



**Università
di Genova**

Ph.D. in Mathematics and Applications

Curriculum

Mathematical Methods for Data Analysis

**Towards explainable radiomics:
stability and interpretability in
computational models for clinical
applications**

Candidate:

Isabella Cama

Supervisors:

Prof. Cristina Campi

Dr. Sara Garbarino

UniGe
DIMA

XXXVII Cycle



OSPEDALE POLICLINICO SAN MARTINO
Sistema Sanitario Regione Liguria

Abstract

Radiomics has been widely used in medical imaging for various tasks, like biomarker prediction, tumor subtype classification, and forecast of disease recurrence and treatment response. However, increasing literature shows that the radiomics approach is rarely explored from a methodological perspective. This thesis focuses on key challenges of radiomic analysis, encompassing segmentation, evaluation of feature stability, construction of explicable prediction models, and image harmonization, with the main scope of investigating radiomics reliability and contributing to the development of interpretable radiomics-based models for prediction purposes in clinical applications. Radiomics reliability is addressed in terms of stability with respect to segmentation variability and stability of the predictive performance. The proposed machine learning models focus on the explainability of feature selection for the prediction of triple-negative breast cancer subtype, and on the interpretability of model architecture for radiomics-based longitudinal prediction of glioblastoma response to treatment. Other contributions concern numerical schemes for medical image segmentation and deep learning Image2Image network for image harmonization, both conceived as a preliminary step to radiomic analysis. While the investigation conducted in this thesis deals with specific applications, the methodology introduced and the findings reported could potentially provide a broader understanding of the principles underlying radiomics-based analyses. Ultimately, this work aspires to contribute to a deeper comprehension of the interplay between feature stability and segmentation variability for model explainability. Moreover, the construction of

interpretable models and harmonization strategies represents a step toward refining the radiomics workflow, fostering the development of more reliable, generalizable, and clinically applicable computational models for diagnosis and prognosis.

Contents

Introduction	1
1 Segmentation	6
1.1 Motivation	8
1.2 Materials and methods	11
1.2.1 Evolution models	11
1.2.2 Numerical schemes for model discretization	13
1.2.3 Pre-processing and post-processing	16
1.2.4 Data collection	20
1.2.5 Benchmark and evaluation metrics	22
1.3 Results	23
1.4 Discussion	28
2 Radiomics reliability	34
2.1 Motivation	35
2.2 Materials and methods	38
2.2.1 Data collection	38
2.2.2 Segmentation	38
2.2.3 Radiomic feature extraction	41
2.2.4 Clinical features	42
2.2.5 Machine Learning models	43
2.2.6 Statistical analysis	45
2.2.7 Radiomics reliability scores	47
2.3 Results	51

2.4	Discussion	59
3	Radiomics for longitudinal prediction	67
3.1	Motivation	69
3.2	Materials and methods	71
3.2.1	Longitudinal approach	71
3.2.2	Data	74
	Lumiere longitudinal dataset	74
	Synthetic longitudinal dataset	77
3.2.3	Benchmark experiments	79
3.2.4	Machine learning models	83
3.3	Results	85
3.4	Discussion	88
4	Image harmonization	93
4.1	Materials and methods	95
4.1.1	Architecture	96
4.1.2	Losses	97
4.1.3	Datasets	102
4.1.4	Pre-processing	103
4.1.5	Benchmarking	103
4.1.6	Evaluation procedure	105
4.2	Results	106
4.2.1	Evaluation of scanner information transfer	107
4.2.2	Evaluation of scanner free space	107
4.2.3	Preservation of anatomical structure	108
4.2.4	Age prediction task	109
4.3	Discussion	111
	Conclusions	114
	Bibliography	120

List of Figures

1.1	Example of abdominal CT slices.	10
1.2	View of the psoas muscle.	21
1.3	Segmentations obtained through the classical evolution model, GMFD 1ord, and GMFD 2ord on the internal cohort.	25
1.3	Figure continued from previous page. Segmentations obtained through the classical evolution model, GMFD 1ord, and GMFD 2ord on the internal cohort.	26
1.4	Post-processing pipeline applied to GMFD 1ord segmentation outcome on the internal cohort.	26
1.5	Segmentations obtained for a representative subject from the CMB-PCA dataset.	27
2.1	First-post-contrast image and segmentation masks of three different patients (manual tumor segmentation, closing 08 mask, closing 07, closing 06, and ellipsoid 04).	40
2.2	Flowchart of the feature selection methodology employed for this study, based on SHAP explainability algorithm.	45
2.3	Scatter plot of feature stability and reliability histogram.	47
2.4	Reliability matrices: quality, consistency, robustness, and instability.	48
2.5	Example of original DCE-MR image, wavelet-HHH filtered image, wavelet-HLH filtered image, and LoG $\sigma = 3$ filtered image, with tumor segmentation.	53

LIST OF FIGURES

2.6	ROC-AUC scores obtained by testing demographical model, biopsy model, baseline models, and best-SHAP models.	53
2.7	ICC and Pearson correlation of the four common best-SHAP features at varying segmentation mask.	56
2.8	Reliability scores of the four common SHAP features across segmentation masks.	57
2.9	Scatter plots of feature stability w.r.t. segmentation variability across segmentation masks, for the four common SHAP features.	58
3.1	Proposed probabilistic approach for the longitudinal prediction of response to treatment.	73
3.2	Lumiere dataset overview.	76
3.3	Eligible patients of Lumiere dataset.	77
3.4	Synthetic longitudinal dataset overview.	80
3.5	Calibration curves for the baseline model to predict the response at first follow-up, in the case of synthetic and Lumiere dataset.	86
3.6	Examples of probability distribution of disease progression at the first follow-up, for patients from synthetic and Lumiere dataset.	92
4.1	Brain structure encoder, scanner encoder, and generator.	98
4.2	Transfer of an image into a target scanner space and generation of an image within the scanner-free space.	99
4.3	Overview of DISARM training procedure.	100
4.4	Average gray distributions in each scanner pre- and post- transfer to Prisma Fit scanner space.	108
4.5	Average gray distributions in each scanner before and after the transfer to the scanner-free space $\mathcal{F}$	109
4.6	Age regression lines, pre- and post-harmonization.	110
4.7	Examples of harmonized images with DISARM and StarGANv2.	113

List of Tables

1.1	Number of iterations, CPU time, parameters p , σ , and number of slices for the three segmentation algorithms.	24
1.2	Evaluation of the segmentation results of classical model, GMFD 1ord, GMFD 2ord, TS-FM, and TS-FRM, compared to manual segmentation on the internal cohort.	27
1.3	Evaluation of the segmentation results of classical model, GMFD 1ord, GMFD 2ord, TS-FM, and TS-FRM compared to manual segmentation on CMB-PCA dataset.	28
2.1	DSC obtained for each modification of the manual segmentation mask, computed across all patients.	51
2.2	List of the best-SHAP features identified as explicative through SHAP method for each model.	52
2.3	Performances of demographical model, biopsy model, baseline models, and best-SHAP models for TNBC prediction.	54
2.4	Results of the Kolmogorov-Smirnov tests to evaluate differences in model performance.	54
3.1	Summary of the performed experiments.	82
3.2	Calibration scores for baseline models to predict the response at the first follow-up, in the case of synthetic and Lumiere dataset.	85
3.3	Performance of the implemented models on the synthetic dataset.	87
3.4	Performance of the implemented models on Lumiere dataset.	87

LIST OF TABLES

4.1	Parameters used for the training of DISARM.	102
4.2	Significance of TGV variable and TGV-scanner interactions in age regression models before and after harmonization.	110

Abbreviations

The following abbreviations have been introduced.

- ADC: Apparent Diffusion Coefficient
- AD: Alzheimer's disease
- ADNI: Alzheimer's Disease Neuroimaging
- AI: Artificial Intelligence
- ALS: Amyotrophic Lateral Sclerosis
- ASSD: Average Symmetric Surface Distance
- BI-RADS: American College of Radiology's Breast Imaging Reporting and Data System
- CC: connected component
- CFL: Courant-Friedrichs-Lewy
- CI: Confidence Interval
- CLSTM: convolutional long short-term memory
- CMB-PCA: Cancer Moonshot Biobank - Prostate Cancer Collection
- CNN: Convolutional Neural Network
- CR: complete response

- CT: X-ray computed tomography
- CV: Cross Validation
- DCE: Dynamic Contrast-Enhanced
- DISARM: disentangled scanner-free image generation via unsupervised image2image translation
- DL: Deep Learning
- DRIT: diverse image2image translation via disentangled representations
- DSC: Dice Similarity Coefficient
- DWI: Diffusion Weighted Imaging
- FDG: 18F-Fluorodeoxyglucose
- FLAIR: Fluid-Attenuated Inversion Recovery
- GAN: Generative Adversarial Network
- GBM: glioblastoma
- GLCM: gray-level co-occurrence matrix
- GLDM: gray-level dependence matrix
- GLRLM: gray-level run-length matrix
- GLSZM: gray-level size zone matrix
- GMFD 1ord: first order geodesic model finite differences
- GMFD 2ord: second order geodesic model finite differences
- HU: Hounsfield Unit
- I2I: Image2Image

- ICC: Intraclass Correlation Coefficient
- IDH1: isocitrate dehydrogenase 1
- G-KDE: Gaussian Kernel Density Estimation
- L3: third lumbar vertebra
- LoG: Laplacian of Gaussian
- MCI: mild cognitive impairment
- MGMT: O6-methylguanine-DNA methyltransferase
- ML: Machine Learning
- MRI: Magnetic Resonance Imaging
- NGTDM: neighboring gray-tone difference matrix
- PCA: Principal Component Analysis
- PD: progressive disease
- PDF: Probability Density Function
- PET: Positron Emission Tomography
- post-op: post-operative
- PPMI: Parkinson's Progression Marker Initiative
- PR: partial response
- pre-op: pre-operative
- RANO: Response Assessment in Neuro-Oncology criteria
- RIN: Italian Neuroimaging Network
- RMSE: Root Mean Squared Error

- ROC-AUC: Area Under the Receiver Operating Characteristic Curve
- ROI: Region of Interest
- SD: stable disease
- SHAP: Shapley Additive Explanations
- SMOTE: Synthetic Minority Oversampling Technique
- SOTA: state-of-the-art
- SSIM: Structural Similarity Index Measure
- TGV: Total Gray Volume
- TNBC: Triple-Negative Breast Cancer
- TP: timepoint
- TS: Total Segmentator
- TS-FM: Total Segmentator Fast Mode
- TS-FRM: Total Segmentator Full Resolution Mode
- w.r.t.: with respect to

Introduction

Radiomics has transformed quantitative radiology in oncology since its introduction, around two decades ago. The rationale beneath radiomics in oncology relies on the assumption that quantitative information about morphological and functional tumor descriptors can be extracted from medical images in the various stages of the diagnostic process [104, 112, 148, 177]. These imaging features allow for the investigation of correlations with tumor-related data obtained from biopsies, clinical assessments, and imaging studies. As a result, radiomic features hold potential as predictors of clinical outcomes and as cost-effective, early surrogates of genomic characteristics, ultimately offering prognostic insights and supplementary information within a personalized diagnostic framework. Radiomic analysis has been applied to several types of tumors, like breast [61], oropharyngeal [109], and pancreatic cancer [199], as well as multiple myeloma [160] and brain tumors [158].

From an operational point of view, a radiomic pipeline is made of three subsequent steps: image segmentation (by means of either automatic algorithms or manual delineation), radiomic feature extraction, and data analysis based on the extracted feature values. Feature computation can be performed by using either pattern recognition algorithms or deep learning (DL) models [97, 185, 200]. The data analysis step relies on artificial intelligence (AI) tools that utilize radiomic features to 1) establish correlation with clinical data (e.g., biopsy, visual assessment, genomic information), 2) predict clinical follow-up, and 3) perform patient clustering [73, 90, 98]. However, the reliability of radiomics as a systematic effective predictor or reliable tool for stratification has been

critically discussed [116, 137, 181]. These studies point out several pitfalls, most criticisms underlining that the stratification/prediction effectiveness of radiomics-based models significantly depends on

- the number of patients enrolled in the study, which is often low with respect to (w.r.t.) the requirements of machine learning (ML) approaches;
- the degree of homogeneity of the images acquired for the study, which depends on different acquisition protocols (imaging modality, image spatial resolution, administration of contrast agents, type of sequence and sequence parameters) and inter-scanner discrepancies (different vendors provide different reconstruction algorithms, with customizable reconstruction parameters and consequent possible variations in individual patients);
- the reliability of the computational pipeline generating radiomic features, including segmentation and feature computation steps.

Despite these issues, radiomics has been used for many different tasks, such as prediction of biomarkers, classification of tumor subtype, as well as forecast of recurrence and response to therapy [166, 192], while increasing literature shows that the radiomics approach is rarely examined and investigated from a methodological perspective [14, 91].

The objective of this thesis is to advance the development of methodologies that enhance the reliability and interpretability in predictive applications of radiomics. To this end, this work addresses several critical components of the radiomics pipeline, including image segmentation, the evaluation of feature reliability, explored in terms of stability with respect to segmentation variability and stability of the predictive performance, the development of explainable and interpretable predictive models, and image harmonization. The aim of these studies is to shed light on the potential benefits and drawbacks associated with the usage of radiomics in predictive models, offering insights into their transferability in clinical applications. Each chapter tackles distinct

yet interconnected aspects of the radiomics pipeline, offering methodological advances from varied perspectives and tailored to different clinical purposes.

Chapter 1 focuses on psoas muscle segmentation, as the psoas is an anatomically and clinically relevant structure in the context of sarcopenia assessment. While the gold standard is to estimate variation of psoas muscle mass on a cross-sectional slice of abdominal X-ray computed tomography (CT), we propose the use of three-dimensional numerical schemes for the volumetric segmentation of the muscle, which have never been used for this application, showing the feasibility of applying the proposed methods to real low-dose CT images. Specifically, we introduce a first-order geodesic model that offers a lower computational burden compared to its second-order counterpart and deep-learning models for medical image segmentation. Although not directly addressed in this thesis, the three-dimensional evaluation of psoas volume could support radiomic analysis and therefore facilitate the identification of feature-based biomarkers. This work is published in [136].

Chapter 2 delves into radiomics reliability. Many studies caution against using radiomic features that are sensitive to contouring variability in predictive models for disease stratification, recommending to guide feature selection based on stability with respect to segmentation accuracy. However, the direct impact of segmentation variability on the performance of predictive models remains underexplored. The study here presented investigates how segmentation variability affects both feature stability and predictive performance, through the use of explainable predictive models, focusing on the radiomics-based classification of Triple-Negative Breast Cancer (TNBC) using breast Magnetic Resonance Imaging (MRI). The insights provided by this study carry interesting implications for the field of radiomics, as they challenge the assumption that precise segmentation is always critical for feature selection and predictive performance, providing evidence that, in the reported case, radiomics tolerates segmentation variability without sacrificing prediction quality. This work has been disseminated in preprint form in [26] and utilizes radiomics reliability metrics as defined and published in [25].

Chapter 3 addresses the problem of longitudinal image analysis, which is crucial for tracking temporal changes in medical images, such as lesion progression or anatomical alterations, thereby supporting disease monitoring, treatment evaluation, and outcome prediction. Analyzing longitudinal images poses several challenges, such as managing variable imaging intervals and handling sparse data. State-of-the-art (SOTA) papers on radiomics-based longitudinal prediction use radiomics at baseline timepoint, or combine radiomics extracted from subsequent timepoints, and use it for longitudinal outcome prediction. However, these approaches may fail in cases where only baseline features are available, such as for prospective studies, or in case of missing follow-up timepoints. For this reason, we develop a probabilistic method for longitudinal outcome prediction, with specific application on predicting glioblastoma longitudinal response to treatment. Our method leverages baseline radiomic data in combination with a probabilistic estimate of the response to treatment at the first follow-up to forecast treatment outcomes at subsequent stages, eliminating the need for radiomics data from intermediate timepoints. This novel approach to longitudinal prediction of treatment response consistently improves the longitudinal prediction, even with reduced dataset size. Furthermore, the proposed methodology holds potential for generalization beyond the domains of oncology and radiomics, extending to a broader range of longitudinal applications. This work is published in preprint form in [24].

The last contribution of this thesis concerns image harmonization. In Chapter 2 we investigated the link between radiomics stability and segmentation variability. Beyond segmentation, radiomics features are subject to an additional level of uncertainty that directly arises from unharmonized imaging acquisition, causing variability in the feature extraction step. To ensure a more generalizable and consistent radiomics pipeline, harmonizing the images themselves, rather than features extracted from unstandardized images, could be a more effective strategy to enable robust and reliable analyses. For this reason, in Chapter 4 we tackle the problem of image harmonization. We propose DISARM, an unsupervised Image2Image (I2I) model for the harmonization

of 3D medical images across various scanners, to remove batch effects, with specific application to T1-weighted brain MRI data. Other than transferring images to other scanner domains, DISARM disentangles image content from scanner-specific information, enabling the generation of a scanner-free image, which is the main novelty introduced by this work. Demonstrating competitive performance with a state-of-the-art network in both harmonization effectiveness and downstream tasks, DISARM shows promise as a preprocessing step to mitigate inter-scanner variability in radiomics- or deep learning-based predictive analyses. This work is published in [22].

Overall, the contributions of this thesis aim to strengthen the conceptual and methodological framework of radiomics, introducing original insights and innovations at multiple stages. Specifically, this research represents a step toward refining the radiomics workflow with a methodological focus on radiomics development and understanding, with the final objective of contributing to the advance of radiomics robustness and reliability for clinical applications.

Chapter 1

Segmentation

Medical imaging is a pivotal diagnostic tool in disease assessment. With a variety of diagnostic imaging techniques such as X-ray computed tomography (CT), positron emission tomography (PET), mammograms, and magnetic resonance imaging (MRI), the clinical information embedded in the imaging data is vital for analyzing structural abnormalities, evaluating the loss of muscular mass, detecting pre-cancerous lesions, lesion segmentation, and diagnosis. Segmentation of biomedical images entails the delineation of pathological regions. In the case of oncological lesions, segmentation involves separating different heterogeneous regions of tumor tissues as the necrotic core, enhancing tumor, and peritumoral edema [7]. For neurodegenerative conditions, e.g., amyotrophic lateral sclerosis (ALS), segmentation allows for skeletal muscle mass evaluation and measurement [11], a valuable practice also employed in oncological diseases where muscle loss is linked to poor overall survival [10]. The manual delineation of the regions of interest is a very demanding and skillful process, hence automated and accurate segmentation has great clinical application value and significance, while remaining a challenging task. Automatic segmentation of cancer lesions, for example, is challenged by the low accuracy of delineation of the infiltration area, large intensity and heterogeneity changes, and low boundary contrast [144]. A common DL framework for medical imaging segmentation is based on U-type network structure that

downsamples the images, extracts different levels of feature information, and combines high- and low-level feature information to identify the voxels of interest within the segmentation [149]. Other segmentation methods are based on the transformer structure [110]. However, in the case of medical imaging applications, the use of three-dimensional DL-based algorithms may be hampered by the computational burden and, sometimes, by the lack of sufficiently large datasets, necessary in the training phase of the networks [43]. Therefore, the introduction of three-dimensional deterministic numerical schemes may be particularly useful for this kind of applications.

This chapter introduces the level set models employed in our study [136] for the three-dimensional segmentation of the psoas muscle, along with the numerical schemes used for their discretization. We present the classical evolution model for segmentation, a standard second-order geodesic model, and a novel first-order modification of the geodesic model that we formulated in [136]. The evaluation of these methods on the segmentation of the psoas muscle serves as an empirical evaluation aimed at demonstrating the feasibility of applying the proposed approaches to real clinical cases, as the psoas is an anatomically and clinically relevant structure in the context of sarcopenia assessment. Particular attention is given to the performance of the first-order model.

The chapter is organized as follows. Section 1.1 outlines the motivation behind the application of these methods to psoas muscle segmentation in low-dose CT images. Section 1.2 presents the evolution models and details the numerical schemes for their discretization. Additionally, the pre- and post-processing pipelines are reported. Finally, Section 1.3 compares the results of the different numerical approaches when applied to clinical CT data, with further discussion provided in Section 1.4.

1.1 Motivation

The psoas muscle holds considerable clinical relevance due to its anatomical position and functional roles in movement and postural regulation. Due to its origin on the lumbar vertebrae, the psoas contributes significantly to lumbar spine stabilization. It helps maintaining the integrity of the lower back during movement and while standing upright, facilitating movements such as walking. Psoas functionality is critical for nearly all lower body motions. Alterations in psoas muscle, such as diminished strength or anatomical asymmetry, have been linked to various musculoskeletal pathologies, among which sarcopenia is a prominent example [51]. Consequently, evaluation of the psoas muscle mass is useful as both a diagnostic and a prognostic indicator in the assessment of this condition, due to its strong correlation with whole-body skeletal muscle mass [68].

Sarcopenia is a pathological condition of the skeletal muscle characterized by loss of mass and depletion of strength and physical performance [44]. Several studies have shown that sarcopenia is associated to specific chronic diseases. Low skeletal muscle mass is associated with post-transplantation death [83, 122], it is a risk factor for poor performances and toxicity of chemotherapy in cancer patients [33, 41, 134, 189, 191], and it has a high probability of co-occurrence with several neurodegenerative conditions such as Parkinson's disease and multiple sclerosis [53, 204]. Despite its relevance as a frailty biomarker, sarcopenia is difficult to quantitatively determine. Most studies in this field estimate variation of psoas muscle mass on specific slices of low resolution abdominal CT scans [51]. The gold standard is to select a cross-sectional image in correspondence with the third lumbar vertebra (L3), to identify the psoas in this image, and to compute the corresponding psoas area and density. To this end, different approaches for automated two-dimensional psoas segmentation have been proposed. For example, anatomy-based manually extracted prior shapes have been used to initialize energy-based iterative schemes [32, 72, 84]. More recently, convolutional neural networks (CNNs)

and Generative Adversarial Networks (GANs) have been tailored to realize psoas segmentation at L3 on CT images [54, 68, 75, 188]. Also, DL algorithms generally designed for muscles and internal organ segmentation [85, 190] can be applied to the specific task of psoas identification.

In this context, the image-based approach to sarcopenia quantification relies on the controversial assumption that a single CT slice is representative of the whole psoas condition [10, 119, 211]. Only few studies have addressed sarcopenia assessment within a three-dimensional segmentation framework [35, 57], the latter one employing MRI rather than CT scans. Both works propose deep learning-based methods for the automated quantification of psoas muscle volume. In particular, the second study highlights the challenges associated with accurately isolating the psoas muscle in a three-dimensional context. To address this, the authors segment the psoas major, the iliacus, and the psoas minor muscles altogether, acknowledging the inherent difficulty of consistently delineating the entire psoas muscle in three-dimensional medical images. Two other studies [10, 11] showed that a complete three-dimensional segmentation of the psoas muscle in CT images of oncological and neurological patients may be exploited to obtain metabolic information about the diseased tissue. Both studies utilized hybrid data made of low-dose CT and PET of patients injected with 18F-Fluorodeoxyglucose (FDG-PET/CT). The segmentation mask obtained on the co-registered CT was multiplied voxel-wise with the FDG-PET volumes in order to extract the metabolic information associated to the psoas tissue. However, the reliability of these metabolic outcomes strongly depend on the accuracy of the segmentation step. In the first study [10], the segmentation approach was based on an extended version of the Hough transform [13], which can be used only if a realistic parametric model for the muscle is at disposal, while the second study [11] used a heuristic approach based on histogram equalization and an α -shape algorithm to identify the region corresponding to the inner muscle.

Despite the increasing potential relevance of this kind of analysis for diagnostic/prognostic applications, no systematic study has been performed



Figure 1.1: Example of abdominal CT slices.

so far on the effectiveness of three dimensional numerical schemes for image segmentation of the psoas muscle in low resolution CT data. Clinical low-dose CT images from PET/CT scanners are generally noisy and lower in resolution compared to high-resolution CT scans. Moreover, the topographical complexity of the psoas district, where organs and muscles with similar Hounsfield Unit (HU) values frequently touch, makes defining its boundaries particularly challenging (see Figure 1.1 for an example of abdominal CT slices).

In the following we describe the computational problem of segmenting the psoas muscle in clinical low-dose CT data, from hybrid PET/CT imaging, by means of three-dimensional numerical schemes originating from the level set methodology [102, 114, 132, 135, 154]. The use of three-dimensional methods is appropriate for assessing sarcopenia because volumetric segmentation naturally exploits the information coming from adjacent regions, leading to improved accuracy of tissue density evaluation [43, 117]. We focused on gradient-based methods that search for an optimal solution via the minimization of an objective function [46]. These methods can incorporate weighted curve length and variation of intensity of the segmented regions in their objective function [71], and can be combined with deep learning techniques [40]. In this study we considered a classical evolution model, a standard second-order geodesic model, and a new first-order modification of the geodesic model. We designed this latter method to retain all the convergence properties of the second-order model, while being computationally lighter, making it potentially more advantageous for the segmentation of 3D medical images. The discretization of these three models was performed by means of a standard

finite-difference scheme. We obtained three segmentation algorithms that require the user to select one point for each psoas in the data volume to initialize the segmentation process. Additionally, we implemented an algorithm for mitigating the artifacts that could affect the segmentation result due to the low resolution of clinical low-dose CT.

1.2 Materials and methods

1.2.1 Evolution models

The general framework considered in the present study assumes that segmentation is realized by letting a three-dimensional wave-front

$$\gamma_t := \{x \in \mathbb{R}^3 : v(t, x) = 0\} \quad (1.1)$$

evolve from an initial point inside the shape to segment until it reaches its boundary. Following the level set method [132], the field $v = v(t, x)$ is assumed to be the solution of the Cauchy problem

$$\begin{cases} v_t + g(x)|Dv| = 0, & (x, t) \in \mathbb{R}^3 \times (0, T) \\ v(0, x) = v_0(x) & x \in \mathbb{R}^3, \end{cases} \quad (1.2)$$

where D is the three-dimensional gradient operator, $g(x)$ is the speed in the normal direction and v_0 is a proper representation of the front γ_0 (i.e. denoting by Ω_0 the region enclosed by the front, v_0 has to be equal to 0 on γ_0 , positive outside Ω_0 and negative inside), and the front γ_t is characterized as the 0-level set of the solution v of the Cauchy problem (1.2). It is well known that, under appropriate assumptions on the velocity field $g(x)$ and the initial condition v_0 , the viscosity solution of (1.2) exists and is unique (see, e.g., [42]). In the following we refer to the eikonal equation (1.2) as the classical model. This approach can be generalized to a second order Hamilton-Jacobi equation of

the form

$$v_t + H(x, Dv, D^2v) = 0, \quad (1.3)$$

which includes more sophisticated definitions of the velocity field g , possibly dependent on the curvature of the surface, as for example in the Mean Curvature Motion. Here we consider the version proposed by Caselles et al. in [28], denominated geodesic model

$$H(x, Dv, D^2v) = -g(x)|Dv| \left(\varepsilon \nabla \cdot \left(\frac{Dv}{|Dv|} \right) - \mu \right) - \eta Dg \cdot Dv, \quad (1.4)$$

where ε, μ, η are positive parameters to appropriately tune. To give a better insight on the choices made in the numerical section below, we recall that these parameters are used to properly weigh the different components of the model. For instance, η is related to the transport driven term, which attracts the front towards the boundaries of the image, and ε is related to the curvature part that regularizes the solution and tends to shorten the front. Finally, the constant velocity μ is added to balance the shrinking effect of the curvature term and allows outward evolution, as in our case, where the front expands from an initial sphere towards the surface of the muscle. We decided to give a normalized value to the constant $\mu = 1$ and adapted the other parameters accordingly. More precisely, the higher the value of η , the slower the expansion toward the boundary, whereas high values for the curvature term enforce the use of restrictive Courant-Friedrichs-Lewy (CFL) conditions to keep the finite difference scheme stable.

The model (1.4) gives the desired result at stationary state, that is for $t \rightarrow +\infty$, but is usually rather heavy if solved numerically, especially when 3D numerical schemes are employed. Therefore we propose a first-order modification of equation (1.4), i.e.,

$$H(x, Dv) = \mu g(x)|Dv| - \eta Dg \cdot Dv. \quad (1.5)$$

This simplified formulation maintains the key characteristics of the second-

order counterpart, though it compromises slightly on smoothening to allow for a lower computational burden.

In order to complete the description of the segmentation model we have to define the velocity $g(x)$ such that it forces the front towards the relevant boundary and then it stops its evolution. A typical definition in gradient-based methods is the following

$$g(x) = \frac{1}{(1 + |\nabla(G * I(x))|^p)}, \quad p \geq 1, \quad (1.6)$$

where I represents the intensity values of the image, and G a Gaussian filter of deviation σ , used to regularize noisy images. This function has the desired properties for edge detection; in fact, it takes values in $[0, 1]$, and it is decreasing with respect to the gradient of the image I . If a spike in the gradient is present, then g is close to 0; otherwise it is equal to 1 in smooth regions. In practice, the filtering process is done once before the evolution, and the values of the smoothed image $G * I$ are used instead of I in the computation of the gradient. High values for the parameter p may enlarge the region where the function is close to 0, reducing the overall accuracy of the scheme. This is why here we used (1.6) with $p = 2$ as in [28].

1.2.2 Numerical schemes for model discretization

We now provide some details about the numerical schemes utilized for the discretization of the first order equations (1.2) and (1.5), which can also be adapted to equation (1.4) with minor changes. For sake of simplicity, the description of these schemes is provided in the two-dimensional setting, although the codes used for data analysis have been implemented for three-dimensional segmentation.

Let $(x_j, y_i) = (j\Delta x, i\Delta y)$, $j, i \in \mathbb{Z}$ be a uniform grid in space, and $t_n = t_0 + n\Delta t$, $n \in [0, N_T]$, with $(N_T - 1)\Delta t < T \leq N_T\Delta t$, a grid in time. We denote the

updating scheme as

$$\begin{cases} u_{i,j}^{n+1} = S_{\Delta}(u^n)_{i,j}, & n \geq 0, \\ u_{i,j}^0 = v_0(x_j, y_i), \end{cases} \quad (1.7)$$

where $u_{i,j}^n = u(x_j, y_i, t_n)$ is an approximation of the solution of the equations and $\Delta = (\Delta x, \Delta y, \Delta t)$. We remark that the convergence of numerical schemes (1.7) for (1.2), and more generally for (1.3), follows from the Barles-Souganidis theorem [9], requiring monotonicity, stability, and consistency of the schemes. Following [105], we implemented the numerical discretization of all three models by means of the standard difference form

$$u_{i,j}^{n+1} = S_{\Delta}(u^n)_{i,j} := u_{i,j}^n - \Delta t h \left(x_j, y_i, D_x^- u_{i,j}^n, D_x^+ u_{i,j}^n, D_y^- u_{i,j}^n, D_y^+ u_{i,j}^n \right), \quad (1.8)$$

where the numerical Hamiltonian is the local Lax-Friedrichs Hamiltonian

$$\begin{aligned} h(x, y, p^-, p^+, q^-, q^+) := & H \left(x, y, \frac{p^+ + p^-}{2}, \frac{q^+ + q^-}{2} \right) \\ & - \frac{\alpha_x(p^-, p^+)}{2} (p^+ - p^-) - \frac{\alpha_y(q^-, q^+)}{2} (q^+ - q^-), \end{aligned} \quad (1.9)$$

with $p^{\pm} = D_x^{\pm} u_{i,j}^n$, $q^{\pm} = D_y^{\pm} u_{i,j}^n$, and

$$\alpha_x(p^-, p^+) := \max_{\substack{x,y,q, \\ p \in I(p^-, p^+)}} |H_p(x, y, p, q)|, \quad \alpha_y(q^-, q^+) := \max_{\substack{x,y,p, \\ q \in I(q^-, q^+)}} |H_q(x, y, p, q)|, \quad (1.10)$$

with $I(a, b) := [\min(a, b), \max(a, b)]$. We point out that $h(x, y, p^-, p^+, q^-, q^+)$ is a Lipschitz continuous function, with

$$D_x^{\pm} u_{i,j}^n := \pm \frac{u_{i,j \pm 1}^n - u_{i,j}^n}{\Delta x}, \quad (1.11)$$

and

$$D_y^{\pm} u_{i,j}^n := \pm \frac{u_{i \pm 1, j}^n - u_{i,j}^n}{\Delta y}. \quad (1.12)$$

Further, in this context consistency holds via the condition

$$h(x, y, a, a, b, b) = H(x, y, a, b), \quad (1.13)$$

whereas monotonicity requires the numerical Hamiltonian h to be non-decreasing with respect to its third and fifth argument and non-increasing with respect to the fourth and the sixth ones. The resulting scheme (1.8)-(1.9) is monotone under the CFL condition $\frac{\Delta t}{\Delta x} \cdot \alpha_x + \frac{\Delta t}{\Delta y} \cdot \alpha_y \leq 1$.

Moreover, as suggested in [133], this scheme can also be used to solve equation (1.4) by first splitting the Hamiltonian into two components

$$H(x, y, Dv, D^2v) = H_0(x, y, Dv) + H_1(k(x, y), Dv),$$

where H_0 includes the eikonal and advection term, whereas H_1 accounts for the dependence on the curvature $k(x, y)$, which is computed as

$$k(x, y) = \nabla \cdot \left(\frac{Dv(x, y)}{|Dv(x, y)|} \right) = \frac{v_{xx}v_y^2 - 2v_xv_yv_{xy} + v_{yy}v_x^2}{(v_x^2 + v_y^2)^{3/2}}.$$

Then, it is enough to discretize the first-order part using the Lax-Friedrichs scheme and H_1 with simple centered finite difference approximations. In this case, the stability of the scheme requires a parabolic-type CFL condition of the form $\Delta t \sim \Delta x^2$.

The initialization of the scheme requires the user to select two points, one for each muscle. These points become the centers of the two spheres of radius $r = 5\Delta x$ used as initial condition for the evolution, which is executed separately for each psoas. The number of iterations required to achieve convergence was determined via grid search.

1.2.3 Pre-processing and post-processing

Before solving the evolution equations, we pre-processed the CT data according to the following process.

1. Hounsfield normalization: given that normal bone tissue in CT has around 160 HU, all HU values above 120 have been set to 0, in order to reduce the relative weight of the bones [141].
2. For image enhancement purposes, a histogram equalization algorithm has been applied, which performed a re-mapping of the image based on the probability distribution of the input gray levels [86]. More specifically, histogram equalization is a contrast enhancement technique that aims to redistribute the intensity values of an image such that the resulting histogram is approximately uniform. This process is particularly effective for improving the visual quality of images with low contrast, where pixel intensities are concentrated within a limited range. The method operates by computing the cumulative distribution function (CDF) of the original image histogram and then applying a transformation that maps each pixel intensity to a new value proportional to the CDF. This ensures that gray levels with a high population of pixels are spread across a broader intensity range, thereby enhancing global contrast. After histogram equalization, the output intensities are redistributed so that each gray level (roughly) corresponds to an equal portion of the pixel population. As a result, the transformed image exhibits a more uniform distribution of intensities, allowing for improved visibility of structural details and features that may have been obscured in the original image.
3. A Gaussian filter with standard deviation σ has been applied to improve the signal-to-noise ratio of each image [130]. The filtering procedure was applied in a heuristic manner. Initially, no filtering was applied; if the segmentation result obtained with unfiltered image was deemed satisfactory, the image was preserved without alteration. Instead, a

Gaussian filter was employed in cases where the segmented profile exhibited physiologically implausible characteristics, such as substantial deviation beyond the anatomical boundaries of the muscle. In such cases, Gaussian filters with σ values of 1 and 2 were applied incrementally. To maintaining the visual fidelity of the original image, we only explored σ values of 1 and 2. While a small σ yields minimal smoothing and preserves fine structural details, larger σ values attenuate not only noise but also small intensity variations. This results in more stable gradients; however, it can also lead to reduced edge localization accuracy and diminished boundary precision in segmentation tasks.

As previously remarked, low-dose CT images from PET/CT scans are typically noisy and exhibit low resolution. Moreover, the complexity of human body in the psoas region often implies the presence of adjacent regions with similar HU values. Therefore, the definition of a clear boundary of the psoas becomes an intricate issue, which impacts the reliability of the numerical scheme in automatically segmenting the image. Specifically, in some cases the wave-front could leave the correct boundary and start detecting objects outside the muscle of interest, thus making the stopping criterion ineffective, and providing unreliable results.

In order to solve this issue, we devised a post-processing routine able to clean the result from most of the external objects and eventually provide an acceptable final result. The procedure is composed of two steps and has a very low computational cost, which makes it efficient even if more than one iteration is needed. While the numerical schemes used in this study are fully three-dimensional, this post-processing routine works slice-wise, while leveraging information from adjacent slices to take advantage of the problem's inherent spatial continuity. In this way, it simulates manual slice-by-slice correction to refine the segmentation, targeting spurious voxels that extend beyond the true boundaries of the psoas muscle. The procedure is summarized in *Algorithm 1*, while more details on the implementation are given in the following items.

- **First step.** The aim of this step is to reduce the number of connected

components (CCs) in each slice, detecting the one belonging to the psoas and discarding the remaining ones. To do so, the algorithm starts from the slice containing the center chosen by the user and selects the correct CC by trivially excluding the ones that do not include the chosen point. Then, the slices are scanned toward the top and the bottom starting from the center. The main loop consists in computing the centroid of the CC detected in the previous slice, and then, in the current slice, selecting the CC containing such point. The result is a segmented muscle in which only the spurious objects directly connected to the boundary may still be present.

