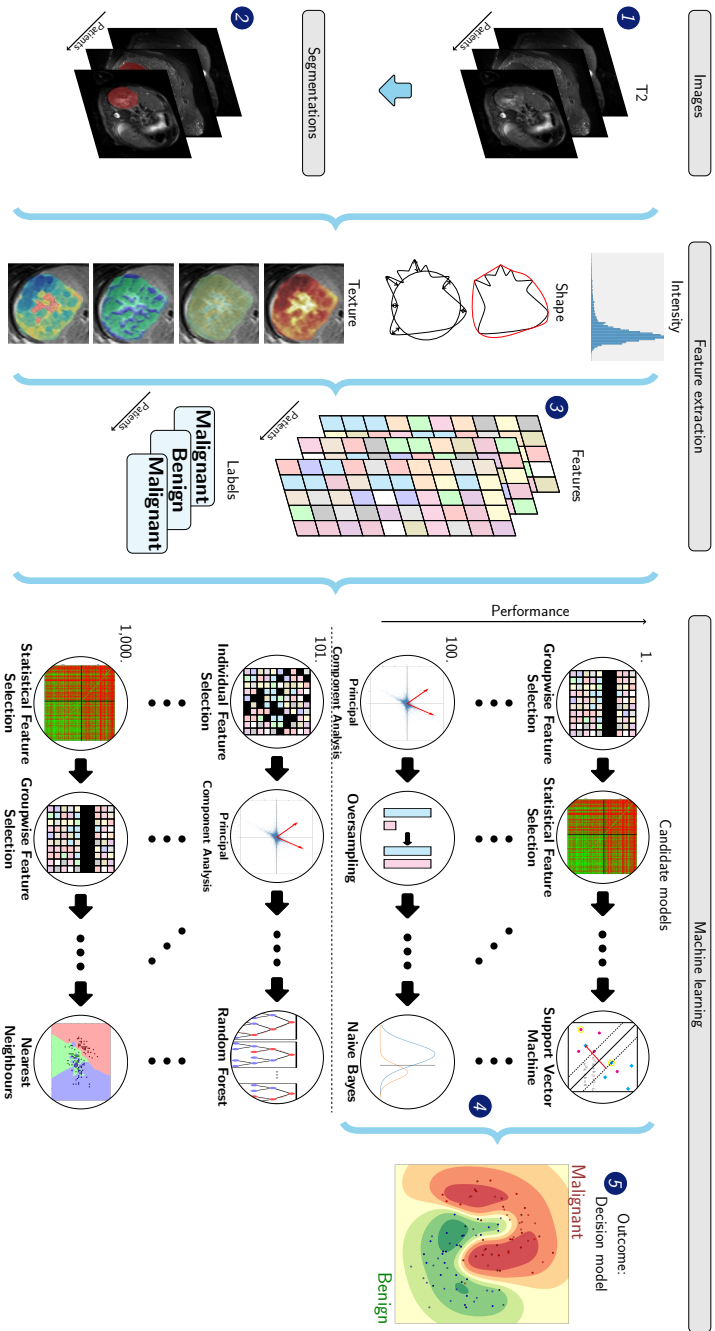


Figure 12.1: Schematic overview of the radionomics approach. Adapted from [72] (i.e., Chapter 5 of this thesis). Input to the algorithm are the T2-weighted MRI scans (1) and the lesion segmentations (2). Processing steps include feature extraction (3) and the creation of a machine learning decision model (5), using an ensemble of the best 100 workflows from 1,000 candidate workflows (4), where the workflows are different combinations of the different analysis steps (e.g., the classifier used).

by training a model on dataset A, and testing it on the unseen datasets B and C, see [Figure 12.A.1 B](#).

Third, as clinicians frequently use age and sex in their decision making, two additional models were externally validated based on: 1) age and sex; and 2) age, sex, and radiomics features.

For both the internal and external validations, model optimization was performed within the training dataset using an internal 5x random-split cross-validation, see [Figure 12.A.1](#). Hence, all optimization was done on the training dataset to eliminate any risk of overfitting on the test dataset.

12.2.5 Performance of the radiologists

To compare the models with clinical practice, the T2-weighted MRI scans were scored by two experienced abdominal radiologists. They were blinded to the diagnosis, but aware of the inclusion and exclusion criteria. Classification of malignancy was made on a four-point scale to indicate the radiologists' certainty:

1. benign, certain
2. benign, uncertain
3. malignant, uncertain
4. malignant, certain

To obtain binary scores, 1 and 2 were converted to benign, 3 and 4 to malignant. Several characteristics used in the decision making were also scored by the radiologists:

1. presence of central scar [[12](#)]
2. presence of liquid
3. presence of atoll sign [[317](#)]
4. degree of heterogeneity (scale 1 - 4 similar to malignancy)

As the radiologists were from centers A and B, scoring was done on dataset C to prevent them from having seen the data previously.

12.2.6 Statistical analysis

To evaluate the difference in clinical characteristics and explore the predictive value of the individual radiomics features between the malignant and benign lesions, per dataset, univariate statistical testing was performed using a Mann-Whitney U test for continuous variables and a Chi-square test for categorical variables. For the clinical characteristics, the statistical significance of the difference between datasets was assessed using a Kruskal-Wallis test for continuous variables, and a Chi-square test for discrete variables. P-values of the clinical characteristics were not corrected for multiple testing as these are purely descriptive: p-values of the radiomics features

were corrected using the Bonferroni correction (i.e., multiplying the p-values by the number of tests).

For all models, the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve, Accuracy, Sensitivity, and Specificity were calculated. ROC confidence bands were constructed using fixed-width bands [67]. The positive class was defined as the malignant lesions.

For the internally validated model, 95% confidence intervals of the performance metrics were constructed using the corrected resampled t-test, thereby taking into account that the samples in the cross-validation splits are not statistically independent [64]. For the externally validated model, 95% confidence intervals were constructed using 1,000x bootstrap resampling of the test dataset and the standard method for normal distributions ([66] table 6, method 1), see Figure 12.A.1 B.

For binary scores, the agreement between radiologists was evaluated using Cohen's κ [318]. For ordinal scores, i.e., degree of heterogeneity and malignancy, the correlation was evaluated using Pearson correlation [319]. The AUCs of the radiomics model and the radiologists were compared using the DeLong test [155], and confusion matrices were used to analyze the agreement.

To gain insight into the radiomics model's decision making, lesions were ranked based on the probability of a lesion being malignant as predicted by the model. Ranking was done as:

- archetypal benign (ground truth benign, probability near 0%)
- pitfall malignant (ground truth malignant, probability near 0%)
- borderline (probability around 50%)
- pitfall benign (ground truth benign, probability near 100%)
- archetypal malignant (ground truth malignant, probability near 100%)

This was done on dataset C to enable comparison with the radiologists.

For all statistical tests, p-values below 0.05 were considered statistically significant.

12.3 Results

12.3.1 Datasets

In total, 486 patients were included (A: 187; B: 98; C: 201). The clinical and imaging characteristics are reported in Table 12.1. As all centers serve as tertiary referral centers, the datasets originated from 159 different scanners (A: 52; B: 21; C: 86), resulting in substantial heterogeneity in the MRI acquisition protocols. Statistically significant differences between datasets A, B, and C included magnetic field strength ($p=0.001$), manufacturer ($p=10^{-4}$), slice thickness ($p=10^{-32}$), repetition time ($p=0.006$), flip angle ($p=0.05$), and use of fat saturation ($p=10^{-17}$).

On the subset that was segmented by two observers, the mean \pm standard deviation of DSC indicated good agreement (B: 0.80 ± 0.21 ; C: 0.81 ± 0.11).

12.3.2 Radiomics

The results of the radiomics model are depicted in [Table 12.2](#). The internal validation on dataset A had a mean AUC of 0.78; the two external validations yielded a similar performance (B: 0.74; C: 0.76). The ROC curves ([Figure 12.2](#)) illustrate that the model trained on dataset A performed similar in each of the three centers.

The age-and-sex-only model had a high AUC in both the internal validation (A: 0.88) and the two external validations (B: 0.93; C: 0.85). Combining age, sex, and the radiomics features yielded an improvement (A: 0.93; B: 0.98; C: 0.91), although not statistically significant. The Accuracy for the age-and-sex-only model (A: 0.83; B: 0.92; C: 0.82) and the combined age, sex, and radiomics model (A: 0.85; B: 0.92; C: 0.83) were similar.

12.3.3 Comparison with radiologists

The performance of the two experienced abdominal radiologists on classifying dataset C is depicted in [Table 12.2](#). The ROC curves ([Figure 12.2 C](#)) were mostly just above the 95% confidence interval of the radiomics model. The AUC of Radiologist 1 (0.87) was statistically significantly better than the radiomics model (DeLong: $p=0.0028$); the differences in AUC between Radiologist 2 (0.83) and the radiomics model and between the two radiologists were not statistically significant. The Accuracy per phenotype is depicted in [Table 12.3](#). The radiomics model had a similar Accuracy in HCC (0.83) and iCCA (0.82), while the performance in FNH (0.66) was slightly better than in HCA (0.54).

Confusion matrices of the predictions on dataset C are depicted in [Figure 12.3](#). The agreement between the radiologists on classifying the lesions as malignant or benign was moderate (Cohen's κ : 0.61) [318]; the two radiologists agreed in 160 of the 201 patients (80%). The agreement between the two radiologists and the radiomics model was weak (Radiologist 1: κ of 0.47; Radiologist 2: κ of 0.42), as reflected by the confusion matrices. For the other characteristics scored by the two radiologists, the agreement was weak for presence of a scar (κ : 0.41) and liquid (κ : 0.52), and strong for presence of the atoll sign (κ : 0.80); the correlation was moderate for heterogeneity (Pearson coefficient: 0.69) and strong for malignancy (Pearson coefficient: 0.70) [319].

12.3.4 Model insight

In dataset A, on which the radiomics model was developed, 45 radiomics features showed statistically significant differences between the malignant and benign lesions with p-values after Bonferroni correction from 9×10^{-10} to 0.049. These included 4 shape features (volume was not significant), 1 orientation feature, and 40 texture features. Statistically significant differences were found for 49 radiomics features in dataset B and 10 in dataset C. Four radiomics features (all texture) showed statistically significant differences in all three datasets. A list of these features and their p-values can be found in [Section 12.A](#). The differences in volume between the three datasets was statistically significant ($p=10^{-10}$).

Table 12.1: Clinical and imaging characteristics of the datasets. The number of patients (N) in each dataset is indicated in the column header. Per dataset, the statistical significance of the difference between the malignant and benign lesions was assessed using a Mann-Whitney U test for continuous variables, and a Chi-square test for discrete variables. The statistical significance of the difference between datasets was assessed using a Kruskal-Wallis test for continuous variables, and a Chi-square test for discrete variables. Statistically significant p-values are displayed in **bold**.

Dataset	A: Erasmus MC (N=187)				B: Maastricht UMC+ (N=98)				C: Beaujon APHP (N=201)				P
	Benign	Malignant	P		Benign	Malignant	P		Benign	Malignant	P		
Patients	93	94			55	43			117	84			
Age in years+	37 [30-46]	62 [25-70]	10 ⁻¹⁹		38 [31-45]	64 [60-71]	10 ⁻¹⁴		38 [31-45]	63 [53-68]	10 ⁻²⁰		0.69
Sex			10 ⁻¹²				10 ⁻⁶				10 ⁻¹⁷		0.22
Male	4	48			3	20			11	55			
Female	89	46			52	23			106	29			
Phenotype													0.003
HCC		81				28				47			
iCCA		13				15							
HCA	48				26				65				
FNH	45				29				52				
Imaging													
Magnetic field strength			0.10				0.45				0.003		0.002
1.0 Tesla	1	4			2	4			1	3			
1.5 Tesla	76	82			48	39			74	68			
3.0 Tesla	16	8			5	0			42	13			
Scanner Manufacturer			10 ⁻⁸				0.03				0.77		10 ⁻¹⁵
Siemens	13	32			21	7			23	17			
Philips	16	38			34	36			62	40			
GE	64	24			0	0			30	24			
Toshiba	0	0			0	0			2	3			
Slice thickness (mm) [†]	6.0 - 8.0	6.0 - 7.0	0.12		5.0 - 6.0	5.0 - 5.0	0.08		5.0 - 6.0	5.0 - 6.0	0.41		10 ⁻³²
Pixel spacing (mm) [†]	0.72 - 0.94	0.73 - 1.19	0.005		0.77 - 1.38	0.77 - 0.99	0.13		0.74 - 1.0	0.75 - 1.07	0.13		0.07
Repetition time (ms) [†]	1348 - 8571	1218 - 4844	0.001		1100 - 2805	1600 - 2961	0.007		1200 - 3884	1512 - 6058	0.14		0.006
Echo time (ms) [†]	89 - 100	80 - 100	10 ⁻³		80 - 112	80 - 90	0.04		80 - 120	80 - 103	0.13		0.62
Flip angle (degree) [†]	90 - 150	90 - 150	0.47		90 - 141	90 - 90	0.01		90 - 140	90 - 134	0.33		0.07
Fat Saturation yes/no	72/21	59 / 35	0.04		35/20	39/4	0.004		98/19	59/25	0.03		10 ⁻¹⁸

* Abbreviations: GE: General Electric; HCC: hepatocellular carcinoma; iCCA: intrahepatic cholangiocarcinoma; HCA: hepatocellular adenoma; FNH: focal nodular hyperplasia; Max: maximum; P: p-value of Mann-Whitney U test for continuous variables, Chi-square for categorical variables.

[†]: median [Quartile 1 - Quartile 3]

[‡]: Quartile 1 - Quartile 3

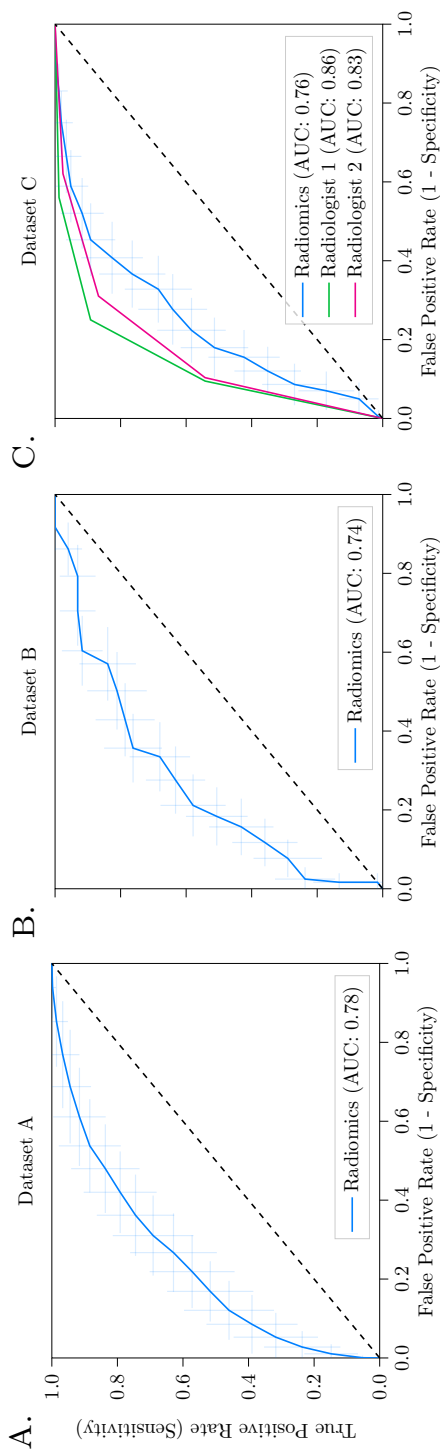


Figure 12.2: Receiver operating characteristic (ROC) curves of the radiomics model and radiologists. For the radiomics model, the curves present the model internally validated on dataset A (A); and trained on dataset A, externally validated on dataset B (B) and dataset C (C). The performance of scoring by the two experienced abdominal radiologists on dataset C is also depicted in (C). For the radiomics model, the crosses identify the 95% confidence intervals of the 100x random-split cross-validation (A) or 1,000x bootstrap resampling (B and C); the bold curves are fit through the means.

Table 12.2: Performance of the radiomics model and the radiologists three datasets (A, B, and C). For the radiomics model, the mean (internal cross-validation) or point estimate (external validation) and 95% confidence intervals are reported.

Evaluation	Internal cross-validation	External validation		Radiologist 1	Radiologist 2
Train set	A [†]	A	A	-	-
Test set	A [†]	B	C	C	C
AUC	0.78 [0.70, 0.85]	0.74 [0.65, 0.84]	0.76 [0.70, 0.83]	0.86	0.83
Accuracy	0.69 [0.62, 0.76]	0.64 [0.54, 0.74]	0.69 [0.62, 0.75]	0.80	0.77
Sensitivity	0.70 [0.57, 0.82]	0.79 [0.67, 0.91]	0.82 [0.74, 0.91]	0.88	0.87
Specificity	0.68 [0.59, 0.78]	0.53 [0.40, 0.66]	0.59 [0.50, 0.68]	0.74	0.69

*Abbreviations: AUC: area under the receiver operating characteristic curve.
†Training and testing within a single dataset was done through a 100x random-split cross-validation.

Table 12.3: Accuracy per phenotype of the radiologists and the radiomics model in the external validation on dataset C. The Accuracy per phenotype represents the percentage of the lesions with that specific phenotype being correctly classified as malignant or benign. The number of lesions per phenotype in dataset C is given between brackets in the first column.

Accuracy	Radiomics	Radiologist 1	Radiologist 2
Train dataset	A	-	-
Test dataset	C	C	C
HCC (47)	0.83	0.85	0.83
iCCA (37)	0.82	0.95	0.92
HCA (65)	0.54	0.69	0.62
FNH (52)	0.66	0.82	0.78

*Abbreviations: HCC: hepatocellular carcinoma; HCA: hep-
atocellular adenoma; FNH: focal nodular hyperplasia; iCCA:
intrahepatic cholangiocarcinoma

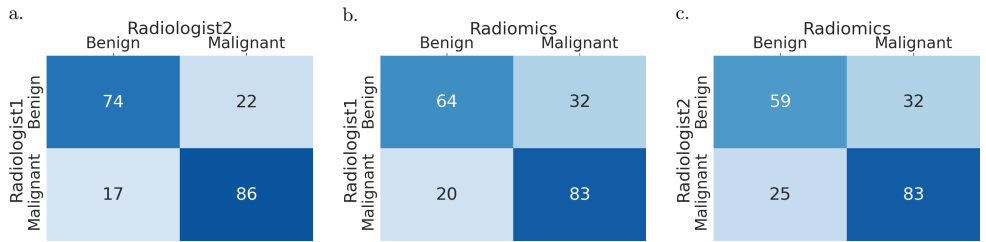


Figure 12.3: Confusion matrices of the predictions by the radiomics model and the two radiologists. The darker the background, the higher the agreement.

Examples of lesions from dataset C ranked as archetypal, borderline, or pitfall by the radiomics model are depicted in [Figure 12.4](#). Visual inspection of the T2-weighted MRI scans of the archetypal or pitfall lesions showed a relation with heterogeneity (archetypal malignant: heterogeneous; archetypal benign: homogeneous), area and volume (archetypal malignant: generally high maximum axial area and high volume), and irregularity of shape on 2-D axial slices (archetypal malignant lesions: irregular; archetypal benign: compact). Pitfall lesions showed the opposite, e.g. pitfall benign: heterogeneous. Borderline lesions, i.e., with an almost equal predicted chance of being malignant or benign, were mostly of medium size and medium heterogeneity.

The predictions by the radiomics model on dataset C were compared to the characteristic scores of Radiologist 1, who had the highest performance. The correlation between the probability of malignancy as predicted by the radiomics model and heterogeneity as scored by Radiologist 1 was moderate (Pearson coefficient: 0.58). Radiologist 1 performed well when lesions had an apparent atoll sign: from the 19 lesions which Radiologist 1 scored as having an atoll sign and therefore classified as benign, 17 were indeed benign and 2 malignant. On the contrary, the radiomics model only classified 11 of these lesions correctly, but these included the 2 malignant lesions misclassified by Radiologist 1.

12.4 Discussion

In this study, we developed a radiomics model to distinguish between malignant and benign primary solid liver lesions based on T2-weighted MRI in patients with non-cirrhotic livers. We showed that our radiomics model can distinguish between these lesions, both in an internal cross-validation and in two external validations.

The substantial increase of radiomics related research in recent years has led to various guidelines, vulnerabilities, and gaps [24, 25, 30, 31]. While several studies have evaluated radiomics for the classification of liver lesions [308, 309, 310], radiomics for primary liver cancer is still in the early stages, and many of these aspects still need to be addressed [313]. One of the most important is external validation, which is crucial to ensure a high level of evidence in a variety of settings [24, 30]. Furthermore, the lack of standard imaging parameters can be problematic as these can affect the appearance of the lesion and thus radiomics [18, 313]. Requiring a comprehensive, standardized set of multiple MRI sequences is hardly feasible in practice. In this study, we therefore only used T2-weighted MRI without strict protocol requirements, and externally validated our model on two multi-center cohorts from different countries to assess the generalizability. The scans of the 486 patients included in this study originated from 159 different MRI scanners, resulting in substantial heterogeneity in the acquisition protocols. In univariate analyses, only four radiomics features showed statistically significant differences in all three datasets. Nevertheless, our method performed well on data from unseen scanners (i.e., not present in the training dataset), indicating good generalizability. Furthermore, we used routinely acquired T2-weighted MRI, increasing the chance that the reported performance can be reproduced in a routine clinical setting. All lesions in our study, except typical FNH [106], were pathologically proven to ensure the ground truth was objective. We also set inclusion criteria to maximize the

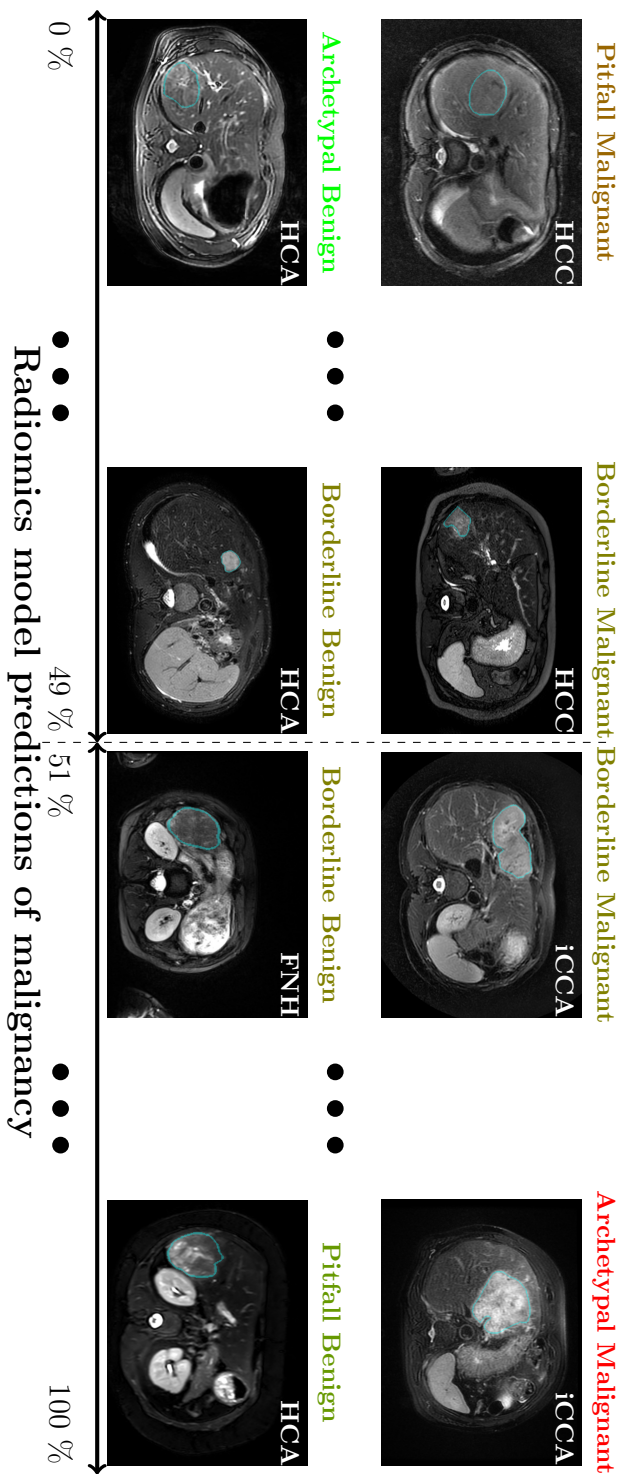


Figure 12.4: Examples of liver lesions on T2-weighted MRI. From left to right, examples of lesions considered by the radiomics model as archetypal (i.e., predicted probability close to extremes and correct), pitfall (i.e., predicted probability close to extremes and incorrect), and borderline (i.e., predicted probability close to border of 50%). * Abbreviations: HCC: hepatocellular carcinoma; iCCA: intrahepatic cholangiocarcinoma; HCA: hepatocellular adenoma; FNH: focal nodular hyperplasia.

relevance to clinical decision making. Usage of a single, widely used sequence and the fact that the lesion phenotypes included in our study present more than 90% of all solid lesions, makes our model widely applicable.

To compare the radiomics model to routine clinical practice, the model's predictions were compared to assessment by two experienced abdominal radiologists. The agreement between radiologists was moderate, indicating some observer variation in the predictions. The characteristics apparently used by the radiomics model to define lesions as archetypal, borderline, and pitfalls, were different than those used in the scoring of the radiologists. This is also illustrated by the moderate correlation in the heterogeneity scored by Radiologist 1 and the radiomics model's score, and their different predictions on lesions with an apparent atoll sign. As these results indicate the potential complementary value of the radiomics model, further research should focus on how the radiologists' and the radiomics model's predictions can be optimally combined to improve clinical decision making.

Our results indicate that assessment of primary solid liver lesions by radiologists can be challenging and is subject to observer dependence. Existing guidelines may aid the radiologist in specific scenarios, such as EASL's guidelines for management of benign liver tumors [107] and HCC [304], or LI-RADS for patients with cirrhotic livers [310]. In this study, inclusion and exclusion criteria were determined to maximize the clinical relevance, covering scenarios not included in these guidelines. Our radiomics model therefore complements these existing initiatives. Radiomics may be especially useful on lesions where there is no consensus between radiologists, or on the pitfalls for radiologists. Additionally, it may serve as a gatekeeper in non-specialized centers, shortening the diagnostic delay by enabling direct referral to an expertise center and reducing the number of missed malignant lesions.

Age and sex are known to be strong predictors for distinguishing malignant from benign liver lesions [12, 301]. In our study, in line with worldwide findings, (young) females represented the majority of benign lesions, while older patients represented the majority of malignant lesions [12, 301]. The models based on age and sex used an age threshold at 49 years. In dataset C, only 19 (17%) of the 114 lesions of patients below 49 years were malignant. Although this therefore yielded a good overall performance, it would lead to missing all malignant lesions in young patients, for whom such a diagnosis is essential as these patients would benefit most from treatment. Simply classifying all lesions below 49 years as benign, regardless of any imaging information, would be unacceptable and cannot be applied to the general population. On the other hand, the radiomics model purely based on T2-weighted MRI does not use any population-based information. The model rather predicts the probability of a lesion being malignant based on the imaging appearance. Our radiomics method could be especially useful in young males to not miss malignant lesions, and in older females to detect benign lesions. Future research should therefore also focus on optimally combining imaging, age, and sex.

Our study has several limitations. First, while the inclusion and exclusion criteria were set to maximize the relevance to clinical decision making, they limit the applicability, as our model cannot be applied to all liver lesions, and may have led to selection biases. Future research should therefore focus on loosening these criteria, for example including patients with smaller lesions (maximum diameter < 3 cm),

liver disease, more typical lesions, i.e., that are routinely not biopsied, and other (rare) phenotypes. Second, the current radiomics approach requires semi-automatic segmentations. While accurate, this process is time consuming and subject to some observer variability, limiting the transition to clinical practice. We do not believe that this has substantially affected the results, as the inter-observer DSC indicated good segmentation reproducibility, and the radiomics model performed similar in the internal and external validations despite training and testing on segmentations of various observers. Automatic segmentation methods, for example with deep learning [164], may help to further automate the method and avoid observer dependence.

On one hand, using a single, widely available (T2-weighted) MRI sequence without strict protocol restrictions is a strength of our model. On the other hand, in real life, radiologists use multiple sequences in their assessment, indicating that a multi-sequence model may lead to an improved performance. EASL's guidelines also describe lesion assessment characteristics based on these other sequences, e.g. wash-out on dynamic contrast enhanced T1-weighted MRI, and diffusion restrictions (low ADCs) [107, 304]. These other sequences may contain additional information to improve the radiomics and radiologists' performance [308]. Especially when extending our work to phenotyping, these sequences may contain essential information for an accurate diagnosis. Main additional challenges for such a multi-sequence model, due to the lack of a standardized protocol in the literature, are the additional heterogeneity, missing data as not all these sequences are acquired by default, and overcoming differences in appearance caused by the variations in contrast agents [320]. We used only T2-weighted MRI, as this sequence suffers less from these disadvantages; is widely available, thus a T2-weighted MRI based radiomics model is feasible to use in routine clinical practice; is relatively simple and thus showing less heterogeneity as e.g. sequences with contrast; is reliable for lesion segmentation; and is minimally sensitive to motion or breathing artefacts; and is informative [107, 304, 311]. The latter is also illustrated by our results, as the two radiologists were already able to distinguish malignant from benign lesions quite accurately using only T2-weighted MRI.

Future research should, besides the points mentioned in the previous paragraphs, focus on extending our work to phenotyping (e.g. HCC, iCCA, HCA, FNH), and possibly even subtyping (e.g. inflammatory HCA, β -catenin activated HCA) to further aid clinical decision making. Furthermore, to gain better insight into the complementary value of radiomics, our model may be compared with more radiologists. In our study, two experienced abdominal radiologists who were trained at the same center scored the patients. Hence, it would be valuable to compare with radiologists from a variety of institutes, also including less experienced and non-academic radiologists. This will also give a better insight into which type of lesions are difficult for radiologists to classify or reach consensus on, and thus where radiomics could have the highest added value.

In conclusion, our radiomics model based on T2-weighted MRI was able to distinguish malignant from benign primary solid liver lesions in patients with non-cirrhotic livers, both in an internal validation and in two external validations on heterogeneous, multi-center data. Pending further optimization and generalization, our model may serve as a robust, non-invasive and low-cost aid to enable quicker

referral and refine patient selection prior to biopsies, and help solve the shortage of radiologists [14].

Data availability statement: Imaging and clinical research data are not available at this time. Programming code is available on Zenodo at DOI <https://doi.org/10.5281/zenodo.5175705>.

Conflict of interest statement: Wiro Niessen is founder, scientific lead, and shareholder of Quantib BV. The other authors do not declare any conflicts of interest.

Financial support statement: No funding sources were involved in the study design, collection, analysis, and interpretation of data, writing the report, nor the decision to submit the article for publication.

Author contributions: M.P.A.S., R.L.M., V.V., M.R., W.J.N., S.K. and M.G.T. provided the conception and design of the study. M.P.A.S., R.L.M., Y.P., J.V., J.I., R.A.d.M., M.D. and M.G.T. acquired the data. M.P.A.S., S.K. and M.G.T. analyzed and interpreted the data. M.P.A.S. created the software. M.P.A.S., S.K. and M.G.T. drafted the article. All authors read and approved the final manuscript.

Acknowledgments: Martijn Starmans acknowledges funding from the research program STRaTeGy (project number 14929-14930), which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO). This work was partially carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

Appendix

Appendix 12.A Pathological examination

In the pathology, a distorted (micro)architecture of liver tissue was the common feature of the included lesions. Histomorphology often combined with (immuno) histochemistry served the final diagnosis. Hepatocellular lesions with loss of portal tracts, cell atypia, thick trabeculae (loss of reticulin fibers), pseudoglandular transformation, isolated small arterial branches, and capillarization of the sinusoidal areas (CD34 positive) with supportive immunohistochemistry (glypican-3, glutamine synthetase, HSP-70), were classified as HCC [321]. Cases where the reticulin fibers were maintained, the pseudoglandular transformation and the cell atypia were absent or minimal, and the immunohistochemistry (glypican-3, HSP-70) was negative, were classified as HCA [321]. Lesions composed of non-organoid arranged glandular structures, localized at the periphery of the second-order bile ducts with an expression of keratin 7 and 19, were classified as iCCA, either conventional or cholangiolocarcinoma [322]. Non-neoplastic lesions, composed of hyperplastic hepatocellular nodules separated by fibrotic septa, creating a microscopic image of “localized cirrhosis” and often centrally a scar, were classified as FNH. Glutamine

synthetase showed the pathognomonic “map-like” pattern of immunohistochemical expression (anastomosing groups of positively stained hepatocytes [323]).

Appendix 12.B Radiomics feature extraction

This appendix is similar to Vos *et al.*, Timbergen *et al.* [72, 73] (i.e., Chapter 5 and Chapter 6 of this thesis), but details relevant for the current study are highlighted.

A total of 564 radiomics features were used in this study. All features were extracted using the defaults for MRI scans from the Workflow for Optimal Radiomics Classification (WORC) [36], which internally uses the PREDICT [51] and PyRadiomics [44] feature extraction toolboxes. An overview of all features is depicted in Table 12.A.2. For details on the mathematical formulation of the features, we refer the reader to Zwanenburg *et al.* [39]. More details on the extracted features can be found in the documentation of the respective toolboxes, mainly the WORC documentation [68].

For MRI scans, the images are by default normalized in WORC as the scans do not have a fixed unit and scale, contrary to e.g. computed tomography (Hounsfield units). Normalization is performed using z-scoring, i.e., subtracting the mean and dividing by the standard deviation. As the datasets used in this study exhibit substantial heterogeneity in the acquisition protocols, the mean and standard deviation were computed based on the segmentation of the regions of interest (ROIs), i.e., the lesions, and not on the full image, as the latter is more sensitive to acquisition variations. The images were not resampled, as this would result in interpolation errors, especially in the axial direction due to the substantial differences in slice thicknesses. The code to extract the features has been published open-source [316].

The features can be divided in several groups. Thirteen intensity features were extracted using the histogram of all intensity values within the ROIs and included several first-order statistics such as the mean, standard deviation and kurtosis. These describe the distribution of intensities within the lesion. Thirty-five shape features were extracted based only on the ROI, i.e. not using the image, and included shape descriptions such as the volume, compactness and circular variance. These describe the morphological properties of the lesion. Nine orientation features were used, describing the orientation of the ROI, i.e. not using the image. Lastly, 507 texture features were extracted using Gabor filters (156 features) [39], Laplacian of Gaussian filters (39 features) [39], vessel (i.e. tubular structures) filters (39 features) [54], the Gray Level Co-occurrence Matrix (144 features) [39], the Gray Level Size Zone Matrix (16 features) [39], the Gray Level Run Length Matrix (16 features) [39], the Gray Level Dependence Matrix (14 features) [39], the Neighbourhood Grey Tone Difference Matrix (5 features) [39], Local Binary Patterns (39 features) [52], and Local Phase filters (39 features) [53, 300]. These features describe more complex patterns within the lesion, such as heterogeneity, presence of blob-like structures, and presence of line patterns.

Most of the texture features include parameters to be set for the extraction. The values of the parameters that will result in features with the highest discriminative power for the classification at hand (i.e., malignant versus benign) are not known beforehand. Including these parameters in the workflow optimization, see Section 12.C,

would lead to repeated computation of the features, resulting in a redundant increase in computation time. Therefore, alternatively, these features are extracted at a range of parameters as is default in WORC. The hypothesis is that the features with high discriminative power will be selected by the feature selection methods and/or the machine learning methods as described in [Section 12.C](#). The parameters used are described in [Table 12.A.2](#).

The variations in the slice thickness due to the heterogeneity in the acquisition protocols may cause feature values to be dependent on the acquisition protocol. Moreover, the slice thickness is substantially larger than the pixel spacing. Hence, extracting robust 3D features may be hampered by these variations, especially for low resolutions. To overcome this issue, all features were extracted per 2D axial slice and aggregated over all slices, which is default in WORC. Afterwards, several first- order statistics over the feature distributions were evaluated and used in the machine learning approach.

Appendix 12.C Radiomics decision model creation

This appendix is similar to Vos *et al.*, Timbergen *et al.* [72, 73] (i.e., [Chapter 5](#) and [Chapter 6](#) of this thesis), but details relevant for the current study are highlighted.

The Workflow for Optimal Radiomics Classification (WORC) toolbox [36] makes use of automated machine learning to create the optimal performing workflow from a variety of algorithms. Besides deciding whether to use an algorithm, most algorithms require hyperparameters, i.e., parameters that need to be set before the actual learning step, to be tuned to enhance the performance. WORC defines a workflow as a specific sequential combination of algorithms and their respective hyperparameters. In WORC, the radiomics workflow is split into the following components: image and segmentation preprocessing, feature extraction, feature and sample preprocessing, and machine learning. For each component, a collection of algorithms and their associated hyperparameters is included. Given this search space, WORC uses automated machine learning to find the optimal solution. The code to use WORC for creating the decision models in this specific study has been published open-source [316].

The workflows could be constructed from the following default search space in WORC, which components can only be combined in the order listed below:

1. Features selection: a group-wise search, in which specific groups of features (i.e., intensity, shape, and the subgroups of texture features as defined in [Section 12.B](#) and [Table 12.A.2](#)) are selected or deleted. To this end, each feature group had an on/off variable which is randomly activated or deactivated, which were all included as hyperparameters in the optimization.
2. Feature imputation: when a feature could not be computed, e.g. a lesion is too small for a specific feature to be extracted, a feature imputation algorithm is used to estimate replacement values for the missing values. Strategies for imputation included 1) the mean; 2) the median; 3) the mode; 4) a constant (default: zero); and 5) a nearest neighbor approach.

3. Feature selection: a variance threshold, in which features with a low variance (<0.01) are removed. This method was always used, as this serves as a feature sanity check with almost zero risk of removing relevant features.
4. Feature scaling was performed to make all features have the same scale, as otherwise the machine learning methods may focus only on those features with large values. This was done through z-scoring, i.e., subtracting the mean value followed by division by the standard deviation, for each individual feature. A robust version of z-scoring was used, in which outliers, i.e., values below the 5th percentile or above the 95th percentile, were excluded from computing the mean and variance.
5. Feature selection: optionally, the RELIEF method [55], which ranks the features according to the differences between neighboring samples. Features with more differences between neighbors of different classes (i.e., malignant versus benign) are considered higher in rank.
6. Feature selection: optionally, features are selected by training a machine learning model and selecting features that are regarded important by the model. Hence the used model should be able to give the features an importance weight. Included model choices are LASSO, logistic regression, and a random forest.
7. Dimensionality reduction: optionally, principal component analysis (PCA) is used, in which either only those linear combinations of features were kept which explained 95% of the variance in the features or a limited number of components (between 10 – 50).
8. Feature selection: optionally, individual feature are selected through univariate testing. To this end, for each feature, a Mann-Whitney U test was performed to test for significant differences in distribution between the labels (i.e., malignant versus benign). Afterwards, only features with a p-value above a certain threshold were selected.
9. Resampling: optionally, a various resampling strategy could be used, which are used to overcome class imbalances and reduce overfitting on specific training samples. These included various methods from the imbalanced-learn toolbox [57]: random over-sampling, random under-sampling, near-miss resampling, the neighborhood cleaning rule, ADASYN, and SMOTE (regular, borderline, Tomek and the edited nearest neighbors variant).
10. Machine learning: lastly, a machine learning method is used to determine a decision rule to distinguish the classes. Methods included were; 1) logistic regression; 2) support vector machines; 3) random forests; 4) naive Bayes; 5) linear discriminant analysis; 6) quadratic discriminant analysis; 7) AdaBoost [61]; and 8) extreme gradient boosting [62].

By default in WORC, all model construction and optimization was performed on the training set in order to prevent overfitting on the test dataset. To prevent overfitting on the training dataset, a 5x random-split stratified cross-validation [63,

64] was performed within the training dataset as well, using 85% for model training and 15% for model validation, see Figure 12.A.1.

