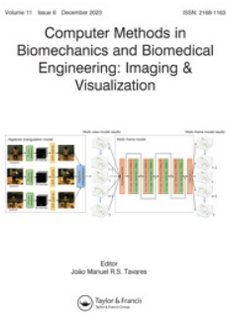


Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization

ISSN: 2168-1163 (Print) 2168-1171 (Online) Journal homepage: www.tandfonline.com/journals/tciv20

Exploring the complexities of radiomics: an in-depth analysis of a machine learning pipeline for predicting rectal cancer therapy response using MRI

Arianna Defeudis , Jovana Panic , Lorenzo Vassallo , Stefano Cirillo , Marco Gatti , Riccardo Faletti , Antonio Esposito , Anna Palmisano , Serena Dell'Aversana , Alfonso Ragozzino , Salvatore Siena , Angelo Vanzulli , Daniele Regge , Samanta Rosati , Gabriella Balestra & Valentina Giannini

To cite this article: Arianna Defeudis , Jovana Panic , Lorenzo Vassallo , Stefano Cirillo , Marco Gatti , Riccardo Faletti , Antonio Esposito , Anna Palmisano , Serena Dell'Aversana , Alfonso Ragozzino , Salvatore Siena , Angelo Vanzulli , Daniele Regge , Samanta Rosati , Gabriella Balestra & Valentina Giannini (2025) Exploring the complexities of radiomics: an in-depth analysis of a machine learning pipeline for predicting rectal cancer therapy response using MRI, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 13:1, 2564261, DOI: 10.1080/21681163.2025.2564261

To link to this article: https://doi.org/10.1080/21681163.2025.2564261

COMPUTER METHODS IN BIOMECHANICS AND BIOMEDICAL ENGINEERING: IMAGING & VISUALIZATION
2025, VOL. 13, NO. 1, 2564261
https://doi.org/10.1080/21681163.2025.2564261

Exploring the complexities of radiomics: an in-depth analysis of a machine learning pipeline for predicting rectal cancer therapy response using MRI

Arianna Defeudis[a], Jovana Panic[b,c,d], Lorenzo Vassallo[e], Stefano Cirillo[f], Marco Gatti[g], Riccardo Faletti[g], Antonio Esposito[h,i], Anna Palmisano[h,i], Serena Dell'Aversana[j], Alfonso Ragozzino[k], Salvatore Siena[l,m], Angelo Vanzulli[l,m], Daniele Regge[a], Samanta Rosati[d], Gabriella Balestra[d] and Valentina Giannini[a,n]

[a]Candiolo Cancer Institute, FPO-IRCCS, Candiolo, Italy; [b]D3 Center, Osaka University, Osaka, Japan; [c]Premium Research Institute for Human Metaverse Medicine (WPI-PRIMe), Osaka University, Osaka, Japan; [d]Department of Electronics and Telecommunications, Polytechnic of Turin, Turin, Italy; [e]Department of Diagnostic Imaging and Radiotherapy, AOU Città della Salute e della Scienza, Turin, Italy; [f]Department of Radiology, A.O. Ordine Mauriziano (Ospedale Umbertol), Turin, Italy; [g]Department of Surgical Science, University of Turin, Turin, Italy; [h]School of Medicine, Vita-Salute San Raffaele University, Milan, Italy; [i]Experimental Imaging Center, IRCCS San Raffaele Scientific Institute, Milano, Italy; [j]Department of Radiology, Santa Maria delle Grazie Hospital, ASL Napoli 2 Nord, Pozzuoli, Italy; [k]Bioinformatics and Biostatistics Lab, IRCCS SYNLAB SDN, School of Engineering, Naples, Italy; [l]Niguarda Cancer Center, Grande Ospedale Metropolitano Niguarda, Milan, Italy; [m]Department of Oncology and Hemato-oncology, University of Milan, Milan, Italy; [n]Department of Oncology, University of Turin, Turin, Italy

ABSTRACT

Developing radiomic biomarkers remains challenging due to the variability in imaging protocols across centres and the lack of standardised methodologies. This study evaluates the impact of different technical decisions in radiomics pipelines using multiparametric magnetic resonance imaging data from six centres for predicting therapy response in rectal cancer. Preprocessing, feature extraction, normalisation strategies, and machine learning (ML) models were assessed for robustness and generalisability. Key findings demonstrated that the preprocessing significantly enhanced the feature reproducibility and the importance of the selected ones over the complexity of the ML classifier. This analysis highlights the necessity of a unique pipeline and the complexity of radiomics biomarkers in real-world settings, particularly when handling highly imbalanced datasets. Several insights and methodologies have been presented that may support towards more conscious decisions when implementing radiomic systems. Future efforts should focus on integrating clinical/genomic/pathomics data to improve the predictive capabilities and facilitate the introduction into clinical practice.

ARTICLE HISTORY
Received 13 January 2025
Accepted 14 September 2025

KEYWORDS
Radiomics guidelines; MRI; therapy response; rectal cancer

1. Introduction

Over the past decade, there has been a notable increase in interest in developing artificial intelligence (AI) systems to assist clinicians in oncological medical imaging. These developments have yielded encouraging outcomes, suggesting the potential use of AI-based techniques for automated image segmentation and detection (McKinney et al. 2020; Lipkova et al. 2022; Wang et al. 2022; Bordron et al. 2022; He et al. 2023). However, more challenging tasks such as developing prognostic imaging biomarkers to characterise tumours or predicting treatment response are still limited for different reasons, including the lack of multicenter and extended external validation (Santinha et al. 2024). One of the primary factors contributing is not only the dearth of international collaborations but also the absence of standardised methods able to cope with inherent differences between images from different centres. Indeed, the inevitable high variability between patients' images acquired with different protocols and equipment strongly constrains the reproducibility, repeatability, and generalisability of the results, introducing inconsistencies that are an obstacle to the interpretability of the outcomes (Mali et al. 2021; Giannini et al. 2023; Panic et al. 2023; Santinha et al. 2024).

Therefore, to develop radiomic biomarkers that are robust and likely to be translatable to the clinic, several important steps must be carefully analysed, and thoughtful decisions must be made

CONTACT Arianna Defeudis arianna.defeudis@ircc.it Candiolo Cancer Institute, FPO-IRCCS, strada Provinciale, Candiolo, TO, Italy
Supplemental data for this article can be accessed online at https://doi.org/10.1080/21681163.2025.2564261

regarding them. These decisions include minimising the unwanted variability that characterises multicenter databases, identifying the most informative and representative data, and developing the most appropriate AI model for each task. However, even if some studies addressed generalisability issues, mostly on deep learning-based segmentation tasks on brain and cardiac magnetic resonance imaging (MRI) (Graves et al. 2020; Ribeiro and Nunesy 2021; De Dumast and Bach Cuadra 2023), there are a noticeable lack of analyses evaluating the impact of various methodological choices on the robustness, reliability, and generalisability of machine learning (ML)-based multicenter biomarkers. Furthermore, while several studies propose standardised radiomics pipelines, they often fail to assess the differences in outcomes and the influence of each methodological decision (Da-Ano et al. 2020; Giannini et al. 2023; Panic et al. 2023).

The objective of this study is to evaluate the impact of various technical decisions at different stages of the radiomics pipeline on the performance of ML-based algorithms, using the development of radiomics biomarkers for predicting therapy response in rectal cancer (RC) as a case study. RC is the third leading cancer-related cause of death worldwide in both men and women (Benson et al. 2022; Siegel et al. 2023). The current treatment plan includes neoadjuvant chemoradiotherapy (nCRT) (Benson et al. 2020); however, 75–85% of patients who either partially respond or experience tumour progression may undergo unnecessary treatments associated with high toxicities. Unfortunately, no biomarkers are currently available to predict the individual RC patient's response.

Therefore, our study aims to propose standard guidelines for the development of an MRI-based radiomic signature for RC's therapy response prediction, assessing and evaluating the impact of multiple technical decisions on model's accuracy. Indeed, we believe it is crucial to define commonly agreed guidelines, by providing suggestions and insights for radiomics applications, facilitating the translation of such systems into clinical practice.