- **Second step.** In the second step, a vector tol is initialized with length equal to P , the set of segmented voxels in the starting slice, where the i -th entry is the minimum distance (computed using voxel indices) between voxel $i \in P$ and the set of voxels in the previous slice P_{prev} , incremented by a fixed constant of 0.75. Then the algorithm loops over the slices, setting $P_{prev} = P$ and first computing the difference set $P_{diff} = P - P_{prev}$, where P and P_{prev} are the sets of voxels in the current and previous slices, respectively; then, for each point $j \in P_{diff}$, the algorithm computes the set distance $dist(j, P_{prev})$ with respect to P_{prev} . This distance is obtained in correspondence of a voxel $i_{min} \in P_{prev}$. If $dist(j, P_{prev}) > tol(i_{min})$, voxel j is excluded from P and hence removed from the segmentation. After all voxels in P_{diff} have been scanned and removed if needed, a new tolerance vector tol is computed using the current cleaned set of voxels P and P_{prev} .

Basically, according to the second step, the distances between the voxels of the current slice and those of the previous slice determine the tolerance criterion for the inclusion of voxels from the next slice. For each potentially spurious voxel to evaluate on the next slice, an extra tolerance margin of 0.75 is allowed in addition to the minimum distance between the closest voxel to the spurious one and the previous slice. The second step, in its initialization phase, assumes that the starting slice and the previous one (according to the scanning direction)

do not need any cleaning. If this is not the case, a second iteration of this step is needed, using a different starting point.

The MATLAB (R2021a) implementation of the three numerical schemes and the cleaning procedure is available at: <https://github.com/theMIDAGroup/psoas>. Experiments and analyses were conducted on a personal computer equipped with an Apple M1 Pro chip and 16 GB of unified memory.

Algorithm 1 Cleaning procedure

Input: final solution u of the segmentation method

Step 1: reduce the number of CCs in each slice

Initialization: Start from the center and select the CC containing it

Main loop: Scan the slice in both directions and keep the CC containing the centroid of the previous slice

Step 2: clean the remaining spurious objects

Initialization: Set an initial slice S_k . List all the voxels in the segmentation of that slice in the set P and all the voxels in the segmentation of the previous slice S_{k-1} in the set P_{prev} . Initialize a vector tol with length equal to cardinality of P . The value of the i -th entry of this vector is given by the minimum distance of voxel $i \in P$ from all the voxels in P_{prev} , plus a fixed value of 0.75.

Main loop on the slices:

1. set $P_{prev} = P$
 2. set $S_k = S_{k+1}$ and list all the voxels in the segmetation of slice S_k in the set P
 3. compute the set difference $P_{diff} = P - P_{prev}$
 4. for each $j \in P_{diff}$:
 - compute its distance $dist(j, P_{prev})$ w.r.t. the set P_{prev} and save the point i_{min} in P_{prev} of minimum distance
 - remove the point j from P if $dist(j, P_{prev}) > tol(i_{min})$
 5. update the tol vector, computing it as in the initialization step.
-

1.2.4 Data collection

The pipeline made of pre-processing, numerical segmentation, and post-processing was applied to low-dose CT images (see Figure 1.2) acquired from nine subjects belonging to a retrospective dataset of patients recruited at the IRCCS Ospedale Policlinico San Martino, Genova, Italy. All subjects underwent FDG-PET/CT and the study was performed according to the Declaration of Helsinki, Good Clinical Practice, and local ethics regulation. All enrolled patients signed a written informed consent at the time of the exam, encompassing the use of anonymized data for retrospective research purposes. In addition, CT of six subjects were collected from the Cancer Moonshot Biobank - Prostate Cancer Collection (CMB-PCA) [27, 38].

Various dataset have been used in literature for the psoas segmentation, performed either in a two- or three-dimensional manner. All these studies propose pipelines for psoas volume evaluation aimed at assessing sarcopenia from standard abdominal CT scans, with the goal of enabling opportunistic screening for frailty. However, a key limitation of such opportunistic datasets is the absence of manual or reference segmentations. In fact, some studies use data where psoas muscle is visible but not the intended target of the imaging procedure [35]. As a result, the psoas muscle must be manually annotated to create training data for supervised segmentation methods. For example, the work [75] uses diverse datasets from The Cancer Imaging Archive (TCIA) [38], including head and neck cancer CT scans [63] ($n = 90$), colon cancer scans [171] ($n = 674$), and a liver tumor dataset ($n = 96$) from the LiTS segmentation challenge [17] for a two-dimensional segmentation at L3 level, based on a U-Net architecture. Another example of paper using a public dataset is [57], using MR images from the UKBB Abdominal Protocol [106] for 3D segmentation of iliopsoas muscle. Additionally, some datasets available on TCIA are annotated for various tasks such as organ segmentation (140 CT scans) [147], muscular tissue segmentation (882 CT scans) [92], and lymph node detection (176 CT scans) [150]. However, specific annotations for the psoas muscle are not included and would need to be manually

added for targeted segmentation tasks. Lastly, the paper [195] deals with the segmentation of 104 anatomic structures in CT images, providing training images with two modalities, CT <https://zenodo.org/records/10047292> and MRI <https://zenodo.org/records/14710732>, along with segmentations of each district, including the psoas muscle. Their dataset is freely available to download.

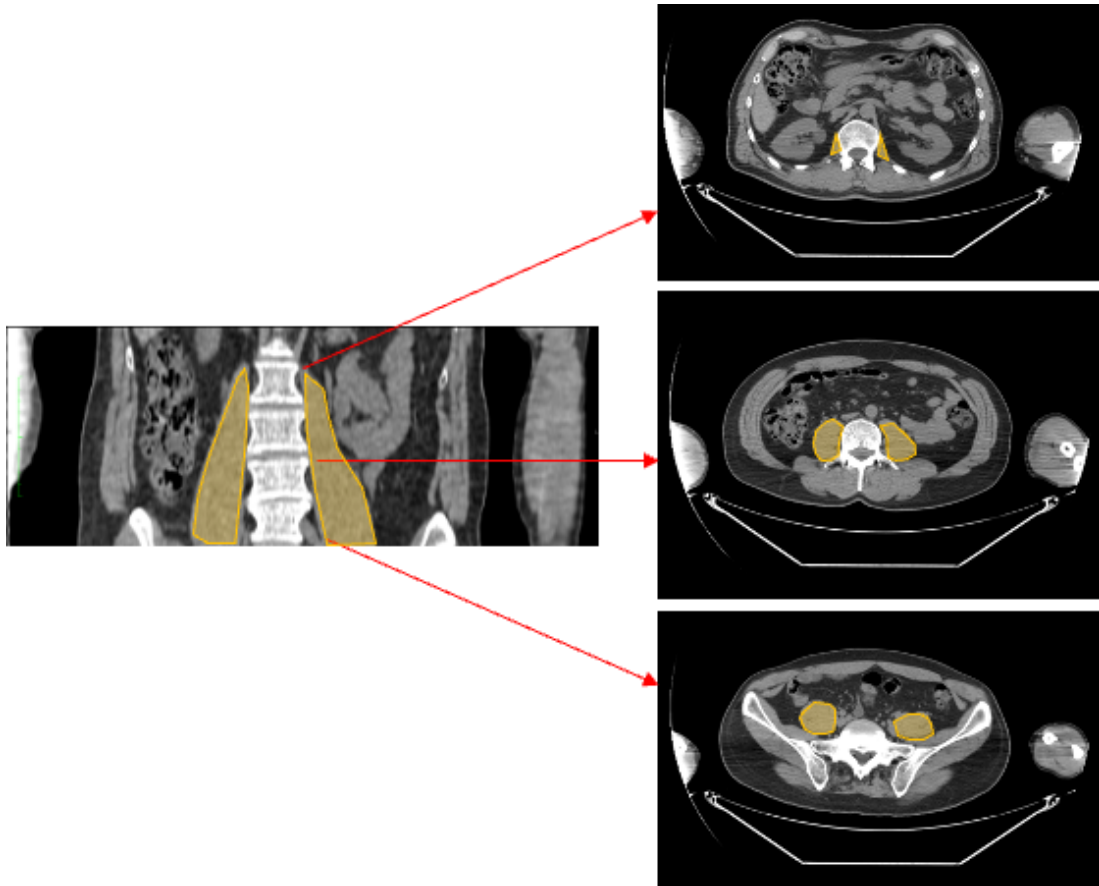


Figure 1.2: Coronal view of the psoas muscles (in yellow, left) and the corresponding axial views at different levels of the abdomen (right).

1.2.5 Benchmark and evaluation metrics

To benchmark the results obtained through the proposed segmentation methods, the 3D Slicer extension TotalSegmentator (TS) [196] was used for psoas segmentation. This is a fully automated total body segmentation deep learning routine based on nnU-Net [74], which has been applied with reliable results for the segmentation of abdominal tissues in the three-dimensional setting [66, 151, 152]. In our analysis, TS was used in both fast (TS-FM) and full-resolution (TS-FRM) modes. A manual segmentation of the psoas muscles was also performed for comparison with the proposed methods.

In order to evaluate the agreement between the manually drawn psoas profiles and the segmentations obtained with the three numerical schemes and the TotalSegmentator tool we used six metrics: the Dice Similarity Coefficient (DSC) [52], the Jaccard index [76], the Hausdorff distance [69], the Average Symmetric Surface Distance (ASSD) [70], the Surface Dice, and the Robust Hausdorff distance (Robust Hausdorff). Surface Dice and Robust Hausdorff have been computed via the packages provided by <https://github.com/google-deepmind/surface-distance/tree/master>. The DSC and the Jaccard index vary between 0 and 1 where 0 represents an empty intersection between the two segmentations while 1 a perfect overlap. The Hausdorff distance and ASSD are indeed distances, so smaller values correspond to closer objects. Specifically, the Hausdorff distance is the greatest distance from a point in one segmentation mask to the closest point in the other mask, while ASSD measures the average of all distances from points on the boundary of the first mask to the boundary of the second one, and viceversa. Similarly to the classical volumetric DSC, the Surface Dice measures the overlap of two surfaces, where a surface element is counted as overlapping when the closest distance to the other surface is less than or equal to the specified tolerance, fixed at 3 mm. The range of the Surface Dice is $[0, 1]$. Robust Hausdorff is computed like the classical Hausdorff distance but it uses a fixed percentile (75th) of the distances instead of the maximum distance.

1.3 Results

The CT images of the patients involved in the study were segmented through the finite difference scheme (1.8)-(1.12) applied to model (1.2) ('classical' from now on), and to the geodesic models (1.4) and (1.5) ('GMFD 1ord' and 'GMFD 2ord' from now on, respectively). For both GMFD 1ord and GMFD 2ord we set $\mu = 1$ and $\eta = 0.25$. For the second order model, $\varepsilon = 0.05$. Furthermore, GMFD 1ord exploits the linear CFL condition $\Delta t = \lambda \Delta x$, with

$$\lambda = (3\|g\| + \|D_x g\| + \|D_y g\| + \|D_z g\|)^{-1}, \quad (1.14)$$

and GMFD 2ord uses the parabolic CFL condition $\Delta t = 0.25\Delta x^2$. For all simulations the spatial discretization is set to $\Delta x = 0.1$ cm.

Table 1.1 contains the number of iterations and CPU time employed to realize the segmentation of the nine volumes via the three numerical schemes. Figure 1.3 contains the results of the segmentation provided by the three methods for six subjects, while Figure 1.4 reports the last three cases, presenting the most significant amount of artifacts, and shows the performance of the post-processing algorithm on these subjects. The segmentations in Figure 1.3 and Figure 1.4 show that the run of the numerical schemes without any post-processing may imply either the persistence of artifacts, typically in the region between the two psoas muscles, or a slight shrinking of the psoas volume. Figure 1.5 shows the segmentation of the psoas on a subject from the CMB-PCA dataset.

The segmentations obtained with the three numerical schemes and TS were compared with the manual segmentations. Table 1.2 reports the average values and standard deviations of the metrics computed across the nine internal subjects, while Table 1.3 shows the metrics obtained for the CMB-PCA dataset.

Table 1.1 indicates that the computational burden required by the classical and GMFD 1ord schemes is an order of magnitude lower than that of GMFD 2ord. DL segmentation in fast mode (TS-FM) is generally completed in the lowest computational time, an order of magnitude lower than the two fastest

level set schemes, even though TS-FM provides lower segmentation reliability (see Tables 1.2 and 1.3).

Table 1.1: Number of iterations, CPU time (s), values of p and σ , and number of slices for the three segmentation algorithms, for the nine internal CT volumes.

Method	Patient	N_{max}	CPU	p	σ	Slices
classical	CTTO01	400	178.48	2	1	49
GMFD 1ord	CTTO01	400	197.59	2	1	49
GMFD 2ord	CTTO01	1900	1955.84	2	1	49
classical	CTTO02	350	133.19	2	0	45
GMFD 1ord	CTTO02	350	147.56	2	0	45
GMFD 2ord	CTTO02	2000	1870.16	2	0	45
classical	CTTO03	450	232.91	2	1	57
GMFD 1ord	CTTO03	450	250.65	2	1	57
GMFD 2ord	CTTO03	2700	3298.88	2	1	57
classical	PZGE01	700	229.96	2	2	33
GMFD 1ord	PZGE01	700	234.94	2	2	33
GMFD 2ord	PZGE01	3500	2654.22	2	2	33
classical	PZGE02	700	232.67	2	2	35
GMFD 1ord	PZGE02	700	246.20	2	2	35
GMFD 2ord	PZGE02	3500	2787.62	2	2	35
classical	PZGE03	700	207.53	2	2	31
GMFD 1ord	PZGE03	700	224.22	2	2	31
GMFD 2ord	PZGE03	3500	2499.52	2	2	31
classical	PZTO01	500	258.00	2	0	57
GMFD 1ord	PZTO01	500	275.68	2	0	57
GMFD 2ord	PZTO01	2500	2991.82	2	0	57
classical	PZTO03	500	229.58	2	0	52
GMFD 1ord	PZTO03	500	250.62	2	0	52
GMFD 2ord	PZTO03	2500	2758.94	2	0	52
classical	PZTO04	450	207.79	2	1	51
GMFD 1ord	PZTO04	450	230.97	2	1	51
GMFD 2ord	PZTO04	2300	2540.63	2	1	51



Figure 1.3: Segmentations provided by the classical evolution model (1.2) (first column), GMFD 1ord (1.5) (second column), and GMFD 2ord (1.4) (third column). Top to bottom: subject CTTO02 (45 slices), subject PZGE01 (33 slices), subject PZGE02 (35 slices), subject PZGE03 (31 slices), subject PZTO03 (52 slices), subject PZTO04 (51 slices). Figure continues on the next page.

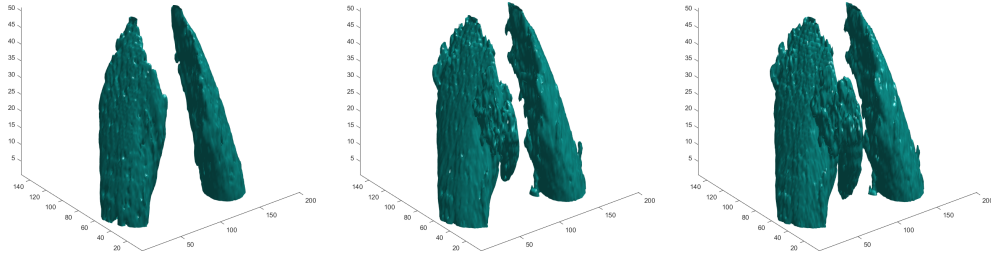


Figure 1.3: Figure continued from previous page.

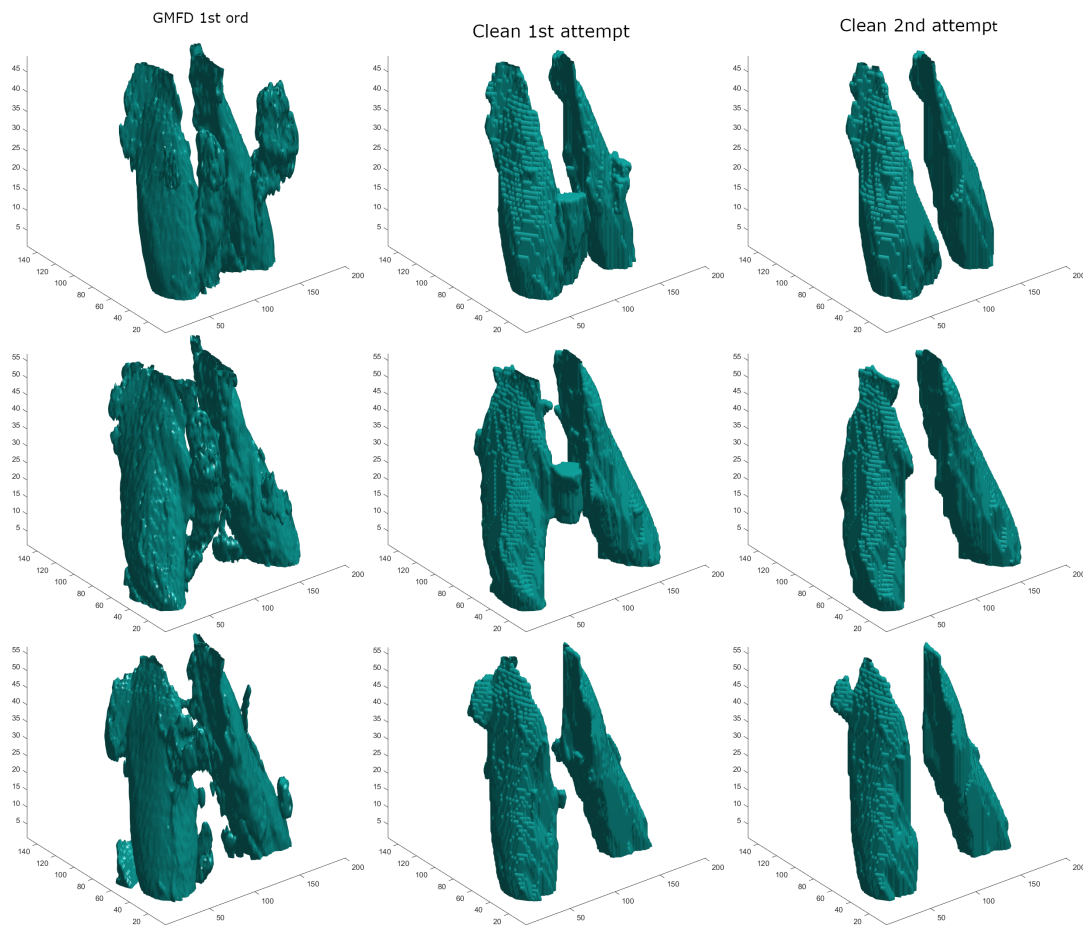


Figure 1.4: Post-processing pipeline for the patients CTTO01, CTTO03, and PZTO01 when applied to the outcomes of GMFD 1ord. First column: starting data. Second column: outcome of the first step. Third column: outcome of the second step.

Table 1.2: Evaluation of the segmentation results produced by the three three-dimensional numerical schemes (classical, GMFD 1ord, and GMFD 2ord) and TotalSegmentator 3D in both fast and full-resolution mode (TS-FM and TS-FRM) when compared to manual segmentation on the internal dataset. Mean score $\pm$ standard deviation.

Metric	classical	GMFD 1ord	GMFD 2ord	TS-FM	TS-FRM
Dice	0.86 ± 0.03	0.87 ± 0.02	0.87 ± 0.03	0.83 ± 0.02	0.85 ± 0.05
Surface Dice	0.88 ± 0.05	0.89 ± 0.02	0.88 ± 0.04	0.83 ± 0.03	0.84 ± 0.07
Jaccard	0.76 ± 0.05	0.77 ± 0.04	0.77 ± 0.04	0.72 ± 0.03	0.75 ± 0.07
Hausdorff (mm)	12.15 ± 3.85	12.95 ± 2.96	12.64 ± 3.76	29.18 ± 25.95	29.55 ± 15.58
Robust Hausdorff (mm)	2.12 ± 0.71	2.00 ± 0.41	2.01 ± 0.52	2.86 ± 0.43	2.77 ± 0.95
ASSD (mm)	0.40 ± 0.10	0.40 ± 0.04	0.42 ± 0.08	0.64 ± 0.35	0.66 ± 0.38

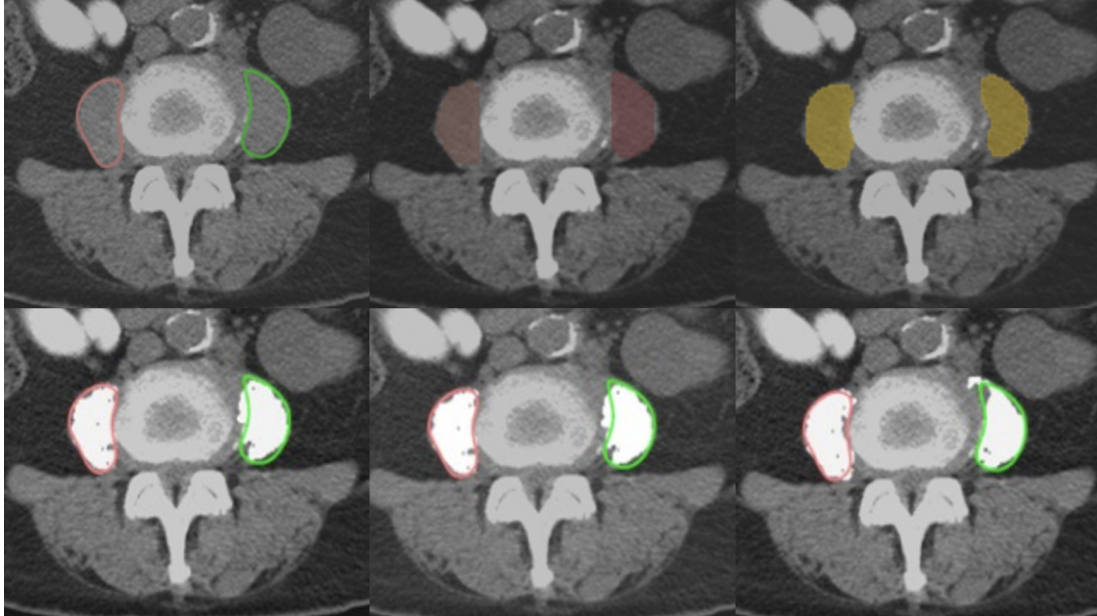


Figure 1.5: Segmentations obtained for a representative subject from the CMB-PCA dataset. First row, from left to right: manual segmentation, TS-FM, and TS-FRM segmentations. Second row, from left to right (in white): classical, GMFD 1ord, and GMFD 2ord segmentations, with the reference manual segmentation superimposed.

Table 1.3: Evaluation of the segmentation results produced by the three three-dimensional numerical schemes (classical, GMFD 1ord, and GMFD 2ord) and TotalSegmentator 3D in both fast and full-resolution mode (TS-FM and TS-FRM) when compared to manual segmentation on CMB-PCA dataset. Mean score $\pm$ standard deviation.

Metric	classical	GMFD 1ord	GMFD 2ord	TS-FM	TS-FRM
DSC	0.84 ± 0.01	0.84 ± 0.01	0.84 ± 0.02	0.86 ± 0.02	0.86 ± 0.02
Surface Dice	0.84 ± 0.05	0.84 ± 0.04	0.84 ± 0.07	0.87 ± 0.06	0.88 ± 0.05
Jaccard	0.72 ± 0.20	0.73 ± 0.02	0.72 ± 0.06	0.75 ± 0.03	0.76 ± 0.03
Hausdorff (mm)	20.56 ± 9.60	20.56 ± 9.60	21.77 ± 3.76	46.03 ± 15.37	48.21 ± 15.16
Robust Hausdorff (mm)	2.40 ± 0.63	2.31 ± 0.60	2.42 ± 0.93	2.29 ± 0.75	2.27 ± 0.67
ASSD (mm)	0.54 ± 0.18	0.53 ± 0.17	0.56 ± 0.25	0.88 ± 0.41	0.93 ± 0.43

1.4 Discussion

In the presented numerical schemes, we tuned the parameters to balance accuracy and stability during segmentation. The curvature parameter ε was set to 0.05, introducing only mild regularization. This choice avoids excessive smoothing of the boundaries, which could otherwise result in over-resolved or artificially flattened segmentations that do not reflect the true anatomy of the psoas muscle. The transport term η was set to 0.25, allowing the evolving contour to move slowly and cautiously. This prevents the segmentation from leaking beyond the actual boundary of the muscle, maintaining anatomical accuracy. Finally, the velocity term μ was set to the normalized value of 1, defined relative to the other parameters. The choice of parameters within the same order of magnitude as those adopted here tend to produce similar outcomes in terms of segmentation quality. The choice of the spatial discretization parameter Δx is related to the voxel size (1 mm per side), whereas the initial spherical shape is defined with a radius that is reasonably contained within the volume of an adult psoas muscle. Larger initial shapes would evolve more quickly toward the boundary, as they are initially closer. However, since the evolution proceeds outward, such larger initial configurations might fail to accurately capture the muscle margins. Regarding the number of iterations required to achieve convergence, a fixed and sufficiently large $N_{\max}$ could have been applied uniformly across all patients. Nevertheless, table 1.1 reports the

smallest $N_{\max}$ value that yielded satisfactory results for each case, determined via grid search. Additionally, the second step of the post-processing procedure includes the tolerance parameter of 0.75, which approximately corresponds to half the diagonal length of a pixel with 1 mm isotropic resolution, serving as a tolerance margin for when we are considering whether to include a spurious voxel or not. This parameter could either be relaxed, with the risk of allowing more false positives, or restricted, with the consequent possibility of cutting out valid parts of the anatomy that are slightly misaligned due to psoas shape variations. The tuning of these parameter was done on a test patient and extended to the others, showing satisfying results for the purposes of the study.

The two implemented geodesic models provide similar segmentation results, while the classical scheme produces over-resolved shapes, see for example patient CTTO02 (Figure 1.3, first column, first row) and patient PZGE02 (Figure 1.3, first column, third row). This effect is not evident for the two geodesic models because the transport term in (1.4)-(1.5) pushes the front toward the external boundary of the muscle. The transport term should help the front stick to the right boundary, reducing the overflow outside the muscle, which happens when some complex structure is present in the data, as shown in Figure 1.3, fifth row.

The presence of segmentation artifacts may impact the assessment of sarcopenia. In fact, an inaccurate evaluation of the psoas dimension implies the extraction of imprecise metabolic information from the PET data, and may result in an over- or under-estimated determination of the FDG uptake of the psoas tissue. The post-processing procedure was tailored on the three-dimensional schemes to limit this problem by cleaning the segmentation output. For lower quality segmentations, the post-processing typically requires more than one iteration, see Figure 1.4.

Overall, the three numerical methods present comparable performances in terms of segmentation accuracy (Tables 1.2 and 1.3). However, the computational cost required by GMFD 1ord to achieve convergence is one order

of magnitude smaller than GMFD 2nd (see Table 1.1). This is due to the parabolic CFL condition, which slows down the evolution of the wave front and forces the second-order algorithm to do a greater number of iterations to reach convergence.

As for the reached segmentation accuracy, the average Hausdorff and ASSD distances computed between the masks produced by our schemes and the manual segmentation are more than 2 times and more than 1.5 times lower than the distances obtained using the nnUNet-based TS masks, respectively (see Tables 1.2 and 1.3). This fact highlights that, even if the Jaccard and the DSC metrics (of both volume and surface) indicate comparable overlap, our segmentation masks present less outliers. Also, for the internal dataset, the Robust Hausdorff distance is 28% lower for the numerical schemes compared to the TotalSegmentator, on average (Table 1.2), while on the CMB-PCA dataset it is comparable across the segmentation methods.

We observe that the TS segmentations were obtained without pre- or post-processing of the images. Indeed, no pre-processing was performed for the training or testing of the TS [195]. Since the CT images are represented in Hounsfield units, the intensity values are inherently standardized across datasets, making the training and testing images quantitatively comparable. This ensures that TS can generalize well to new data without requiring intensity normalization or other preprocessing steps. In our case, although the images remain quantitatively comparable, we developed a different segmentation method: it is not based on a trained neural network but rather on a numerical schemes driven by the evolution of a three-dimensional wave front using image gradient. As such, pre-processing steps, such as denoising, contrast enhancement, or image filtering, may play a more significant role in improving robustness and accuracy in our approach. Furthermore, no post-processing was applied to the segmentation outputs generated by the TS. Upon visual and quantitative inspection, we observed that the resulting segmentations were already highly regular and smooth, often exhibiting over-resolved anatomical shapes. Due to this inherent smoothness, there was no need to apply additional

filtering or morphological operations to further refine the segmentation.

While the primary aim of this study was to evaluate the applicability and effectiveness of geodesic models for three-dimensional segmentation of the psoas muscle, following more recent developments on the automatic assessment of sarcopenia it may be worth considering alternative deep learning methodologies for psoas muscle segmentation, particularly those designed to perform well under limited data conditions. Techniques like transfer learning [29], knowledge distillation [207], and teacher-student networks [67] have shown their ability to perform well even when large labeled datasets are unavailable. Transfer learning, for instance, involves leveraging pre-trained models, typically trained on large-scale datasets, and fine-tuning them on a smaller, domain-specific dataset. This can significantly reduce the data and computational requirements while still achieving competitive performance [29]. However, pre-trained models may carry biases from the source domain that do not align well with the target domain, leading to suboptimal generalization. Similarly, the teacher-student paradigm includes various strategies where a smaller, simpler model (the student) is trained to replicate the behavior of a larger, more complex model (the teacher) so that the student can perform well, even with fewer parameters or less training data [67, 207]. The teacher model is a source of guidance, it might be pre-trained, trained with more data, or trained using a different strategy. The student model is being trained to learn from the teacher, either directly (by mimicking) or indirectly (e.g., through pseudo-labels or self-training). Different types of teacher-student setups are the i) knowledge distillation, where the student network learns teacher’s reasoning by learning the soft labels produced by the teacher (i.e., the probability distributions of the output classes instead of the hard labels); ii) self-distillation, where the student is the same model architecture as the teacher (knowledge is transferred within layers of the same model or between different stages of the same model); iii) semi-supervised teacher-student, where the teacher labels unlabeled data (pseudo-labeling), and the student learns from

both labeled and pseudo-labeled examples; iv) cross-modal distillation, used when transferring knowledge from one modality (e.g., text) to another (e.g., images). Nevertheless, the effectiveness of this approach often relies on access to a sufficiently large dataset during the initial training of the teacher model. As discussed in Section 1.2.4, while numerous public datasets are referenced in the literature, only a limited number provide reference segmentations of the psoas muscle. To address this, iterative annotation techniques have been utilized [195]. Also, pre-trained models for 3D psoas segmentation are available [57, 195], offering opportunities for transfer learning or use as teacher models.

Our study presents some limitations. First, we recall that this work serves as an empirical evaluation intended to demonstrate the feasibility of applying the proposed approaches to real low-dose CT images, rather than as a fully optimized method for large-scale dataset segmentation. Consequently, parameter selection was conducted in a heuristic manner, guided by insights from the literature, grid search, and empirical testing. For improved efficiency and broader applicability, both the numerical scheme parameters and the post-processing tolerance should ideally be selected in an adaptive and automated fashion. Future work should also focus on defining a stopping criterion that quantitatively accounts for the number of slices in the CT volume, the overall image quality, and the resulting segmentation mask after each iteration. Furthermore, the initialization of both the numerical schemes and the post-processing step currently relies on the manual selection of the initial slice. One possible improvement to avoid robustness issues could be done by automatically identifying the center of mass of a volume containing the psoas (or the segmented psoas, in case of the post-processing), and initiating the schemes from there. Alternatively, deep learning techniques could be utilized to automatically identify the slice corresponding to the L3 vertebral level, as in [75, 188], which could then serve as the starting point for the post-processing analysis, that operates slice-wise and whose result strictly depends on the initial choice. Moreover, the impact of different post-processing pipelines on

the quality of segmentation results is worth considering for future developments. Finally, this approach to segmentation of medical imaging data could be extended to the case of other muscular districts using CT scans, while applications to other structural imaging modalities, such as MRI, is restricted to high-contrast areas.

Chapter 2

Radiomics reliability

Radiomics-based data analysis employs ML or DL techniques to stratify patients and predict their clinical follow-up using radiomic features extracted from regions of interest in the images [73, 90, 98]. Most of the criticism of this approach [116, 137, 181] emphasizes that the effectiveness of these methods significantly depends on the reliability of the computational pipeline generating radiomic features, particularly regarding the steps of segmentation and feature extraction. In this context, studying the relationship between image segmentation and radiomic features is particularly important. Many papers [5, 142, 174, 201] use the Intraclass Correlation Coefficient (ICC) [93] to evaluate radiomic features' reproducibility, as a synonym of stability and reliability, across different segmentation methods, and to perform feature selection. However, the ICC-based approach does not consider the possibility of significant quantitative differences in the various segmentation outcomes, implicitly assuming that the segmentation agreement is always high. Moreover, when analyzing feature stability exclusively with respect to segmentation, i.e., outside the context of prediction, specific features may exhibit a lack of robustness, which would be mostly mitigated during prediction, due to the feature scaling required by ML methods.

This chapter presents a study investigating how segmentation variability affects the predictive performance of radiomics-based ML models designed to

differentiate Triple-Negative Breast Cancer (TNBC) from other molecular subtypes based on MRI-derived features [26]. The stability of the feature selected by these models is reported via ICC, for comparison with existing literature, Pearson correlation, and through the method we described in [25], which introduces four quantitative scores measuring feature stability in terms of consistency, robustness, instability, and quality of feature computation and explicitly accounting for possible variability between results of the segmentation process.

The plan of the chapter is as follows. Section 2.1 illustrates the motivation for the study of radiomics stability with respect to segmentation. Section 2.2 details the methods used in this work, describing the data, the segmentation, and feature selection steps, the machine learning models, and the stability analysis performed in this study. Section 2.3 shows the results obtained on a public cohort of breast cancer patients [156]. Section 2.4 compares the results provided by the different experiments.

2.1 Motivation

Breast cancer is a complex and heterogeneous disease, with multiple molecular subtypes that hamper accurate prediction of disease evolution and the development of targeted treatments [82]. TNBC is defined by the absence of estrogen receptor (ER), progesterone receptor (PR), and HER2 overexpression, accounting for approximately 15–20% of all breast cancers and disproportionately affecting younger and African American women [16, 59]. It is known to present a higher grade, earlier recurrence, and worse overall prognosis compared to other subtypes [59], thus representing the most challenging breast cancer subtype to treat. Subtype determination has always relied on methods such as immunohistochemistry, staining, and fluorescence in situ hybridization [82]. Recent advances in literature indicate the potential for predicting molecular subtypes using radiomics through machine learning techniques [100, 165, 173].

However, the widespread adoption of radiomics in clinics is hampered by issues related to features stability [23, 25]. Factors that could affect radiomic features computation can be found in the image acquisition and reconstruction phase, in the image pre-processing steps, and in the segmentation of the region of interest from which radiomic features are extracted. Scaco et al. [159] reported a list of papers that evaluated the effects of segmentation on radiomics stability, with just one of them considering applications to breast cancer imaging. Granzier et al. [61] studied the robustness of radiomic features, extracted by two software tools, with respect to variability in manual segmentation of breast tumor on MRI. Although a threshold value of 0.90 for the ICC was chosen to determine feature robustness, its significance for radiomic models in predicting patient outcomes was not investigated. Other studies addressed the problem of assessing the robustness of radiomic features by segmentation perturbation [183, 202, 214] for various applications, and various type of imaging (CT, PET, and MRI), for different diseases (lung cancer, head and neck squamous cell carcinoma, glioblastoma). All these studies call for caution in the use of predictive models involving radiomic features implicated with contouring variability within the context of disease stratification and risk assessment, although no prediction experiment is reported. Kothari et al. [94] highlighted that the selection of robust features from masks delineated by different clinicians allows survival models to retain their prognostic ability. Among the papers addressing segmentation variability in a more systematic way, along with predictive tasks, Poirot et al. [142] studied how differences among segmentations affect radiomic features in neuroimaging, in a cohort of T1-weighted and diffusion tensor images of sleep-deprived patients. The robustness and reproducibility of radiomic features were assessed using ICC for descriptive purposes, while the performance of the predictive model is only expressed in terms of accuracy, and no statistical analysis is reported to evaluate feature robustness in relation to predictive performance. Liu et al. [109] showed that for CT images of oropharyngeal cancer, radiomic features vary a lot when the regions of interest (ROIs) are not well segmented (ICC).

No statistical test is reported for the significance of the results, and no quantification of the segmentation variability is presented. Moreover, the authors do not examine the behavior of the same group of features across variations in masks, nor do they attempt to generate a new prognostic signature; instead, they train univariate models for prediction.

In this study, we investigated how feature stability and prediction performance are impacted by segmentation variability [26]. We focused on identifying radiomic features that are consistently predictive of TNBC across a population of breast cancer patients, and we explored the prediction performances and the stability of the most significant radiomic features across different segmentation masks. In contrast to previous studies, our work systematically addresses segmentation variability, as assessing radiomics stability across different segmentations is essential for ensuring the reliability and generalizability of radiomics-based models. Unlike the studies that perform feature selection based on segmentation variability, we focus on selecting features that are both predictive and explainable, prior to evaluating their stability with respect to segmentation. Our feature selection step is based on explainability, ensuring that we identify predictive and interpretable features relevant to the task of distinguishing TNBC from other cancer subtypes. This approach enables a critical post hoc analysis to determine whether there is a meaningful connection between the predictive features and segmentation variability, and how this relationship influences predictive performance. While this investigation focuses on a specific application, its findings could potentially elucidate the general behavior of radiomic features, providing a broader understanding of the principles underlying radiomics-based analyses.