WORC states the radiomics workflow as a combined algorithm selection and hyperparameter optimization problem (CASH), as algorithm selection and hyperparameter optimization are often not independent [34]. Within the training dataset, CASH optimization is performed by testing thousand pseudo-randomly generated radiomics workflows from the above search space. These are trained on the five training datasets in the 5x random-split training-validation cross-validation, and ranked according to their mean performance on the five validation datasets. As performance metric, the weighted F1-score is used, which is the harmonic average of the precision and recall.

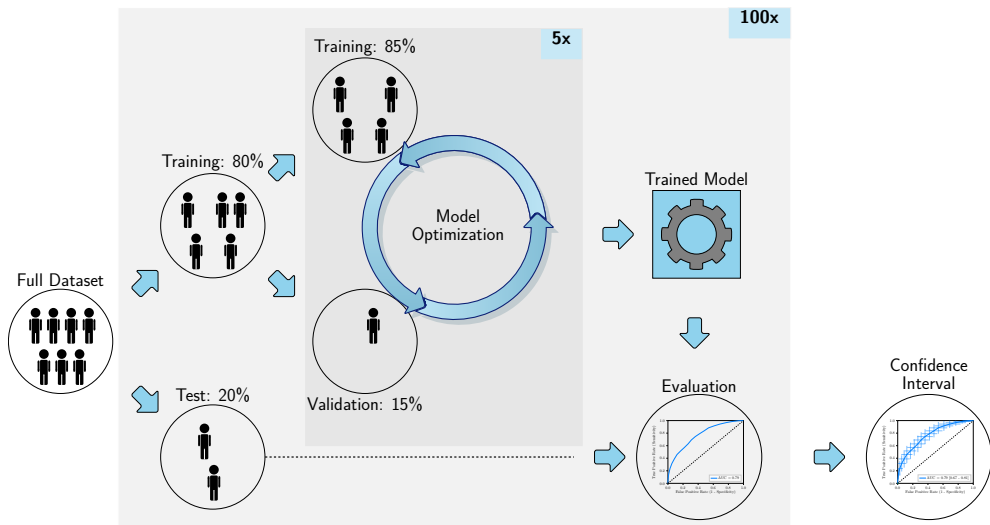
Using only the single workflow that on average performs best on the validation datasets may result in poor generalization due to overfitting on the validation datasets. Hence, an ensemble was constructed by combining the workflows that perform best on the validation datasets. Ensembling was done using the default of WORC by averaging the posteriors of the 100 best workflows.

The following pseudo code illustrates the algorithm of WORC:

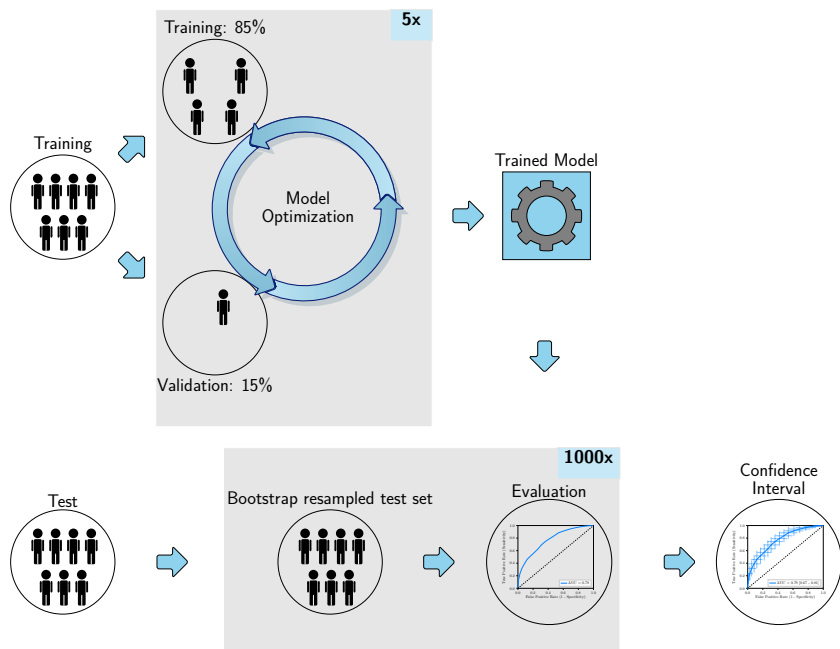
- **For** each 100x random-split training-test cross-validation iteration:
 - **Do:** Construct the training dataset by randomly selecting 80% of the patients.
 - **Do:** On this training dataset, define 5x random-split cross-validation splits, selecting in each iteration 85% of the patients for training and 15% for validation.
 - **Do:** Pseudo-randomly sample 1,000 workflows from the search space.
 - **For** each of the 1,000 sampled workflows:
 - * **Do:** Train the workflow on the five training datasets in the 5x random-split cross-validation.
 - * **Do:** Compute the mean weighted F1-score on the corresponding five validation datasets in the 5x random-split cross-validation.
 - **Do:** Rank the 1,000 workflows, retrain the best 100 workflows on the full training dataset and combine them in an ensemble.
 - **Do:** Evaluate “the model”, i.e., the ensemble of the best 100 workflows as trained on the training dataset, on the test dataset, i.e., the remaining 20% of the patients that were not included in the training dataset.

The largest experiments in this study consists of executing 500,000 workflows (1,000 pseudo-randomly generated workflows, times a 5x train-validation cross-validation, times 100x train-test cross-validation for the internal validation), which can be parallelized. The computation time of training or testing a single workflow is on average less than a second, depending on the size of the dataset both in terms of samples (i.e. patients) and features. The largest experiment in this study, i.e. the internal validation on dataset A, had a computation time of approximately 24 hours on a 32 CPU core machine. The contribution of the feature extraction to the

computation time was negligible. The code for the radiomics feature extraction and model creation, including more details, has been published open-source [316].



A. Internal validation



B. External validation

Figure 12.A.1: Visualization of evaluation setups. (A) The 100x random-split cross-validation used in the internal validation; (B) and the 1,000x bootstrap resampling in the external validations. Both include an internal random-split cross-validation within the training dataset for the model optimization.

Table 12.A.1: Overview of univariate testing of radiomics features. Per dataset (A, B, and C), the statistical significance of the difference between the malignant and benign lesions was assessed using a Mann-Whitney U test for continuous variables, and a Chi-square test for discrete variables. Only the features that showed statistically significant differences in dataset A are include. All p-values were corrected for multiple testing by multiplying the p-values with the total number of tests (564). Statistically significant p-values and names of that showed statistically significant differences in all three datasets are given in **bold**.

Feature name	p-value A	p-value B	p-value C
tf_kurtosis_sigma1	9.26×10^{-10}	1.80×10^{-5}	1.00
tf_mean_sigma1	1.06×10^{-8}	1.27×10^{-4}	1.00
tf_LBP_std_R3_P12	3.19×10^{-8}	8.60×10^{-4}	1.00
tf_LBP_quartile_range_R8_P24	1.56×10^{-7}	0.0026	7.77×10^{-5}
tf_peak_sigma1	2.74×10^{-7}	0.0028	1.00
tf_median_sigma1	8.27×10^{-7}	0.0035	1.00
tf_LBP_skewness_R8_P24	1.33×10^{-6}	0.0067	2.16×10^{-4}
tf_LBP_kurtosis_R8_P24	1.51×10^{-6}	0.013	9.44×10^{-5}
tf_LBP_mean_R8_P24	1.53×10^{-6}	0.014	2.20×10^{-4}
tf_LBP_skewness_R15_P36	5.18×10^{-5}	0.13	8.70×10^{-4}
tf_LBP_mean_R15_P36	6.18×10^{-5}	0.14	9.92×10^{-4}
tf_mean_sigma10	8.40×10^{-5}	0.94	1.00
tf_LBP_kurtosis_R15_P36	9.02×10^{-5}	0.19	4.75×10^{-4}
tf_LBP_median_R3_P12	1.29×10^{-4}	0.086	0.27
sf_area_min_2D	2.29×10^{-4}	0.67	1.00
tf_Gabor_std_F0.2_A0.79	3.65×10^{-4}	6.42×10^{-5}	0.50
tf_Gabor_kurtosis_F0.05_A0.79	6.24×10^{-4}	1.00	1.00
tf_LBP_skewness_R3_P12	8.44×10^{-4}	0.28	0.17
tf_Gabor_quartile_range_F0.2_A0.79	0.001	9.71×10^{-5}	0.11
tf_median_sigma10	0.001	0.51	1.00
tf_Gabor_quartile_range_F0.2_A1.57	0.001	3.79×10^{-5}	1.00
tf_Gabor_max_F0.2_A0.79	0.004	2.26×10^{-4}	0.24
tf_Gabor_std_F0.2_A0.0	0.006	0.001	0.18
tf_Gabor_std_F0.2_A1.57	0.008	0.002	1.00
tf_Gabor_range_F0.2_A0.79	0.008	6.92×10^{-4}	0.69
tf_Gabor_quartile_range_F0.2_A2.36	0.009	6.92×10^{-4}	0.12
tf_LBP_mean_R3_P12	0.012	1.00	0.45
tf_kurtosis_sigma10	0.012	0.73	1.00
tf_std_sigma1	0.014	0.44	1.00
tf_LBP_std_R15_P36	0.015	1.00	0.002
tf_Gabor_max_F0.2_A1.57	0.015	0.001	1.00
tf_Gabor_median_F0.5_A0.0	0.015	1.00	1.00
sf_area_avg_2D	0.015	0.010	1.00
tf_Gabor_min_F0.2_A0.79	0.017	0.005	1.00
tf_LBP_std_R8_P24	0.019	1.00	8.80×10^{-4}
tf_LBP_quartile_range_R15_P36	0.020	1.00	0.084
tf_Gabor_quartile_range_F0.2_A0.0	0.020	5.80×10^{-4}	0.090
sf_area_max_2D	0.023	0.011	1.00
sf_shape_Flatness	0.038	1.00	1.00
of_COM_y	0.038	0.69	1.00
tf_Frangi_inner_energy_SR(1.0. 10.0)_SS2.0	0.042	0.033	1.00
tf_GLDM_SmallDependenceHighGrayLevelEmphasis	0.046	0.020	1.00
tf_max_sigma10	0.046	1.00	1.00
tf_Frangi_edge_energy_SR(1.0. 10.0)_SS2.0	0.049	0.081	1.00
tf_Frangi_full_energy_SR(1.0. 10.0)_SS2.0	0.049	0.081	1.00

*Abbreviations: tf: texture feature; sf: shape features; of: orientation feature.

Table 12.A.2: Overview of the 564 features used in this study. GLCM features were calculated in four different directions (0, 45, 90, 135 degrees) using 16 gray levels and pixel distances of 1 and 3. LBP features were calculated using the following three parameter combinations: 1 pixel radius and 8 neighbours, 2 pixel radius and 12 neighbours, and 3 pixel radius and 16 neighbours. Gabor features were calculated using three different frequencies (0.05, 0.2, 0.5) and four different angles (0, 45, 90, 135 degrees). LoG features were calculated using three different widths of the Gaussian (1, 5 and 10 pixels). Vessel features were calculated using the full mask, the edge, and the inner region. Local phase features were calculated on the monogenic phase, phase congruency and phase symmetry.

Histogram (13 features)	LoG (13*3=39 features)	Vessel (12*3=39 features)	GLCM (MS) (6*3*4*2=144 features)	Gabor (13*3*4=156 features)	NGTDM (5 features)	LBP (13*3=39 features)
min	min	min	contrast (normal, MS mean + std)	min	busyness	min
max	max	max	dissimilarity (normal, MS mean + std)	max	coarseness	max
mean	mean	mean	homogeneity (normal, MS mean + std)	mean	complexity	mean
median	median	median	angular second moment (ASM) (normal, MS mean + std)	median	contrast	median
std	std	std	energy (normal, MS mean + std)	std	strength	std
skewness	skewness	skewness	correlation (normal, MS mean + std)	skewness		skewness
kurtosis	kurtosis	kurtosis		kurtosis		kurtosis
peak	peak	peak		peak		peak
peak position	peak position	peak position		peak position		peak position
range	range	range		range		range
energy	energy	energy		energy		energy
quartile range	quartile range	quartile range		quartile range		quartile range
entropy	entropy	entropy		entropy		entropy
GLSZM (16 features)	GLRM (16 features)	GLDM (14 features)	Shape (35 features)	Orientation (9 features)	Local phase (13*3=39 features)	
Gray Level Non Uniformity	Gray Level Non Uniformity	Dependence Entropy	compactness (mean + std)	theta_x	min	
Gray Level Non Uniformity Normalized	Gray Level Non Uniformity Normalized	Dependence Non-Uniformity	radial distance (mean + std)	theta_y	max	
Gray Level Variance	Gray Level Variance	Dependence Non-Uniformity Normalized	roughness (mean + std)	theta_z	mean	
High Gray Level Zone Emphasis	High Gray Level Run Emphasis	Dependence Variance	convexity (mean + std)	COM index x	median	
Large Area Emphasis	Long Run Emphasis	Gray Level Non-Uniformity	circular variance (mean + std)	COM index y	std	
Large Area High Gray Level Emphasis	Long Run High Gray Level Emphasis	Gray Level Variance	principal axes ratio (mean + std)	COM index z	skewness	
Large Area Low Gray Level Emphasis	Long Run Low Gray Level Emphasis	High Gray Level Emphasis	elliptic variance (mean + std)	COM x	kurtosis	
Low Gray Level Zone Emphasis	Low Gray Level Run Emphasis	Large Dependence Emphasis	solidity (mean + std)	COM y	peak	
SizeZoneNonUniformity	RunEntropy	Large Dependence High Gray Level Emphasis	area (mean, std, min + max)	COM z	peak position	
SizeZoneNonUniformityNormalized	RunLengthNonUniformity	Large Dependence Low Gray Level Emphasis	volume (total, mesh, volume)		range	
SmallAreaEmphasis	RunLengthNonUniformityNormalized	Low Gray Level Emphasis	elongation		energy	
SmallAreaHighGrayLevelEmphasis	RunPercentage	Small Dependence Emphasis	flatness		quartile	
SmallAreaLowGrayLevelEmphasis	RunVariance	Small Dependence High Gray Level Emphasis	least axis length		entropy	
ZoneEntropy	ShortRunEmphasis	Small Dependence Low Gray Level Emphasis	major axis length			
ZonePercentage	ShortRunHighGrayLevelEmphasis		minor axis length			
ZoneVariance	ShortRunLowGrayLevelEmphasis		maximum diameter 3D			
			maximum diameter 2D (rows, columns, slices)			
			sphericity			
			surface area			
			surface volume ratio			

* Abbreviations: COM: center of mass; GLCM: gray level co-occurrence matrix; MS: multi slice; NGTDM: neighborhood gray tone difference matrix; GLSZM: gray level size zone matrix; GLRLM: gray level run length matrix; LBP: local binary patterns; LoG: Laplacian of Gaussian; std: standard deviation.

General discussion and summary



13.

Discussion

General discussion

To assist clinicians in adapting healthcare to each patient's unique characteristics, the paradigm shift to personalized medicine has led to an increased need for biomarkers that reflect the health or disease status of a person. Medical imaging holds much potential to provide such biomarkers, but methods are required to extract quantitative and objective biomarkers. Radiomics, i.e., the use of quantitative imaging features and machine learning, has shown many successes in various clinical applications to identify and extract such biomarkers. However, this field faces several challenges: 1) it is challenging to find the optimal radiomics method from the wide variety of available options; 2) there is a need for publicly sharing data to facilitate reproducibility, to develop accurate biomarkers, and to validate the performance and generalization of biomarkers; 3) there is a lack of image acquisition standardization; and 4) there is a lack of reproducibility of both radiomics methods and biomarkers. Overcoming these barriers is vital for the translation of radiomics models to clinical practice.

In this thesis, these challenges are addressed in order to streamline radiomics research, facilitate the reproducibility of radiomics methods and biomarkers, and ultimately simplifying the use of radiomics in (new) clinical applications. [Figure 13.1](#) provides a schematic overview of the biomarkers and the adaptive radiomics framework developed in this thesis. In [Chapter 2](#), we introduced radiomics, described its potential and several of its challenges. In [Chapter 3](#), to overcome these challenges, we exploited advances in automated machine learning to automatically construct and optimize the radiomics workflow per application. We validated the resulting radiomics method in twelve different, independent clinical applications to evaluate its generalization across clinical applications. We evaluated in depth our adaptive framework in eight clinical applications to develop quantitative imaging biomarkers in [Chapters 5, 6, 7, 8, 9, 10, 11, and 12](#). Lastly, we publicly released six datasets consisting of 930 patients in total as described in [Chapter 4](#), the WORC toolbox (open-source), and for each study the code to reproduce our experiments. This database facilitates reproducibility, enables researchers to use this data for improved training or external validation, and facilitates public benchmarking.

13.1 Contributions and impact

13.1.1 Methodological

Analysis of quantitative medical image features has been performed for several decades. However, when in 2012 the term “radiomics” was coined [15], the field quickly gained more attention. When I started my PhD in 2016 on radiomics, the idea was to start with a straight-forward radiomics study to get familiar with the field. However, while radiomics had just started to gain popularity, there already was a proliferation of radiomics methods. Hence I wondered which method would work best for my specific application. An additional challenge was the fact that only a small percentage of the radiomics studies publicly released their software. Thus, besides finding a suitable method, I would also have to reimplement the method, which can be challenging.

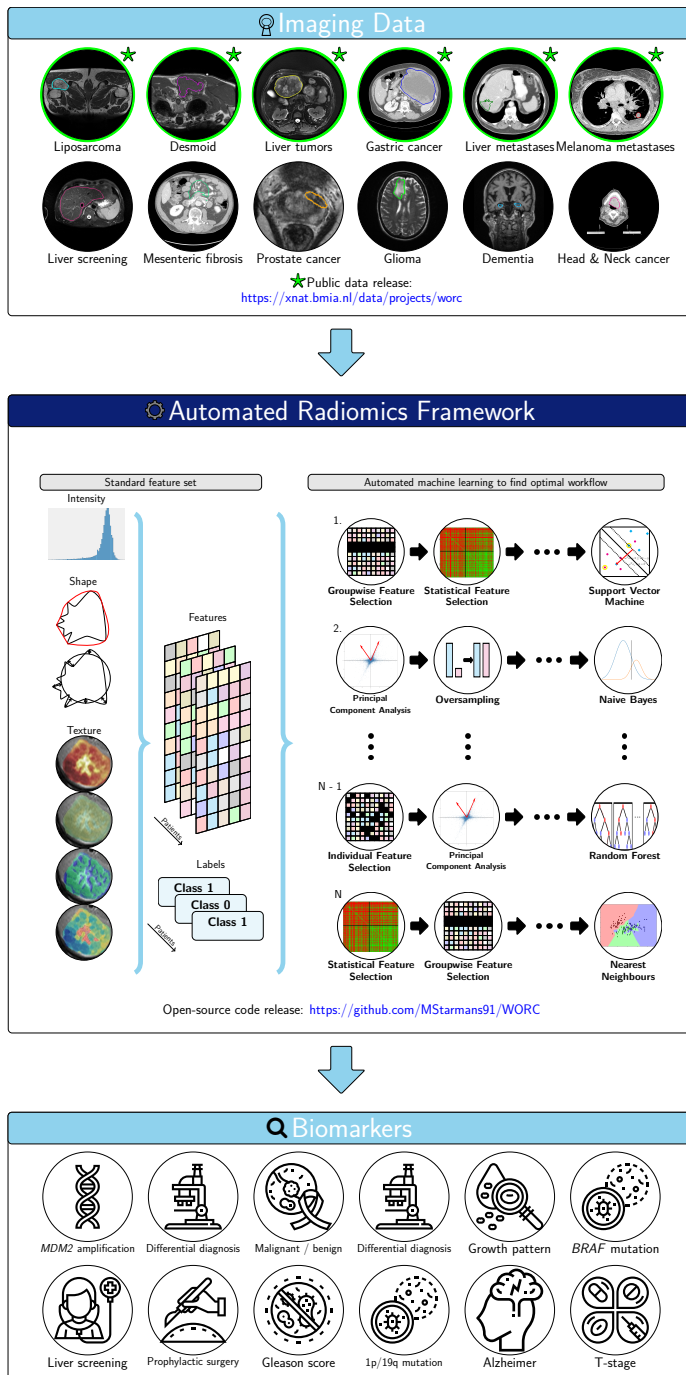


Figure 13.1: Schematic overview of the work presented in this thesis. We have performed radiomics studies on various clinical applications based on imaging data, which we have fed into the adaptive radiomics framework that we presented, resulting in a multiple biomarkers relating quantitative imaging features to a specific clinical label or outcome. We have publicly released a database composed of six datasets and the toolbox for our adaptive radiomics framework open source.

The aim of my PhD was therefore to create one radiomics framework in which multiple approaches could easily be integrated, and in which the construction of radiomics models using these approaches would automatically be optimized. In this way, radiomics could be generalized across clinical applications.

The main contribution of this thesis is to exploit recent advances in automated machine learning (AutoML) to automatically construct and optimize complete radiomics workflows. The construction of a radiomics workflow was formulated as a Combined Algorithm Selection and Hyperparameter (CASH) optimization problem, thereby enabling the use of AutoML. In order to formulate radiomics as a CASH problem, a modular approach was used, defining radiomics as a *workflow*, i.e., a specific combination of algorithms and their associated hyperparameters. A wide variety of algorithms were included in the framework. The original CASH mathematical formulation from the field of machine learning was extended to encompass the complete radiomics workflow. The formulation allows straight-forward integration of additional algorithms into the search space. As optimization strategy, a straight-forward random search algorithm was used, as it is efficient and often performs well. Hyperensembles were introduced to combine different workflows into a single model, improving both the performance of the resulting radiomics model and the stability of the workflow optimization. The resulting framework was extensively validated in twelve applications, which showed the generalization of the method across clinical applications. These aspects are described in [Chapter 3](#).

From an engineering perspective, I implemented my method as a software package in Python, resulting in the WORC (Workflow for Optimal Classification) toolbox. The WORC toolbox enables users to conduct a validated, standardized radiomics baseline for any study with minimal effort. In this way, for a new application, one can quickly probe a dataset for potential biomarkers. This has enabled me to conduct a large number of clinical radiomics studies. The WORC toolbox was released as an open-source software package, including extensive documentation [68], and tutorials¹. In this way, my work enables others in their radiomics studies to efficiently probe datasets for radiomics biomarkers by conducting a validated, standardized baseline with one press of a button.

In one study, and in the secondary goals of two of the other studies, we were not able to find a biomarker that performed well: the mutation stratification of lung metastases of melanoma ([Chapter 9](#)), and the secondary goals in DTF ([Chapter 6](#)) and GIST ([Chapter 7](#)). Coincidentally, these are three of the four studies (the fourth is the 1p19q mutation stratification from [Chapter 3](#)) involving the prediction of genetics using radiomics, also coined “radiogenomics”. In the DTF and GIST studies, this may be attributed to the relatively small sample sizes of the mutation analysis, see the descriptions in the respective chapters. However, in the melanoma study ([Chapter 9](#)) (169 lesions from 103 patients), to our knowledge, the dataset is currently the largest CT-based radiomics study on the *BRAF* mutation status in patients with metastatic melanoma. We concluded in this chapter that there is no relation between CT-based imaging features and the *BRAF* mutation. The study thus had a negative result. Currently, there is a positive publication bias in radiomics, with as few as

¹<https://github.com/MStarmans91/WORCTutorial>

6% of the studies between 2015 and 2018 showing negative results as reported in a recent study [88]. The authors of this study indicate that, to overcome this bias, sound methodology, robustness, reproducibility, and standardization are key. Hence, by addressing these hurdles of publishing negative results in this thesis (mainly Chapter 3), I hope to contribute to overcoming the positive publication bias [88].

13.1.2 Clinical

From a clinical perspective, in this thesis, we have shown the following:

- Chapter 5: radiomics **can** distinguish between well differentiated liposarcomas and lipomas on MRI.
- Chapter 6: radiomics **can** distinguish desmoid-type fibromatosis (DTF) from non-DTF tumors in the DTF differential diagnosis on MRI, but **could not** predict genetic mutations in DTF.
- Chapter 7: radiomics **can** distinguish gastrointestinal stromal tumors (GISTs) from non-GIST tumors in the GIST differential diagnosis on CT, but **could not** predict genetic mutations in GISTs.
- Chapter 8: radiomics **can** classify high grade versus low grade prostate cancer on multi-parametric MRI.
- Chapter 9: radiomics **cannot** determine the *BRAF* P.V600E mutation status of melanoma lung metastases on CT.
- Chapter 10: radiomics **can** predict symptomatic mesenteric mass in small intestinal neuroendocrine tumors on CT.
- Chapter 11: radiomics **can** distinguish pure replacement from pure desmoplastic HGP of colorectal liver metastases on CT.
- Chapter 12: radiomics **can** distinguish malignant from benign primary solid liver lesions on MRI.

Additionally, in Chapter 3, we validated that:

- radiomics **can** distinguish livers in which no hepatocellular carcinoma (HCC) developed from livers with HCC at first detection during screening on MRI.
- radiomics **can** predict the 1p/19q co-deletion in patients with presumed low-grade glioma.
- radiomics **can** distinguishing patients with Alzheimer's disease from cognitive normals on MRI.
- radiomics **can** predict the T-stage in patients with head-and-neck cancer.

Note that historically, the WORC method has evolved over time, thus leading to differences within the method between the papers described in [Part II](#). These include differences in the extracted features, the WORC WORC search space, the WORC optimization algorithm (specifically the number of random search iterations and ensemble size), and the ratio's used in the cross-validations. However, in the paper presenting WORC as described in [Chapter 3](#), the experiments of the studies in [Part II](#) were all repeated using the exact same WORC version, i.e., the version released at the moment [Chapter 3](#) was submitted. Although this version was different from those used in the papers described in [Part II](#), the results were similar.

For all but one study (the melanoma study described in [Chapter 9](#)), we have thus found a successful radiomics biomarker that predicts a clinical label or outcome based on medical imaging. The resulting biomarkers are not yet ready for usage in clinical practice, see [Subsection 13.2.5](#) on how to overcome the remaining challenges. However, these biomarkers can currently be used in research, for example in clinical trials for inclusion, screening, and monitoring, by radiologists to compare their scoring to radiomics, or to detect previously unknown patterns in imaging features which can be studied to uncover relations between underlying biological processes and outcomes. An example of this is the MINIMALIST clinical trial, in which new methods to minimize the invasiveness of liposarcoma treatment are studied [324]. In this study, the radiomics model will be used to aid in the monitoring of patients to detect changes in the tumor phenotype. Another radiomics research use-case, is to aid in predictions for which there currently exists no alternative. For example, in [Chapter 11](#), we used radiomics on pre-operative imaging to predict histopathological growth patterns (HGP) in colorectal liver metastases. Currently, these HGP can only be determined post-operatively on resection specimen. Being able to predict the HGP pre-operatively enables a whole new direction of research, for example to analyze the relation between HGP and the response to pre-operative chemotherapy treatment. Currently, this is not possible, as chemotherapy may alter the HGP [104].

The clinical applications evaluated in this thesis vary in the disease studied, body area, imaging modality, and image acquisition (e.g. scanners from Philips, Siemens, General Electric, and Toshiba). Moreover, we used radiomics for various prediction tasks: phenotyping, prognosis, genetic mutations, differential diagnosis, diagnosis of malignant versus benign, future symptoms, grading, and HGP. We thus showed the broad scope of our radiomics method, both in terms of clinical area and outcome variables. These insights can help both technical and clinical researchers to identify promising directions for new radiomics studies.

Most datasets used in this thesis were gathered in the Erasmus MC, which generally serves as a tertiary referral center, resulting in multi-center imaging datasets. As we imposed little to no restrictions on the included acquisition protocols, the datasets show substantial differences in the imaging hardware and acquisition parameters, and thus show heterogeneity in the image appearance. In spite of these differences, our radiomics method was able to successfully develop biomarkers. We thus showed that overcoming these differences is possible by training on multi-center, heterogeneous datasets, and thus that radiomics models can be used on routinely acquired clinical scans, paving the way for the translation to clinical practice. However, to enhance the applicability of our models, further generalization is required, see

Subsection 13.2.6.

In most applications described in this thesis, the performance was compared with clinical practice, i.e., scoring by a clinician, primarily radiologists. The clinicians were asked to perform the same task as the radiomics model: based on the same image as the radiomics model had access to (or including additional images), predict the same outcome as the radiomics model was tasked with. In all studies in which clinicians visually scored the images, there was substantial variation in their predictions. This indicates that in these applications, assessing these images is not trivial and subject to observer variability. The scoring by the clinicians gives insight into which characteristics have the highest predictive value, which cases can be considered pitfalls or easy to score, and on which cases there is consensus and or substantial disagreement. Concluding, I showed the importance of comparing radiomics to current clinical practice, as this aids in assessing when the performance of a radiomics model is sufficient to serve as a clinical aid. Moreover, this helps to identify the cases where the complementary value of radiomics is the highest. I recommend to always include multiple observers to score when comparing radiomics with scoring by clinicians.

In several applications, the radiomics model performed substantially or even statistically significantly better than the scoring by the clinicians. Combined with the substantial disagreement in the scoring, this indicates the potential complementary value of radiomics with respect to visual assessment. Besides assessing the performance, we also analyzed which image features contributed to the resulting radiomics models. This provides clinicians insights into which imaging characteristics and associated biological phenomena relate to the outcome of interest. For example, in [Chapter 5](#), we showed that volume alone was a good predictor of whether a lesion was a lipoma or well differentiated liposarcoma (WDLPS). While it was already known that WDLPS were generally larger than lipoma [113], this model performed better than the scoring by the clinicians. Hence, if the clinicians would have solely based their decision on volume, their accuracy would have improved. Such radiomics derived insights may help improve the decision making by clinicians; help in the acceptance of radiomics, as they may show that part of the radiomics models is based on characteristics that were clinically already known to be relevant; and improve the interpretability of radiomics models.

Caution should be taken when drawing strong conclusions on the comparison of radiomics with scoring by clinicians however. Not all of the visual scores the clinicians were tasked with in this thesis are routinely performed in clinical practice. Instead, for most of the described tasks, biopsies are the gold standard. Moreover, while our radiomics models were tasked to predict pathological outcome, the primary added value of radiomics to clinical practice would be to provide quantitative information about imaging data. The hypothesis of radiomics in this context is that the quantitative information relates to the pathological outcome, but there may be discrepancies. Similar discrepancies may therefore arise when comparing the prediction of a radiomics model which predicts pathological outcomes to visual scoring of clinicians who quantify imaging properties.

Machine learning models are only as good as the data they were trained on [4, 29]. Hence, if a dataset includes a bias, the resulting model may learn this bias

to leverage its performance. In this thesis, we evaluated (potential) biases in our datasets relating to: volume (Chapter 5), selection due to specific inclusion and exclusion criteria (Chapters 5, 6, 12, and 11), age and sex (all chapters in Part II), class imbalance (all chapters in Part II), tumor location (Chapters 6, 7, and 10), data originating from specific centers (Chapters 8 and 12), usage of contrast during image acquisition (Chapter 10), and segmentations by various observers (Chapters 6, 7, 11, and 12). We showed that some of our datasets contained biases, which our radiomics models in several studies leveraged to increase the overall performance when allowed to use these biases. These findings emphasize the importance of assessing potential biases in any radiomics study to prevent the development and use of biased biomarkers. Simply classifying patients based on one of these biases towards clinical characteristics, e.g., only using age, sex, tumor location, or purely based on acquisition parameters or from which hospital a scan originated, regardless of any imaging information, would be unacceptable and cannot be applied in clinical practice.

13.1.3 Open science

Open science has become increasingly important in recent years [325]. Open science is crucial for reproducibility, which is a defining feature of science. Besides providing the data, software and guides enabling replication of a study are essential to guarantee reproducibility [87]. However, open science requires substantial additional efforts by researchers, such as addressing legal practices, ethics, secure data storage and infrastructure, usage instructions, and covering the associated additional costs. While these efforts may not directly be rewarded, it has been shown that open science ultimately brings substantial benefits to the researcher [326]. Moreover, large-scale open science efforts such as OpenML [87], an initiative from the machine learning community to build an online ecosystem including data sets, associated scientifically tasks, training routines, and trained models, and <https://grand-challenge.org> [327], a platform to host and facilitate the organization of challenges in medical imaging, have led to substantial advancements in their respective fields.

All studies conducted in this thesis mainly rely on the methods implemented in the WORC toolbox, which I published open-source [36]. The WORC toolbox enables users to automatically conduct a validated radiomics baseline with minimal effort. Users simply need to provide their imaging data, segmentations, and labels to be predicted.

Moreover, for all studies conducted in this thesis, the code to reproduce the experiments has been released open-source [37, 134, 153, 184, 214, 244, 272, 295, 316]. Together with the open-source release of the WORC toolbox, this facilitates reproducibility of our studies. Moreover, researchers may use the provided code to repeat our experiments on other datasets to externally validate our methods.

In medical image processing, there is a need for providing access to large, multi-center datasets, to improve the training of radiomics method, to benchmark radiomics methods, and especially for external validation [1, 17, 20, 24, 25, 27, 30, 31]. Examples of existing initiatives to address this need are the Cancer Imaging Archive (TCIA), which hosts a variety of medical imaging datasets [328], ADNI [329],

and grand challenges such as CADDementia [330], LITS [110], BRATS [331], and the Medical Decathlon challenge [93]. These initiatives have created an enormous impact in their respective fields.

In this thesis, in [Chapter 4](#), we described the WORC database, which we publicly released, consisting of data from 930 patients from six radiomics studies. The public release of this dataset facilitates reproducibility of our studies. The database has been collected in routine clinical care at multiple centers, thus representing real-life variability and heterogeneity of imaging data. This database can be used to validate, benchmark, or develop radiomics methods, but also for automated segmentation methods. For most of the clinical applications included (WDLPS, DTF, GIST, primary liver tumors), to our knowledge, these are the first (large) datasets to be publicly released.

13.1.4 Collaborations

The core of radiomics methods involves medical image processing and machine learning. However, it is an interdisciplinary field, in which combining domain knowledge from various expertise is essential for success. The work as presented in this thesis is the result of many collaborative efforts between anesthesiologists, clinical technologists, computer scientists, endocrinologists, engineers, gastroenterologists, hepatologists, internists, medical physicists, oncologists, pathologists, physicists, programmers, radiologists, surgeons, and urologists. For example, the study described in [Chapter 3](#) is a collaborative effort by 46 authors associated with thirteen different departments. The key to success in radiomics therefore lies in the convergence of these technical and clinical disciplines to combine their respective knowledge and skills towards one common goal. This thesis is an example of reaping the benefits of convergence, in which I have contributed to establishing the links between the different departments and various networks, within the Erasmus Medical Center, nationally, and internationally.

13.2 Roadmap for future research and vision

13.2.1 Expanding and extending the horizon of radiomics applications

This work in streamlining and automating the construction and optimization of the radiomics workflow facilitates quick and standardized probing of datasets for developing radiomics biomarkers. New directions for research and collaborations have already been initiated: I am currently using WORC in various new studies including malignant peripheral nerve sheath tumors [332], retro-peritoneal sarcoma, bladder cancer [333], liquid biopsies in melanoma, liver cancer screening [78], colorectal liver metastases screening and prognosis, hypertrophic cardiomyopathy [334], magnetoencephalography [335], a clinical trial towards minimally invasive liposarcoma treatment [324], breast cancer chemotherapy response [336], and complex regional pain syndrome [337]. These projects involve many new collaborations, including other researchers both from the Erasmus MC using my framework and external users from other institutes or companies. I would like to invite others to try out my

WORC framework on their clinical applications to assess whether these hold potential radiomics biomarkers.

Besides expanding the scope of radiomics to other clinical applications, WORC can also be applied to other modalities. This thesis concentrates on using MRI and CT, as these were the modalities of choice in routine clinical practice in the applications included. For example, WORC may perform well in other radiological imaging modalities such as PET or US: part of the default features included in WORC have already been shown to result in radiomics biomarkers in other PET and US studies [99, 100]. It would be interesting to see on which types of imaging data WORC can also successfully be applied to find imaging biomarkers, and on which ones different approaches are required.

In this thesis, I have focused on binary classification problems. In the WORC toolbox, methods for multiclass classification and regression are also included, and a start has been made for multilabel and survival predictions. Future research involves the expansion of the WORC landscape to these types of problems.

Lastly, in this thesis, per patient, only a single time point has been analyzed. Future research involves analyzing multiple timepoints through radiomics, also coined “delta radiomics”. Delta radiomics may be used to make predictions at various time points. Additionally, delta imaging features, i.e., features that are defined using data from multiple time points, may be used to create models that can make predictions over time, e.g. at future time points.

13.2.2 Methodological innovations

In this thesis, as described in [Chapter 3](#), the construction and optimization of the radiomics workflow was formulated as a CASH optimization problem. As optimization strategy, we used a straight-forward random search algorithm, as it is efficient and often performs well [48], as also observed in this thesis. However, a wide variety of other optimization strategies exist, see [338, 339, 340, 341] for more detailed overviews. Examples include particle swarm optimization [342], genetic programming [343], tree based optimization [91, 344]), and covariance matrix adaptation evaluation strategies [345]. Employing these optimization strategies instead of the random search may improve the prediction accuracy, but may result in a higher computational burden and additional complexity.

One of the most popular and most promising optimization strategies is Sequential Model-Based Optimization (SMBO) [346], specifically applying Bayesian optimization [34, 340, 341, 347]. In the work of one of my students [348], we compared Bayesian optimization to the default random search in WORC. Summarizing this work, in the performed experiments, while more computationally demanding, the Bayesian optimization approach did not result in a better predictive accuracy than the random search strategy. While the performance of the resulting radiomics workflows was higher in the validation sets, which is what the optimization strategies are optimizing, this increase in performance did not generalize to the test set. These results thus validate the use of randomized search as an optimization strategy. Although other optimization strategies might show an improvement over the randomized

search, this study clearly showed the limits of optimization and the importance of generalization.

I believe that automated machine learning is the future to overcome the hurdles and deficits of manual tuning in medical image processing. The field of radiomics and thus the number of proposed methods is still rapidly expanding. Hence manual tuning, in addition to the many disadvantages, is quickly becoming infeasible altogether. However, besides optimization, at least equal attention should be given to the generalization of the resulting models to prevent overfitting. To this end, I think three factors are key. First, to prevent overfitting on a single dataset, I propose to learn from previous problems which solutions worked best instead of starting from scratch. To this end, we can learn from the public datasets as presented in [Chapter 4](#). To learn from these previous problems, a variety of solutions from the field of meta-learning can be used [\[338\]](#). Options range from simple solutions such as warm starting the optimization with workflows that worked well on previous studies, or favoring workflows that have shown to generalize well in previous studies. Second, instead of optimizing for validation performance, I suggest to create an ensemble of complementary solutions using multi-objective or Pareto optimization [\[349, 350\]](#). Instead of simply optimizing a single objective function focused on the performance on a validation set, multi-objective optimization can additionally include metrics to ensure the generalization of the found solutions. Lastly, other ensembling strategies could be employed [\[50\]](#). In this thesis, we have shown that ensembling is a powerful strategy to improve the performance and stability of machine learning solutions. The used ensembling strategy was effective, but also simple. Hence I suggest to exploit ensemble learning in the multi-objective optimization.

Radiomics can be seen as a collection of various disciplines from computer science and medical image processing. Beyond the default machine learning used radiomics, there are various disciplines from both the fields of image processing and artificial intelligence which are interesting to integrate into radiomics. In my opinion, three of these are key to improve radiomics models. Firstly, the interpretation of radiomics models is vital for its acceptance and translation to clinical practice [\[351\]](#). Whereas a segmentation algorithm is relatively easy to verify, i.e., the result can be visually inspected, the prediction of a radiomics model is not, and thus interpretability is crucial. Over the last years, as a result of the steep increase in radiomics studies, interest in radiomics' interpretability has increased. One could argue that models should be interpretable in itself instead of trying to explain black box models [\[352\]](#). I believe that we have to accept that some of the predictions we are trying to achieve with radiomics may be too complex to fully grasp. Medicine already makes use of black box concepts: for example, while someone may not fully understand PET physics, Fourier transformations used in MRI, or reconstruction of CT, that person may still have a deep understanding of how to use these images to enhance their (clinical) performance [\[25\]](#). However, it is important to make radiomics models as interpretable as possible. The benefit of using conventional radiomics features is that these are to a certain extent interpretable, i.e., the decision models are based on a very specific set of features. Moreover, insight into the decision models can be gained by looking at the individual predictive values of single features or feature groups as we did in most chapters in [Part II](#). However, some features are complex

to understand in itself. For example, while the mean of the intensities in a tumor is simple to understand, exactly determining what kind of patterns a feature such as “the cluster shade of the gray level co-occurrence matrix after application of a Laplacian of Gaussian filter” quantifies is complex. The final decision rules are often complex as well, especially when using higher-order non-linear decision rules, which are commonly used. Hence, while the standardized set of features and their individual predictive values give some insight into the interpretation of radiomics models, the mentioned factors limit the current interpretability of radiomics models. Two areas that I think are promising for future research to interpret these black box models, are sensitivity analysis [353] and game theory [354]. These methods are model agnostic, scalable, and the concepts used are relatively simple and common, and thus interpretable in itself.