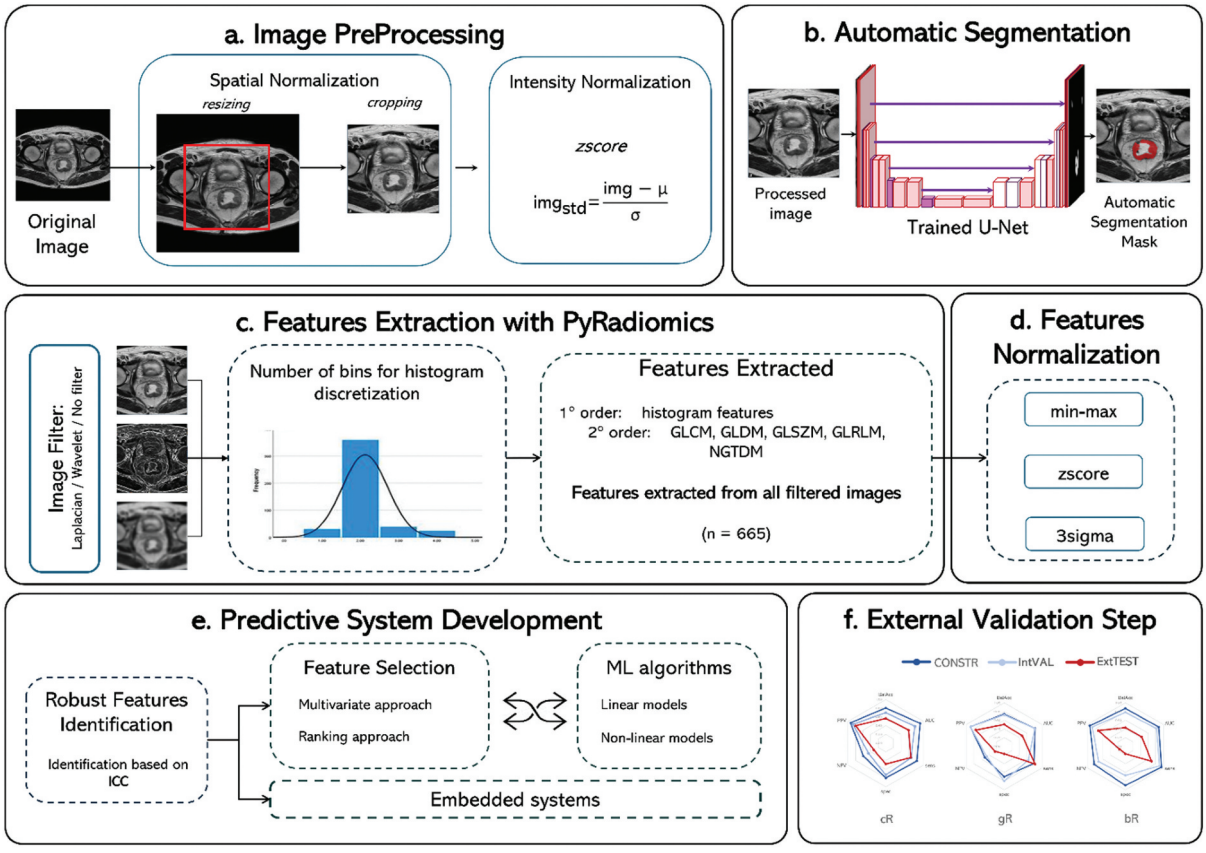


Figure 1. Flowchart of the study.

2. Materials and methods

The steps that will be addressed in this study are depicted in Figure 1. More in detail, we first present the preprocessing methods applied to reduce the intrinsic image variability (Figure 1(a)) and the automatic segmentation algorithm (Figure 1b), then we focus on the technical decisions related to the feature extraction (Figure 1c) and normalization (Figure 1d), and the impact of different Feature selection and machine learning (ML) approaches for the predictive system development (Figure 1e). Finally, we evaluate the performance on the external validation set (Figure 1f).

2.1. Dataset

In this study, multiparametric MRIs of patients with pathologically proven RC were retrospectively collected from six different Italian institutions from November 2000 to September 2019.

The inclusion criteria were: (a) biopsy confirmed RC; (b) MRI exam performed before nCRT, including at least the axial T2-weighted (T2w) sequence and the Apparent Diffusion Coefficient (ADC) map (Da-Ano et al. 2020); (c) assessment of the response to therapy. This multicenter retrospective study was approved by the institutional review boards (IRBs) of each institution, with a waiver for the requirement of signed informed consent, as de-identified data were used. Table 1 shows acquisition parameters and vendors per centre. Table Supp 1 shows the number of patients according to the pathological tumor regression grade (pTRG) and clinical response.

2.2. Reference standard

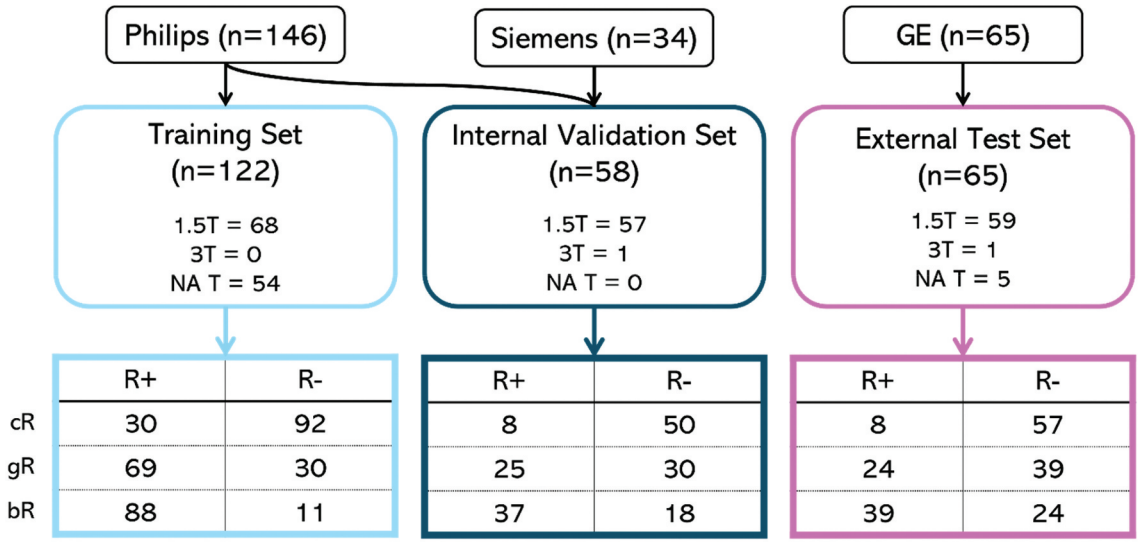
For the development of the predictive systems, the reference standard was either the pTRG ($n = 207$) or the clinical response to therapy ($n = 38$). pTRG was evaluated by experienced pathologists, blinded to clinical information and MRI findings from the resected tumour, and assessed using the Mandard classification (Mandard et al. 1994). Three different clinical aims were considered:

- *Prediction of the complete Responder (cR)*: Patients were classified as responders (R+) if they had a pTRG = 1 or, in case of missing pTRG, achieved a complete clinical response; otherwise, they were classified as non-responders (*R-*).
- *Prediction of the good Responder (gR)*: R+ patients were defined as those with pTRG ≤ 2 or, in case of a missing pTRG, those reaching a clinical complete response, and *R-* those with pTRG > 2. Patients for whom the pTRG was missing and not achieving a clinical complete response, were discarded for this and the following clinical aim.
- *Prediction of the bad Responder (bR)*: patients were dichotomised as R+ if they had a pTRG ≤ 3 or if they reached a complete clinical response, and as *R-* if pTRG was ≥ 4.

Table 1. Dataset characteristics according to the centres and scanner characteristics.

	Scanner Vendor											
	Construction Set						ExtTEST					
	Philips			Siemens			GE					
Center	1.5T	3T	NA	1.5T	3T	NA	1.5T	3T	NA	Pixel resolution (M-IQR) (mm)	Slice thickness (M-IQR) (mm)	FOV (M-IQR) (mm)
01	/	/	1	/	/	/	57	/	5	0.45 (0.43–0.45)	4.40 (4.00–4.40)	230 (220–230)
02	24*	/	/	/	/	/	/	/	/	0.47 (0.47–0.74)	3.50 (3.50–4.00)	240 (240–240)
03	37	/	/	/	/	/	/	/	/	0.49 (0.49–0.49)	3.00 (3.00–3.00)	250 (205–250)
04	31	/	/	4*	1*	/	2	/	/	0.55 (0.54–0.55)	5.00 (4.00–5.00)	280 (280–280)
05	/	/	/	29*	/	/	/	1	/	0.63 (0.63–0.63)	3.00 (3.00–3.00)	200 (200–200)
06	/	/	53	/	/	/	/	/	/	0.72 (0.72–0.77)	4.00 (4.00–4.02)	405 (385–405)

Note: ExtTEST: External Validation Set; NA: Not Available Magnetic Field Strength; M: median; IQR: Inter-Quartile Range; FOV: quadratic Field of View, since both axes have the same dimensions, * sequences included in the Internal Validation.