2.2 Materials and methods

2.2.1 Data collection

Consent or waiver for data usage was not required since all data were obtained in de-identified form from the publicly available Duke-Breast-Cancer-MRI dataset [156], hosted on TCIA [38]. For each patient, the Duke-Breast-Cancer-MRI dataset contains dynamic contrast-enhanced (DCE) MR images from multiple time points, capturing both pre-contrast and post-contrast phases. Imaging data were collected using various scanner models from GE MEDICAL SYSTEMS, MPTronic, and SIEMENS, including the Avanto, Optima MR450w, SIGNA EXCITE, SIGNA HDx, SIGNA HDxt, Skyra, Trio, and TrioTim. The dataset also contains demographic, clinical, pathology and treatment information, and outcomes of patients (e.g., response to treatment, recurrence, follow-up). This study focuses only on the pre-operative first post-contrast images from DCE-MR. The patients involved correspond to the 251 selected by Caballo et al. [20]. Seven patients were excluded due to issues related to DICOM metadata. Therefore, a cohort of 244 patients was analyzed for this study, including 71 with TNBC and 173 non-TNBC (30% vs. 70% of the dataset, respectively). The dataset was preliminary split into a 70%-30% stratified proportion for training and testing of the ML models, for 130 random data splits.

2.2.2 Segmentation

Manual segmentation of MR images was provided by Caballo et al [20]. To assess the stability of the features with respect to segmentation accuracy, we introduced variability in the manual segmentation mask by simulating four other annotations per case. Our aim was to obtain an average mean DSC across all segmentations ranging from 0.4 to 0.8 with respect to the original mask. This was achieved via the morphological operation of closing. Using the ball structuring element from the `skimage.morphology` library [187] with varying

kernel sizes, we gradually enlarged the manual segmentation mask to achieve various mean DSC, including more and more portions of the peritumoral region. Specifically, we applied the following operations:

- dilation and erosion operations, using a kernel size of 5, to obtain mean DSC 0.8;
- dilation and erosion, using a kernel size of 9, to obtain mean DSC 0.7;
- dilation with a kernel size of 11 and erosion with a kernel size of 9, to obtain mean DSC 0.6;
- finally, the manual segmentation was replaced by the ellipsoid contained in the ROI box provided by the dataset along with the images, to obtain mean DSC 0.4.

The implemented modifications of the manual masks allowed for consideration of a broad range of potential contour variations, as the segmentation obtained through morphological operations exhibited significant variability. For instance:

- certain tumors were originally located near the skin, causing the modified mask to extend beyond the patient's body;
- segmentation of tumors with multiple lesions resulted to be grouped into a single mask (Figure 2.1, central column);
- spike-like structures were enhanced (Figure 2.1, in orange).

Additionally, the ellipsoid mask served as a geometric approximation of tumor segmentation (DSC 0.4), fulfilling the same function as a ROI box by providing a highly simplified approximation, while offering a more natural shape, closer to a potential manual segmentation.

In the rest of the document the manual mask and its modifications are referred to as manual, closing 08, closing 07, closing 06, and ellipsoid 04, respectively (see Figure 2.1), referring to the mean DSC value reached by each morphological operation applied to the original mask.

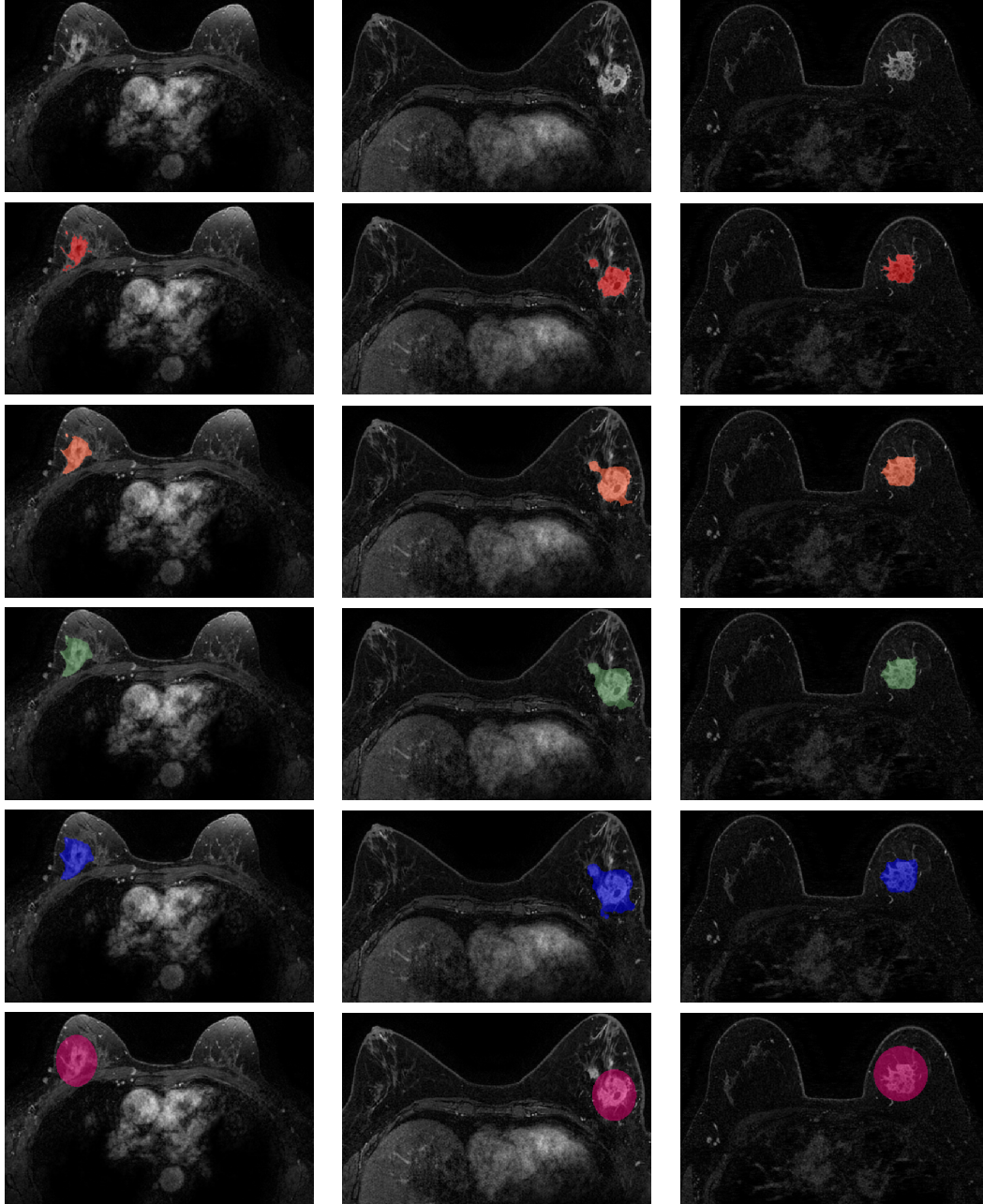


Figure 2.1: Left to right: MR slices of three different patients (first-post-contrast image). Top to bottom: original image, manual tumor segmentation (red), closing 08 mask (orange), closing 07 (green), closing 06 (blue), and ellipsoid 04 (magenta). For the definition of ‘closing mask’ see Section 2.2.2.

2.2.3 Radiomic feature extraction

Radiomic features were extracted from the segmentation masks on the pre-operative first post-contrast DCE-MR images, as in [100, 101, 155, 203]. Subtraction of imaging temporal phases or other sequences were not considered for feature extraction. Image z-score normalization and fixed-bin count discretization with 50 bins were applied, and the extraction was performed using PyRadiomics 3.1.0 library [62]. We extracted 1130 radiomic features including shape, first-order, matrix-based, wavelet, and Laplacian of Gaussian (LoG) features.

Shape-based features describe the geometric properties of a region of interest, such as volume, surface area, and compactness. First-order statistics capture the distribution of voxel intensities within the image, summarizing global intensity characteristics. Gray Level Co-occurrence Matrix (GLCM) features assess spatial relationships between intensity values, reflecting textural uniformity and contrast. Gray Level Run Length Matrix (GLRLM) features quantify the length and distribution of consecutive voxel intensity runs, characterizing texture smoothness. Gray Level Size Zone Matrix (GLSZM) features evaluate the size and distribution of homogeneous intensity zones. Neighbouring Gray Tone Difference Matrix (NGTDM) features measure texture strength and contrast based on intensity differences with neighboring voxels. Gray Level Dependence Matrix (GLDM) features capture the degree to which voxels depend on neighboring intensities, quantifying image granularity and texture complexity.

All these features were extracted from the original and filtered images. Wavelet-HHH, wavelet-HLH, and wavelet-LLL filters are types of transform filters [209] where HHH applies high-pass filtering across all three dimensions to capture high-frequency features, HLH applies a combination of high and low-pass filtering in different dimensions to extract mixed frequency information, and LLL applies low-pass filtering across all dimensions to capture the overall low-frequency components. On the other hand, LoG [50] is an edge detection method where σ values denote the scale of Gaussian smoothing

applied before computing the Laplacian to detect edges at varying levels of detail. Various parameter choices have been explored in the literature for the LoG features. For example, in [167], the authors employed $\sigma = 1, 3, 5$. Their feature selection process revealed that, for the classification task distinguishing TNBC from non-TNBC, the most relevant feature was extracted using a LoG filter with $\sigma = 1$. In the task of differentiating HER2+ from HER2- tumors, features corresponding to $\sigma = 3$ and $\sigma = 5$ were selected. Similarly, for the classification of luminal vs. non-luminal subtypes, one feature derived from LoG filters with $\sigma = 1$, three with $\sigma = 3$, and one with $\sigma = 5$ were identified as significant. The work [80] studying lung cancer brain metastases on MRI used $\sigma = 2, 3, 4, 5$, while [79] used $\sigma = 1, 2, 3$ for repeatability and reproducibility study on non-small cell lung cancer on CT. In general, the choice of sigma parameters could be interpreted as a grid search strategy designed to identify the most informative LoG features (the same might be said for the combination of high- and low-pass filtering used in the extraction of wavelet features, as in [210]). Larger values of σ tend to capture coarser, large-scale edges and offer greater robustness to noise, whereas smaller σ values retain finer details and edges within the image. In this study, we selected $\sigma = 1, 2, 3$ to span a reasonable range of both small and moderately large scales for the kernel used in the filtering process. Extensive documentation of radiomics features can be found in [62].

For the analysis, features were normalized via z-score. Then, they were harmonized to limit the potential bias introduced by the differences in signal-to-noise ratio caused by the manufacturer model. Harmonization was performed using the parametric version of the ComBat method [131, 198], which yields a transformation of the feature distributions according to the variable being tested (manufacturer model) using additive and multiplicative batch effects.

2.2.4 Clinical features

Clinical features were utilized for prediction purposes to enable comparisons with radiomics-based models. Specifically, this work considered demographi-

cal variables, i.e., age, menopause at diagnosis, ethnicity, and metastatic state at presentation, and biopsy variables, including tubule formation, nuclear grade, and mitotic rate. These biopsy variables are clinically used for breast cancer grading and are strongly associated with tumor aggressiveness. While they do not directly determine tumor subtype, they may be naturally linked to it, as higher histologic grades often correlate with more aggressive subtypes. For this reason, the model based on biopsy variables could serve as a reference for comparison in subtype prediction.

2.2.5 Machine Learning models

For each data split, the imbalanced data issue (TNBC vs. non-TNBC) was addressed prior to the training process via the synthetic minority oversampling technique (SMOTE) [31], a widely used method for random oversampling of tabular data, like radiomic features. A preliminary selection of the 50 most informative features was performed through ANOVA F-values computation, used to rank features based on their relevance to the target variable. Given the total cohort size of 244 patients, this dimensionality reduction was adopted to maintain a reasonable feature-to-sample ratio within the training subsets used for model development, thereby minimizing overfitting risk while retaining a meaningful subset of relevant variables. Then, a Logistic Regression model with L1-norm penalty [127] was trained for the classification task TNBC vs. non-TNBC with 5-fold cross-validation. The regularization parameter C , to control feature selection strength, was set to 1. L1-norm penalty is known to promote sparsity by selecting only the most relevant features for the model, thus identifying the radiomic signature of each lesion.

After the training process, we used SHAP algorithm (SHapley Additive exPlanations [113]) to collect the most explicative features for each model (each trained on a different data split). SHAP is a feature importance tool, based on a game theoretic approach, used in ML for explaining the output of a model by quantifying the importance of each feature. SHAP identifies the most relevant features that contribute to the model's predictions by calculating

SHAP values for each feature in the dataset: features with higher SHAP values are considered more influential in the model’s predictions, while features with lower SHAP values have less impact. Specifically, for each trained model, we identified the top 10 features selected by SHAP. Since SHAP analysis is performed iteratively across multiple data splits, retaining the top 10 features each time allows to capture variability in feature relevance across different splits (the number 10 was empirically chosen). From the aggregated set of top SHAP-selected features across all models, we further selected those with the highest frequency of occurrence, that appeared at least 15 times overall (the ‘best-SHAP features’ from now on). This threshold was chosen to maintain approximately 20 features, aligning with the number of features retained by the L1-regularized logistic regression models. These features can be considered collectively predictive, as they contributed to the predictive performance across the whole dataset. A diagram outlining the SHAP feature selection methodology employed in this study is shown in Figure 2.2, while detailed implementation is reported in Algorithm 2. The best-SHAP features were then used to train Logistic Regression models, one for each data split, for comparison with the baseline models, described at the beginning of this section.

This procedure was repeated for each feature set, extracted from the segmentation masks described in Section 2.2.2. In the following, ‘baseline-manual’, ‘baseline-closing 08’, ‘baseline-closing 07’, ‘baseline-closing 06’, and ‘baseline-ellipsoid 04’ will refer to baseline models trained with features extracted from the corresponding mask, while ‘best-SHAP-manual’, ‘best-SHAP-closing 08’, ‘best-SHAP-closing 07’, ‘best-SHAP-closing 06’, and ‘best-SHAP-ellipsoid 04’ will indicate models trained with best-SHAP features derived from the corresponding baseline model.

All ML models were developed in Python (v3.12.2). Experiments and analyses were conducted on a personal computer equipped with an Apple M1 Pro chip and 16 GB of unified memory.

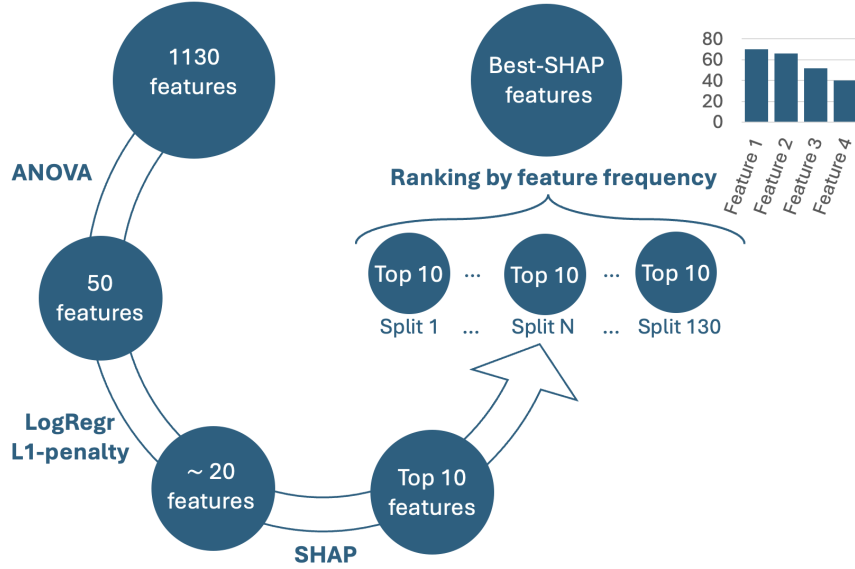


Figure 2.2: Flowchart of the feature selection methodology employed for this study, based on SHAP explainability algorithm.

2.2.6 Statistical analysis

Univariate statistical analysis was performed to evaluate the potential significance of individual features in discriminating TNBC vs. non-TNBC cases. Continuous variables were analyzed with two-sided Kolmogorov-Smirnov test [15], while discrete variables with the Pearson's chi-squared test. Bonferroni correction for multiple comparisons was applied.

Prediction models, leveraging clinical, radiomic and the best-SHAP features, underwent rigorous evaluation for various performance metrics: accuracy, recall, specificity, balanced accuracy (which accounts for dataset imbalance by computing the mean of recall and specificity), and Area Under the Receiver Operating Characteristic Curve (ROC-AUC). The performances reported in this manuscript are averaged across data splits and 95% confidence intervals (CI) are reported. Kolmogorov-Smirnov test was used to evaluate the difference in score distributions from different ML models, with Bonferroni correction for multiple comparisons.

To check feature stability/robustness across different segmentation masks,

Algorithm 2 Best-SHAP feature selection

Input: radiomic features extracted from the manual segmentation

Step 1: preliminary feature selection and model training

For each data split (training set):

1. compute ANOVA F-values to rank features based on their relevance to the target (TNBC)
2. select the first 50 features
3. train the predictive model with such features (each LogReg model further selects approximately 20 features due to L1 penalty)
4. collect the top 10 features considered explainable for the current model by SHAP method

Step 2: best-SHAP feature selection

From the aggregated set of SHAP-selected features identified in Step 1:

1. count the frequency of occurrence of each feature
2. rank them in descending order
3. select the ones that appeared at least 15 times overall: these are the best-SHAP features for the manual segmentation

Repeat steps 1 and 2 for each feature set (closing 08, 07, 06, ellipsoid 04).

the most relevant features for the models were evaluated via ICC [93] and Pearson correlation coefficient. The ICC measures agreement between measurements and, in radiomics, is often used to assess feature reproducibility across different segmentation masks (representing the raters). This study utilized a two-way random effects model, generally denoted as ICC(2), which is appropriate when all raters evaluate all subjects and both raters and subjects are considered random effects. The computation accounts for variability between subjects and raters, as well as error variance, which may result from differences in rater performance or other unaccounted noise. ICC and Pearson correlation were computed pairwise using the features extracted from the manual mask and each of its modifications (DSC 08, 07, 06, 04). Other scores

for the evaluation of feature stability are introduced in the next section.

2.2.7 Radiomics reliability scores

In this section we introduce a method that aims to generate quantitative scores measuring the numerical relationship between feature stability and tumor segmentation. It is based on a quantitative assessment of segmentation variability and the relative error on feature values, enabling the definition of four reliability scores for radiomic features, where 'reliability' refers to feature stability with respect to segmentation variability.

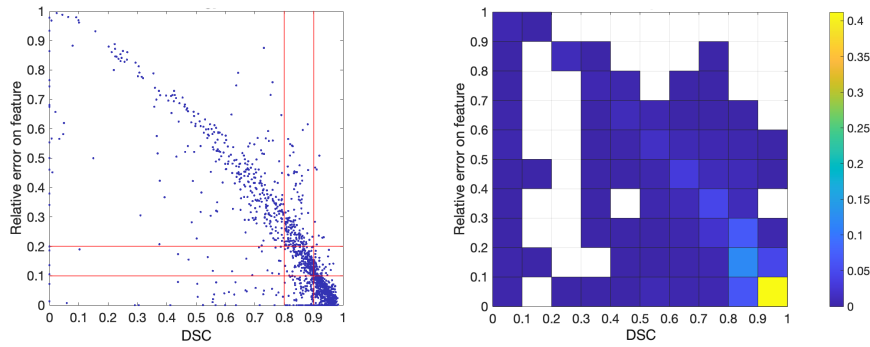


Figure 2.3: Left panel: scatter plot of the relative error (in absolute value) on a specific feature against DSC. Reference lines at 0.8 and 0.9 DSC, 0.1 and 0.2 relative error. Right panel: two-dimensional reliability histogram H of the same feature. Each bin contains the rate of slices falling in the corresponding area of the scatter plot.

For this method, a segmentation mask is selected as reference, while another mask is selected for the stability analysis. These could be, for example, the manual segmentation and one of its modifications, respectively. For each patient, the DSC is calculated between the two masks and, leveraging the fact that each radiomic feature is extracted from both segmentations, the relative error on feature value is computed. Then, for each feature, a scatter plot is generated to show the relative error plotted against the corresponding DSC. Subsequently, each scatter plot is discretized by means of a 10×10 uniform

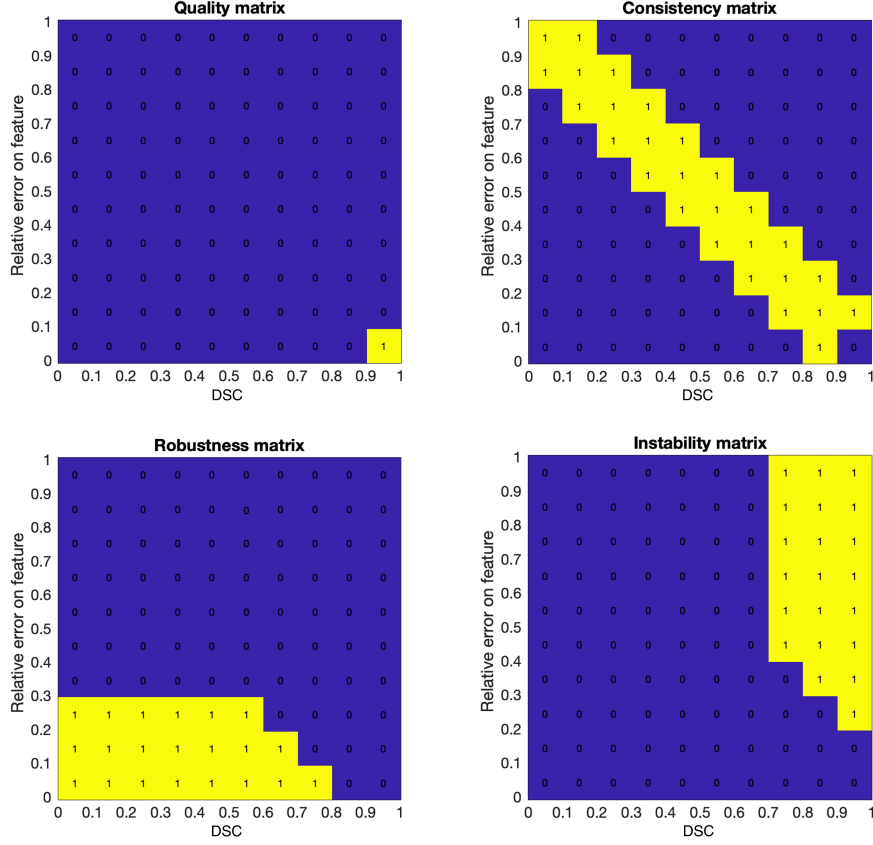


Figure 2.4: Reliability matrices. Top left panel: quality matrix. Top right panel: consistency matrix. Bottom left panel: robustness matrix. Bottom right panel: instability matrix. Each entry contains the binary weight we assign to the corresponding bin of the reliability histogram.

grid to construct the *reliability histogram* H that counts the rate of patients falling into each bin of the grid. The bin width is 0.1 and each bin is of the form $[DSC_1, DSC_2) \times [rel. err._1, rel. err._2)$, $0 \leq DSC \leq 1$, $0 \leq rel. err. \leq 1$. Figure 2.3 contains an example of scatter plot and of the corresponding reliability histogram, for the same specific feature.

Four 10×10 *reliability matrices* having entry values in $\{0, 1\}$ are then introduced. Coherently with the reliability histogram, the rows and columns in the reliability matrices correspond to the relative error and the DSC, respectively. The *quality matrix* Q (Fig 2.4, top left panel) promotes high-quality features,

indicating when segmentation agreement and feature computation accuracy are simultaneously high, as

$$Q_{ij} = \begin{cases} 1, & \text{if } 0 \leq \text{rel. err.} \leq \lceil 1 - \text{DSC} \rceil \text{ and } 0.9 \leq \text{DSC} \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

The *consistency matrix* C (Fig. 2.4, top right) promotes the anti-linear correlation between the relative error and the DSC, i.e., indicates when feature stability varies anti-linearly with segmentation variation:

$$C_{ij} = \begin{cases} 1, & \text{if } \lfloor 1 - \text{DSC} - 0.2 \rfloor \leq \text{rel. err.} < \lceil 1 - \text{DSC} + 0.1 \rceil \cap 0 \leq \text{rel. err.} \leq 1 \\ & \text{and } 0 \leq \text{DSC} < 0.9, \\ 1, & \text{if } \lceil 1 - \text{DSC} \rceil \leq \text{rel. err.} < \lceil 1 - \text{DSC} + 0.1 \rceil \\ & \text{and } 0.9 \leq \text{DSC} < 1, \\ 0, & \text{otherwise.} \end{cases}$$

The *robustness matrix* R (Fig. 2.4, bottom left) promotes high-robustness radiomics features, showing when feature values remain stable with any segmentation variation:

$$R_{ij} = \begin{cases} 1, & \text{if } 0 \leq \text{rel. err.} < \lfloor 1 - \text{DSC} - 0.1 \rfloor \cap 0 \leq \text{rel. err.} < 0.3 \\ & \text{and } 0 \leq \text{DSC} < 0.8, \\ 0, & \text{otherwise.} \end{cases}$$

Finally, we introduce an *instability matrix* A (Fig. 2.4, bottom right) to promote features that vary largely within small changes in the segmentation mask:

$$A_{ij} = \begin{cases} 1, & \text{if } \lceil 1 - \text{DSC} + 0.1 \rceil \leq \text{rel. err.} \leq 1 \text{ and } 0.7 \leq \text{DSC} < 1, \\ 0, & \text{otherwise.} \end{cases}$$

By exploiting the reliability matrices, the following four ‘reliability scores’ are finally introduced:

- the *quality score*

$$q = \sum_{i=1}^{10} \sum_{j=1}^{10} Q_{ij} H_{ij}; \quad (2.1)$$

- the *consistency score*

$$c = \sum_{i=1}^{10} \sum_{j=1}^{10} C_{ij} H_{ij}; \quad (2.2)$$

- the *robustness score*

$$r = \sum_{i=1}^{10} \sum_{j=1}^{10} R_{ij} H_{ij}; \quad (2.3)$$

- the *instability score*

$$a = \sum_{i=1}^{10} \sum_{j=1}^{10} A_{ij} H_{ij}. \quad (2.4)$$

These scores can serve as quantitative parameters for a reliability/stability assessment process in radiomics. Specifically, for each feature, the scores provide measures of the collective behavior of that feature on a group of patients, by quantifying the proportion of patients with specific trend of feature stability. However, their interpretation is tied to the distribution of the DSC along the x-axis. Ideally, patients would be evenly distributed along the entire x-axis, but this scenario is rarely achieved, as segmentation methods for clinical applications typically exhibit consistent performance across most patients, with only a few outliers. Consequently, patient distribution on the x-axis is usually uneven and the DSC distribution should be considered when interpreting the reliability scores. In this study, the DSC was deliberately adjusted to achieve specific mean values, concentrating the analyses on specific portions of the x-axis for each case.

2.3 Results

Table 2.1 displays the means and standard deviations of the DSC obtained for the various modifications of the manual segmentation mask, computed across the patients involved in the study.

The univariate feature analysis, performed on the whole dataset, showed that no individual feature independently exhibited a significant difference between the TNBC and non-TNBC groups based on the Kolmogorov-Smirnov and Chi-squared tests ($p\text{-value} > 0.05$).

Table 2.2 lists the best-SHAP-features for each model, identified as the most overall explicative features for each segmentation mask, as described in Section 2.2.5. Only four of these features were commonly selected by the models and are reported in bold in Table 2.2: High-Gray-Level-Zone Emphasis (from LoG $\sigma = 3$ image, GLSZM matrix), Large-Dependence High-Gray-Level Emphasis (from original image and GLDM matrix), Zone Entropy (from wavelet-HHH image and GLSZM matrix), and Skewness (from wavelet-HLH image, first order features). Figure 2.5 reports an example of wavelet-HHH, wavelet-HLH, and LoG $\sigma = 3$ filtered images, for a visual reference of the filtering effect.

Figure 2.6 illustrates the performance of the prediction models using demographical variables (boxplot 1) and biopsy variables (boxplot 2), as well as baseline models (boxplots 3-7) and best-SHAP models (boxplots 8-12). Performance is shown in terms of ROC-AUC. Table 2.3 reports detailed statistics, including skill-score means and confidence intervals obtained for each ML model. The prediction using solely demographical variables is random. On

Table 2.1: Mean $\pm$ standard deviation of the DSC obtained for each modification of the manual segmentation mask, computed across all patients.

Segmentation mask	Mean DSC $\pm$ standard deviation
Closing 08	0.81 ± 0.13
Closing 07	0.69 ± 0.12
Closing 06	0.60 ± 0.12
Ellipsoid 04	0.40 ± 0.13

2.3 Results

average, predictions obtained without SMOTE augmentation showed a 4% decrease compared to those presented in Table 2.3.

Table 2.2: List of the best-SHAP features identified as explicative through SHAP method for each model. '✓' indicates the model for which the feature has been selected as explicative. Features that are selected by all the models are highlighted in bold.

Image Filter	Matrix	Feature Name	Manual	Closing 08	Closing 07	Closing 06	Ellipsoid 04
LoG $\sigma = 2$	glcm	Cluster Prominence	✓	-	-	-	-
LoG $\sigma = 3$	glrlm	Run Entropy	✓	✓	-	-	-
LoG $\sigma = 3$	glszm	High-Gray-Level-Zone Emphasis	✓	✓	✓	✓	✓
LoG $\sigma = 3$	glszm	Low-Gray-Level-Zone Emphasis	-	-	-	✓	-
LoG $\sigma = 3$	glszm	Small-Area-High-Gray-Level Emphasis	✓	✓	-	-	-
original	firstorder	Median	-	-	✓	-	-
original	firstorder	Minimum	-	✓	✓	✓	✓
original	gldm	Large-Dependence-High-Gray-Level Emphasis	✓	✓	✓	✓	✓
wavelet-HHH	firstorder	Median	-	-	-	-	✓
wavelet-HHH	firstorder	Uniformity	-	-	✓	-	-
wavelet-HHH	glcm	Cluster Tendency	-	-	-	✓	-
wavelet-HHH	glcm	Inverse Variance	✓	-	-	-	-
wavelet-HHH	glcm	Joint Energy	-	✓	-	-	-
wavelet-HHH	glcm	MCC	✓	-	✓	-	-
wavelet-HHH	glcm	Maximum Probability	✓	✓	-	-	-
wavelet-HHH	glcm	Sum Entropy	✓	✓	✓	✓	-
wavelet-HHH	gldm	Dependence Entropy	✓	✓	-	✓	✓
wavelet-HHH	glszm	Gray-Level Non-Uniformity Normalized	✓	-	-	-	✓
wavelet-HHH	glszm	Gray-Level Variance	-	-	✓	✓	-
wavelet-HHH	glszm	Zone Entropy	✓	✓	✓	✓	✓
wavelet-HHH	ngtdm	Complexity	✓	-	✓	-	✓
wavelet-HHL	firstorder	Skewness	-	-	✓	-	-
wavelet-HHL	glcm	MCC	-	✓	✓	-	✓
wavelet-HLH	firstorder	Skewness	✓	✓	✓	✓	✓
wavelet-HLH	glcm	Cluster Shade	-	-	✓	-	-
wavelet-HLH	gldm	Dependence Entropy	✓	-	-	-	✓
wavelet-HLH	glrlm	Long-Run-High-Gray-Level Emphasis	-	✓	✓	✓	-
wavelet-HLH	glrlm	Short-Run-High-Gray-Level Emphasis	-	✓	-	-	-
wavelet-HLH	glszm	Low-Gray-Level-Zone Emphasis	-	-	✓	✓	✓
wavelet-HLH	glszm	Small-Area-High-Gray-Level Emphasis	-	-	✓	-	-
wavelet-HLH	glszm	Small-Area-Low-Gray-Level Emphasis	-	-	-	✓	-
wavelet-HLH	glszm	Zone Entropy	-	-	-	-	✓
wavelet-HLL	glszm	Small-Area-High-Gray-Level Emphasis	-	-	-	-	✓
wavelet-LHH	firstorder	Median	-	✓	-	-	-
wavelet-LHH	glcm	Autocorrelation	-	✓	-	-	-
wavelet-LHH	glszm	Zone Entropy	✓	✓	✓	-	✓
wavelet-LHL	firstorder	Interquartile Range	-	-	-	✓	-
wavelet-LHL	firstorder	Mean	-	✓	✓	✓	-
wavelet-LHL	firstorder	Median	-	-	-	✓	-
wavelet-LHL	firstorder	Skewness	-	✓	-	-	-
wavelet-LHL	glcm	Correlation	-	✓	✓	✓	✓
wavelet-LHL	glrlm	Long-Run-Low-Gray-Level Emphasis	-	-	-	✓	-
wavelet-LLH	glcm	Imc1	-	-	-	✓	-
wavelet-LLH	glcm	MCC	✓	✓	✓	-	-
wavelet-LLL	firstorder	10th Percentile	✓	-	-	-	-
wavelet-LLL	firstorder	Median	-	-	-	✓	-
wavelet-LLL	glcm	Cluster Shade	-	✓	✓	-	-
wavelet-LLL	glcm	MCC	-	-	-	-	✓
wavelet-LLL	gldm	Large-Dependence-High-Gray-Level Emphasis	✓	-	-	-	✓
wavelet-LLL	glrlm	Long-Run-High-Gray-Level Emphasis	-	✓	-	-	-
wavelet-LLL	glszm	Small-Area-Low-Gray-Level Emphasis	-	-	-	✓	-

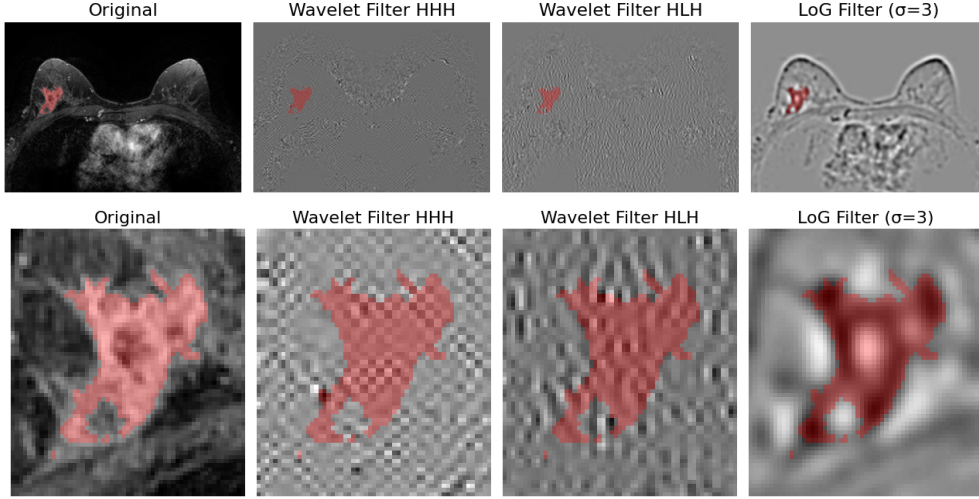


Figure 2.5: Top row, from left to right: example slice of original DCE-MR image, wavelet-HHH filtered image, wavelet-HLH filtered image, and LoG $\sigma = 3$ filtered image. Bottom row, from left to right: zoomed view of the manual segmentation mask on the original and filtered images. Tumor segmentation is shown in red.

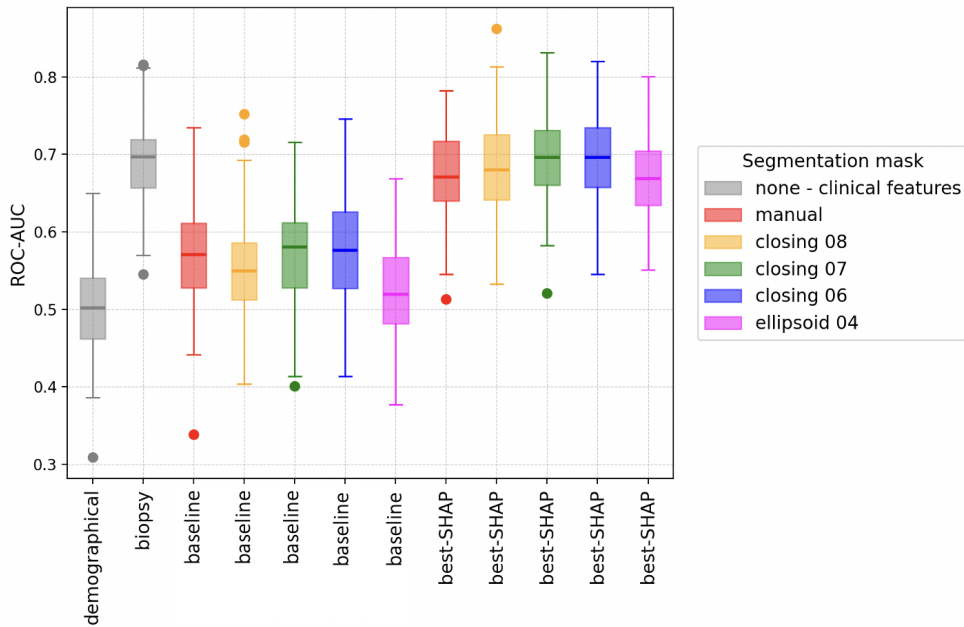


Figure 2.6: ROC-AUC scores obtained by testing demographical model (boxplot 1), biopsy model (boxplot 2), baseline models (boxplots 3-7), and best-SHAP models (boxplots 8-12).