Secondly, closely connected to interpretability, is the uncertainty of radiomics models. I hypothesize that a correct estimation of a model’s uncertainty may both add to the interpretability of the model and improve the overall model performance. For example, when applying a radiomics model, one could decide to reject the model’s prediction if the model is highly uncertain. Future research should thus include improved uncertainty estimates and taking uncertainty into account in the model predictions to improve the performance and interpretability. A relatively simple first step could be to look at the variations in the used ensembles [50], which is also model agnostic and scalable. Afterwards, value-of-information analysis could be used to evaluate the value of such information on a model’s uncertainty [355].

Thirdly, in radiomics based on multiple imaging modalities (e.g. CT and MRI) or multiple image sequences (e.g. T1w MRI, T2w MRI, and DWI), one has to deal with additional challenges. First, missing data is common due to the lack of image acquisition standardization. Second, the sequences are generally not spatially aligned, which may hamper radiomics methods. In this thesis, we used relatively simple methods to deal with these challenges, using imputation to estimate the missing feature values for the missing sequences, and conventional image registration methods [127] to spatially align the sequences. However, recent advances in the use of deep learning both for dealing with missing data and for medical image registration may provide better alternatives. For example, to deal with missing data, image synthesis may be used to replace the missing sequences itself. A promising solution for image synthesis is the use of cyclic generative adversarial neural networks, as these can deal with unpaired training data, and thus with missing data [88]. To deal with spatial misalignment, VoxelMorph [356], a deep learning based registration framework, may not only improve the registration accuracy, but also the registration speed, which is crucial for translation to clinical practice. Alternatively, deformable convolutions can be used to directly incorporate spatial transformations in a classification neural network [357].

13.2.3 Conventional radiomics versus deep learning

In this thesis, I have focused on radiomics using conventional machine learning. In theory, deep learning holds much more potential than conventional radiomics methods, as deep learning can reach higher levels of complexity, for example learn-

ing which features work best for a specific application instead of using predefined features. The field of medical deep learning faces several similar challenges as conventional machine learning [21, 22, 23, 26], including a wide variety of available algorithms and the need for model selection and hyperparameter tuning per application. The same problem thus persists: on a given application, from all available deep learning algorithms, how to find the optimal (combination of) workflows? Future research may therefore include a similar approach to WORC to facilitate construction and optimization of deep learning workflows. In the field of computer science, automatic deep learning model selection is addressed in Neural Architecture Search (NAS) [94]. Simultaneous optimization of the network and model selection hyperparameters using gradient descent based optimization is highly efficient. However, the usage of gradient descent based optimization is complicated due to the large number of categorical or discrete choices to be made in the method selection. I therefore believe that formulating the optimization as a CASH problem and using associated optimization strategies is the most promising approach, including adoption of the meta-learning and multi-objective optimization approaches I previously suggested. Conventional radiomics models may complement deep learning models [358], especially on small datasets where learning features is difficult. Future research should therefore include optimally combining both approaches in a hybrid solution.

In this thesis, I have focused on binary classification problems. Clinical applications may also include other tasks, such as multi-class or multi-label classification, regression, or survival. Additionally, while in this thesis segmentation was merely a means to an end, i.e., to be able to perform radiomics, segmentation can also be a task on its own. For all these tasks, I would suggest to use the same CASH approach as introduced in this thesis to optimize the construction of radiomics workflows from the relevant algorithms. The WORC framework already includes multi-class classification, which was performed to address the secondary aims of Chapter 6 and Chapter 7, and regression. As clinical applications may require multiple of these tasks to be performed, I suggest to combine the relevant tasks in multi-objective optimization to exploit the various forms of information during training (e.g., simultaneously using classification labels and survival labels).

13.2.4 Integrated diagnostics

In personalized medicine, integrating complementary information from different sources is key to aid clinical decision making. In this thesis, we primarily used radiological imaging (MRI and CT) to create biomarkers. In most studies, we combined imaging with other clinical or patient characteristics, such as age and sex (all studies in Part II), visual features manually scored by clinicians (Chapter 5: tumor depth, lobularity, atypical appearance; Chapter 9: Lung Image Database Consortium (LIDC) criteria [245]), tumor location (Chapters 6, 7, and 10), primary tumor information while looking at secondary tumors (Chapters 10, 11), various biological characteristics obtained from urine and blood samples (Chapters 10, 11), and disease progression markers obtained from e.g. blood and urine samples (Chapters 10, 11). Future research should focus on integrating all relevant information into one comprehensive model. Besides relatively simple clinical characteristics, such as the

ones mentioned and (basic) clinical information such as medication use, lab values, and medical history, four data sources seem most promising.

First, integration of radiological data with histopathological data could substantially improve classification performance. Histopathology-derived biomarkers often serve as the ground truth for radiomics, which also holds for most of the clinical application included in this thesis. Although the usage of machine learning on histopathology data has been around for decades, the term pathomics, i.e., radiomics for histopathological data, has emerged in recent years [359]. Integrating radiomics and pathomics is in theory relatively straight-forward, as radiomics methods can also be used on histopathological data. However, due to the difference in the data dimensions (i.e., 2D slices in pathomics versus 3D images in radiomics), the amount of detail, and the type of relevant features, the optimal radiomics and pathomics methods probably differ. Integrating radiomics and pathomics is not relevant in all applications, for example when histopathological data is not available at the time point of interest, e.g. before surgery or treatment.

Second, in some applications, radiologists outperform radiomics methods while using exactly the same data, as also shown in this thesis. To learn from radiologists, their visual scores may be used to improve radiomics methods. While both radiomics and the radiologist make use of the same radiological data, radiologists may use different information when visually scoring, relying on different prior knowledge and being able to quantify specific characteristics, compared to radiomics features extracted from the raw imaging data. Characteristics that are commonly visually scored by radiologists such as necrosis, lobularity, invasion, heterogeneity, presence of fat, presence of liquid, enhancement, and interaction with vena are generic, interpretable, and used in clinical practice to predict a wide variety of outcomes. Training radiomics methods to quantify these characteristics could greatly enhance their performance. However, scoring of these characteristics is generally not standardized, subjective, and the reporting unstructured. Adopting guidelines for standardized scoring and advances in structured reporting [360], both in radiology and in other fields, are therefore key to be able to incorporate this data in radiomics models.

Third, genetic data is currently one of the primary data sources for biomarkers in personalized medicine [1, 2, 3]. The field of genome wide association studies (GWAS) has substantially grown and successfully identified common traits in a wide variety of diseases [361]. In recent years, advances in imaging genetics have shown that the combination of genetic information and radiological imaging can lead to improved prediction models, both in terms of performance and interpretability [362]. Future research should therefore aim at combining these domains, either on the input level by developing methods to combined the data into a single biomarker, or by simply combining different biomarkers from the two domains into one decision model [363].

Lastly, more data is readily becoming available through the Internet of Things (IoT) [364]. In healthcare, these include wearables, actuators, communication technology, providing information on e.g. activity, nutrition, medication use, symptom detection. While relatively new, steadily more of these devices are becoming part of daily live, e.g. smartphones are nowadays able to record a majority of this information through a wide variety of apps. However, most of the data is not yet

incorporated in medicine or even accessible. Once these hurdles are however tackled, to deal with this “big data”, the usage of machine learning, and thus also radiomics to combine IoT data with imaging, holds much potential to create comprehensive biomarkers.

13.2.5 Translation to clinical practice

Radiomics holds much potential as an imaging biomarker for clinical practice, as it can be used at various stages in healthcare: as a gatekeeper in peripheral hospitals, as a selection method for biopsies and in some cases as an alternative to biopsies, as an aid to include patients in clinical trials, in follow-up monitoring, in assessing therapy response, and many more. Despite the steep increase in radiomics studies, the translation to clinical practice is relatively slow, posing numerous challenges [1, 5, 7, 18, 20, 24, 25, 27, 30, 31, 365, 366, 367, 368]. Based on the literature and my own experience, I have identified five key factors for the translation to clinical practice.

First, while this may seem trivial, radiomics studies should address clinically relevant problems. Radiomics researchers may define synthetic problems in radiomics studies, for example to demonstrate the validity of a method, or because of the availability of specific datasets. I define synthetic problems as problems including predictions that are not relevant for clinical practice, but also problems using inclusion or exclusion criteria that are not feasible to apply in clinical practice. While these studies are definitely valuable for research on radiomics, the resulting biomarkers are not relevant for clinical research. Additionally, not in all clinical applications will a radiomics biomarker add additional value to the current gold standard. In many applications, histopathology derived markers obtained from biopsies serve as a gold standard. Depending on the clinical application, the biopsy may be relatively non-invasive, risk free, and easy to conduct. Alternatively, some clinical problems can be easily solved by radiologists, as also indicated in some studies in this thesis, e.g. the differentiation between de-differentiated and well differentiated liposarcoma in [Chapter 5](#). It could be useful to automate routine tasks, but it may sometimes take more time verifying the output of a (radiomics) biomarker than performing the actual task. Hence, generally, in these applications, although it may be interesting to evaluate the potential of a radiomics biomarker from a research perspective, clinically, these add little value. Thus, for each clinical application, one should both estimate the potential value of radiomics complementary to that of existing approaches, and the required efficacy in order to contribute to clinical practice (e.g. diagnostic accuracy, robustness to acquisition variations) [294].

Second, there is a high need for large-scale multi-center datasets to develop and validate biomarkers developed through radiomics. Radiomics algorithms are only as good as their data [4, 29]. Hence, models trained on datasets containing specific demographics, acquisition protocols, or biases will not generalize. Moreover, as models might only have a high (or low) performance on such specific datasets, as a third party, valuing a radiomics biomarker is difficult without having access to the datasets. International datasets for standardized benchmarking of radiomics methods and biomarkers are therefore key for the validation and acceptance of these methods.

Third, assessment of radiomics studies is hampered, as it is difficult and subjective. This may be attributed due to the proliferation of radiomics methods and to the lack of reproducibility of radiomics methods and biomarkers. The latter is primarily a result of the lack of benchmarking data, and the fact that a substantial part of radiomics studies not publishing code containing the implementation of their methods or biomarkers. Standardized evaluation guidelines are thus required to objectively assess the quality of radiomics studies. Recently, there has been a trend in creating guidelines for the assessment of radiomics studies, from journals [25, 369], radiomics communities [30], or clinical domains [370]. These efforts are key to be able to properly assess, compare, and validate radiomics models, and thus decide when radiomics models are suitable for clinical practice. However, if these initiatives continue independently, this will result in a proliferation of evaluation guidelines, hampering not only the translation to clinical practice but also radiomics research. Moreover, caution should be taken when developing such guidelines: if the guidelines are too restrictive or demanding, they may substantially limit the scope of radiomics research.

Fourth, implementation of radiomics models in clinical practice will require a dynamic approach, as these models will probably not work perfectly upon first implementation. Ideally, besides detecting radiomics models detecting their own mistakes, e.g. using uncertainty as I suggested in [Subsection 13.2.2](#), they should be able to learn from their mistakes and thus improve over time. A promising solution to achieve this is the use of continuous or active learning (AL), which predicts which data should be labeled and added to the training dataset of a model in order to achieve a high improvement in performance [371]. Additionally, AL contains updating routines to take these new data-points into account. To maximally profit from AL, AL should already be taken into account in the radiomics model construction, as not all radiomics methods are suitable for AL. One open challenge of employing AL in clinical practice is how to perform quality assurance and verification of the results, e.g. in obtaining FDA and CE marking, which can only be solved by incorporating dynamic validation routines which can be repeated upon changing the radiomics models.

Fifth, and most important, is collaboration and convergence between the various disciplines involved in radiomics. To address clinical relevance, develop large-scale datasets, integrate standardization and evaluation guidelines, and continuously improve existing models, i.e., the four key factors discussed previously, researchers from various clinical and technical domains, and the clinicians who would ultimately use these radiomics models, have to collaborate. Without the support of clinicians, especially radiologists, radiomics will not make the transition to clinical practice. Recent surveys on AI in radiology in over 1,000 radiologists indicated fear of replacement, a lack of knowledge, lack of acceptance, and therefore various hurdles for the translation to clinical practice [367, 368]. Hence, to overcome these hurdles radiomics researchers should make an effort to actively reach out to clinicians: to identify which clinical problems are key; to provide education on radiomics to clinicians, and vice versa, to provide education about the clinical background and problems to engineers; and to increase the interpretability of radiomics models, and thus their acceptance. An example of such an initiative addressing these factors is the

“convergence alliance”, consisting of the Delft University of Technology, Erasmus Medical Center, and Erasmus University Rotterdam. To bridge the boundaries between clinical, technical, economical, and ethical sciences, The convergence alliance plans to create a technical university medical center [372, 373].

I find it highly unlikely that radiomics will replace radiologists. Rather, radiomics can serve as an aid for radiologists, replacing tedious, simple routine tasks, identifying patients or regions of interest to help in prioritizing tasks, and thus leave the radiologists more time to deal with more complex tasks (where radiomics can also serve as an aid), and free up time for research. To this end, radiomics should be seamlessly integrated into the radiologist’s workflow. Hence, more research should also be conducted on how radiologists can use radiomics models in a clinical setting.

13.2.6 Generalization

One of the most important aspects in future research on radiomics is the generalization: clinically, technically, methodologically, and demographically. As identified in the previous section, convergence and collaboration are key for generalization and thus for the translation of radiomics to clinical practice. While radiomics studies in clinical niches conducted in single centers are highly important to provide proof-of-concepts, large scale initiatives are required to bring radiomics to the next level. To illustrate this, I will shortly discuss various generalization initiatives in which I am involved and how such initiatives contribute to generalization.

Most radiomics studies target very specific clinical problems, including the studies described in [Part II](#) of this thesis. While in these studies, we set the inclusion and exclusion criteria to maximize clinical relevance, these limit the generalizability. Besides generalization to various types of a disease, generalization to different demographics and various types of image acquisition protocols is highly relevant for the translation to clinical practice. Gathering sufficient data to facilitate and evaluate generalizability can only be achieved in large-scale collaborative efforts between clinical centers. I have co-founded two of such initiatives. The first is coined the “SAI” (Sarcoma Artificial Intelligence) consortium [374], a collaborative effort between six medical centers and associations from different countries (The Netherlands, Poland, United Kingdom, and the United States of America). Our aim is to develop a comprehensive soft tissue tumor diagnostic radiomics model, both for phenotyping and grading, trained and validated in a large, multi-center cohort, and evaluated in a clinical setting. Second, I co-founded the “LAI” (Liver Artificial Intelligence) consortium [375], a collaborative effort between eighteen medical centers, patient and clinician associations, and companies, originating from eight different countries (The Netherlands, Belgium, France, Sweden, Spain, Italy, Austria, and the United States of America). Our aim is to create a benchmark MRI data collection, and a comprehensive MRI-based liver lesion phenotyping radiomics model. By combining domain knowledge from different fields, building large, multi-center, comprehensive benchmark datasets, and focus on clinical applicability and interpretability, we aim to contribute to the generalizability of radiomics.

The SAI and LAI consortia are primarily focused on clinical generalization, thus mostly including clinical experts. Complementary to these initiatives, the

European Cancer Imaging (EuCanImage) consortium aims to develop a cancer imaging platform to facilitate AI in oncology [376]. EuCanImage thus focuses on the technical, methodological, engineering, and infrastructure parts. The infrastructure and methods developed in EuCanImage to anonymize, curate, and store data, as well as to train and validate radiomics methods, could therefore substantially contribute to clinical initiatives such as the SAI and LAI consortia. Convergence of such complementary clinical and technical initiatives are key for the generalization of quantitative imaging biomarkers.

The question raises whether true generalization is feasible. In this thesis, we have shown generalization in dealing with data with substantial heterogeneity from different centers. However, this was mostly done on data from the Netherlands. Hence, to truly make these models generalizable, we should widen the scope, not only in terms of imaging heterogeneity, but also to incorporate variations in for example demographics and ethnicity. As machine learning models are only as good as the data on which they are trained, it is difficult to make hard claims relating to generalizability to unseen instances without comprehensively training (or testing) of these models on all expected data variations. However, a dataset containing sufficient samples of all expected data variations is an utopia. Therefore, a more realistic scenario is to locally adapt these models to the local data variations. To this end, the AL routines I suggested in [Subsection 13.2.5](#) can be employed.

I hope more of these initiatives will be founded in the future, expanding in scale and scope. Moreover, I hope that these initiatives will generalize even more, and one day into one big initiative in science to collaborate on a large scale in an open fashion.

13.2.7 Overcoming the hurdles towards open science

For radiomics, sharing of data is crucial for algorithm development and reproducibility, benchmarking, external validation, and therefore the translation to clinical practice. A recent study [28] even warned that radiomics research must achieve “higher evidence levels” to avoid a reproducibility crisis such as the recent one in psychology [87]. As open science has been shown to be beneficial for the careers of researchers [326], nothing seems to be in the way of open science. However, still relatively few radiomics studies publicly publish their data and make their code available open-source. I have identified two key factors to overcome this lack of open science.

Firstly, the current recognizing and rewarding of open science poses a paradox. On one hand, open science has been shown to contribute to the career of researchers [326] and is valued by the field, e.g. reproducibility is highly valued. Hence implicitly, open science is recognized and rewarded. On the other hand, open science is generally not explicitly recognized and rewarded, and thus researchers initially may not see or gain the added value and will not go through the additional efforts required by open science for free. The recent AI guidelines published by *Radiology*, one of the highest impact radiology journals, include that AI algorithms should be publicly available [25]. The recent guidelines from another high impact journal, *European Radiology*, do not mention sharing of algorithms [369]. Open science should

thus be explicitly recognized and rewarded. Journals should explicitly mention the sharing of data and algorithms in their submission guidelines, and incorporate evaluating these in their reviewing process. Institutes should encourage researchers to publish datasets and software implementations in dedicated journals. Grants should value open science: a good example is the NWO Open Science fund that was launched recently in 2020 [377]. Moreover, institutes and research communities should facilitate the processes involved in open science, such as providing standard data transfer agreements, ethical committee protocols, anonymization protocols, and data storing and sharing protocols and infrastructure.

Open science can be conducted in multiple ways. Instead of sharing data between different centers, algorithms can be shared such that the data does not have to leave the centers, which is known as “federated learning” [378]. Sharing algorithms is often easier from a legal and ethical perspective than sharing patient data. Federated learning is relatively new and faces various challenges [379], including the need for local expertise for data curation, need for local hardware and expertise to run the algorithms, and the need to ensure no leakage of privacy. Two major disadvantages of federated learning are that; 1) it negatively affects interpretability and bias evaluation, as researchers do not have access to the data; and 2) it limits the (radiomics) methods that can be applied, as not all central learning methods are suitable for federated learning. For these reasons, I think that federated learning brings more disadvantages than benefits, and that open science has more added value to science.

13.3 Conclusion

To streamline quantitative imaging biomarker development, this thesis describes an adaptive framework to automatically construct and optimize the radiomics workflow using automated machine learning. To facilitate reproducibility and provide a large, multi-center, clinically relevant database for benchmarking, improved training, and external validation, I contributed to open science by publicly releasing a large imaging database, releasing the code of my WORC radiomics toolbox open-source, as well as the code to reproduce all the experiments described in this thesis. The framework was validated and its generalization evaluated in a large number of clinical applications. In this thesis, in eight of these applications, the evaluation of my framework to develop biomarkers is described in detail: 1) liposarcoma and lipoma; 2) desmoid-type fibromatosis; 3) gastrointestinal stromal tumors; 4) prostate cancer; 5) melanoma lung metastases; 6) mesenteric fibrosis; 7) colorectal liver metastases; and 8) primary solid liver lesions. The resulting biomarkers give insight into the (absence of a) relation between quantitative imaging features and various clinical labels or outcomes, and can be used in a research context to discover more about these specific diseases to improve patient care. The adaptive framework can be used to efficiently probe datasets for biomarkers by conducting a validated, standardized radiomics baseline with minimal effort. I have identified several promising directions for future research, including methodological innovations, integrated diagnostics, how to facilitate the translation to clinical practice, and improve the generalization of radiomics biomarkers. Hopefully, one or more of these biomarkers can make the

translation to clinical practice in order to contribute to personalized medicine and to improvement of patient care.

Summary

The shift in medicine towards personalized medicine has led to an increased need for quantitative and objective biomarkers. These biomarkers can be used to relate patient data to a biological state, outcome or condition. Radiomics, i.e., the use of quantitative imaging features and machine learning, has shown many successes in various clinical applications to identify and extract such biomarkers from medical imaging. However, the field faces various challenges, including how to find the optimal method per application, a lack of public, large, multi-center cohorts, a lack of image acquisition standardization, and a lack of reproducibility for both radiomics methods and biomarkers in a routine clinical setting. Overcoming these barriers is vital for the translation of radiomics models to clinical practice.

In this thesis, I have addressed these major challenges in radiomics. In [Part I](#), I have made various contributions to the methodology of radiomics, including the development of an adaptive radiomics framework, and have contributed to the publicly available data for radiomics. In [Part II](#), in collaboration with clinical researchers from various disciplines, I used the developed framework to obtain novel insights into the use of radiomics-based quantitative imaging biomarkers in eight clinical applications. Each of these aspects is summarized below.

Part I - Adaptive radiomics framework

In [Chapter 2](#), a general introduction to radiomics is presented. This chapter covers the complete spectrum of a typical radiomics study, starting from the acquisition and preparation of the data and the segmentation of regions of interest, to the actual radiomics methods for feature extraction and data mining to create predictions models based on the radiomics features. Additionally, insights into the design of a radiomics study and an overview of the required infrastructure components are provided.

In [Chapter 3](#), I developed an adaptive radiomics to generalize radiomics across clinical applications. To this end, we state radiomics as a modular workflow, i.e., a specific combination of various algorithms and their associated hyperparameters. Examples of components included in this workflow are feature extraction, feature selection, dimensionality reduction, dataset resampling, and machine learning. For each component, the framework includes a large collection of commonly used algorithms in radiomics. To automatically construct and optimize the complete radiomics

workflow from this search space of algorithms and hyperparameters, I exploited recent advances in automated machine learning. To this end, the construction of a radiomics workflow was formulated as a Combined Algorithm Selection and Hyperparameter (CASH) optimization problem. Instead of selecting the single best workflow, hyperensembles were introduced to combine different workflows into a single model, improving both the performance of the resulting radiomics model and the stability of the workflow optimization. The resulting framework was extensively validated in twelve clinical applications, which showed the generalization of the method across applications. I implemented my framework in the WORC software toolbox, which was released open-source.

In [Chapter 4](#), I describe the publicly released WORC* database, consisting of imaging data, segmentations, and ground truth labels of in total 930 patients from six clinical applications. The database facilitates reproducibility and allows others to use the data for training, validation, and benchmarking of both radiomics and segmentation methods.

Part II - Novel radiomics biomarkers in clinical applications

In [Chapter 3](#), I validated my adaptive radiomics framework in twelve clinical applications. In this chapter, for each of these applications, a single, relatively simple experiment was performed. In eight of these applications, we studied the use of my framework in more detail to evaluate the potential of radiomics as a quantitative imaging biomarker in the specific application. While clinically independent, these studies were performed in a similar way. Each study includes multi-center, routinely acquired imaging data with little restrictions on the image acquisition protocols, resulting in heterogeneity in the imaging data. Next, my adaptive radiomics framework was used to develop biomarkers. For each study, the code to apply the WORC toolbox to perform the experiments was released open-source. Six of these datasets (those from chapters [5](#), [6](#), [7](#), [9](#), [11](#) and [12](#)) were included in the public dataset described in [Chapter 4](#).

In [Chapter 5](#), we evaluated the use of my radiomics framework to distinguish between well differentiated liposarcomas and lipomas on magnetic resonance imaging (MRI). Currently, the distinction is obtained using a biopsy and determining the *MDM2* amplification status using the obtained tissue. Our results showed that there is a relationship between radiomics features and the *MDM2* amplification status, and that radiomics models based on T1-weighted (T1w) and T1w + T2-weighted (T2w) MRI both outperformed three radiologists. Additionally, we showed that a strong relation between the tumor volume and the *MDM2* amplification status, but also that other radiomics features do provide additional predictive value.

In [Chapter 6](#), we evaluated the use of my radiomics framework for the differential diagnosis and mutation stratification of desmoid-type fibromatosis on MRI. Currently, the diagnosis is obtained using a biopsy followed by β -catenin staining of the obtained tissue and *CTNNB1* mutational analysis. Our results showed that there is a relationship between radiomics features and the differential diagnosis, and that the radiomics model based on T1-w MRI performed similar to two radiologists. Adding T2w or T1w post-contrast MRI did not substantially improve the model, indicating

that a plain T1w MRI scan may be sufficient to make the distinction. However, our results indicated that there is no relationship between the radiomics features and the *CTNNB1* mutational status, which is in line with the absence of literature linking DTF MRI appearance to the *CTNNB1* mutation.

In [Chapter 7](#), we evaluated the use of my radiomics framework for the differential diagnosis and mutation stratification of gastro-intestinal stromal tumors (GISTs) on computed tomography (CT). Currently, the diagnosis is obtained using a biopsy followed by pathological analysis of the obtained tissue. Moreover, as treatment planning of GISTs is also based on their molecular profile, the *c-KIT* mutational status or mitotic index (MI) are determined. Our results showed that there is a relationship between radiomics features and the differential diagnosis, and the radiomics model based on CT performed similar to three radiologists. However, our results indicated that there is no relationship between the radiomics features and the *c-KIT* mutational status or MI. As this may be attributed to the small sample size, we invite others to validate or improve upon our results on larger datasets or using different methods.

In [Chapter 8](#), we evaluated the use of my radiomics framework to distinguish high grade from low grade prostate cancer based on multiparametric MRI. Currently, the diagnosis is obtained using a biopsy followed by determination of the Gleason score. Our results showed that there is a relationship between radiomics features and the Gleason score, and that radiomics models based on multiparametric MRI outperformed two radiologists. Moreover, we evaluated the generalizability by training and externally validating the model in a multi-center, multi-vendor cohort. We showed that the models trained on data from a single-center showed a substantial drop in performance upon external validation, indicating major differences between centers in imaging of prostate cancer. However, the models trained on multi-center data showed a similar performance in the internal and external validation, indicating that training on multi-center data can aid in overcoming these differences. To make the transition to clinical practice, training and validation of radiomics models should thus be performed in multi-center scenarios with data representative of the population on which the model will be applied.

In [Chapter 9](#), we evaluated the use of my radiomics framework for the mutation stratification of melanoma lung metastases on CT. Specifically, we evaluated radiomics to predict the *BRAF* P.V600E mutation status, as this is used for treatment planning. Currently, the mutation status is obtained using a biopsy followed by using a polymerase chain reaction based assay or next generation sequencing. Our results showed that there is no relationship between radiomics features and the *BRAF* P.V600E mutational status of melanoma lung metastases, nor the visual scores of a radiologist. This validates my radiomics framework, showing that it does not invent a relation when one does not exist.

In [Chapter 10](#), we evaluated the use of my radiomics framework to predict symptomatic mesenteric mass in small intestinal neuroendocrine tumors on CT. Currently, there is no method to make this diagnosis: whether a mass is symptomatic is determined during follow-up, as some patients develop severe complications. Our results showed that there is a relationship between CT-based radiomics features and whether a mesenteric mass is symptomatic or asymptomatic. We showed that a radiomics model based on these features performed similar to five clinicians and

a multi-disciplinary tumor board. Moreover, our results indicated that radiomics features of the surrounding mesentery were most predictive, which was confirmed by the radiomics models.

In [Chapter 11](#), we evaluated the use of my radiomics framework to distinguish pure histopathological growth patterns (HGP) of colorectal liver metastases (CRLMs) on preoperative CT. Currently, preoperative HGP assessment is not possible, as assessment requires pathology slices of resection specimens to be reviewed with a light microscope. Our results showed that there is a relationship between radiomics features and the pure replacement and desmoplastic HGPs, and the radiomics model based on CT showed potential for automatic distinction of these two pure HGPs. Moreover, we combined the automatic classification of the radiomics model with a convolutional neural network for automatic segmentation of the CRLMs, and showed that the resulting model is observer independent and robust to segmentation variations.

In [Chapter 12](#), we evaluated the use of my radiomics framework to differentiate malignant and benign primary solid liver lesions on MRI. Currently in clinical practice, a first assessment is commonly made by the radiologist based on MRI. However, as the diagnosis is challenging, a biopsy is often performed followed by pathological examination to make the final diagnosis. In this study, we trained a radiomics model on a dataset from our hospital (i.e., the Erasmus MC, Rotterdam, the Netherlands) and externally validated the model in two datasets from the Maastricht UMC+ (Maastricht, the Netherlands) and Hôpital Beaujon (Paris, France). Our results showed that there is a relationship between radiomics features and the differentiation between malignant and benign liver lesions, and the externally validated radiomics model based on MRI performed similar to two radiologists.

Conclusion

In summary, I have developed an adaptive radiomics framework to streamline the development of quantitative imaging biomarkers. To this end, I generalized radiomics across applications by exploiting recent advances in automated machine learning. The framework was validated and its generalizability evaluated in twelve different clinical applications. I publicly released a large database to facilitate reproducibility and which others can use to improve the training of, externally validate, and benchmark radiomics and segmentation methods. Hence, my framework may be used to streamline the construction and optimization of radiomics workflows on new applications, and thus for probing datasets for radiomics signatures.

In collaboration with clinical researchers from various disciplines, I used the developed adaptive radiomics framework to obtain novel insights into the use of radiomics-based quantitative imaging biomarkers in eight clinical applications: 1) liposarcoma and lipoma; 2) desmoid-type fibromatosis; 3) gastrointestinal stromal tumors; 4) prostate cancer; 5) melanoma lung metastases; 6) mesenteric fibrosis; 7) colorectal liver metastases; and 8) primary liver cancer. The insights into the relation between quantitative radiomics features and these various diseases, and the additional predictive radiomics models, could be invaluable for transition of quantitative imaging biomarkers to clinical practice. The performances of the various

radiomics models reported in my studies and the comparison to scoring by various clinicians indicate that these biomarkers have the potential to improve the diagnostic work-up and treatment planning of patients in the future.

Future research can build upon my work by using my open-source software and public datasets. Moreover, future research could benefit by taking into account my recommendations presented in this thesis relating to the expansion of the horizon of radiomics applications, extensions of the automated machine learning approach including meta-learning and (multi-objective) optimization strategies, how to improve the interpretability of these models and leverage their uncertainty, the extension of radiomics with deep-learning-based approaches, integrated diagnostics, the transition to clinical practice, generalization, and open science.

Nederlandse samenvatting

In geneeskunde is er een verschuiving van een uniforme aanpak naar *personalized medicine*, i.e., een gepersonaliseerde, op maat gemaakte aanpak op basis van de unieke kenmerken van een patiënt. Hierdoor neemt de vraag naar objectieve biomarkers toe, welke gebruikt kunnen worden om data van patiënten te relateren aan bepaalde biologische processen, uitkomstmaten of aandoeningen. In radiomics worden kwantitatieve karakteristieken of *features* op basis van medische beelden berekend en in combinatie met machine learning gebruikt om biomarkers te identificeren en ontwikkelen. Radiomics is met veel succes in verschillende klinische toepassingen gebruikt is. Er zijn echter een aantal uitdagingen, waaronder het vinden van de optimale methode per toepassing, gebrek aan grote, publieke, multi-center cohorten, gebrek aan standaardisatie in het maken van de medische beelden, en de lage reproduceerbaarheid van zowel de radiomics methodes als de biomarkers in een routine klinische setting. Het overkomen van deze barrières is cruciaal voor de translatie van radiomics modellen naar de klinische praktijk.

In dit proefschrift heb ik mij gericht op deze uitdagingen in radiomics. In [Deel 1](#) heb ik verschillende methodologische bijdragen op het gebied van radiomics geleverd, waaronder het ontwikkelen van een adaptieve radiomics methode, en bijgedragen aan de publiek beschikbare data voor radiomics. In [Deel 2](#) heb ik, in samenwerking met klinische onderzoekers van verschillende disciplines, deze methode in acht verschillende klinische gebieden toegepast om nieuwe inzichten te verkrijgen in het gebruik van kwantitatieve biomarkers op basis van medische beelden en radiomics. Deze twee bijdragen zijn hieronder samengevat.

Deel 1 - Adaptieve radiomics methode

[Hoofdstuk 2](#) geeft een algemene introductie tot het veld van radiomics. Dit hoofdstuk beschrijft het complete spectrum van een typische radiomics studie, beginnend bij het verzamelen en voorbereiden van de data en het intekenen van de belangrijkste regio's in de beelden, tot de daadwerkelijke radiomics methodes voor het berekenen van features en methodes om op basis van deze features voorspellende modellen te ontwikkelen. Daarnaast wordt het ontwerpen van een radiomics studie behandeld en een overzicht gegeven van de benodigde infrastructuurcomponenten.

In [Hoofdstuk 3](#) heb ik een adaptieve radiomics methode ontwikkelt om radiomics te generaliseren naar verschillende klinische toepassingen. Om dit mogelijk te maken

beschouwen we radiomics als een zogenoemde modulaire workflow: een specifieke combinatie van bepaalde algoritmes en de bijbehorende hyperparameters. Voorbeelden van onderdelen die deze workflow bevat zijn het berekenen van features, het selecteren van features, het reduceren van het aantal dimensies, resampling van de dataset, en machine learning. Voor elk van deze onderdelen bevat de methode een grote collectie van algoritmes die in radiomics veel gebruikt worden. Om het construeren van de complete radiomics workflow uit deze verzameling van algoritmes en hyperparameters te automatiseren, en dit proces ook te optimaliseren, maak ik gebruik van recente ontwikkelingen uit het gebied van de automatische machine learning. Hiervoor formuleer ik het selecteren van algoritmes en hyperparameters om een radiomics workflow te construeren als een gecombineerd optimalisatieprobleem. In plaats van enkel de beste workflow te selecteren, maak ik gebruik van hyperensembles waarin verschillende workflows tot één model gecombineerd worden, wat zowel de prestaties van de resulterende modellen als de stabiliteit van de workflow optimalisatie bevordert. De resulterende methode is uitgebreid gevalideerd in twaalf klinische toepassingen, wat aantoont dat de methode generaliseert over verschillende toepassingen. Ik heb deze methode geïmplementeerd in de WORC toolbox, die ik beschikbaar heb gesteld als open source software.

In [Hoofdstuk 4](#) beschrijf ik de WORC* database, welke publiek beschikbaar is gemaakt. Deze database bestaat uit medische beelddata, tekeningen, en gouden standaard labels van in totaal 930 patiënten uit zes klinische toepassingen. Deze database bevordert de reproduceerbaarheid en staat anderen toe de data te gebruiken voor het trainen, valideren, en benchmarken van zowel radiomics methodes als methodes om tekeningen te genereren.

Deel 2 - Nieuwe radiomics biomarkers in klinische applicaties

In [Hoofdstuk 3](#) heb ik mijn adaptieve radiomics methode in twaalf verschillende klinische toepassingen gevalideerd. In dat hoofdstuk heb ik voor elk van deze toepassingen een enkel, relatief simpel experiment uitgevoerd. In acht van deze toepassingen hebben we het gebruik van mijn methode in meer detail geëvalueerd om zo de mogelijkheden voor het gebruik van radiomics als kwantitatieve biomarker voor medische afbeeldingen beter te analyseren. Hoewel deze studies vanuit een klinisch oogpunt onafhankelijk van elkaar zijn, hebben we deze op een vergelijkbare manier uitgevoerd. In elke studie wordt gebruik gemaakt van multi-center, routine verzamelde beelddata met weinig restricties op de acquisitieprotocollen van de beelden. Vervolgens hebben we mijn adaptieve radiomics methode toegepast om biomarkers te ontwikkelen. Voor elke studie is de code om de experimenten met de WORC toolbox uit te voeren open source beschikbaar gesteld. Zes van deze datasets, namelijk die uit [Hoofdstukken 5, 6, 7, 9, 11 en 12](#), vormen samen de publieke dataset die in [Hoofdstuk 4](#) beschreven is.

In [Hoofdstuk 5](#) hebben we het gebruik van mijn radiomics methode om onderscheid te maken tussen goed gedifferentieerde liposarcomen en lipomen op basis van magnetic resonance imaging (MRI) geëvalueerd. Momenteel wordt deze diagnose bepaald door afname van een biopsie en bepaling van de *MDM2* amplificatie. Onze resultaten laten zien dat er een relatie is tussen radiomics features en de *MDM2* am-

plificatie status, en dat radiomics modellen op basis van T1-gewogen (T1w) en T1w + T2-gewogen (T2w) MRI beiden beter presteerden dan drie radiologen. We hebben daarnaast aangetoond dat er een sterke relatie is tussen het volume van de tumor en de *MDM2* amplificatie, maar dat ook andere radiomics features voorspellende waarde bevatten.

In [Hoofdstuk 6](#) hebben we het gebruik van mijn radiomics methode voor de differentiaal diagnose en het onderscheiden van verschillende mutaties in desmoïd tumoren op basis van MRI geëvalueerd. Momenteel wordt deze diagnose bepaald door afname van een biopt, gevolgd door β -catenine kleuring en bepaling van de *CTNNB1* mutatie. Onze resultaten laten zien dat er een relatie is tussen radiomics features en de differentiaal diagnose, en dat het radiomics model op basis van T1w MRI vergelijkbaar presteert als twee radiologen. Het toevoegen van T2w of T1w MRI na contrast leidde niet tot substantiële verbeteringen in het model, wat aangeeft dat een gewone T1w MRI scan mogelijk voldoende is om dit onderscheid te maken. Echter, onze resultaten gaven aan dat er geen relatie bestaat tussen radiomics features en de *CTNNB1* mutatie, wat in lijn is met de afwezigheid van literatuur die de presentatie van desmoïd tumoren op MRI linkt aan de *CTNNB1* mutatie.

In [Hoofdstuk 7](#) hebben we het gebruik van mijn radiomics methode voor de differentiaal diagnose en het onderscheiden van verschillende mutaties in gastro intestinale stroma tumoren (GISTs) op basis van computed tomography (CT) geëvalueerd. Momenteel wordt deze diagnose bepaald door afname van een biopt, gevolgd door een histopathologische analyse van het verkregen weefsel. Aangezien het behandelplan van GISTs gebaseerd wordt op het moleculaire profiel worden daarnaast de *c-KIT* mutatie status en de mitose index (MI) bepaald. Onze resultaten laten zien dat er een relatie bestaat tussen radiomics features en de differentiaal diagnose, en dat het radiomics model op basis van CT vergelijkbaar presteert als drie radiologen. Echter, onze resultaten gaven aan dat er geen relatie bestaat tussen radiomics features en de *c-KIT* mutatie status of de MI. Omdat dit mogelijk toegeschreven kan worden aan de kleine omvang van de dataset, nodigen we anderen uit om onze resultaten te valideren of te verbeteren op grotere datasets of door het gebruik van andere methodes.

