Impact of MRI Deep Learning reconstruction on radiomic feature stability and discriminative power: a phantom study

Gabriel Ballout[1], Navid Rabbani[2,3], Bruno Pereira [4], Lucie Cassagnes[1,2], Pascal Chabrot[1,2], Louis Boyer[1,2], Jean-Marc Garcier[1,2], Eleonore Pouget[1], Adrien Bartoli[2,3], Benoit Magnin[1,2,3]*

1 Radiology Department, Clermont-Ferrand University Hospital, Clermont-Ferrand, France

2 Institut Pascal, UMR 6602 CNRS, Université Clermont Auvergne, Clermont-Ferrand, France

3 DIA2M, DRCI, Clermont-Ferrand University Hospital, Clermont-Ferrand, France

4 Biostatistics Unit, DRCI, Clermont-Ferrand University Hospital, Clermont-Ferrand, France

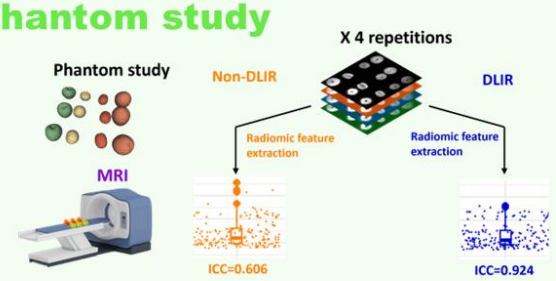
* corresponding author

Benoît Magnin
Radiology Department CHU Estaing
1 place Lucie Aubrac 63100 Clermont-Ferrand
bmagnin@chu-clermontferrand.fr

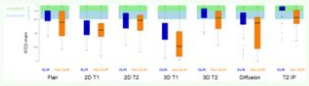
Impact of MRI Deep Learning reconstruction on radiomic feature stability and discriminative power: a phantom study

EUROPEAN SOCIETY OF RADIOLOGY

- DLIR significantly improves MRI radiomic feature reproducibility and repeatability over non-DLIR.
- DLIR shows potential to enhance the discriminative capabilities of radiomic features
- Results from this single-vendor phantom study warrant future clinical validation



Stability assessement



Discriminative power assessement

Classification accuracy

	Mean	SD	p-value	
DLIR	0.688	0.337	0.567	PyRadiomics
non-DLIR	0.667	0.317		
DLIR	0.840	0.161	0.015	MIRP
non-DLIR	0.797	0.170		

Deep Learning-based reconstruction can improve radiomic stability in MRI studies.

European Radiology EXPERIMENTAL

Eur Radiol Exp (2026), Ballout et al. DOI: XXX

Impact of MRI Deep Learning reconstruction on radiomic feature stability and discriminative power: a phantom study

Abstract

Background

Radiomics depends on image acquisition and reconstruction. While MRI Deep-Learning image reconstructions (DLIR) reduce noise and enhance image quality, their impact on radiomic features remains underexplored. This study compares the reproducibility, repeatability and discriminative power of MRI radiomic features between traditional non-DLIR and DLIR.

Methods

An MRI protocol comprising seven sequences was conducted using 12 pieces of fruit as phantom models. Each sequence was acquired with four distinct protocols, repeated four times. Images were reconstructed using both non-DLIR and DLIR. After 3D semi-automated segmentation of each piece of fruit, 93 radiomic features were extracted with PyRadiomics and 152 with MIRP. Features were defined as reproducible if the intraclass correlation coefficient (ICC(2,1)) exceeded 0.9 and repeatable if the overall concordance correlation coefficient (OCCC) exceeded 0.9. A classification task, assessing the type of fruit from the radiomic features, was performed to compare the discriminative power of the reconstructions.

Results

DLIR yielded 19.51% and 34.10% reproducible features, compared to 9.06% and 30.74% with non-DLIR, i.e., 10.45% and 3.36% improvement, for PyRadiomics and MIRP respectively ($p<0.01$). The difference in repeatability between reconstructions was 8.45% in favor of DLIR (47.47% for DLIR and 39.02% for non-DLIR) for PyRadiomics and 2.81% in favor of DLIR for MIRP (46.57% for DLIR and

43.76% for non-DLIR) ($p<0.01$). The classification accuracies were higher using DLIR features for MIRP ($p=0.015$).

Conclusions

In our single-vendor fruit phantom study, DLIR improves radiomic feature reproducibility and repeatability and suggests potential gains in discriminative power, pointing towards its suitability for MRI radiomic studies.

Key points

- DLIR significantly improves MRI radiomic feature reproducibility and repeatability over non-DLIR.

- DLIR shows potential to enhance the discriminative capabilities of radiomic features.

- Results from this single-vendor phantom study warrant future clinical validation.

Critical Relevance Statement

MRI deep learning reconstruction demonstrates improved radiomic feature reproducibility and repeatability over conventional methods. These findings, along with potential gains in discriminative power, point towards its suitability for future MRI radiomic applications.