	R+	R-	R+	R-	R+	R-
cR	30	92	8	50	8	57
gR	69	30	25	30	24	39
bR	88	11	37	18	39	24

Figure 2. Dataset division into CONSTR, IntVAL, and ExtTEST according to the different clinical aims.

2.3. Dataset division

We split the dataset by MRI vendors to address the challenges of a vendor-agnostic strategy. Table 1 presents the MRI characteristics stratified by centre and scanner, while Figure 2 shows the dataset division. The construction set included sequences from both Philips and Siemens scanners, whereas the external test set (ExtTEST) comprised only GE exams. The construction set was further divided into a training set (TR) and an internal validation set (IntVAL). TR, used for algorithm training, contained only exams acquired with a Philips scanner. IntVAL, that was used to select the best-performing model for each clinical aim, included all Siemens exams and Philips exams from *center 02*. This partition was designed to assess the impact of both centre-specific and manufacturer-related variations.

2.4. Image preprocessing

The images in the dataset showed high variability in terms of both spatial resolution and signal intensity, characteristics, highly dependent on the scanner and acquisition parameters (Scalco and Rizzo 2017), and on the field of view, i.e. since different anatomical structures can be included This resulted in very different histograms (Figure Supp 1). To reduce the impact of noise, variability, and heterogeneity of the images, we defined a preprocessing step that included both spatial and intensity normalisation approaches, as previously reported (Giannini et al. 2023; Panic et al. 2023). First, we resampled all images to match the highest in-plane resolution found in the construction set. Then, we cropped them to the smallest FOV among all vendors and centres in the same set (Table 1). As a result, all images have a standardised FOV of 18 cm and an in-plane resolution of 0.47×0.47 mm. This step is performed to obtain images representing the same anatomical structures and in which each voxel contains the same amount of information. The same spatial normalisation was applied to the masks as well (Figure 1a). Finally, to standardise image intensity, we applied z-score normalisation, which has been shown to reduce overall variability in histogram shape, intensity ranges, and feature distribution differences across multi-centre MRI datasets more effectively than other methods (e.g. min-max, p1p99, or no normalisation). Specifically, z-score normalisation increased the proportion of statistically similar features from 5% to 11% (Panic et al. 2023). Moreover, according to a meta-analysis (Giannini et al. 2023), *z-score* combined with resampling and resize methods allows to yield an area under the ROC curve (AUC) in predicting response to therapy from 0.60 (95% Confidence Interval (CI): 0.49–0.71) to 0.79 (95% CI = 0.76–0.82).

2.5. Automatic segmentation

To move towards a fully automated system for rectal cancer detection and prognosis, for this study we relied on automatic segmentation masks instead of manual annotations. These masks were obtained using an already developed and validated fully convolutional network, e.g. U-Net, trained on manually segmented masks provided by expert radiologists as ground truth (Figure 1b). This multi-centre and multi-vendor network was trained on 441 patients and externally tested on 381 exams, reaching a dice similarity coefficient (DSC) of 0.68 (Interquartile Range (IQR): 0.57–0.78), precision of 0.55 (IQR: 0.41–0.70), and recall of 0.95 (IQR: 0.87–0.98) on the external test set (Panic et al. 2023).

2.6. Feature extraction

When extracting features, different parameters should be properly set. However, there is a paucity of conclusive evidence in the literature to guide the choice of these parameters. Concerning MRI, it is unclear whether the proper number of bins to discretise will produce more stable and robust results, given that the intensity values of the tumoural volumes (of both normalised and non-normalised images) range widely across centres. Indeed, PyRadiomics, an IBSI-compliant open-source platform (Defeudis et al. 2022) suggests that the *fixedBinCount* parameter should range between 30 and 130 (Breiding 2014; Zwanenburg et al. 2020); however, this choice is strictly dependent on the signal intensities of the structure we want to analyse. Considering the above issues and our intensity tumoural values, we evaluated the impact of 16, 32, and 64 bins.

To avoid interpolated isotropic voxels, the feature extraction was performed in 2.5D (Zwanenburg A et al. 2016), and the distance between two neighbouring voxels was considered equal to one. Using PyRadiomics, we extracted all the group features (listed in supplementary materials 15.1) from the original, wavelet, and Laplacian-filtered images using both manual and automatic segmentation masks, obtaining a total of 665 radiomic features.

2.7. Feature normalization

Another important parameter, crucial for reducing the multicenter databases variability, is feature normalisation (Figure 1d). Following the insights obtained in previous studies (Giannini et al. 2023; Panic et al. 2023), we decided to assess the influence of normalising the features using the *z-score* and *3sigma* methods in comparison to the *no-norm* scenario. In this study, the normalisation parameters (mean and standard deviation) were calculated considering both TR and IntVAL,and then applied to the ExtTEST.

2.8. Predictive model development

2.8.1. Robust feature identification

Feature reproducibility is of key importance for developing a strong and reliable radiomics model. To identify the robust features, we evaluated the intraclass correlation coefficient (ICC), considering the automatic segmentations which yielded a DSC higher than 0.75, thus emulating the variability between two readers. Referring to the Koo&Li guidelines (Koo and Li 2016), we set the ICC threshold at 0.90, to specifically identify an *excellent* feature reliability (Koo and Li 2016). We examine the behaviour of features within each group (shape, first-order, and texture) for both original and filtered images in terms of their robustness, using ICC to evaluate whether certain feature groups are more reliable than others. We assessed the impact on the feature reproducibility by considering all combinations of image normalisation, number of bins, and feature normalisation. The combination that provided the highest number was selected for the development of the classifiers.

2.8.2. Feature selection and model development

Once the most robust features were selected, different feature selection (FS) approaches were implemented and compared to identify the most suitable subgroup for the three classification aims. For *multivariate FS approaches*, we analysed: a) minimum redundancy maximum relevance (mRMR) and b)

affinity propagation (AP) (Defeudis et al. 2022). While as *Ranking approach*, we customised three FS approaches, by linking the Spearman correlation and the Area Under the ROC Curve (AUC) values related to the outcome: a) *Corr&Corr* in which we excluded the redundant feature, defined as highly correlated ones by Kuhn et al. (Kuhn and Johnson 0000) (threshold = 0.90), and the least correlated ones with the outcome, with a correlation lower than the 25^{th} percentile of the correlation of all features; b) *Corr&AUC* in which similarly we excluded the redundant features and those with AUC < 0.55, to exclude any AUC data that could result from random chance; c) *Corr&AUC_Sequential* in which we ordered the features by AUC values selected by the *Corr&AUC* method and inserted them into the ML model one at a time, evaluating the overfitting as the point where the AUCs between the TR and the IntVAL strongly differed, selecting the best suitable number of input features. All these feature selection methods were applied in combination with the ML models and compared with no-feature selection (NoFS).

All these feature selection (FS) methods were combined with the following machine learning (ML) algorithms:

a) Logistic linear regression (LR) with a normal distribution, b) Bayesian classifier, and c) support vector machine (SVM) utilising both linear (SVM_l) and Gaussian (SVM_g) kernel functions, with a box constraint (C) set to 1. Additionally, we evaluated the impact on the embedded systems (ES), which automatically apply FS and train the ML algorithms. The performance of the following models was assessed: d) least absolute shrinkage and selection operator (LASSO) with alpha = 1 for L1 optimisation, and cross-validation (CV) to estimate the mean squared error at a value of 3, e) ElasticNet, which combines LASSO feature elimination and coefficient reduction; we set alpha = 0.5 for ElasticNet optimisation, f) random forest (RF), with the default number of trees set to 100 ($n = 100$) to ensure a balance between performance and computational efficiency, g) ensemble learning, analysing two different adaptive boosting algorithms: AdaBoost (EL_AdaBoost) and GentleBoost (EL_GentleBoost), with 100 learning cycles and tree-based weak learners, and h) stepwise regression, using both linear (SW_l) and Gaussian (SW_g) distributions, with the deviance criterion for adding/removing terms and *PEnter* and *PRemove* values set to 0.05 and 0.1, respectively.