In [Hoofdstuk 8](#) hebben we het gebruik van mijn radiomics methode om hooggradig van laaggradig prostaatkanker te onderscheiden op basis van multiparametrische MRI geëvalueerd. Momenteel wordt de diagnose bepaald door afname van een biopt, gevolgd door bepaling van de Gleason score. Onze resultaten laten zien dat er een relatie bestaat tussen radiomics features en de Gleason score, en dat een radiomics model gebaseerd op multiparametrische MRI beter presteerde dan twee radiologen. Daarnaast hebben we geëvalueerd hoe ons model generaliseert door het te trainen en extern te valideren in een cohort uit meerdere ziekenhuizen en met scanners van verschillende fabrikanten. We hebben laten zien dat modellen die getraind waren op data van een enkel ziekenhuis substantieel minder goed presteerden in externe validatie, wat aangeeft dat er grote verschillen zijn in de MRI scans tussen de ziekenhuizen. De modellen die getraind waren op data van meerdere ziekenhuizen presteerden daarentegen vergelijkbaar in de interne en externe validatie. Dit laat zien dat het trainen op data uit verschillende ziekenhuizen kan helpen om de verschillen tussen de MRI scans te overbruggen. Om de transitie

naar de klinische praktijk te maken, zou het trainen en valideren van radiomics modellen dus gedaan moeten worden in datasets vanuit meerdere ziekenhuizen met data die representatief is voor de gehele populatie waarop het model toegepast zal worden.

In [Hoofdstuk 9](#) hebben we het gebruik van mijn radiomics methode om de verschillende mutaties van uitzaaiingen van melanomen in de long te onderscheiden op CT geëvalueerd. We hebben specifiek gekeken of radiomics de *BRAF P.V600E* mutatie status kan voorspellen, omdat deze gebruikt wordt in bepaling van de behandeling. Momenteel wordt de mutatie status bepaald door afname van een bipt, gevolgd door het gebruik van een reeks op basis van een polymerasekettingreactie of “next generation sequencing”. Onze resultaten laten zien dat er noch een relatie is tussen radiomics features en de *BRAF P.V600E* mutatie status van uitzaaiingen van melanomen in de long, noch tussen de visuele scores van een radioloog en de mutatie status. Dit valideert mijn radiomics methode, omdat dit laat zien dat mijn methode geen relatie verzint als deze niet bestaat.

In [Hoofdstuk 10](#) hebben we het gebruik van mijn radiomics methode om op CT te voorspellen of een mesenteriale massa van kleine neuro-endocriene tumoren symptomatisch is geëvalueerd. Momenteel is er geen methode om deze diagnose te stellen: of een massa symptomatisch is wordt bepaald tijdens follow-up wanneer bij sommige patiënten complicaties optreden. Onze resultaten laten zien dat er een relatie is tussen radiomics features op basis van CT en of een mesenteriale massa symptomatisch of asymptomatisch is. We hebben laten zien dat een radiomics model gebaseerd op deze features vergelijkbaar presteerde als vijf artsen en een multidisciplinair overleg. Daarnaast geven onze resultaten aan dat radiomics features van het omliggende mesenterium het meest voorspellend waren, wat bevestigd wordt door de radiomics modellen.

In [Hoofdstuk 11](#) hebben we het gebruik van mijn radiomics methode om pure histopathologische groeipatronen (HGP's) van colorectale levermetastasen (CRLMs) op preoperatieve CT geëvalueerd. Momenteel is het niet mogelijk om preoperatief het HGP te bepalen, omdat hiervoor pathologie coupes verkregen uit een resectie beoordeeld moeten worden met een microscoop. Onze resultaten laten zien dat er een relatie is tussen radiomics features en pure “replacement” en “desmoplastische” HGP's, en dat een radiomics model gebaseerd op CT potentie heeft om automatisch deze twee pure HGP's van elkaar te onderscheiden. Daarnaast hebben we het automatische classificatie model op basis van radiomics gecombineerd met een convolutioneel neurale netwerk om automatisch de CRLMs te segmenteren. We laten zien dat het resulterende model niet afhangt van een waarnemer en robuust is voor variatie in de segmentaties.

In [Hoofdstuk 12](#) hebben we het gebruik van mijn radiomics methode om maligne van benigne primaire solide levertumoren op basis van MRI te onderscheiden geëvalueerd. Momenteel wordt in de klinische praktijk een eerste inschatting gedaan door een radioloog op basis van MRI. Echter, omdat deze diagnose complex is, wordt er vaak alsnog een bipt afgenomen gevolgd door een histopathologische analyse om de uiteindelijke diagnose te bepalen. In deze studie hebben we radiomics modellen getraind op een dataset uit ons ziekenhuis (het Erasmus MC, Rotterdam), en extern gevalideerd in twee datasets uit het Maastricht UMC+ (Maastricht) en Hôpital

Beaujon (Parijs, Frankrijk). Onze resultaten laten zien dat er een relatie bestaat tussen radiomics features en het onderscheid tussen maligne en benigne tumoren, en dat het extern gevalideerde radiomics model op basis van MRI vergelijkbaar presteert als twee radiologen.

Conclusie

Samenvattend heb ik een adaptieve radiomics methode ontwikkeld om het ontwikkelen van kwantitatieve biomarkers op basis van beeldvorming te stroomlijnen. Om dit mogelijk te maken heb ik radiomics gegeneraliseerd door recente ontwikkelingen op het gebied van automatische machine learning te gebruiken. Deze methode is gevalideerd en de generaliseerbaarheid geëvalueerd in twaalf verschillende klinische toepassingen. Ik heb een grote dataset publiek beschikbaar gemaakt om de reproduceerbaarheid te bevorderen. Deze dataset kan door anderen gebruikt worden om het trainen van modellen te verbeteren, deze extern te valideren, en om radiomics en segmentatie methodes te benchmarken. Mijn methode kan gebruikt worden om het ontwikkelen en optimaliseren van radiomics methodes in nieuwe applicaties te stroomlijnen, en om in datasets op efficiënte wijze te zoeken naar nieuwe radiomics patronen.

In samenwerking met klinische onderzoekers van verschillende disciplines heb ik de ontwikkelde adaptieve radiomics methode gebruikt om nieuwe inzichten te krijgen in het gebruik van kwantitatieve biomarkers op basis van medische beelden en radiomics in acht klinische applicaties: 1) liposarcomen en lipomen; 2) desmoïde tumoren; 3) gastro intestinale stroma tumoren; 4) prostaatkanker; 5) long metastasen van melanomen; 6) mesenteriale fibrose; 7) colorectale levermetastasen; en 8) primaire levertumoren. De inzichten in de relatie tussen kwantitatieve radiomics features en deze verschillende ziektes, en daarnaast de voorspellende radiomics modellen, zouden van onschatbare waarde kunnen zijn voor de transitie van kwantitatieve biomarkers op basis van beeldvorming naar de klinische praktijk. De prestaties van de verschillende radiomics modellen gerapporteerd in mijn studies, en de vergelijking met het scoren door verschillende artsen, laten zien dat deze biomarkers potentie hebben om het diagnostische traject en de behandelplannen van patiënten in de toekomst te verbeteren.

Toekomstig onderzoek kan voortborduren op mijn werk door het gebruik van mijn open source software en publieke datasets. Daarnaast kan toekomstig werk profiteren van mijn aanbevelingen in deze thesis met betrekking tot het verbreden van de horizon van radiomics applicaties, het uitbreiden van de automatische machine learning benadering door het gebruik van meta-learning en (multi-objective) optimalisatie strategieën, het verbeteren van de interpreteerbaarheid van deze modellen en het benutten van hun onzekerheid, het integreren van verschillende vormen van diagnostiek, de transitie naar de klinische praktijk, generalisatie, en open wetenschap.

Acknowledgements

I could start by stating that writing this section is the final part of wrapping up my PhD trajectory. However, since I am currently staying in academia and on a topic in line with the work of my PhD, it does not feel like something truly ended. This is strengthened by the fact that being a PhD student is not merely a job or a “study” but rather adoption of a lifestyle, i.e., the academic lifestyle. I wonder if the feeling of finishing something will come after my defense, but I doubt it. I rather have the feeling that this is just the beginning, which is fine, as I’ve thoroughly enjoyed the past five years.

Like most people, I started writing this section as the final part of my thesis. However, after writing such a heavy document as a PhD thesis, you get a bit saturated with writing and lose part of your creativity. Which is a pity, as the acknowledgments is one of the most important sections. Especially in this area of science which heavily depends on teamwork. I would thus recommend everyone who needs to write a thesis to start with this section. Luckily, I’ve had ample opportunity to collaborate with many great people, which does result in a long list of people I have to thank for their contributions to this work. I will start with the people that were most crucial, my supervisors.

Stefan, thank you for your open yet critical way of supervision. From the start, you have given me the space to develop my own ideas and vision, and allowed me to venture in the directions of my interests. While being a great motivator, you also have quite a critical attitude, thereby always challenging me to defend my ideas and looking at them from all possible angles, leading to substantial improvements of my research. As my project and the number of papers grew a bit out of hand, you have spend quite some time reading my papers and providing feedback, which I am very thankful for. You are also very versatile, alternating between a philosophical person, a down-to-earth pragmatic person, my greatest supporter, my greatest critic, a leader with a vision, and even my secretary when I was taking on far too many projects than I should. I think we share quite some personality traits, including the ability to keep adding ideas and criticizing every detail that is not thoroughly motivated. This has led to many brainstorm spiraling out of control and lasting for hours, which I’m quite fond of as they always led to great ideas. However, we are also quite different or even opposites in some aspects, in which we nicely complement each other. I’ve thoroughly our collaboration and am thus very grateful that we will continue working together in the future!

Wiro, thank you for your modesty, trust in giving me a lot of freedom, and for the superb environment for PhD students you have created and are still improving. While Stefan and me have the urge to quickly dive into details, you were always able to zoom out and put our research in a bigger perspective. I have sometimes jokingly referred to you as the “head of PR” of our group due to the incredible amount of networks and connections you have. Due to this, you could always connect my work to existing initiatives, have given me the opportunity to present my work in various international communities, and are a great visionary advisor in general. I look forward to being a part of your large scale initiatives on convergence, and hope to attend many more conference (dance) parties together.

Jan-Jaap, thank you for being the clinical counterpart in my supervision, the many opportunities you have given me, and your entrepreneurial attitude. While only officially becoming one of my co-promotors in my final year, we have collaborated since the second month of my PhD. While starting out my PhD without a clear clinical project, you quickly provided me with a steady inflow of clinicians interested in conducting radiomics studies. You always kept in mind what the clinic itself would require of the solutions we were creating: I hope that one day one of our tools can truly make the transition.

Besides my promotor and co-promotors, I would like to thank the other members of my thesis committee: **Meike Vernooij**, **André Dekker** and **Karim Lekadir** thanks for taking the time to read and approve my thesis. Additionally, I would like to thank the other members of my defense committee: **Valérié Vilgrain** and **Kees Verhoef**, thanks for taking the time to join my defense and critically appraising my work.

Next, I would like to thank my clinical counterparts, the clinical PhD students without whom this work would not have been possible and who have shared the workload of various studies with me: **Melissa Vos**, **Milea**, **Lindsay**, **Anela**, **Florian Buisman**, and **Anne-Rose**. It was great having a sparring partner to continuously evaluate the research and debate about which directions were most promising to take. Also, writing papers becomes much easier when you can freely iterate with a co-author. When a reviewer is very critical, being able to handle the criticism together is much easier. Hence, thank you for all your time spent on our collaborations! And thank you for all the hours you have spend on segmenting CT or MRI scans: I hope it was worth it :). Good luck on your specialization journeys, and I sincerely hope we can collaborate again in the future!

I started out out this project as a physicist with very limited experience in medicine. Therefore, I had to heavily rely on clinical partners to both come up with clinically relevant questions to address and to provide me insight in the various clinical application I have worked on. I would therefore like to thank all my clinical collaborators, and specifically some of the main driving forces behind the projects in this thesis. **Kees**, **Dirk**, and **Stefan Sleijfer**, thank you for all your project ideas, your wisdom, your accessibility, and your endless enthusiasm. I’m very satisfied with the projects we have conducted together, from which we have reaped quite some benefits. As we all have many more ideas to follow-up our original research questions, I hope these projects can serve as a basis for expanding our collaboration in future projects, of which the Hanarth project is a good example. **Maarten Thomeer**, thank you for

giving me a thorough introduction in the radiology of liver cancer, your enthusiasm, and the many brainstorming and ideas. I must admit, you can be a bit chaotic at moments, but behind that is a keen mind with a lot of specialized knowledge. I've enjoyed many different moments with you, from going to ECR with you and deciding to start "scouting" for our new consortium, to going to Paris back and forth on a single day for a meeting and actually missing the Thalys on the way back. I look forward to expanding our research in our future collaborations! Related to this, I would also like to thank **Razvan** for all your help with this project, your kindness, and your optimism. **Astrid**, thank you for all your insights, your endless dedication to research, and your cheerfulness. I don't think I've met someone else who is more dedicated to improving patient care. We have a funny track record, because both our studies produced negative results, but we apparently conducted them in a proper way because they both got published more easily than the average positive study in this thesis. I hope we can continue our collaborations in the future, let's get that first positive result :). Last but not least, **Cecile** and **Else**, thanks for introducing me to an area I would have never expected to venture in during my PhD, and the co-supervision of many students. We have had a steady stream of students conducting internships on our project, which I enjoyed supervising with you. I think we helped many students in their education and started very interesting and unique direction for research.

From day one of my PhD, education has been one of my passions. Hence I'm very glad to have had the opportunity to participate in various educational activities. For that, I mostly want to thank **Jifke** for giving me this opportunity, teaching me about teaching, and giving me the freedom and trust to develop myself in this aspect. Thank you for all the effort, love, and dedication you are putting into education. **Marius**, thank you for setting up the "Medische Beeldverwerking" course, and the nice collaborative effort in teaching it over the years. The course was set up in such an amazing way that I could just jump straight in. I like to think that in this way, we have contributed a bit in the education of many students and hopefully have made several of them enthusiastic about image processing. **Thomas Höllt**, thank you for the joint teaching in the Advanced Image Processing course. Although a large part of the visualization topic you taught is still a bit of a mystery to me, I've enjoyed our collaboration, and admire your ability to make comprehensive slides and explain them at just the right pace. **Oleh**, thank you for the joint teaching in several courses. You were always available to jump in and help out on many aspects.

There are several key factors to a successful and enjoyable PhD trajectory. One of them is the environment in which you conduct your PhD. Hence, I am very thankful to have arrived in BIGR with such wonderful colleagues. You have all helped me in many sparring sessions to improve my ideas, solving bugs, sharing frustrations when something didn't work, but also socializing and relaxing to take my mind off work. Knowing you're not alone in the common struggles that come with doing a PhD can be a great relief.

Jose, you have been a close colleague and friend to me for many years. We have been to many courses, conferences, writing weeks, and summer schools together all over the world. I jokingly sometimes boast about the number of times we have shared a bed :). We have helped develop two courses and taught these together for

years, which I really enjoyed. Since we have become office roommates, I could daily spar with you about my ideas, ask you for your insights, and enjoyed many coffee breaks with you. Your enthusiasm and dedication make you a great researcher, your humor, kindness, and humility an even greater person. I am very grateful for all of these aspects, and that you agreed to be my paranymp and also help me out with this last aspect of my PhD.

Antonio, I'm still not sure how our friendship exactly started out. I guess naturally due to the fact that we were the early risers of the group, and thus had a daily coffee break ritual at 08.00. This was always a nice quiet moment to catch-up and start the day. Thank you for being a good listener, your calmness, sharing our "getting married in COVID-19 times" troubles, and the nice double-dates together with our (now) wives.

Sebastian, we started out our PhDs more or less at the same time and on similar topics. However, since you started a couple of months before me, you were always just one step ahead. Hence, whenever I encountered some issue, you were often just done solving it and had developed some nifty tool that you were always willing to share such that others did not have to go through the troubles you had been through. Hence, thank you for all your tools and help, and the nice discussions and sparring along the way of our PhDs. Additionally, your skills in organizing social activities such as the board game nights is only surpassed by your infectious enthusiasm for participating in these yourself. Hence, thank you for organizing so many activities and in doing so creating many wonderful memories.

Zahra, you are such a unique and wonderful person. Comparing you on the first time I met you (on my first day of my PhD) to now is like night and day. You have literally and figuratively come out of your shell, which I greatly admire, showing what a powerful person you are. What didn't change were your constant kindness, caring, and your brightening smile, for which I would like to deeply thank you. It sometimes almost felt like you were the mother figure of BGR. And please, don't scare me again by squeezing three whole limes in one bowl of soup.

Florian Dubost, you are one of the most dedicated researchers I have encountered. I still remember how you sometimes came to the office in the evening when you "finally managed to get away from your girlfriend to continue your work". You have a strong personality and are not afraid to express your opinions and diving in a firm debate when someone disagrees. I have enjoyed many discussions with you, both professionally and on a more personal level. Thank you for inspiring me with your dedication and attitude, and bringing sparks to boring days with your lively discussions.

Vikram, you are in some aspects the complete opposite of me. I admire how deliberate, patient, and calm you can be, which are nice and quite unique traits to encounter in such a stressful and competitive environment. You always think your answers and methods through, making you a formidable scientist. I have enjoyed our conversations over the years, admire your work, and am 100% sure you will succeed in your further scientific career.

Gerda, thanks for cracking many jokes together over the years. But also thanks for challenging me in our discussions as you are not someone to back of easy when making an argument :). You are always an enjoyable person at any party and

celebration. I still remember our first MICCAI, being one of the few people not afraid to go on the dance floor of the big party in a museum, and up close observing multiple professors conducting dance battles with students.

Shuai, thank you your majesty. Thanks for all the fun moments we had together, you seem to be smiling every moment of the day.

Gijs, in the category “deep waters still grounds”, you often seemed to have an extensive amount of knowledge hidden behind a calm, humble and patient person. Thank you for all the great discussions, and your help and insights on various topics.

Karin, thank you for together supervising many students over the years. Your realistic attitude and insights have protected many students from all my crazy ideas :). You are also one of those people always up for a coffee break or a social activity, so thanks for all the great moments!

Kim, I’m still not sure whether I was really your BIGH buddy or whether we acted as buddies for each other. Thanks for all the moments sharing our PhD struggles, but also discussing new ideas. You always put other people’s needs before your own and are eager to organize fun activities. To name a few, thank you for organizing and joining the BIGH outing and various writing weeks!

Thomas, thanks for all the discussions about WORC, your programming help, and all your other IT skills that have helped me out on numerous occasions. We could sometimes spend hours on hacking new stuff into WORC, discussing how certain routines could be improved, or finding bugs that I seemed to have hidden on purpose for you to discover. I’m glad to be allowed to now help out in the supervision of your PhD and look forward to diving into even more detail with you on your projects!

Wietske, thank you for being such a joyful person. You are very easy going, have always something to talk about or a good story to tell. I admire your dedication to the cause of your project, and that you are one of the few people actually working on image registration in the image registration group. It’s a pity we were not able to collaborate much in our PhD’s, but I sincerely hope we can do so one day in the future.

Robin, thank you for all the discussions we had on Dutch versus French habits and customs. We often mockingly made fun of each others culture, or how the Dutch ruined yet another French product or tradition, but you know that I actually admire the French language and especially the music. You really have a programmers mind, even using git in your personal live, which allows you to efficiently optimize any daily routine. I look forward to many more early riser coffee breaks and continuing our debates :).

Hakim, I have made much use of your `fastR` toolbox and frequently asked you for help, to add features, or fix the occasional bug. You introduced me to the “borrel committee” and we have organized various drinks and outings together afterwards. Hence, thank you for all your programming help, and organization of various social activities.

Marcel, thank you for your work on various infrastructure projects which I have benefited from, your happiness, but also your pragmatic attitude. We have shared many frustrations about political power struggles obstructing our projects, for which

I'm glad to have been able to discuss them with you. I look forward to continuing our pragmatic problem solving in EUCanImage!

Esther, thanks for being a cheerful person and a great example. I could always come to you to discuss the “next step” issues in academia and you were always willing to share your insights and experience.

Douwe, thanks for taking the risk of being the first PhD student co-supervised by me and for continuing my work. You nicely complement Stefan and me, and are not afraid to state your own opinion and bring in your own ideas. Which is definitely good with two supervisors who are also not holding back their own :). I'm impressed with how much you have already achieved and learned at this point. Hence, I look forward to the rest of the trajectory and what ideas and insights we will discover!

I would also like to express my thanks to the members of the EuCanImage team I have not mentioned yet: **Aad, Eline, Alexander, Ivan, and Andrea**. As I was focused on finishing my PhD, I was not able to do that much in the first year of the project. Thank you for taking charge and all your achievements in the first year, I hope I can do my part in the coming years!

Thank you to all other BIGR colleagues which I've shared many great moments with: **Marleen, Theo, Annegreet, Kasper, Hua, Arno, Kostas, Riway, Emanoel, Abdullah, Chaoping, Wyke, Gena, Henri, Adriaan, Willem, Yuan Yuan, Pierre, Dirk, Bo, Shengnan, Mohammed, Gokhan, Jiahang, Lau, Ruisheng, Tareq, Jukka, Danilo, Luisa**, all the past students of BIGR which are far too many to all name, and I'm sure I'm forgetting some people.

Laurens, Jeffrey Langerak and Mart, thanks for all your help in data management and IT related issues over the years.

I would also like to thanks the students I have (co-)supervised: **Dai Rui, Paul, Anthony, Michel, Wouter, Florian Calvet, Melissa van Gaalen, Sanne, Alice, Timo, Sybren, Theodoros, Lisa, Anna, Laura, Marit, Emma, Koen, Marijn, Mitchell, Jeffrey Visser, Myrthe, Lucy, Teun, Li Shen, Coen, Iris and Amber**. Thank you for all your hard work, which has also contributed to this thesis, the nice discussions, and all your ideas. Supervising you brought me a lot of joy and inspiration.

Thanks to the partners from the STRaTeGy consortium which I have collaborated with over the years: **Albi, Len, Petros, Zhenwei, Ivan, René, Ronald, and Elli**.

Luckily, I have been able to collaborate with a lot of different people. That does mean that I don't have enough space to thank you all in detail, and probably have forgotten some people. Hence, last but not least from my “professional” connections, I would like to thank all other people that have contributed to the work in this thesis, i.e., my co-authors.

Without friends and family to support me along the way and take my mind off my work, this thesis would never have succeeded in such a manner.

Jort, you have been a close friend to me for years. I've enjoyed many different events with you over the years, mostly organized by you, together with the friend groups we share, but also with our partners. But I could also always drop by or give you a call about more serious things, and you always showed interest in what I was doing in my PhD. Thank you for all the fun and your support, and for being one of my paranymphs!

My “Jaarclub” Icarus: **Alienor, Denise van Bavel, Jim, Maaïke, Ken, Susan, Thomas Hartog, Tim, and Yuup**. I am very grateful to have such a close group of friends to share many of the important moments in life with. The diversity in our group always leads to interesting discussions and a variety of activities we do together, which were both great to take my mind off work. I really hope we will stay together in this way for many years to come! Additionally, **Susan**, thank you for helping designing the cover of this thesis and the chapter pages. I could have never achieved this result on my own.

The others from B’jubeld: **Thomas Horstink, Bram, Bianca and Karel**. Thank you for the nice get-away weekends and holidays, the more serious discussions about work, and providing me with a lot of fun moments over the years.

The others from mAKsimaal: **Thomas Janssens, Teddy and Raymond**. Thank you for all the fun board game evenings, the nostalgic weekends in cheap bungalow houses, and laughing a lot.

My “oldest” friends: **Fengwei and Thijs**. Although we got a bit out of touch while studying, you are those types of friends for whom that doesn’t matter. We just went on where we left of. I’m very thankful to have you both, and for the many adventures over the years, such as randomly deciding to go to a concert in France and unsuspectingly meeting your future wife, or drinking beer on a private rooftop terrace at a rowing club and looking at boats floating by.

Bill and Mirthe, thank you for all the couple activities over the years. Hanging out with you is always a guarantee for a relaxing time with a lot of laughter, which I hope we will continue to do for a long time!

Esmé, it’s very nice to be able to share your experiences with someone who is not exactly in your position, but very close. Thank you for sharing the joys and challenges of doing a PhD, and the fun activities together with our partners.

My basketball team **Baros H4**, thanks for the fun and exercise. As physical exercise brings mental relaxation, playing basketball has helped me a lot, and it’s especially great to do so in a fun team.

All my aunts, uncles, and cousins and my in-laws: thank you for all the family activities. I am thankful to have families with a lot of fun activities to enjoy.

Freek, Marieke, Simone and Milan, thank you for being my family-in-law. You are so easy going, being in Kapelle often felt like a small holiday. I always enjoy the hiking and strolls we frequently do, both in Zeeland and occasionally when you came to visit us on weekends away.

Opa en Oma de Wit, Oma Starmans, dank voor al jullie support. Jullie toonden altijd interesse in mijn werk en luisterden aandachtig als ik probeerde in begrijpelijke taal uit te leggen wat ik allemaal uitvoerde. Jullie hielden het nieuws goed in de gaten of er iets technisch voorbij kwam wat mij zou interesseren, wat resulteerde in menig krantenknipsel in mijn brievenbus. Dank jullie dat ik altijd langs mocht komen om even te genieten van de rust en jullie gezelschap.

My sisters and brother, **Eline, Walter, and Katinka**: I am very thankful to have such close siblings. Over the years, I’ve come to realize more and more how many similarities we share with each other, and with our parents. From board game nights to ski vacations to Christmas dinners with tofurkey, I’ve enjoyed many activities

with you. Moreover, I could always reach out to discuss anything with you. I'm proud of all of you to follow your dreams and hope we stay this close.

Papa, thank you for always being there for me no matter what. You always put the need of your children above your own. You have helped me out on many occasions when something needed repairing or reworking in the house, maybe such that I could focus on my thesis :). I'm very proud of you for raising our family for the last twelve years as a single father, which must not have been easy. I cannot express in words how much you mean to me: thank you for making me the way I am.

Lastly, **Denise**, at the start of my PhD my girlfriend, and now my wife. I feel very lucky to have someone who complements me in all aspects. Being with you always allows me to forget everything else and simply enjoy your company. You supported me when needed, you motivated me when things were tough, and slowed me down when I was (again) trying to take on more tasks than I could handle. You helped me tremendously in separating work from my personal life in a very natural way. Without you, I seriously doubt whether I could have completed this PhD without neglecting myself and my health. Thank you for being my significant other.

About the author

Martijn Pieter Anton Starmans was born in Velsen, the Netherlands on the 22nd of December 1991. After finishing secondary school with the predicate “Cum Laude” in 2010, he started a Bachelor Applied Physics at the Delft University of Technology. He next followed a Master Applied Physics specialized in Imaging Physics and a Quantum Mechanics honors track. He conducted his thesis in collaboration with the Erasmus MC on “deformable registration in 3D breast ultrasound scans” and finished his Master in August 2016 with an internship at Philips Healthtech.



Afterwards, Martijn started his PhD at the Biomedical Imaging Group Rotterdam (BIGR), Erasmus MC, in October 2016. His PhD trajectory titled “Streamlined Quantitative Imaging Biomarker Development” resulting in the current thesis was supervised by Stefan Klein, Wiro Niessen and Jan-Jaap Visser. Following his passion to efficiently and automatically optimize routines, he developed an adaptive radiomics framework using automated machine learning. He collaborated with a large number of clinicians to develop radiomics biomarkers in a wide variety of clinical applications. As supporter of open science, he released the software for all his studies open source. He was also very active in education, including co-founding two courses in the Master Technical Medicine and (co-)supervising 28 students.

In October 2020, Martijn continued his work as a postdoctoral researcher at BIGR, Erasmus MC. He is expanding his work on radiomics, deep learning, and automated machine learning to improve the diagnostic work-up in oncology. He successfully co-applied for a grant from the Hanarth foundation on AI for the grading and phenotyping of soft-tissue tumors, is part of the EuCanImage consortium, and is co-supervising two PhD students.

Publications

Journal Papers

E. J. Bijl^{*}, **M. P. A. Starmans^{*}**, J. M. Mostert, S. Klein, F. J. P. M. Huygen, and C. C. de Vos, "Automatic quantification of complex regional pain syndrome using radiomics and deep learning based on thermography images," *In Preparation*.

J. M. Castillo T^{*}, M. Arif^{*}, **M. P. A. Starmans**, W. J. Niessen, C. H. Bangma, I. Schoots, and J. F. Veenland, "Prostate cancer classification on multi-parametric MRI: A validation study comparing deep learning and radiomics," *Accepted*.

F. Dubost, P. Yilmaz, K. van Wijnen, H. Adams, T. Evans, **M. P. A. Starmans**, G. Bortsova, M. A. Ikram, W. J. Niessen, M. W. Vernooij, and M. de Bruijne, "Visual versus automated detection of enlarged perivascular spaces and their mimics," *Submitted*.

M. P. A. Starmans^{*}, M. J. M. Timbergen^{*}, M. Vos, M. Renckens, D. J. Grünhagen, G. J. L. H. van Leenders, R. S. Dwarkasing, F. E. J. A. Willemsen, W. J. Niessen, C. Verhoef, S. Sleijfer, J. J. Visser, and S. Klein, "Differential diagnosis and molecular stratification of gastrointestinal stromal tumors on CT images using a radiomics approach," *Accepted*. arXiv: [2010.06824](https://arxiv.org/abs/2010.06824).

M. P. A. Starmans^{*}, L. S. Ho^{*}, F. Smits, N. Bije, I. de Kruijff, J. de Jong, D. Somford, E. R. Boevé, E. te Slaa, E. C. C. Cauberg, S. Klaver, A. G. van der Heijden, C. Wijburg, A. C. van de Luitgaarden, H. H. van Melick, E. Cauffman, P. de Vries, R. Jacobs, W. J. Niessen, J. J. Visser, S. Klein, J. Boormans, and A. A. M. van der Veldt, "Optimization of preoperative lymph node staging in patients with muscle-invasive bladder cancer using radiomics on computed tomography," *Under Revision*.

M. P. A. Starmans, R. L. Miclea, V. Vilgrain, M. Ronot, Y. Purcell, J. Verbeek, W. J. Niessen, J. N. Ijzermans, R. A. de Man, M. Doukas, S. Klein^{*}, and M. G. Thomeer^{*}, "Automated differentiation of malignant and benign primary solid liver lesions

on MRI: An externally validated radiomics model," *Submitted*. medrxiv: [2021.08.10.21261827](https://doi.org/10.21261827).

L. Angus*, **M. P. A. Starmans***, A. Rajicic, A. E. Odink, M. Jalving, W. J. Niessen, J. J. Visser, S. Sleijfer, S. Klein, and A. A. M. van der Veldt, "The BRAF P.V600E mutation status of melanoma lung metastases cannot be discriminated on computed tomography by LIDC criteria nor radiomics using machine learning," *Journal of Personalized Medicine*, vol. 11, no. 4, p. 257, 4 Apr. 2021. doi: [10.3390/jpm11040257](https://doi.org/10.3390/jpm11040257).

A. Blazevic*, **M. P. A. Starmans***, T. Brabander, R. S. Dwarkasing, R. A. H. van Gils, J. Hofland, G. J. H. Franssen, R. A. Feelders, W. J. Niessen, S. Klein, and W. W. de Herder, "Predicting symptomatic mesenteric mass in small intestinal neuroendocrine tumors using radiomics," *Endocrine-Related Cancer*, vol. 28, no. 8, pp. 529–539, 8 Aug. 2021. doi: [10.1530/erc-21-0064](https://doi.org/10.1530/erc-21-0064).

J. M. Castillo T, **M. P. A. Starmans**, M. Arif, W. J. Niessen, S. Klein, C. H. Bangma, I. G. Schoots, and J. F. Veenland, "A multi-center, multi-vendor study to evaluate the generalizability of a radiomics model for classifying prostate cancer: High grade vs. low grade," *Diagnostics*, vol. 11, no. 2, p. 369, 2 Feb. 2021. doi: [10.3390/diagnostics11020369](https://doi.org/10.3390/diagnostics11020369).

M. P. A. Starmans, M. J. M. Timbergen, M. Vos, G. A. Padmos, D. J. Grünhagen, C. Verhoef, S. Sleijfer, G. J. L. H. van Leenders, F. E. Buisman, F. E. J. A. Willemssen, B. G. Koerkamp, L. Angus, A. A. M. van der Veldt, A. Rajicic, A. E. Odink, M. Renckens, M. Doukas, R. A. de Man, J. N. M. IJzermans, R. L. Miclea, P. B. Vermeulen, M. G. Thomeer, J. J. Visser, W. J. Niessen, and S. Klein, "The WORC* database: MRI and CT scans, segmentations, and clinical labels for 930 patients from six radiomics studies," *Submitted*, 2021. medRxiv: [2021.08.19.21262238](https://doi.org/2021.08.19.21262238).

M. P. A. Starmans, S. R. van der Voort, T. Phil, M. J. M. Timbergen, M. Vos, G. A. Padmos, W. Kessels, D. Hanff, D. J. Grünhagen, C. Verhoef, S. Sleijfer, M. J. van den Bent, M. Smits, R. S. Dwarkasing, C. J. Els, F. Fiduzi, G. J. L. H. van Leenders, A. Blazevic, J. Hofland, T. Brabander, R. van Gils, G. J. H. Franssen, R. A. Feelders, W. W. de Herder, F. E. Buisman, F. E. J. A. Willemssen, B. Groot Koerkamp, L. Angus, A. A. M. van der Veldt, A. Rajicic, A. E. Odink, M. Deen, J. M. Castillo T, J. F. Veenland, I. Schoots, M. Renckens, M. Doukas, R. A. de Man, J. N. M. IJzermans, R. L. Miclea, P. B. Vermeulen, E. E. Bron, M. G. Thomeer, J. J. Visser, W. J. Niessen, and S. Klein, "Reproducible radiomics through automated machine learning validated on twelve clinical applications," *Submitted*, 2021. arXiv: [2108.08618](https://arxiv.org/abs/2108.08618).

M. P. A. Starmans^{*}, F. E. Buisman^{*}, M. Renckens, F. E. J. A. Willemssen, S. R. van der Voort, B. Groot Koerkamp, D. J. Grünhagen, W. J. Niessen, P. B. Vermeulen, C. Verhoef, J. J. Visser, and S. Klein, "Distinguishing pure histopathological growth patterns of colorectal liver metastases on CT using deep learning and radiomics: A pilot study," *Clinical & Experimental Metastasis*, 2021. DOI: [10.1007/s10585-021-10119-6](https://doi.org/10.1007/s10585-021-10119-6).

P. Kalendralis, Z. Shi, A. Traverso, A. Choudhury, M. Sloep, I. Zhovannik, **M. P. A. Starmans**, D. Grittner, P. Feltens, R. Monshouwer, S. Klein, R. Fijten, H. Aerts, A. Dekker, J. Soest, and L. Wee, "FAIR-compliant clinical, radiomics and DICOM metadata of RIDER, interobserver, Lung1 and head-Neck1 TCIA collections," *Medical Physics*, vol. 47, no. 11, pp. 5931–5940, 11 Nov. 2020. DOI: [10.1002/mp.14322](https://doi.org/10.1002/mp.14322).

M. J. M. Timbergen^{*}, **M. P. A. Starmans**^{*}, G. A. Padmos, D. J. Grünhagen, G. J. L. H. van Leenders, D. F. Hanff, C. Verhoef, W. J. Niessen, S. Sleijfer, S. Klein, and J. J. Visser, "Differential diagnosis and mutation stratification of desmoid-type fibromatosis on MRI using radiomics," *European Journal of Radiology*, vol. 131, p. 109 266, Oct. 2020. DOI: [10.1016/j.ejrad.2020.109266](https://doi.org/10.1016/j.ejrad.2020.109266).

P. Kalendralis, A. Traverso, Z. Shi, I. Zhovannik, R. Monshouwer, **M. P. A. Starmans**, S. Klein, E. Pfaehler, R. Boellaard, A. Dekker, and L. Wee, "Multicenter CT phantoms public dataset for radiomics reproducibility tests," *Medical Physics*, vol. 46, no. 3, pp. 1512–1518, 3 Mar. 2019. DOI: [10.1002/mp.13385](https://doi.org/10.1002/mp.13385).

S. R. van der Voort, F. Incekara, M. M. J. Wijnenga, G. Kapas, M. Gardeniers, J. W. Schouten, **M. P. A. Starmans**, R. N. Tewarie, G. J. Lycklama, P. J. French, H. J. Dubbink, M. J. van den Bent, A. J. P. E. Vincent, W. J. Niessen, S. Klein, and M. Smits, "Predicting the 1p/19q codeletion status of presumed low-grade glioma with an externally validated machine learning algorithm," *Clinical Cancer Research*, vol. 25, no. 24, pp. 7455–7462, 24 Dec. 2019. DOI: [10.1158/1078-0432.ccr-19-1127](https://doi.org/10.1158/1078-0432.ccr-19-1127).

M. Vos^{*}, **M. P. A. Starmans**^{*}, M. J. M. Timbergen, S. R. van der Voort, G. A. Padmos, W. Kessels, W. J. Niessen, G. J. L. H. van Leenders, D. J. Grünhagen, S. Sleijfer, C. Verhoef, S. Klein, and J. J. Visser, "Radiomics approach to distinguish between well differentiated liposarcomas and lipomas on MRI," *British Journal of Surgery*, vol. 106, no. 13, pp. 1800–1809, Dec. 2019. DOI: [10.1002/bjs.11410](https://doi.org/10.1002/bjs.11410).

Book Chapters

M. P. A. Starmans^{*}, S. R. van der Voort^{*}, J. M. Castillo T, J. F. Veenland, S. Klein, and W. J. Niessen, "Radiomics: Data mining using quantitative medical

image features,” in *Handbook of Medical Image Computing and Computer Assisted Intervention*, S. K. Zhou, D. Rueckert, and G. Fichtinger, Eds. Academic Press, 2020, ch. 18, pp. 429–456. DOI: [10.1016/B978-0-12-816176-0.00023-5](https://doi.org/10.1016/B978-0-12-816176-0.00023-5).

Conference Papers

K. B. de Raad[†], K. A. van Garderen, M. Smits, S. R. van der Voort, F. Incekara, E. H. G. Oei, J. Hirvasniemi, S. Klein, and **M. P. A. Starmans**, “The effect of preprocessing on convolutional neural networks for medical image segmentation,” in *International Symposium on Biomedical Imaging (ISBI 2021)*, Apr. 2021. DOI: [10.1109/ISBI48211.2021.9433952](https://doi.org/10.1109/ISBI48211.2021.9433952).

J. M. Castillo T[†], **M. P. A. Starmans**, W. J. Niessen, I. Schoots, S. Klein, and J. F. Veenland, “Classification of prostate cancer: High grade versus low grade using a radiomics approach,” in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, Institute of Electrical and Electronics Engineers (IEEE), Apr. 2019, pp. 1319–1322. DOI: [10.1109/isbi.2019.8759217](https://doi.org/10.1109/isbi.2019.8759217).

M. P. A. Starmans[†], R. L. Miclea, S. R. van der Voort, W. J. Niessen, M. G. Thomeer, and S. Klein, “Classification of malignant and benign liver tumors using a radiomics approach,” in *Medical Imaging 2018: Image Processing*, E. D. Angelini and B. A. Landman, Eds., vol. 10574, SPIE-Intl Soc Optical Eng, Mar. 2018, pp. 343–349. DOI: [10.1117/12.2293609](https://doi.org/10.1117/12.2293609).

Software

M. P. A. Starmans, *CLMRadiomics*, <https://github.com/MStarmans91/CLMRadiomics>, Zenodo, 2021. DOI: [10.5281/zenodo.4392829](https://doi.org/10.5281/zenodo.4392829).

M. P. A. Starmans, *LiverRadiomics*, <https://github.com/MStarmans91/LiverRadiomics>, Zenodo, 2021. DOI: [10.5281/zenodo.5175705](https://doi.org/10.5281/zenodo.5175705).

M. P. A. Starmans, *MelaRadiomics*, <https://github.com/MStarmans91/MelaRadiomics>, 2021. DOI: [10.5281/zenodo.4644067](https://doi.org/10.5281/zenodo.4644067).

M. P. A. Starmans, *MesentericRadiomics*, <https://github.com/MStarmans91/MesentericRadiomics>, Zenodo, 2021. DOI: [10.5281/zenodo.4916317](https://doi.org/10.5281/zenodo.4916317).

M. P. A. Starmans, *WORCDatabase*, <https://github.com/MStarmans91/WORCDatabase>, Zenodo, 2021. DOI: [10.5281/zenodo.5119040](https://doi.org/10.5281/zenodo.5119040).