Keywords

radiomics reproducibility; Deep Learning reconstruction; MRI harmonization; phantom study

Abbreviations

CNR: Contrast-to-Noise Ratio

DLIR: Deep Learning-based image reconstruction

GLCM: gray-level co-occurrence matrix

GLDM: gray-level dependence matrix

GLDZM: gray-level distance zone matrix

GLRLM: gray-level run-length matrix

GLSZM: gray-level size zone matrix

LOFO: Leave-One-Fruit-Out

LoG: Laplacian of Gaussian

OCCC: Overall concordance correlation coefficient

IBSI: Image Biomarker Standardization Initiative

ICC: Intraclass correlation coefficient

NGLDM: neighbouring gray-level dependence matrix

NGTDM: neighborhood gray-tone difference matrix

SNR: Signal-to-Noise Ratio

Background

Radiomics aims at extracting quantitative information, known as radiomic features, from medical images [1]. Radiomic features can be used either alone or in combination with other data (clinical, biological, genetic) for clinical purposes, such as diagnosis, treatment response or prognostic

prediction. Numerous studies aimed to produce a predictive model for a given pathology, based on radiomic parameters extracted from a training population [2, 3].

However, despite the growing number of publications on radiomics, its practical application remains limited [4], due to the difficulty in reproducing results. One factor contributing to this lack of reproducibility is the complexity of the radiomics study methodology. To address this, efforts have been made toward standardizing the extraction methodologies, in particular through the Image Biomarker Standardization Initiative (IBSI) [5].

Methodological studies have highlighted the factors influencing the stability of radiomic parameters in CT scans [6], in PET imaging, and to a lesser extent in MRI. It was shown that the stability of MRI-derived radiomic features is influenced by the scanner and its manufacturer [7, 8], the acquisition parameters [8, 9], image post-processing [10, 11] and the observer [10].

Deep Learning-based image reconstruction (DLIR), which is becoming increasingly prevalent in both CT and MRI, offers superior noise reduction, shorter scans, and improved lesion detection [12–19], though performance varies by sequence and training data. To date, the understanding of how DLIR algorithms impact radiomic analyses remains incomplete. Existing literature has consistently established that DLIR significantly alters the values of extracted radiomic features when compared to traditional reconstruction techniques, a phenomenon documented in both CT [20–23] and MRI [24–26]. In the context of CT, studies have demonstrated that despite these feature alterations, DLIR actually improves the overall stability and reproducibility of the radiomic data [20–23]. However, whether this beneficial effect on feature robustness translates to MRI remains currently underexplored. Finally, the downstream clinical implications, specifically how DLIR-derived radiomics affect the diagnostic or prognostic performance of predictive models, have yet to be thoroughly investigated. Because DLIR algorithms are non-linear and proprietary, their impact on biomarker robustness across acquisition settings is difficult to predict. Addressing this uncertainty is a critical clinical necessity. Feature stability is indeed a fundamental prerequisite for intended

clinical use cases, such as large-scale multicenter model development, reliable longitudinal patient monitoring, and protocol harmonization across diverse imaging networks. Without first establishing whether DLIR improves or degrades MRI feature reproducibility, the generalizability and clinical reliability of any downstream predictive radiomic model remain fundamentally questionable.

The objective of our study was to evaluate the influence of DLIR on the stability and discriminative power of radiomic features in MRI. More specifically, we want to assess the stability by evaluating the repeatability (comparison under constant condition) and reproducibility (comparison under varying conditions) of radiomic features using DLIR compared to conventional non-DLIR. Our secondary objectives were to evaluate the impact of DLIR on their discriminative power and to quantify the alteration of radiomic features induced by DLIR.

Methods

The design of the study is shown in Figure 1.

No institutional review board was required as this is a phantom study.

Phantom

We chose to use a phantom composed of pieces of fruit, as organic components are known to have features close to human tissues and lesions; we selected 3 oranges, 3 kiwis, 3 onions, and 3 apples, as was also proposed by several teams [11, 27].

Image acquisition

All images were acquired on a 1.5T Signa Artist MRI (GE Healthcare). We performed an MRI protocol consisting of 7 sequences commonly used during routine scans: 2D T1, 2D T2, 3D T1, 3D T2, 2D FLAIR, Diffusion, T2 In Phase.

In order to introduce variability and simulate inter-scanner differences, each sequence was acquired using a reference protocol and then repeated with three modified protocols. Each of these protocols involved minor variations to the acquisition settings, targeting changes in spatial resolution, slice thickness, or TE, while acknowledging that altering these primary targets simultaneously affects other related imaging parameters. The acquisition parameters are given in Table 1.

The images were acquired twice consecutively; the phantom was then removed and repositioned, and another series of 2 acquisitions were performed (scan-rescan-reposition-scan-rescan sequence). This resulted in 4 repeated acquisitions, each composed of 7 sequences, each sequence acquired with 4 parameter variations. In total we obtained 16 different acquisitions by sequence. All images were reconstructed using both traditional reconstruction (non-DLIR) and a vendor-provided DLIR algorithm (AIR Recon DL, GE Healthcare). Briefly, this algorithm includes a deep CNN that takes raw k-space data as its input and generates high fidelity images as its output [28] and has no user-adjustable settings.