2.9. Multiparametric approach

Since diffusion-weighted sequences and their ADCs are frequently acquired at the diagnostic level, we sought to assess the potential of a multiparametric approach including both ADC and T2w features. Before extracting the features from ADC maps, we applied an elastic registration to align these images to the T2w, followed by the same image resampling and normalisation. Then radiomics features were extracted with the same parameters used for the T2w images, and the robust subset of features was identified through the ICC. To evaluate the impact of adding characteristics of ADC, we re-trained the combination of FS and ML models that yielded the best results on the IntVAL in the mono-parametric approach for each classification aim.

Since ADC might suffer from strong artefacts and low quality, to evaluate the impact of image quality on classification performance, a centralised expert radiologist provided a quality score (ranging between 1: bad and 3: good) of the ADC maps of ExtTEST.

2.10. Statistical analysis

Balanced accuracy (BalAcc), AUC, sensitivity (sens), specificity (spec), positive and negative predictive values (PPV and NPV) of the predictive system were evaluated on TR, IntVAL, and ExtTEST. Only those models that achieved an AUC ≥ 0.60 in the TR were considered for further validation analyses. The best-performing models were defined as those with AUC > 0.60 in IntVAL and the smallest difference between sens and spec. To statistically compare the ICC distributions between the feature groups we first applied the Kruskal-Wallis test for the multiple comparisons, and then the Mann-Whiteny for the pairwise ones. To compare the differences between the validation sets, we performed the chi-squared (comparison of proportions) analysis. Statistical significance was established at the *pvalue* < 0.05. All analyses were performed using Python 3.7, Matlab (R2023a), and MedCalc Software Ltd.

3. Results and discussion

3.1. Dataset

A total of 245 patients were collected: 180 were used to train (n = 122, TR) and internally validate (n = 58, IntVAL) the algorithms, and 65 to externally validate them (ExtTEST). Figure 2 shows the composition of the dataset according to vendors and reference standard. No statistical differences between training and validation sets for each aim for R+ and *R*- patients were found (*pvalue* < 0.01).

3.2. Radiomic feature processing

Both *z-score* and *3sigma* allow to increase the number of robust features regardless of the number of bins (Table 2). Moreover, the robust features were consistent among the different combinations, showing the percentage of shared features ranging between 98% and 100% (Table Supp 2). Finally, we identified 16 bin and *z-score* normalisation as the configuration providing the highest number of robust features (*n* = 39). Among them, 23 were first-order features, 11 from GLCM, 2 from GLRLM and 3 from GLSZM. Of those, 1 was obtained from the original image, while the remaining 38 were extracted from the 4 wavelet deconvolutions. Therefore, we confirmed, as other studies (Giannini et al. 2023; Panic et al. 2023), that features extracted from filtered images are more robust. The full list of robust features is reported in Table Supp 3.

The boxplot graph in Figure 3 illustrates the distribution of ICC values across three feature groups (shape, first-order, and texture) while comparing original and filtered images (wavelet, and Laplacian). Shape features exhibit a significantly lower ICC compared to other feature types. This

Table 2. Number of robust features for each combination of image normalisation, number of bins, and feature normalisation. The highest number of robust features is in bold.

Number of bins								
16 bins			32 bins			64 bins		
Feature normalisation								
no-norm	z-score	3sigma	no-norm	z-score	3sigma	no-norm	z-score	3sigma
32	**39**	38	29	36	35	32	38	36

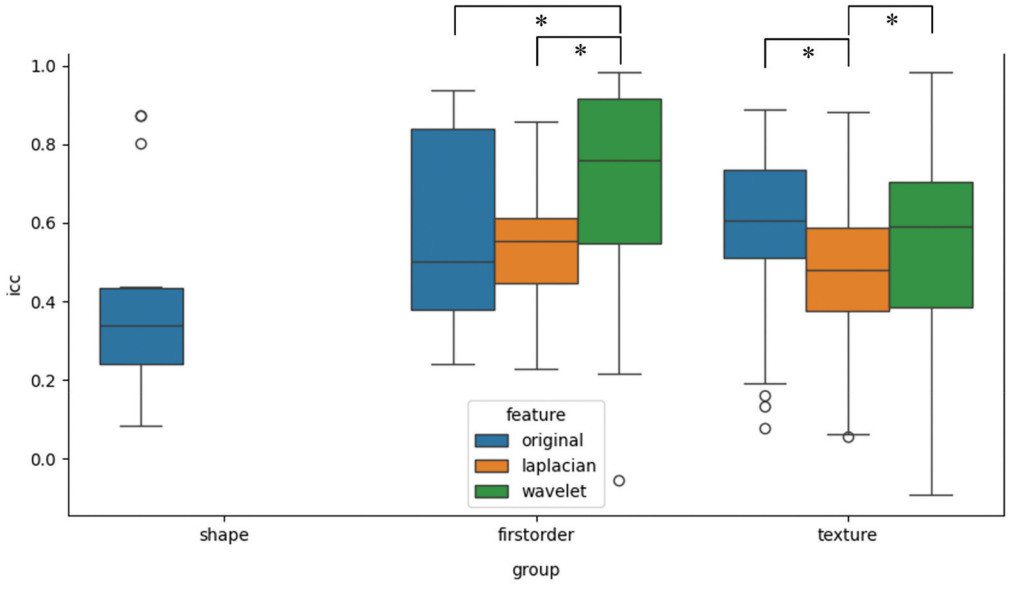


Figure 3. Box plots representing the ICC distributions of the different feature groups, divided between those extracted from original images, and filters (Laplacian, and wavelet). The asterisks indicate statistically significance (*p*value < 0.05).

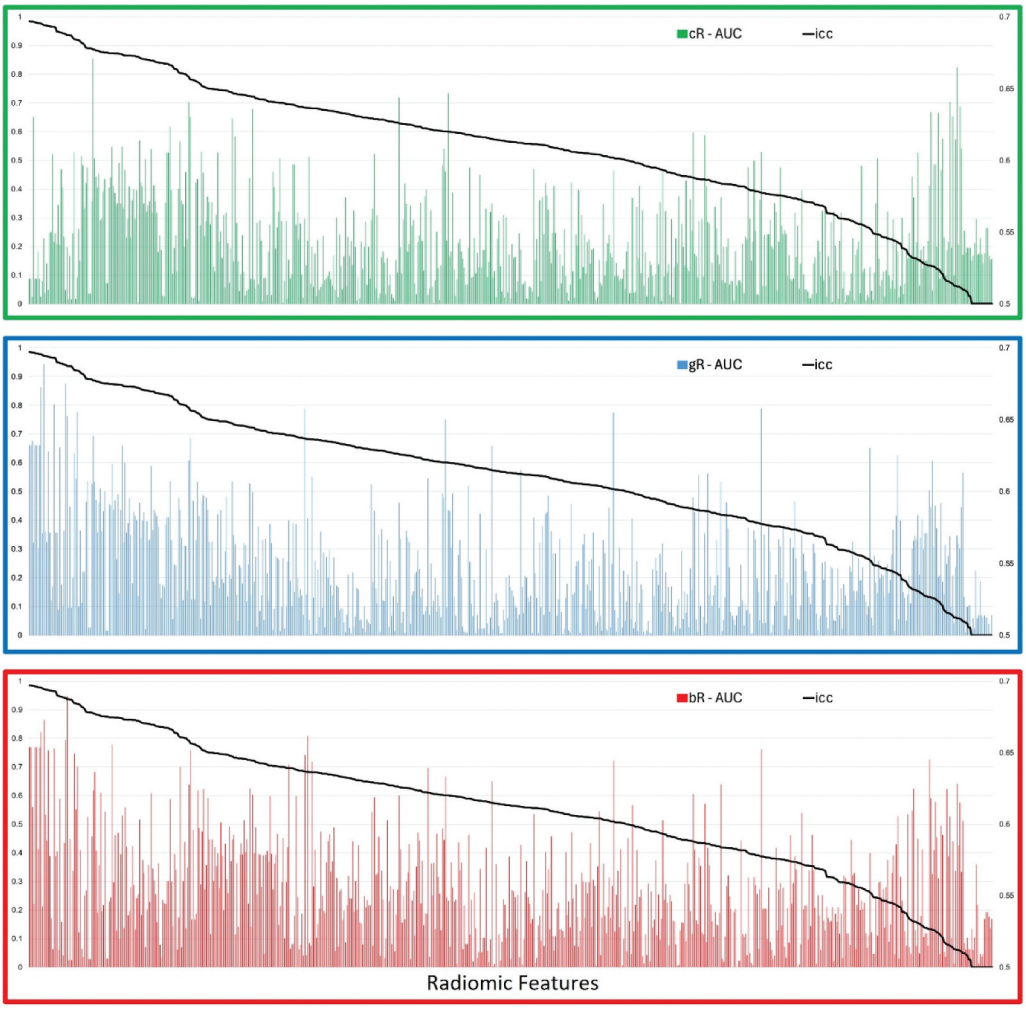


Figure 4. Relationship between ICC and AUC for each feature for the three clinical aims. The secondary axis refers to AUC values.