M. P. A. Starmans, *DMRadiomics*, <https://github.com/MStarmans91/DMRadiomics>, Zenodo, 2020. DOI: [10.5281/zenodo.4017190](https://doi.org/10.5281/zenodo.4017190).

M. P. A. Starmans, *GISTRadiomics*, <https://github.com/MStarmans91/GISTRadiomics>, Zenodo, 2020. DOI: [10.5281/zenodo.3839322](https://doi.org/10.5281/zenodo.3839322).

M. P. A. Starmans, *LipoRadiomicsFeatures*, <https://github.com/MStarmans91/LipoRadiomicsFeatures>, Zenodo, 2019. DOI: [10.5281/zenodo.3950040](https://doi.org/10.5281/zenodo.3950040).

M. P. A. Starmans, S. R. Van der Voort, T. Phil, and S. Klein, *Workflow for optimal radiomics classification (WORC)*, <https://github.com/MStarmans91/WORC>, Zenodo, 2018. DOI: [10.5281/zenodo.3840534](https://doi.org/10.5281/zenodo.3840534).

S. R. van der Voort* and **M. P. A. Starmans***, *Predict: A radiomics extensive digital interchangeable classification toolkit (PREDICT)*, <https://github.com/Svdvoort/PREDICTFastr>, Zenodo, 2018. DOI: [10.5281/zenodo.3854839](https://doi.org/10.5281/zenodo.3854839).

Conference Abstracts

A. Blazevic[†], **M. P. A. Starmans**, T. Brabander, J. Hofland, G. J. H. Franssen, R. A. Feelders, W. J. Niessen, S. Klein, and W. W. de Herder, "Prediction of symptomatic mesenteric mass in patients with small intestinal neuroendocrine tumors using a CT radiomics approach," in *Neuroendocrinology*, Abstracts of the 17th Annual ENETS Conference for the Diagnosis and Treatment of Neuroendocrine Tumor Disease, vol. 110, 2020, pp. 1–312.

M. P. A. Starmans, S. R. van der Voort, H. C. Achterberg[†], T. Phil, M. J. M. Timbergen, M. Vos, G. A. Padmos, W. Kessels, D. Hanff, D. J. Grünhagen, C. Verhoef, S. Sleijfer, M. J. van den Bent, M. Smits, R. S. Dwarkasing, C. J. Els, F. Fiduzi, G. J. L. H. van Leenders, A. Blazevic, J. Hofland, T. Brabander, R. van Gils, G. J. H. Franssen, R. A. Feelders, W. W. de Herder, F. E. Buisman, F. E. J. A. Willemssen, B. Groot Koerkamp, L. Angus, A. A. M. van der Veldt, A. Rajjicic, A. E. Odink, M. Deen, J. M. Castillo T, J. F. Veenland, I. Schoots, M. Renckens, M. Doukas, R. A. de Man, J. N. M. Ijzermans, R. L. Miclea, P. B. Vermeulen, E. E. Bron, M. G. Thomeer, J. J. Visser, W. J. Niessen, and S. Klein, "Fully automatic construction of optimal radiomics workflows," in *Health-RI Conference*, 2020.

M. J. M. Timbergen[†], **M. P. A. Starmans**, M. Vos, M. Renckens, D. J. Grünhagen, G. J. L. H. van Leenders, W. J. Niessen, C. Verhoef, S. Sleijfer, S. Klein, and J. J. Visser, "Radiomics of gastrointestinal stromal tumors; risk classification based on computed tomography images – a pilot study," 2, vol. 46, Elsevier BV, Feb. 2020, p. e6. DOI: [10.1016/j.ejso.2019.11.011](https://doi.org/10.1016/j.ejso.2019.11.011).

P. Kalendralis[†], A. Traverso, Z. Shi, I. Zhovannik, R. Monshouwer, **M. P. A. Starmans**, S. Klein, P. Elisabeth, R. Boellaard, A. Dekker, and L. Wee, "Multicenter CT phantoms public dataset for radiomics reproducibility studies," vol. 133, Elsevier BV, Apr. 2019, p. S1030. doi: [10.1016/s0167-8140\(19\)32315-1](https://doi.org/10.1016/s0167-8140(19)32315-1).

T. Theodoridis[†], **M. P. A. Starmans**, S. Klein, and E. E. Bron, "Radiomics features for use in dementia diagnosis," in *36th Annual Scientific Meeting of the ESMRMB*, 2019.

M. J. M. Timbergen[†], **M. P. A. Starmans**, M. Vos, G. A. Padmos, D. J. Grünhagen, G. J. L. H. van Leenders, C. Verhoef, W. J. Niessen, S. Sleijfer, S. Klein, and J. J. Visser, "Mutation stratification of desmoid-type fibromatosis using a radiogenomics approach," 2, vol. 45, Elsevier BV, Feb. 2019, p. e16. doi: [10.1016/j.ejso.2018.10.084](https://doi.org/10.1016/j.ejso.2018.10.084).

M. J. M. Timbergen[†], **M. P. A. Starmans**, M. Vos, M. Renckens, D. J. Grünhagen, G. J. L. H. van Leenders, W. J. Niessen, C. Verhoef, S. Sleijfer, S. Klein, and J. J. Visser, "Radiomics of gastrointestinal stromal tumours, risk classification based on computed tomography images: A pilot study," Supplement 5, vol. 30, Elsevier BV, Oct. 2019, pp. v699–v700. doi: [10.1093/annonc/mdz283.040](https://doi.org/10.1093/annonc/mdz283.040).

A. Traverso[†], I. Zhovannik, Z. Shi, P. Kalendralis, R. Monshouwer, **M. P. A. Starmans**, S. Klein, E. Pfaehler, R. Boellaard, A. Dekker, and L. Wee, "Are quality assurance phantoms useful to assess radiomics reproducibility? a multi-center study," vol. 133, Elsevier BV, Apr. 2019, pp. S515–S516. doi: [10.1016/s0167-8140\(19\)31373-8](https://doi.org/10.1016/s0167-8140(19)31373-8).

M. Vos[†], **M. P. A. Starmans**, M. J. M. Timbergen, S. R. van der Voort, G. A. Padmos, W. Kessels, W. J. Niessen, G. J. L. H. van Leenders, D. J. Grünhagen, S. Sleijfer, C. Verhoef, S. Klein, and J. J. Visser, "Differentiating well-differentiated liposarcomas from lipomas using a radiomics approach," in *Annals of Oncology*, Abstract Book of the 44th ESMO Congress (ESMO 2019), vol. 30, Elsevier BV, Oct. 2019, p. v700. doi: [10.1093/annonc/mdz283.041](https://doi.org/10.1093/annonc/mdz283.041).

Presentations

M. P. A. Starmans[†], *Multicentre studies for more robust radiomics signatures*, Presented at the ECR 2021, Mar. 2021.

M. P. A. Starmans[†] and M. Koek[†], *Reproducible radiomics through automated machine learning validated on twelve clinical applications*, Presented at the Euro-Bioimaging User Forum 2021, Jun. 2021.

M. P. A. Starmans[†], *Multicentre studies for more robust radiomics signatures*, Presented at the ECR 2020, Jul. 2020.

M. P. A. Starmans[†], C. J. Els, F. Fiduzi, W. J. Niessen, S. Klein, and R. S. Dwarkasing, "Radiomics model to predict hepatocellular carcinoma on liver MRI of high-risk patients in surveillance: A proof-of-concept study," in *Insights into Imaging*, ECR 2020 Book of Abstracts, Presented at the ECR 2020, vol. 11, May 2020, p. 419. doi: [10.1186/s13244-020-00851-0](https://doi.org/10.1186/s13244-020-00851-0).

M. P. A. Starmans[†], M. J. M. Timbergen, G. A. Padmos, D. J. Grünhagen, G. J. L. H. van Leenders, D. F. Hanff, S. Sleijfer, J. J. Visser, and S. Klein, "Distinguishing desmoid-type fibromatosis from soft tissue sarcoma on MRI using a radiomics approach," in *Insights into Imaging*, ECR 2020 Book of Abstracts, Presented at the ECR 2020, vol. 11, May 2020, p. 236. doi: [10.1186/s13244-020-00851-0](https://doi.org/10.1186/s13244-020-00851-0).

M. P. A. Starmans[†], M. J. M. Timbergen, M. Vos, M. Renckens, D. J. Grünhagen, G. J. L. H. van Leenders, S. Sleijfer, J. J. Visser, and S. Klein, "Differential diagnosis and mutation stratification of gastrointestinal stromal tumours on CT images using a radiomics approach," in *Insights into Imaging*, ECR 2020 Book of Abstracts, Presented at the ECR 2020, vol. 11, May 2020, p. 308. doi: [10.1186/s13244-020-00851-0](https://doi.org/10.1186/s13244-020-00851-0).

M. P. A. Starmans[†], F. E. Buisman, F. Willemssen, S. R. van der Voort, D. J. Grünhagen, P. B. Vermeulen, C. Verhoef, S. Klein, and J. J. Visser, "Prediction of histopathological growth patterns by radiomics and CT-imaging in patients with operable colorectal liver metastases: A proof-of-concept study," in *Insights into Imaging*, ECR 2020 Book of Abstracts, Presented at the ECR 2020, vol. 11, May 2020, p. 419. doi: [10.1186/s13244-020-00851-0](https://doi.org/10.1186/s13244-020-00851-0).

M. P. A. Starmans[†], M. Vos, M. J. M. Timbergen, S. R. van der Voort, D. J. Grünhagen, S. Sleijfer, C. Verhoef, J. J. Visser, and S. Klein, "Distinguishing well-differentiated liposarcomas from lipomas on MR images using a radiomics approach," in *Insights into Imaging*, ECR 2020 Book of Abstracts, Presented at the ECR 2020, vol. 11, May 2020, p. 235. doi: [10.1186/s13244-020-00851-0](https://doi.org/10.1186/s13244-020-00851-0).

M. P. A. Starmans[†], S. van der Voort, R. L. Miclea, M. Vos, F. Incekara, M. J. M. Timbergen, M. M. J. Wijnenga, G. A. Padmos, W. Kessels, G. J. L. H. van Leenders, G. Kapsas, M. J. Van den Bent, A. J. P. E. Vincent, D. J. Grünhagen, C. Verhoef, S. Sleijfer, J. J. Visser, M. Smits, M. G. Thomeer, W. J. Niessen, and S. Klein, "Fully automatic construction of optimal radiomics workflows," in *7th Dutch Bio-Medical Engineering (BME) Conference*, Presented at the BME Conference 2019, 2019.

M. P. A. Starmans[†], A. Blazevic, T. Brabander, J. Hofland, W. J. Niessen, W. W. de Herder, and S. Klein, "Prediction of surgery requirement in mesenteric fibrosis on CT using a radiomics approach," in *Insights into Imaging*, ECR 2019: Book of Abstracts, Presented at the ECR 2019, vol. 10, Feb. 2019, p. S457. doi: [10.1186/s13244-019-0713-y](https://doi.org/10.1186/s13244-019-0713-y).

M. P. A. Starmans[†], R. Miclea, S. R. van der Voort, W. J. Niessen, S. Klein, and M. G. Thomeer, "Classification of malignant and benign liver tumours using a radiomics approach," in *Insights into Imaging*, ECR 2019: Book of Abstracts, Presented at the ECR 2019, vol. 10, Feb. 2019, p. S200. doi: [10.1186/s13244-019-0713-y](https://doi.org/10.1186/s13244-019-0713-y).

M. P. A. Starmans[†], S. R. van der Voort, M. Vos, F. Incekara, J. J. Visser, M. Smits, M. G. Thomeer, W. J. Niessen, and S. Klein, "Fully automatic construction of optimal radiomics workflows," in *Insights into Imaging*, ECR 2019: Book of Abstracts, Presented at the ECR 2019, vol. 10, Feb. 2019, p. S379. doi: [10.1186/s13244-019-0713-y](https://doi.org/10.1186/s13244-019-0713-y).

M. P. A. Starmans[†], S. R. van der Voort, R. L. Miclea, M. Vos, F. Incekara, M. J. M. Timbergen, M. M. J. Wijnenga, G. A. Padmos, G. J. L. H. van Leenders, G. Kapsas, M. J. van den Bent, A. J. P. E. Vincent, D. J. Grünhagen, C. Verhoef, S. Sleijfer, J. J. Visser, M. Smits, M. G. Thomeer, W. J. Niessen, and S. Klein, "Harmonizing radiomics among applications through adaptive workflow optimization," in *European Society of Medical Imaging Informatics (EuSoMII) Annual Meeting*, Presented at the EuSoMII Annual Meeting 2018, 2018.

M. P. A. Starmans[†], S. R. van der Voort, R. L. Miclea, W. J. Niessen, M. G. Thomeer, and S. Klein, *Radiomics and liver tumors*, Presented at the Current and Future Perspectives in Primary Liver Tumors Symposium 2017, Aug. 2017.

M. P. A. Starmans[†], S. R. van der Voort, W. J. Niessen, and S. Klein, *A radiomics approach for colorectal liver metastases survival prediction*, Presented at the MICCAI 2017 - CPM Colorectal Liver Metastases Challenge, Sep. 2017.

M. P. A. Starmans[†], F. Buisman, S. R. van der Voort, M. Renckens, B. Galjart, P. Nierop, W. J. Niessen, C. Verhoef, J.-J. Visser, and S. Klein, "Prediction of histopathological growth patterns in colorectal liver metastases using a radiomics approach," in *Dutch Society for Pattern Recognition (NVPBHV) Spring Meeting*, Presented at the NVPBHV Spring Meeting 2017, 2017.

* indicates equal contributions

† indicates presenting author

PhD portfolio

Courses	Year	ECTS
Advanced Pattern Recognition <i>ASCII Research School, NL</i>	2017	4.0
Computer Vision by Learning <i>ASCII Research School, NL</i>	2017	4.0
NFBIA Summer School <i>NFBIA, NL</i>	2017	3.0
English Biomedical Writing and Communication <i>Erasmus MC, NL</i>	2018	3.0
Bayesian Statistics and JASP <i>Erasmus MC, NL</i>	2018	0.3
EGSL - Academic Integrity & Responsible Research <i>Erasmus MC, NL</i>	2018	0.3
Big Data for Imaging Winter School <i>Maastricht UMC+, NL</i>	2018	3.0
WS776: Writing for Publication <i>Houston, USA</i>	2018	0.3
ESHCC - BKO (English: UTQ - University Teaching Qualification) <i>Erasmus MC, NL</i>	2018	5.0
Machine Learning Summer School (MLSS) <i>Moscow, RU</i>	2019	3.0
Total		25.9

International and local research meetings	Year	ECTS
Medical Informatics research lunch meeting (biweekly) <i>Erasmus MC, NL</i>	2016 – 2020	1.0
Biomedical imaging group seminars (biweekly) <i>Erasmus MC, NL</i>	2016 – 2021	1.0
Radiomics research meetings (monthly) <i>Erasmus MC, NL</i>	2016 – 2021	0.5
NVPHBV <i>Eindhoven, NL</i>	2017	0.3
Medical Imaging Symposium for PhD students <i>Utrecht, NL</i>	2017	0.3
Current and Future Perspectives in Primary Liver Tumors Symposium (1x invited talk) <i>Rotterdam, NL</i>	2017	1.1
Medical Image Computing and Computer Assisted Intervention (MICCAI) 2017 (1x oral workshop presentation) <i>Quebec, CA</i>	2017	2.0
European Society of Medical Imaging Informatics (EuSoMII) Yearly Meeting <i>Erasmus MC, NL</i>	2018	0.8
Bayer First Liver MRI Workshop <i>Rotterdam, NL</i>	2018	0.6
SPIE Medical Imaging 2018 (1x oral presentation) <i>Houston, USA</i>	2018	2.0
BioMedical Engineering Conference (1x oral presentation) <i>Egmond aan Zee, NL</i>	2019	1.4
European Congress of Radiology (ECR) (3x oral presentation) <i>Vienna, AT</i>	2019	3.0

European Society of Medical Imaging Informatics (EuSoMII) Yearly Meeting (1x oral presentation) <i>Erasmus MC , NL</i>	2019	0.8
European Congress of Radiology (ECR) (5x oral presentation, 1x invited talk) <i>Vienna, AT</i>	2020	4.5
European Congress of Radiology (ECR) (1x invited talk) <i>Vienna, AT</i>	2021	2.0
Total		14.8

Teaching activities	Year	ECTS
KT3352 - Medical Image Processing (responsible for $\approx \frac{1}{2}$ of the course) <i>Bachelor Technical Medicine</i>	2017 - 2021	12.0
TM11005 - Advanced Image Processing (co-founder, responsible for $\approx \frac{1}{4}$ of the course) <i>Master Technical Medicine</i>	2017 - 2021	12.0
TM10007 - Machine Learning (co-founder, responsible for $\approx \frac{1}{4}$ of the course) <i>Master Technical Medicine</i>	2019 - 2021	6.0
Total		30.0

Student (co-)supervision	Year	ECTS
<i>Missing feature in radiomics: application of the 1p/19q status in presumed low-grade glioma</i> Dai Rui, Internship MSc Electronics, ENSEEIHT, Toulouse, FR	2017	0.5
<i>Robustness of Radiomics features to segmentation variation</i> Paul De Almeida, Internship MSc Electronics, ENSEEIHT, Toulouse, FR	2017	0.5
<i>Radiomics in desmoid-type fibromatosis</i> Guillaume Padmos, Thesis MSc Medicine, Erasmus MC, Rotterdam, NL	2017	1.0
<i>Radiomics for the prediction of histopathological growth patterns of colorectal liver metastases</i> Michel Renckens, Internship MSc Medicine, Erasmus MC, Rotterdam, NL	2017	0.5
<i>Radiomics for the mutation stratification of gastrointestinal stromal tumors</i> Michel Renckens, Thesis MSc Medicine, Erasmus MC, Rotterdam, NL	2018	1.0
<i>Classification of Lipomatous Tumours</i> Wouter Kessels, Thesis BSc Applied Physics, TU Delft, Delft, NL	2018	1.0
<i>Segmentation of colorectal liver metastases on CT using convolutional neural networks and shape constraints</i> Florian Calvet, Thesis MSc Physics, Centrale Marseille, Marseille, FR	2018	3.0
<i>Objective analysis of video thermography used as diagnostic tool in patients with Complex Regional Pain Syndrome</i> Melissa van Gaalen, Internship MSc Technical Medicine, TU Delft, Delft, NL	2018	0.5
<i>Radiomics as a potential surrogate for histopathological growth patterns and as a prognostic biomarker in patients with resectable colorectal liver metastases</i> Sanne Hazen, MSc Thesis Medicine, Erasmus MC, Rotterdam, NL	2019	1.0
<i>Primary Liver Tumor Classification on MRI using Deep Learning</i> Alice Duddle, Thesis MSc Physics, ETH, Zurich, CH	2019	3.0
<i>Classification of Complex Regional Pain Syndrome based on temperature asymmetry in the lower extremity: using Infrared Thermography and a Radiomics approach</i> Timo Oosterveer, Internship MSc Technical Medicine, TU Delft, Delft, NL	2019	0.5
<i>Objective analysis of thermography images of patients with CRPS</i> Sybren van Hal, Internship MSc Technical Medicine, TU Delft, Delft, NL	2019	0.5

<i>Optimal workflows for computer-aided dementia diagnosis</i> Theodoros Theodoridis, Thesis MSc Applied Medical Sciences, VUMC, Amsterdam, NL	2019	3.0
<i>Automatic segmentation of hepatocellular carcinoma with deep learning</i> Lisa Klaassen, Internship MSc Technical Medicine, TU Delft, Delft, NL	2019	0.5
<i>Analysis of videothermography images of patients with Complex Regional Pain Syndrome</i> Anna Walstra, Laura Artz, Marit Verboom and Emma Gommers, Thesis BSc Technical Medicine, TU Delft, Delft, NL	2020	1.0
<i>The Effect of Preprocessing on Convolutional Neural Networks for Medical Image Segmentation</i> Koen de Raad, Thesis MSc Data Science and Entrepreneurship, Jheronimus Academy of Data Science, 's Hertogenbosch, NL	2020	3.0
<i>Objective analysis of vasomotor disturbances in CRPS: automatic segmentation and classification of thermography images</i> Marijn Mostert, Internship MSc Technical Medicine, TU Delft, Delft, NL	2020	0.5
<i>Automatic algorithm selection and hyperparameter optimization for medical image classification</i> Mitchell Deen, Thesis MSc Computer Science, TU Delft, Delft, NL	2020	3.0
<i>Objective analysis of vasomotor disturbances in CRPS: validation of the classification model</i> Jeffrey Visser, Internship MSc Technical Medicine, TU Delft, Delft, NL	2020	0.5
<i>Predicting the development of colorectal liver metastases based on the liver parenchyma on CT scans using radiomics</i> Myrthe van Haften, Internship MSc Technical Medicine, TU Delft, Delft, NL	2020	0.5
<i>Evaluation of the differences between thermograms of CRPS patients acquired by different cameras</i> Lucy Knops, Internship MSc Technical Medicine, TU Delft, Delft, NL	2020	0.5
<i>AUTOMONAI: Towards automatic tuning of medical image segmentation networks</i> Teun Tanis, Thesis MSc Data Science and Entrepreneurship, Jheronimus Academy of Data Science, 's Hertogenbosch, NL	2021	3.0
<i>Optimization of preoperative lymph node staging in patients with muscle-invasive bladder cancer using radiomics on CT</i> Li Shen Ho, Thesis MSc Medicine, Erasmus MC, Rotterdam, NL	2021	3.0

<i>Automated Machine Learning in Medical Image Segmentation</i> Coen van Gruijthuisen, Thesis MSc Mechanical Engineering, TU Delft, Delft, NL	2021	3.0
<i>Feature selection for an objective analysis of thermograms of patients with CRPS</i> Iris Huele, Internship MSc Technical Medicine, TU Delft, Delft, NL	2021	0.5
<i>Prediction of hypertrophic cardiomyopathy using radiomics</i> Amber Heijdra, Thesis MSc Biomedical Engineering, TU Delft, Delft, NL	2021	3.0
Total		38.0

Grants & Awards	Year
Pilot grant SURFsara (500,000 billing units)	2017 – 2019
NVidia GPU grant (together with Dr. Stefan Klein)	2018
Pilot grant SURFsara (500,000 billing units)	2019 – 2021
Colorectal liver metastases survival prediction challenge Hosted at Medical Image Computing and Computer Assisted Intervention (MICCAI) 2017. <i>1st Place</i>	2017
Employee of the Year Department of Radiology and Nuclear Medicine, Erasmus MC, NL <i>Honorable Mention</i>	2019
Hanarth Fonds (400,000 euro) Automatic grading and phenotyping of soft-tissue tumors through machine learning to guide personalized cancer treatment. <i>Co-applicant</i> (not officially mentioned as co-applicant due to formalities. Reference of formal applicant available upon request.)	2020
Open Research Award Convergence Health and Technology	2021

Committees	Year
Radiomics meeting <i>Organizing member</i>	2016 – 2020
MISP committee <i>Organizing member</i>	2018
Cross-pollination meeting <i>Organizing member</i>	2016 – 2020

Acronyms

5-HIAA 5-hydroxyindoleacetic acid.

AD Alzheimer's disease.

AdaBoost adaptive boosting.

ADASYN adaptive synthetic sampling.

ADC apparent diffusion coefficient.

ADNI Alzheimer's disease neuroimaging initiative.

AFS anterior fibromuscular stroma.

AI artificial intelligence.

AL active learning.

API application programming interface.

AUC area under the curve.

AutoML automated machine learning.

BCA balanced classification accuracy.

BCR Balanced Classification Rate.

BRAF-mt BRAF mutated.

BRAF-wt BRAF wild type.

BraTS brain tumor segmentation challenge.

BSc bachelor of science.

CA chromogranin A.

CASH combined algorithm selection and hyperparameter optimization.

CEA carcinoembryonic antigen.

CI confidence interval.

cl centiliter.

cm centimeter.

CN cognitively normal.

CNN convolutional neural network.

CoLIAGe co-occurrence of local anisotropic gradient orientations.

COM center of mass.

CRLM colorectal liver metastases.

CT computed tomography.

CTP clinical trial processor.

DCE dynamical contrast enhanced.

DDLPS dedifferentiated liposarcoma.

dHGP desmoplastic histopathological growth pattern.

DICOM digital imaging and communications in medicine.

DSC Dice similarity coefficient.

DTF desmoid-type fibromatosis.

DWI diffusion-weighted imaging.

EASL European association for the study of the liver.

EGFR epidermal growth factor receptor.

EMC Erasmus MC, university medical center, Rotterdam, the Netherlands.

ENETS European neuroendocrine tumor society.

EuCanImage European cancer imaging.

FFE fast field echo.

FISH fluorescence in situ hybridization.

FNH focal nodular hyperplasia.

FPR false positive rate.

FS fat saturation.

- GD** gadolinium contrast.
- GE** general electric.
- GIST** gastrointestinal stromal tumor.
- GLCM** gray level co-occurrence matrix.
- GLDM** gray level dependence matrix.
- GLRLM** gray level run length matrix.
- GLSZM** gray level size zone matrix.
- GPU** graphics processing unit.
- GS** gleason score.
- GTV-1** first gross tumor volume.
- GUI** graphical user interface.
- GWAS** genome wide association studies.
- HCA** hepatocellular adenoma.
- HCC** hepatocellular carcinoma.
- hf** histogram feature.
- HGP** histopathological growth pattern.
- HPF** high power fields.
- HU** Hounsfield units.
- IBSI** imaging biomarker standardization initiative.
- ICC** intra-class correlation coefficient.
- iCCA** intrahepatic cholangiocarcinoma.
- IoT** internet of things.
- IQR** interquartile range.
- IR** inversion recovery.
- KNN** k-nearest neighbors.
- KVP** kilovoltage peak.
- LAI** liver artificial intelligence.

LASSO least absolute shrinkage and selection operator.

LBP local binary pattern.

LDA linear discriminant analysis.

LIDC lung image database consortium.

LITS liver tumor segmentation.

LoG Laplacian of Gaussian.

LR logistic regression.

MCI mild cognitive impairment.

METC medische ethische toetsings commissie.

MI mitotic index.

MM mesenteric mass.

mm millimeter.

mpMRI multi-parametric magnetic resonance imaging.

MR magnetic resonance.

MRI magnetic resonance imaging.

MS multi slice.

ms milliseconds.

MSc master of science.

MSD medical segmentation decathlon.

MTB multidisciplinary tumor board.

NAS neural architecture search.

NGTDM neighbourhood grey tone difference matrix.

NLP normal liver parenchyma.

NPV negative predictive value.

NWO Netherlands organization for scientific research.

of orientation feature.

PC postcontrast.

PCA principal component analysis.

PCa prostate cancer.

PCMM Prostate cancer molecular medicine.

PDw proton density weighted.

PET positron emission tomography.

PIRADS prostate imaging reporting and data system.

PPV positive predictive value.

PSA prostatic specific antigen.

PVP portal venous phase.

PZ peripheral zone.

QDA quadratic discriminant analysis.

RADISTAT radiomic spatial textural descriptor.

RBF radial basis function.

RECIST response evaluation criteria in solid tumors.

RF random forest.

rHGP replacement histopathological growth pattern.

ROC receiver operating characteristic.

ROI region of interest.

RQS radiomics quality score.

RSNA radiological society of North America.

S45F serine 45.

SAI sarcoma artificial intelligence.

SD standard deviation.

sf shape features.

SI-NET small intestinal neuroendocrine tumor.

SM surrounding mesentery.

SMA superior mesenteric artery.

SMBO sequential model-based optimization.

SMOTE synthetic minority over-sampling technique.

SPAIR spectral attenuated inversion recovery.

SPIR spectral presaturation with inversion recovery.

std standard deviation.

STIR short τ inversion recovery.

STS soft tissue sarcoma.

SVM support vector machine.

T tesla.

T1w T1-weighted.

T2w T2-weighted.

T41A threonine 41.

TCIA the cancer imaging archive.

tf texture feature.

TIRM turbo inversion recovery magnitude.

TPOT tree based optimization tool.

TPR true positive rate.

TS transition zone.

US ultrasound.

WDLPS well differentiated liposarcoma.

WHO world health organization.

WORC workflow for optimal radiomics classification.

WT wild type.

XGBoost extreme gradient boosting.

Bibliography

- [1] L. Hood and S. H. Friend, "Predictive, personalized, preventive, participatory (P4) cancer medicine," *Nature Reviews Clinical Oncology*, vol. 8, no. 3, pp. 184–187, 3 Mar. 2011. doi: [10.1038/nrclinonc.2010.227](https://doi.org/10.1038/nrclinonc.2010.227).
- [2] I. S. Chan and G. S. Ginsburg, "Personalized medicine: Progress and promise," *Annual Review of Genomics and Human Genetics*, vol. 12, no. 1, pp. 217–244, 1 Sep. 2011. doi: [10.1146/annurev-genom-082410-101446](https://doi.org/10.1146/annurev-genom-082410-101446).
- [3] M. A. Hamburg and F. S. Collins, "The path to personalized medicine," *New England Journal of Medicine*, vol. 363, no. 4, pp. 301–304, 4 Jul. 2010. doi: [10.1056/nejmp1006304](https://doi.org/10.1056/nejmp1006304).
- [4] E. M. Cahan, T. Hernandez-Boussard, S. Thadaney-Israni, and D. L. Rubin, "Putting the data before the algorithm in big data addressing personalized healthcare," *npj Digital Medicine*, vol. 2, no. 1, p. 78, 1 Dec. 2019. doi: [10.1038/s41746-019-0157-2](https://doi.org/10.1038/s41746-019-0157-2).
- [5] N. Goossens, S. Nakagawa, X. Sun, and Y. Hoshida, "Cancer biomarker discovery and validation," *Translational cancer research*, vol. 4, no. 3, pp. 256–269, 2015, ISSN: 2218-676X 2219-6803.
- [6] H. J. W. L. Aerts, "The potential of radiomic-based phenotyping in precision medicine," *JAMA Oncology*, vol. 2, no. 12, p. 1636, 12 Dec. 2016. doi: [10.1001/jamaoncol.2016.2631](https://doi.org/10.1001/jamaoncol.2016.2631).
- [7] J. P. B. O'Connor, E. O. Aboagye, J. E. Adams, *et al.*, "Imaging biomarker roadmap for cancer studies," *Nature Reviews Clinical Oncology*, vol. 14, no. 3, pp. 169–186, 3 Mar. 2017. doi: [10.1038/nrclinonc.2016.162](https://doi.org/10.1038/nrclinonc.2016.162).
- [8] E. Eisenhauer, P. Therasse, J. Bogaerts, *et al.*, "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)," *European Journal of Cancer*, vol. 45, no. 2, pp. 228–247, 2 Jan. 2009. doi: [10.1016/j.ejca.2008.10.026](https://doi.org/10.1016/j.ejca.2008.10.026).
- [9] K. M. Elsayes, A. Z. Kielar, V. Chernyak, *et al.*, "LI-RADS: A conceptual and historical review from its beginning to its recent integration into AASLD clinical practice guidance," *Journal of Hepatocellular Carcinoma*, vol. Volume 6, pp. 49–69, Feb. 2019. doi: [10.2147/jhc.s186239](https://doi.org/10.2147/jhc.s186239).
- [10] J. C. Weinreb, J. O. Barentsz, P. L. Choyke, *et al.*, "PI-RADS prostate imaging – reporting and data system: 2015, version 2," *European Urology*, vol. 69, no. 1, pp. 16–40, 1 Jan. 2016. doi: [10.1016/j.eururo.2015.08.052](https://doi.org/10.1016/j.eururo.2015.08.052).

- [11] D. N. Louis, A. Perry, P. Wesseling, *et al.*, "The 2021 WHO classification of tumors of the central nervous system: A summary," *Neuro-Oncology*, vol. 23, no. 8, pp. 1231–1251, 8 Aug. 2021. doi: [10.1093/neuonc/noab106](https://doi.org/10.1093/neuonc/noab106).
- [12] I. D. Nagtegaal, R. D. Odze, D. Klimstra, V. Paradis, M. Rugge, P. Schirmacher, K. M. Washington, F. Carneiro, I. A. Cree, and the WHO Classification of Tumours Editorial Board, "The 2019 WHO classification of tumours of the digestive system," *Histopathology*, vol. 76, no. 2, pp. 182–188, 2 Jan. 2020. doi: [10.1111/his.13975](https://doi.org/10.1111/his.13975).
- [13] J. C. Vilanova, "WHO classification of soft tissue tumors," in *Imaging of Soft Tissue Tumors*, F. M. Vanhoenacker, P. M. Parizel, and J. L. Gielen, Eds. Springer Science and Business Media LLC, 2017, pp. 187–196. doi: [10.1007/978-3-319-46679-8_11](https://doi.org/10.1007/978-3-319-46679-8_11).
- [14] Radiological Society of North America (RSNA), *International radiology societies tackle radiologist shortage*, <https://www.rsna.org/en/news/2020/February/International-Radiology-Societies-And-Shortage>, 2020.
- [15] P. Lambin, E. Rios-Velazquez, R. Leijenaar, *et al.*, "Radiomics: Extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, vol. 48, no. 4, pp. 441–446, 4 Mar. 2012. doi: [10.1016/j.ejca.2011.11.036](https://doi.org/10.1016/j.ejca.2011.11.036).
- [16] S. S. F. Yip and H. J. W. L. Aerts, "Applications and limitations of radiomics," *Physics in Medicine and Biology*, vol. 61, no. 13, R150–R166, 13 Jul. 2016. doi: [10.1088/0031-9155/61/13/r150](https://doi.org/10.1088/0031-9155/61/13/r150).
- [17] S. Rizzo, F. Botta, S. Raimondi, D. Origgi, C. Fanciullo, A. G. Morganti, and M. Bellomi, "Radiomics: The facts and the challenges of image analysis," *European Radiology Experimental*, vol. 2, no. 1, p. 36, 1 Dec. 2018. doi: [10.1186/s41747-018-0068-z](https://doi.org/10.1186/s41747-018-0068-z).
- [18] A. Traverso, L. Wee, A. Dekker, and R. Gillies, "Repeatability and reproducibility of radiomic features: A systematic review," *International Journal of Radiation Oncology*Physics*, vol. 102, no. 4, pp. 1143–1158, 4 Nov. 2018. doi: [10.1016/j.ijrobp.2018.05.053](https://doi.org/10.1016/j.ijrobp.2018.05.053).
- [19] M. P. A. Starmans, S. R. van der Voort, J. M. Castillo T, J. F. Veenland, S. Klein, and W. J. Niessen, "Radiomics: Data mining using quantitative medical image features," in *Handbook of Medical Image Computing and Computer Assisted Intervention*, S. K. Zhou, D. Rueckert, and G. Fichtinger, Eds. Academic Press, 2020, ch. 18, pp. 429–456. doi: [10.1016/B978-0-12-816176-0.00023-5](https://doi.org/10.1016/B978-0-12-816176-0.00023-5).
- [20] M. Sollini, L. Antunovic, A. Chiti, and M. Kirienko, "Towards clinical application of image mining: A systematic review on artificial intelligence and radiomics," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 46, no. 13, pp. 2656–2672, 13 Dec. 2019. doi: [10.1007/s00259-019-04372-x](https://doi.org/10.1007/s00259-019-04372-x).
- [21] P. Afshar, A. Mohammadi, K. N. Plataniotis, A. Oikonomou, and H. Benali, "From handcrafted to deep-learning-based cancer radiomics: Challenges and opportunities," *IEEE Signal Processing Magazine*, vol. 36, no. 4, pp. 132–160, 4 Jul. 2019. doi: [10.1109/msp.2019.2900993](https://doi.org/10.1109/msp.2019.2900993).

- [22] V. S. Parekh and M. A. Jacobs, "Deep learning and radiomics in precision medicine," *Expert Review of Precision Medicine and Drug Development*, vol. 4, no. 2, pp. 59–72, 2 Mar. 2019. doi: [10.1080/23808993.2019.1585805](https://doi.org/10.1080/23808993.2019.1585805).
- [23] Z. Bodalal, S. Trebeschi, T. D. L. Nguyen-Kim, W. Schats, and R. Beets-Tan, "Radiogenomics: Bridging imaging and genomics," *Abdominal Radiology*, vol. 44, no. 6, pp. 1960–1984, 6 Jun. 2019. doi: [10.1007/s00261-019-02028-w](https://doi.org/10.1007/s00261-019-02028-w).
- [24] J. Song, Y. Yin, H. Wang, Z. Chang, Z. Liu, and L. Cui, "A review of original articles published in the emerging field of radiomics," *European Journal of Radiology*, vol. 127, p. 108991, Jun. 2020. doi: [10.1016/j.ejrad.2020.108991](https://doi.org/10.1016/j.ejrad.2020.108991).
- [25] D. A. Bluemke, L. Moy, M. A. Bredella, B. B. Ertl-Wagner, K. J. Fowler, V. J. Goh, E. F. Halpern, C. P. Hess, M. L. Schiebler, and C. R. Weiss, "Assessing radiology research on artificial intelligence: A brief guide for authors, reviewers, and readers—from the radiology editorial board," *Radiology*, vol. 294, no. 3, pp. 487–489, 2019. doi: [10.1148/radiol.2019192515](https://doi.org/10.1148/radiol.2019192515).
- [26] M. Avanzo, L. Wei, J. Stancanella, M. Vallières, A. Rao, O. Morin, S. A. Mattonen, and I. E. Naqa, "Machine and deep learning methods for radiomics," *Medical Physics*, vol. 47, no. 5, pp. e185–e202, 5 May 2020. doi: [10.1002/mp.13678](https://doi.org/10.1002/mp.13678).
- [27] J. Guiot, A. Vaidyanathan, L. Deprez, *et al.*, "A review in radiomics: Making personalized medicine a reality via routine imaging," *Medicinal Research Reviews*, vol. n/a, no. n/a, med.21846, Jul. 2021. doi: [10.1002/med.21846](https://doi.org/10.1002/med.21846).
- [28] D. P. dos Santos, M. Dietzel, and B. Baessler, "A decade of radiomics research: Are images really data or just patterns in the noise?" *European Radiology*, vol. 31, no. 1, pp. 1–4, 1 Jan. 2021. doi: [10.1007/s00330-020-07108-w](https://doi.org/10.1007/s00330-020-07108-w).
- [29] M. Hosseini, M. Powell, J. Collins, C. Callahan-Flintoft, W. Jones, H. Bowman, and B. Wyble, "I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data," *Neuroscience & Biobehavioral Reviews*, vol. 119, pp. 456–467, Dec. 2020. doi: [10.1016/j.neubiorev.2020.09.036](https://doi.org/10.1016/j.neubiorev.2020.09.036).
- [30] P. Lambin, R. T. Leijenaar, T. M. Deist, *et al.*, "Radiomics: The bridge between medical imaging and personalized medicine," *Nature Reviews Clinical Oncology*, vol. 14, no. 12, pp. 749–762, 12 Dec. 2017. doi: [10.1038/nrclinonc.2017.141](https://doi.org/10.1038/nrclinonc.2017.141).
- [31] M. L. Welch, C. McIntosh, B. Haibe-Kains, *et al.*, "Vulnerabilities of radiomic signature development: The need for safeguards," *Radiotherapy and Oncology*, vol. 130, pp. 2–9, Jan. 2019. doi: [10.1016/j.radonc.2018.10.027](https://doi.org/10.1016/j.radonc.2018.10.027).
- [32] M. Feurer and F. Hutter, "Hyperparameter optimization," in *Automated Machine Learning, The Springer Series on Challenges in Machine Learning*, F. Hutter, L. Kotthoff, and J. Vanschoren, Eds., ser. The Springer Series on Challenges in Machine Learning. Cham: Springer Science and Business Media LLC, 2019, pp. 3–33. doi: [10.1007/978-3-030-05318-5_1](https://doi.org/10.1007/978-3-030-05318-5_1).
- [33] F. Hutter, L. Kotthoff, and J. Vanschoren, Eds., *Automated Machine Learning : Methods, Systems, Challenges*. Springer Nature, 2019.