Segmentation

A 3D semi-automatic segmentation was performed by a radiology resident (G.B.) of each piece of fruit on every sequence of the different protocols using the open-source software 3D Slicer (https://www.slicer.org) [29]. This segmentation was performed using the level tracing function, taking advantage of the high signal gradient between each piece of fruit and the adjacent void. A visual assessment was conducted for each slice, with manual correction applied where necessary. The segmentation was realized on the DLIR images, the mask was then propagated between paired reconstructions. Images and segmentations examples are provided in Figure 2.

Radiomic feature extraction

A first radiomic feature extraction was performed with PyRadiomics [30] (version 3.0.1) for each piece of fruit. N4 bias field correction was applied, intensity image normalization was performed using the intensity-normalization package in Python 3.11, a 1x1x1 mm resampling was performed with a sitkBSpline interpolator, a fixed bin width of 25 was used. A total of 93 3D radiomic features were extracted including 18 first-order histogram features and 75 second-order or higher-order radiomic features (24 from the gray-level co-occurrence matrix (GLCM), 16 from the gray-level run-length matrix (GLRLM), 16 from the gray-level size zone matrix (GLSZM), 14 from the gray-level dependence matrix (GLDM) and 5 from the neighborhood gray-tone difference matrix (NGTDM)), all will be referred as "higher-order" features in contrast to first-order). PyRadiomics was used as it is one of the most commonly used radiomic platforms. It suffers however from a lack of reproducibility [31].

A second radiomic features extraction was performed with MIRP [32], a fully IBSI compliant platform. N4 bias field correction was applied, intensity image normalization was performed with a standardization method, a fixed bin number of 32 was used. A total of 152 3D radiomic features were extracted including 57 first-order histogram features (2 local intensity features, 18 intensity-based statistical features, 23 intensity histogram features and 14 intensity-volume histogram features, all will be referred to as "first-order" features) and 95 second-order or higher-order radiomic features (25 GLCM, 16 GLRLM, 16 GLSZM, 16 from the gray-level distance zone matrix (GLDZM), 5 NGTDM and 16 from neighbouring gray-level dependence matrix (NGLDM)), all will be referred to as "higher-order" features in contrast to first-order).

For both radiomic platforms, a second extraction was made applying a Laplacian of Gaussian (LoG) filter (sigma =1).

Shape features from PyRadiomics and morphological features from MIRP were not analyzed as they reflect mostly the segmentation and not the image signal.

Discriminative power

The discriminative power of the reconstructions was evaluated by training a Random Forest classifier to predict fruit type based on the extracted radiomic features. To prevent data leakage and address potential optimistic bias related to both specific physical specimens and scanner conditions, a dual-level validation strategy was implemented. This strategy combined a Leave-One-Fruit-Out (LOFO) cross-validation with an independent hold-out acquisition protocol (designated as the "reference" protocol). During the iterative LOFO process, for each unique fruit specimen, the classifier was trained on data from all remaining specimens using only the non-held-out acquisition protocols. The trained model was subsequently tested on the left-out specimen using exclusively its images from the held-out reference protocol.

To robustly evaluate classification performance, a dual-tier analytical approach was adopted to accommodate the unique structure of the LOFO design. First, overall accuracy was calculated at the fold level (per-fruit), with statistical differences between reconstruction methods assessed using a paired t-test. Second, because individual test folds contained only a single true class, a condition that mathematically artificially deflates fold-level macro-averaged metrics, all test predictions across the LOFO iterations were pooled for each reconstruction method. Balanced accuracy and macro-averaged F1-score (macro-F1) were then computed directly on these pooled predictions to provide a descriptive evaluation of class-balanced performance.

Feature alteration

We evaluated the alteration of radiomic features induced by DLIR by quantifying the number of features significantly different between the reconstructions. This difference was also tested after compensation of the differences between the reconstructions by using the ComBat correction

procedure [33]. We applied the neuroCombat algorithm, defining the reconstruction type as the batch variable. Specifically, a mean-only model was utilized (adjusting only for location shifts without altering the variance of the features) to account for the specificities of our dataset.

Image quality assessment

To objectively evaluate the impact of the reconstruction algorithms on image quality, the Signal-to-Noise Ratio (SNR) and Contrast-to-Noise Ratio (CNR) were calculated for each phantom object across all sequences and acquisitions as $SNR = \frac{\mu_{fruit}}{\sigma_{bg}}$ and $CNR = \frac{\mu_{fruit} - \mu_{bg}}{\sigma_{bg}}$, where μ_{fruit} is the mean signal intensity within the 3D semi-automated segmentation mask of each piece of fruit, and $where\ \mu_{bg}$ and σ_{bg} are respectively the mean and standard deviation of background signal. The resulting SNR and CNR values were then compared between DLIR and non-DLIR images using a paired t-test for each sequence.

Statistical analysis

The reproducibility and repeatability of radiomic features were evaluated with several statistical tests in the literature and little statistical justification. We chose to evaluate repeatability and reproducibility each with a main statistical test, the most commonly used in the literature, and with two secondary statistical tests, to evaluate the differences between these statistical methods.