2.3 Results

Table 2.3: Performances of demographical model, biopsy model, baseline models, and best-SHAP models for TNBC prediction. Mean skill-scores and their 95% confidence interval (within brackets) are reported. In bold, the best mean result for each skill-score.

Experiment	Accuracy	Balanced Accuracy	Recall	Specificity	ROC-AUC
demographical	0.482 (0.473, 0.490)	0.492 (0.483, 0.501)	0.517 (0.497, 0.537)	0.467 (0.454, 0.480)	0.499 (0.488, 0.509)
biopsy	0.607 (0.598, 0.615)	0.651 (0.642, 0.661)	0.760 (0.739, 0.782)	0.543 (0.528, 0.557)	0.687 (0.677, 0.697)
baseline - manual	0.588 (0.579, 0.596)	0.553 (0.543, 0.563)	0.468 (0.447, 0.488)	0.638 (0.628, 0.649)	0.575 (0.563, 0.586)
baseline - closing 08	0.567 (0.558, 0.576)	0.532 (0.523, 0.542)	0.446 (0.428, 0.464)	0.619 (0.607, 0.631)	0.552 (0.542, 0.563)
baseline - closing 07	0.591 (0.581, 0.600)	0.553 (0.543, 0.564)	0.462 (0.443, 0.481)	0.645 (0.632, 0.658)	0.573 (0.562, 0.584)
baseline - closing 06	0.591 (0.581, 0.601)	0.555 (0.544, 0.566)	0.468 (0.448, 0.488)	0.643 (0.629, 0.656)	0.578 (0.566, 0.590)
baseline - ellipsoid 04	0.560 (0.551, 0.569)	0.516 (0.506, 0.526)	0.407 (0.390, 0.425)	0.625 (0.612, 0.637)	0.524 (0.514, 0.535)
best-SHAP - manual	0.643 (0.635, 0.651)	0.624 (0.615, 0.633)	0.577 (0.560, 0.595)	0.671 (0.660, 0.683)	0.673 (0.664, 0.683)
best-SHAP - closing 08	0.647 (0.638, 0.655)	0.633 (0.624, 0.642)	0.598 (0.582, 0.615)	0.667 (0.655, 0.679)	0.685 (0.675, 0.695)
best-SHAP - closing 07	0.655 (0.648, 0.663)	0.638 (0.629, 0.647)	0.595 (0.575, 0.615)	0.681 (0.670, 0.692)	0.695 (0.686, 0.705)
best-SHAP - closing 06	0.671 (0.663, 0.678)	0.649 (0.641, 0.657)	0.595 (0.577, 0.614)	0.703 (0.690, 0.715)	0.696 (0.686, 0.706)
best-SHAP - ellipsoid 04	0.639 (0.632, 0.647)	0.618 (0.610, 0.627)	0.566 (0.549, 0.583)	0.671 (0.660, 0.681)	0.671 (0.662, 0.680)

Kolmogorov-Smirnov tests provided evidence against equal model performance for all the comparisons between baseline models and best-SHAP models (for visual reference see Figure 2.6). No statistical significance is observed between the performances of biopsy-based model and best-SHAP models, nor is any difference detected in the pairwise comparison of the best-SHAP models across all masks, except between best-SHAP-closing 07 and best-SHAP-ellipsoid 04 models. Table 2.4 provides a summary of these key comparisons.

Table 2.4: Results of the Kolmogorov-Smirnov tests to evaluate differences in model performance. Bonferroni correction for multiple comparisons was applied. Significance column: ‘*’ indicates statistical evidence against equal model performance after Bonferroni correction; ‘-’ means no statistical significance.

Model 1	Model 2	Significance
baseline model (any mask)	best-SHAP model (any mask)	*
biopsy model	best-SHAP model (any mask)	-
best-SHAP model (any mask)	best-SHAP model (any mask)	- [†]

[†] Exception: statistical significance between best-SHAP-closing 07 and best-SHAP-ellipsoid 04.

As for the stability analysis, we focused on the four commonly selected best-SHAP features cited above. Figure 2.7 displays the values of ICC and Pearson correlation for each of these features at varying segmentation mask. The dashed lines indicate the median ICC and Pearson correlation across all features at varying mask, for comparison with the scores of the four common features. These features exhibited decreasing ICC values as the segmentation accuracy declined (mean ICC across features for decreasing DSC: 0.91, 0.82, 0.76, 0.65) and high Pearson correlation coefficient (mean Pearson correlation across features for decreasing DSC: 0.91, 0.82, 0.92, 0.76).

Figure 2.8 reports the reliability/stability scores of the four common SHAP features across mask modification, and Figure 2.9 shows the scatter plots of feature stability w.r.t. segmentation variability for each of these features, for each modification of the manual mask. By definition, the scores may not sum to 1 if some patients fall outside the predefined ranges of quality, consistency, robustness, and instability. This is particularly evident for the wavelet-HLH-firstorder Skewness feature (Figure 2.9, last row), as decreasing accuracy of the segmentation mask leads to higher and higher relative error in the feature value for many patients (up to 60% of the patients, see bottom right panel of Figure 2.8), causing them to fall outside the predefined ranges, and highlighting overall feature instability.

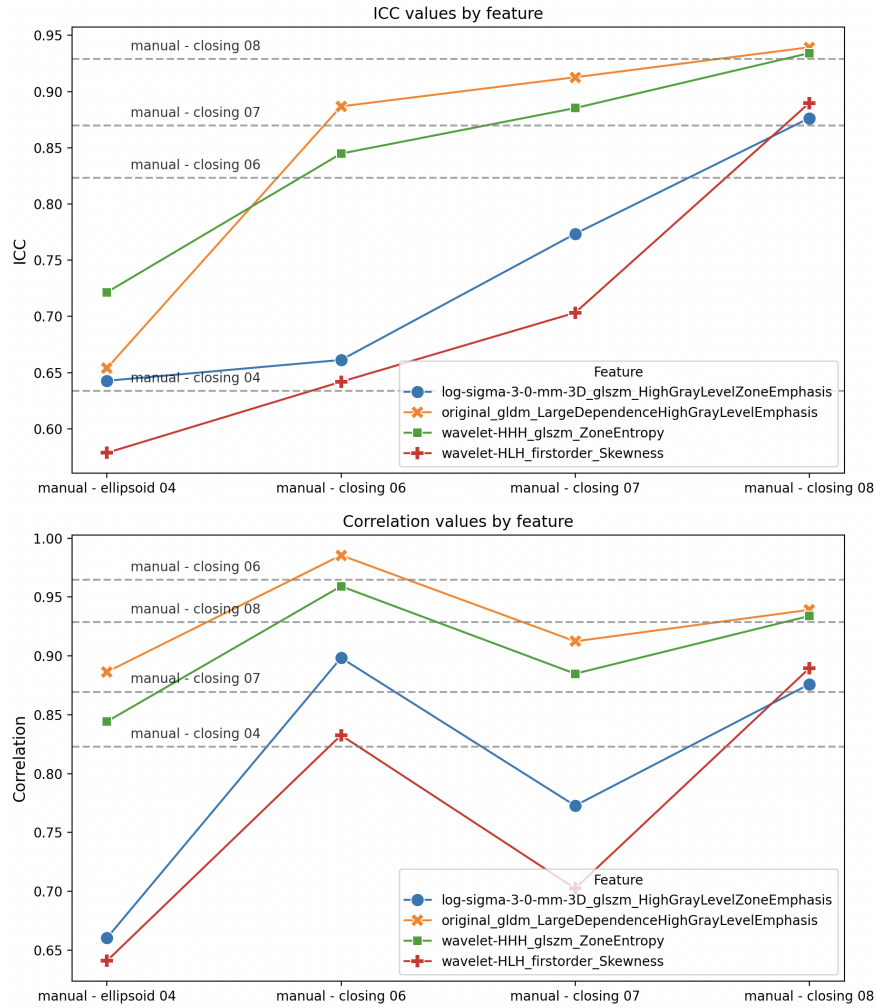


Figure 2.7: Top panel: ICC of the four common best-SHAP features at varying segmentation mask; the dashed lines indicate median ICC on all features, summarizing the reproducibility between features extracted from the manual mask and each of its modifications. Bottom panel: Pearson correlation of the four common features at varying mask; the dashed lines indicate median correlation on all features, summarizing the correlation between features extracted from the manual mask and each of its modifications.

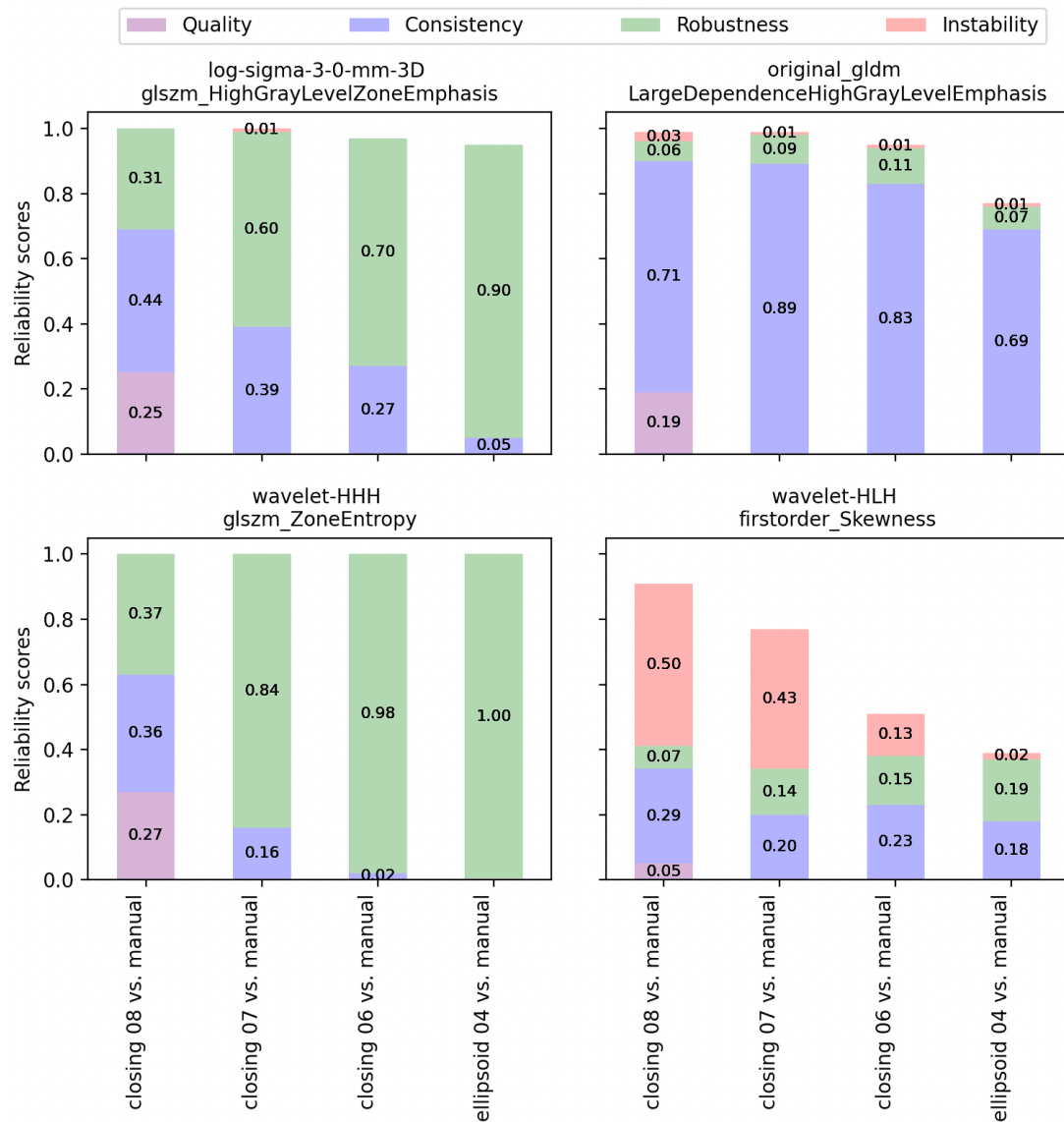


Figure 2.8: Reliability scores of the four common SHAP features, computed for each modification of the manual segmentation (manual vs. closing 08, 07, 06, and ellipsoid 04).

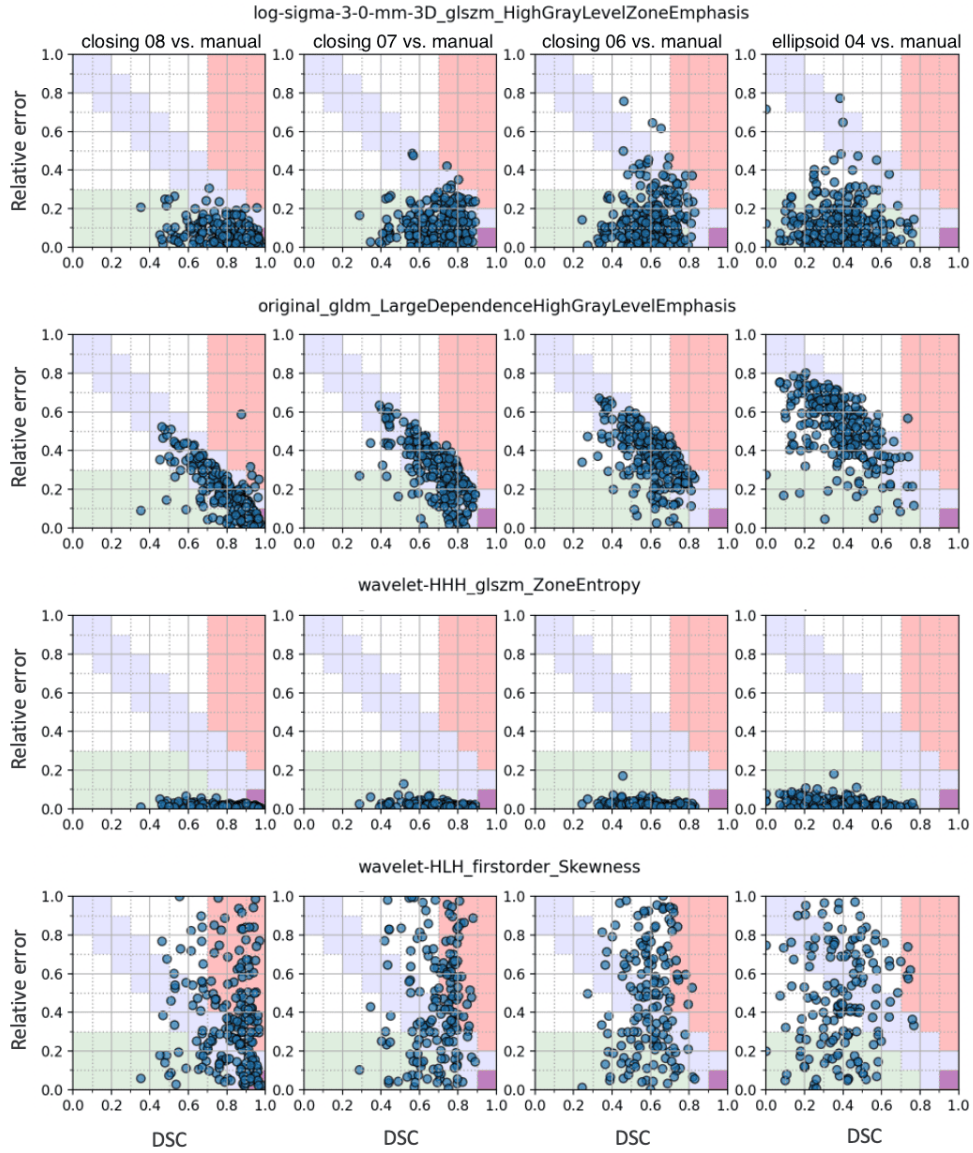


Figure 2.9: Scatter plots of relative error on feature value against DSC, for the four common SHAP features, across segmentation masks (from left to right: manual vs. closing 08, 07, 06, and ellipsoid 04).

2.4 Discussion

We developed and evaluated ML models for the investigation of the role of radiomic features in predicting breast cancer subtypes (TNBC vs. non-TNBC) across various segmentation masks. We focused on radiomic features extracted from the segmentation of the pre-operative first post-contrast DCE-MRI, while other imaging phases and sequences were not considered for this step. We used both an extended set of radiomic features generated by a preliminary selection, and a more refined set (best-SHAP features) provided by the SHAP method. Additional models were trained with demographical and biopsy features.

Statistical tests show that no individual feature is able to differentiate TNBC and non-TNBC groups, which remarks the challenge of relying solely on radiomics for the automatic differentiation of breast cancer subtypes.

The results of the ML models show that using a selection of very explicative radiomic features (the best-SHAP features) leads to a prediction performance that is comparable to using biopsy variables, for the binary prediction of tumor subtypes. Interestingly, prediction results of the best-SHAP models do not exhibit significant statistical difference across segmentation masks, i.e., models trained with features extracted from the manual segmentation and its variations produce comparable performance (see Figure 2.6 and Table 2.4). The prediction obtained for the ellipsoid 04 mask (an ellipsoid over the region of interest provided by the dataset) is particularly surprising. Not only this segmentation mask is extremely poor in terms of DSC, but it also represents a very rough approximation of the tumor shape, missing all contour details and incorporating contrast and irregularity of the surrounding regions of the tumor, like veins, necrosis, edema, and fat tissue (see Figure 2.1, last row).

These results motivated the need to investigate the best-SHAP features selected by each model. We focused our attention on the features that were commonly selected by the five models, as they could elucidate behaviors regarding their stability and robustness across segmentation masks, and offer

insights into the factors influencing the prediction results.

Figure 2.7 compares the ICC and the Pearson correlation of each common feature, both computed between features extracted from the manual mask and its variations (closing 08, 07, 06, and ellipsoid 04). As a general trend, the curves associated to original-gldm Large-Dependence High-Gray-Level Emphasis and wavelet-HHH-glszm Zone Entropy (orange and green) are above the ones associated to LoG $\sigma = 3$ -glszm High-Gray-Level-Zone Emphasis and wavelet-HLH-firstorder Skewness (blue and red), both for ICC and correlation coefficient. We can notice that, according to thresholds reported in literature [21, 61], LoG $\sigma = 3$ -glszm High-Gray-Level-Zone Emphasis and wavelet-HLH-firstorder Skewness would have been discarded by the feature selection step ($ICC < 0.9$) from the prediction model based on the manual mask. However, despite their low ICC values (e.g., lower than 0.75), these features are still considered important and are consistently selected in all predictive models. This observation suggests that relying solely on ICC for feature selection may not be the most effective approach. We recall that the ICC(2) evaluates how much of the total variability in feature values is due to differences between subjects, compared to variability introduced by different segmentations (raters) and residual error. The ICC is high if the radiomic feature values are consistent across segmentations, and most of the variance arises from the differences between subjects. If different raters (segmentation masks) produce significantly different feature values for the same tumor, the inter-rater variance increases, and the ICC drops. On the other hand, Pearson correlation measures only the strength and direction of the linear association between two sets of feature values (e.g., Rater A vs Rater B), but ignores the magnitude of disagreement between raters, and just checks if features move together. Therefore, features derived from different segmentations with high Pearson correlation, i.e., following a similar trend across subjects, might report a low ICC because the ICC punishes systematic differences between raters. Specifically, when examining the Pearson correlation, we find cases where a low ICC does not necessarily correspond to a low correlation coefficient. For example, in Figure 2.7, the

feature `original_gldm_LargeDependenceHighGrayLevelEmphasis` (orange line) has ICC around 0.65 and Pearson correlation around 0.88, when computed between manual vs. ellipsoid 04 mask. The observed discrepancy between ICC and Pearson correlation is likely due to the fact that, for some features, their values change consistently across subjects in response to segmentation variability, resulting in high Pearson correlation and low absolute agreement (ICC). Importantly, the high correlation of a predictive feature with its corresponding values extracted from another segmentation mask suggests that both these features are capturing meaningful information relevant to the target, despite the poor reproducibility rated by ICC.

Additionally, the computation of the reliability scores for the common features, in Figure 2.8, highlights that there is no shared numerical relationship between feature stability and segmentation variability among the predictive features. However, segmentation variability does not disrupt feature stability pattern of individual features, as robust, consistent, and unstable patterns remain well-defined across mask modifications, as displayed in Figure 2.9.

These considerations suggest that the feature selection process in predictive models is not inherently linked to feature stability with respect to segmentation variability, neither from a numerical nor from a reproducibility perspective (ICC), but rather to general patterns captured by the distribution of features across patients. This underscores why feature selection approaches based on ICC or reliability scores may fail to identify the most predictive features for a specific task, highlighting that exclusive reliance on ICC or stability metrics for feature selection may inadvertently discard informative features. On the other hand, the high correlation between the same predictive features extracted from different segmentation masks helps explaining why the best-SHAP models yield similar performance.

Overall, the results here presented suggest that high segmentation accuracy may not be an imprescindible requirement for radiomics applications, and that including partial peritumoral imaging information through various segmentation masks does not hamper feature predictive power. However,

these findings are context-specific (TNBC prediction using the Duke dataset on first post-contrast MRI) and do not imply that segmentation precision is unimportant in all radiomic tasks; for certain applications, accurate delineation may still be essential to achieve reliable predictions. Nonetheless, this study carries important implications for the field of radiomics, as it challenges the assumption that precise segmentation is always critical for feature selection and predictive performance. It opens the possibility that, in some scenarios, particularly predictive modeling, radiomics may tolerate greater segmentation variability than expected, without sacrificing outcome quality. By showing that model performance can be preserved despite segmentation differences, this work supports the development of more flexible and time-efficient radiomics pipelines, potentially enhancing their feasibility and scalability for clinical integration.

To enhance our analysis, less deterministic methods could be considered to simulate variations in segmentation, such as introducing contour noise or incorporating voids within the segmented region. While a comprehensive evaluation of the entire tumor is desirable for accurate histological characterization [64, 186], these modifications would more closely simulate the histological examinations commonly used to determine tumor subtypes, as they are typically performed on localized and partial tumoral regions. Conversely, making finer adjustments with targeted deformations of the tumor contour, like artificially adding spiculated parts, may be of limited relevance or even counterproductive, as irregularities along the tumor margins are generally linked to Luminal B and HER2-enriched subtypes [103, 162].

To position our results within the current state-of-the-art for subtype prediction with radiomics on MR imaging, Leithner et al. [101] used linear discriminant analysis followed by k-nearest neighbor classification for separation of molecular subtypes. The study included 143 with biopsy-proven breast cancer who underwent CE-MRI. The reported accuracy for classification of TNBC vs others is 0.74, obtained with a leave-one-out testing method. By

combining radiomics data extracted from both first post-contrast T1-weighted imaging and Apparent Diffusion Coefficient (ADC) mapping they reached an accuracy of 0.85 (0.85, 0.91) and ROC-AUC of 0.86 (0.77, 0.92) [100], however no independent test set was used. In [155] the authors used the Duke dataset, including 922 patients with pre-contrast, first post-contrast, and subtraction imaging, with various masks for each patient: tumor, breast mask, and fibroglandular tissue mask. The feature selection was based on literature review breast MR image processing [157]. By using a machine learning-based multivariate models they obtain a classification performance of TNBC reaching ROC-AUC of 0.654. In [167], the authors used DCE-MRI T1 images of 190 patients (32 TNBC) to train random forest, logistic regression, support vector machine, naïve Bayes, and eXtreme Gradient Boosting. Lasso and T-tests were used for feature selection. The reported results are averaged across 5 data splits. EXtreme Gradient Boosting model showed the best performance in the TNBC vs non-TNBC classification task, with AUC 0.903 but specificity of 0.444, showing that the model fails to classify TNBC class correctly, as it is less represented. The logistic regression model reaches 0.707 ROC-AUC, which is comparable to our performance. In our analysis, all the other models exhibited significant overfitting despite an extensive hyperparameter search. Consequently, we opted to use logistic regression, which, while being a linear classifier, offers the advantage of inherent feature selection through the L1 regularization penalty.

Among the paper using deep learning techniques for the identification of TNBC, [203] used a pre-trained ResNet18 obtaining 0.920 (0.863, 0.977) ROC-AUC with T1 first post-contrast image, 0.851 (0.770, 0.932) with ADC, and 0.697 (0.588, 0.806) with T2, indicating superior performances compared to the ones provided by radiomics-based ML models. In another work [206], a CNN and a recurrent CNN were used to differentiate three breast cancer molecular subtypes on MRI. The DCE-MRI series consisted of one pre-contrast and five post-contrast frames. The convolutional long short-term memory (CLSTM), was applied to track the temporal information of the changed signal intensity

in the DCE time sequence. For TNBC vs. non-TNBC (10 TNBC vs 89 non TNBC), the models yielded AUC 0.84 for CNN and 0.89 for CLSTM on the training set, while the performances on an independent test set are shown in the case of 3-class prediction of TNBC vs. HER2+ vs. HR+/HER2-, yielding accuracy of 0.74-0.82.

Regarding mammography, this study [115] aimed to investigate whether quantitative radiomic features extracted from digital mammogram images are associated with molecular subtypes of breast cancer (331 Chinese patients, 29 TNBC). The Naive Bayes machine learning scheme was employed for the classification, and the least absolute shrink age and selection operator method was used to select the most predictive features for the classifiers. The model used a combination of both the craniocaudal and the mediolateral oblique view images, achieving the overall best performance than using either of the two views alone, yielding ROC-AUC of 0.865 and accuracy of 0.796 for triple-negative vs non-triple-negative. Another work [173] conducted a study to predict the molecular subtype using clinical and radiomics data (221 patients, 62 TNBC). Clinical data such as patient age, lesion size, mammographic features based on the American College of Radiology's Breast Imaging Reporting and Data System (BI-RADS) criteria [45], and radiomic features were used. An elastic-net logistic regression model was used to create the radiomics signature of each lesion. They found that predictions using a combination of radiomics and clinical data outperformed those using clinical data alone, yielding an AUC of 0.838 in distinguishing TNBC vs non-TNBC, but no averaging across data split is reported. Finally, [205] employs a multi-modal deep learning framework for predicting the molecular subtypes of breast cancer directly from mammography and ultrasound images of 3360 patients. Various ResNet models and Inception models are trained, along with a multi-modal deep learning with intra- and inter-modality attention model for 4-class classification task. The accuracy of the trained models ranges from 81.1 to 88.5, while the confusion matrices show that 66.7% to 71.7% of TNBC was correctly classified, across the proposed models.

To enhance predictive performance in TNBC classification, incorporating additional established risk factors, such as demographic information (e.g., ethnicity) and breast-specific characteristics (e.g., breast density, background parenchymal enhancement, and fibroglandular tissue [124]), could enhance model performance, as these are recognized contributors to TNBC risk [96, 123]. These measures of risk are typically evaluated by radiologists using mammography or MRI, following the BI-RADS guidelines [1], or through automated methods [81]. However, while the integration of these variables may boost model performance, the key advantage of radiomics lies in its potential to yield valuable insights independent of conventional radiological assessments. Furthermore, recent advancements [4, 49, 120, 208] have demonstrated that leveraging both radiomics and deep learning could advance risk prediction in breast cancer beyond what is achievable through manual or visual assessments alone, by capturing spatial patterns and subtle textures. This capability could potentially serve as a proxy for variables like breast density and parenchymal complexity. This underscores the importance of systematically validating the radiomics approach in medical imaging, which is why the primary contribution of this study.

This study presents some limitations. First, the variability range of the ROC-AUC is not particularly narrow, a concern previously highlighted by Montoya-del-Angel et al. [125] in other radiomics-based applications for breast cancer. Second, the dataset we used was obtained from a single institution, therefore no independent testing was conducted. Although cross-validation helps reducing overfitting on specific test sets, future work should involve validating the method with larger, multi-institutional datasets to better assess reproducibility and generalizability of the results. MAMA-MIA dataset [60] could be used for this purpose as it is a curated collection of pre- and first post-contrast imaging from ISPY1 [47], ISPY2 [126], NACT [48], Duke MRI [156] datasets, the last one being the one employed for this work, including expert segmentations of first post-contrast phase. The integration of such cohorts

within the training set could sharpen the prediction result, ultimately strengthening the robustness of the models used in this critical analysis of radiomics stability. Additionally, utilizing a larger dataset could establish a benchmark for the prediction model within the clinical context of subtype prediction, facilitating a robust comparison with the ambitious performances shown by DL and reported in the previous paragraph. However, the set of extracted first-post-contrast texture features and tumor morphology features may not adequately capture the full spectrum of complexities involved in tumor subtype classification. Future investigations should incorporate time-dependent texture descriptors and heterogeneity features related to enhancement kinetics into the ML models, as illustrated in Caballo et al. [20] for other prediction purposes. Future work may also include features extracted from other MRI sequences, e.g., Diffusion Weighted Imaging (DWI) and ADC map, to further improve cancer characterization.

Chapter 3

Radiomics for longitudinal prediction

In oncology, as well as for neurogenerative diseases and other fields of medicine, medical imaging plays a vital role in diagnosis, treatment planning, and therapy execution. Longitudinal image analysis enables the tracking of temporal changes, facilitating disease monitoring, treatment assessment, and outcome prediction. However, analyzing longitudinal images poses several challenges, such as managing variable imaging intervals and handling sparse data [180]. The use of radiomics for longitudinal prediction has been explored in literature for various applications. The two primary approaches to incorporate longitudinal information across timepoints include feature concatenation and delta radiomics. The first one involves extracting radiomic features from multiple timepoints and combining them to construct a comprehensive radiomics-based representation of disease evolution. This method enables to capture longitudinal changes in tissue characteristics that may not be evident from a single timepoint analysis. However, a significant drawback is the increased dimensionality of the model, which can introduce complexity and the risk of overfitting. The second approach, delta radiomics, quantifies the relative net change in radiomic features between different imaging timepoints, providing a characterization of tumor alterations over time. Even so,

both feature concatenation and delta radiomics face challenges in determining optimal regions for radiomic feature extraction at subsequent timepoints, for example after tumor resection, making it difficult to accurately characterize disease evolution. Additionally, missing data across time points can limit the applicability of such models for longitudinal disease characterization. For these reasons, some studies rely exclusively on pre-operative or onset radiomics to forecast successive follow-ups, even though this last approach is not intrinsically longitudinal.

The aforementioned longitudinal radiomics techniques have been applied in diverse clinical contexts. Baseline radiomics extracted from bone CT has been used to predict the 1-, 3-, and 4-year risk of osteoporotic vertebral fractures [193]. Feature concatenation, as well as 3D ResNet, has been used to predict tuberculosis treatment outcomes at six months, using features from pre-treatment and second-month follow-up CT [129]. In the field of oncology, a study showed that concatenating pre- and post-neoadjuvant chemotherapy CT radiomics, along with clinical and pathological data, could predict the pathological complete response of metastasized axillary lymph nodes in breast cancer using a Lasso model [192]. Another study developed longitudinal XGBoost models to predict progressive brain metastases after the third post-surgery timepoint [34], using radiomic features from MR images of four sequential timepoints. Regarding delta radiomics, it has been used in DCE-MRI to predict response to neoadjuvant chemotherapy in breast cancer [55], in CT imaging for early recurrence detection of hepatocellular carcinoma [166], and in CT and 18F-FDG-PET imaging for outcome prediction in head and neck cancer using pre-treatment and in-treatment scans [175]. Typically, delta radiomics is applied to data from two consecutive timepoints, but some studies extend this approach to three or more [180], where pairwise delta features are computed between timepoints and combined in different ways within predictive models [2]. To capture deeper longitudinal patterns, some works perform a linear regression on radiomic features, and use the slope parameter as a new feature that captures longitudinal information. This method was

employed in a study on CT imaging for outcome prediction in non-small cell lung cancer [182]. Other studies have integrated multiple longitudinal radiomics approaches within a single framework, both in oncology [56] and neurodegenerative diseases [146].

This chapter presents a study on the longitudinal response to treatment of patients affected by glioblastoma (GBM) [24]. We introduce a longitudinal model that is capable of enhancing the predictive power of baseline radiomics in forecasting future responses, eliminating the need for radiomics data from intermediate timepoints. To achieve this, we leverage pre-operative radiomics to model the probability distribution of treatment response at an intermediate timepoint, and later combine baseline radiomics and first response information to predict subsequent response to treatment.

The structure of this chapter is as follows. Section 3.1 discusses the rationale for investigating radiomics-based longitudinal prediction of GBM treatment response. Section 3.2 details the proposed longitudinal radiomics approach, describes the custom-designed synthetic longitudinal dataset and the real-world dataset [176], and outlines the machine learning models. Section 3.3 presents the results of longitudinal prediction experiments. Finally, Section 3.4 provides a comparative analysis of the experimental findings.

3.1 Motivation

GBM is one of the most aggressive primary brain neoplasms, with an average annual incidence of 1-5 per 100,000 individuals [179]. The median survival rate is approximately 15 months, as these tumors can infiltrate the surrounding healthy tissues with consequent predisposition to recurrence [95]. Despite surgical resection, adjuvant radiotherapy, and chemotherapy, recurrence is frequent, leading to overall poor survival rates. Early recurrence detection can facilitate timely therapeutic intervention and treatment that may consequently delay the relapse [37]. The treatment response is routinely evaluated on MRI, according to the Response Assessment in Neuro-Oncology criteria

(RANO)[197]. The quantitative assessment of RANO is concerned with measuring the active tumor appearing on contrast-enhanced T1-weighted MRI, while a qualitative evaluation is based on checking abnormalities on the T2-weighted and fluid-attenuated inversion recovery (FLAIR) MRI. Potentially, radiomic features can provide deeper insights into tumor changes, serving as early indicators of GBM recurrence and treatment response. However, predicting treatment response from pre-operative MRI alone may be suboptimal, as the time interval between initial diagnosis and postoperative assessments varies significantly, leading to considerable variations in tumor phenotypic characterization [37]. This highlights the need to analyze not only preoperative but also multiple postoperative timepoints, to comprehensively map GBM progression.

Only few studies have explored longitudinal MRI analysis to track GBM progression [37, 153], and none have specifically focused on treatment response. The work in [37] conducted a longitudinal radiomics analysis of 29 subjects with three or more follow-ups and multimodal MRI (FLAIR, ADC, and T1CE), analyzing if pre-recurrence radiomics could predict GBM recurrence patterns. Paper [153] used 3D nnU-Net models to assess changes on tumor tissue by confronting the morphological composition of pre- and post-treatment diffuse glioma, including GBM, using T1 and FLAIR imaging of 298 patients. Notably, the review [180] indicates that between 2011 and 2023 no oncological longitudinal radiomics studies with more than 100 patients focused on glioblastoma. This is likely due to the fact that most of the publicly available GBM MRI datasets lack extensive longitudinal follow-up or treatment response [6, 8, 118]. The Lumiere dataset [176] is the first publicly available longitudinal MRI dataset with expert RANO treatment response ratings and clinical data, including demographics, O6-methylguanine-DNA methyltransferase (MGMT) promoter methylation and isocitrate dehydrogenase 1 (IDH1) mutation status.

In this study, we introduce a method to advance longitudinal radiomic analysis for GBM response prediction. Our approach mitigates a critical

limitation of state-of-the-art radiomics-based longitudinal prediction methods, which typically require feature extraction from multiple timepoints, as explained in the introduction of this chapter. This requirement limits their applicability in prospective settings or in presence of incomplete follow-up data, hindering their ability to effectively handle sparse longitudinal datasets. Our method utilizes baseline radiomics but also additionally incorporates a probabilistic framework to forecast future response that quantifies uncertainty of intermediate outcomes, eliminating the need for radiomics data from intermediate follow-ups, and ultimately enhancing the longitudinal prediction. More specifically, we build a predictive model that uses radiomic data from the pre-operative timepoint in combination with a probabilistic estimate of the first post-operative treatment response to forecast treatment outcomes at the subsequent stage. We validate our model on both a synthetic longitudinal dataset and on radiomics data from GBM patients of Lumiere dataset [176]. Numerical results show that our probabilistic model for longitudinal prediction of treatment response performs similarly to SOTA references [138], consistently improving the longitudinal prediction solely based on baseline features, even with reduced dataset size. By developing this longitudinal model, our work aims to advance longitudinal radiomic analysis from a methodological perspective. Moreover, this approach might have potential orienting treatment decisions based on a dynamic understanding of a patient’s disease profile, and can be generalizable beyond oncology and radiomics to other longitudinal applications.

3.2 Materials and methods

3.2.1 Longitudinal approach

This section presents a novel, probabilistic approach to longitudinal prediction of treatment response, integrating baseline features with a probabilistic

estimate of intermediate responses, to forecast the response at successive follow-ups. The longitudinal method we propose is applicable to any number of timepoints. However, for simplicity and without loss of generality, in the following we refer specifically to the first and second follow-ups after the baseline timepoint. In the context of predicting GBM response to treatment, baseline features can be interpreted as pre-operative radiomic features, first follow-up features as those extracted from the first post-operative imaging, while the response at the first and second follow-up refer to the response to treatment at the first and second post-operative follow-up, respectively.

Let $\mathcal{C} = \{0, 1\}$ be the target label space, and $\mathcal{F} \subset \mathbb{R}^M$ be the space of radiomic features extracted at the baseline timepoint, with M number of features. The subscript i indicates the i -th data split.