- [34] C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2013, pp. 847–855, ISBN: 1-4503-2174-7.
- [35] M. P. A. Starmans, M. J. M. Timbergen, M. Vos, *et al.*, "The WORC* database: MRI and CT scans, segmentations, and clinical labels for 930 patients from six radiomics studies," *medRxiv*, 2021.08.19.21262238, Aug. 2021. doi: [10.1101/2021.08.19.21262238](https://doi.org/10.1101/2021.08.19.21262238).
- [36] M. P. A. Starmans, S. R. van der Voort, T. Phil, and S. Klein, *Workflow for optimal radiomics classification (WORC)*, <https://github.com/MStarmans91/WORC>, Zenodo, 2018. doi: [10.5281/zenodo.3840534](https://doi.org/10.5281/zenodo.3840534).
- [37] M. P. A. Starmans, *WORCDatabase*, <https://github.com/MStarmans91/WORCDatabase>, Zenodo, 2021. doi: [10.5281/zenodo.5119040](https://doi.org/10.5281/zenodo.5119040).
- [38] V. Parekh and M. A. Jacobs, "Radiomics: A new application from established techniques," *Expert Review of Precision Medicine and Drug Development*, vol. 1, no. 2, pp. 207–226, 2 Mar. 2016. doi: [10.1080/23808993.2016.1164013](https://doi.org/10.1080/23808993.2016.1164013).
- [39] A. Zwanenburg, M. Vallières, M. Abdalah, *et al.*, "The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping," *Radiology*, vol. 295, no. 2, p. 191 145, 2020. doi: [10.1148/radiol.2020191145](https://doi.org/10.1148/radiol.2020191145).
- [40] P. M. Szczypiński, M. Strzelecki, A. Materka, and A. Klepaczko, "Mazda—a software package for image texture analysis," *Computer Methods and Programs in Biomedicine*, vol. 94, no. 1, pp. 66–76, 1 Apr. 2009. doi: [10.1016/j.cmpb.2008.08.005](https://doi.org/10.1016/j.cmpb.2008.08.005).
- [41] Y.-H. D. Fang, C.-Y. Lin, M.-J. Shih, H.-M. Wang, T.-Y. Ho, C.-T. Liao, and T.-C. Yen, *Development and evaluation of an open-source software package "CGITA" for quantifying tumor heterogeneity with molecular images*. doi: [10.1155/2014/248505](https://doi.org/10.1155/2014/248505).
- [42] A. P. Apte, A. Iyer, M. Crispin-Ortuzar, *et al.*, "Technical note: Extension of CERR for computational radiomics: A comprehensive MATLAB platform for reproducible radiomics research," *Medical Physics*, vol. 45, no. 8, pp. 3713–3720, 8 Aug. 2018. doi: [10.1002/mp.13046](https://doi.org/10.1002/mp.13046).
- [43] L. Zhang, D. V. Fried, X. J. Fave, L. A. Hunter, J. Yang, and L. E. Court, "Ibex: An open infrastructure software platform to facilitate collaborative work in radiomics," *Medical Physics*, vol. 42, no. 3, pp. 1341–1353, 3 Mar. 2015. doi: [10.1118/1.4908210](https://doi.org/10.1118/1.4908210).
- [44] J. J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, 21 Nov. 2017. doi: [10.1158/0008-5472.can-17-0339](https://doi.org/10.1158/0008-5472.can-17-0339).

- [45] S. Rathore, S. Bakas, S. Pati, *et al.*, “Brain cancer imaging phenomics toolkit (brain-CaPTk): An interactive platform for quantitative analysis of glioblastoma,” *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries Lecture Notes in Computer Science*, vol. 10670, pp. 133–145, 2018. doi: [10.1007/978-3-319-75238-9_12](https://doi.org/10.1007/978-3-319-75238-9_12).
- [46] C. Nioche, F. Orlhac, S. Boughdad, S. Reuzé, J. Goya-Outi, C. Robert, C. Pellot-Barakat, M. Soussan, F. Frouin, and I. Buvat, “LIFEx: A freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity,” *Cancer Research*, vol. 78, no. 16, pp. 4786–4789, 16 Aug. 2018. doi: [10.1158/0008-5472.can-18-0125](https://doi.org/10.1158/0008-5472.can-18-0125).
- [47] E. Pfaehler, A. Zwanenburg, J. R. de Jong, and R. Boellaard, “RaCaT: An open source and easy to use radiomics calculator tool,” *PLOS ONE*, vol. 14, no. 2, Y. Wang, Ed., 2 Feb. 2019. doi: [10.1371/journal.pone.0212223](https://doi.org/10.1371/journal.pone.0212223).
- [48] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012, ISSN: 1532-4435.
- [49] M. Feurer, A. Klein, K. Egensperger, J. T. Springenberg, M. Blum, and F. Hutter, “Auto-sklearn: Efficient and robust automated machine learning,” in *Automated Machine Learning, The Springer Series on Challenges in Machine Learning*, Springer Science and Business Media LLC, 2019, pp. 113–134. doi: [10.1007/978-3-030-05318-5_6](https://doi.org/10.1007/978-3-030-05318-5_6).
- [50] C. Zhang and Y. Ma, Eds., *Ensemble Machine Learning*. New York: Springer Science and Business Media LLC, 2012, p. 332. doi: [10.1007/978-1-4419-9326-7](https://doi.org/10.1007/978-1-4419-9326-7).
- [51] S. R. van der Voort and M. P. A. Starmans, *Predict: A radiomics extensive digital interchangeable classification toolkit (PREDICT)*, Zenodo, 2018. doi: [10.5281/zenodo.3854839](https://doi.org/10.5281/zenodo.3854839).
- [52] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 7 Jul. 2002. doi: [10.1109/tpami.2002.1017623](https://doi.org/10.1109/tpami.2002.1017623).
- [53] P. Kovesei, “Phase congruency detects corners and edges,” in *The Australian pattern recognition society conference: DICTA*, 2003.
- [54] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, “Multiscale vessel enhancement filtering,” in *Medical Image Computing and Computer-Assisted Intervention — MICCAI’98, Lecture Notes in Computer Science*, W. M. Wells, A. Colchester, and S. Delp, Eds., Springer Science and Business Media LLC, 1998, pp. 130–137. doi: [10.1007/bfb0056195](https://doi.org/10.1007/bfb0056195).
- [55] R. J. Urbanowicz, R. S. Olson, P. Schmitt, M. Meeker, and J. H. Moore, “Benchmarking relief-based feature selection methods for bioinformatics data mining,” *Journal of Biomedical Informatics*, vol. 85, pp. 168–188, Sep. 2018. doi: [10.1016/j.jbi.2018.07.015](https://doi.org/10.1016/j.jbi.2018.07.015).
- [56] V. Fonti and E. N. Belitser, “Feature selection using LASSO,” *VU Amsterdam Research Paper in Business Analytics*, 2017.

- [57] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, 2017, ISSN: 1532-4435.
- [58] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Lecture Notes in Computer Science, Advances in Intelligent Computing*, D.-S. Huang, X.-P. Zhang, and G.-B. Huang, Eds., Springer Science and Business Media LLC, 2005, pp. 878–887. DOI: [10.1007/11538059_91](https://doi.org/10.1007/11538059_91).
- [59] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Institute of Electrical and Electronics Engineers (IEEE), Jun. 2008, pp. 1322–1328. DOI: [10.1109/ijcnn.2008.4633969](https://doi.org/10.1109/ijcnn.2008.4633969).
- [60] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011, ISSN: ISSN 1533-7928.
- [61] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1 Aug. 1997. DOI: [10.1006/jcss.1997.1504](https://doi.org/10.1006/jcss.1997.1504).
- [62] T. Chen, T. He, M. Benesty, V. Khotilovich, and Y. Tang, "Xgboost: Extreme gradient boosting," *R package version 0.4-2*, pp. 1–4, 2015.
- [63] R. R. Picard and R. D. Cook, "Cross-validation of regression models," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 575–583, 387 Sep. 1984. DOI: [10.1080/01621459.1984.10478083](https://doi.org/10.1080/01621459.1984.10478083).
- [64] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Machine Learning*, vol. 52, no. 3, pp. 239–281, 2003, ISSN: 1573-0565.
- [65] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. ahead-of-print, pp. 168–192, 1 Jan. 2021. DOI: [10.1016/j.aci.2018.08.003](https://doi.org/10.1016/j.aci.2018.08.003).
- [66] B. Efron and R. Tibshirani, "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy: Rejoinder," *Statistical Science*, vol. 1, no. 1, pp. 54–75, 1 Feb. 1986. DOI: [10.1214/ss/1177013817](https://doi.org/10.1214/ss/1177013817).
- [67] S. A. Macskassy, F. Provost, and S. Rosset, "ROC confidence bands: An empirical evaluation," in *Proceedings of the 22nd international conference on Machine learning*, ACM, 2005, pp. 537–544. DOI: [10.1145/1102351.1102419](https://doi.org/10.1145/1102351.1102419).
- [68] M. P. A. Starmans, *Workflow for optimal radiomics classification (WORC) documentation*, <https://worc.readthedocs.io>, 2018.
- [69] H. C. Achterberg, M. Koek, and W. J. Niessen, "Fastr: A workflow engine for advanced data flows in medical image analysis," *Frontiers in ICT*, vol. 3, p. 15, Aug. 2016. DOI: [10.3389/fict.2016.00015](https://doi.org/10.3389/fict.2016.00015).

- [70] A. B. Yoo, M. A. Jette, and M. Grondona, "SLURM: Simple linux utility for resource management," *Job Scheduling Strategies for Parallel Processing Lecture Notes in Computer Science*, vol. 2862, pp. 44–60, 2003. doi: [10.1007/10968987_3](https://doi.org/10.1007/10968987_3).
- [71] D. S. Marcus, T. R. Olsen, M. Ramaratnam, and R. L. Buckner, "The extensible neuroimaging archive toolkit," *Neuroinformatics*, vol. 5, no. 1, pp. 11–33, 1 Mar. 2007. doi: [10.1385/ni:5:1:11](https://doi.org/10.1385/ni:5:1:11).
- [72] M. Vos, M. P. A. Starmans, M. J. M. Timbergen, *et al.*, "Radiomics approach to distinguish between well differentiated liposarcomas and lipomas on MRI," *British Journal of Surgery*, vol. 106, no. 13, pp. 1800–1809, 13 Nov. 2019. doi: [10.1002/bjs.11410](https://doi.org/10.1002/bjs.11410).
- [73] M. J. Timbergen, M. P. Starmans, G. A. Padmos, *et al.*, "Differential diagnosis and mutation stratification of desmoid-type fibromatosis on MRI using radiomics," *European Journal of Radiology*, vol. 131, p. 109 266, Oct. 2020. doi: [10.1016/j.ejrad.2020.109266](https://doi.org/10.1016/j.ejrad.2020.109266).
- [74] M. P. A. Starmans, R. L. Miclea, V. Vilgrain, *et al.*, "Automated differentiation of malignant and benign primary solid liver lesions on MRI: An externally validated radiomics model," 2021, *Submitted*. medrxiv: [2021.08.10.21261827](https://doi.org/2021.08.10.21261827).
- [75] M. P. A. Starmans, M. J. M. Timbergen, M. Vos, *et al.*, "Differential diagnosis and molecular stratification of gastrointestinal stromal tumors on CT images using a radiomics approach," 2020, *Submitted*. arXiv: [2010.06824](https://arxiv.org/abs/2010.06824).
- [76] M. P. A. Starmans, F. E. Buisman, M. Renckens, *et al.*, "Distinguishing pure histopathological growth patterns of colorectal liver metastases on CT using deep learning and radiomics: A pilot study," *Clinical & Experimental Metastasis*, Sep. 2021. doi: [10.1007/s10585-021-10119-6](https://doi.org/10.1007/s10585-021-10119-6).
- [77] L. Angus, M. P. A. Starmans, A. Rajicic, A. E. Odink, M. Jalving, W. J. Niessen, J. J. Visser, S. Sleijfer, S. Klein, and A. A. M. van der Veldt, "The *braf* P.V600E mutation status of melanoma lung metastases cannot be discriminated on computed tomography by LIDC criteria nor radiomics using machine learning," *Journal of Personalized Medicine*, vol. 11, no. 4, p. 257, 4 Apr. 2021. doi: [10.3390/jpm11040257](https://doi.org/10.3390/jpm11040257).
- [78] M. P. A. Starmans, C. J. Els, F. Fiduzi, W. J. Niessen, S. Klein, and R. S. Dwarkasing, "Radiomics model to predict hepatocellular carcinoma on liver MRI of high-risk patients in surveillance: A proof-of-concept study," in *Insights into Imaging*, ECR 2020 Book of Abstracts, Presented at the ECR 2020, vol. 11, May 2020, p. 419. doi: [10.1186/s13244-020-00851-0](https://doi.org/10.1186/s13244-020-00851-0).
- [79] A. Blazevic, M. P. A. Starmans, T. Brabander, *et al.*, "Predicting symptomatic mesenteric mass in small intestinal neuroendocrine tumors using radiomics," *Endocrine-Related Cancer*, vol. 28, no. 8, pp. 529–539, 8 Aug. 2021. doi: [10.1530/erc-21-0064](https://doi.org/10.1530/erc-21-0064).

- [80] J. M. Castillo T, M. P. A. Starmans, W. J. Niessen, I. Schoots, S. Klein, and J. F. Veenland, "Classification of prostate cancer: High grade versus low grade using a radiomics approach," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, Institute of Electrical and Electronics Engineers (IEEE), Apr. 2019, pp. 1319–1322. doi: [10.1109/isbi.2019.8759217](https://doi.org/10.1109/isbi.2019.8759217).
- [81] S. R. van der Voort, F. Incekara, M. M. Wijnenga, *et al.*, "Predicting the 1p/19q codeletion status of presumed low-grade glioma with an externally validated machine learning algorithm," *Clinical Cancer Research*, vol. 25, no. 24, pp. 7455–7462, 24 Dec. 2019. doi: [10.1158/1078-0432.ccr-19-1127](https://doi.org/10.1158/1078-0432.ccr-19-1127).
- [82] E. E. Bron, S. Klein, J. M. Papma, *et al.*, "Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of alzheimer's disease," *NeuroImage: Clinical*, vol. 31, p. 102712, 2021. doi: [10.1016/j.nicl.2021.102712](https://doi.org/10.1016/j.nicl.2021.102712).
- [83] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, no. 1, p. 4006, 1 Sep. 2014. doi: [10.1038/ncomms5006](https://doi.org/10.1038/ncomms5006).
- [84] C. R. Jack, J. Barnes, M. A. Bernstein, *et al.*, "Magnetic resonance imaging in alzheimer's disease neuroimaging initiative 2," *Alzheimer's & Dementia*, vol. 11, no. 7, pp. 740–756, 7 Jul. 2015. doi: [10.1016/j.jalz.2015.05.002](https://doi.org/10.1016/j.jalz.2015.05.002).
- [85] S. R. van der Voort, F. Incekara, M. Wijnenga, *et al.*, *Data belonging to predicting the 1p/19q co-deletion status of presumed low grade glioma with an externally validated machine learning algorithm*, 2019. doi: [10.17632/rssf5nxxby.1](https://doi.org/10.17632/rssf5nxxby.1).
- [86] E. E. Bron, R. M. Steketee, G. C. Houston, *et al.*, "Diagnostic classification of arterial spin labeling and structural MRI in presenile early stage dementia," *Human Brain Mapping*, vol. 35, no. 9, pp. 4916–4931, 9 Sep. 2014. doi: [10.1002/hbm.22522](https://doi.org/10.1002/hbm.22522).
- [87] Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, aac4716–aac4716, 6251 Aug. 2015. doi: [10.1126/science.aac4716](https://doi.org/10.1126/science.aac4716).
- [88] I. Buvat and F. Orlhac, "The dark side of radiomics: On the paramount importance of publishing negative results," *Journal of Nuclear Medicine*, vol. 60, no. 11, pp. 1543–1544, 11 Nov. 2019. doi: [10.2967/jnumed.119.235325](https://doi.org/10.2967/jnumed.119.235325).
- [89] M. Hatt, M. Vallieres, D. Visvikis, and A. Zwanenburg, "IBSI: An international community radiomics standardization initiative," *Journal of Nuclear Medicine*, vol. 59, no. supplement 1, pp. 287–287, 2018, issn: 0161-5505, 2159-662X.
- [90] R. S. Olson and J. H. Moore, *TPOT: A tree-based pipeline optimization tool for automating machine learning*, 2019. doi: [10.1007/978-3-030-05318-5_8](https://doi.org/10.1007/978-3-030-05318-5_8).
- [91] X. Su, N. Chen, H. Sun, *et al.*, "Automated machine learning based on radiomics features predicts H3 K27M mutation in midline gliomas of the brain," *Neuro-Oncology*, Sep. 2019. doi: [10.1093/neuonc/noz184](https://doi.org/10.1093/neuonc/noz184).

- [92] H. Sun, H. Qu, L. Chen, W. Wang, Y. Liao, L. Zou, Z. Zhou, X. Wang, and S. Zhou, "Identification of suspicious invasive placentation based on clinical MRI data using textural features and automated machine learning," *European Radiology*, vol. 29, no. 11, pp. 6152–6162, 11 Nov. 2019. doi: [10.1007/s00330-019-06372-9](https://doi.org/10.1007/s00330-019-06372-9).
- [93] M. Antonelli, A. Reinke, S. Bakas, *et al.*, "The medical segmentation decathlon," 2021. arXiv: [2106.05735](https://arxiv.org/abs/2106.05735).
- [94] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," 2018. arXiv: [1808.05377](https://arxiv.org/abs/1808.05377).
- [95] H. J. Escalante, Q. Yao, W.-W. Tu, N. Pillay, R. Qu, Y. Yu, and N. Houlsby, "Guest editorial: Automated machine learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 9, pp. 2887–2890, 9 Sep. 2021. doi: [10.1109/tpami.2021.3077106](https://doi.org/10.1109/tpami.2021.3077106).
- [96] Y. Mao, G. Zhong, Y. Wang, and Z. Deng, *Differentiable light-weight architecture search*, Jul. 2021. doi: [10.1109/icme51207.2021.9428132](https://doi.org/10.1109/icme51207.2021.9428132).
- [97] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2 Feb. 2021. doi: [10.1038/s41592-020-01008-z](https://doi.org/10.1038/s41592-020-01008-z).
- [98] H. Bonab and F. Can, "Less is more: A comprehensive framework for the number of components of ensemble classifiers," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2735–2745, 9 Sep. 2019. doi: [10.1109/tnnls.2018.2886341](https://doi.org/10.1109/tnnls.2018.2886341).
- [99] B. Yang, J. Zhong, J. Zhong, *et al.*, "Development and validation of a radiomics nomogram based on 18F-Fluorodeoxyglucose positron emission tomography/computed tomography and clinicopathological factors to predict the survival outcomes of patients with non-small cell lung cancer," *Frontiers in Oncology*, vol. 10, p. 1042, Jul. 2020. doi: [10.3389/fonc.2020.01042](https://doi.org/10.3389/fonc.2020.01042).
- [100] F.-H. Yu, J.-X. Wang, X.-H. Ye, J. Deng, J. Hang, and B. Yang, "Ultrasound-based radiomics nomogram: A potential biomarker to predict axillary lymph node metastasis in early-stage invasive breast cancer," *European Journal of Radiology*, vol. 119, p. 108658, Oct. 2019. doi: [10.1016/j.ejrad.2019.108658](https://doi.org/10.1016/j.ejrad.2019.108658).
- [101] M. P. A. Starmans, S. R. van der Voort, T. Phil, *et al.*, "Reproducible radiomics through automated machine learning validated on twelve clinical applications," 2021, *Submitted*. arXiv: [2108.08618](https://arxiv.org/abs/2108.08618).
- [102] M. P. A. Starmans, S. R. v. der Voort, T. Phil, *et al.*, "Reproducible radiomics through automated machine learning validated on twelve clinical applications," 2021. arXiv: [2108.08618](https://arxiv.org/abs/2108.08618).
- [103] X. Li, P. S. Morgan, J. Ashburner, J. Smith, and C. Rorden, "The first step for neuroimaging data analysis: DICOM to NIfTI conversion," *Journal of Neuroscience Methods*, vol. 264, pp. 47–56, May 2016. doi: [10.1016/j.jneumeth.2016.03.001](https://doi.org/10.1016/j.jneumeth.2016.03.001).

- [104] S. Frentzas, E. Simoneau, V. L. Bridgeman, *et al.*, "Vessel co-option mediates resistance to anti-angiogenic therapy in liver metastases," *Nature Medicine*, vol. 22, no. 11, pp. 1294–1302, 11 Nov. 2016. DOI: [10.1038/nm.4197](https://doi.org/10.1038/nm.4197).
- [105] M. P. A. Starmans, S. Klein, S. R. van der Voort, M. G. Thomeer, R. L. Miclea, and W. J. Niessen, *Classification of malignant and benign liver tumors using a radiomics approach*, E. D. Angelini and B. A. Landman, Eds., Mar. 2018. DOI: [10.1117/12.2293609](https://doi.org/10.1117/12.2293609).
- [106] V. Vilgrain, "Focal nodular hyperplasia," *European Journal of Radiology*, vol. 58, no. 2, pp. 236–245, 2 May 2006. DOI: [10.1016/j.ejrad.2005.11.043](https://doi.org/10.1016/j.ejrad.2005.11.043).
- [107] E. A. f. t. S. o. t. L. (EASL), "EASL clinical practice guidelines on the management of benign liver tumours," *Journal of Hepatology*, vol. 65, no. 1600-0641 (Electronic), pp. 386–398, 2 Aug. 2016. DOI: [10.1016/j.jhep.2016.04.001](https://doi.org/10.1016/j.jhep.2016.04.001).
- [108] P. B. Vermeulen, C. Colpaert, R. Salgado, R. Royers, H. Hellemans, E. V. den Heuvel, G. Goovaerts, L. Y. Dirix, and E. V. Marck, "Liver metastases from colorectal adenocarcinomas grow in three patterns with different angiogenesis and desmoplasia," *The Journal of Pathology*, vol. 195, no. 3, pp. 336–342, 3 Oct. 2001. DOI: [10.1002/path.966](https://doi.org/10.1002/path.966).
- [109] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2663–2674, 12 Dec. 2018. DOI: [10.1109/tmi.2018.2845918](https://doi.org/10.1109/tmi.2018.2845918).
- [110] P. Bilic, P. F. Christ, E. Vorontsov, *et al.*, "The liver tumor segmentation benchmark (LiTS)," *arxiv:1901.04056*, 2019. arXiv: [1901.04056](https://arxiv.org/abs/1901.04056).
- [111] C. Fletcher, J. Bridge, P. Hogendoorn, F. Mertens, and I. A. f. R. o. C. World-HealthOrganization, *WHO Classification of Tumours of Soft Tissue and Bone*. Lyon: IARC Press, 2013.
- [112] M. Brisson, T. Kashima, D. Delaney, R. Tirabosco, A. Clarke, S. Cro, A. M. Flanagan, and P. O'Donnell, "MRI characteristics of lipoma and atypical lipomatous tumor/well-differentiated liposarcoma: Retrospective comparison with histology and MDM2 gene amplification," *Skeletal Radiology*, vol. 42, no. 5, pp. 635–647, 5 May 2013. DOI: [10.1007/s00256-012-1517-z](https://doi.org/10.1007/s00256-012-1517-z).
- [113] M. J. Kransdorf, L. W. Bancroft, J. J. Peterson, M. D. Murphey, W. C. Foster, and H. T. Temple, "Imaging of fatty tumors: Distinction of lipoma and well-differentiated liposarcoma," *Radiology*, vol. 224, no. 1, pp. 99–104, 1 Jul. 2002. DOI: [10.1148/radiol.224101113](https://doi.org/10.1148/radiol.224101113).
- [114] P. Gupta, T. A. Potti, S. D. Wuertzer, L. Lenchik, and D. A. Pacholke, "Spectrum of fat-containing soft-tissue masses at MR imaging: The common, the uncommon, the characteristic, and the sometimes confusing," *RadioGraphics*, vol. 36, no. 3, pp. 753–766, 3 May 2016. DOI: [10.1148/rg.2016150133](https://doi.org/10.1148/rg.2016150133).
- [115] A. Drevelegas, M. Pilavaki, and D. Chourmouzi, "Lipomatous tumors of soft tissue: MR appearance with histological correlation," *European Journal of Radiology*, vol. 50, no. 3, pp. 257–267, 3 Jun. 2004. DOI: [10.1016/j.ejrad.2004.01.022](https://doi.org/10.1016/j.ejrad.2004.01.022).

- [116] P. W. O'Donnell, A. M. Griffin, W. C. Eward, A. Sternheim, L. M. White, J. S. Wunder, and P. C. Ferguson, "Can experienced observers differentiate between lipoma and well-differentiated liposarcoma using only MRI?" *Sarcoma*, vol. 2013, pp. 1–6, 2013. doi: [10.1155/2013/982784](https://doi.org/10.1155/2013/982784).
- [117] M. J. Kransdorf, J. M. Meis, and J. S. Jelinek, "Dedifferentiated liposarcoma of the extremities: Imaging findings in four patients.," *American Journal of Roentgenology*, vol. 161, no. 1, pp. 127–130, 1 Jul. 1993. doi: [10.2214/ajr.161.1.8517290](https://doi.org/10.2214/ajr.161.1.8517290).
- [118] U. Tateishi, T. Hasegawa, Y. Beppu, M. Satake, and N. Moriyama, "Primary dedifferentiated liposarcoma of the retroperitoneum. prognostic significance of computed tomography and magnetic resonance imaging features," *J Comput Assist Tomogr*, vol. 27, no. 5, pp. 799–804, 2003. doi: [10.1097/00004728-200309000-00019](https://doi.org/10.1097/00004728-200309000-00019).
- [119] J. S. Yun, H. W. Chung, J. S. Song, S. H. Lee, M. H. Lee, and M. J. Shin, "Dedifferentiated liposarcoma of the musculoskeletal system: Expanded MR imaging spectrum from predominant fatty mass to non-fatty mass," *Acta Radiologica*, vol. 60, pp. 1474–1481, 11 Nov. 2019. doi: [10.1177/0284185119833060](https://doi.org/10.1177/0284185119833060).
- [120] M. D. Murphey, L. K. Arcara, and J. Fanburg-Smith, "From the archives of the AFIP: Imaging of musculoskeletal liposarcoma with radiologic-pathologic correlation," *Radiographics*, vol. 25, no. 5, pp. 1371–95, 2005. doi: [10.1148/rg.255055106](https://doi.org/10.1148/rg.255055106).
- [121] E. S. N. W. Group, "Soft tissue and visceral sarcomas: ESMO clinical practice guidelines for diagnosis, treatment and follow-up," *Annals of Oncology*, vol. 25, pp. iii102–iii112, Sep. 2014. doi: [10.1093/annonc/mdl254](https://doi.org/10.1093/annonc/mdl254).
- [122] K. Thway, J. Wang, J. Swansbury, T. Min, and C. Fisher, "Fluorescence In Situ Hybridization for MDM2 Amplification as a routine ancillary diagnostic tool for suspected well-differentiated and dedifferentiated liposarcomas: Experience at a tertiary center," *Sarcoma*, vol. 2015, pp. 1–10, 2015. doi: [10.1155/2015/812089](https://doi.org/10.1155/2015/812089).
- [123] H. Kimura, Y. Dobashi, T. Nojima, H. Nakamura, N. Yamamoto, H. Tsuchiya, H. Ikeda, S. Sawada-Kitamura, T. Oyama, and A. Ooi, "Utility of fluorescence in situ hybridization to detect MDM2 amplification in liposarcomas and their morphological mimics," *Int J Clin Exp Pathol*, vol. 6, no. 7, pp. 1306–16, 2013, issn: 1936-2625.
- [124] J. Junttu, J. Sijbers, S. D. Backer, J. Rajan, and D. V. Dyck, "Machine learning study of several classifiers trained with texture analysis features to differentiate benign from malignant soft-tissue tumors in T1-MRI images," *Journal of Magnetic Resonance Imaging*, vol. 31, no. 3, pp. 680–689, 3 Mar. 2010. doi: [10.1002/jmri.22095](https://doi.org/10.1002/jmri.22095).
- [125] V. D. Corino, E. Montin, A. Messina, P. G. Casali, A. Gronchi, A. Marchianò, and L. T. Mainardi, "Radiomic analysis of soft tissues sarcomas can distinguish intermediate from high-grade lesions," *Journal of Magnetic Resonance Imaging*, vol. 47, no. 3, pp. 829–840, 3 Mar. 2018. doi: [10.1002/jmri.25791](https://doi.org/10.1002/jmri.25791).

- [126] M. Vallières, C. R. Freeman, S. R. Skamene, and I. E. Naqa, "A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities," *Physics in Medicine and Biology*, vol. 60, no. 14, pp. 5471–5496, 14 Jul. 2015. doi: [10.1088/0031-9155/60/14/5471](https://doi.org/10.1088/0031-9155/60/14/5471).
- [127] S. Klein, M. Staring, K. Murphy, M. Viergever, and J. Pluim, "Elastix: A toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 1 Jan. 2010. doi: [10.1109/tmi.2009.2035616](https://doi.org/10.1109/tmi.2009.2035616).
- [128] M. R. Sanchez, F. M. Golomb, J. A. Moy, and J. R. Potozkin, "Giant lipoma: Case report and review of the literature," *Journal of the American Academy of Dermatology*, vol. 28, no. 2 Pt 1, pp. 266–268, 2 Feb. 1993. doi: [10.1016/s0190-9622\(08\)81151-6](https://doi.org/10.1016/s0190-9622(08)81151-6).
- [129] C. A. Smith, S. R. Martinez, W. H. Tseng, R. M. Tamurian, R. J. Bold, D. Borys, and R. J. Canter, "Predicting survival for well-differentiated liposarcoma: The importance of tumor location," *Journal of Surgical Research*, vol. 175, no. 1, pp. 12–17, 1 Jun. 2012. doi: [10.1016/j.jss.2011.07.024](https://doi.org/10.1016/j.jss.2011.07.024).
- [130] R. E. Thornhill, M. Golfam, A. Sheikh, G. O. Cron, E. A. White, J. Werier, M. E. Schweitzer, and G. D. Primio, "Differentiation of lipoma from liposarcoma on MRI using texture and shape analysis," *Academic Radiology*, vol. 21, no. 9, pp. 1185–1194, 9 Sep. 2014. doi: [10.1016/j.acra.2014.04.005](https://doi.org/10.1016/j.acra.2014.04.005).
- [131] S. Echegaray, O. Gevaert, R. Shah, A. Kamaya, J. Louie, N. Kothary, and S. Napel, "Core samples for radiomics features that are insensitive to tumor segmentation: Method and pilot study using CT images of hepatocellular carcinoma," *Journal of Medical Imaging*, vol. 2, no. 4, p. 041011, 4 Oct. 2015. doi: [10.1117/1.jmi.2.4.041011](https://doi.org/10.1117/1.jmi.2.4.041011).
- [132] L. Lu, W. Lv, J. Jiang, J. Ma, Q. Feng, A. Rahmim, and W. Chen, "Robustness of radiomic features in [11C]Choline and [18F]FDG PET/CT imaging of nasopharyngeal carcinoma: Impact of segmentation and discretization," *Molecular Imaging and Biology*, vol. 18, no. 6, pp. 935–945, 6 Dec. 2016. doi: [10.1007/s11307-016-0973-6](https://doi.org/10.1007/s11307-016-0973-6).
- [133] C. Parmar, E. R. Velazquez, R. Leijenaar, *et al.*, "Robust radiomics feature quantification using semiautomatic volumetric segmentation," *PLoS ONE*, vol. 9, no. 7, G. E. Woloschak, Ed., p. e102107, 7 Jul. 2014. doi: [10.1371/journal.pone.0102107](https://doi.org/10.1371/journal.pone.0102107).
- [134] M. P. A. Starmans, *LipoRadiomicsFeatures*, <https://github.com/MStarmans91/LipoRadiomicsFeatures>, 2019.
- [135] D. L. M. van Broekhoven, D. J. Grünhagen, M. A. den Bakker, T. van Dalen, and C. Verhoef, "Time trends in the incidence and treatment of extra-abdominal and abdominal aggressive fibromatosis: A population-based study," *Annals of Surgical Oncology*, vol. 22, no. 9, pp. 2817–2823, 9 Sep. 2015. doi: [10.1245/s10434-015-4632-y](https://doi.org/10.1245/s10434-015-4632-y).

- [136] J. J. Reitamo, P. Häyry, E. Nykyri, and E. Saxen, "The desmoid tumor. i.: Incidence, sex-, age- and anatomical distribution in the finnish population," *American Journal of Clinical Pathology*, vol. 77, no. 6, pp. 665–673, 6 Jun. 1982. doi: [10.1093/ajcp/77.6.665](https://doi.org/10.1093/ajcp/77.6.665).
- [137] M. Braschi-Amirfarzan, A. R. Keraliya, K. M. Krajewski, S. H. Tirumani, A. B. Shinagare, J. L. Hornick, E. H. Baldini, S. George, N. H. Ramaiya, and J. P. Jagannathan, "Role of imaging in management of desmoid-type fibromatosis: A primer for radiologists," *RadioGraphics*, vol. 36, no. 3, pp. 767–782, 3 May 2016. doi: [10.1148/rg.2016150153](https://doi.org/10.1148/rg.2016150153).
- [138] E. A. Walker, M. E. Fenton, J. S. Salesky, and M. D. Murphey, "Magnetic resonance imaging of benign soft tissue neoplasms in adults," *Radiologic Clinics of North America*, vol. 49, no. 6, pp. 1197–1217, 6 Nov. 2011. doi: [10.1016/j.rcl.2011.07.007](https://doi.org/10.1016/j.rcl.2011.07.007).
- [139] T. L. Ng, A. M. Gown, T. S. Barry, M. C. U. Cheang, A. K. W. Chan, D. A. Turbin, F. D. Hsu, R. B. West, and T. O. Nielsen, "Nuclear beta-catenin in mesenchymal tumors," *Modern Pathology*, vol. 18, no. 1, pp. 68–74, 1 Jan. 2005. doi: [10.1038/modpathol.3800272](https://doi.org/10.1038/modpathol.3800272).
- [140] M. V. Enzo, P. Cattelan, M. Rastrelli, A. Tosi, C. R. Rossi, U. Hladnik, and D. Segat, "Growth rate and myofibroblast differentiation of desmoid fibroblast-like cells are modulated by TGF- β signaling," *Histochemistry and Cell Biology*, vol. 151, no. 2, pp. 145–160, 2 Feb. 2019. doi: [10.1007/s00418-018-1718-1](https://doi.org/10.1007/s00418-018-1718-1).
- [141] C. Colombo, R. Miceli, A. J. Lazar, *et al.*, "CTNNB1 45F mutation is a molecular prognosticator of increased postoperative primary desmoid tumor recurrence: An independent, multicenter validation study," *Cancer*, vol. 119, no. 20, pp. 3696–702, 2013. doi: [10.1002/cncr.28271](https://doi.org/10.1002/cncr.28271).
- [142] A. J. Lazar, D. Tuvin, S. Hajibashi, S. Habeeb, S. Bolshakov, E. Mayordomo-Aranda, C. L. Warneke, D. Lopez-Terrada, R. E. Pollock, and D. Lev, "Specific mutations in the β -catenin gene (CTNNB1) correlate with local recurrence in sporadic desmoid tumors," *The American Journal of Pathology*, vol. 173, no. 5, pp. 1518–1527, 5 Nov. 2008. doi: [10.2353/ajpath.2008.080475](https://doi.org/10.2353/ajpath.2008.080475).
- [143] D. L. M. van Broekhoven, C. Verhoef, D. J. Grünhagen, J. M. H. H. van Gorp, M. A. den Bakker, J. W. J. Hinrichs, C. M. A. de Voijs, and T. van Dalen, "Prognostic value of CTNNB1 gene mutation in primary sporadic aggressive fibromatosis," *Annals of Surgical Oncology*, vol. 22, no. 5, pp. 1464–1470, 5 May 2015. doi: [10.1245/s10434-014-4156-x](https://doi.org/10.1245/s10434-014-4156-x).
- [144] A. M. Crago, J. Chmielecki, M. Rosenberg, *et al.*, "Near universal detection of alterations in CTNNB1 and Wnt pathway regulators in desmoid-type fibromatosis by whole-exome sequencing and genomic analysis," *Genes, Chromosomes and Cancer*, vol. 54, no. 10, pp. 606–615, 10 Oct. 2015. doi: [10.1002/gcc.22272](https://doi.org/10.1002/gcc.22272).
- [145] M. J. M. Timbergen, C. Colombo, M. Renckens, *et al.*, "The prognostic role of beta-catenin mutations in desmoid-type fibromatosis undergoing resection only: A meta-analysis of individual patient data," *Ann Surg*, 2019, ISSN: 1528-1140.

- [146] B. Alman, S. Attia, C. Baumgarten, *et al.*, "The management of desmoid tumours: A joint global consensus-based guideline approach for adult and paediatric patients," *European Journal of Cancer*, vol. 127, pp. 96–107, Mar. 2020. doi: [10.1016/j.ejca.2019.11.013](https://doi.org/10.1016/j.ejca.2019.11.013).
- [147] A. M. Rutman and M. D. Kuo, "Radiogenomics: Creating a link between molecular diagnostics and diagnostic imaging," *European Journal of Radiology*, vol. 70, no. 2, pp. 232–241, 2 May 2009. doi: [10.1016/j.ejrad.2009.01.050](https://doi.org/10.1016/j.ejrad.2009.01.050).
- [148] M. A. Mazurowski, "Radiogenomics: What it is and why it is important," *Journal of the American College of Radiology*, vol. 12, no. 8, pp. 862–866, 8 Aug. 2015. doi: [10.1016/j.jacr.2015.04.019](https://doi.org/10.1016/j.jacr.2015.04.019).
- [149] H. Smith, D. Tzanis, C. Messiou, C. Benson, J. van der Hage, M. Fiore, S. Bonvalot, and A. Hayes, "The management of soft tissue tumours of the abdominal wall," *European Journal of Surgical Oncology*, vol. 43, no. 9, pp. 1647–1655, 9 Sep. 2017. doi: [10.1016/j.ejso.2017.04.009](https://doi.org/10.1016/j.ejso.2017.04.009).
- [150] K. H. Zou, S. K. Warfield, A. Bharatha, C. M. Tempany, M. R. Kaus, S. J. Haker, W. M. Wells, F. A. Jolesz, and R. Kikinis, "Statistical validation of image segmentation quality based on a spatial overlap index1," *Academic Radiology*, vol. 11, no. 2, pp. 178–189, 2 Feb. 2004. doi: [10.1016/s1076-6332\(03\)00671-8](https://doi.org/10.1016/s1076-6332(03)00671-8).
- [151] M. P. A. Starmans, S. R. van der Voort, M. Vos, F. Incekara, J. J. Visser, M. Smits, M. G. Thomeer, W. J. Niessen, and S. Klein, "Fully automatic construction of optimal radiomics workflows," in *Insights into Imaging*, ECR 2019: Book of Abstracts, Presented at the ECR 2019, vol. 10, Feb. 2019, p. S379. doi: [10.1186/s13244-019-0713-y](https://doi.org/10.1186/s13244-019-0713-y).
- [152] D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Machine Learning*, vol. 45, no. 2, pp. 171–186, 2001. doi: [10.1023/A:1010920819831](https://doi.org/10.1023/A:1010920819831).
- [153] M. P. A. Starmans, *DMRadiomics*, <https://github.com/MStarmans91/DMRadiomics>, Zenodo, 2020. doi: [10.5281/zenodo.4017190](https://doi.org/10.5281/zenodo.4017190).
- [154] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2 Jun. 2016. doi: [10.1016/j.jcm.2016.02.012](https://doi.org/10.1016/j.jcm.2016.02.012).
- [155] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, p. 837, 3 Sep. 1988. doi: [10.2307/2531595](https://doi.org/10.2307/2531595).
- [156] N. Penel, A. L. Cesne, S. Bonvalot, *et al.*, "Surgical versus non-surgical approach in primary desmoid-type fibromatosis patients: A nationwide prospective cohort from the french sarcoma group," *European Journal of Cancer*, vol. 83, pp. 125–131, Sep. 2017. doi: [10.1016/j.ejca.2017.06.017](https://doi.org/10.1016/j.ejca.2017.06.017).