Main stability statistical tests

To evaluate reproducibility, which is the set of variations occurring under varying conditions, the features were evaluated on all 16 acquisitions (4 repetitions times of 4 different acquisition parameters) with ICC(2,1) as defined in [34] used as the main test, following [35, 36], and denoted ICC2-main.

To evaluate repeatability, which is the set of variations occurring under constant condition, the features were evaluated on the 4 repetitions (scan-rescan-reposition-scan-rescan) with the same acquisition parameters using overall concordance correlation coefficient (OCCC), following [11, 22,

35, 37, 38], denoted OCCC-main. OCCC is a generalization of twofold concordance correlation coefficient [39].

A feature was considered stable (reproducible or repeatable) if the value of the criterion, whether ICC or OCCC, was greater than 0.9, following [40].

Secondary stability statistical tests

The secondary tests to evaluate reproducibility were ICC(3,1) and OCCC, denoted ICC3-2ndary and OCCC-2ndary and the secondary tests to evaluate repeatability were ICC(1,1) and ICC(2,1) [34], denoted ICC1-2ndary and ICC2-2ndary. A secondary threshold of 0.75 was also tested.

Other statistical tests

To compare the impact of the reconstruction methods on feature stability, the proportions of reproducible and repeatable features were analyzed as paired nominal data. Because the exact same radiomic features were evaluated under both DLIR and non-DLIR conditions, reproducibility (and repeatability) was treated as a binary outcome for each feature. The statistical significance of the difference in the proportion of reproducible (and repeatable) features between the two reconstruction methods was assessed using an exact McNemar's test. Furthermore, 95% CIs for the absolute differences in these paired proportions were calculated using the asymptotic Wald method to quantify the magnitude of the stability shift provided by DLIR.

The difference of features between the reconstructions was evaluated with a Kruskall-Wallis test with Benjamini-Hochberg correction for multiple comparisons, using p=0.05 as a threshold for significance.

Statistical analysis was performed in R (version 4.3.2, R Core Team, Vienna, Austria [41]) using the psych package for ICC calculation, the epiR package for OCCC and the neuroCombat package.

Results

Overall results

The main results on all sequences for reproducibility and repeatability are presented in Table 2.

The overall percentages of reproducible features extracted with PyRadiomics and MIRP were significantly higher with DLIR (19.51% and 34.10% respectively) compared to non-DLIR (9.06% and 30.74% respectively) with a difference of 10.45%, (corresponding to 68 features), and 3.36% (35 features) respectively ($p<0.01$)

The overall percentages of repeatable features extracted with PyRadiomics and MIRP were significantly higher with DLIR (47.47% and 46.57% respectively) compared to non-DLIR (39.02% and 43.76% respectively) with a difference of 8.45%, (corresponding to 220 features), and 2.81% (104 features) respectively ($p<0.01$).

Reproducibility results

The number of reproducible features (using ICC2-main>0.9) by feature category and sequences are shown in Table 3.

By feature categories

The detailed results by feature categories among the higher-order features are shown in Table 4. First-order and higher-order features showed greater reproducibility in DLIR compared to non-DLIR. The difference with PyRadiomics was 10.32% (13 features) for first-order features and 10.48% (55 features) for higher-order features, with MIRP the difference was 1.57% for first-order features (6 features) and 4.41% (29 features) for higher-order features. An example of the distribution of 4 features is presented in Figure 3.

By sequences

Figure 4 provides a visual representation of the distribution of ICC by sequences. A visual overview comparing the reproducibility of each feature by sequence between the two reconstruction methods is provided in Additional File 1.

For both PyRadiomics and MIRP, a vast majority of sequences yielded more reproducible features with DLIR (Figure 4). The most significant differences were observed in the 3DT2 sequence, with a difference of 15.05% (14 features) and 10.81% (16 features) favoring DLIR with PyRadiomics and MIRP respectively, and in the T2 In phase sequence for PyRadiomics, with a difference of 53.76% (50 features). However, PyRadiomics and MIRP 2DT2, as well as MIRP 3DT1 were more reproducible with non-DLIR.

For the majority of sequences, the results predominantly favored DLIR in higher-order features, except for PyRadiomics and MIRP 2DT2. For first-order features, the impact of DLIR was more heterogeneous.

Secondary criteria for reproducibility

Reproducibility was also assessed using two other methods (ICC3-2ndary and OCCC-2ndary), which revealed consistent trends (Additional File 2 and Additional File 3A). Additionally, applying a threshold of 0.75 identified a greater number of features meeting the criteria, but the overall trend remained unchanged (Additional File 3B).

Image filtering

The main results of reproducibility with or without LoG filtering are shown in Additional File 4.

Key findings indicate that LoG filtering increases the difference of reproducible features between reconstructions with PyRadiomics (10.45% to 20.58%). The application of LoG filtering within the MIRP framework results in a decline in reproducibility across both reconstruction methods, with the most pronounced impact observed on DLIR; the MIRP features from LoG filtered images are more reproducible with non-DLIR.

Repeatability results

The number of repeatable features by feature category (using OCCC-main>0.9) is shown in Table 5.