is expected, as we are computing ICC between manual and automatic segmentations with a Dice overlap > 0.75, which implies that the segmentations are not perfectly overlapping. Since shape features are highly dependent on segmentation accuracy, they are the most affected by these variations. A similar effect has been reported in previous studies (Zwanenburg et al. 2020). The Laplacian filter presented the lower ICC values for both first-order and texture features, suggesting that this filtering method may introduce additional variability in these groups. Conversely, wavelet decompositions led to improved ICC values. Notably, in this group, the ICC of first-order features was significantly higher than that obtained from the original images ($p < 0.05$). These differences suggest that wavelet decompositions may enhance feature stability, offering greater robustness and reliability compared to Laplacian and original features, which tend to be less consistent across repeated measures.

Additionally, in Figure 4 we observed that ICC values were not correlated to AUC for any clinical aims, meaning that a high number of relevant features (i.e. more correlated with the output) were discarded due to their low ICC. This is particularly evident for cR (green bars), and it might impact the overall performance of the radiomic model.

These insights may suggest the need to explore alternative approaches for the selection of reliable and predictive features.

Table 3. Total number of features selected by each features selection algorithm for all clinical aims.

Features Selection		#features cR	#features gR	#features bR
NoFS		39	39	39
mRMR		10	9	12
AP		3	3	3
Corr&Corr		6	7	9
Corr&AUC		21	21	18
Corr&AUC_Sequential	*Bayesian*	21	4	6
	LR	21	3	6
	SVM_l	20	7	16
	SVM_g	3	12	4
Embedded Systems	*RF*	7	7	7
	AdaBoost	39	39	39
	GentleBoost	39	39	39
	SW_l	9	19	15

3.3. Features selection

Table 3 shows the number of features selected by each FS algorithm for each clinical aim. In general, the number of features differs considerably between FS methods, whereas it remains relatively consistent across clinical aims. AdaBoost and GentleBoost algorithms do not discard any robust features, i.e. the number of features is the same as noFS, making these methods vulnerable to potential overfitting (Figure 5). AP is the method that selects the lowest number of features ($n = 3$) for all clinical aims. In contrast, the biggest differences in number of features between clinical aims are observed with *Corr&AUC_Sequential*, where in some cases were selected three times the number of features for cR and bR. FS strongly impacts the training of the models for all clinical aims (Figure 5). ES methods reached the highest performances in the training set for all clinical aims regardless of the classifier used, even those trained with a strongly unbalanced dataset (cR and bR). Although these methods obtain consistent results in IntVAL, they are not as capable of generalising on ExtTEST, which means they may suffer from overfitting more than other methods. However, even if two of them used the same number of features as NoFS, ES algorithms were less prone to overfitting as they reached higher results in both validation sets. AP leads to the most variable results depending on the classifier that was used for both cR and gR, however, most of them obtained poor performances in all datasets (Table Supp 5), including the TR, probably due to the meagre number of features selected (Table 3). Except for bR, *Corr&Corr* reaches stable results between different datasets regardless of the classifier used. In particular, for gR, this method maintains good results in both IntVAL and ExtTEST.

3.4. Model development

Most models (20, 27, and 25, respectively, for cR, gR, and bR) exceed the established performance thresholds (AUC ≥0.6) within the TR set. Figure 6 shows the sens, spec, and AUC values in TR, IntVAL, and ExtTESTfor all models and clinical aims. In general, the impact of FS on performance was significantly greater than the type of classifier. This could highlight the importance of features over the complexity of the classifier itself, as confirmed by other studies (Santinha et al. 2024; Mylona et al. 2024).

More consistent results were obtained when patients were dichotomised between pTRG 1 and 2 vs pTRG > 2, i.e. prediction of good response. Indeed, in this setting some models, i.e. SVM_l, SVM_g, and Bayesian classifiers obtained AUC > 0.60 in both IntVAL and ExtTEST, regardless of FS used. SVM_l and Bayesian were the best performing for the other two clinical aims, particularly when combined with FS methods based on correlation. It is important to note that while the AUC values achieved by the LR classifier were comparable to those of other classifiers, they resulted in a significant imbalance between sens and spec, particularly for the clinical aims affected by a high-class imbalance, namely cR, and bR. Embedded classifiers are prone to overfitting, as evidenced by the fact that none of the employed methods were able to maintain the high-performance levels observed in the training set when tested on the two external validation sets. It is noteworthy that our results indicated that the performance of the classifiers trained to distinguish bR did not reach satisfactory results, primarily due to the very low number of cases in the *R-* class within the training set (n = 8 vs 88 in R+). In general, the performance for the automatic cR prediction are not strongly affected by the different ML models

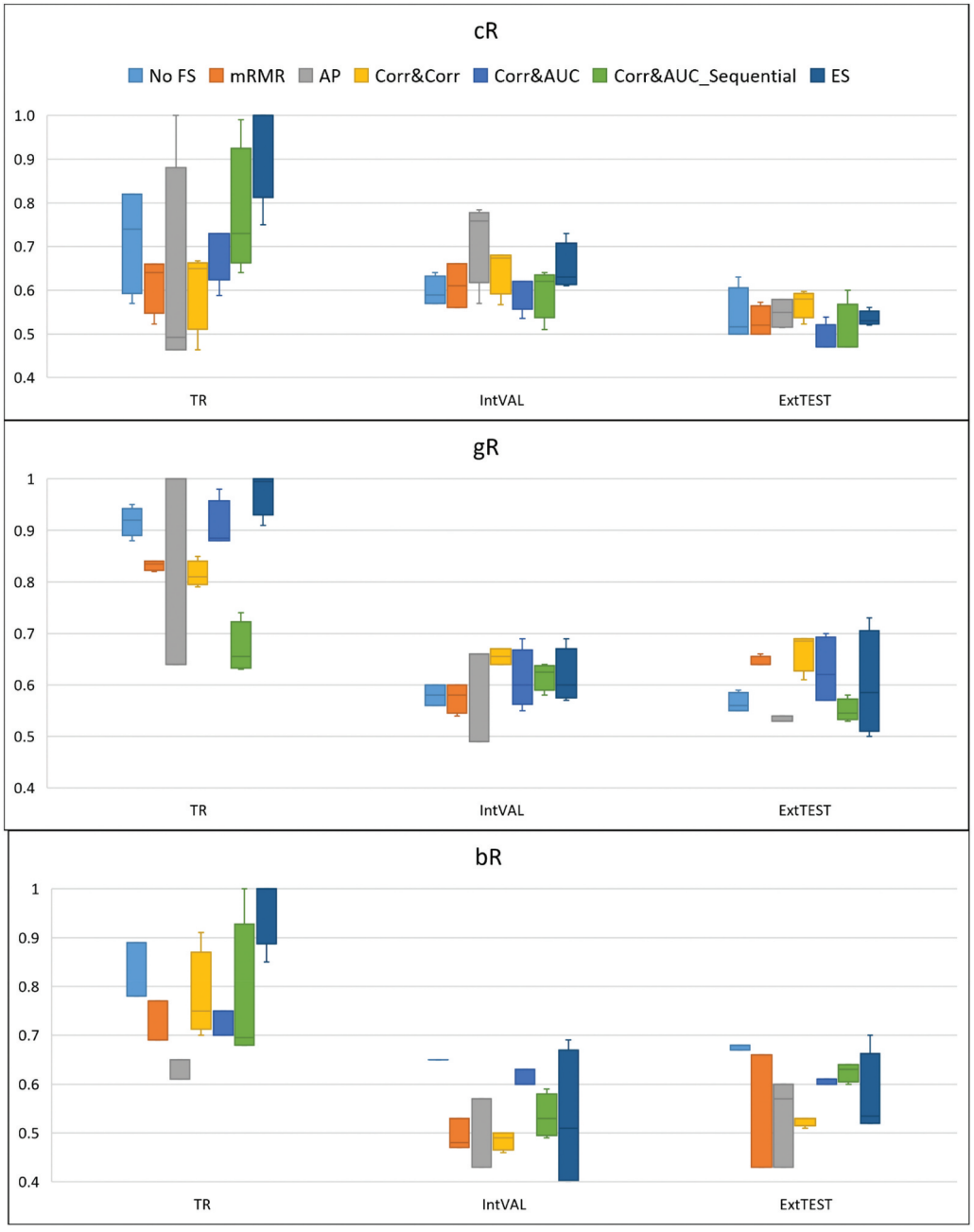


Figure 5. Results of the models grouped for features selection methods and clinical aim.

developed. None of the models trained show both sens and spec higher than 0.6 in both IntVAL and ExtTEST, despite the higher AUC values. This could be due to the high imbalance between the *R*- and R+ classes (8vs50 in IntVAL and 8vs57 in ExtTEST). Indeed, when considering classifiers trained to detect gR, we observed that their performance was higher, probably due to the lower imbalance between the R+ and *R*- classes. (25vs30 and 24vs39 for IntVAL and ExtTESTrespectively). Results obtained by all classifiers for all clinical aims are reported in Table Supp 5, while their ROC curves in Figure Supp 3.