For the i -th data split, $f_1^i, f_2^i : \mathcal{F} \rightarrow \mathcal{C}$ are the ML models to predict the response at the first and second follow-up, respectively, using baseline radiomic features, and $i = 1, \dots, \text{\#splits}$. These models, f_1^i and f_2^i , will be referred to as *baseline models*. For the j -th patient of the i -th data split, $x^j \in \mathcal{F}$, $f_1^i(x^j) = \text{proba}_1^{i,j}$ and $f_2^i(x^j) = \text{proba}_2^{i,j}$ are the probabilities of belonging to class 1 outputted by the models, while $y_1^{i,j} = \mathbb{1}_{\{\text{proba}_1^{i,j} \geq 0.5\}}$ and $y_2^{i,j} = \mathbb{1}_{\{\text{proba}_2^{i,j} \geq 0.5\}}$ are the predicted class labels for response to treatment at first and second follow-up, with $y_1^{i,j}, y_2^{i,j} \in \mathcal{C}$.

An additional classifier is designed to predict the response at the second follow-up using the combination of baseline radiomic features and the response at the first follow-up, as input feature, to explicitly account for the intermediate treatment outcome in the training phase. For the i -th data split, $f_2^{L,i} : \mathcal{F} \times \mathcal{C} \rightarrow \mathcal{C}$ constitutes the *longitudinal model*.

In the experimental setup, the true response labels at intermediate follow-up are available, allowing the model $f_2^{L,i}$ to be tested using baseline radiomics and actual treatment responses $y_{1,true}^{i,j}$. However, this is not the case for prospective studies, in which only baseline features are available (e.g., for newly enrolled patients who only have pre-operative radiomics from their scans), or in case

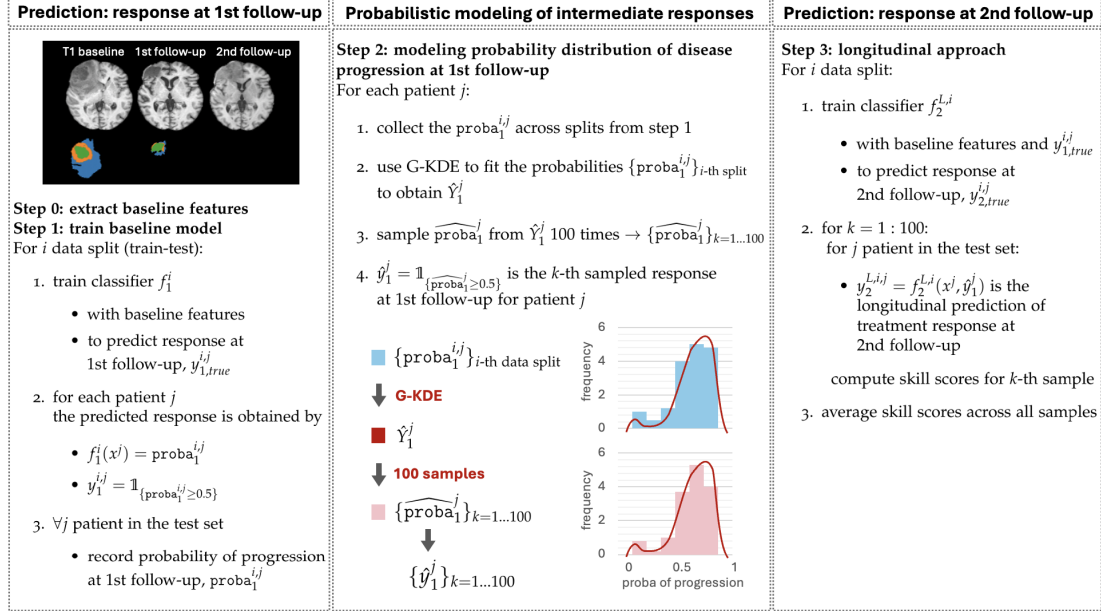


Figure 3.1: Proposed probabilistic approach for the longitudinal prediction of response to treatment.

of missing data. If treatment response is assumed to be unavailable during the model's testing phase, modeling the feature representing the intermediate response becomes crucial to benefit from the longitudinality of the model itself. In this work, the feature 'response at 1st follow-up' is modeled in the following ways:

- by using the predicted label $y_1^{i,j}$;
- by using the label $\hat{y}_1^j$, obtained by sampling from the probability distribution $\hat{Y}_1^j$ and thresholding, where $\hat{Y}_1^j$ is an approximation of the ideal probability distribution Y_1^j of presenting response of class 1 (progressive disease) at the first follow-up for each patient.

The last option allows to incorporate in $f_2^{L,i}$ the uncertainty of the prediction outcomes at the first follow-up ,i.e., the intermediate timepoint, derived from the ML model f_1^i .

To approximate the distribution Y_1^j , we leverage the prediction outcome of the models f_1^i across multiple data splits. Whenever a patient x^j from the

i -th split is included in the test set, we record their probability $\text{proba}_1^{i,j}$ of presenting class 1 response at the first follow-up, as predicted by the current trained model f_1^i . This procedure is repeated across all splits. As a result, each patient has a number of probabilities corresponding to the number of times they appear in the test set of a split. Then, for each patient, Gaussian Kernel Density Estimation (G-KDE) [163, 169] is used to model the distribution Y_1^j by fitting the probabilities $\{\text{proba}_1^{i,j}\}_{i\text{-th data split}}$, thus obtaining $\hat{Y}_1^j$. For testing the longitudinal model $f_2^{L,i}$, $\widehat{\text{proba}}_1^j$ is sampled from $\hat{Y}_1^j$ one hundred times, the probabilities are thresholded as $\hat{y}_1^j = \mathbb{1}_{\{\widehat{\text{proba}}_1^j \geq 0.5\}}$, and the response labels at first follow-up $\hat{y}_1^j$ are assigned to each patient. For the j -th patient of the i -th data split, $y_2^{L,i,j} = f_2^{L,i}(x^j, \hat{y}_1^j)$ is the longitudinal prediction of the response at the second follow-up. In this way, the longitudinal model is tested one hundred times, one for each sampling, and the prediction results can be averaged. A step-by-step explanation of our probabilistic approach to longitudinal prediction of treatment response is reported in Figure 3.1.

3.2.2 Data

To evaluate the performance of the longitudinal approach introduced in Section 3.2.1, based on the probabilistic modeling of intermediate treatment response, we apply the described method to a synthetic tabular dataset consisting of three timepoints - baseline, first, and second follow-up - as well as to the Lumiere dataset, a publicly available radiomic dataset on GBM progression [176].

Lumiere longitudinal dataset

Lumiere dataset [176] is a single-center longitudinal GBM MRI dataset with expert ratings of selected follow-up studies according to RANO criteria: progressive disease (PD), stable disease (SD), partial response (PR), or complete response (CR). This collection includes MRI data of 91 GBM patients with a total of 638 study dates and 2487 images (age at surgery: mean 62.4 years, std 10.3

years, min 39.7 years, max 80.4 years; sex: 44 female, 47 male), having different and follow-up intervals (overall survival: mean 589.1 days, std 334.0 days, min 25 days, max 1412 days; time to progression: mean 305.4 days, std 266.1 days, min 25 days, max 1537 days). Figure 3.2 illustrates all the patients included in Lumiere dataset, along with the response assessment at each timepoint across available follow-ups. For a subset of patients, pathology information regarding MGMT methylation (11 not available) and IDH1 are present (23 not available). The dataset includes T1-weighted pre- and post-contrast, T2-weighted, and FLAIR MRI. Segmentations obtained through the automated segmentation tools DeepBraTumIA (<https://www.nitrc.org/projects/deepbratumia>) and HD-GLIO-AUTO [74, 88] (<https://github.com/CCI-Bonn/HD-GLIO>) are included, both based on U-Net deep learning architecture [149]. Both methods include a co-registration of the four MRI sequences. HD-GLIO-AUTO registers all sequences to a reference image chosen automatically during processing, and DeepBraTumIA registers to an atlas such that all studies may be analyzed in the same space. HD-GLIO-AUTO provides segmentations for the contrast-enhancing tumor and the T2-signal abnormality. DeepBraTumIA outputs labels for necrosis, contrast enhancement, and edema. For this study, DeepBraTumIA segmentations were used.

Subject to data availability, radiomic features are provided by Lumiere dataset, extracted from the co-registered and resampled images for each segmentation label (PyRadiomics version 3.0.1). Z-score normalization, scaling by a factor of 100, and intensity shifting by 300 are used to harmonize the value range. Feature types include first-order statistics, 3D shape, GLCM, GLRLM, GLSZM, NGTDM, and GLDM features, for a total of 110 features for each mask label. No image filtering is considered for feature extraction.

To obtain a homogeneous longitudinal dataset suitable for testing our methodology, we curated Lumiere dataset to include patients with a minimum of three timepoints (baseline, first follow-up, and second follow-up). Patients who exhibited complete or partial treatment responses were excluded, as

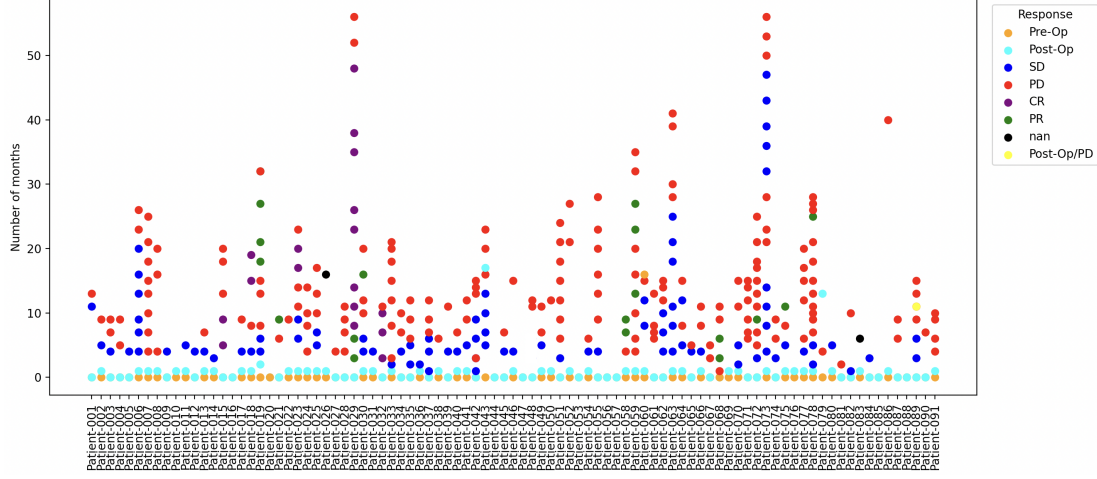


Figure 3.2: All the patients included in the Lumiere dataset, with the response assessment at each timepoint. Pre-Op and Post-Op indicate pre- and post-operative timepoint; SD stable disease; PD progressive disease; CR complete response; PR partial response; nan indicates missing information about the timepoint; Post-Op/PD indicates second post-operative progression.

their numerosity across all timepoints was insufficient to support training a four-class model (CR vs. PR vs. PD vs. SD). Consequently, our experiments focused on distinguishing between stable disease ($SD = 0$) and progressive disease ($PD = 1$). Given the variability in follow-up intervals among patients in the Lumiere dataset, we introduced an inclusion criterion based on the maximum time from baseline (pre-operative stage) to the first and second post-operative follow-up. Specifically, the first follow-up was required to occur within 20 months from the baseline timepoint, and the second follow-up within 20 months from the first. Patients whose first follow-up occurred more than 20 months after surgery, or whose second follow-up exceeded 20 months after the first, were excluded from the analysis.

For training our longitudinal model we required radiomics from all segmentation labels, i.e., contrast enhancement, necrosis, and edema, at the baseline timepoint on T1-weighted imaging. Following this selection, 47 patients were deemed eligible. Additional selection was applied for the benchmark experiments, described in Section 3.2.3, involving first follow-up radiomics, requiring

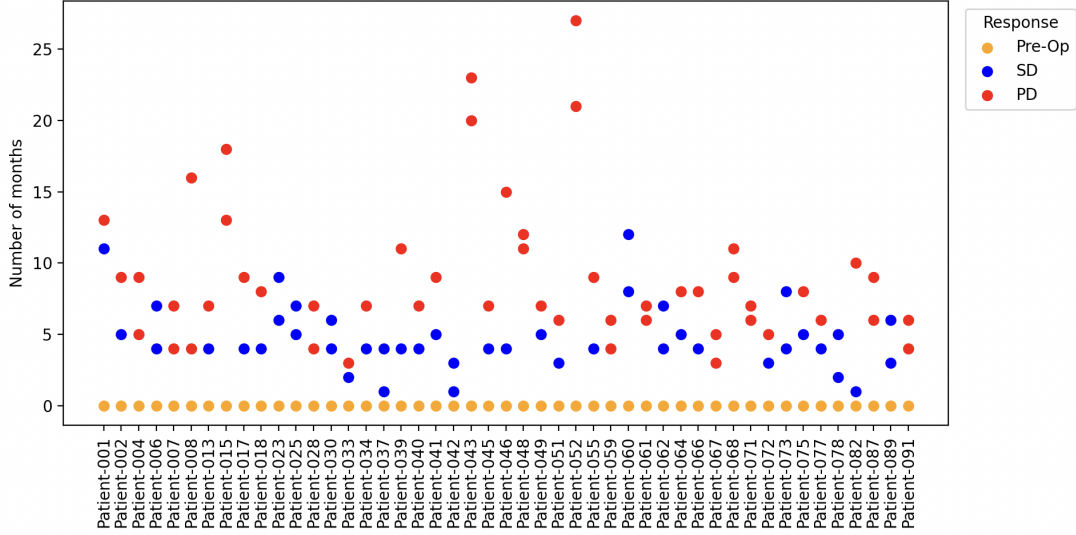


Figure 3.3: Subset of Lumiere dataset meeting the inclusion criteria, with response assessment at the first and second follow-up after baseline (pre-op label in the legend). For the benchmark experiment based on radiomics at the first follow-up, exclude patients 2, 7, 23, 30, 39, 60, 62, and 89.

radiomics from the complete set of segmentation labels on T1 images at the first follow-up. After this step, 38 patients remained eligible for the analyses. Figure 3.3 displays patients meeting the inclusion criteria.

Synthetic longitudinal dataset

Following the structure of Lumiere longitudinal dataset, we construct a synthetic version that preserves the same proportions of outcome labels at the first and second timepoints (stable vs. progressive disease). Our objective is not only to replicate the outcome distribution but also to ensure that the generated features follow a coherent logic. Specifically, we aim to create a longitudinal dataset where we know a feature-based separation is possible, in contrast to the real-world scenario, where such separation may not necessarily exist, or be very difficult to achieve, and use it to validate our methodology for longitudinal prediction in a simulated setting. With this objective, we used a linear combination of baseline features to simulate the probability of progression at the first follow-up timepoint. Subsequently, we simulate a second

follow-up by applying the transition probabilities derived from the Lumiere dataset, calculated based on the observed frequencies of transitions from stable to progressive and vice versa.

To construct the synthetic longitudinal dataset, we utilized existing radiomic features extracted from the original T1 images of the Lumiere dataset, randomly chosen among the available features: shape Surface-Volume Ratio (x_1), first order Median (x_2), first order Entropy (x_3), glcm Difference Entropy (x_4), and glcm Maximum Probability (x_5). Such features were considered as baseline features for the synthetic longitudinal dataset. To construct the first follow-up response, we followed this procedure. First, we drew the distribution of these features using the Kernel Density Estimation (G-KDE) method [163, 169] and sampled 300 data points from the distributions. Next, we assigned a coefficient to each feature. The vector of coefficients is denoted as α . The features were then standardized within the range $[-1, 1]$, therefore $x = [x_1, x_2, x_3, x_4, x_5]$, $x \in [-1, 1]^{300 \times 5}$. Then, by multiplying the matrix x with α^T , we obtained a vector $p \in \mathbb{R}^{300 \times 1}$ containing the scoring of each response at the first follow-up and representing the likelihood of having progressive disease. These scores were then converted into a binary column with values 0 and 1, based on whether their value was non-negative (assigned 1) or negative (assigned 0). The obtained labels represented the outcome of the first timepoint of the synthetic dataset. We chose $\alpha = [0.4, 0.8, 0.5, 0.7, 0.2]$ through a trial-and-error approach, ensuring that the linear combination of the selected features would result in balanced outcome labels at the first follow-up, according to the outcome proportions provided by Lumiere dataset at the first follow-up.

To simulate the second follow-up response, we introduced transition probabilities inspired by the Lumiere dataset, but we deliberately emphasized them to obtain a clearer stratification of patients and a more homogeneous dataset, accentuating transitions which are underrepresented in the original data, such as the shift from progressive to stable disease. Specifically, we identified four score ranges, $(-\infty, -1)$, $[-1, 0)$, $[0, 1]$, and $(1, +\infty)$, obtained for the first synthetic follow-up response. All cases with extreme positive scores, indicating

progression, remain progressive. The majority of patients who were stable at the first follow-up, i.e., with extreme negative score at the first follow-up, transition to a progressive state. These align with the trends observed in the Lumiere dataset. Additionally, we introduced the possibility that uncertain or borderline progressive cases (i.e., those with low positive progression scores at the first follow-up) may revert to a stable condition. Therefore, the labeling for the second follow-up response was conducted as follows.

- All the patients having scores in $(1, +\infty)$ were labeled as 1, reflecting the persistence of a severe condition (i.e., progressive patients remaining progressive).
- 90% of patients were sampled from the extreme low scores $(-\infty, -1)$. These patients were labeled as 1, reflecting a transition of patient condition from the first to the second follow-up (i.e., stable patients becoming progressive).
- All patients having scores in $[-1, 0]$ were labeled as 0, simulating stabilization of patient condition between first and second follow-up (i.e., stable patients remain stable).
- 50% of patients were sampled from the range of moderate positive scores $[0, 1]$ and labeled as 0, simulating partial improvement in patient condition between first and second follow-up (i.e., transition from progressive to stable disease).

The Principal Component Analysis (PCA) representation of the synthetic dataset is showed in Figure 3.4.

3.2.3 Benchmark experiments

In order to benchmark the proposed probabilistic model for longitudinal prediction of treatment response, we conducted a series of experiments to predict the second follow-up response using radiomics from Lumiere dataset and

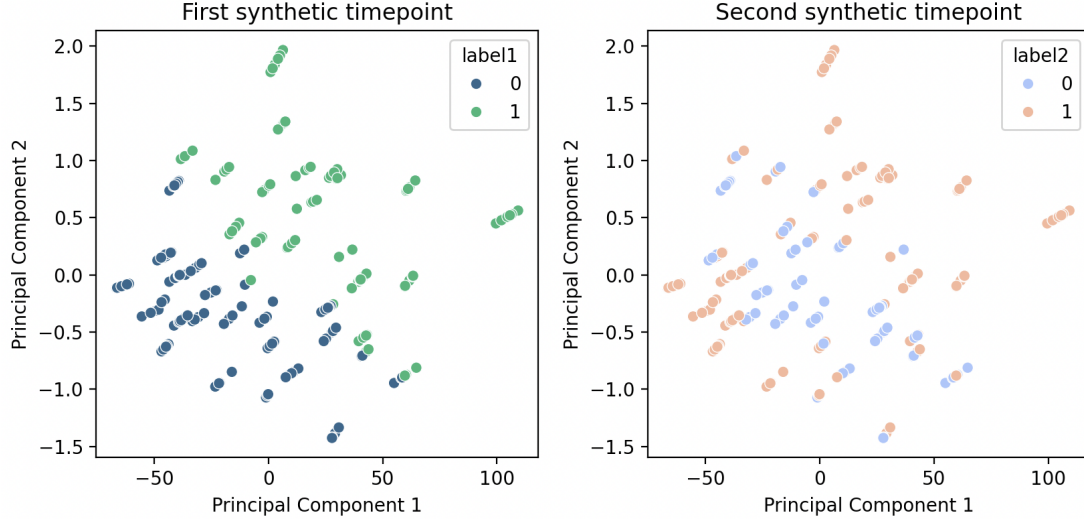


Figure 3.4: Synthetic longitudinal dataset overview. Left panel: PCA representation of the synthetic dataset colored by the target label at first timepoint. Right panel: PCA representation of the synthetic dataset colored by the target label at second timepoint.

the custom-designed synthetic dataset. Our goal was to assess whether our longitudinal approach could outperform existing radiomics-based methods for longitudinal prediction. Specifically, we trained the following models to predict the second follow-up response to treatment (model names derive from the set of features used for training). With the Lumiere dataset:

- prediction of the second follow-up response using baseline radiomic features, ‘baseline radiomics’ model (also referred to as f_2^i , in Section 3.2.1);
- prediction of the second follow-up response using the true response labels at the first follow-up (i.e., RANO labels), ‘true labels at 1st follow-up’ model;
- prediction of the second follow-up response using radiomics extracted from the first follow-up imaging, when available, ‘radiomics at 1st follow-up’ model;

- prediction of the second follow-up response with delta radiomics (relative change between features at first follow-up and baseline), 'delta radiomics' model.

A model based on radiomic feature concatenation at the first and second follow-ups was not considered due to the excessive number of features relative to the limited number of patients.

Similarly, for the synthetic dataset:

- prediction of the second follow-up response using baseline synthetic features, 'baseline synthetic' model (also referred to as f_2^i , in Section 3.2.1);
- prediction of the second follow-up response using the synthetic response labels at the first follow-up, 'synthetic labels at 1st follow-up' model.

The performances of these models were compared with the ones of the longitudinal model $f_2^{L,i}$, described in Section 3.2.1. We recall that this model was trained with baseline features (radiomic or synthetic features) and the feature 'response at 1st follow-up'. The testing phase, to evaluate model performance, was performed with various approaches to modeling the feature 'response at 1st follow-up'. The longitudinal models were named according to the way this feature was modeled. Specifically, in both the Lumiere and synthetic case, the feature 'response at 1st follow-up' was modeled:

- by using the true response labels at the first follow-up, $y_{1,true}^{i,j}$, leading to 'baseline + true labels' and 'baseline + synthetic labels' models, using Lumiere and synthetic dataset, respectively;
- by using the predicted response labels at the first follow-up, $y_1^{i,j}$, obtaining the 'baseline + predicted labels' model (we use the same model name for both Lumiere and synthetic case);
- by using the sampled response label $\hat{y}_1^j$, obtained as described in Section 3.2.1, producing the 'baseline + G-KDE labels' model, that is our proposed

method (we use the same model name for both Lumiere and synthetic case).

To compute the predicted response labels at the first follow-up, $y_1^{i,j}$ and the sampled response label $\hat{y}_1^j$, we trained a model for the prediction of the first follow-up response with baseline features, referred to as ‘baseline radiomics’ or ‘baseline synthetic’ model depending on the dataset used for training (either Lumiere or synthetic), also indicated with f_1^i in Section 3.2.1. The performance of this model is reported among the results; however, it should not be directly compared to the longitudinal prediction at the second follow-up. Instead, it may serve as a reference point for understanding the difference in the predictive power of baseline radiomics for forecasting the response to treatment at the first versus the second follow-up.

All the models in this study that were trained with Lumiere radiomics included additional demographics features, age at surgery and sex, along with the feature ‘number of months to the first (second) post-operative follow-up’, to account for variations in follow-up intervals across patients when predicting the first (second) post-operative treatment response. A summary of the models employed for this study is reported in Table 3.1.

Table 3.1: Summary of the performed experiments. The ‘Model’ column indicates the model names. The ‘Longitudinality’ column specifies whether the model leverages information from multiple timepoints, indicating a truly longitudinal approach. The ‘Synthetic’ and ‘Lumiere’ columns indicate whether the models were trained on the respective datasets.

Model	Longitudinality	Synthetic	Lumiere
Baseline {radiomics, synthetic}	✗	✓	✓
{True, synthetic} labels at 1st follow-up	✗	✓	✓
Radiomics at 1st follow-up	✗	✗	✓
Delta radiomics	✓	✗	✓
Baseline + {true, synthetic} labels	✓	✓	✓
Baseline + predicted labels	✓	✓	✓
Baseline + G-KDE labels	✓	✓	✓

3.2.4 Machine learning models

The dataset was initially divided into stratified 60%-40% splits for training and testing, with stratification based on the responses to treatment at the first and second follow-ups (40 splits for Lumiere dataset, 230 splits for the synthetic dataset).

The imbalanced data issue was addressed via inverse frequency weighting-class strategy for prediction of the first follow-up response, when the dataset is more balanced (32 SD, 15 PD). SMOTE was used for the prediction of the second follow-up response (11 SD and 36 PD when using baseline radiomics, 6 SD and 33 PD when using radiomics from 1st follow-up imaging). For the synthetic dataset, inverse frequency weighting-class strategy is used for both predictions, for first and second follow-up response. Features were normalized between 0 and 1 before training.

All classifiers were implemented as Logistic Regression models with L1-norm penalty and the regularization parameter C was set to 1. Leave-one-out cross-validation was used. Baseline and longitudinal prediction models were evaluated via the following performance metrics: accuracy, recall, specificity, balanced accuracy, and ROC-AUC.

To check if that the model f_1^i could generate reliable probability estimates for the response at the first follow-up, we evaluated the calibration of this model. A well-calibrated classifier should produce probability estimates that can be directly interpreted as confidence levels, meaning that if the model predicts a probability of 0.8, about 80% of those cases should be positive. In this study, model calibration is useful to consider, as the construction of the proposed longitudinal model, $f_2^{L,i}$, relies on the probabilities associated with the predicted response at the first follow-up, generated by the model f_1^i . Log loss [18] and Brier score [19] are two widely-used metrics for evaluating the calibration of probabilistic models. Log loss, also known as cross-entropy loss,

is defined as

$$\text{LogLoss} = -\frac{1}{N} \sum_{i=1}^N (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)),$$

where y_i is the true label (0 or 1) and p_i is the predicted probability for sample i . Log loss penalizes confident but wrong predictions very heavily, and rewards confident correct predictions, making it particularly sensitive to miscalibrations in high-confidence regions. The Brier score is defined as

$$\text{Brier score} = \frac{1}{N} \sum_{i=1}^N (p_i - y_i)^2,$$

and measures the mean squared difference between predicted probabilities and the actual outcomes. It treats overconfident and underconfident errors more symmetrically than log loss, and is generally less sensitive to extreme mistakes. The calibration of the baseline model f_1^i was assessed using the Brier score and the Log Loss, both on the original model and after isotonic calibration, for comparison purposes. Calibrating a classifier consists of fitting a regressor (calibrator) that maps the output of the classifier to a calibrated probability. The calibrator tries to predict the conditional event probability $\mathbb{P}(y_i = 1|p_i)$. We employed the isotonic regression [128], a non-parametric method, to calibrate our model f_1^i . In isotonic calibration, we seek a function g that maps the uncalibrated predicted probabilities p_i to calibrated probabilities $\hat{p}_i$, such that $\hat{p}_i = g(p_i)$, where g is a monotonic non-decreasing mapping. The function g is found by minimizing the squared error between adjusted probabilities and actual outcomes:

$$\sum_{i=1}^N (g(p_i) - y_i)^2.$$

Isotonic regression captures complex, non-smooth calibration patterns, and is suited for correcting arbitrary calibration errors. Instead, for smooth and global miscalibration, Platt's scaling calibration [140] would be more appropriate.

The models were developed in Python (v3.12.2) on a personal computer equipped with an Apple M1 Pro chip and 16 GB of unified memory.

3.3 Results

The eligible subjects of Lumiere dataset present variable follow-up intervals between subsequent post-operative timepoints (mean interval between the 1st and 2nd post-operative follow-up 108.7 days, std 66.4 days, min 21 days, max 343 days). At the first follow-up, there are 32 SD and 15 PD responses, while at the second follow-up 11 SD and 36 PD cases are registered. Specifically, 21 patients transition from SD to PD, 11 remain SD, 0 transition from PD to SD, and 15 PD cases remain PD.

The results presented on Lumiere dataset are obtained from pre-contrast T1-weighted radiomics, as models based on FLAIR, T2, and post-contrast T1 yield random baseline predictions at the first follow-up, therefore are not suitable for meaningful analysis or longitudinal assessments.

Table 3.2 presents the Brier scores and Log Loss for the baseline models to predict the response at the first follow-up, f_1^i , trained on synthetic and Lumiere datasets. For comparison, the same metrics are provided for the corresponding models after isotonic calibration. Additionally, Figure 3.5 illustrates examples of calibration curves of the original baseline models f_1^i (not calibrated), obtained with random data splits.

Table 3.2: Calibration scores for baseline models to predict the response at the first follow-up, before and after isotonic calibration, in the case of synthetic and Lumiere dataset. Mean and standard deviations are computed across all splits.

Model	Brier Score	Log Loss
Baseline synthetic	0.05 ± 0.01	0.18 ± 0.02
Baseline synthetic + isotonic calibration	0.03 ± 0.02	0.41 ± 0.48
Baseline radiomics	0.22 ± 0.02	0.64 ± 0.05
Baseline radiomics + isotonic calibration	0.22 ± 0.06	4.13 ± 2.70

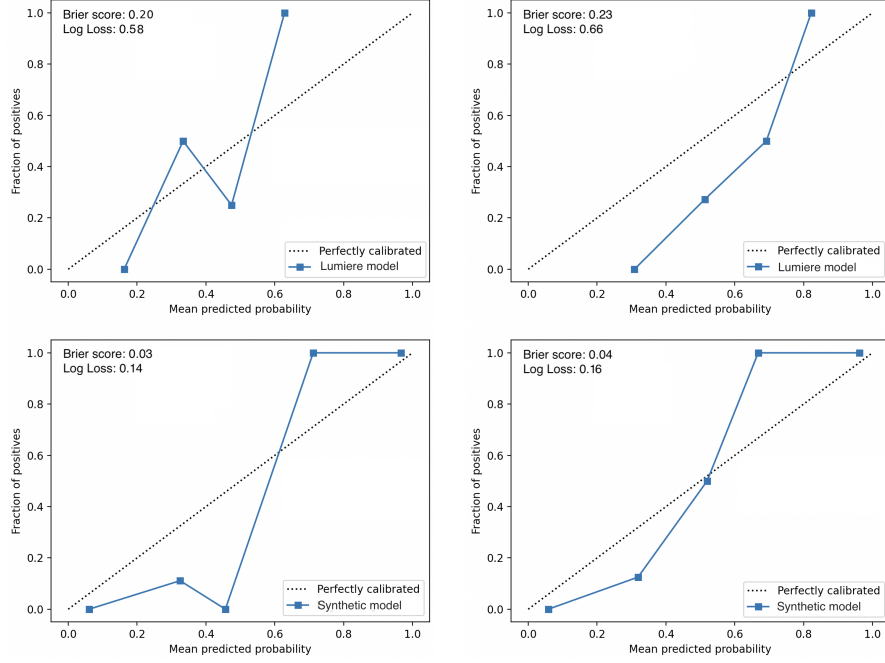


Figure 3.5: Calibration curves for the baseline model to predict the response at first follow-up. The curves are generated by grouping predicted probabilities into bins and then plotting the average predicted probability for each bin against the corresponding observed frequency (fraction of positive cases). Top row: calibration curves of two models trained on Lumiere dataset. Bottom row: calibration curves of two models trained on the synthetic dataset.

Figure 3.6 (see last page of the chapter) depicts the probability distribution of disease progression (class 1) at the first follow-up for various patients, provided by the baseline models f_1^i , in both the synthetic and Lumiere dataset case. It also shows the fit provided by G-KDE, $\hat{Y}_1^i$, along with data samples drawn from $\hat{Y}_1^i$.

Tables 3.3 and 3.4 report the mean skill-scores and their 95% confidence interval obtained by testing each ML model on the synthetic and Lumiere dataset, respectively. A summary of the implemented models is in Table 3.1. SMOTE augmentation did not have a noticeable impact on the predictive performance with the synthetic dataset and for predicting response at first follow-up using Lumiere dataset. In contrast, for the prediction of the second follow-up response with the Lumiere dataset, inverse frequency class weighting

3.3 Results

alone was insufficient to address class imbalance, resulting in performance close to random. The application of SMOTE in this case led to the results reported in Table 3.4.

Table 3.3: Performance of the implemented models on the synthetic dataset. Mean skill-scores and their 95% confidence interval (within brackets) are reported. Longitudinal models are differentiated by the modeling of 'response at 1st follow-up' feature: 'baseline + synthetic labels' for the true synthetic responses $y_{1,true}^{i,j}$, 'baseline + predicted labels' for predicted responses $y_1^{i,j}$, and 'baseline + G-KDE labels' for sampled responses $\hat{y}_1^j$. The performance of proposed longitudinal approach ('baseline + G-KDE labels') is reported in bold.

Model	Accuracy	Balanced accuracy	Recall	Specificity	ROC-AUC
Prediction of 1st follow-up response					
Baseline synthetic	0.947 (0.944, 0.950)	0.950 (0.947, 0.953)	0.967 (0.963, 0.971)	0.933 (0.928, 0.937)	0.994 (0.993, 0.994)
Prediction of 2nd follow-up response					
<i>Benchmark models</i>					
Baseline synthetic	0.603 (0.598, 0.608)	0.620 (0.614, 0.626)	0.540 (0.533, 0.548)	0.699 (0.686, 0.712)	0.707 (0.703, 0.711)
Synthetic labels at 1st follow-up	0.628 (0.627, 0.628)	0.655 (0.654, 0.655)	0.527 (0.527, 0.528)	0.782 (0.781, 0.783)	0.655 (0.654, 0.655)
<i>Proposed longitudinal models</i>					
Baseline + synthetic labels	0.731 (0.728, 0.734)	0.739 (0.736, 0.741)	0.704 (0.699, 0.709)	0.773 (0.771, 0.775)	0.801 (0.798, 0.803)
Baseline + predicted labels	0.697 (0.694, 0.700)	0.696 (0.693, 0.700)	0.699 (0.694, 0.704)	0.693 (0.689, 0.698)	0.744 (0.740, 0.748)
Baseline + G-KDE labels	0.696 (0.693, 0.699)	0.696 (0.692, 0.699)	0.697 (0.692, 0.702)	0.694 (0.690, 0.699)	0.743 (0.739, 0.747)

Table 3.4: Performance of the implemented models on Lumiere dataset. Mean skill-scores and their 95% confidence interval (within brackets) are reported. Longitudinal models are differentiated by the modeling of 'response at 1st follow-up' feature: 'baseline + true labels' for actual treatment responses $y_{1,true}^{i,j}$, 'baseline + predicted labels' for the predicted responses $y_1^{i,j}$, and 'baseline + G-KDE labels' for sampled responses $\hat{y}_1^j$. The performance of proposed longitudinal approach ('baseline + G-KDE labels') is reported in bold.

Model	Accuracy	Balanced accuracy	Recall	Specificity	ROC-AUC
Prediction of 1st follow-up response					
Baseline radiomics	0.721 (0.693, 0.749)	0.682 (0.660, 0.704)	0.575 (0.524, 0.626)	0.788 (0.736, 0.841)	0.721 (0.697, 0.746)
Prediction of 2nd follow-up response					
<i>Benchmark models</i>					
Baseline radiomics	0.680 (0.653, 0.708)	0.561 (0.524, 0.599)	0.767 (0.735, 0.798)	0.356 (0.282, 0.430)	0.591 (0.547, 0.634)
True labels at 1st follow-up	0.662 (0.637, 0.687)	0.620 (0.606, 0.634)	0.676 (0.639, 0.713)	0.564 (0.499, 0.629)	0.639 (0.625, 0.652)
Radiomics at 1st follow-up	0.742 (0.731, 0.754)	0.493 (0.478, 0.508)	0.825 (0.811, 0.839)	0.161 (0.130, 0.191)	0.451 (0.425, 0.478)
Delta radiomics	0.755 (0.742, 0.768)	0.526 (0.506, 0.546)	0.832 (0.818, 0.846)	0.219 (0.181, 0.257)	0.601 (0.572, 0.631)
<i>Proposed longitudinal models</i>					
Baseline + true labels	0.707 (0.677, 0.736)	0.626 (0.587, 0.666)	0.765 (0.727, 0.803)	0.487 (0.405, 0.570)	0.688 (0.646, 0.729)
Baseline + predicted labels	0.678 (0.650, 0.705)	0.590 (0.555, 0.624)	0.742 (0.708, 0.776)	0.438 (0.370, 0.505)	0.635 (0.598, 0.672)
Baseline + G-KDE labels	0.689 (0.663, 0.715)	0.609 (0.575, 0.643)	0.747 (0.715, 0.780)	0.470 (0.401, 0.540)	0.639 (0.599, 0.679)

3.4 Discussion

In this work we introduced a longitudinal method for predicting GBM response to treatment at the second (post-operative) follow-up, using baseline (pre-operative) radiomics and the response to treatment at the first follow-up, the last one modeled in a probabilistic manner. The probability distribution of having progressive disease at the first follow-up was estimated based on the output probabilities produced by a model designed to predict the response at the first follow-up using baseline features. Sample responses were drawn from this distribution and used for testing the longitudinal model at the second follow-up. This approach leverages the estimated probability distribution of the response at the first follow-up to handle the intrinsic uncertainty of the longitudinal prediction.

This methodology facilitates the prediction of longitudinal responses for newly enrolled patients with only baseline radiomic data and no successive follow-ups. Indeed, by utilizing an ensemble of pre-trained models, the probability distribution of disease progression at the first follow-up could be inferred, thereby enabling a longitudinal assessment at the second follow-up for unlabeled patients, through the pre-trained longitudinal model.