The effect of reconstruction was not uniform across the sequences. Notably, 3DT1 favored non-DLIR for both radiomic platforms, whereas diffusion favored DLIR.

Secondary criteria for repeatability

Repeatability was also assessed using two other methods (ICC1-2ndary and ICC2-2ndary), which revealed consistent trends (see Additional File 5A). Additionally, when applying a threshold of 0.75, a greater number of features meeting the criteria were identified, and MIRP DLIR features appeared more repeatable (see Additional File 5B).

Discriminative power

To assess the discriminative power of each reconstruction, a classifier was trained to predict fruit type based on the radiomic features, with a Leave-One-Fruit-Out and a hold-out acquisition protocol constraint. The results of the classifications are presented in Table 6.

Feature alteration

A total of 77.11% and 32.82% of features were statistically different between DLIR and non-DLIR for PyRadiomics and MIRP respectively. The difference was greater when performing ComBat (77.57% and 40.54%). Detailed results and graphical example are shown in Additional File 6.

Image quality assessment

Results of image quality measurement are presented in Additional File 7A, showing significantly higher SNR and CNR with DLIR. A graphical example is displayed in Additional File 7B.

Discussion

Our study aimed to evaluate the impact of a Deep Learning image reconstruction method increasingly integrated into routine radiologic practice on radiomic feature stability, a critical factor for the clinical application of radiomics. DLIR offers the benefits of noise reduction and enhanced image quality but also impacts radiomic features, potentially altering their stability.

Concerning our main objective, our findings indicate that DLIR notably enhances the reproducibility of radiomic features, with a statistically significant improvement of 10.45% of features with PyRadiomics and 3.36% with MIRP, leading to more features deemed reproducible compared to non-DLIR. In addition, the repeatability of radiomic features is also significantly enhanced by DLIR (+8.45% of features with PyRadiomics and +2.81% with MIRP). These findings are consistent with equivalent studies conducted in CT [21–23]. In MRI, to the best of our knowledge, only one previous publication has partially addressed this issue [42], reporting similar reproducibility between the two reconstruction methods. However, their study design introduced substantial segmentation variability, which constituted the primary source of radiomic variations and likely masked the specific effects attributable to the reconstruction algorithms. By using a controlled setup, our study overcomes this limitation and clearly isolates the positive impact of DLIR on radiomic stability in MRI.

Concerning feature class analysis, it was expected that higher-order features would be less reproducible than first-order ones [22, 43], but it was not initially clear that reproducibility differences would arise between DLIR and non-DLIR. Non-stationary and textured noise within images degrades the overall quality and contributes to low reproducibility of radiomic features [22]. DLIR reduces this noise (Additional File 7A) [13], which might explain the observed improvements in feature reproducibility. When using LoG-filtered image with lower noise, the contribution of the reconstruction methods is more pronounced with PyRadiomics but negative with MIRP. These contrasting results suggest that the denoising capability of DLIR may play a role in radiomic feature reproducibility, although this mechanism appears to be platform-dependent.

Sequence-specific analysis revealed substantial heterogeneity. The T1 acquisitions showed low reproducibility compared to other acquisitions. This finding is similar to previous MRI phantom studies [9, 11]. A first explanation could be that T1-weighted gradient-echo sequences are highly sensitive to field instabilities and subtle flip angle variations between repeated scans. A second explanation is that T1 relaxation times are particularly susceptible to minor thermal fluctuations within the phantom during extended scanning sessions. Additionally, 2D T2 features demonstrated superior reproducibility with non-DLIR across both platforms. This departure from the general trend favoring DLIR aligns with existing evidence that feature stability varies across different protocols [9]. Such variability suggests that DLIR performance is contingent upon the specific imaging sequence and anatomical region, likely reflecting the characteristics of the underlying training datasets. Consequently, these sequence-dependent differences in image reconstruction can differentially impact the robustness of extracted radiomic features.

Despite the lack of consensus and varying definitions regarding statistical methods for radiomics stability [6, 34, 44, 45], we empirically tested the most common approaches. We found that while absolute counts of stable features varied by method, the relative ranking of DLIR versus non-DLIR remained consistent. A detailed technical comparison of these statistical frameworks is beyond our scope; however, our primary conclusion regarding the two reconstruction techniques remains robust regardless of the chosen approach.

Regarding discriminative power, our findings demonstrate that DLIR preserves, and depending on the extraction methodology, may even enhance the informative content of radiomic features. Indeed, it could have been feared that DLIR might remove relevant information about the objects or add image characteristics irrelevant to the imaged objects, thereby improving reproducibility at the expense of informative content. Our validation framework revealed a pipeline-dependent behavior: while classifiers trained on DLIR features extracted via the MIRP pipeline demonstrated superior discriminative performance, those utilizing PyRadiomics features showed comparable accuracy to standard reconstructions. Importantly, these findings demonstrate that DLIR does not result in a loss

of relevant information. At the very least, the informative content enabling the accurate identification of the fruit type is fully preserved, and under certain extraction conditions, it can even be significantly enhanced. The interpretation of these results is yet subject to a limitation: specifically, the small number of fruits relative to the number of radiomic features could lead to overfitting and performance instability. Therefore, while this phantom-based classification task does not provide a validation as robust as a clinical predictive model, it effectively serves as a proof-of-concept confirming the absence of information loss.