Considering the aforementioned facts, the most successful models that were selected for the multi-parametric approach were:

(1) cR: Corr&Corr+Bayesian, that showed an AUC of 0.65, 0.68 and 0.58, sens 0.65, 0.56 and 0.73, spec 0.53, 0.11 and 0.44 in the training IntVAL and ExtTESTdatasets, respectively.

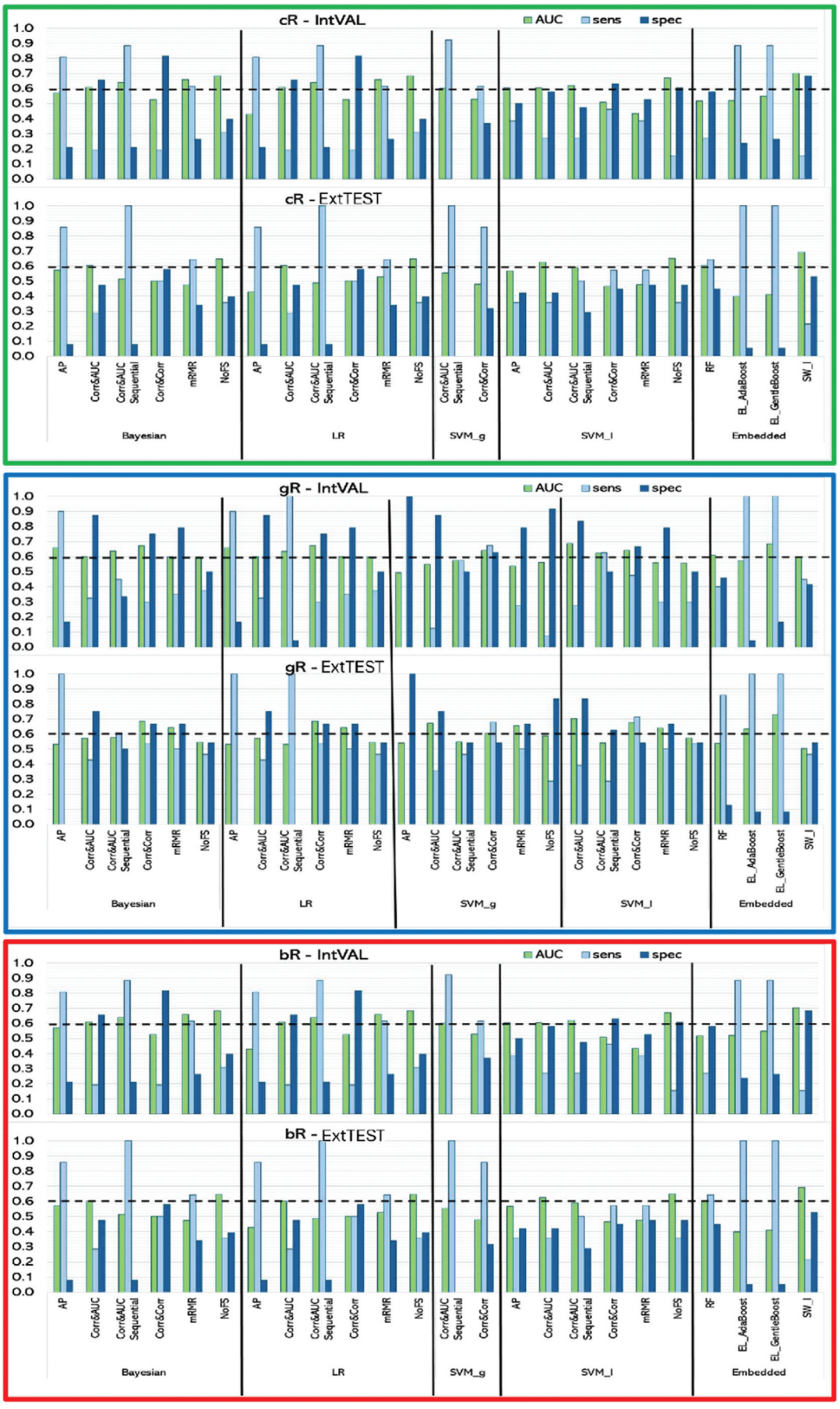


Figure 6. Results in terms of sensitivity (sens), specificity (spec) and Area Under the ROC curve (AUC) on IntVAL and ExtTEST.

(2) gR: Corr&Corr+SVM_g, which yielded an AUC of 0.85, 0.64, and 0.61, respectively, in the training set and in ExtTESTand IntVAL, with a good balance between sens and spec in both IntVAL (0.68 and 0.62), ExtTEST(0.68 and 0.54) and in the training set (0.81 and 0.76).
(3) bR: Corr&Corr + SVM_l, because it presented the better balance between performance on IntVAL and ExtTEST, the discrepancies between sets are reduced, even if lower than 0.6 in terms of BalAcc. The model selected yielded AUC of 0.70, 0.46, and 0.51, sens equal to 0.80, 0.57, and 0.46, and spec of 0.55, 0.45, and 0.63 in the training IntVAL and ExtTESTdatasets, respectively.

3.4.1. Multiparametric approach

The robust features obtained combining both T2w and ADC sequences were 100:39 from T2w and 61 from ADC. The combinations chosen were as FS the *Corr&Corr* approach and as ML algorithm the Bayesian for cR (n selected features = 43), the SMV_g for gR ($n = 55$), and the SVM_l for bR ($n = 37$). In Figure 7, it is possible to notice a slight worsening of the overall performance in both IntVAL and ExtTEST, even if the trend of the metrics is similar for each aim.

The poorer results could be due to several reasons. First, we used real-world images, so imaging protocols varied among centres, leading to ADC maps generated with different b-values (e.g. b0 or b50 as the minimum and b800, b1000, or b1200 as the maximum). Although ADC is a quantitative image derived from the signal decay between two b-values, studies show that texture features in healthy tissues systematically vary with b-values selection (Becker et al. 2017). This variability could compromise the reproducibility of ADC measurements.

Also, the quality of the ADC images may have affected performance. Indeed, among the 65 ADC of the ExtTEST, 27 were classified as *poor*, 19 as *medium*, and 17 as *high* quality. Image quality seems to have worsened the results mainly for cR aim (Figure 8), where the overall number of wrongly predicted samples is 30, while the correct ones are 35 against 22 and 43 of the only T2w based approach. Most of the wrongly predicted samples ($n = 68\%$) were classified as *poor* quality, while 78% of the correctly classified samples were characterised by *high* and *medium* quality.