- [157] P. G. Teixeira, A. Chanson, J.-L. Verhaeghe, S. Lecocq, M. Louis, G. Hossu, and A. Blum, "Correlation between tumor growth and hormonal therapy with MR signal characteristics of desmoid-type fibromatosis: A preliminary study," *Diagnostic and Interventional Imaging*, vol. 100, no. 1, pp. 47–55, 1 Jan. 2019. DOI: [10.1016/j.diii.2018.06.007](https://doi.org/10.1016/j.diii.2018.06.007).
- [158] G. Castellazzi, D. Vanel, A. L. Cesne, C. L. Pechoux, H. Caillet, F. Perona, and S. Bonvalot, "Can the MRI signal of aggressive fibromatosis be used to predict its behavior?" *European Journal of Radiology*, vol. 69, no. 2, pp. 222–229, 2 Feb. 2009. DOI: [10.1016/j.ejrad.2008.10.012](https://doi.org/10.1016/j.ejrad.2008.10.012).
- [159] P. J. Sheth, S. del Moral, B. A. Wilky, J. C. Trent, J. Cohen, A. E. Rosenberg, H. T. Temple, and T. K. Subhawong, "Desmoid fibromatosis: MRI features of response to systemic therapy," *Skeletal Radiology*, vol. 45, no. 10, pp. 1365–1373, 10 Oct. 2016. DOI: [10.1007/s00256-016-2439-y](https://doi.org/10.1007/s00256-016-2439-y).
- [160] M. R. Cassidy, R. A. Lefkowitz, N. Long, *et al.*, "Association of MRI T2 signal intensity with desmoid tumor progression during active observation: A retrospective cohort study," *Ann Surg*, vol. 271, no. 4, pp. 748–755, 2018. DOI: [10.1097/SLA.0000000000003073](https://doi.org/10.1097/SLA.0000000000003073).
- [161] N. Tuncbilek, H. M. Karakas, and O. O. Okten, "Dynamic contrast enhanced MRI in the differential diagnosis of soft tissue tumors," *European Journal of Radiology*, vol. 53, no. 3, pp. 500–505, 3 Mar. 2005. DOI: [10.1016/j.ejrad.2004.04.012](https://doi.org/10.1016/j.ejrad.2004.04.012).
- [162] K. Oka, T. Yakushiji, H. Sato, T. Fujimoto, T. Hirai, Y. Yamashita, and H. Mizuta, "Usefulness of diffusion-weighted imaging for differentiating between desmoid tumors and malignant soft tissue tumors," *Journal of Magnetic Resonance Imaging*, vol. 33, no. 1, pp. 189–193, 1 Jan. 2011. DOI: [10.1002/jmri.22406](https://doi.org/10.1002/jmri.22406).
- [163] M. Khanna, S. Ramanathan, A. S. Kambal, M. Al-Berawi, S. Yadav, D. Kumar, and N. Schieda, "Multi-parametric (mp) MRI for the diagnosis of abdominal wall desmoid tumors," *European Journal of Radiology*, vol. 92, pp. 103–110, Jul. 2017. DOI: [10.1016/j.ejrad.2017.04.010](https://doi.org/10.1016/j.ejrad.2017.04.010).
- [164] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017. DOI: [10.1016/j.media.2017.07.005](https://doi.org/10.1016/j.media.2017.07.005).
- [165] K. Søreide, O. M. Sandvik, J. A. Søreide, V. Giljaca, A. Jureckova, and V. R. Bulusu, "Global epidemiology of gastrointestinal stromal tumours (GIST): A systematic review of population-based cohort studies," *Cancer Epidemiology*, vol. 40, pp. 39–46, Feb. 2016. DOI: [10.1016/j.canep.2015.10.031](https://doi.org/10.1016/j.canep.2015.10.031).
- [166] A. J. Verschoor, The PALGA group, J. V. M. G. Bovée, L. I. H. Overbeek, P. C. W. Hogendoorn, and H. Gelderblom, "The incidence, mutational status, risk classification and referral pattern of gastro-intestinal stromal tumours in the netherlands: A nationwide pathology registry (PALGA) study," *Virchows Archiv*, vol. 472, no. 2, pp. 221–229, 2 Feb. 2018. DOI: [10.1007/s00428-017-2285-x](https://doi.org/10.1007/s00428-017-2285-x).

- [167] M. Miettinen and J. Lasota, "Gastrointestinal stromal tumors: Review on morphology, molecular pathology, prognosis, and differential diagnosis," *Arch Pathol Lab Med*, vol. 130, no. 10, pp. 1466–78, 2006, issn: 1543-2165.
- [168] S. Lau, K. Tam, C. Kam, C. Lui, C. Siu, H. Lam, and K. Mak, "Imaging of gastrointestinal stromal tumour (GIST)," *Clinical Radiology*, vol. 59, no. 6, pp. 487–498, 6 Jun. 2004. doi: [10.1016/j.crad.2003.10.018](https://doi.org/10.1016/j.crad.2003.10.018).
- [169] G. D. Demetri, M. von Mehren, C. R. Antonescu, *et al.*, "NCCN task force report: Update on the management of patients with gastrointestinal stromal tumors," *Journal of the National Comprehensive Cancer Network*, vol. 8, pp. S–1–S–41, Suppl 2 Apr. 2010. doi: [10.6004/jnccn.2010.0116](https://doi.org/10.6004/jnccn.2010.0116).
- [170] P. Rudolph, K. Gloeckner, R. Parwaresch, D. Harms, and D. Schmidt, "Immunophenotype, proliferation, DNA ploidy, and biological behavior of gastrointestinal stromal tumors: A multivariate clinicopathologic study," *Human Pathology*, vol. 29, no. 8, pp. 791–800, 8 Aug. 1998. doi: [10.1016/s0046-8177\(98\)90447-6](https://doi.org/10.1016/s0046-8177(98)90447-6).
- [171] P. A. Cassier, E. Fumagalli, P. Rutkowski, *et al.*, "Outcome of patients with platelet-derived growth factor receptor alpha-mutated gastrointestinal stromal tumors in the tyrosine kinase inhibitor era," *Clinical Cancer Research*, vol. 18, no. 16, pp. 4458–4464, 16 Aug. 2012. doi: [10.1158/1078-0432.ccr-11-3025](https://doi.org/10.1158/1078-0432.ccr-11-3025).
- [172] A. Ba-Ssalamah, D. Muin, R. Schernthaner, C. Kulinna-Cosentini, N. Bastati, J. Stift, R. Gore, and M. E. Mayerhoefer, "Texture-based classification of different gastric tumors at contrast-enhanced CT," *European Journal of Radiology*, vol. 82, no. 10, pp. e537–e543, 10 Oct. 2013. doi: [10.1016/j.ejrad.2013.06.024](https://doi.org/10.1016/j.ejrad.2013.06.024).
- [173] T. Chen, Z. Ning, L. Xu, *et al.*, "Radiomics nomogram for predicting the malignant potential of gastrointestinal stromal tumours preoperatively," *European Radiology*, vol. 29, no. 3, pp. 1074–1082, 3 Mar. 2019. doi: [10.1007/s00330-018-5629-2](https://doi.org/10.1007/s00330-018-5629-2).
- [174] C. Feng, F. Lu, Y. Shen, A. Li, H. Yu, H. Tang, Z. Li, and D. Hu, "Tumor heterogeneity in gastrointestinal stromal tumors of the small bowel: Volumetric CT texture analysis as a potential biomarker for risk stratification," *Cancer Imaging*, vol. 18, no. 1, p. 46, 1 Dec. 2018. doi: [10.1186/s40644-018-0182-4](https://doi.org/10.1186/s40644-018-0182-4).
- [175] Y. Kurata, K. Hayano, G. Ohira, K. Narushima, T. Aoyagi, and H. Matsubara, "Fractal analysis of contrast-enhanced CT images for preoperative prediction of malignant potential of gastrointestinal stromal tumor," *Abdominal Radiology*, vol. 43, no. 10, pp. 2659–2664, 10 Oct. 2018. doi: [10.1007/s00261-018-1526-z](https://doi.org/10.1007/s00261-018-1526-z).
- [176] R. Liu, H. Elhalawani, A. S. R. Mohamed, B. Elgohari, L. Court, H. Zhu, and C. D. Fuller, "Stability analysis of CT radiomic features with respect to segmentation variation in oropharyngeal cancer," *Clinical and Translational Radiation Oncology*, vol. 21, pp. 11–18, Mar. 2020. doi: [10.1016/j.ctro.2019.11.005](https://doi.org/10.1016/j.ctro.2019.11.005).

- [177] Z. Ning, J. Luo, Y. Li, S. Han, Q. Feng, Y. Xu, W. Chen, T. Chen, and Y. Zhang, "Pattern classification for gastrointestinal stromal tumors by integration of radiomics and deep convolutional features," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, pp. 1181–1191, 3 May 2019. doi: [10.1109/jbhi.2018.2841992](https://doi.org/10.1109/jbhi.2018.2841992).
- [178] F. Xu, X. Ma, Y. Wang, Y. Tian, W. Tang, M. Wang, R. Wei, and X. Zhao, "CT texture analysis can be a potential tool to differentiate gastrointestinal stromal tumors without KIT exon 11 mutation," *European Journal of Radiology*, vol. 107, pp. 90–97, Oct. 2018. doi: [10.1016/j.ejrad.2018.07.025](https://doi.org/10.1016/j.ejrad.2018.07.025).
- [179] L. Yang, D. Dong, M. Fang, Y. Zhu, Y. Zang, Z. Liu, H. Zhang, J. Ying, X. Zhao, and J. Tian, "Can CT-based radiomics signature predict KRAS/NRAS/BRAF mutations in colorectal cancer?" *European Radiology*, vol. 28, no. 5, pp. 2058–2067, 5 May 2018. doi: [10.1007/s00330-017-5146-8](https://doi.org/10.1007/s00330-017-5146-8).
- [180] C. Zhou, X. Duan, X. Zhang, H. Hu, D. Wang, and J. Shen, "Predictive features of CT for risk stratifications in patients with primary gastrointestinal stromal tumour," *European Radiology*, vol. 26, no. 9, pp. 3086–3093, 9 Sep. 2016. doi: [10.1007/s00330-015-4172-7](https://doi.org/10.1007/s00330-015-4172-7).
- [181] T. Zhuo, X. Li, and H. Zhou, "Combining radiomics and CNNs to classify benign and malignant GIST," in *Proceedings of the 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018)*, Chongqing, China: Atlantis Press, 2018. doi: [10.2991/ncce-18.2018.44](https://doi.org/10.2991/ncce-18.2018.44).
- [182] H. C. Kang, C. O. Menias, A. H. Gaballah, S. Shroff, M. W. Taggart, N. Garg, and K. M. Elsayes, "Beyond the GIST: Mesenchymal tumors of the stomach," *RadioGraphics*, vol. 33, no. 6, pp. 1673–1690, 6 Oct. 2013. doi: [10.1148/rg.336135507](https://doi.org/10.1148/rg.336135507).
- [183] M. Miettinen and J. Lasota, "Gastrointestinal stromal tumors: Pathology and prognosis at different sites," *Seminars in Diagnostic Pathology*, vol. 23, no. 2, pp. 70–83, 2 May 2006. doi: [10.1053/j.semdp.2006.09.001](https://doi.org/10.1053/j.semdp.2006.09.001).
- [184] M. P. A. Starmans, *GISTRadiomics*, <https://github.com/MStarmans91/GISTRadiomics>, Zenodo, 2021. doi: [10.5281/zenodo.3839322](https://doi.org/10.5281/zenodo.3839322).
- [185] J.-P. Fortin, D. Parker, B. Tunç, *et al.*, "Harmonization of multi-site diffusion tensor imaging data," *NeuroImage*, vol. 161, pp. 149–170, Nov. 2017. doi: [10.1016/j.neuroimage.2017.08.047](https://doi.org/10.1016/j.neuroimage.2017.08.047).
- [186] F. Orlhac, F. Frouin, C. Nioche, N. Ayache, and I. Buvat, "Validation of a method to compensate multicenter effects affecting CT radiomics," *Radiology*, vol. 291, no. 1, pp. 53–59, 1 Apr. 2019. doi: [10.1148/radiol.2019182023](https://doi.org/10.1148/radiol.2019182023).
- [187] F. J. Maldonado, S. P. Sheedy, V. R. Iyer, S. L. Hansel, D. H. Bruining, C. H. McCollough, W. S. Harmsen, J. M. Barlow, and J. G. Fletcher, "Reproducible imaging features of biologically aggressive gastrointestinal stromal tumors of the small bowel," *Abdominal Radiology*, vol. 43, pp. 1567–1574, 7 Jul. 2018. doi: [10.1007/s00261-017-1370-6](https://doi.org/10.1007/s00261-017-1370-6).

- [188] K. Akahoshi, M. Oya, T. Koga, and Y. Shiratsuchi, "Current clinical management of gastrointestinal stromal tumor," *World Journal of Gastroenterology*, vol. 24, no. 26, pp. 2806–2817, 26 Jul. 2018. DOI: [10.3748/wjg.v24.i26.2806](https://doi.org/10.3748/wjg.v24.i26.2806).
- [189] S. Liu, X. Pan, R. Liu, *et al.*, "Texture analysis of CT images in predicting malignancy risk of gastrointestinal stromal tumours," *Clinical Radiology*, vol. 73, no. 3, pp. 266–274, 3 Mar. 2018. DOI: [10.1016/j.crad.2017.09.003](https://doi.org/10.1016/j.crad.2017.09.003).
- [190] C. D. M. Fletcher, J. J. Berman, C. Corless, *et al.*, "Diagnosis of gastrointestinal stromal tumors: a consensus approach," *International Journal of Surgical Pathology*, vol. 10, no. 2, pp. 81–89, 2 Apr. 2002. DOI: [10.1177/106689690201000201](https://doi.org/10.1177/106689690201000201).
- [191] H. Joensuu, "Risk stratification of patients diagnosed with gastrointestinal stromal tumor," *Human Pathology*, vol. 39, no. 10, pp. 1411–1419, 10 Oct. 2008. DOI: [10.1016/j.humpath.2008.06.025](https://doi.org/10.1016/j.humpath.2008.06.025).
- [192] R. L. Jones, "Practical aspects of risk assessment in gastrointestinal stromal tumors," *Journal of Gastrointestinal Cancer*, vol. 45, no. 3, pp. 262–267, 3 Sep. 2014. DOI: [10.1007/s12029-014-9615-x](https://doi.org/10.1007/s12029-014-9615-x).
- [193] B. Milliron, P. K. Mittal, J. C. Camacho, A. Datir, and C. C. Moreno, "Gastrointestinal stromal tumors: Imaging features before and after treatment," *Current Problems in Diagnostic Radiology*, vol. 46, no. 1, pp. 17–25, 1 Jan. 2017. DOI: [10.1067/j.cpradiol.2015.08.001](https://doi.org/10.1067/j.cpradiol.2015.08.001).
- [194] H. Xu, L. Chen, Y. Shao, D. Zhu, X. Zhi, Q. Zhang, F. Li, J. Xu, X. Liu, and Z. Xu, "Clinical application of circulating tumor DNA in the genetic analysis of patients with advanced GIST," *Molecular Cancer Therapeutics*, vol. 17, no. 1, pp. 290–296, 1 Jan. 2018. DOI: [10.1158/1535-7163.mct-17-0436](https://doi.org/10.1158/1535-7163.mct-17-0436).
- [195] P. Kovesi, "Symmetry and asymmetry from local phase," in *Tenth Australian joint conference on artificial intelligence*, vol. 190, Citeseer, 1997, pp. 2–4.
- [196] P. Rawla, "Epidemiology of prostate cancer," *World Journal of Oncology*, vol. 10, no. 2, pp. 63–89, 2 2019. DOI: [10.14740/wjon1191](https://doi.org/10.14740/wjon1191).
- [197] N. Mottet, J. Bellmunt, E. Briers, M. Bolla, P. Cornford, and M. DeSantis, *European association of urology prostate cancer guidelines*, 2016.
- [198] H. U. Ahmed, A. E.-S. Bosaily, L. C. Brown, *et al.*, "Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): A paired validating confirmatory study," *The Lancet*, vol. 389, no. 10071, pp. 815–822, 10071 Feb. 2017. DOI: [10.1016/s0140-6736\(16\)32401-1](https://doi.org/10.1016/s0140-6736(16)32401-1).
- [199] X. Min, M. Li, D. Dong, *et al.*, "Multi-parametric MRI-based radiomics signature for discriminating between clinically significant and insignificant prostate cancer: Cross-validation of a machine learning method," *European Journal of Radiology*, vol. 115, pp. 16–21, Jun. 2019. DOI: [10.1016/j.ejrad.2019.03.010](https://doi.org/10.1016/j.ejrad.2019.03.010).
- [200] A. H. Dinh, R. Souchon, C. Melodelima, F. Bratan, F. Mege-Lechevallier, M. Colombel, and O. Rouviere, "Characterization of prostate cancer using T2 mapping at 3T: a multi-scanner study," *Diagnostic and Interventional Imaging*, vol. 96, no. 4, pp. 365–372, 4 Apr. 2015. DOI: [10.1016/j.diii.2014.11.016](https://doi.org/10.1016/j.diii.2014.11.016).

- [201] A. Chaddad, M. Kucharczyk, and T. Niazi, "Multimodal radiomic features for the predicting gleason score of prostate cancer," *Cancers*, vol. 10, no. 8, p. 249, 8 Aug. 2018. doi: [10.3390/cancers10080249](https://doi.org/10.3390/cancers10080249).
- [202] J. M. Castillo T, M. Arif, W. J. Niessen, I. G. Schoots, and J. F. Veenland, "Automated classification of significant prostate cancer on MRI: A systematic review on the performance of machine learning applications," *Cancers*, vol. 12, no. 6, p. 1606, 6 Jun. 2020. doi: [10.3390/cancers12061606](https://doi.org/10.3390/cancers12061606).
- [203] A. Stanzone, M. Gambardella, R. Cuocolo, A. Ponsiglione, V. Romeo, and M. Imbriaco, "Prostate MRI radiomics: A systematic review and radiomic quality score assessment," *European Journal of Radiology*, vol. 129, p. 109095, Aug. 2020. doi: [10.1016/j.ejrad.2020.109095](https://doi.org/10.1016/j.ejrad.2020.109095).
- [204] S. Transin, R. Souchon, C. Gonindard-Melodelima, R. de Rozario, P. Walker, M. F. de la Vega, R. Loffroy, L. Cormier, and O. Rouvière, "Computer-aided diagnosis system for characterizing ISUP grade ≥ 2 prostate cancers at multiparametric MRI: A cross-vendor evaluation," *Diagnostic and Interventional Imaging*, vol. 100, no. 12, pp. 801–811, 12 Dec. 2019. doi: [10.1016/j.diii.2019.06.012](https://doi.org/10.1016/j.diii.2019.06.012).
- [205] G. Penzias, A. Singanamalli, R. Elliott, *et al.*, "Identifying the morphologic basis for radiomic features in distinguishing different gleason grades of prostate cancer on MRI: Preliminary findings," *PLOS ONE*, vol. 13, no. 8, A. Ahmad, Ed., p. e0200730, 8 Aug. 2018. doi: [10.1371/journal.pone.0200730](https://doi.org/10.1371/journal.pone.0200730).
- [206] A. H. Dinh, C. Melodelima, R. Souchon, *et al.*, "Characterization of prostate cancer with gleason score of at least 7 by using quantitative multiparametric MR imaging: Validation of a computer-aided diagnosis system in patients referred for prostate biopsy," *Radiology*, vol. 287, no. 2, pp. 525–533, 2 May 2018. doi: [10.1148/radiol.2017171265](https://doi.org/10.1148/radiol.2017171265).
- [207] F. Orlhac, S. Boughdad, C. Philippe, H. Stalla-Bourdillon, C. Nioche, L. Champion, M. Soussan, F. Frouin, V. Frouin, and I. Buvat, "A postreconstruction harmonization method for multicenter radiomic studies in PET," *Journal of Nuclear Medicine*, vol. 59, no. 8, pp. 1321–1328, 8 Aug. 2018. doi: [10.2967/jnumed.117.199935](https://doi.org/10.2967/jnumed.117.199935).
- [208] T. A. Ozkan, A. T. Eruyar, O. O. Cebeci, O. Memik, L. Ozcan, and I. Kuskonmaz, "Interobserver variability in gleason histological grading of prostate cancer," *Scandinavian Journal of Urology*, vol. 50, no. 6, pp. 420–424, 6 Nov. 2016. doi: [10.1080/21681805.2016.1206619](https://doi.org/10.1080/21681805.2016.1206619).
- [209] B. Nilsson, L. Egevad, B. Sundelin, A. Glaessgen, H. Hamberg, and C.-G. Pihl, "Interobserver reproducibility of modified gleason score in radical prostatectomy specimens," *Virchows Archiv*, vol. -1, no. 1, pp. 1–1, 1 Jun. 2003. doi: [10.1007/s00428-004-1034-0](https://doi.org/10.1007/s00428-004-1034-0).
- [210] S. E. Viswanath, P. V. Chirra, M. C. Yim, N. M. Rofsky, A. S. Purysko, M. A. Rosen, B. N. Bloch, and A. Madabhushi, "Comparing radiomic classifiers and classifier ensembles for detection of peripheral zone prostate tumors on T2-weighted MRI: A multi-site study," *BMC Medical Imaging*, vol. 19, no. 1, p. 22, 1 Dec. 2019. doi: [10.1186/s12880-019-0308-6](https://doi.org/10.1186/s12880-019-0308-6).

- [211] Y. Artan, A. Oto, and I. S. Yetik, "Cross-device automated prostate cancer localization with multiparametric MRI," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 5385–5394, 12 Dec. 2013. doi: [10.1109/tip.2013.2285626](https://doi.org/10.1109/tip.2013.2285626).
- [212] Y. Peng, Y. Jiang, T. Antic, M. L. Giger, S. E. Eggener, and A. Oto, "Validation of quantitative analysis of multiparametric prostate MR images for prostate cancer detection and aggressiveness assessment: A cross-imager study," *Radiology*, vol. 271, no. 2, pp. 461–471, 2 May 2014. doi: [10.1148/radiol.14131320](https://doi.org/10.1148/radiol.14131320).
- [213] MeVisLab, *MeVisLab*.
- [214] J. M. Castillo T, *Josemanuel097/PCaclassificationgeneralizability*, https://github.com/josemanuel097/PCa_classification_generalizability, 2021.
- [215] K. Buch, H. Kuno, M. M. Qureshi, B. Li, and O. Sakai, "Quantitative variations in texture analysis features dependent on MRI scanning parameters: A phantom model," *Journal of Applied Clinical Medical Physics*, vol. 19, no. 6, pp. 253–264, 6 Nov. 2018. doi: [10.1002/acm2.12482](https://doi.org/10.1002/acm2.12482).
- [216] M. Schwier, J. van Griethuysen, M. G. Vangel, S. Pieper, S. Peled, C. Tempany, H. J. W. L. Aerts, R. Kikinis, F. M. Fennessy, and A. Fedorov, "Repeatability of multiparametric prostate MRI radiomics features," *Scientific Reports*, vol. 9, no. 1, p. 9441, 1 Dec. 2019. doi: [10.1038/s41598-019-45766-z](https://doi.org/10.1038/s41598-019-45766-z).
- [217] L. Rundo, C. Militello, G. Russo, A. Garufi, S. Vitabile, M. Gilardi, and G. Mauri, "Automated prostate gland segmentation based on an unsupervised fuzzy c-means clustering technique using multispectral T1w and T2w MR imaging," *Information*, vol. 8, no. 2, p. 49, 2 Jun. 2017. doi: [10.3390/info8020049](https://doi.org/10.3390/info8020049).
- [218] M. Arif, I. G. Schoots, J. M. Castillo T, C. H. Bangma, G. P. Krestin, M. J. Roobol, W. Niessen, and J. F. Veenland, "Clinically significant prostate cancer detection and segmentation in low-risk patients using a convolutional neural network on multi-parametric MRI," *European Radiology*, vol. 30, no. 12, pp. 6582–6592, 12 Dec. 2020. doi: [10.1007/s00330-020-07008-z](https://doi.org/10.1007/s00330-020-07008-z).
- [219] A. H. Dinh, C. Melodelima, R. Souchon, J. Lehaire, F. Bratan, F. Mege-Lechevallier, A. Ruffion, S. Crouzet, M. Colombel, and O. Rouviere, "Quantitative analysis of prostate multiparametric MR images for detection of aggressive prostate cancer in the peripheral zone: A multiple imager study," *Radiology*, vol. 280, no. 1, pp. 117–127, 1 Jul. 2016. doi: [10.1148/radiol.2016151406](https://doi.org/10.1148/radiol.2016151406).
- [220] R. MacKie, A. Hauschild, and A. Eggermont, "Epidemiology of invasive cutaneous melanoma," *Annals of Oncology*, vol. 20, pp. vi1–vi7, Aug. 2009. doi: [10.1093/annonc/mdp252](https://doi.org/10.1093/annonc/mdp252).
- [221] A. H. Shain and B. C. Bastian, "Author correction: From melanocytes to melanomas," *Nature Reviews Cancer*, vol. 20, no. 6, pp. 355–355, 6 Jun. 2020. doi: [10.1038/s41568-020-0269-7](https://doi.org/10.1038/s41568-020-0269-7).

- [222] J. Ferlay, M. Colombet, I. Soerjomataram, T. Dyba, G. Randi, M. Bettio, A. Gavin, O. Visser, and F. Bray, "Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018," *European Journal of Cancer*, vol. 103, pp. 356–387, Nov. 2018. doi: [10.1016/j.ejca.2018.07.005](https://doi.org/10.1016/j.ejca.2018.07.005).
- [223] D. C. Whiteman, A. C. Green, and C. M. Olsen, "The growing burden of invasive melanoma: Projections of incidence rates and numbers of new cases in six susceptible populations through 2031," *Journal of Investigative Dermatology*, vol. 136, no. 6, pp. 1161–1171, 6 Jun. 2016. doi: [10.1016/j.jid.2016.01.035](https://doi.org/10.1016/j.jid.2016.01.035).
- [224] J. J. Luke, K. T. Flaherty, A. Ribas, and G. V. Long, "Targeted agents and immunotherapies: Optimizing outcomes in melanoma," *Nature Reviews Clinical Oncology*, vol. 14, no. 8, pp. 463–482, 8 Aug. 2017. doi: [10.1038/nrclinonc.2017.43](https://doi.org/10.1038/nrclinonc.2017.43).
- [225] H. Davies, G. R. Bignell, C. Cox, *et al.*, "Mutations of the BRAF gene in human cancer," *Nature*, vol. 417, no. 6892, pp. 949–54, 2002, ISSN: 0028-0836.
- [226] J. A. Curtin, J. Fridlyand, T. Kageshita, *et al.*, "Distinct sets of genetic alterations in melanoma," *New England Journal of Medicine*, vol. 353, no. 20, pp. 2135–2147, 20 Nov. 2005. doi: [10.1056/nejmoa050092](https://doi.org/10.1056/nejmoa050092).
- [227] M. Colombino, M. Capone, A. Lissia, *et al.*, "BRAF/NRAS mutation frequencies among primary tumors and metastases in patients with melanoma," *Journal of Clinical Oncology*, vol. 30, no. 20, pp. 2522–2529, 20 Jul. 2012. doi: [10.1200/jco.2011.41.2452](https://doi.org/10.1200/jco.2011.41.2452).
- [228] K. T. Flaherty, I. Puzanov, K. B. Kim, *et al.*, "Inhibition of mutated, activated BRAF in metastatic melanoma," *New England Journal of Medicine*, vol. 363, no. 9, pp. 809–819, 9 Aug. 2010. doi: [10.1056/nejmoa1002011](https://doi.org/10.1056/nejmoa1002011).
- [229] G. V. Long, D. Stroyakovskiy, H. Gogas, *et al.*, "Combined BRAF and MEK inhibition versus BRAF inhibition alone in melanoma," *New England Journal of Medicine*, vol. 371, no. 20, pp. 1877–1888, 20 Nov. 2014. doi: [10.1056/nejmoa1406037](https://doi.org/10.1056/nejmoa1406037).
- [230] G. Long, K. Flaherty, D. Stroyakovskiy, *et al.*, "Dabrafenib plus trametinib versus dabrafenib monotherapy in patients with metastatic BRAF V600E/K-mutant melanoma: Long-term survival and safety analysis of a phase 3 study," *Annals of Oncology*, vol. 30, no. 7, p. 1848, 11 Nov. 2019. doi: [10.1093/annonc/mdz221](https://doi.org/10.1093/annonc/mdz221).
- [231] J. Larkin, P. A. Ascierto, B. Dréno, *et al.*, "Combined vemurafenib and cobimetinib in BRAF-Mutated melanoma," *New England Journal of Medicine*, vol. 371, no. 20, pp. 1867–1876, 20 Nov. 2014. doi: [10.1056/nejmoa1408868](https://doi.org/10.1056/nejmoa1408868).
- [232] P. A. Ascierto, G. A. McArthur, B. Dréno, *et al.*, "Cobimetinib combined with vemurafenib in advanced BRAFV600-mutant melanoma (coBRIM): Updated efficacy results from a randomised, double-blind, phase 3 trial," *The Lancet Oncology*, vol. 17, no. 9, pp. 1248–1260, 9 Sep. 2016. doi: [10.1016/s1470-2045\(16\)30122-x](https://doi.org/10.1016/s1470-2045(16)30122-x).

- [233] O. Michielin, A. van Akkooi, P. Ascierto, R. Dummer, and U. Keilholz, "Cutaneous melanoma: ESMO clinical practice guidelines for diagnosis, treatment and follow-up," *Annals of Oncology*, vol. 30, no. 12, pp. 1884–1901, 12 Dec. 2019. doi: [10.1093/annonc/mdz411](https://doi.org/10.1093/annonc/mdz411).
- [234] E. R. Velazquez, C. Parmar, Y. Liu, *et al.*, "Somatic mutations drive distinct imaging phenotypes in lung cancer," *Cancer Research*, vol. 77, no. 14, pp. 3922–3930, 14 Jul. 2017. doi: [10.1158/0008-5472.can-17-0122](https://doi.org/10.1158/0008-5472.can-17-0122).
- [235] O. Gevaert, J. Xu, C. D. Hoang, A. N. Leung, Y. Xu, A. Quon, D. L. Rubin, S. Napel, and S. K. Plevritis, "Non-small cell lung cancer: Identifying prognostic imaging biomarkers by leveraging public gene expression microarray data—methods and preliminary results," *Radiology*, vol. 264, no. 2, pp. 387–396, 2 Aug. 2012. doi: [10.1148/radiol.12111607](https://doi.org/10.1148/radiol.12111607).
- [236] Y. Huang, Z. Liu, L. He, X. Chen, D. Pan, Z. Ma, C. Liang, J. Tian, and C. Liang, "Radiomics signature: A potential biomarker for the prediction of disease-free survival in early-stage (i or II) non-small cell lung cancer," *Radiology*, vol. 281, no. 3, pp. 947–957, 3 Dec. 2016. doi: [10.1148/radiol.2016152234](https://doi.org/10.1148/radiol.2016152234).
- [237] V. S. Nair, O. Gevaert, G. Davidzon, S. Napel, E. E. Graves, C. D. Hoang, J. B. Shrager, A. Quon, D. L. Rubin, and S. K. Plevritis, "Prognostic PET 18F-FDG uptake imaging features are associated with major oncogenomic alterations in patients with resected non-small cell lung cancer," *Cancer Research*, vol. 72, no. 15, pp. 3725–3734, 15 Aug. 2012. doi: [10.1158/0008-5472.can-11-3943](https://doi.org/10.1158/0008-5472.can-11-3943).
- [238] C. Parmar, R. T. H. Leijenaar, P. Grossmann, E. R. Velazquez, J. Bussink, D. Rietveld, M. M. Rietbergen, B. Haibe-Kains, P. Lambin, and H. J. Aerts, "Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer," *Scientific Reports*, vol. 5, p. 11044, 1 Sep. 2015. doi: [10.1038/srep11044](https://doi.org/10.1038/srep11044).
- [239] M. Scrivener, E. E. C. de Jong, J. E. van Timmeren, T. Pieters, B. Ghaye, and X. Geets, "Radiomics applied to lung cancer: A review," *Translational Cancer Research*, vol. 5, no. 4, pp. 398–409, 4 Aug. 2016. doi: [10.21037/tcr.2016.06.18](https://doi.org/10.21037/tcr.2016.06.18).
- [240] J. J. E. van Timmeren, "Longitudinal radiomics for prognosis in non-small cell lung cancer," PhD Thesis. doi: [10.26481/dis.20190705jt](https://doi.org/10.26481/dis.20190705jt).
- [241] S. Yamamoto, R. L. Korn, R. Oklu, C. Migdal, M. B. Gotway, G. J. Weiss, A. J. Iafrate, D.-W. Kim, and M. D. Kuo, "ALKMolecular phenotype in non-small cell lung cancer: CT radiogenomic characterization," *Radiology*, vol. 272, no. 2, pp. 568–576, 2 Aug. 2014. doi: [10.1148/radiol.14140789](https://doi.org/10.1148/radiol.14140789).
- [242] S. Trebeschi, S. Drago, N. Birkbak, *et al.*, "Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers," *Annals of Oncology*, vol. 30, pp. 998–1004, 6 Jun. 2019. doi: [10.1093/annonc/mdz108](https://doi.org/10.1093/annonc/mdz108).
- [243] C. Menzer, A. M. Menzies, M. S. Carlino, *et al.*, "Targeted therapy in advanced melanoma with rare BRAF mutations," *Journal of Clinical Oncology*, vol. 37, no. 33, pp. 3142–3151, 33 Nov. 2019. doi: [10.1200/jco.19.00489](https://doi.org/10.1200/jco.19.00489).

- [244] M. P. A. Starmans, *MelaRadiomics*, <https://github.com/MStarmans91/MelaRadiomics>, Zenodo, 2020. doi: [10.5281/zenodo.4644067](https://doi.org/10.5281/zenodo.4644067).
- [245] P. Opulencia, D. S. Channin, D. S. Raicu, and J. D. Furst, "Mapping LIDC, RadLex™, and lung nodule image features," *Journal of Digital Imaging*, vol. 24, no. 2, pp. 256–270, 2 Apr. 2011. doi: [10.1007/s10278-010-9285-6](https://doi.org/10.1007/s10278-010-9285-6).
- [246] C. Robert, A. Ribas, J. Schachter, *et al.*, "Pembrolizumab versus ipilimumab in advanced melanoma (KEYNOTE-006): post-hoc 5-year results from an open-label, multicentre, randomised, controlled, phase 3 study," *The Lancet Oncology*, vol. 20, no. 9, pp. 1239–1251, 9 Sep. 2019. doi: [10.1016/s1470-2045\(19\)30388-2](https://doi.org/10.1016/s1470-2045(19)30388-2).
- [247] C. Durot, S. Mulé, P. Soyer, A. Marchal, F. Grange, and C. Hoeffel, "Metastatic melanoma: Pretreatment contrast-enhanced CT texture parameters as predictive biomarkers of survival in patients treated with pembrolizumab," *European Radiology*, vol. 29, no. 6, pp. 3183–3191, 6 Jun. 2019. doi: [10.1007/s00330-018-5933-x](https://doi.org/10.1007/s00330-018-5933-x).
- [248] R. Sun, E. J. Limkin, M. Vakalopoulou, *et al.*, "A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: An imaging biomarker, retrospective multicohort study," *The Lancet Oncology*, vol. 19, no. 9, pp. 1180–1191, 9 Sep. 2018. doi: [10.1016/s1470-2045\(18\)30413-3](https://doi.org/10.1016/s1470-2045(18)30413-3).
- [249] H. Saadani, B. van der Hiel, E. A. Aalbersberg, I. Zavrakidis, J. B. Haanen, O. S. Hoekstra, R. Boellaard, and M. P. Stokkel, "Metabolic biomarker-based BRAFV600 mutation association and prediction in melanoma," *Journal of Nuclear Medicine*, vol. 60, no. 11, pp. 1545–1552, 11 Nov. 2019. doi: [10.2967/jnumed.119.228312](https://doi.org/10.2967/jnumed.119.228312).
- [250] M. P. A. Starmans, F. E. Buisman, F. Willemssen, S. R. van der Voort, D. J. Grünhagen, P. B. Vermeulen, C. Verhoef, S. Klein, and J. J. Visser, "Prediction of histopathological growth patterns by radiomics and CT-imaging in patients with operable colorectal liver metastases: A proof-of-concept study," in *Insights into Imaging*, ECR 2020 Book of Abstracts, Presented at the ECR 2020, vol. 11, May 2020, p. 419. doi: [10.1186/s13244-020-00851-0](https://doi.org/10.1186/s13244-020-00851-0).
- [251] J. Chen, H. Cheung, L. Milot, and A. L. Martel, "AMINN: Autoencoder-based multiple instance neural network for outcome prediction of multifocal liver metastases," *arXiv:2012.06875*, 2020. arXiv: [2012.06875](https://arxiv.org/abs/2012.06875).
- [252] A. Manca, Melanoma Unit of Sassari (MUS), P. Paliogiannis, *et al.*, "Mutational concordance between primary and metastatic melanoma: A next-generation sequencing approach," *Journal of Translational Medicine*, vol. 17, no. 1, p. 289, 1 Dec. 2019. doi: [10.1186/s12967-019-2039-4](https://doi.org/10.1186/s12967-019-2039-4).
- [253] E. Riveiro-Falkenbach, C. A. Villanueva, M. C. Garrido, Y. Ruano, R. M. García-Martín, E. Godoy, P. L. Ortiz-Romero, J. J. Ríos-Martín, A. Santos-Briz, and J. L. Rodríguez-Peralto, "Intra- and inter-tumoral homogeneity of BRAF V600E mutations in melanoma tumors," *Journal of Investigative Dermatology*, vol. 135, no. 12, pp. 3078–3085, 12 Dec. 2015. doi: [10.1038/jid.2015.229](https://doi.org/10.1038/jid.2015.229).

- [254] A. Valachis and G. J. Ullenhag, "Discrepancy in BRAF status among patients with metastatic malignant melanoma: A meta-analysis," *European Journal of Cancer*, vol. 81, pp. 106–115, Aug. 2017. doi: [10.1016/j.ejca.2017.05.015](https://doi.org/10.1016/j.ejca.2017.05.015).
- [255] G. Rossi, E. Barabino, A. Fedeli, *et al.*, "Radiomic detection of EGFR mutations in NSCLC," *Cancer Research*, vol. 81, no. 3, pp. 724–731, 3 Feb. 2021. doi: [10.1158/0008-5472.can-20-0999](https://doi.org/10.1158/0008-5472.can-20-0999).
- [256] L. Wu, J. Fu, L. Wan, J. Pan, S. Lai, J. Zhong, D. C. Chung, and L. Wang, "Survival outcomes and surgical intervention of small intestinal neuroendocrine tumors: A population based retrospective study," *Oncotarget*, vol. 8, no. 3, pp. 4935–4947, 3 Jan. 2017. doi: [10.18632/oncotarget.13632](https://doi.org/10.18632/oncotarget.13632).
- [257] M. Pavel, D. O'Toole, F. Costa, *et al.*, "ENETS consensus guidelines update for the management of distant metastatic disease of intestinal, pancreatic, bronchial neuroendocrine neoplasms (NEN) and NEN of unknown primary site," *Neuroendocrinology*, vol. 103, no. 2, pp. 172–185, 2 Apr. 2016. doi: [10.1159/000443167](https://doi.org/10.1159/000443167).
- [258] B. Niederle, U.-F. Pape, F. Costa, *et al.*, "ENETS consensus guidelines update for neuroendocrine neoplasms of the jejunum and ileum," *Neuroendocrinology*, vol. 103, no. 2, pp. 125–138, 2 Apr. 2016. doi: [10.1159/000443170](https://doi.org/10.1159/000443170).
- [259] A. Blažević, W. T. Zandee, G. J. H. Franssen, J. Hofland, M.-L. F. van Velthuisen, L. J. Hofland, R. A. Feelders, and W. W. de Herder, "Mesenteric fibrosis and palliative surgery in small intestinal neuroendocrine tumours," *Endocrine-Related Cancer*, vol. 25, no. 3, pp. 245–254, 3 Mar. 2018. doi: [10.1530/erc-17-0282](https://doi.org/10.1530/erc-17-0282).
- [260] K. Daskalakis, A. Karakatsanis, O. Hessman, H. C. Stuart, S. Welin, E. T. Janson, K. Öberg, P. Hellman, O. Norlén, and P. Stålberg, "Association of a prophylactic surgical approach to stage IV small intestinal neuroendocrine tumors with survival," *JAMA Oncology*, vol. 4, no. 2, p. 183, 2 Feb. 2018. doi: [10.1001/jamaoncol.2017.3326](https://doi.org/10.1001/jamaoncol.2017.3326).
- [261] A. Blažević, J. Hofland, L. J. Hofland, R. A. Feelders, and W. W. de Herder, "Small intestinal neuroendocrine tumours and fibrosis: An entangled conundrum," *Endocrine-Related Cancer*, vol. 25, no. 3, R115–R130, 3 Mar. 2018. doi: [10.1530/erc-17-0380](https://doi.org/10.1530/erc-17-0380).
- [262] M. Druce, N. Bharwani, S. Akker, W. Drake, A. Rockall, and A. Grossman, "Intra-abdominal fibrosis in a recent cohort of patients with neuroendocrine ('carcinoid') tumours of the small bowel," *QJM*, vol. 103, no. 3, pp. 177–185, 3 Mar. 2010. doi: [10.1093/qjmed/hcp191](https://doi.org/10.1093/qjmed/hcp191).
- [263] F. Laskaratos, N. Cox, W. L. Woo, M. Khalifa, M. Ewang, S. Navalkisoor, A. M. Quigley, D. Mandair, M. Caplin, and C. Toumpanakis, "Assessment of changes in mesenteric fibrosis (MF) after peptide receptor radionuclide therapy (PRRT) in midgut neuroendocrine tumours (NETs)," *Neuroendocrinology*, vol. 108, p. 217, 2019, ISSN: 0028-3835.