Concerning feature alteration, as expected, we found that MRI DLIR radiomic features differ from non-DLIR features, with 32.82% to 77.11% of features significantly different between reconstructions. This is less than in Dana et al. [25], who reported 54 to 98% depending on prostate region and sequence but higher than the findings of Rey et al. [42], who observed lower values ranging from 0.66% to 38.16% in the liver. Together, these wide variations across studies demonstrate that the alteration effect of DLIR is highly specific and strongly depends on both the imaging sequence and the underlying anatomical structure. In an exploratory analysis, we tested whether ComBat correction could address this difference. Although effective in reducing inter-center MRI variability [33] and CT reconstruction algorithms differences [46], ComBat failed to compensate for the difference between the two MR reconstruction algorithms and may even have inadvertently increased these discrepancies in certain cases. Therefore, this specific finding should be interpreted cautiously, as ComBat harmonization may not be optimally suited for intra-scanner algorithmic variations without further methodological adaptations.

The use of two radiomics platforms, PyRadiomics and MIRP, was motivated by a dual objective. PyRadiomics is the most widely used platform in the field, and it therefore appeared logical to assess the impact of DLIR on the most commonly adopted tool. We used a fixed bin width, which is the default setting of the 3DSlicer PyRadiomics interface, which is commonly used. However, PyRadiomics is known to have reproducibility issues [31], which led us to also use a second platform - MIRP - that is fully compliant with the IBSI standards and offers improved reproducibility [31, 47],

as well as a high level application programming interface. This approach yielded two sets of results: one based on PyRadiomics and the other on MIRP. Although the results are not identical, both tend to indicate greater radiomic feature stability when using DLIR. It should be noted that any direct comparison between the outputs of these two platforms must account for their different preprocessing methodologies used in the article, such as binning and normalization. Consequently, the variations between their results can reflect these methodological choices rather than purely software-related differences.

The first limitation of our study is that using a phantom that lacks human anatomy cannot fully replicate a clinical patient study. However, this methodological approach is widely used in methodological radiomics, as conducting repeated acquisitions on human subjects is hardly feasible and as phantoms provide signal characteristics similar to human tissues [36, 48]. Moreover, numerous methodological radiomic studies have used fruits as phantoms, in particular to investigate preprocessing influence [9, 11, 36]. The second limitation we need to point is the use of a single imaging device, which limits the generalizability of our findings to DLIR implementations from other vendors, different scanner models and magnetic field strengths, as well as future iterations of the same algorithm. Furthermore, DLIR algorithms rely on proprietary non-linear reconstruction behaviors, making it difficult to fully interpret the exact mathematical transformations applied to the signal. Third, the segmentation was performed by a single observer, thus preventing the assessment of inter-observer variability and introducing a potential source of bias. This is partially mitigated by the standardized semi-automatic approach used in a controlled phantom setting. Finally, regarding the classification analysis, there is a risk of possible validation leakage and overfitting due to the dataset size, which we attempted to mitigate by using rigorous grouped-validation and pooled predictions.

Building on this proof-of-concept, future studies must expand beyond phantom models to include clinical patient datasets across multi-vendor and multi-scanner settings. Furthermore, these *in vivo* evaluations will need to be conducted against established quality metrics and reporting frameworks,

such as the IBSI reporting guidelines [49] and radiomic scoring systems [1, 50], to ensure broad reproducibility and predictive reliability.

In conclusion, in this single vendor fruit phantom study we demonstrated that MRI Deep Learning-based reconstruction enhances reproducibility and repeatability of radiomic features. Furthermore, it suggests a potential improvement in discriminative power, subject to sequence- and platform-dependent effects. While these results establish a promising foundation for the future of DLIR, multi-scanner and clinical validation are required before recommending DLIR as a standard approach for MRI radiomics.

References

1. Lambin P, Leijenaar RTH, Deist TM, et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology 14:749–762. https://doi.org/10.1038/nrclinonc.2017.141

2. Temperley HC, O'Sullivan NJ, Waters C, et al (2024) Radiomics; Contemporary Applications in the Management of Anal Cancer; A Systematic Review. Am Surg 90:445–454. https://doi.org/10.1177/00031348231216494

3. Deng K, Chen T, Leng Z, et al (2024) Radiomics as a tool for prognostic prediction in transarterial chemoembolization for hepatocellular carcinoma: a systematic review and meta-analysis. Radiol Med 129:1099–1117. https://doi.org/10.1007/s11547-024-01840-9

4. Pinto Dos Santos D, Dietzel M, Baessler B (2021) A decade of radiomics research: are images really data or just patterns in the noise? Eur Radiol 31:1–4. https://doi.org/10.1007/s00330-020-07108-w

5. Zwanenburg A, Vallières M, Abdalah MA, et al (2020) The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology 295:328–338. https://doi.org/10.1148/radiol.2020191145