The estimation of the probability distribution for progressive disease relies on the assumption that the probabilities generated by the baseline model to predict the response at the first follow-up can be directly interpreted as event likelihoods, i.e., a predicted value of 0.7 truly reflects a 70% probability of the event occurring. For such reason we evaluated model calibration via Brier score and Log Loss. Table 3.2 shows that both scores resulted lower for models trained on the synthetic dataset than for Lumiere-based models, indicating a better overall calibration in the synthetic setting. This could happen because the Lumiere training size is too small, overfitting case, or because the model is misspecified, e.g. if the true decision boundary of the dataset is a highly non-linear function of the input features. Although applying isotonic calibra-

tion to the models slightly lowered the mean Brier score for both dataset, the Log Loss resulted to increase significantly. This deteriorates model predictive effectiveness, as the logistic regression classifiers are trained by minimizing the Log Loss, undermining the purpose of calibration itself, and is likely due to data scarcity, as the calibration dataset (in this case, the test set) does not properly match the real distribution of the data. Figure 3.5 shows examples of calibration curves for the baseline model, to predict the response at first follow-up, having under-confident trend, for both synthetic and Lumiere dataset.

Baseline models f_1^i for the prediction of the response at the first follow-up exhibit the highest overall performance across the trained prediction models (see first row of Tables 3.3 and 3.4). The highest performance is achieved on the synthetic dataset, where the models produce more extreme probability outputs (Figure 3.6, first column), while baseline radiomics maintains good predictive power at the first follow-up for the Lumiere dataset, although with larger uncertainty (Figure 3.6, second column). We remark that the probabilities outputted by these models are then used to build the distribution $\hat{Y}_1^j$, described in Section 3.2.1, that is the core of our probabilistic method for longitudinal prediction of treatment response. Of note, the performance of the baseline models for the prediction at the first follow-up should not be directly compared to the prediction at the second follow-up. Instead, it can serve as a reference for comparing the predictive power of pure baseline radiomics in forecasting treatment response at the first and second follow-ups. Examining the benchmark models based on baseline radiomics reveals that baseline features are less predictive of treatment response at the second follow-up compared to the first (mean ROC-AUC on the synthetic dataset: 0.707 vs. 0.994; mean ROC-AUC on the Lumiere dataset: 0.591 vs. 0.721).

In the synthetic setting, the proposed longitudinal model 'baseline + G-KDE labels' performs similar to the 'baseline + predicted labels' model, using the predicted labels at the first follow-up, but slightly worse than the one using the true synthetic labels, 'baseline + synthetic labels' model. This is expected, since

the true synthetic labels do not account for uncertainty in the response at the first follow-up. G-KDE also performs better than the model trained with the 1st follow-up labels alone, 'synthetic labels at 1st follow-up' model, indicating that incorporating baseline radiomics in the model provides valuable information to the longitudinal outcome (0.655 vs. 0.696 mean balanced accuracy, 0.655 vs. 0.743 mean ROC-AUC). Additionally, G-KDE outperforms the 'baseline synthetic' model, which uses pure baseline synthetic features to forecast the response at the second follow-up (0.620 vs. 0.696 mean balanced accuracy, 0.707 vs. 0.743 mean ROC-AUC).

As for Lumiere dataset, 'baseline + G-KDE labels' performance results lower compared to the longitudinal model based on true labels, 'baseline + true labels', but demonstrates superior performance compared to 'baseline + predicted labels' model in all the skill-scores, indicating an improved capability in handling predictive uncertainty through sampling from the probability distribution of the treatment response at the first follow-up, $\hat{Y}_1^j$ (see last three rows of Table 3.4). The benchmark model based on delta radiomics is more effective in predicting disease progression, with a mean recall of 0.832 compared to 0.747 for our G-KDE-based method. However, G-KDE shows a substantial improvement in mean specificity, 0.470 versus 0.219 for delta radiomics, showing better ability to predict the stable disease cases, which are severely under-represented at the second follow-up. This results in the enhancement of the balanced accuracy obtained with G-KDE model (0.609 vs. 0.526) and in the increase of the ROC-AUC (0.639 vs. 0.601), see Table 3.4. G-KDE also enhances balanced accuracy and ROC-AUC compared to 'baseline radiomics' and 'radiomics at 1st follow-up' models, respectively based on baseline and 1st follow-up radiomics only. Finally, G-KDE demonstrates comparable performance to the model trained with the true response labels at the first follow-up, 'true labels at 1st follow-up' model.

Our longitudinal study fits within the context of GBM response prediction. The study [138] provides a reference for radiomics-based prediction of GBM

tumor progression after chemoradiotherapy from the baseline timepoint. This work included 76 patients, radiomics from contrast-enhanced T1, T2-weighted, and ADC MRI, clinical characteristics and MGMT promoter methylation status. Outcomes were assessed using baseline radiomics up to six months post-chemoradiotherapy. Key predictive features included age and MGMT promoter methylation status, a known discriminator of overall survival [176]. The model including MGMT achieved a ROC-AUC of 0.80 with 0.67 specificity, demonstrating the benefit of combining radiomics with clinical and molecular markers. In contrast, MGMT and age alone reached a ROC-AUC of 0.66, while radiomics-based models alone presented ROC-AUCs between 0.46 and 0.69. The performance of our probabilistic model for longitudinal prediction of treatment response aligns with this reference. Moreover, we remark that our longitudinal approach improves the prediction of the baseline model, based on pure baseline radiomics, even with reduced dataset size. This indicates room for improved performance leveraging bigger longitudinal datasets [6, 8, 118]. Indeed, a bigger dataset would allow the inclusion of informative clinical biomarkers in the predictive models, like the MGMT, the consideration of an extended radiomic feature set (wavelet and LoG features), and a multimodal evaluation. A larger longitudinal dataset would also enable more comprehensive validation of the methodology and facilitate the establishment of a benchmark for longitudinal prediction models. Ideally, a highly-accurate well-calibrated model for predicting the first follow-up response would produce reliable probability distribution of disease progression at the first follow-up, $\hat{Y}_1^j$, enabling the training of the longitudinal model with estimated response samples, thus eliminating the need for intermediate treatment evaluation. This would be useful in case of missing intermediate timepoint. With a nested approach, our methodology might also enable evaluation beyond the second follow-up response, thereby enhancing longitudinal assessment for various applications, including those related to neurodegenerative diseases.

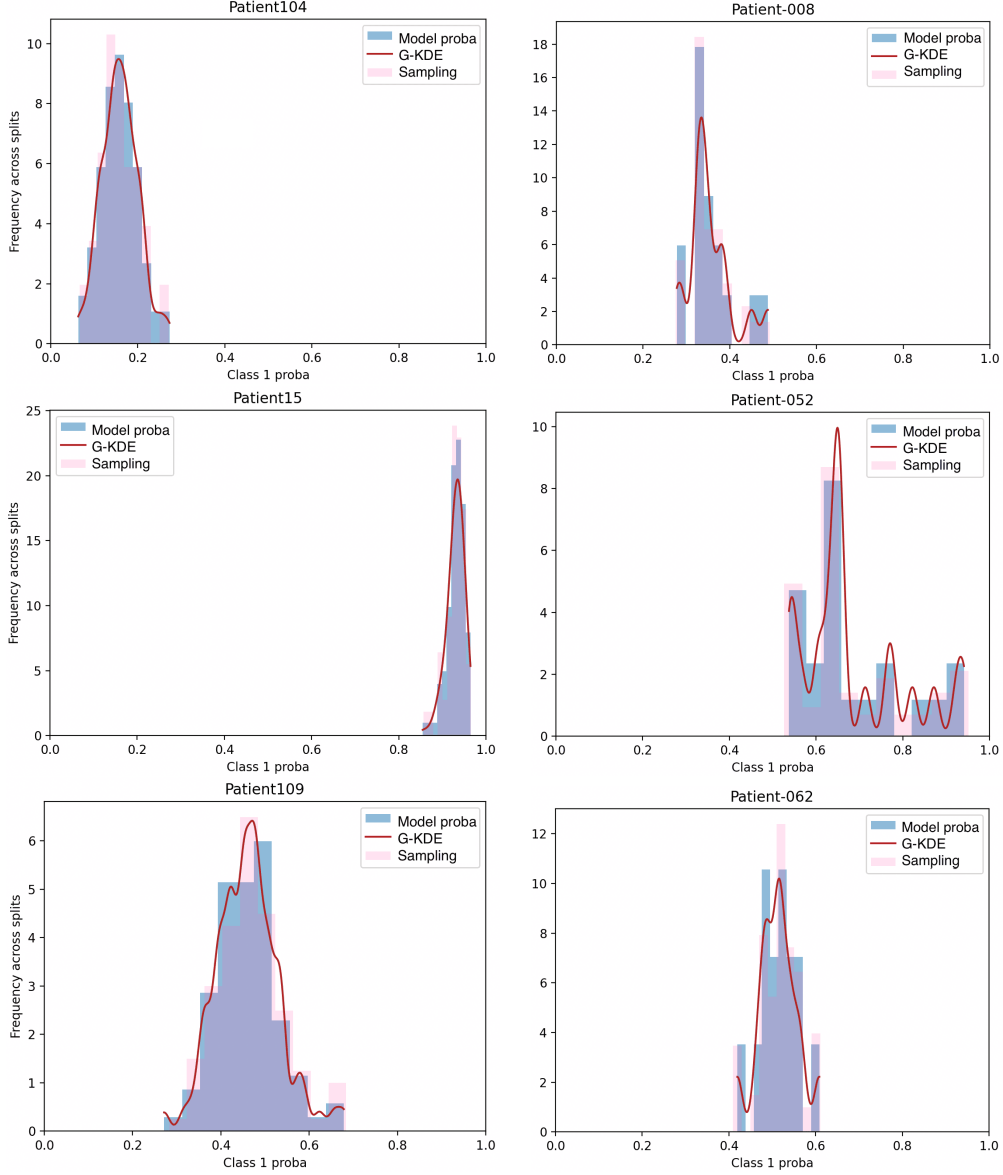


Figure 3.6: Examples of probability distribution of disease progression (class 1) at the first follow-up. Light blue: normalized histogram (area 1) of $\{\text{proba}_1^{i,j}\}_i$ obtained across splits, i , for each patient, j . Red: probability density function (PDF) of PD distribution estimated via G-KDE, $\hat{Y}_1^j$, for the j -th patient. Pink: samples drawn from the G-KDE PDF, $\{\widehat{\text{proba}}_1^j\}$. First column: examples from the synthetic dataset. Second column: examples from Lumiere dataset.

Chapter 4

Image harmonization

With the rapid growth of MRI datasets collected from various centers, the power of data-derived insights is becoming increasingly fundamental for the understanding of specific diseases, and can significantly enhance medical practice by providing robust statistical evidence. However, the reliability of radiomics remains an area of investigation. In Chapter 2, we tackled the challenge of radiomics stability with respect to segmentation variability in the context of breast cancer subtype prediction. Leveraging explainability methods (SHAP) we identified a set of predictive features, and we used radiomics-based ML models to evaluate the stability of the predictive performance with respect to segmentation variability. In addition, the relationship between image segmentation and radiomic features was examined from a numerical and reproducibility perspective.

Beyond segmentation, radiomics features are subject to an additional level of uncertainty that directly arises from imaging acquisition. In fact, the MRI images used in radiomics, and more in general in multicentric studies, are often acquired in centers that follow unharmonized acquisition protocols, involving diverse scanners and centers, leading to substantial variability in the extracted radiomic features. MRI scanners can produce images with different contrast, brightness, and spatial characteristics due to variations in hardware, software, calibration, and other factors, including discrepancies that arise

from machine maintenance, protocols, environmental conditions, and operator expertise [87]. In Chapter 2, we addressed this issue by harmonizing the extracted features using the ComBat method, which mitigates the effects of confounding variables, such as scanner manufacturer models. Despite this, feature variability alone may not fully reflect the variability introduced by differences in imaging acquisition protocols. To ensure a more generalizable and consistent radiomics pipeline, harmonizing the images themselves, rather than features extracted from unstandardized images, could be a more effective strategy to enable robust and reliable analyses [168, 178, 213].

Several methodologies have been proposed for harmonizing images and, specifically, multicenter MRI data. I2I translation family includes methods that aim to map an input image into an output image while preserving certain semantic properties. Specifically, these generative models produce translated images, that appear to be drawn from the distribution of the target domain. They are categorized according to the type of supervision, i.e., supervised (directional or bidirectional) or unsupervised (cyclic consistency, autoencoders, disentangler), or to the type of translation, i.e., one-to-one, one-to-many, many-to-many [3]. Transformers [65] are models based on self-attention mechanism with the power of capturing long-range context dependencies. Still, they have been shown to overlook important high-frequency features that may be the site for biology- and pathology-related information [194]. Style transfer is a fully unsupervised deep learning framework [107], wherein image harmonization is treated as a style transfer problem rather than a domain transfer problem.

Only few attempts have been brought out in the medical domain, among all NeuroHarmonize [143], Calamiti [212], MURD [111] and ComBat [145, 184]. At the crossroad between style transfer and I2I translation families, StarGANv2 [36, 108] stands out as a single framework that addresses both the diversity of generated images and scalability across multiple domains, so far demonstrating significantly improved results compared to baseline methods. While showing promising results in the harmonization task, this approach necessitates extensive pre-processing, including skull-stripping of the

T1 images [164], nonuniformity correction [170], and registration to a standard space, making the end-to-end process heavy.

This chapter presents DISARM (disentangled scanner-free image generation via unsupervised Image2Image translation) [22], a model we designed to harmonize 3D brain MR images by disentangling anatomical structure and scanner-specific information, requiring reduced imaging pre-processing. The model is designed to produce *scanner-free* images while preserving the original anatomical structure and biologically informative data as to achieve robust generalizability across various scanners. The aim of the presented method is to overcome imaging inter-scanner variability as a preliminary step to any radiomics- or DL-based prediction analysis.

The chapter is organized as follows. Section 4.1 reports details on DISARM model, including the architecture and the losses, pre-processing, benchmarking model, and the statistical evaluation procedure, and describes the datasets employed for the study. Section 4.2 shows the results of scanner information transfer, scanner-free harmonization, along with an evaluation of the predictive performance of a regression model designed to assess the effectiveness of harmonization in enhancing biological information. The discussion is in Section 4.3.

4.1 Materials and methods

The model here proposed, DISARM, is a harmonization model for 3D medical images extending upon DRIT++ network architecture [99], with specific application to T1-weighted brain MRI data. Particularly, this model targets the scanner-free generation of 3D MR images to clean out clinical images from batch effects. The final objective is to ensure robust generalizability of the harmonization method across various scanners, including those not encountered during the training phase. Compared to DRIT++, DISARM is specifically designed to process 3D medical images, whereas DRIT++ operates on 2D natural images. While DRIT++ focuses on multi-domain Image2Image

translation to generate diverse outputs, DISARM aims to harmonize images across different scanner domains, producing consistent and comparable results. Other than transferring images to other scanner domains, DISARM disentangles image content from scanner-specific information, enabling the generation of a scanner-free image. This is obtained by representing the scanner-free domain with a zero vector, rather than a conventional one-hot domain code, and by sampling the scanner-free attributes from a standardized normal distribution (see Section 4.1.1). Additional differences between DISARM and DRIT++ are reflected in the formulation of the total loss, as detailed at the end of Section 4.1.2.

Let $\mathcal{X} = \bigcup_{k=1}^N \mathcal{X}_k \in \mathbb{R}^{1 \times H \times W \times D}$ be a set of MR images and denote with $\mathcal{X}_k$ the set of images acquired from the k -th scanner domain among N distinct domains. We assume that the imaging information can be disentangled into two latent spaces $(\mathcal{B}, \mathcal{S})$, being $\mathcal{B}$ the space encoding the information related to the brain anatomical structure and being $\mathcal{S}$ the scanner space, specifically targeting the scanner effects. An image x_k drawn from $\mathcal{X}_k$ can be thus obtained as the combination of information from $\mathcal{B}$ and $\mathcal{S}$. The scanner-free harmonization task consists of noising out the scanner-specific information by combining the anatomical structure of the image with a random Gaussian noise $\mathcal{N}(0, 1)$, instead of the scanner distribution in $\mathcal{S}$.

4.1.1 Architecture

The backbone of the architecture includes two encoders and a generator. The brain structure encoder $E_b : \mathcal{X} \rightarrow \mathcal{B}$ maps an input image $x_k \in \mathcal{X}_k$ into a lower-dimensional space $\mathcal{B}$ encoding the information related to the brain structure $z_{x_k}^b$, see Figure 4.1, top left panel, while the scanner encoder $E_s : (\mathcal{X}, \mathcal{C}) \rightarrow \mathcal{S}$ takes an input image $x_k \in \mathcal{X}_k$ and its associated scanner label $c_k \in \mathcal{C}$ and operates as a variational autoencoder, producing a parametric distribution of the scanner effect, as represented in Figure 4.1, top right panel. The latent

vector of the k -th scanner effect $\mathbf{z}_k^s$ is thus sampled as:

$$\mathbf{z}_k^s = \sigma_k \cdot \varepsilon + \mu_k, \quad \mathbf{z}_k^s \in \mathcal{S}_k, \quad \varepsilon \sim \mathcal{N}(0, 1).$$

The generator $G : (\mathcal{B}, \mathcal{S}, \mathcal{C}) \rightarrow \mathcal{X}$ produces an image $x_k \in \mathcal{X}_k$ that preserves a specified brain structure within the space $\mathcal{B}$ while incorporating a scanner attribute from the space $\mathcal{S}$ associated to its label within the space $\mathcal{C}$ (Figure 4.1, bottom panel). To transfer the scanner attribute of scanner $\mathcal{X}_h$ to another image $x_k \in \mathcal{X}_k$, the scanner attribute $\mathbf{z}_h^s$ is plugged into the generator together with its label $c_h \in \mathcal{C}_h$ and the brain attribute $\mathbf{z}_{x_k}^b$, obtaining the image x_k as it was acquired by the scanner h , $v_h = G(\mathbf{z}_{x_k}^b, \mathbf{z}_h^s, c_h)$. This passage is represented in Figure 4.2, top panel. Upon this strategy, we can achieve a scanner-free harmonization of images by sampling the latent vector of the scanner effect from $\varepsilon \sim \mathcal{N}(0, 1)$ and plugging it into G together with x_k , as shown in Figure 4.2, bottom panel. The scanner-free space is denoted as $\mathcal{F}$ and the encoded scanner label is defined as a vector of zeros.

4.1.2 Losses

The training process is shown in Figure 4.3 and is carried out on pairs of image (x_k, c_k) and (x_h, c_h) from randomly selected scanner domains (K and H) among a pool of N domains. Each image is simultaneously fed into its own E_b and E_s , yielding the corresponding latent representations $\mathbf{z}_l^s = E_s(x_l, c_l)$ and $\mathbf{z}_{x_l}^b = E_b(x_l)$, for $l = \{h, k\}$. The training follows the minimization of a set of losses governing the different subnetworks in the architecture.

The Brain Structure Adversarial Loss L_{adv}^b aims to ensure that the brain encoders E_b generate scanner-independent brain embeddings, that is, map images into a shared space $\mathcal{B}$ where no domain membership is distinguishable. To do this, the latent representations $\mathbf{z}_{x_k}^b$ and $\mathbf{z}_{x_h}^b$ are fed into a discriminator D_b which tries to identify their domain membership as follows:

$$L_{\text{adv}}^b = \frac{1}{2} \mathbb{E}_{x_k} \left[\log \left[D_b(\mathbf{z}_{x_k}^b) (1 - D_b(\mathbf{z}_{x_k}^b)) \right] \right] + \frac{1}{2} \mathbb{E}_{x_h} \left[\log \left[D_b(\mathbf{z}_{x_h}^b) (1 - D_b(\mathbf{z}_{x_h}^b)) \right] \right].$$

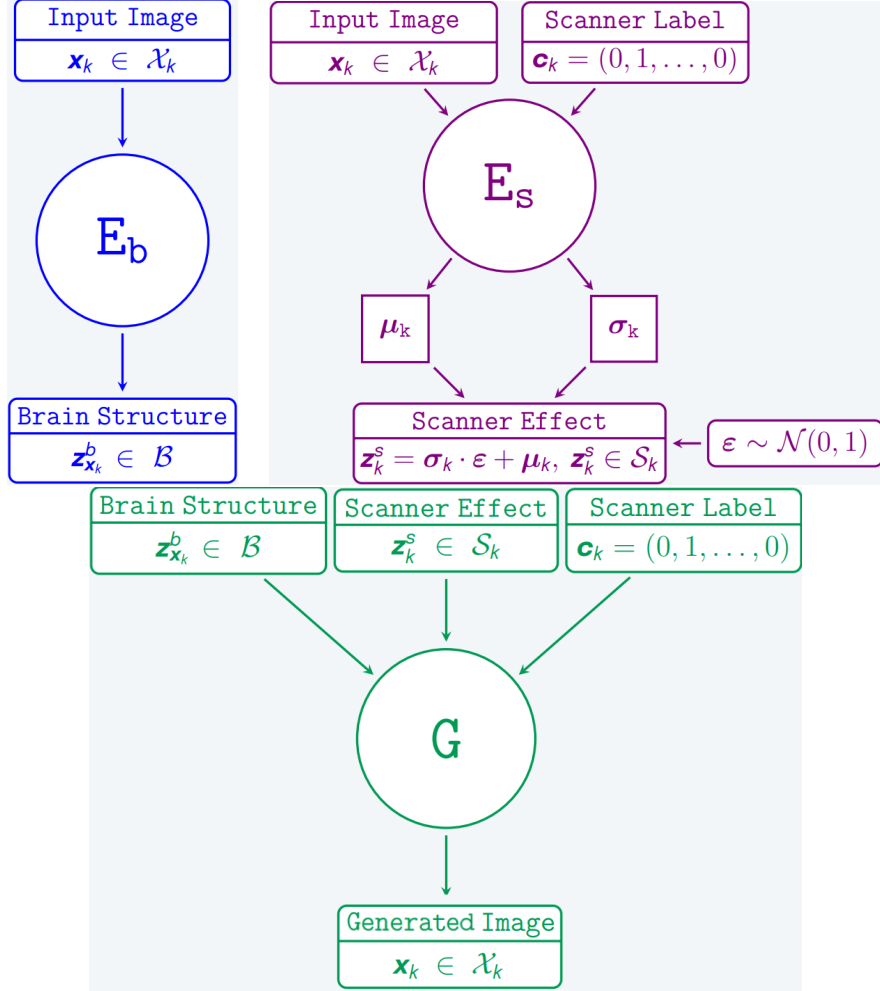


Figure 4.1: Top row: brain structure encoder E_b (left), scanner encoder E_s (right). Bottom row: generator G .

The Scanner Adversarial Loss L_{adv}^s aims to force the generators G to produce realistic images. To this purpose, the generated images $\mathbf{u}_k$ and $\mathbf{v}_h$ are fed into a discriminator D_s which tries to distinguish real images from generated images:

$$L_{\text{adv}}^s = \frac{1}{2} \mathbb{E}_{x_k} \left[\log \left[D_s(x_k) (1 - D_s(u_k)) \right] \right] + \frac{1}{2} \mathbb{E}_{x_h} \left[\log \left[D_s(x_h) (1 - D_s(v_h)) \right] \right].$$

The Scanner Classification Loss L_{cls}^s is defined as follows. The discriminator D_c is trained to predict the scanner label of the generated images, to ensure

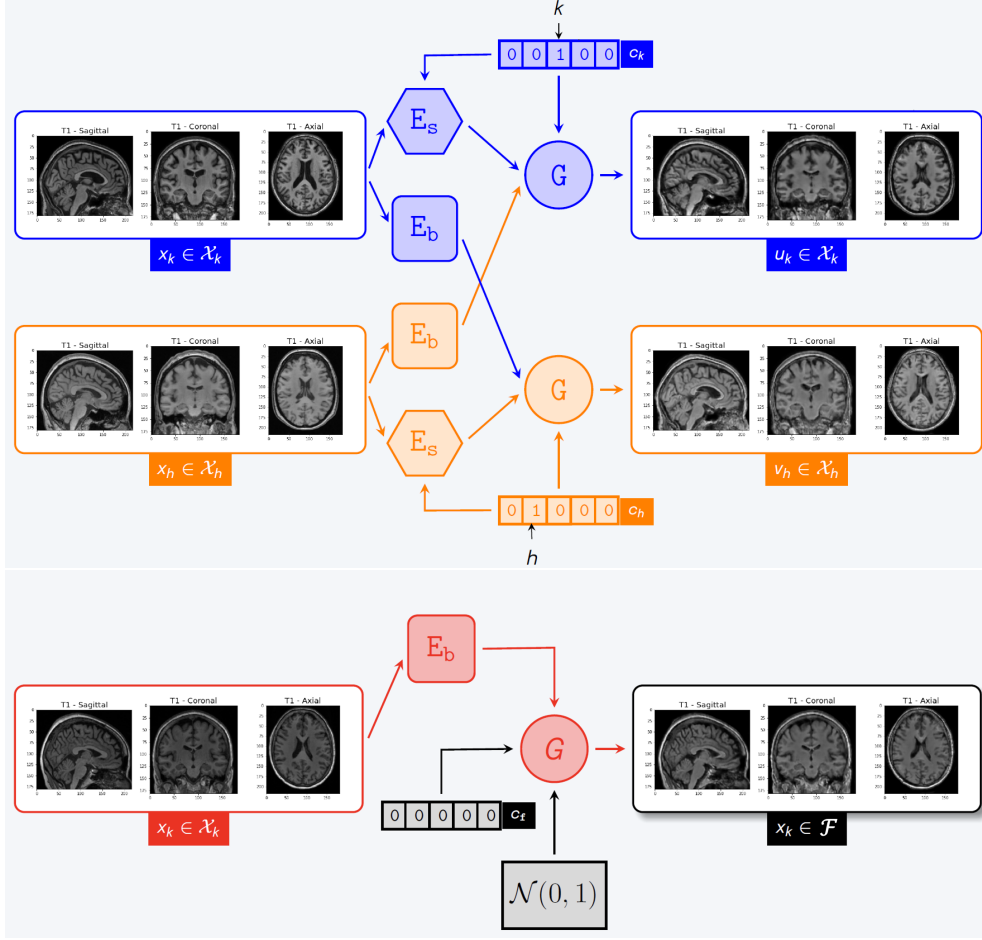


Figure 4.2: Top panel: transfer of an image into a target scanner space. Bottom panel: generation of an image within the scanner-free space $\mathcal{F}$.

that its scanner-related information is compliant with the desired one (i.e., the one plugged in):

$$L_{\text{cls}}^s = \mathbb{E}_{x_k} \left[-\log [D_c(c_k | u_k)] \right] + \mathbb{E}_{x_h} \left[-\log [D_c(c_h | v_h)] \right].$$

The Cycle Consistency Loss L_{cc} is introduced to maintain cross-cycle consistency. u_k and v_h undergo a backward translation process where they are re-generated in their own original scanner domain. The reconstructed images $\hat{x}_k = G(E_b(v_h), E_s(u_k, c_k), c_k)$ and $\hat{x}_h = G(E_b(u_k), E_s(v_h, c_h), c_h)$ need to be as

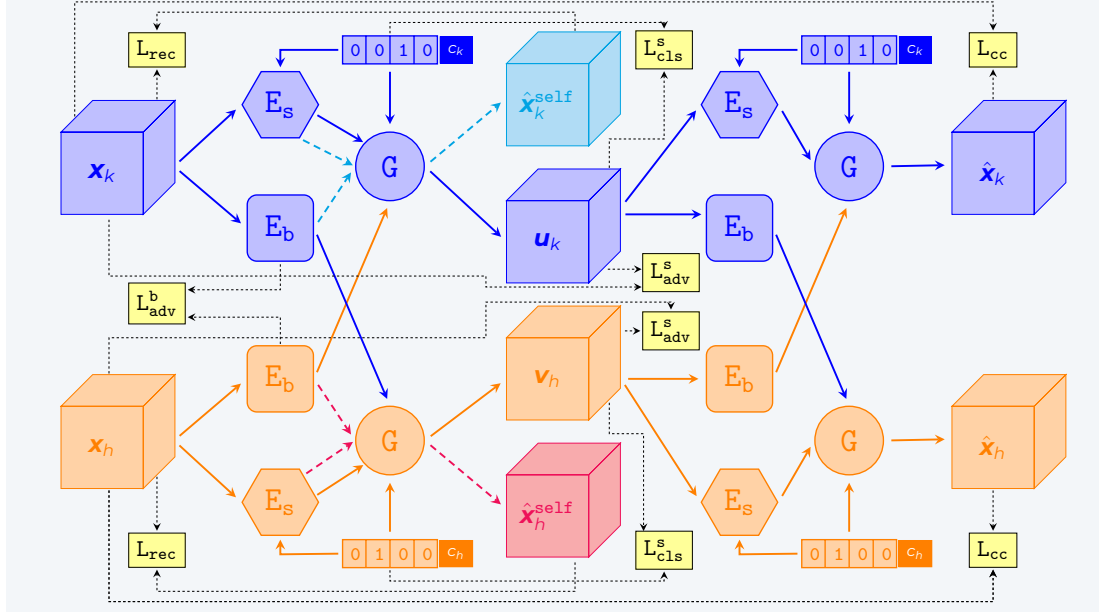


Figure 4.3: High level functional diagram of the training procedure for DISARM network. Discriminators, L_{lat} , and L_{KL} have not been displayed due to space reasons.

maximally similar to the original x_k and x_h in terms of:

$$L_{cc} = \mathbb{E}_{x_k, x_h} \left[\|\hat{x}_k - x_k\| + \|\hat{x}_h - x_h\| \right].$$

The Self Reconstruction Loss L_{rec} ensures that the reconstructed image $\hat{x}_l^{\text{self}} = G(E_b(x_l), E_s(x_l), c_l)$ closely matches the original image x_l for $l = h, k$. To do this, x_k and x_h are passed through the encoders E_s and E_b , and the generator G , and are reconstructed as x_k^{self} and x_h^{self} . The loss thus computes:

$$L_{\text{rec}} = \mathbb{E}_{x_k, x_h} \left[\|\hat{x}_k^{\text{self}} - x_k\| + \|\hat{x}_h^{\text{self}} - x_h\| \right].$$

To encourage the generator G and the encoder E_s to preserve the latent space after generating and encoding an image (invertible mapping between

the image and the latent space), a Latent Regression Loss L_{lat} is used:

$$L_{\text{lat}} = \mathbb{E}_{x_k, x_h} \left[\left\| z_k^s - E_s(G(z_{x_h}^b, z_k^s, c_k), c_k) \right\| + \left\| z_h^s - E_s(G(z_{x_k}^b, z_h^s, c_h), c_h) \right\| \right].$$

Finally, the Kullback–Leibler Loss L_{KL} aims to align the scanner effect representations with a normal Gaussian distribution.

The overall model loss is defined as follows:

$$L_{\text{tot}} = \lambda_{\text{cc}} L_{\text{cc}} + \lambda_{\text{rec}} L_{\text{rec}} + \lambda_{\text{lat}} L_{\text{lat}} + \lambda_{\text{KL}} L_{\text{KL}} - \lambda_{\text{adv}}^b L_{\text{adv}}^b - \lambda_{\text{cls}}^s L_{\text{cls}}^s - \lambda_{\text{adv}}^s L_{\text{adv}}^s.$$

Compared to DRIT++, we removed the Mode Seeking Loss because the objectives of our framework differ fundamentally. In DRIT++, the mode seeking loss is introduced to explicitly promote diversity among generated outputs by maximizing the ratio between the distance of generated images and the distance between corresponding random latent codes. This mechanism ensures that small perturbations in the latent space lead to significant variations in the generated outputs, which is desirable for tasks requiring extensive multimodal synthesis. In our case, however, the primary objective is not to maximize output diversity but to enable consistent sampling of scanner-related features and still produce consistent and comparable results, within each scanner space. Moreover, with the main objective of building a scanner-free representation of the images by sampling attributes from a standard normal distribution $\varepsilon \sim \mathcal{N}(0, 1)$, encouraging excessive diversity through the Mode Seeking Loss would be counterproductive, as it would introduce unnecessary variability when we actually want image representations to be stable and domain-invariant in the scanner-free space.

For training, we employed the Adam optimizer [89] with a learning rate of 0.0001 and batch size of 1. The parameters used are reported in Table 4.1. The DRIT++ hyperparameters were taken as a reference and adjusted to fit our context. The key modification was adjusting the size of the brain encoder E_b . While the content encoder in DRIT++ reduces image dimensions to one-quarter, a strategy that works well for 2D natural images, preserving fine details in

Table 4.1: Parameters used for the training of DISARM.

λ_{cc}	λ_{rec}	λ_{cls}^s	λ_{adv}^s	λ_{lat}	λ_{KL}	λ_{adv}^b	$\dim(\mathcal{B})$	$\dim(\mathcal{S})$	Iterations
10	10	3	10	8	0.01	1	(91, 109, 91)	8	90000

the brain structure latent space is more critical for MR images. Therefore, the brain encoder is set to reduce the original image size (182, 218, 182) by a half, resulting in $\dim(\mathcal{B}) = (91, 109, 91)$. The dimension of the scanner latent space is set to 8, as in DRIT++ implementation. The other parameters were tuned by balancing the model’s ability to transfer scanner-specific effects versus the ability to reconstruct the anatomical structure. With certain parameter settings, the model produced a good anatomical reconstruction but failed to modify the scanner effect. Therefore, we adjusted them to encourage changes related to the scanner effect without compromising the anatomical structure. Compared to DRIT++ implementation, we tuned the following hyperparameters: λ_{cls}^s was increased from 1 to 3, to strengthen the emphasis on scanner-related information and ensure that the generated scanner domain matches the desired one; λ_{adv}^s was increased from 1 to 10, to enhance the scanner adversarial loss, encouraging the generator to produce more realistic images within each specific scanner domain; λ_{lat} was reduced from 10 to 8, to preserve the scanner latent space representation in a generated image.

DISARM implementation is available at https://github.com/luca2245/DISARM_Harmonization/tree/main. The model training was performed in Python (v3.12.2) on a high-performance computing system equipped with 4 GPUs (40 GB each, CUDA 11), 128 CPU cores, and 512 GB of RAM, enabling efficient large-scale training and parallel processing.

4.1.3 Datasets

DISARM methodology was validated using T1-weighted MR images from healthy controls, with a slice thickness of 1 mm, obtained on a 3T scanner. In particular, DISARM was trained on a dataset comprising 335 images from Alzheimer’s Disease Neuroimaging dataset (ADNI3 study) [39], aged 55-89

years with mean 73.3 ± 8.5 years, and from Parkinson’s Progression Marker Initiative dataset (PPMI) [121], aged 60-80 years with mean age 68.4 ± 5.4 years, satisfying the inclusion criteria cited above. The training images came from $N = 5$ different scanners (163 from Prisma Fit, 64 from Prisma, 30 from Achieva dStream, 38 from Skyra, and 40 from TrioTim). We tested the results on a clinical dataset from the Italian Neuroimaging Network (RIN), which we will refer to as the RIN dataset. This dataset comprises 117 images acquired with 7 scanners (17 from Achieva dStream, 20 from DISCOVERY MR750, 23 from Ingenia CX, 40 from Prisma, 2 from Signa HDxt, 12 from Skyra, 3 from Skyra Fit). The RIN cohort age range is 49 to 80 years, with average age of 64.5 ± 8.6 years.

4.1.4 Pre-processing

All the images were 1) normalized to the standard orientation to facilitate further steps, 2) applied a bias-field correction to mitigate the magnetic field variations in the MRI scanners [77, 172], and 3) registered to the reference Standard MNI152-T1-1mm [78], widely used in neuroimaging softwares like FSL [172]. After preprocessing, the output images had dimensions $182 \times 218 \times 182$. Moreover, data augmentation was performed through dense random elastic deformation [139].

4.1.5 Benchmarking

StarGANv2 was selected as a benchmark because it represents the state of the art in multi-domain image translation and was pre-trained specifically for the same clinical objective of T1-weighted brain MR image harmonization [108], making it directly comparable within our application context. DISARM and StarGANv2 are both multi-domain Image2Image translation frameworks capable of performing many-to-many mappings within a single unified model, without requiring paired data. However, they differ in their architectural design and approach to style modeling. StarGANv2 considers image har-

monization as a pure style transfer problem rather than a domain transfer problem, considering every single image as a unique "domain" with its own style, and the styles can be learned using an adversarial approach. StarGANv2 proposes that the style information needed for harmonization can be encoded from a single reference MR image directly. Instead, DISARM explicitly disentangles image representations into a shared content space and domain-specific style spaces, using separate encoders to extract each component. This separation allows DISARM to learn the distribution of styles for each domain and sample diverse outputs through latent style codes that reflect domain-level variation. In contrast, StarGANv2 does not perform explicit disentanglement; instead, it conditions the generator on a style code that is either sampled from a latent space via a domain-aware mapping network or extracted from a single reference image using a style encoder. While both frameworks allow for diverse image generation, DISARM focuses on learning interpretable, disentangled style distributions per domain, whereas StarGANv2 emphasizes flexible reference-guided or latent-guided style transfer without modeling the domain's style space explicitly. Despite these architectural and methodological differences, both models support multi-domain translation, utilize adversarial losses, and enable style diversity through random sampling mechanisms. To enforce content preservation in unpaired Image2Image translation, StarGANv2 uses a standard cycle consistency loss, enforcing that the translation of an image from domain A to domain B, and then back to A should result in the original image. DISARM employs a cross-cycle consistency loss, assuming that image representations are disentangled into separate content and style components. It operates by swapping the style codes between images from different domains, translating them accordingly, and then re-swapping to reconstruct the original image. This loss ensures that the content remains invariant under style manipulation and is essential for maintaining a consistent and disentangled representation across domains. Moreover, DISARM has a scanner classification loss, to ensure that the scanner related information is compliant with the desired one. Additionally, StarGANv2 uses the Style

Diversification Loss, which is not present in DISARM for the reasons explained in Section 4.1.2.