- [264] I. M. Modlin, B. I. Gustafsson, M. Pavel, B. Svejda, B. Lawrence, and M. Kidd, "A nomogram to assess small-intestinal neuroendocrine tumor ('carcinoid') survival," *Neuroendocrinology*, vol. 92, no. 3, pp. 143–157, 3 2010. doi: [10.1159/000319784](https://doi.org/10.1159/000319784).
- [265] C. Fang, W. Wang, X. Feng, *et al.*, "Nomogram individually predicts the overall survival of patients with gastroenteropancreatic neuroendocrine neoplasms," *British Journal of Cancer*, vol. 117, no. 10, pp. 1544–1550, 10 Nov. 2017. doi: [10.1038/bjc.2017.315](https://doi.org/10.1038/bjc.2017.315).
- [266] S. Pusceddu, F. Barretta, A. Trama, *et al.*, "A classification prognostic score to predict OS in stage IV well-differentiated neuroendocrine tumors," *Endocr Relat Cancer*, vol. 25, no. 6, pp. 607–618, 2018, ISSN: 1479-6821.
- [267] C. A. Karlo, P. L. D. Paolo, J. Chaim, A. A. Hakimi, I. Ostrovnya, P. Russo, H. Hricak, R. Motzer, J. J. Hsieh, and O. Akin, "Radiogenomics of clear cell renal cell carcinoma: Associations between CT imaging features and mutations," *Radiology*, vol. 270, no. 2, pp. 464–471, 2 Feb. 2014. doi: [10.1148/radiol.13130663](https://doi.org/10.1148/radiol.13130663).
- [268] R. Canellas, K. S. Burk, A. Parakh, and D. V. Sahani, "Prediction of pancreatic neuroendocrine tumor grade based on CT features and texture analysis," *American Journal of Roentgenology*, vol. 210, no. 2, pp. 341–346, 2 Feb. 2018. doi: [10.2214/ajr.17.18417](https://doi.org/10.2214/ajr.17.18417).
- [269] L. Pantongrag-Brown, P. C. Buetow, N. J. Carr, J. E. Lichtenstein, and J. L. Buck, "Calcification and fibrosis in mesenteric carcinoid tumor: CT findings and pathologic correlation," *American Journal of Roentgenology*, vol. 164, no. 2, pp. 387–391, 2 Feb. 1995. doi: [10.2214/ajr.164.2.7839976](https://doi.org/10.2214/ajr.164.2.7839976).
- [270] U. Öhrvall, B. Eriksson, C. Juhlin, S. Karacagil, J. Rastad, P. Hellman, and G. Åkerström, "Method for dissection of mesenteric metastases in mid-gut carcinoid tumors," *World Journal of Surgery*, vol. 24, no. 11, pp. 1402–1408, 11 Nov. 2000. doi: [10.1007/s002680010232](https://doi.org/10.1007/s002680010232).
- [271] S. Lardière-Deguelte, L. de Mestier, F. Appéré, *et al.*, "Toward a preoperative classification of lymph node metastases in patients with small intestinal neuroendocrine tumors in the era of intestinal-sparing surgery," *Neuroendocrinology*, vol. 103, no. 5, pp. 552–559, 5 Aug. 2016. doi: [10.1159/000441423](https://doi.org/10.1159/000441423).
- [272] M. P. A. Starmans, *MesentericRadiomics*, <https://github.com/MStarmans91/MesentericRadiomics>, Zenodo, 2021. doi: [10.5281/zenodo.4916317](https://doi.org/10.5281/zenodo.4916317).
- [273] R. Bakeman and V. Quera, *Sequential Analysis and Observational Methods for the Behavioral Sciences*. Cambridge: Cambridge University Press (CUP), 2011. doi: [10.1017/cbo9781139017343](https://doi.org/10.1017/cbo9781139017343).
- [274] H. J. Jang, H. K. Lim, S. J. Lee, W. J. Lee, E. Y. Kim, and S. H. Kim, "Acute diverticulitis of the cecum and ascending colon: The value of thin-section helical CT findings in excluding colonic carcinoma," *AJR Am J Roentgenol*, vol. 174, no. 5, pp. 1397–402, 2000, ISSN: 0361-803X.

- [275] A. Sundin, R. Arnold, E. Baudin, *et al.*, "ENETS consensus guidelines for the standards of care in neuroendocrine tumors: Radiological, nuclear medicine and hybrid imaging," *Neuroendocrinology*, vol. 105, no. 3, pp. 212–244, 3 Sep. 2017. DOI: [10.1159/000471879](https://doi.org/10.1159/000471879).
- [276] C. Kratochwil, P. Flechsig, T. Lindner, *et al.*, "⁶⁸Ga-FAPI PET/CT: Tracer uptake in 28 different kinds of cancer," *Journal of Nuclear Medicine*, vol. 60, no. 6, pp. 801–805, 6 Jun. 2019. DOI: [10.2967/jnumed.119.227967](https://doi.org/10.2967/jnumed.119.227967).
- [277] S. B. Montesi, P. Désogère, B. C. Fuchs, and P. Caravan, "Molecular imaging of fibrosis: Recent advances and future directions," *Journal of Clinical Investigation*, vol. 129, no. 1, pp. 24–33, 1 Jan. 2019. DOI: [10.1172/jci122132](https://doi.org/10.1172/jci122132).
- [278] C. Schmidkonz, S. Rauber, A. Atzinger, *et al.*, "Disentangling inflammatory from fibrotic disease activity by fibroblast activation protein imaging," *Annals of the Rheumatic Diseases*, vol. 79, no. 11, pp. 1485–1491, 11 Nov. 2020. DOI: [10.1136/annrheumdis-2020-217408](https://doi.org/10.1136/annrheumdis-2020-217408).
- [279] S. Manfredi, C. Lepage, C. Hatem, O. Coatmeur, J. Faivre, and A.-M. Bouvier, "Epidemiology and management of liver metastases from colorectal cancer," *Annals of Surgery*, vol. 244, no. 2, pp. 254–259, 2 Aug. 2006. DOI: [10.1097/01.sla.0000217629.94941.cf](https://doi.org/10.1097/01.sla.0000217629.94941.cf).
- [280] J. S. Tomlinson, W. R. Jarnagin, R. P. DeMatteo, Y. Fong, P. Kornprat, M. Gonen, N. Kemeny, M. F. Brennan, L. H. Blumgart, and M. D'Angelica, "Actual 10-Year survival after resection of colorectal liver metastases defines cure," *Journal of Clinical Oncology*, vol. 25, no. 29, pp. 4575–4580, 29 Oct. 2007. DOI: [10.1200/jco.2007.11.0833](https://doi.org/10.1200/jco.2007.11.0833).
- [281] P.-J. van Dam, E. P. van der Stok, L.-A. Teuwen, *et al.*, "International consensus guidelines for scoring the histopathological growth patterns of liver metastasis," *British Journal of Cancer*, vol. 117, no. 10, pp. 1427–1441, 10 Nov. 2017. DOI: [10.1038/bjc.2017.334](https://doi.org/10.1038/bjc.2017.334).
- [282] F. Buisman, E. van der Stok, B. Galjart, *et al.*, "Histopathological growth patterns as a guide for adjuvant systemic chemotherapy in patients with resected colorectal liver metastases," *European Journal of Surgical Oncology*, vol. 45, no. 2, p. e10, 2 Feb. 2019. DOI: [10.1016/j.ejso.2018.10.069](https://doi.org/10.1016/j.ejso.2018.10.069).
- [283] F. E. Buisman, E. P. van der Stok, B. Galjart, *et al.*, "Histopathological growth patterns as biomarker for adjuvant systemic chemotherapy in patients with resected colorectal liver metastases," *Clinical & Experimental Metastasis*, vol. 37, no. 5, pp. 593–605, 5 Oct. 2020. DOI: [10.1007/s10585-020-10048-w](https://doi.org/10.1007/s10585-020-10048-w).
- [284] B. Galjart, P. M. H. Nierop, E. P. van der Stok, R. R. J. C. van den Braak, D. J. Höppener, S. Daelemans, L. Y. Dirix, C. Verhoef, P. B. Vermeulen, and D. J. Grünhagen, "Angiogenic desmoplastic histopathological growth pattern as a prognostic marker of good outcome in patients with colorectal liver metastases," *Angiogenesis*, vol. 22, no. 2, pp. 355–368, 2 May 2019. DOI: [10.1007/s10456-019-09661-5](https://doi.org/10.1007/s10456-019-09661-5).

- [285] E. Latacz, P.-J. van Dam, C. Vanhove, *et al.*, “Can medical imaging identify the histopathological growth patterns of liver metastases?” *Seminars in Cancer Biology*, vol. 71, pp. 33–41, Jun. 2021. doi: [10.1016/j.semcan.2020.07.002](https://doi.org/10.1016/j.semcan.2020.07.002).
- [286] S.-X. Rao, D. M. Lambregts, R. S. Schnerr, *et al.*, “CT texture analysis in colorectal liver metastases: A better way than size and volume measurements to assess response to chemotherapy?” *United European Gastroenterology Journal*, vol. 4, no. 2, pp. 257–263, 2 Apr. 2016. doi: [10.1177/2050640615601603](https://doi.org/10.1177/2050640615601603).
- [287] S.-X. Rao, D. M. Lambregts, R. S. Schnerr, *et al.*, “Whole-liver CT texture analysis in colorectal cancer: Does the presence of liver metastases affect the texture of the remaining liver?” *United European Gastroenterology Journal*, vol. 2, no. 6, pp. 530–538, 6 Dec. 2014. doi: [10.1177/2050640614552463](https://doi.org/10.1177/2050640614552463).
- [288] R. C. Beckers, D. M. Lambregts, R. S. Schnerr, *et al.*, “Whole liver CT texture analysis to predict the development of colorectal liver metastases—a multi-centre study,” *European Journal of Radiology*, vol. 92, pp. 64–71, Jul. 2017. doi: [10.1016/j.ejrad.2017.04.019](https://doi.org/10.1016/j.ejrad.2017.04.019).
- [289] F. Fiz, L. Viganò, N. Gennaro, *et al.*, “Radiomics of liver metastases: A systematic review,” *Cancers*, vol. 12, no. 2881, p. 2881, 10 Oct. 2020. doi: [10.3390/cancers12102881](https://doi.org/10.3390/cancers12102881).
- [290] J. Cheng, J. Wei, T. Tong, *et al.*, “Prediction of histopathologic growth patterns of colorectal liver metastases with a noninvasive imaging method,” *Annals of Surgical Oncology*, vol. 26, pp. 4587–4598, 13 Dec. 2019. doi: [10.1245/s10434-019-07910-x](https://doi.org/10.1245/s10434-019-07910-x).
- [291] M. L. Belli, M. Mori, S. Broggi, *et al.*, “Quantifying the robustness of [18 F]FDG-PET/CT radiomic features with respect to tumor delineation in head and neck and pancreatic cancer patients,” *Physica Medica*, vol. 49, pp. 105–111, May 2018. doi: [10.1016/j.ejmp.2018.05.013](https://doi.org/10.1016/j.ejmp.2018.05.013).
- [292] R. T. H. Leijenaar, S. Carvalho, E. R. Velazquez, *et al.*, “Stability of FDG-PET radiomics features: An integrated analysis of test-retest and inter-observer variability,” *Acta Oncologica*, vol. 52, no. 7, pp. 1391–1397, 7 Oct. 2013. doi: [10.3109/0284186x.2013.812798](https://doi.org/10.3109/0284186x.2013.812798).
- [293] R. Berenguer, M. del RosarioPastor-Juan, J. Canales-Vázquez, M. Castro-García, M. V. Villas, F. M. Legorburo, and S. Sabater, “Radiomics of CT features may be nonreproducible and redundant: Influence of CT acquisition parameters,” *Radiology*, vol. 288, no. 2, pp. 407–415, 2 Aug. 2018. doi: [10.1148/radiol.2018172361](https://doi.org/10.1148/radiol.2018172361).
- [294] D. G. Fryback and J. R. Thornbury, “The efficacy of diagnostic imaging,” *Medical Decision Making*, vol. 11, no. 2, pp. 88–94, 2 Jun. 1991. doi: [10.1177/0272989x9101100203](https://doi.org/10.1177/0272989x9101100203).
- [295] M. P. A. Starmans, *CLMRadiomics*, <https://github.com/MStarmans91/CLMRadiomics>, Zenodo, 2021. doi: [10.5281/zenodo.43928](https://doi.org/10.5281/zenodo.43928).

- [296] A. Mikolajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, Institute of Electrical and Electronics Engineers (IEEE), May 2018, pp. 117–122. doi: [10.1109/iiphdw.2018.8388338](https://doi.org/10.1109/iiphdw.2018.8388338).
- [297] C. M. Bishop, *Pattern recognition and machine learning*, ser. Information science and statistics. New York: Springer, 2006, p. 738, ISBN: 0387310738 (hd.bd.) 9780387310732.
- [298] S. Bipat, M. S. van Leeuwen, E. F. I. Comans, M. E. J. Pijl, P. M. M. Bossuyt, A. H. Zwinderman, and J. Stoker, "Colorectal liver metastases: CT, MR imaging, and PET for diagnosis—meta-analysis," *Radiology*, vol. 237, no. 1, pp. 123–131, 1 Oct. 2005. doi: [10.1148/radiol.2371042060](https://doi.org/10.1148/radiol.2371042060).
- [299] A. S. Becker, M. A. Schneider, M. C. Wurnig, M. Wagner, P. A. Clavien, and A. Boss, "Radiomics of liver MRI predict metastases in mice," *European Radiology Experimental*, vol. 2, no. 1, p. 11, 1 Dec. 2018. doi: [10.1186/s41747-018-0044-7](https://doi.org/10.1186/s41747-018-0044-7).
- [300] P. Kovesi, "Image features from phase congruency," *Videre: Journal of computer vision research*, vol. 1, no. 3, pp. 1–26, 1999.
- [301] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 3 May 2021. doi: [10.3322/caac.21660](https://doi.org/10.3322/caac.21660).
- [302] J. Balogh, D. Victor, E. H. Asham, S. G. Burroughs, M. Boktour, A. Saharia, X. Li, M. Ghobrial, and H. Monsour, "Hepatocellular carcinoma: A review," *Journal of Hepatocellular Carcinoma*, vol. Volume 3, pp. 41–53, Oct. 2016. doi: [10.2147/jhc.s61146](https://doi.org/10.2147/jhc.s61146).
- [303] L. Grazioli, M. P. Bondioni, H. Haradome, U. Motosugi, R. Tinti, B. Frittoli, S. Gambarini, F. Donato, and S. Colagrande, "Hepatocellular adenoma and focal nodular hyperplasia: Value of gadoxetic acid-enhanced MR imaging in differential diagnosis," *Radiology*, vol. 262, no. 2, pp. 520–529, 2 Feb. 2012. doi: [10.1148/radiol.11101742](https://doi.org/10.1148/radiol.11101742).
- [304] P. R. Galle, A. Forner, J. M. Llovet, V. Mazzaferro, F. Piscaglia, J.-L. Raoul, P. Schirmacher, and V. Vilgrain, "EASL clinical practice guidelines: Management of hepatocellular carcinoma," *Journal of Hepatology*, vol. 69, no. 1, pp. 182–236, 1 Jul. 2018. doi: [10.1016/j.jhep.2018.03.019](https://doi.org/10.1016/j.jhep.2018.03.019).
- [305] B. K. Barth, O. F. Donati, M. A. Fischer, E. J. Ulbrich, C. A. Karlo, A. Becker, B. Seifert, and C. S. Reiner, "Reliability, validity, and reader acceptance of LI-RADS: An in-depth analysis," *Academic Radiology*, vol. 23, no. 9, pp. 1145–1153, 9 Sep. 2016. doi: [10.1016/j.acra.2016.03.014](https://doi.org/10.1016/j.acra.2016.03.014).
- [306] M. A. Silva, B. Hegab, C. Hyde, B. Guo, J. A. C. Buckels, and D. F. Mirza, "Needle track seeding following biopsy of liver lesions in the diagnosis of hepatocellular cancer: A systematic review and meta-analysis," *Gut*, vol. 57, no. 11, pp. 1592–1596, 11 Nov. 2008. doi: [10.1136/gut.2008.149062](https://doi.org/10.1136/gut.2008.149062).

- [307] A. Saini, I. Breen, Y. Pershad, S. Naidu, M. Knuttinen, S. Alzubaidi, R. Sheth, H. Albadawi, M. Kuo, and R. Oklu, "Radiogenomics and radiomics in liver cancers," *Diagnostics*, vol. 9, no. 1, p. 4, 1 Mar. 2019. doi: [10.3390/diagnostics9010004](https://doi.org/10.3390/diagnostics9010004).
- [308] M. J. A. Jansen, H. J. Kuijf, W. B. Veldhuis, F. J. Wessels, M. A. Viergever, and J. P. W. Pluim, "Automatic classification of focal liver lesions based on MRI and risk factors," *PLOS ONE*, vol. 14, no. 5, T. M. Deserno, Ed., p. e0217053, 5 May 2019. doi: [10.1371/journal.pone.0217053](https://doi.org/10.1371/journal.pone.0217053).
- [309] I. Gatos, S. Tsantis, M. Karamezini, S. Spiliopoulos, D. Karnabatidis, J. D. Hazle, and G. C. Kagadis, "Focal liver lesions segmentation and classification in nonenhanced T2-weighted MRI," *Medical Physics*, vol. 44, no. 7, pp. 3695–3705, 7 Jul. 2017. doi: [10.1002/mp.12291](https://doi.org/10.1002/mp.12291).
- [310] S.-h. Zhen, M. Cheng, Y.-b. Tao, *et al.*, "Deep learning for accurate diagnosis of liver tumor based on magnetic resonance imaging and clinical data," *Frontiers in Oncology*, vol. 10, p. 680, May 2020. doi: [10.3389/fonc.2020.00680](https://doi.org/10.3389/fonc.2020.00680).
- [311] American College of Radiology, *Liver reporting & data system (LI-RADS)*, <https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/LI-RADS>.
- [312] Y. Kim, A. Furlan, A. A. Borhani, and K. T. Bae, "Computer-aided diagnosis program for classifying the risk of hepatocellular carcinoma on MR images following liver imaging reporting and data system (LI-RADS)," *Journal of Magnetic Resonance Imaging*, vol. 47, pp. 710–722, 3 Mar. 2018. doi: [10.1002/jmri.25772](https://doi.org/10.1002/jmri.25772).
- [313] T. Wakabayashi, F. Ouhmich, C. Gonzalez-Cabrera, *et al.*, "Radiomics in hepatocellular carcinoma: A quantitative review," *Hepatology International*, vol. 13, no. 5, pp. 546–559, 5 Sep. 2019. doi: [10.1007/s12072-019-09973-0](https://doi.org/10.1007/s12072-019-09973-0).
- [314] H. Oka, N. Kurioka, K. Kim, T. Kanno, T. Kuroki, Y. Mizoguchi, and K. Kobayashi, "Prospective study of early detection of hepatocellular carcinoma in patients with cirrhosis," *Hepatology*, vol. 12, no. 4 Pt 1, pp. 680–687, 4 Oct. 1990. doi: [10.1002/hep.1840120411](https://doi.org/10.1002/hep.1840120411).
- [315] A. S. Befeler and A. M. di Bisceglie, "Hepatocellular carcinoma: Diagnosis and treatment," *Gastroenterology*, vol. 122, no. 6, pp. 1609–1619, 6 May 2002. doi: [10.1053/gast.2002.33411](https://doi.org/10.1053/gast.2002.33411).
- [316] M. P. A. Starmans, *LiverRadiomics*, <https://github.com/MStarmans91/LiverRadiomics>, Zenodo, 2021. doi: [10.5281/zenodo.5175705](https://doi.org/10.5281/zenodo.5175705).
- [317] S. M. van Aalten, M. G. J. Thomeer, T. Terkivatan, R. S. Dwarkasing, J. Verheij, R. A. de Man, and J. N. M. IJzermans, "Hepatocellular adenomas: Correlation of MR imaging findings with pathologic subtype classification," *Radiology*, vol. 261, no. 1, pp. 172–181, 1 Oct. 2011. doi: [10.1148/radiol.11110023](https://doi.org/10.1148/radiol.11110023).
- [318] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012. doi: [10.11613/bm.2012.031](https://doi.org/10.11613/bm.2012.031).

- [319] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesthesia & Analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018. doi: [10.1213/ANE.0000000000002864](https://doi.org/10.1213/ANE.0000000000002864).
- [320] B. E. V. Beers, C. M. Pastor, and H. K. Hussain, "Primovist, eovist: What to expect?" *Journal of Hepatology*, vol. 57, no. 2, pp. 421–429, 2 Aug. 2012. doi: [10.1016/j.jhep.2012.01.031](https://doi.org/10.1016/j.jhep.2012.01.031).
- [321] L. D. Tommaso, A. Destro, J. Y. Seok, *et al.*, "The application of markers (HSP70 GPC3 and GS) in liver biopsies is useful for detection of hepatocellular carcinoma," *Journal of Hepatology*, vol. 50, no. 4, pp. 746–754, 4 Apr. 2009. doi: [10.1016/j.jhep.2008.11.014](https://doi.org/10.1016/j.jhep.2008.11.014).
- [322] J. M. Banales, J. J. G. Marin, A. Lamarca, *et al.*, "Cholangiocarcinoma 2020: The next horizon in mechanisms and management," *Nature Reviews Gastroenterology & Hepatology*, vol. 17, no. 9, pp. 557–588, 9 Sep. 2020. doi: [10.1038/s41575-020-0310-z](https://doi.org/10.1038/s41575-020-0310-z).
- [323] P. Bioulac-Sage, H. Laumonier, A. Rullier, G. Cubel, C. Laurent, J. Zucman-Rossi, and C. Balabaud, "Over-expression of glutamine synthetase in focal nodular hyperplasia: A novel easy diagnostic tool in surgical pathology," *Liver International*, vol. 29, no. 3, pp. 459–465, 3 Mar. 2009. doi: [10.1111/j.1478-3231.2008.01849.x](https://doi.org/10.1111/j.1478-3231.2008.01849.x).
- [324] MINIMALIST trial (NL8738), <https://www.kanker.nl/trials/1131-minimalist---studie-wekedelentumoren>, 2020.
- [325] S. Beck, C. Bergenholtz, M. Bogers, *et al.*, "The open innovation in science research field: A collaborative conceptualisation approach," *Industry and Innovation*, pp. 1–50, Aug. 2020. doi: [10.1080/13662716.2020.1792274](https://doi.org/10.1080/13662716.2020.1792274).
- [326] E. C. McKiernan, P. E. Bourne, C. T. Brown, *et al.*, "How open science helps researchers succeed," *eLife*, vol. 5, p. e16800, Jul. 2016. doi: [10.7554/elife.16800](https://doi.org/10.7554/elife.16800).
- [327] B. van Ginneken, S. Kerkstra, and J. Meakin, *Grand challenge*, <https://grand-challenge.org/>.
- [328] K. Clark, B. Vendt, K. Smith, *et al.*, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *Journal of Digital Imaging*, vol. 26, no. 6, pp. 1045–1057, 6 Dec. 2013. doi: [10.1007/s10278-013-9622-7](https://doi.org/10.1007/s10278-013-9622-7).
- [329] R. C. Petersen, P. S. Aisen, L. A. Beckett, *et al.*, "Alzheimer's disease neuroimaging initiative (ADNI): Clinical characterization," *Neurology*, vol. 74, no. 3, pp. 201–209, 3 Jan. 2010. doi: [10.1212/wnl.0b013e3181cb3e25](https://doi.org/10.1212/wnl.0b013e3181cb3e25).
- [330] E. E. Bron, M. Smits, W. M. van der Flier, *et al.*, "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge," *NeuroImage*, vol. 111, pp. 562–579, May 2015. doi: [10.1016/j.neuroimage.2015.01.048](https://doi.org/10.1016/j.neuroimage.2015.01.048).

- [331] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, and R. Wiest, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014, issn: 0278-0062.
- [332] E. Martin, R. T. J. Geitenbeek, J. H. Coert, D. F. Hanff, L. H. Graven, D. J. Grünhagen, C. Verhoef, and W. Taal, "A bayesian approach for diagnostic accuracy of malignant peripheral nerve sheath tumors: A systematic review and meta-analysis," *Neuro-Oncology*, vol. 23, no. 4, pp. 557–571, 4 Apr. 2021. doi: [10.1093/neuonc/noaa280](https://doi.org/10.1093/neuonc/noaa280).
- [333] N. Beije, I. E. de Kruijff, J. de Jong, *et al.*, "Circulating tumor cell-driven use of neoadjuvant chemotherapy in patients with muscle-invasive bladder cancer," *Journal of Clinical Oncology*, vol. 39, no. 15 Supplement, pp. 4523–4523, Supplement 15 May 2021. doi: [10.1200/jco.2021.39.15_suppl.4523](https://doi.org/10.1200/jco.2021.39.15_suppl.4523).
- [334] J. I. van Waning, K. Caliskan, M. Michels, *et al.*, "Cardiac phenotypes, genetics, and risks in familial noncompaction cardiomyopathy," *Journal of the American College of Cardiology*, vol. 73, no. 13, pp. 1601–1611, 13 Apr. 2019. doi: [10.1016/j.jacc.2018.12.085](https://doi.org/10.1016/j.jacc.2018.12.085).
- [335] B. Witjes, S. Baillet, M. Roy, R. Oostenveld, F. J. Huygen, and C. C. de Vos, "Magnetoencephalography reveals increased slow-to-fast alpha power ratios in patients with chronic pain," *PAIN Reports*, vol. 6, no. 2, p. e928, 2 2021. doi: [10.1097/pr9.0000000000000928](https://doi.org/10.1097/pr9.0000000000000928).
- [336] R. Granzier, T. van Nijmegen, H. Woodruff, M. Smidt, and M. Lobbes, "Exploring breast cancer response prediction to neoadjuvant systemic therapy using MRI-based radiomics: A systematic review," *European Journal of Radiology*, vol. 121, p. 108736, Dec. 2019. doi: [10.1016/j.ejrad.2019.108736](https://doi.org/10.1016/j.ejrad.2019.108736).
- [337] E. J. Bijl, M. P. A. Starmans, J. M. Mostert, S. Klein, F. J. P. M. Huygen, and C. C. de Vos, "Automatic quantification of complex regional pain syndrome using radiomics and deep learning based on thermography images," *In Preparation*.
- [338] F. Hutter, L. Kotthoff, and J. Vanschoren, Eds., *Automated Machine Learning: Methods, Systems, Challenges*, ser. The Springer Series on Challenges in Machine Learning. Cham: Springer, 2019, isbn: 3-030-05318-0.
- [339] G. Luo, "A review of automatic selection methods for machine learning algorithms and hyper-parameter values," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 5, no. 1, p. 18, 1 Dec. 2016. doi: [10.1007/s13721-016-0125-6](https://doi.org/10.1007/s13721-016-0125-6).
- [340] J. Waring, C. Lindvall, and R. Umeton, "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare," *Artificial Intelligence in Medicine*, vol. 104, p. 101822, Apr. 2020. doi: [10.1016/j.artmed.2020.101822](https://doi.org/10.1016/j.artmed.2020.101822).
- [341] Q. Yao, M. Wang, Y. Chen, W. Dai, Y.-F. Li, W.-W. Tu, Q. Yang, and Y. Yu, "Taking human out of learning applications: A survey on automated machine learning," 2018.


- [342] H. J. Escalante, M. Montes, and E. Sucar, "Ensemble particle swarm model selection," *The 2010 International Joint Conference on Neural Networks (IJCNN)*, vol. 10, pp. 405–440, Jul. 2010. DOI: [10.1109/ijcnn.2010.5596915](https://doi.org/10.1109/ijcnn.2010.5596915).
- [343] W. Banzhaf, P. Nordin, R. E. Keller, and F. D. Francone, *Genetic programming: an introduction: on the automatic evolution of computer programs and its applications*. Morgan Kaufmann Publishers Inc., 1998, ISBN: 1-55860-510-X.
- [344] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, "The (un)reliability of saliency methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Lecture Notes in Computer Science*. Springer Science and Business Media LLC, 2019, pp. 267–280. DOI: [10.1007/978-3-030-28954-6_14](https://doi.org/10.1007/978-3-030-28954-6_14).
- [345] N. Hansen, "The CMA evolution strategy: A comparing review," in *Towards a New Evolutionary Computation, Studies in Fuzziness and Soft Computing*, J. A. Lozano, P. Larrañaga, I. Inza, and E. Bengoetxea, Eds. Berlin, Heidelberg: Springer Science and Business Media LLC, pp. 75–102. DOI: [10.1007/3-540-32494-1_4](https://doi.org/10.1007/3-540-32494-1_4).
- [346] C. Antonio, "Sequential model based optimization of partially defined functions under unknown constraints," *Journal of Global Optimization*, vol. 79, no. 2, pp. 281–303, 2 Feb. 2021. DOI: [10.1007/s10898-019-00860-4](https://doi.org/10.1007/s10898-019-00860-4).
- [347] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: A review of bayesian optimization," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148–175, 1 Jan. 2016. DOI: [10.1109/jproc.2015.2494218](https://doi.org/10.1109/jproc.2015.2494218).
- [348] M. Deen, "Automatic algorithm selection and hyperparameter optimization for medical image classification," MSc Thesis, 2021.
- [349] R. Marler and J. Arora, "Survey of multi-objective optimization methods for engineering," *Structural and Multidisciplinary Optimization*, vol. 26, no. 6, pp. 369–395, 6 Apr. 2004. DOI: [10.1007/s00158-003-0368-6](https://doi.org/10.1007/s00158-003-0368-6).
- [350] P. Ngatchou, A. Zarei, and A. El-Sharkawi, "Pareto multi objective optimization," in *Proceedings of the 13th International Conference on, Intelligent Systems Application to Power Systems*, Institute of Electrical and Electronics Engineers (IEEE), pp. 84–91. DOI: [10.1109/isap.2005.1599245](https://doi.org/10.1109/isap.2005.1599245).
- [351] A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care," *Neural Computing and Applications*, vol. 32, no. 24, pp. 18 069–18 083, 24 Dec. 2020. DOI: [10.1007/s00521-019-04051-w](https://doi.org/10.1007/s00521-019-04051-w).
- [352] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 5 May 2019. DOI: [10.1038/s42256-019-0048-x](https://doi.org/10.1038/s42256-019-0048-x).
- [353] P. Cortez and M. J. Embrechts, "Using sensitivity analysis and visualization techniques to open black box data mining models," *Information Sciences*, vol. 225, pp. 1–17, Mar. 2013. DOI: [10.1016/j.ins.2012.10.039](https://doi.org/10.1016/j.ins.2012.10.039).

- [354] L. Merrick and A. Taly, "The explanation game: Explaining machine learning models using shapley values," in *Lecture Notes in Computer Science, Machine Learning and Knowledge Extraction*, Machine Learning and Knowledge Extraction, Springer Science and Business Media LLC, 2020, pp. 17–38. doi: [10.1007/978-3-030-57321-8_2](https://doi.org/10.1007/978-3-030-57321-8_2).
- [355] E. C. F. Wilson, "A practical guide to value of information analysis," *PharmacoEconomics*, vol. 33, no. 2, pp. 105–121, 2 Feb. 2015. doi: [10.1007/s40273-014-0219-x](https://doi.org/10.1007/s40273-014-0219-x).
- [356] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxel-Morph: A learning framework for deformable medical image registration," *IEEE Transactions on Medical Imaging*, vol. 38, pp. 1788–1800, 8 Aug. 2019. doi: [10.1109/tmi.2019.2897538](https://doi.org/10.1109/tmi.2019.2897538).
- [357] M. P. Heinrich, O. Oktay, and N. Bouteldja, "OBELISK-Net: Fewer layers to solve 3D multi-organ segmentation with sparse deformable convolutions," *Medical Image Analysis*, vol. 54, pp. 1–9, May 2019. doi: [10.1016/j.media.2019.02.006](https://doi.org/10.1016/j.media.2019.02.006).
- [358] M.-A. Schulz, B. T. T. Yeo, J. T. Vogelstein, J. Mourao-Miranada, J. N. Kather, K. Kording, B. Richards, and D. Bzdok, "Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets," *Nature Communications*, vol. 11, no. 1, p. 4238, 1 Dec. 2020. doi: [10.1038/s41467-020-18037-z](https://doi.org/10.1038/s41467-020-18037-z).
- [359] R. Gupta, T. Kurc, A. Sharma, J. S. Almeida, and J. Saltz, "The emergence of pathomics," *Current Pathobiology Reports*, vol. 7, no. 3, pp. 73–84, 3 Sep. 2019. doi: [10.1007/s40139-019-00200-x](https://doi.org/10.1007/s40139-019-00200-x).
- [360] D. Ganeshan, P.-A. T. Duong, L. Probyn, L. Lenchik, T. A. McArthur, M. Retrouvey, E. H. Ghobadi, S. L. Desouches, D. Pastel, and I. R. Francis, "Structured reporting in radiology," *Academic Radiology*, vol. 25, no. 1, pp. 66–73, 1 Jan. 2018. doi: [10.1016/j.acra.2017.08.005](https://doi.org/10.1016/j.acra.2017.08.005).
- [361] A. Buniello, J. A. L. MacArthur, M. Cerezo, *et al.*, "The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019," *Nucleic Acids Research*, vol. 47, no. D1, pp. D1005–D1012, D1 Jan. 2019. doi: [10.1093/nar/gky1120](https://doi.org/10.1093/nar/gky1120).
- [362] G. Roshchupkin, "Imaging genetics : Methodological approaches to overcoming high dimensional barriers," Ph.D. thesis, 2018.
- [363] O. Morin, M. Vallières, S. Braunstein, *et al.*, "An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication," *Nature Cancer*, vol. 2, no. 7, pp. 709–722, 7 Jul. 2021. doi: [10.1038/s43018-021-00236-2](https://doi.org/10.1038/s43018-021-00236-2).
- [364] F. Sadoughi, A. Behmanesh, and N. Sayfour, "Internet of things in medicine: A systematic mapping study," *Journal of Biomedical Informatics*, vol. 103, p. 103383, Mar. 2020. doi: [10.1016/j.jbi.2020.103383](https://doi.org/10.1016/j.jbi.2020.103383).

- [365] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, "Artificial intelligence in radiology," *Nature Reviews Cancer*, vol. 18, no. 8, pp. 500–510, 8 Aug. 2018. doi: [10.1038/s41568-018-0016-5](https://doi.org/10.1038/s41568-018-0016-5).
- [366] K. G. van Leeuwen, S. Schalekamp, M. J. C. M. Rutten, B. van Ginneken, and M. de Rooij, "Artificial intelligence in radiology: 100 commercially available products and their scientific evidence," *European Radiology*, vol. 31, no. 6, pp. 3797–3804, 6 Jun. 2021. doi: [10.1007/s00330-021-07892-z](https://doi.org/10.1007/s00330-021-07892-z).
- [367] M. Huisman, E. Ranschaert, W. Parker, *et al.*, "An international survey on AI in radiology in 1,041 radiologists and radiology residents part 1: Fear of replacement, knowledge, and attitude," *European Radiology*, Mar. 2021. doi: [10.1007/s00330-021-07781-5](https://doi.org/10.1007/s00330-021-07781-5).
- [368] M. Huisman, E. Ranschaert, W. Parker, *et al.*, "An international survey on AI in radiology in 1041 radiologists and radiology residents part 2: Expectations, hurdles to implementation, and education," *European Radiology*, May 2021. doi: [10.1007/s00330-021-07782-4](https://doi.org/10.1007/s00330-021-07782-4).
- [369] S. Halligan, Y. Menu, and S. Mallett, "Why did european radiology reject my radiomic biomarker paper? how to correctly evaluate imaging biomarkers in a clinical setting," *European Radiology*, May 2021. doi: [10.1007/s00330-021-07971-1](https://doi.org/10.1007/s00330-021-07971-1).
- [370] P. P. Sengupta, S. Shrestha, B. Berthon, *et al.*, "Proposed requirements for cardiovascular imaging-related machine learning evaluation (PRIME): A checklist," *JACC: Cardiovascular Imaging*, vol. 13, no. 9, pp. 2017–2035, 9 Sep. 2020. doi: [10.1016/j.jcmg.2020.07.015](https://doi.org/10.1016/j.jcmg.2020.07.015).
- [371] E. Capobianco and J. Deng, "Radiomics at a glance: A few lessons learned from learning approaches," *Cancers*, vol. 12, no. 9, p. 2453, 9 Sep. 2020. doi: [10.3390/cancers12092453](https://doi.org/10.3390/cancers12092453).
- [372] Delft University of Technology and Erasmus University Medical Center, *Fundamental and applied scientists, engineers, doctors and ethicists collaborate to improve health*, <https://www.nature.com/articles/d42473-019-00143-2>.
- [373] C. Alliance, *Delft university of technology erasmus university rotterdam erasmus medical center*, <https://convergencealliance.nl/>.
- [374] S. Klein and M. P. A. Starmans, *Automatic grading and phenotyping of soft-tissue tumors through machine learning to guide personalized cancer treatment*, <https://www.hanarthfonds.nl/en/stefan-klein>, 2020.
- [375] M. P. A. Starmans, M. G. Thomeer, and S. Klein, *The liver artificial intelligence (LAI) consortium: A benchmark dataset and optimized machine learning methods for MRI-based diagnosis of solid appearing liver lesions*, Grant Under Review, 2021.
- [376] *EUCanImage: Towards a european cancer imaging platform for enhanced artificial intelligence in oncology*, <https://eucanimage.eu/>.
- [377] *NWO open science fund*, <https://www.nwo.nl/en/researchprogrammes/open-science/open-science-fund>, 2020.

- [378] K. Bonawitz, H. Eichner, W. Grieskamp, *et al.*, "Towards federated learning at scale: System design," 2019. arXiv: [1902.01046](https://arxiv.org/abs/1902.01046).
- [379] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 3 May 2020. doi: [10.1109/msp.2020.2975749](https://doi.org/10.1109/msp.2020.2975749).

Due to the paradigm shift in health care towards personalized medicine, there is an increased demand for biomarkers. Radiomics leverages quantitative medical imaging features and machine learning to create biomarkers based on medical imaging. While many radiomics methods have been described in the literature, these are generally designed for a single application. The overall aim of this thesis is to streamline radiomics research, facilitate its reproducibility, and simplify its application. In this thesis, we exploit recent advances in automated machine learning to develop an adaptive radiomics framework and demonstrate its use to develop radiomics biomarkers in eight different, independent clinical applications.

A thick, white, wavy line that starts from the left edge of the page and curves upwards and then downwards towards the right edge, creating a smooth, flowing shape.