6. Teng X, Wang Y, Nicol AJ, et al (2024) Enhancing the Clinical Utility of Radiomics: Addressing the Challenges of Repeatability and Reproducibility in CT and MRI. Diagnostics 14:1835. https://doi.org/10.3390/diagnostics14161835

7. Peerlings J, Woodruff HC, Winfield JM, et al (2019) Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. Sci Rep 9:4800. https://doi.org/10.1038/s41598-019-41344-5

8. Bianchini L, Santinha J, Loução N, et al (2021) A multicenter study on radiomic features from T $_2$ -weighted images of a customized MR pelvic phantom setting the basis for robust radiomic models in clinics. Magnetic Resonance in Med 85:1713–1726. https://doi.org/10.1002/mrm.28521

9. Baeßler B, Weiss K, Pinto dos Santos D (2019) Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. Investigative Radiology 54:221–228. https://doi.org/10.1097/RLI.0000000000000530

10. Traverso A, Kazmierski M, Shi Z, et al (2019) Stability of radiomic features of apparent diffusion coefficient (ADC) maps for locally advanced rectal cancer in response to image pre-processing. Physica Medica 61:44–51. https://doi.org/10.1016/j.ejmp.2019.04.009

11. Wichtmann BD, Harder FN, Weiss K, et al (2023) Influence of Image Processing on Radiomic Features From Magnetic Resonance Imaging. Invest Radiol 58:199–208. https://doi.org/10.1097/RLI.0000000000000921

12. van Stiphout JA, Driessen J, Koetzier LR, et al (2021) The effect of deep learning reconstruction on abdominal CT densitometry and image quality: a systematic review and meta-analysis. Eur Radiol 32:2921–2929. https://doi.org/10.1007/s00330-021-08438-z

13. Tsuboyama T, Onishi H, Nakamoto A, et al (2022) Impact of Deep Learning Reconstruction Combined With a Sharpening Filter on Single-Shot Fast Spin-Echo T2-Weighted Magnetic

Resonance Imaging of the Uterus. Invest Radiol 57:379–386. https://doi.org/10.1097/RLI.0000000000000847

14. Ueda T, Yamamoto K, Yazawa N, et al (2024) Efficacy of compressed sensing and deep learning reconstruction for adult female pelvic MRI at 1.5 T. Eur Radiol Exp 8:103. https://doi.org/10.1186/s41747-024-00506-5

15. Rastogi A, Brugnara G, Foltyn-Dumitru M, et al (2024) Deep-learning-based reconstruction of undersampled MRI to reduce scan times: a multicentre, retrospective, cohort study. The Lancet Oncology 25:400–410. https://doi.org/10.1016/S1470-2045(23)00641-1

16. Lyu P, Liu N, Harrawood B, et al (2022) Is it possible to use low-dose deep learning reconstruction for the detection of liver metastases on CT routinely? Eur Radiol 33:1629–1640. https://doi.org/10.1007/s00330-022-09206-3

17. Lee DH, Lee JM, Lee CH, et al (2024) Image Quality and Diagnostic Performance of Low-Dose Liver CT with Deep Learning Reconstruction versus Standard-Dose CT. Radiology: Artificial Intelligence 6:e230192. https://doi.org/10.1148/ryai.230192

18. Caruso D, De Santis D, Del Gaudio A, et al (2023) Low-dose liver CT: image quality and diagnostic accuracy of deep learning image reconstruction algorithm. Eur Radiol. https://doi.org/10.1007/s00330-023-10171-8

19. Kanan A, Pereira B, Hordonneau C, et al (2024) Deep learning CT reconstruction improves liver metastases detection. Insights Imaging 15:167. https://doi.org/10.1186/s13244-024-01753-1

20. Xue G, Liu H, Cai X, et al (2023) Impact of deep learning image reconstruction algorithms on CT radiomic features in patients with liver tumors. Front Oncol 13:1167745. https://doi.org/10.3389/fonc.2023.1167745

21. Yang B, Chen X, Yuan S, et al (2022) Deep learning improves image quality and radiomics reproducibility for high-speed four-dimensional computed tomography reconstruction. Radiotherapy and Oncology 170:184–189. https://doi.org/10.1016/j.radonc.2022.02.034

22. Michallek F, Genske U, Niehues SM, et al (2022) Deep learning reconstruction improves radiomics feature stability and discriminative power in abdominal CT imaging: a phantom study. Eur Radiol 32:4587–4595. https://doi.org/10.1007/s00330-022-08592-y

23. Zhong J, Xia Y, Chen Y, et al (2022) Deep learning image reconstruction algorithm reduces image noise while alters radiomics features in dual-energy CT in comparison with conventional iterative reconstruction algorithms: a phantom study. Eur Radiol. https://doi.org/10.1007/s00330-022-09119-1

24. Li H, Alves VV, Pednekar A, et al (2024) Impact of Emerging Deep Learning–Based MR Image Reconstruction Algorithms on Abdominal MRI Radiomic Features. J Comput Assist Tomogr. https://doi.org/10.1097/RCT.0000000000001648