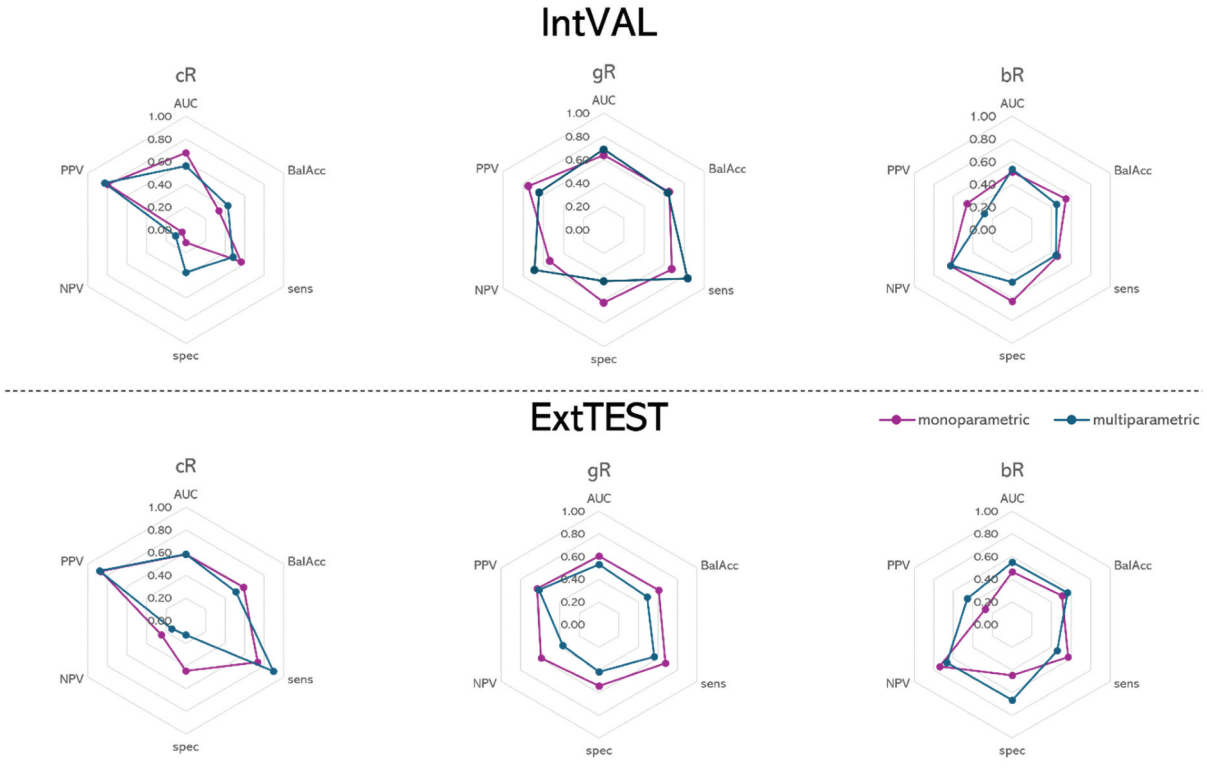


Figure 7. Comparison of radar graphs for the three aims on IntVAL and ExtTEST.

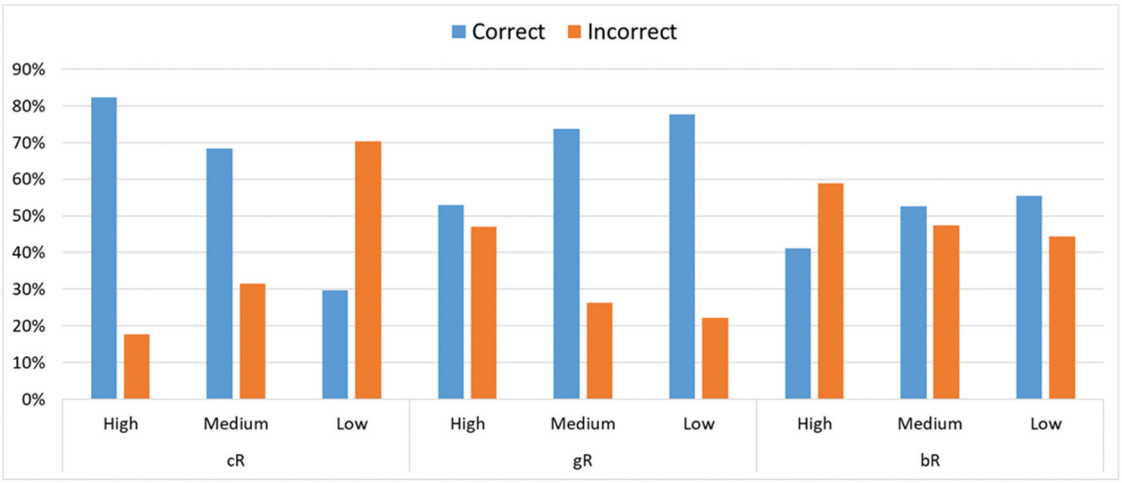


Figure 8. Percentage of correctly and wrongly classified patients for each group of quality and each clinical aim.

This behaviour is not clearly visible for the other clinical aims, but in these cases the performance was higher than cR, meaning that the algorithms, and in particular, the features used, were more robust across image quality.

We also observed that the accuracy of the segmentation has affected the performance of the gR classifier (Figure Supp 4), probably because it is the only model trained with a balanced dataset. In particular, with DSC > 0.60, the percentage of correctly classified was 78% (38/49), while with DSC < 0.60 the percentage of wrongly classified was 79% (11/14). Concerning the other two aims, no assumptions could be evaluated.

4. Conclusion

To the best of our knowledge, this is the first study to offer an analysis of the impact of different choices on the construction of an MRI-based radiomic signature for RC's therapy response prediction, which focuses not only on comparing the performance of different models, feature selection and pre-processing techniques but also on assessing the impact of each of these choices.

Several insights have been collected. Concerning the preprocessing step, we evaluated the importance of filtering the images for improving the overall robustness of the radiomic features. Meanwhile, we also proved that ICC values were not correlated to strong predictive power, highlighting the need for novel selection approaches. Indeed, it has been proven that the Feature selection methods affected the models' performances more than the ML algorithms, so it is of key importance to identify the subgroup of features which satisfy both the robustness and predictive power requirements. Despite the promising results given by the multi-parametric approaches in literature, we assessed the need to carefully manage the different medical data, suggesting a suitable preprocessing for each MRI sequence. In our case, we did not reach higher results adding the ADC sequences, probably due to the image quality, which may have affected performance as many images presented a low quality. Finally, we observed that the best results were achieved for *gR*, characterised by the lowest class imbalance. These findings underscore the need to carefully define the most representative training sets to address the inherent clinical complexities. Properly balancing the dataset is essential for improving model generalisability and avoiding overfitting, ensuring that the model is both accurate and robust across various clinical scenarios. Furthermore, since the highly unbalanced datasets are typical of clinical studies, it is crucial to pay attention to the enrolment and management of the data.

However, this study has also two main limitations. First, we decided to work with real-world data, trying to develop a vendor-agnostic model, to obtain a generalisable model. Unfortunately, it led to a high data complexity that was not entirely solved. So, switching from a vendor-agnostic approach to a vendor-specific one could be a first step towards a better understanding of some technical issues. Finally, we did not assess the impact of integrating information from different sources, i.e. clinical information (Bordron et al. 2022, Nardone et al. 2022, Song et al. 2022), genomics (O'Sullivan et al. 2023), and digital pathology (Feng et al.

2022), which have proven promising results, as shown by Feng et al. (2022) (AUC of 0.87 in the external validation set).

In conclusion, we presented a comprehensive and preliminary analysis of the impact of multiple decisional development parameters in the ML pipeline. This remains a critical topic among researchers due to the lack of standardised and commonly agreed guidelines for the development of suitable, robust and generalisable non-invasive tools for supporting clinicians in personalised medicine (Santinha et al. 2024). Our insights demonstrate that some decision-making choices have a greater impact than others; thus, radiomics alone is not sufficient to solve the complexity of this clinical task, especially when using real-world multicenter datasets. By evaluating all those steps and different strategies, this study contributes to the implementation of more reliable and interpretable AI-driven tools, that may support clinicians in decision-making and better understanding the pathology complexities.

Author contributions

CRediT: **Arianna Defeudis:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing; **Jovana Panic:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing; **Lorenzo Vassallo:** Data curation, Investigation, Resources; **Stefano Cirillo:** Data curation, Resources; **Marco Gatti:** Data curation, Resources; **Riccardo Faletti:** Data curation, Resources; **Antonio Esposito:** Data curation, Resources; **Anna Palmisano:** Data curation, Resources; **Serena Dell'Aversana:** Data curation, Resources; **Alfonso Ragozzino:** Data curation, Resources; **Salvatore Siena:** Data curation, Resources; **Angelo Vanzulli:** Data curation, Resources; **Daniele Regge:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing; **Samanta Rosati:** Conceptualization, Supervision; **Gabriella Balestra:** Conceptualization, Supervision; **Valentina Giannini:** Conceptualization, Project administration, Supervision, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The research leading to these results has received funding from AIRC under 5 per Mille 2018 - ID. 21091 program – P.I. Bardelli Alberto, G.L. Regge Daniele.