To benchmark DISARM model, the performances of the pre-trained StarGANv2 [108] were evaluated on the RIN dataset. The required pre-processing included that the images were skull-stripped [164] and corrected for nonuniformity using the N3 method [170] in Freesurfer. Furthermore, the images needed to be linearly registered to a standard MNI symmetric template (ICBM 152 Nonlinear Symmetric atlas [58]) and resized to isotropic 1 mm³ voxels, yielding dimensions of $256 \times 256 \times 256$. According to StarGANv2 implementation, harmonization was performed on a sliding window of 3 image slices with a stride of 1. The final T1 harmonized MRI 3D volumes were then reconstructed by combining these harmonized partial volumes. Since the StarGANv2 model does not support scanner-free harmonization, we selected the Prisma Fit scanner domain as the reference space for harmonization, enabling a direct comparison with the proposed method.

4.1.6 Evaluation procedure

DISARM model underwent various evaluations to assess its performance. The first evaluation was supposed to assess whether the proposed model could accurately attribute the effects of one scanner to images captured by another scanner. For this purpose, the mean gray level distributions of the RIN images harmonized on Prisma Fit scanner space by DISARM and StarGANv2 were compared. Second, the differences in gray level distributions of the output images harmonized on the scanner-free space $\mathcal{F}$ by DISARM were tested. The mean gray level distribution of each scanner was computed by calculating the voxelwise mean across the images of the considered scanner, followed by average pooling. In both cases, K-sample Anderson-Darling test [161] was employed for comparison of the distributions. K-sample Anderson-Darling test is a non-parametric statistical test used to determine whether multiple (K) independent samples come from the same (unspecified) continuous distribution. The test assesses differences among scanners simultaneously.

To evaluate the preservation of anatomical structures after harmonization, the structural similarity index measure (SSIM) was used for the structural comparison between the original image and its harmonized version.

Lastly, the predictive performance in predicting patients' ages was evaluated, to test the power of the harmonization model to enhance the biological information. For this step, a regression model was employed, fed with the total gray volumes TGV [12, 30] extracted from the RIN MR images before and after harmonization. Since scanner differences could impact TGV computation and consequently interfere with age regression, the interactions with scanner variable was included in the analysis, evaluating the significance of the scanner information in terms of p-values of the β regression coefficients. The scanner Prisma was selected as reference category, meaning that the influence of each scanner in the relationship between TGV and outcome (age) was compared to the effect of scanner Prisma. The root mean squared error (RMSE) was computed before and after harmonization between the real and the predicted age, both with DISARM and StarGANv2 across all scanners. For this downstream analysis, scanners Signa HDxt and Skyra Fit were discarded due to low numerosity, and "before harmonization" refers to before the pre-processing pipeline.

4.2 Results

The evaluation process involved harmonizing the 117 healthy controls from the RIN dataset into both the Prisma Fit scanner space, using DISARM and StarGANv2, and the scanner-free space $\mathcal{F}$, only using DISARM. Figure 4.7 (see last page of the chapter) provides examples of output images for both models. The images presented for DISARM and StarGANv2 harmonization originate from the same patients. The differences in anatomical appearance result from the pre-processing registration on two different reference templates.

4.2.1 Evaluation of scanner information transfer

To test the harmonization task, the average grayscale distribution across all images for each scanner was computed, before and after transferring the images to the Prisma Fit scanner space, both with DISARM and StarGANv2. Figure 4.4 displays the boxplot of the average gray level distributions per scanner before and after harmonization by DISARM and StarGANv2. Of note, the harmonization of DISARM relies on the whole target scanner domain, whereas StarGANv2 relies on a single image as a reference. Therefore, the reference distribution (shown in gray) for DISARM is the average gray distribution of images acquired with Prisma Fit, whereas the StarGANv2 reference distribution is the one of a selected Prisma Fit image. The two plots in the top row of Figure 4.4 display results for DISARM, showing a massive alignment of the distributions after harmonization: K-sample Anderson-Darling test yields $p\text{-value} = 0.39$, indicating no statistically significant difference among the distributions of the tested scanners after DISARM harmonization. The two plots in the bottom row show results for StarGANv2. We remark that the pre-harmonization boxplots highlight the outcomes of the different pre-processing methods used for DISARM and StarGANv2, where the latter is much stronger. In the post-harmonization for StarGANv2 (bottom right in Figure 4.4), the output distributions do not perfectly align with the reference image: the K-sample Anderson-Darling test yields a $p\text{-value}$ of 0.03, indicating a statistically significant difference among the distributions of the scanners ($p < 0.05$). This suggests that at least one scanner has a distribution that differs from the others, after StarGANv2 harmonization.

4.2.2 Evaluation of scanner free space

The harmonization performance in the scanner-free space $\mathcal{F}$ was tested. Figure 4.5 compares the average gray distributions for each scanner before and after the transfer to $\mathcal{F}$. The average post-harmonization distributions tend to align across all scanners, indicating successful harmonization ($p = 0.36$).

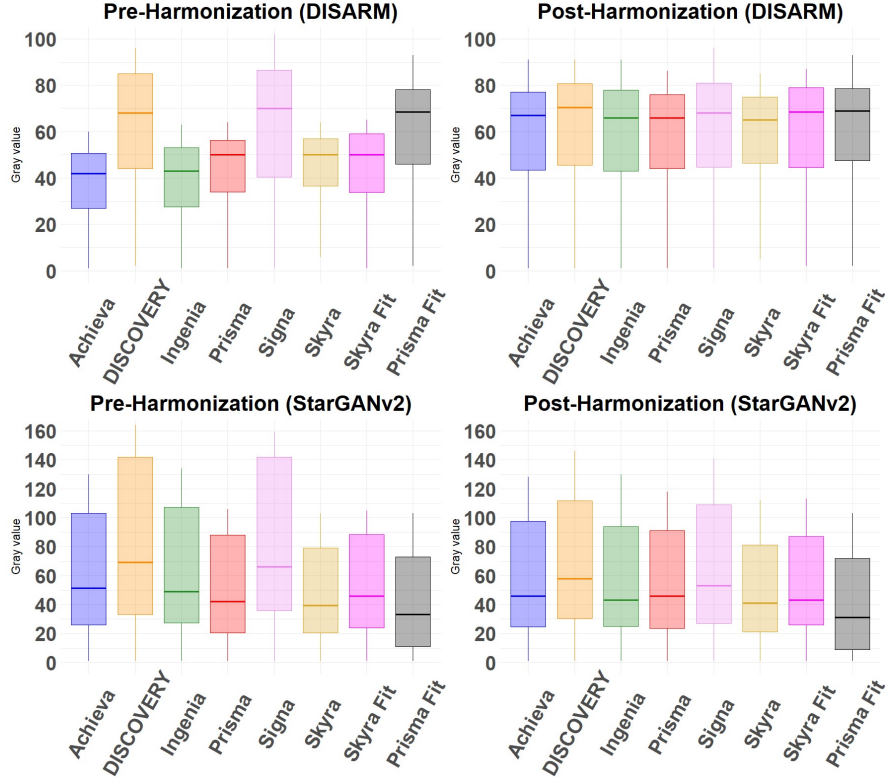


Figure 4.4: Average gray distributions in each scanner pre- and post-transfer to Prisma Fit scanner space. DISARM result (top row, from left to right): distributions after pre-processing and distributions post-harmonization. StarGANv2 result (bottom row, from left to right): distributions after pre-processing and distributions post-harmonization.

4.2.3 Preservation of anatomical structure

For a set of 100 random image pairs from the RIN dataset, which exhibit different anatomical structures, the average SSIM value is 0.700 ± 0.070 . The average SSIM between all 117 test images of the RIN dataset and their harmonized versions in the scanner-free space, obtained using DISARM, is 0.961 ± 0.007 . When harmonization is performed using the Prisma Fit scanner as reference, the average SSIM value is 0.955 ± 0.008 .

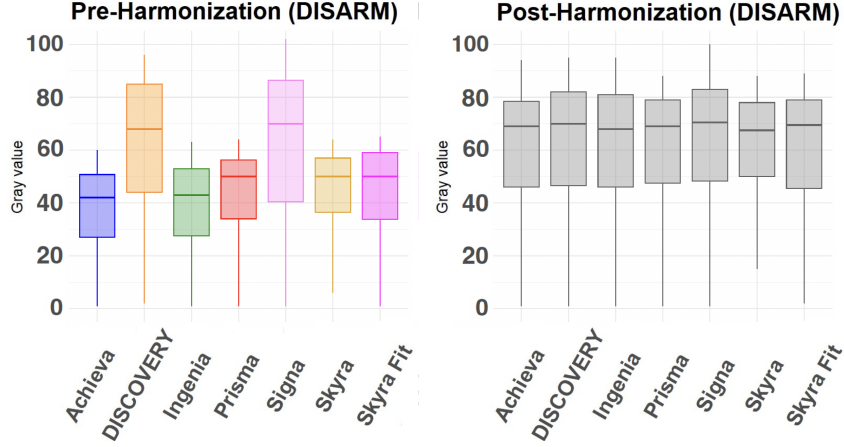


Figure 4.5: Average gray distributions in each scanner before (left) and after (right) the transfer to the scanner-free space $\mathcal{F}$.

4.2.4 Age prediction task

The resulting regression lines are displayed in Figure 4.6. Table 4.2 shows the results of the regression models before and after harmonization both via DISARM and StarGANv2 models. ‘TGV-Achieva’ etc. indicate if the relationship between TGV and age differs for each scanner compared to scanner Prisma. Before harmonization, there are significant interactions with different scanners ($p < 0.05$). After DISARM harmonization, the interaction with the scanner is no longer significant ($p > 0.05$), whereas after StarGANv2 harmonization, it remains weakly significant for two scanners ($p < 0.10$). This reflects the findings of the first experiment, highlighting how the scanner-related information still impacts the age regression task even after the harmonization procedure and sensibly contributes to the prediction. The RMSE between real and predicted age before harmonization is 7.19, after harmonization with DISARM is 7.05, and after harmonization with StarGANv2 is 7.45.

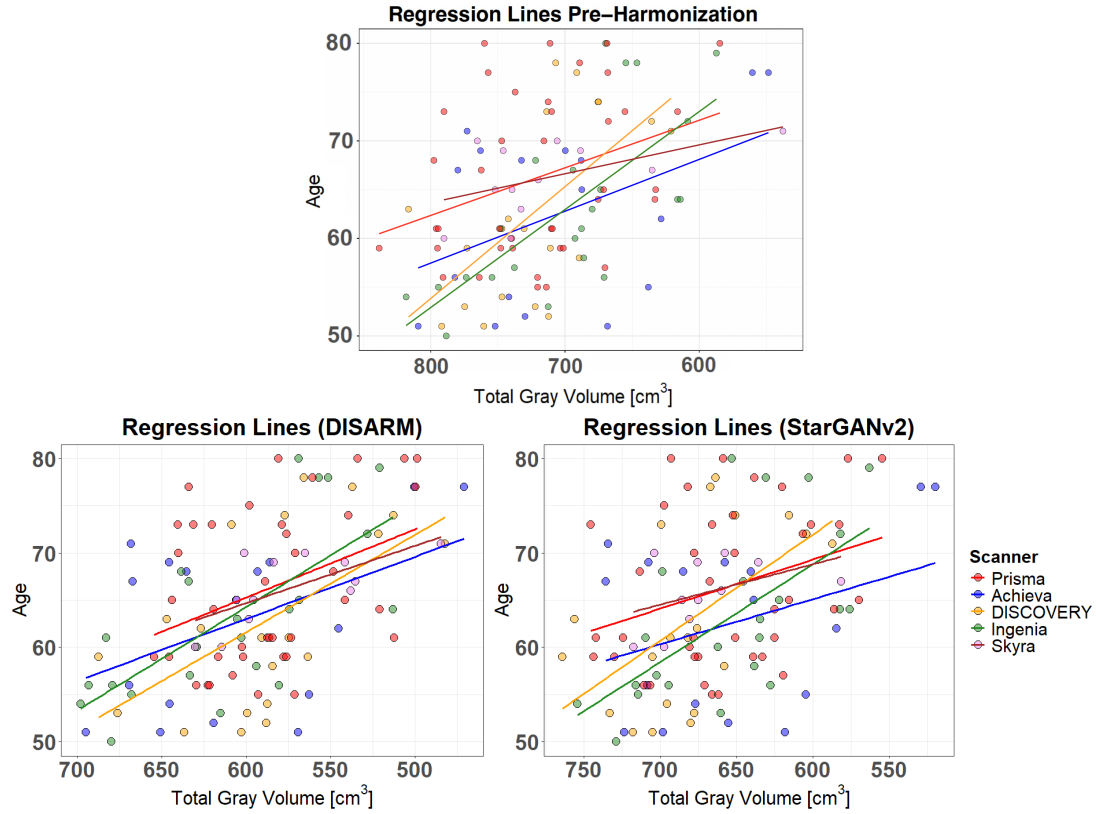


Figure 4.6: Comparison of age regression lines pre- (top panel) and post-harmonization (bottom row, DISARM on the left and StarGANv2 on the right).

Table 4.2: Significance of TGV variable and TGV-scanner interactions in age regression models before and after DISARM and StarGANv2 harmonization.

	Pre-Harmonization	DISARM	StarGANv2
Variables	P-value	P-value	P-value
Intercept	< 0.0001	< 0.0001	< 0.0001
TGV	< 0.0001	< 0.0001	< 0.0001
TGV-Achieva	0.03	0.38	0.08
TGV-DISCOVERY	0.08	0.10	0.33
TGV-Ingenia	0.02	0.52	0.08
TGV-Skyra	0.90	0.69	0.90

4.3 Discussion

DISARM is a novel unsupervised deep learning approach to harmonize T1-weighted MR images from various scanners, aimed at achieving three primary objectives: 1) creating a scanner-free space that enables the uniform transfer of images across different scanners in a denoised setting, 2) mapping images to the space of one of the scanners used in the model's training, embedding the unique characteristics of the selected scanner into the transferred image, 3) avoiding time-consuming pre-processing of MR images. Ultimately, DISARM should contribute to the reproducibility of MR images from different scanners, which is crucial in multicenter studies, as scanner-induced variability is known to impact the results significantly. The methodology of DISARM is based on disentangling image information into two distinct spaces: one encoding the scanner-specific information and one capturing the anatomical/biological structure of the image. The model was trained on two open-source datasets (ADNI and PPMI) of T1-weighted brain MR images of healthy control patients, and its performances were compared to a state-of-the-art model by testing the two methods on a real-world clinical dataset from the Italian Neuroimaging Network. The selection of a dataset that was not included in the training of either network ensures a fair comparison between the two. The results show superior harmonization power of the proposed model, demonstrating its effectiveness in achieving consistent and reproducible harmonization of the MR images across unseen scanning environments. This emerges both from the comparison of the mean gray level distributions pre- and post-harmonization and from the SSIM, indicating good preservation of the anatomical structure of the brain after harmonization. The downstream analysis on age prediction shows that DISARM eliminates scanner interactions with the variable TGV, and the RMSE shows improvement compared to the pre-harmonization regression results. Additionally, DISARM eliminates the need for time-consuming pre-processing steps like skull-stripping, which can be flawed and might remove parts of the brain or leave non-brain tissue. This is particularly beneficial

for applications requiring analysis of the entire head image, such as studies involving head trauma or skull deformities.

Further investigation involves assessing DISARM harmonization procedure on images of neurodegenerative patients, such as those affected by Alzheimer's disease (AD), for which the anatomical structures are typically compromised, particularly in the advanced stages of the disease. An evaluation of the changes in volume and cortical thickness of different anatomical regions after harmonization would also be useful to assess the level of preservation of structural integrity. Moreover, DISARM methodology could be extended to other imaging modalities to create a comprehensive framework for brain imaging harmonization.

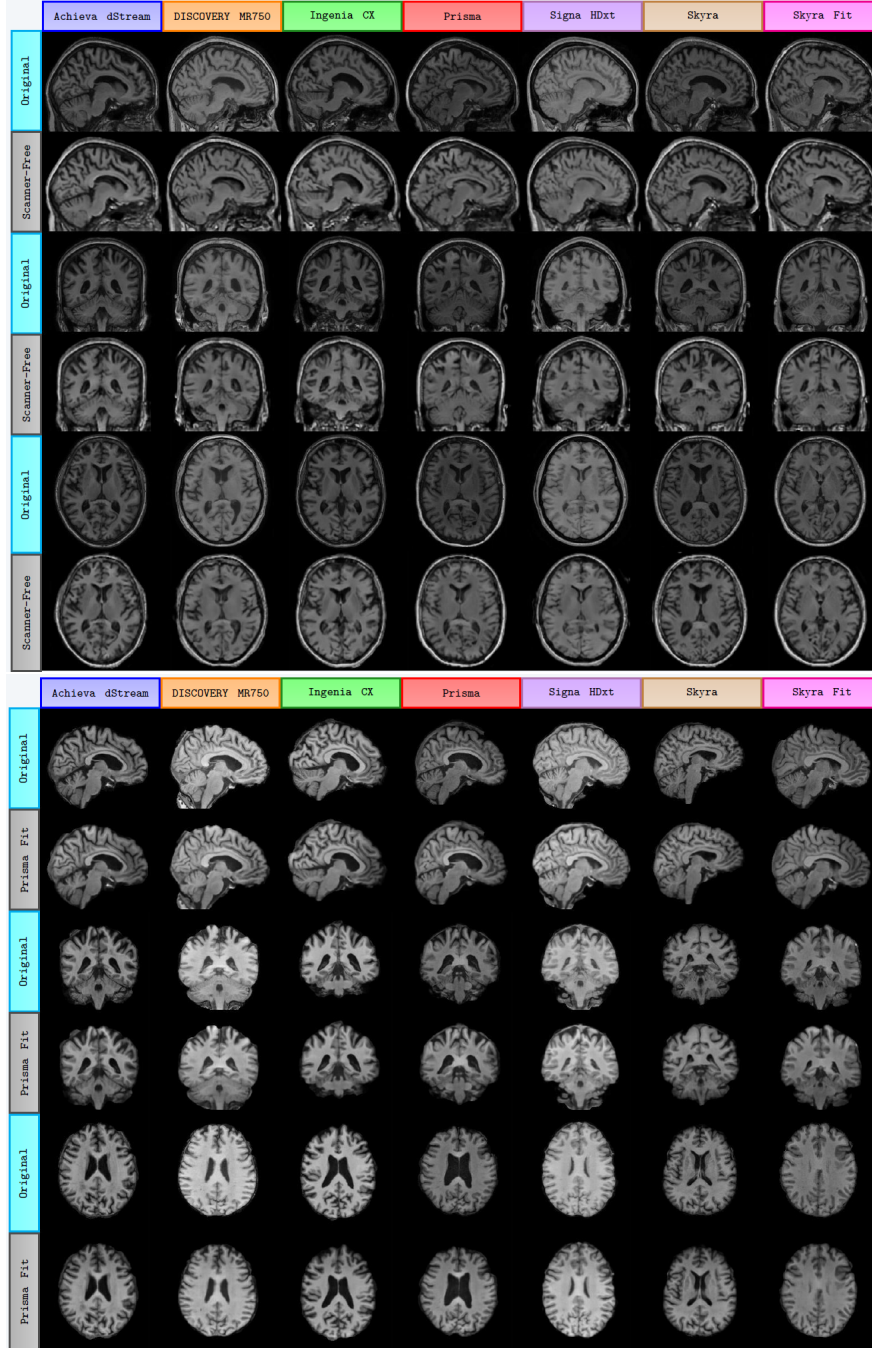


Figure 4.7: Rows 1, 3, and 5 both in top and bottom image display sagittal, coronal, and axial slices of the original images after the pre-processing step required by DISARM and StarGANv2, respectively (one image for each test scanner). Rows 2, 4, and 6 show the corresponding slices of the harmonized images: DISARM harmonization in the scanner-free space $\mathcal{F}$ (top image), and StarGANv2 harmonization to a reference image from Prisma Fit scanner (bottom image).

Conclusions

My doctoral research focused on key challenges of radiomic analysis, encompassing segmentation, evaluation of feature stability, construction of explicable prediction models for various clinical applications, and image harmonization. This thesis provides an overview of the conducted studies, including those that, while not directly applied to radiomics, contribute to the broader conceptual and methodological framework. Chapter 1 focuses on three dimensional numerical schemes for psoas muscle segmentation. Chapters 2 and 3 delve into multiple aspects of radiomics explainability, from investigating feature reliability with respect to segmentation variability, to studying the impact of radiomics stability on the predictive performance, as well as developing interpretable models for diverse applications. In particular, in Chapter 2 the predictive models are informed by features that are specifically selected based on their explainability for the breast cancer application at hand. In Chapter 3, an intuitive longitudinal model is built, where the contribution of pre-operative radiomics is used to produce a probabilistic estimate of the response to treatment at the first post-operative timepoint, with the final scope of forecasting eventual GBM progression at successive follow-ups. Finally, Chapter 4 introduces a deep-learning-based image harmonization approach aimed at overcoming imaging variability that hampers medical imaging studies, including radiomic ones. The experiments and analyses conducted in this thesis required different computational resources. The experiments described in Chapters 1, 2, and 3 were implemented on a personal computer, as they entailed moderate computational demands. In contrast, the model training

in Chapter 4 necessitated high-performance computing resources due to the intensive demands of the I2I harmonization framework.

While segmentation accuracy is essential for the assessment of muscle mass in neurodegenerative diseases such as ALS, as described in Chapter 1, segmentation impact on radiomics-based predictive tasks appears to be limited. According to the findings reported in Chapter 2, the prediction of TNBC subtype results to be independent across variable breast cancer segmentation masks. This suggests that achieving precise segmentation accuracy may not be a necessary prerequisite for radiomic applications. Moreover, results on reproducibility metric and reliability scores computed for the most predictive features suggest that applying feature selection techniques based on reproducibility across segmentations, or aimed at rewarding feature stability w.r.t. segmentation accuracy, may lead to the exclusion of predictive features from the model input. It is possible that, if radiomics exhibited greater predictive power, the impact of segmentation variability on prediction performance would be more evident. However, radiomics predictive capability remains modest, partly due to limited data but mainly because predictive models struggle to consistently identify universally predictive features across different data splits of the same dataset without overfitting. This limitation underscores the importance of using methods for collective explainable feature selection.

Similar issues arise from the prediction results of the longitudinal models for forecasting GBM response to treatment, presented in Chapter 3. Nevertheless, the proposed methodology establishes a framework for interpretable longitudinal prediction, where the use of radiomics is restricted to the pre-operative stage, thereby limiting the growth of problem dimensionality and mitigating the resultant loss of modeling intuition, that would result from indiscriminate aggregation of longitudinal features for model training.

Tackling the challenge of radiomics stability and explainability involves consideration of image harmonization, to reduce the impact of confounding variables, such as differences in scanner models. To this end, DISARM is in-

roduced, an I2I network developed to harmonize T1-weighted MR images by enabling image translation either to a specific scanner domain or to a scanner-free space. The model demonstrates robust generalization to scanners external to the training data. A downstream analysis on healthy subjects reveals that image harmonization improves the prediction of patient age, suggesting its potential for enhancing other biological information as well.

Our future endeavors involve the validation and generalization of the proposed methodologies, leveraging insights gained from various and larger datasets. Specifically, in relation to Chapter 1 on psoas muscle segmentation, we aim to enhance both the efficiency and applicability of the method by introducing adaptive and automated selection of parameters, both for the three-dimensional numerical scheme and the post-processing steps. This advancement could facilitate a more straightforward extension of the approach to other muscular regions in CT imaging and position the method as a viable alternative to deep learning frameworks for large-scale segmentation tasks. Regarding Chapter 2, on radiomics reliability, expanding the dataset could support the development of a benchmark for explainable prediction models in the clinical setting of tumor subtype classification, reinforcing the robustness of the models used in this critical evaluation of radiomics stability. Additionally, less deterministic approaches for simulating segmentation variability may offer a more accurate assessment of model reproducibility and generalizability across different clinical tasks and contexts. Further investigations should also consider integrating time-dependent features into the predictive models, as well as incorporating features extracted from additional MRI sequences. With regard to Chapter 3, on longitudinal radiomics, the availability of a larger longitudinal dataset would allow for a more comprehensive validation of the proposed methodology for longitudinal prediction of glioblastoma treatment response. In this chapter, the proposed probabilistic framework is used to model the first follow-up response in the test cohort, simulating missing longitudinal data in the target population for model application. Future work will investigate

incorporating the probabilistic modeling of intermediate responses into the training phase, thereby removing the dependency on intermediate treatment evaluations during model development. Furthermore, a nested modeling approach will be investigated to enable prediction beyond the second follow-up. Finally, the proposed longitudinal framework should be evaluated in diverse clinical contexts, including neurodegenerative diseases, which inherently require longitudinal analysis. As for Chapter 4, on image harmonization, future work will involve assessing the performance of the proposed harmonization model on neurodegenerative patients, adapting the model to preserve structural integrity in case of compromised anatomical structures. Additionally, the evaluation could be enhanced by analyzing scanner-related effects using traveling head subjects. Furthermore, additional downstream tasks should be explored to evaluate the impact of harmonization on enhancing the consistency and robustness of MRI-based biomarkers. These may include classification of AD versus healthy controls, as well as diagnosis prediction tasks such as distinguishing mild cognitive impairment (MCI) due to AD from AD dementia.

In conclusion, radiomics requires extensive quality control and appears highly application-specific. Despite its widespread adoption in various field of medical imaging, radiomics as a method for disease characterization remains a significant source of uncertainty, raising concerns about its actual predictive capability. If the challenges arising from the lack of reliability in radiomics are not rigorously addressed, they may compromise experimental reproducibility and radiomics clinical integration, potentially leading to inflated model performance expectations and unreliable decision-making in medical applications. While the investigation conducted in this thesis focuses on specific applications, the methodology introduced and the findings reported could potentially provide a broader understanding of the principles underlying radiomics-based analyses. Ultimately, this work aspires to contribute to a deeper understanding of the intricate interplay between radiomic feature stability, segmentation variability, model explainability, and image harmonization, fostering the devel-

opment of more reliable, generalizable, and clinically applicable computational models for diagnosis and prognosis.

Bibliography

- [1] R. J. Acciavatti et al. “Beyond Breast Density: Risk Measures for Breast Cancer in Multiple Imaging Modalities”. In: *Radiology* 306.3 (2023), e222575. DOI: 10.1148/radiol.222575.
- [2] S. S. Alahmari et al. “Delta Radiomics Improves Pulmonary Nodule Malignancy Prediction in Lung Cancer Screening”. In: *IEEE Access* 6 (2018), pp. 77796–77806. DOI: 10.1109/ACCESS.2018.2884126.
- [3] A. Alotaibi. “Deep generative adversarial networks for image-to-image translation: A review”. In: *Symmetry* 12.10 (2020), p. 1705. DOI: 10.3390/sym12101705.
- [4] D. Arefan et al. “Deep learning modeling using normal mammograms for predicting breast cancer risk”. In: *Medical Physics* 47.1 (2020), pp. 110–118. DOI: 10.1002/mp.13886.
- [5] B. Baeßler, K. Weiss, and D. P. Dos Santos. “Robustness and reproducibility of radiomics in magnetic resonance imaging: a phantom study”. In: *Investig. radiol.* 54.4 (2019), pp. 221–228. DOI: 10.1097/RLI.0000000000000530.
- [6] S. Bakas et al. “Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features”. In: *Scientific Data* 4.1 (2017), p. 170117. DOI: 10.1038/sdata.2017.117.
- [7] S. Bakas et al. “Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall

- Survival Prediction in the BRATS Challenge". In: *arXiv* (2019). DOI: 10.48550/arXiv.1811.02629.
- [8] S. Bakas et al. "The University of Pennsylvania glioblastoma (UPenn-GBM) cohort: advanced MRI, clinical, genomics, & radiomics". In: *Scientific Data* 9.1 (2022), p. 453. DOI: 10.1038/s41597-022-01560-7.
- [9] G. Barles and P. E. Souganidis. "Convergence of approximation schemes for fully nonlinear second order equations". In: *Asymptotic analysis* 4.3 (1991), pp. 271–283. DOI: 10.1109/CDC.1990.204046.
- [10] M. Bauckneht et al. "Opportunistic skeletal muscle metrics as prognostic tools in metastatic castration-resistant prostate cancer patients candidates to receive Radium-223". In: *Annals of Nuclear Medicine* 36.4 (2022), pp. 373–383. DOI: 10.1007/s12149-022-01716-w.
- [11] M. Bauckneht et al. "Spinal cord hypermetabolism extends to skeletal muscle in amyotrophic lateral sclerosis: a computational approach to [18F]-fluorodeoxyglucose PET/CT images". In: *EJNMMI research* 10.1 (2020), pp. 1–10. DOI: 10.1186/s13550-020-0607-5.
- [12] T. Bauer et al. "Subcortical grey matter volume and asymmetry in the long-term course of Rasmussen's encephalitis". In: *Brain Communications* 5.6 (2023), fcad324. DOI: 10.1093/braincomms/fcad324.
- [13] M. C. Beltrametti, A. M. Massone, and M. Piana. "Hough transform of special classes of curves". In: *SIAM Journal on Imaging Sciences* 6.1 (2013), pp. 391–412. DOI: 10.1137/120863794.
- [14] K. Bera et al. "Predicting cancer outcomes with radiomics and artificial intelligence in radiology". In: *Nature Reviews Clinical Oncology* 19.2 (2022), pp. 132–146. DOI: 10.1038/s41571-021-00560-7.
- [15] V. W. Berger and Y. Zhou. "Kolmogorov–Smirnov Test: Overview". In: *Wiley StatsRef: Statistics Reference Online*. 2014. DOI: 10.1002/9781118445112.stat06558.

-
- [16] G. Bianchini et al. "Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease". In: *Nature reviews Clinical oncology* 13.11 (2016), pp. 674–690. DOI: 10.1038/nrclinonc.2016.66.
 - [17] P. Bilic et al. "The Liver Tumor Segmentation Benchmark (LiTS)". In: *Medical Image Analysis* 84 (2023), p. 102680. DOI: 10.1016/j.media.2022.102680.
 - [18] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Vol. 4. 4. Berlin, Heidelberg: Springer-Verlag, 2006. DOI: 10.5555/1162264.
 - [19] G. W. Brier. "Verification of forecasts expressed in terms of probability". In: *Monthly Weather Review* 78.1 (1950), pp. 1–3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.
 - [20] M. Caballo et al. "Four-Dimensional Machine Learning Radiomics for the Pretreatment Assessment of Breast Cancer Pathologic Complete Response to Neoadjuvant Chemotherapy in Dynamic Contrast-Enhanced MRI". In: *Journal of Magnetic Resonance Imaging* 57.1 (2023), pp. 97–110. DOI: 10.1002/jmri.28273.
 - [21] M. Caballo et al. "Multi-marker quantitative radiomics for mass characterization in dedicated breast CT imaging". In: *Medical Physics* 48.1 (2021), pp. 313–328. DOI: 10.1002/mp.14610.
 - [22] L. Caldera et al. "DISARM: Disentangled Scanner-Free Image Generation via Unsupervised Image2Image Translation". In: *Machine Learning in Clinical Neuroimaging*. Ed. by D. R. Bathula et al. Cham: Springer Nature Switzerland, 2025, pp. 102–112. DOI: 10.1007/978-3-031-78761-4_10.
 - [23] I. Cama et al. "A study on the role of radiomics feature stability in predicting breast cancer subtypes". In: *17th International Workshop on Breast Imaging (IWBI 2024)*. Vol. 13174. SPIE. 2024, pp. 418–427. DOI: 10.1117/12.3027015.

- [24] I. Cama et al. "Probabilistic approach to longitudinal response prediction: application to radiomics from brain cancer imaging". In: *arXiv* (2025). DOI: 10.48550/arXiv.2505.07973.
- [25] I. Cama et al. "Segmentation agreement and the reliability of radiomics features". In: *Advances in Computational Science and Engineering* 1.2 (2023), pp. 202–217. DOI: 10.3934/acse.2023009.
- [26] I. Cama et al. "Segmentation variability and radiomics stability for predicting Triple-Negative Breast Cancer subtype using Magnetic Resonance Imaging". In: *arXiv* (2025). DOI: 10.48550/arXiv.2504.01692.
- [27] *Cancer Moonshot Biobank - Prostate Cancer Collection (CMB-PCA) (Version 5) [dataset]*. *The Cancer Imaging Archive*. 2022. DOI: 10.7937/25T7-6Y12.
- [28] V. Caselles, R. Kimmel, and G. Sapiro. "Geodesic active contours". In: *International journal of computer vision* 22 (1997), pp. 61–79. DOI: 10.1023/A:1007979827043.
- [29] J. Castiglione et al. "Automated Segmentation of Abdominal Skeletal Muscle on Pediatric CT Scans Using Deep Learning". In: *Radiology: Artificial Intelligence* 3.2 (2021), e200130. DOI: 10.1148/ryai.2021200130.
- [30] C. Charroud and L. Turella. "Subcortical grey matter changes associated with motor symptoms evaluated by the Unified Parkinson's disease Rating Scale (part III): A longitudinal study in Parkinson's disease". In: *NeuroImage: Clinical* 31 (2021), p. 102745. DOI: 10.1016/j.nicl.2021.102745.
- [31] N. V. Chawla et al. "SMOTE: Synthetic Minority Over-sampling Technique". In: *The Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357. DOI: 10.1613/jair.953.
- [32] B. Chen et al. "A fractional order derivative based active contour model for inhomogeneous image segmentation". In: *Applied Mathematical Modelling* 65 (2019), pp. 120–136. DOI: 10.1016/j.apm.2018.08.009.

- [33] J. Chindapasirt. "Sarcopenia in cancer patients". In: *Asian Pacific Journal of Cancer Prevention* 16.18 (2015), pp. 8075–7. DOI: 10.7314/apjcp.2015.16.18.8075.
- [34] S. J. Cho et al. "Prediction of treatment response after stereotactic radiosurgery of brain metastasis using deep learning and radiomics on longitudinal MRI data". In: *Scientific Reports* 14.1 (2024), p. 11085. DOI: 10.1038/s41598-024-60781-5.
- [35] W. Choi et al. "Development and validation of a reliable method for automated measurements of psoas muscle volume in CT scans using deep learning-based segmentation: a cross-sectional study". In: *BMJ Open* 14.5 (2024), e079417. DOI: 10.1136/bmjopen-2023-079417.
- [36] Y. Choi et al. "StarGAN v2: Diverse Image Synthesis for Multiple Domains". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 8185–8194. DOI: 10.1109/CVPR42600.2020.00821.
- [37] T. Chougule et al. "Radiomics signature for temporal evolution and recurrence patterns of glioblastoma using multimodal magnetic resonance imaging". In: *NMR in Biomedicine* 35.3 (2022), e4647. DOI: 10.1002/nbm.4647.
- [38] K. Clark et al. "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository". In: *Journal of Digital Imaging* 26.6 (2013), pp. 1045–1057. DOI: 10.1007/s10278-013-9622-7.
- [39] R. J. Clifford et al. "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods". In: *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 27.4 (2008), pp. 685–691. DOI: 10.1002/jmri.21049.
- [40] J. R. Clough et al. "A topological loss function for deep-learning based image segmentation using persistent homology". In: *IEEE transactions on pattern analysis and machine intelligence* 44.12 (2020), pp. 8766–8778. DOI: 10.1109/TPAMI.2020.3013679.

- [41] J. Collins et al. "The assessment and impact of sarcopenia in lung cancer: a systematic literature review". In: *BMJ open* 4.1 (2014), e003697. DOI: 10.1136/bmjopen-2013-003697.
- [42] M. G. Crandall and P.-L. Lions. "Viscosity solutions of Hamilton-Jacobi equations". In: *Transactions of the American mathematical society* 277.1 (1983), pp. 1–42. DOI: 10.1090/S0002-9947-1983-0690039-8.
- [43] L. Crespi, D. Loiacono, and P. Sartori. "Are 3D better than 2D Convolutional Neural Networks for Medical Imaging Semantic Segmentation?" In: *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2022, pp. 1–8. DOI: 10.1109/IJCNN55064.2022.9892850.
- [44] A. J. Cruz-Jentoft et al. "Sarcopenia: revised European consensus on definition and diagnosis". In: *Age and ageing* 48.1 (2019), pp. 16–31. DOI: 10.1093/ageing/afy169.
- [45] C. D'Orsi et al. *2013 ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*. American College of Radiology, 2014. URL: <https://books.google.it/books?id=nhWSjwEACAAJ>.
- [46] M. S. Daoud et al. "Gradient-based optimizer (GBO): a review, theory, variants, and applications". In: *Archives of Computational Methods in Engineering* 30.4 (2023), pp. 2431–2449. DOI: 10.1007/s11831-022-09872-y.
- [47] N. H. David Newitt. *Multi-center breast DCE-MRI data and segmentations from patients in the I-SPY 1/ACRIN 6657 trials [Data set]*. The Cancer Imaging Archive. 2016. DOI: 10.7937/K9/TCIA.2016.HdHpgJLK.
- [48] N. H. David Newitt. *Single site breast DCE-MRI data and segmentations from patients undergoing neoadjuvant chemotherapy (version 3) [Data set]*. The Cancer Imaging Archive. 2016. DOI: 10.7937/K9/TCIA.2016.QHsyhJKy.