References

Becker AS, Wagner MW, Wurnig MC, Boss A. 2017. Diffusion-weighted imaging of the abdomen: impact of b-values on texture analysis features. NMR Biomed. 30(1). doi: 10.1002/nbm.3669.

Benson AB, et al. 2020. Rectal cancer, version 6.2020: featured updates to the NCCN guidelines. JNCCN J Natl Compr Cancer Netw. 18(7):807–815. doi: 10.6004/jnccn.2020.0032.

Benson AB, Venook AP, Al-Hawary MM, Azad N, Chen Y-J, Ciombor KK, Cohen S, Cooper HS, Deming D, Garrido-Laguna I, et al. 2022. Rectal cancer, version 2.2022. JNCCN J Natl Compr Cancer Netw. 20(10):1139–1167. doi: 10.6004/jnccn.2022.0051.

Bordron A, Rio E, Badic B, Miranda O, Pradier O, Hatt M, Visvikis D, Lucia F, Schick U, Bourbonne V. 2022. External validation of a radiomics model for the prediction of complete response to neoadjuvant chemoradiotherapy in rectal cancer. Cancers (Basel). 14(4):1079. doi: 10.3390/cancers14041079.

Breiding MJ. 2014. Computational radiomics system to decode the radiographic phenotype. Physiol Behav. 63(8):1–18. doi: 10.1158/0008-5472.CAN-17-0339.Computational.

Da-Ano R, Visvikis D, Hatt M. 2020. Harmonization strategies for multicenter radiomics investigations. Phys Med Biol. 65 (24):24TR02. doi: 10.1088/1361-6560/aba798.

De Dumast P, Bach Cuadra M. 2023. Domain generalization in fetal brain MRI segmentation with multi-reconstruction augmentation. Proceedings - International Symposium on Biomedical Imaging, Cartagena de Indias, Colombia, IEEE Computer Society. doi: 10.1109/ISBI53787.2023.10230402.

Defeudis A, Mazzetti S, Panic J, Micilotta M, Vassallo L, Giannetto G, Gatti M, Faletti R, Cirillo S, Regge D, et al. 2022. Mri-based radiomics to predict response in locally advanced rectal cancer: comparison of manual and automatic segmentation on external validation in a multicentre study. Eur Radiol Exp. 6(1). doi: 10.1186/s41747-022-00272-2.

Feng L, et al. 2022. Development and validation of a radiopathomics model to predict pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: a multicentre observational study. www.thelancet.com/.

Giannini V, Panic J, Regge D, Balestra G, Rosati S. 2023. Could normalization improve robustness of abdominal MRI radiomic features? Biomed Phys Eng Express. 9(5):055002. doi: 10.1088/2057-1976/ace4ce.

Graves CV, Moreno RA, Rebelo MS, Nomura CH, Gutierrez MA. 2020. Improving the generalization of deep learning methods to segment the left ventricle in short axis MR images. 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) - July 20-24, 2020 via the EMBS Virtual Academy. IEEE doi: 10.1109/EMBC44109.2020.9175256.

He X, Liu X, Zuo F, Shi H, Jing J. 2023. Artificial intelligence-based multi-omics analysis fuels cancer precision medicine. Academic Press. 88:187–200. doi: 10.1016/j.semcancer.2022.12.009.

Koo TK, Li MY. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 15(2):155–163. doi: 10.1016/j.jcm.2016.02.012 .

Kuhn M, Johnson K. Applied predictive modeling.

Lipkova J, Chen RJ, Chen B, Lu MY, Barbieri M, Shao D, Vaidya AJ, Chen C, Zhuang L, Williamson DFK, et al. 2022. Artificial intelligence for multimodal data integration in oncology. Cell Press. 40(10):1095–1110. doi: 10.1016/j.ccell.2022.09.012.

Mali SA, Ibrahim A, Woodruff HC, Andrearczyk V, Müller H, Primakov S, Salahuddin Z, Chatterjee A, Lambin P. 2021. Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods. MDPI. 11(9):842. doi: 10.3390/jpm11090842.

Mandard A-M, Dalibard F, Mandard J-C, Marnay J, Michel Henry-Amar M, Petiot J-F, Roussel A, Jacob J-H, Segol P, Samama G, Ollivier J-M, Bonvalot S, Gignoux M. 1994. Pathologic assessment of tumor regression after preoperative chemoradiotherapy of esophageal carcinoma. Clinicopathologic correlations. Cancer. 73(11):2680–2686. doi: 10.1002/1097-0142(19940601)73:11<2680:AID-CNCR2820731105>3.0.CO;2-C.

McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GS, Darzi A, et al. 2020. International evaluation of an AI system for breast cancer screening. Nature. 577(7788):89–94. doi: 10.1038/s41586-019-1799-6.

Mylona E, Zaridis DI, Kalantzopoulos CN, Tachos NS, Regge D, Papanikolaou N, Tsiknakis M, Marias K, Mylona E, Zaridis D, et al. 2024. Optimizing radiomics for prostate cancer diagnosis: feature selection strategies, machine learning classifiers, and MRI sequences. Insights Imaging. 15(1). doi: 10.1186/s13244-024-01783-9.

Nardone V, Reginelli A, Grassi R, Vacca G, Giacobbe G, Angrisani A, Clemente A, Danti G, Correale P, Carbone SF, et al. 2022. Ability of delta radiomics to predict a complete pathological response in patients with loco-regional rectal cancer addressed to neoadjuvant chemo-radiation and surgery. Cancers (Basel). 14(12):3004. doi: 10.3390/cancers14123004.

O'Sullivan NJ, Hugo CT, Michelle HT, Alison C, Brian JM, John OL, Paul HM, Dara OK, FMJ Meaney, Michael EK. 2023. Radiogenomics: contemporary applications in the management of rectal cancer. Multidiscip Digit Publishing Inst (MDPI). doi: 10.3390/cancers15245816.

Panic J, Balestra G, Defeudis A, Rosati S, Regge D, Giannini V. 2023. IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE). Comparison between different approaches for the creation of the training set: how clustering and dimensionality impact the performance of a deep learning model". 10.1109/BIBE60311.2023.00070.

Panic J, Defeudis A, Balestra G, Giannini V, Rosati S. 2023. Normalization strategies in multi-center radiomics abdominal MRI: systematic review and meta-analyses. IEEE Open J Eng Med Biol. 4:67–76. doi: 10.1109/OJEMB.2023.3271455.

Ribeiro MAO, Nunesy FLS. 2021. Evaluating the pre-processing impact on the generalization of deep learning networks for left ventricle segmentation. Proceedings - 2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021, Institute of Electrical and Electronics Engineers Inc. 3505–3512. doi: 10.1109/BIBM52615.2021.9669630.

Santinha J, Pinto dos Santos D, Laqua F, Visser JJ, Groot Lipman KBW, Dietzel M, Klontzas ME, Cuocolo R, Gitto S, Akinci D'Antonoli T. 2024. Esr essentials: radiomics—practice recommendations by the European Society of Medical Imaging Informatics. Springer Sci Bus Media Dtschl GmbH. 35(3):1122–1132. doi: 10.1007/s00330-024-11093-9.

Scalco E, Rizzo G. 2017. Texture analysis of medical images for radiotherapy applications. Br J Radiol. 90(1070). doi: 10.1259/bjr.20160642.

Siegel RL, Miller KD, Wagle NS, Jemal A. 2023. Cancer statistics, 2023. CA Cancer J Clin. 73(1):17–48. doi: 10.3322/caac.21763.

Song M, Li S, Wang H, Hu K, Wang F, Teng H, Wang Z, Liu J, Jia AY, Cai Y, et al. 2022. Mri radiomics independent of clinical baseline characteristics and neoadjuvant treatment modalities predicts response to neoadjuvant therapy in rectal cancer. Br J Cancer. 127(2):249–257. doi: 10.1038/s41416-022-01786-7.

Wang CW, Khalil MA, Firdi NP. 2022. A survey on deep learning for precision oncology. MDPI. 12(6):1489. doi: 10.3390/diagnostics12061489.

Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, Ashrafinia S, Bakas S, Beukinga RJ, Boellaard R, et al. 2020. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology. 295(2):328–338. doi: 10.1148/radiol.2020191145.

Zwanenburg A LS, Leger S, Vallières M. 2016. Image biomarker standardization initiative. Arxiv preprint. ArXiv:1612.07003, 10.17195/candat.2016.08.1.