- [49] K. Dembrower et al. "Comparison of a Deep Learning Risk Score and Standard Mammographic Density Score for Breast Cancer Risk Prediction". In: *Radiology* 294.2 (2019), pp. 265–272. DOI: 10.1148/radiol.2019190872.
- [50] A. Demircioğlu. "The effect of preprocessing filters on predictive performance in radiomics". In: *European Radiology Experimental* 6.1 (2022), p. 40. DOI: 10.1186/s41747-022-00294-w.
- [51] B. A. Derstine et al. "Skeletal muscle cutoff values for sarcopenia diagnosis using T10 to L5 measurements in a healthy US population". In: *Scientific reports* 8.1 (2018), p. 11369. DOI: 10.1038/s41598-018-29825-5.
- [52] L. R. Dice. "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26.3 (1945), pp. 297–302. DOI: 10.2307/1932409.
- [53] M. Drey et al. "Associations between early markers of Parkinson's disease and sarcopenia". In: *Frontiers in Aging Neuroscience* 9 (2017), p. 53. DOI: 10.3389/fnagi.2017.00053.
- [54] F. Duong et al. "Automated major psoas muscle volumetry in computed tomography using machine learning algorithms". In: *International Journal of Computer Assisted Radiology and Surgery* 17 (2 2022), pp. 355–361. DOI: 10.1007/s11548-021-02539-2.
- [55] M. Fan et al. "Radiomics of Tumor Heterogeneity in Longitudinal Dynamic Contrast-Enhanced Magnetic Resonance Imaging for Predicting Response to Neoadjuvant Chemotherapy in Breast Cancer". In: *Frontiers in Molecular Biosciences* 8 (2021). DOI: 10.3389/fmolb.2021.622219.
- [56] B. Farina et al. "Integration of longitudinal deep-radiomics and clinical data improves the prediction of durable benefits to anti-PD-1/PD-L1 immunotherapy in advanced NSCLC patients". In: *Journal of Translational Medicine* 21.1 (2023), p. 174. DOI: 10.1186/s12967-023-04004-x.
- [57] J. A. Fitzpatrick et al. "Large-scale analysis of iliopsoas muscle volumes in the UK Biobank". In: *Scientific Reports* 10.1 (2020), p. 20215. DOI: 10.1038/s41598-020-77351-0.

- [58] V. S. Fonov et al. "Unbiased nonlinear average age-appropriate brain templates from birth to adulthood". In: *NeuroImage* 47 (2009), S102. DOI: 10.1016/S1053-8119(09)70884-5.
- [59] W. D. Foulkes, I. E. Smith, and J. S. Reis-Filho. "Triple-negative breast cancer". In: *New England Journal of Medicine* 363.20 (2010), pp. 1938–1948. DOI: 10.1056/NEJMr1001389.
- [60] L. Garrucho et al. "MAMA-MIA: A Large-Scale Multi-Center Breast Cancer DCE-MRI Benchmark Dataset with Expert Segmentations". In: *arXiv* (2024). DOI: 10.48550/arXiv.2406.13844.
- [61] R. W. Y. Granzier et al. "MRI-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability". In: *Scientific Reports* 10.1 (2020), p. 14163. DOI: 10.1038/s41598-020-70940-z.
- [62] J. J. M. van Griethuysen et al. "Computational Radiomics System to Decode the Radiographic Phenotype". In: *Cancer Research* 77.21 (2017), e104–e107. DOI: 10.1158/0008-5472.CAN-17-0339.
- [63] A. Grossberg et al. *Data from Head and Neck Cancer CT Atlas (Version 2) [Data set]*. The Cancer Imaging Archive. 2017. DOI: 10.7937/K9/TCIA.2017.umz8dv6s.
- [64] L. Guo et al. "Breast cancer heterogeneity and its implication in personalized precision therapy". In: *Experimental Hematology & Oncology* 12.1 (2023), p. 3. DOI: 10.1186/s40164-022-00363-1.
- [65] Z. Guo et al. "Transformer for image harmonization and beyond". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022). DOI: 10.1109/TPAMI.2022.3207091.
- [66] D. Gut et al. "Benchmarking of deep architectures for segmentation of medical images". In: *IEEE Transactions on Medical Imaging* 41.11 (2022), pp. 3231–3241. DOI: 10.1109/TMI.2022.3180435.

- [67] D. Hao et al. "A self-training teacher-student model with an automatic label grader for abdominal skeletal muscle segmentation". In: *Artificial Intelligence in Medicine* 132 (2022), p. 102366. DOI: 10.1016/j.artmed.2022.102366.
- [68] F. Hashimoto et al. "Automated segmentation of 2D low-dose CT images of the psoas-major muscle using deep convolutional neural networks". In: *Radiological Physics and Technology* 12 (2 2019), pp. 210–215. DOI: 10.1007/s12194-019-00512-y.
- [69] F. Hausdorff. *Grundzuge der Mengenlehre, Leipzig*. 1914.
- [70] T. Heimann et al. "Comparison and Evaluation of Methods for Liver Segmentation From CT Datasets". In: *IEEE Transactions on Medical Imaging* 28.8 (2009), pp. 1251–1265. DOI: 10.1109/TMI.2009.2013851.
- [71] B. Hell et al. "An approach toward fast gradient-based image segmentation". In: *IEEE Transactions on Image Processing* 24.9 (2015), pp. 2633–2645. DOI: 10.1109/TIP.2015.2419078.
- [72] T. Inoue et al. "Psoas Major Muscle Segmentation Using Higher-Order Shape Prior". In: *Medical Computer Vision: Algorithms for Big Data*. Ed. by B. Menze et al. Cham: Springer International Publishing, 2016, pp. 116–124. DOI: 10.1007/978-3-319-42016-5_11.
- [73] F. Isensee et al. "Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge". In: *International MICCAI Brain lesion Workshop*. Springer. 2017, pp. 287–297. DOI: 10.1007/978-3-319-75238-9_25.
- [74] F. Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* 18.2 (2021), pp. 203–211. DOI: 10.1038/s41592-020-01008-z.
- [75] S. Islam et al. "Fully automated deep-learning section-based muscle segmentation from CT images for sarcopenia assessment". In: *Clinical Radiology* 77.5 (2022), e363–e371. DOI: 10.1016/j.crad.2022.01.036.

- [76] P. Jaccard. “Étude comparative de la distribution florale dans une portion des Alpes et des Jura”. In: *Bulletin de la Société Vaudoise des Sciences Naturelles* 37 (1901), pp. 547–579. DOI: 10.5169/seals-266450.
- [77] M. Jenkinson et al. “FSL”. In: *Neuroimage* 62.2 (2012), pp. 782–790. DOI: 10.1016/j.neuroimage.2011.09.015.
- [78] M. Jenkinson et al. “Improved optimization for the robust and accurate linear registration and motion correction of brain images”. In: *Neuroimage* 17.2 (2002), pp. 825–841. DOI: 10.1006/nimg.2002.1132.
- [79] A. K. Jha et al. “Repeatability and reproducibility study of radiomic features on a phantom and human cohort”. In: *Scientific Reports* 11.1 (2021), p. 2055. DOI: 10.1038/s41598-021-81526-8.
- [80] Z. Jiang et al. “Multimodality MRI-based radiomics approach to predict the posttreatment response of lung cancer brain metastases to gamma knife radiosurgery”. In: *European Radiology* 32.4 (2022), pp. 2266–2276. DOI: 10.1007/s00330-021-08368-w.
- [81] X. Jing et al. “Automated Breast Density Assessment in MRI Using Deep Learning and Radiomics: Strategies for Reducing Inter-Observer Variability”. In: *Journal of Magnetic Resonance Imaging* 60.1 (2024), pp. 80–91. DOI: 10.1002/jmri.29058.
- [82] K. S. Johnson, E. F. Conant, and M. S. Soo. “Molecular subtypes of breast cancer: a review for breast radiologists”. In: *Journal of Breast Imaging* 3.1 (2021), pp. 12–24. DOI: 10.1093/jbi/wbaa110.
- [83] T. Kaido et al. “Impact of sarcopenia on survival in patients undergoing living donor liver transplantation”. In: *American Journal of Transplantation* 13.6 (2013), pp. 1549–1556. DOI: 10.1111/ajt.12221.
- [84] N. Kamiya et al. “Automated segmentation of psoas major muscle in X-ray CT images by use of a shape model: preliminary study”. In: *Radiological Physics and Technology* 5 (1 2012), pp. 5–14. DOI: 10.1007/s12194-011-0127-0.

- [85] N. Kamiya et al. "Fully automatic segmentation of paraspinal muscles from 3D torso CT images via multi-scale iterative random forest classifications". In: *International Journal of Computer Assisted Radiology and Surgery* 13 (11 2018), pp. 1697–1706. DOI: 10.1007/s11548-018-1852-1.
- [86] M. Kaur, J. Kaur, and J. Kaur. "Survey of contrast enhancement techniques based on histogram equalization". In: *International Journal of Advanced Computer Science and Applications* 2.7 (2011). DOI: 10.14569/IJACSA.2011.020721.
- [87] K. E. Keenan et al. "Quantitative magnetic resonance imaging phantoms: A review and the need for a system phantom". In: *Magnetic Resonance in Medicine* 79.1 (2018), pp. 48–61. DOI: 10.1002/mrm.26982.
- [88] P. Kickingereder et al. "Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study". In: *The Lancet Oncology* 20.5 (2019), pp. 728–740. DOI: 10.1016/S1470-2045(19)30098-1.
- [89] D. P. Kingma and J. Ba. *Adam: A Method for Stochastic Optimization*. 2017. DOI: 10.48550/arXiv.1412.6980.
- [90] H. C. Kniep et al. "Radiomics of Brain MRI: Utility in Prediction of Metastatic Tumor Type". In: *Radiology* 290.2 (2019), pp. 479–487. DOI: 10.1148/radiol.2018180946.
- [91] B. Kocak et al. "Quality of radiomics research: comprehensive analysis of 1574 unique publications from 89 reviews". In: *European Radiology* (2024). DOI: 10.1007/s00330-024-11057-z.
- [92] S. Koitka et al. *SAROS - A large, heterogeneous, and sparsely annotated segmentation dataset on CT imaging data (SAROS) (Version 2)[Data set]*. The Cancer Imaging Archive. 2023. DOI: 10.25737/SZ96-ZG60.
- [93] T. K. Koo and M. Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research". In: *Journal of Chiropractic Medicine* 15 (2 June 2016), pp. 155–163. DOI: 10.1016/J.JCM.2016.02.012.

- [94] G. Kothari et al. "The impact of inter-observer variation in delineation on robustness of radiomics features in non-small cell lung cancer". In: *Scientific Reports* 12.1 (2022), p. 12822. DOI: 10.1038/s41598-022-16520-9.
- [95] D. Krex et al. "Long-term survival with glioblastoma multiforme". In: *Brain* 130.10 (Sept. 2007), pp. 2596–2606. DOI: 10.1093/brain/awm204.
- [96] N. Kumar et al. "The unique risk factor profile of triple-negative breast cancer: a comprehensive meta-analysis". In: *JNCI: Journal of the National Cancer Institute* 116.8 (2024), pp. 1210–1219. DOI: 10.1093/jnci/djae056.
- [97] P. Lambin et al. "Radiomics: extracting more information from medical images using advanced feature analysis". In: *European Journal of Cancer* 48.4 (2012), pp. 441–446. DOI: 10.1016/j.ejca.2011.11.036.
- [98] J. Lao et al. "A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme". In: *Scientific Reports* 7.1 (2017), pp. 1–8. DOI: 10.1038/s41598-017-10649-8.
- [99] H. Y. Lee et al. "DRIT++: Diverse image-to-image translation via disentangled representations". In: *International Journal of Computer Vision* 128 (2020), pp. 2402–2417. DOI: 10.1007/s11263-019-01284-z.
- [100] D. Leithner et al. "Non-invasive assessment of breast cancer molecular subtypes with multiparametric magnetic resonance imaging radiomics". In: *Journal of clinical medicine* 9.6 (2020), p. 1853. DOI: 10.3390/jcm9061853.
- [101] D. Leithner et al. "Radiomic signatures with contrast-enhanced magnetic resonance imaging for the assessment of breast cancer receptor status and molecular subtypes: initial results". In: *Breast Cancer Research* 21.1 (2019), p. 106. DOI: 10.1186/s13058-019-1187-z.
- [102] H. Li et al. "A level set method for topological shape optimization of 3D structures with extrusion constraints". In: *Computer Methods in Applied Mechanics and Engineering* 283 (2015), pp. 615–635. DOI: 10.1016/j.cma.2014.10.006.

- [103] N. Li et al. "Correlation between the CEM imaging characteristics and different molecular subtypes of breast cancer". In: *The Breast* 72 (2023), p. 103595. DOI: 10.1016/j.breast.2023.103595.
- [104] E. J. Limkin et al. "Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology". In: *Annals of Oncology* 28.6 (2017), pp. 1191–1206. DOI: 10.1093/annonc/mdx034.
- [105] P. Lions. "Two approximations of solutions of Hamilton-Jacobi equations". In: *Mathematics of Computation* 43 (1984), pp. 1–19. DOI: 10.1090/s0025-5718-1984-0744921-8.
- [106] T. J. Littlejohns et al. "The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions". In: *Nature Communications* 11.1 (2020), p. 2624. DOI: 10.1038/s41467-020-15948-9.
- [107] M. Liu et al. "Style transfer generative adversarial networks to harmonize multisite MRI to a single reference image to avoid overcorrection". In: *Human Brain Mapping* 44.14 (2023), pp. 4875–4892. DOI: 10.1002/hbm.26422.
- [108] M. de Bruijne et al., eds. *Style Transfer Using Generative Adversarial Networks for Multi-site MRI Harmonization*. Springer International Publishing, 2021, pp. 313–322. DOI: 10.1007/978-3-030-87199-4_30.
- [109] R. Liu et al. "Stability analysis of CT radiomic features with respect to segmentation variation in oropharyngeal cancer". In: *Clinical and Translational Radiation Oncology* 21 (2020), pp. 11–18. DOI: 10.1016/j.ctro.2019.11.005.
- [110] R. Liu et al. "End-to-end Lane Shape Prediction with Transformers". In: *arXiv* (2020). DOI: 10.48550/arXiv.2011.04233.
- [111] S. Liu and P.-T. Yap. "Learning multi-site harmonization of magnetic resonance images without traveling human phantoms". In: *Communications Engineering* 3.1 (2024), p. 6. DOI: 10.1038/s44172-023-00140-w.

BIBLIOGRAPHY

- [112] P. Lohmann et al. "Radiomics in neuro-oncology: Basics, workflow, and applications". In: *Methods* 188 (2021), pp. 112–121. DOI: 10.1016/j.ymeth.2020.06.003.
- [113] S. M. Lundberg and S.-I. Lee. "A Unified Approach to Interpreting Model Predictions". In: *arXiv* (2017). DOI: 10.48550/arXiv.1705.07874.
- [114] X. Luo et al. "Semi-supervised medical image segmentation through dual-task consistency". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 10. 2021, pp. 8801–8809. DOI: 10.1609/aaai.v35i10.17066.
- [115] W. Ma et al. "Breast Cancer Molecular Subtype Prediction by Mammographic Radiomic Features". In: *Academic Radiology* 26.2 (2019), pp. 196–201. DOI: 10.1016/j.acra.2018.01.023.
- [116] D. Mackin et al. "Measuring computed tomography scanner variability of radiomics features". In: *Investigative Radiology* 50 (11 2015). DOI: 10.1097/RLI.0000000000000180.
- [117] D. V. C. Mai et al. "A systematic review of automated segmentation of 3D computed-tomography scans for volumetric body composition analysis". In: *Journal of Cachexia, Sarcopenia and Muscle* 14.5 (2023), pp. 1973–1986. DOI: 10.1002/jcsm.13310.
- [118] K.-C. J. Mammonov AB. *QIN GBM Treatment Response. [Data set]*. The Cancer Imaging Archive. 2016. DOI: 10.7937/k9/tcia.2016.nQF4gpn2.
- [119] T. Manabe et al. "Usefulness of the Measurement of Psoas Muscle Volume for Sarcopenia Diagnosis in Patients with Liver Disease". In: *Diagnostics* 13.7 (2023), p. 1245. DOI: 10.3390/diagnostics13071245.
- [120] A. Manduca et al. "Texture features from mammographic images and risk of breast cancer". In: *Cancer Epidemiology Biomarkers & Prevention* 18.3 (2009), pp. 837–845. DOI: 10.1158/1055-9965.EPI-08-0631.

BIBLIOGRAPHY

- [121] K. Marek et al. "The Parkinson progression marker initiative (PPMI)". In: *Progress in neurobiology* 95.4 (2011), pp. 629–635. DOI: 10.1016/j.pneurobio.2011.09.005.
- [122] T. Masuda et al. "Sarcopenia is a prognostic factor in living donor liver transplantation". In: *Liver Transplantation* 20.4 (2014), pp. 401–407. DOI: 10.1002/lt.23811.
- [123] A. M. McCarthy et al. "Relationship of established risk factors with breast cancer subtypes". In: *Cancer Medicine* 10.18 (2021), pp. 6456–6467. DOI: 10.1002/cam4.4158.
- [124] E. Mema et al. "The relationship of breast density in mammography and magnetic resonance imaging in women with triple negative breast cancer". In: *European Journal of Radiology* 124 (2020), p. 108813. DOI: 10.1016/j.ejrad.2020.108813.
- [125] R. Montoya-del-Angel et al. "Added value of feature uncertainty in a radiomic analysis of contrast-enhanced digital mammography boosted by deep learning". In: *17th International Workshop on Breast Imaging (IWBI 2024)*. Ed. by M. L. Giger et al. Vol. 13174. International Society for Optics and Photonics. SPIE, 2024, p. 1317403. DOI: 10.1117/12.3027034.
- [126] D. Newitt et al. *ACRIN 6698/I-SPY2 Breast DWI [Data set]*. The Cancer Imaging Archive. 2021. DOI: 10.7937/TCIA.KK02-6D95.
- [127] T. G. Nick and K. M. Campbell. "Logistic Regression". In: *Topics in Biostatistics*. Ed. by W. T. Ambrosius. Totowa, NJ: Humana Press, 2007, pp. 273–301. DOI: 10.1007/978-1-59745-530-5_14.
- [128] A. Niculescu-Mizil and R. Caruana. *Predicting good probabilities with supervised learning*. Jan. 2005, pp. 625–632. DOI: 10.1145/1102351.1102430.
- [129] M. Nijati et al. "Deep learning and radiomics of longitudinal CT scans for early prediction of tuberculosis treatment outcomes". In: *European Journal of Radiology* 169 (2023), p. 111180. DOI: 10.1016/j.ejrad.2023.111180.

-
- [130] M. Nixon and A. Aguado. *Feature extraction and image processing for computer vision*. Academic press, 2019. DOI: 10.1016/C2011-0-06935-1.
- [131] F. Orlhac et al. "A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies". In: *Journal of Nuclear Medicine* 63.2 (2022), pp. 172–179. DOI: 10.2967/jnumed.121.262464.
- [132] S. Osher and R. P. Fedkiw. "Level set methods: an overview and some recent results". In: *Journal of Computational physics* 169.2 (2001), pp. 463–502. DOI: 10.1006/jcph.2000.6636.
- [133] S. Osher and J. A. Sethian. "Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations". In: *Journal of computational physics* 79.1 (1988), pp. 12–49. DOI: 10.1016/0021-9991(88)90002-2.
- [134] F. Pamoukdjian et al. "Prevalence and predictive value of pre-therapeutic sarcopenia in cancer patients: a systematic review". In: *Clinical Nutrition* 37.4 (2018), pp. 1101–1113. DOI: 10.1016/j.clnu.2017.07.010.
- [135] S. Pan and B. M. Dawant. "Automatic 3D segmentation of the liver from abdominal CT images: a level-set approach". In: *Medical Imaging 2001: Image Processing*. Vol. 4322. SPIE. 2001, pp. 128–138. DOI: 10.1117/12.431019.
- [136] G. Paolucci et al. "Three-dimensional numerical schemes for the segmentation of the psoas muscle in X-ray computed tomography images". In: *BMC Medical Imaging* 24.1 (2024), p. 251. DOI: 10.1186/s12880-024-01423-0.
- [137] J. E. Park et al. "A systematic review reporting quality of radiomics research in neuro-oncology: toward clinical utility and quality improvement using high-dimensional imaging features". In: *BMC Cancer* 20.1 (2020), p. 29. ISSN: 1471-2407. DOI: 10.1186/s12885-019-6504-5.
- [138] M. Patel et al. "Machine learning-based radiomic evaluation of treatment response prediction in glioblastoma". In: *Clinical Radiology* 76.8 (2021), 628.e17–628.e27. DOI: 10.1016/j.crad.2021.03.019.

- [139] F. Pérez-García, R. Sparks, and S. Ourselin. “TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning”. In: *Computer Methods and Programs in Biomedicine* 208 (2021), p. 106236. DOI: 10.1016/j.cmpb.2021.106236.
- [140] J. Platt. “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods”. In: *Advances in Large Margin Classifiers* 10 (2000), pp. 61–74.
- [141] E. B. Podgorsak. *Radiation oncology physics*. 2005.
- [142] M. G. Poiriot et al. “Robustness of radiomics to variations in segmentation methods in multimodal brain MRI”. In: *Scientific Reports* 12 (2022), p. 16712. DOI: 10.1038/s41598-022-20703-9.
- [143] R. Pomponio et al. “Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan”. In: *NeuroImage* 208 (2020), p. 116450. DOI: 10.1016/j.neuroimage.2019.116450.
- [144] C. Qin et al. “Joint Transformer and Multi-scale CNN for DCE-MRI Breast Cancer Segmentation”. In: *Soft Computing* 26.17 (2022), pp. 8317–8334. DOI: 10.1007/s00500-022-07235-0.
- [145] J. Radua et al. “Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA”. In: *Neuroimage* 218 (2020), p. 116956. DOI: 10.1016/j.neuroimage.2020.116956.
- [146] A. Rahmim et al. “Improved prediction of outcome in Parkinson’s disease using radiomics analysis of longitudinal DAT SPECT images”. In: *NeuroImage: Clinical* 16 (2017), pp. 539–544. DOI: 10.1016/j.nicl.2017.08.021.
- [147] B. Rister et al. *CT-ORG: A Dataset of CT Volumes With Multiple Organ Segmentations (Version 1) [Data set]*. The Cancer Imaging Archive. 2019. DOI: 10.7937/tcia.2019.tt7f4v7o.

- [148] S. Rizzo et al. "Radiomics: the facts and the challenges of image analysis". In: *Eur. Radiol. Exp.* 2.1 (2018), p. 36. ISSN: 2509-9280. DOI: 10.1186/s41747-018-0068-z.
- [149] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by N. Navab et al. Cham: Springer International Publishing, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.
- [150] H. Roth et al. *A new 2.5 D representation for lymph node detection in CT[Data set]*. The Cancer Imaging Archive. 2015. DOI: 10.7937/K9/TCIA.2015.AQIIDCNM.
- [151] M. Rozynek et al. "Body composition radiomic features as a predictor of survival in patients with non-small cellular lung carcinoma: A multicenter retrospective study". In: *Nutrition* 120 (2024), p. 112336. DOI: 10.1016/j.nut.2023.112336.
- [152] M. Rozynek et al. "Fully automated 3D body composition analysis and its association with overall survival in head and neck squamous cell carcinoma patients". In: *Frontiers in Oncology* 13 (2023). DOI: 10.3389/fonc.2023.1176425.
- [153] J. D. Rudie et al. "Longitudinal Assessment of Posttreatment Diffuse Glioma Tissue Volumes with Three-dimensional Convolutional Neural Networks". In: *Radiology: Artificial Intelligence* 4.5 (2022), e210243. DOI: 10.1148/ryai.210243.
- [154] M. Rumpf and T. Preusser. "A level set method for anisotropic geometric diffusion in 3D image processing". In: *SIAM Journal on Applied Mathematics* 62.5 (2002), pp. 1772–1793. DOI: 10.1137/S0036139901384662.
- [155] A. Saha et al. "A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features". In: *British journal of cancer* 119.4 (2018), pp. 508–516.

- [156] A. Saha et al. *Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [Data set]*. The Cancer Imaging Archive. 2021. DOI: 10.7937/TCIA.e3sv-re93.
- [157] A. Saha et al. "Effects of MRI scanner parameters on breast cancer radiomics". In: *Expert Systems with Applications* 87 (2017), pp. 384–391. DOI: 10.1016/j.eswa.2017.06.029.
- [158] N. Saltybaeva et al. "Robustness of radiomic features in magnetic resonance imaging for patients with glioblastoma: Multi-center study". In: *Physics and Imaging in Radiation Oncology* 22 (2022), pp. 131–136. DOI: 10.1016/j.phro.2022.05.006.
- [159] E. Scalco, G. Rizzo, and A. Mastropietro. "The stability of oncologic MRI radiomic features and the potential role of deep learning: a review". In: *Physics in Medicine & Biology* 67.9 (2022), 09TR03. DOI: 10.1088/1361-6560/ac60b9.
- [160] D. Schenone et al. "Radiomics and Artificial Intelligence for Outcome Prediction in Multiple Myeloma Patients Undergoing Autologous Transplantation: A Feasibility Study with CT Data". In: *Diagnostics* 11.10 (2021). ISSN: 2075-4418. DOI: 10.3390/diagnostics11101759.
- [161] F. W. Scholz and M. A. Stephens. "K-sample Anderson–Darling tests". In: *Journal of the American Statistical Association* 82.399 (1987), pp. 918–924. DOI: 10.2307/2288805.
- [162] J. G. Schopp et al. "Imaging Challenges in Diagnosing Triple-Negative Breast Cancer". In: *RadioGraphics* 43.10 (2023), e230027. DOI: 10.1148/rg.230027.
- [163] D. W. Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015. DOI: 10.1002/9780470316849.
- [164] F. Ségonne et al. "A hybrid approach to the skull stripping problem in MRI". In: *Neuroimage* 22.3 (2004), pp. 1060–1075. DOI: 10.1016/j.neuroimage.2004.03.032.

- [165] Y. Sha and J. Chen. "MRI-based radiomics for the diagnosis of triple-negative breast cancer: a meta-analysis". In: *Clinical Radiology* 77.9 (2022), pp. 655–663. DOI: 10.1016/j.crad.2022.04.015.
- [166] J. X. Shen et al. "Longitudinal radiomics algorithm of posttreatment computed tomography images for early detecting recurrence of hepatocellular carcinoma after resection or ablation". In: *Translational Oncology* 14.1 (2021), p. 100866. DOI: 10.1016/j.tranon.2020.100866.
- [167] W. Sheng et al. "Invasive ductal breast cancer molecular subtype prediction by MRI radiomic and clinical features based on machine learning". In: *Frontiers in Oncology* Volume 12 - 2022 (2022). DOI: 10.3389/fonc.2022.964605.
- [168] R. T. Shinohara et al. "Volumetric analysis from a harmonized multisite brain MRI study of a single subject with multiple sclerosis". In: *American Journal of Neuroradiology* 38.8 (2017), pp. 1501–1509. DOI: 10.3174/ajnr.A5254.
- [169] B. W. Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018. DOI: 10.1201/9781315140919.
- [170] J. G. Sled, A. P. Zijdenbos, and A. C. Evans. "A nonparametric method for automatic correction of intensity nonuniformity in MRI data". In: *IEEE Transactions on Medical Imaging* 17.1 (1998), pp. 87–97. DOI: 10.1109/42.668698.
- [171] K. Smith et al. *Data From CT COLONOGRAPHY [Data set]*. The Cancer Imaging Archive. 2015. DOI: 10.7937/K9/TCIA.2015.NWTESAY1.
- [172] S. M. Smith et al. "Advances in functional and structural MR image analysis and implementation as FSL". In: *Neuroimage* 23 (2004), S208–S219. DOI: 10.1016/j.neuroimage.2004.07.051.
- [173] J. Son et al. "Prediction of breast cancer molecular subtypes using radiomics signatures of synthetic mammography from digital breast tomosynthesis". In: *Scientific reports* 10.1 (2020), p. 21566. DOI: 10.1038/s41598-020-78681-9.

- [174] A. Stanzione et al. "Prostate MRI radiomics: A systematic review and radiomic quality score assessment". In: *Eur. J. of Radiol.* 129 (2020), p. 109095. DOI: 10.1016/j.ejrad.2020.109095.
- [175] S. Starke et al. "Longitudinal and Multimodal Radiomics Models for Head and Neck Cancer Outcome Prediction". In: *Cancers* 15.3 (2023). DOI: 10.3390/cancers15030673.
- [176] Y. Suter et al. "The LUMIERE dataset: Longitudinal Glioblastoma MRI with expert RANO evaluation". In: *Scientific Data* 9.1 (2022), p. 768. DOI: 10.1038/s41597-022-01881-7.
- [177] A. S. Tagliafico et al. "Overview of radiomics in breast cancer diagnosis and prognostication". In: *The Breast* 49 (2020), pp. 74–80. DOI: 10.1016/j.breast.2019.10.018.
- [178] H. Takao, N. Hayashi, and K. Ohtomo. "Effect of scanner in longitudinal studies of brain volume changes". In: *Journal of Magnetic Resonance Imaging* 34.2 (2011), pp. 438–444. DOI: 10.1002/jmri.22636.
- [179] J. P. Thakkar et al. "Epidemiologic and Molecular Prognostic Review of Glioblastoma". In: *Cancer Epidemiology, Biomarkers & Prevention* 23.10 (2014), pp. 1985–1996. DOI: 10.1158/1055-9965.EPI-14-0275.
- [180] J. E. van Timmeren et al. "Longitudinal Image Data for Outcome Modeling". In: *Clinical Oncology* (2024), p. 103610. DOI: 10.1016/j.clon.2024.06.053.
- [181] J. E. van Timmeren et al. "Radiomics in medical imaging—"how-to" guide and critical reflection". In: *Insights into Imaging* 11.1 (2020), p. 91. ISSN: 1869-4101. DOI: 10.1186/s13244-020-00887-2.
- [182] J. E. van Timmeren et al. "Longitudinal radiomics of cone-beam CT images from non-small cell lung cancer patients: Evaluation of the added prognostic value for overall survival and locoregional recurrence". In: *Radiotherapy and Oncology* 136 (2019), pp. 78–85. DOI: 10.1016/j.radonc.2019.03.032.

- [183] F. Tixier et al. "Reliability of tumor segmentation in glioblastoma: Impact on the robustness of MRI-radiomic features". In: *Medical Physics* 46.8 (2019), pp. 3582–3591. DOI: 10.1002/mp.13624.
- [184] M. E. Torbati et al. "A multi-scanner neuroimaging data harmonization using RAVEL and ComBat". In: *Neuroimage* 245 (2021), p. 118703. DOI: 10.1016/j.neuroimage.2021.118703.
- [185] A. Traverso et al. "Repeatability and Reproducibility of Radiomic Features: A Systematic Review". In: *International Journal of Radiation Oncology, Biology, Physics* 102 (4 Nov. 2018), pp. 1143–1158. DOI: 10.1016/J.IJROBP.2018.05.053.
- [186] G. Turashvili and E. Brogi. "Tumor Heterogeneity in Breast Cancer". In: *Frontiers in Medicine* Volume 4 (2017). DOI: 10.3389/fmed.2017.00227.
- [187] S. Van der Walt et al. "scikit-image: image processing in Python". In: *PeerJ* 2 (2014), e453. DOI: 10.7717/peerj.453.
- [188] D. Van Erck et al. "Evaluation of a Fully Automatic Deep Learning-Based Method for the Measurement of Psoas Muscle Area". In: *Frontiers in Nutrition* 9 (2022). DOI: 10.3389/fnut.2022.781860.
- [189] O. Vergara-Fernandez, M. Trejo-Avila, and N. Salgado-Nesme. "Sarcopenia in patients with colorectal cancer: A comprehensive review". In: *World Journal of Clinical Cases* 8.7 (2020), p. 1188. DOI: 10.12998/wjcc.v8.i7.1188.
- [190] B. Villarini et al. "3D Deep Learning for Anatomical Structure Segmentation in Multiple Imaging Modalities". In: *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. 2021, pp. 166–171. DOI: 10.1109/cbms52027.2021.00066.
- [191] A. Villaseñor et al. "Prevalence and prognostic effect of sarcopenia in breast cancer survivors: the HEAL Study". In: *Journal of Cancer Survivorship* 6 (2012), pp. 398–406. DOI: 10.1007/s11764-012-0234-x.

- [192] J. Wang et al. "The use of longitudinal CT-based radiomics and clinicopathological features predicts the pathological complete response of metastasized axillary lymph nodes in breast cancer". In: *BMC Cancer* 24.1 (2024), p. 549. DOI: 10.1186/s12885-024-12257-y.
- [193] M. Wang et al. "A Computed Tomography-based Radiomics Nomogram for Predicting Osteoporotic Vertebral Fractures: A Longitudinal Study". In: *The Journal of Clinical Endocrinology & Metabolism* 108.6 (Dec. 2022), e283–e294. DOI: 10.1210/clinem/dgac722.
- [194] P. Wang et al. "Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice". In: *arXiv* (2022). DOI: 10.48550/arXiv.2203.05962.
- [195] J. Wasserthal et al. "TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images". In: *Radiology: Artificial Intelligence* 5.5 (2023). DOI: 10.1148/ryai.230024.
- [196] J. Wasserthal et al. "TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images". In: *Radiology: Artificial Intelligence* 5.5 (2023), e230024.
- [197] P. Y. Wen et al. "Updated Response Assessment Criteria for High-Grade Gliomas: Response Assessment in Neuro-Oncology Working Group". In: *Journal of Clinical Oncology* 28.11 (2010), pp. 1963–1972. DOI: 10.1200/JCO.2009.26.3541.
- [198] H. M. Whitney et al. "Harmonization of radiomic features of breast lesions across international DCE-MRI datasets". In: *Journal of Medical Imaging* 7.1 (2020), p. 012707. DOI: 10.1117/1.JMI.7.1.012707.
- [199] J. Wong et al. "Effects of interobserver and interdisciplinary segmentation variabilities on CT-based radiomics for pancreatic cancer". In: *Sci. Rep.* 11.1 (2021), p. 16328. DOI: 10.1038/s41598-021-95152-x.
- [200] G. Wu et al. "Sparse representation-based radiomics for the diagnosis of brain tumors". In: *IEEE Transactions on Medical Imaging* 37.4 (2017), pp. 893–905. DOI: 10.1109/TMI.2017.2776967.

- [201] C. Xue et al. "Radiomics feature reliability assessed by intraclass correlation coefficient: a systematic review". In: *Quantitative Imaging in Medicine and Surgery* 11.10 (2021), p. 4431. DOI: 10.21037/qims-21-86.
- [202] F. Yang et al. "Impact of contouring variability on oncological PET radiomics features in the lung". In: *Scientific Reports* 10.1 (2020), p. 369. DOI: 10.1038/s41598-019-57171-7.
- [203] H. Yin et al. "Noninvasive assessment of breast cancer molecular subtypes on multiparametric MRI using convolutional neural network with transfer learning". In: *Thoracic Cancer* 13.22 (2022), pp. 3183–3191. DOI: 10.1111/1759-7714.14673.
- [204] H. Yuksel et al. "Sarcopenia in patients with multiple sclerosis". In: *Multiple Sclerosis and Related Disorders* 58 (2022), p. 103471. DOI: 10.1016/j.msard.2021.103471.
- [205] T. Zhang et al. "Predicting breast cancer types on and beyond molecular level in a multi-modal fashion". In: *npj Breast Cancer* 9.1 (2023), p. 16. DOI: 10.1038/s41523-023-00517-2.
- [206] Y. Zhang et al. "Prediction of breast cancer molecular subtypes on DCE-MRI using convolutional neural network with transfer learning between two centers". In: *European Radiology* 31.4 (2021), pp. 2559–2567. DOI: 10.1007/s00330-020-07274-x.
- [207] L. Zhao et al. "MSKD: Structured knowledge distillation for efficient medical image segmentation". In: *Computers in Biology and Medicine* 164 (2023), p. 107284. DOI: 10.1016/j.combiomed.2023.107284.
- [208] Y. Zheng et al. "Parenchymal texture analysis in digital mammography: A fully automated pipeline for breast cancer risk assessment". In: *Medical Physics* 42.7 (2015), pp. 4149–4160. DOI: 10.1118/1.4921996.
- [209] J. Zhou et al. "Predicting the response to neoadjuvant chemotherapy for breast cancer: wavelet transforming radiomics in MRI". In: *BMC Cancer* 20.1 (2020), p. 100. DOI: 10.1186/s12885-020-6523-2.

- [210] J. Zhou et al. "Predicting the response to neoadjuvant chemotherapy for breast cancer: wavelet transforming radiomics in MRI". In: *BMC Cancer* 20.1 (2020), p. 100. DOI: 10.1186/s12885-020-6523-2.
- [211] D. Zopfs et al. "Single-slice CT measurements allow for accurate assessment of sarcopenia and body composition". In: *European radiology* 30 (2020), pp. 1701–1708. DOI: 10.1007/s00330-019-06526-9.
- [212] L. Zuo et al. "Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory". In: *NeuroImage* 243 (2021), p. 118569. DOI: 10.1016/j.neuroimage.2021.118569.
- [213] X.-N. Zuo, T. Xu, and M. P. Milham. "Harnessing reliability for neuroscience research". In: *Nature human behaviour* 3.8 (2019), pp. 768–771. DOI: 10.1038/s41562-019-0655-x.
- [214] A. Zwanenburg et al. "Assessing robustness of radiomic features by image perturbation". In: *Scientific Reports* 9.1 (2019), p. 614. DOI: 10.1038/s41598-018-36938-4.