25. Dana J, McNabb E, Castro J, et al (2025) Deep learning reconstruction in biparametric prostate MRI: Impact on qualitative and radiomics analyses. Res Diagn Interv Imaging 14:100059. https://doi.org/10.1016/j.redii.2025.100059

26. Zhong J, Xing Y, Hu Y, et al (2025) Assessment of Robustness of MRI Radiomic Features in the Abdomen: Impact of Deep Learning Reconstruction and Accelerated Acquisition. J Digit Imaging Inform med. https://doi.org/10.1007/s10278-025-01503-9

27. Bernatz S, Zhdanovich Y, Ackermann J, et al (2021) Impact of rescanning and repositioning on radiomic features employing a multi-object phantom in magnetic resonance imaging. Sci Rep 11:14248. https://doi.org/10.1038/s41598-021-93756-x

28. Lebel RM (2020) Performance characterization of a novel deep learning-based MR image reconstruction pipeline

29. Fedorov A, Beichel R, Kalpathy-Cramer J, et al (2012) 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging 30:1323–1341. https://doi.org/10.1016/j.mri.2012.05.001

30. van Griethuysen JJM, Fedorov A, Parmar C, et al (2017) Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 77:e104–e107. https://doi.org/10.1158/0008-5472.CAN-17-0339

31. Bettinelli A, Marturano F, Avanzo M, et al (2022) A Novel Benchmarking Approach to Assess the Agreement among Radiomic Tools. Radiology 303:533–541. https://doi.org/10.1148/radiol.211604

32. Zwanenburg A, Löck S (2024) MIRP: A Python package for standardised radiomics. JOSS 9:6413. https://doi.org/10.21105/joss.06413

33. Orlhac F, Lecler A, Savatovski J, et al (2021) How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. Eur Radiol 31:2272–2280. https://doi.org/10.1007/s00330-020-07284-9

34. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86:420–428. https://doi.org/10.1037//0033-2909.86.2.420

35. Hertel A, Tharmaseelan H, Rotkopf LT, et al (2023) Phantom-based radiomics feature test–retest stability analysis on photon-counting detector CT. Eur Radiol 33:4905–4914. https://doi.org/10.1007/s00330-023-09460-z

36. Bologna M, Tenconi C, Corino VDA, et al (2023) Repeatability and reproducibility of MRI-radiomic features: A phantom experiment on a 1.5 T scanner. Medical Physics 50:750–762. https://doi.org/10.1002/mp.16054

37. Li Y, Reyhan M, Zhang Y, et al (2022) The impact of phantom design and material-dependence on repeatability and reproducibility of CT-based radiomics features. Med Phys 49:1648–1659. https://doi.org/10.1002/mp.15491

38. Muenzfeld H, Nowak C, Riedlberger S, et al (2021) Intra-scanner repeatability of quantitative imaging features in a 3D printed semi-anthropomorphic CT phantom. European Journal of Radiology 141:109818. https://doi.org/10.1016/j.ejrad.2021.109818

39. Barnhart HX, Haber M, Song J (2002) Overall Concordance Correlation Coefficient for Evaluating Agreement Among Multiple Observers. Biometrics 58:1020–1027. https://doi.org/10.1111/j.0006-341X.2002.01020.x

40. Koo TK, Li MY (2016) A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine 15:155–163. https://doi.org/10.1016/j.jcm.2016.02.012

41. R Core Team R: A Language and Environment for Statistical Computing

42. Rey L, Rabbani N, Chauveau B, et al (2026) Reproducibility and discriminative power of MRI liver radiomics: Impact of deep learning-based image reconstruction. European Journal of Radiology 113140. https://doi.org/10.1016/j.ejrad.2026.113140

43. Cattell R, Chen S, Huang C (2019) Robustness of radiomic features in magnetic resonance imaging: review and a phantom study. Vis Comput Ind Biomed Art 2:19. https://doi.org/10.1186/s42492-019-0025-6

44. Traverso A, Wee L, Dekker A, Gillies R (2018) Repeatability and Reproducibility of Radiomic Features: A Systematic Review. International Journal of Radiation Oncology*Biology*Physics 102:1143–1158. https://doi.org/10.1016/j.ijrobp.2018.05.053

45. McGraw KO, Wong SP (1996) Forming inferences about some intraclass correlation coefficients. Psychological Methods 1:30–46. https://doi.org/10.1037/1082-989X.1.1.30

46. Orlhac F, Frouin F, Nioche C, et al (2019) Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics. Radiology 291:53–59. https://doi.org/10.1148/radiol.2019182023

47. Bettinelli A, Marturano F, Sarnelli A, et al (2022) The ImSURE phantoms: a digital dataset for radiomic software benchmarking and investigation. Sci Data 9:695. https://doi.org/10.1038/s41597-022-01715-6

48. Mueller-Lisse UG, Murer S, Mueller-Lisse UL, et al (2017) Everyman's prostate phantom: kiwi-fruit substitute for human prostates at magnetic resonance imaging, diffusion-weighted imaging and magnetic resonance spectroscopy. Eur Radiol 27:3362–3371. https://doi.org/10.1007/s00330-016-4706-7

49. Zwanenburg A, Leger S, Vallières M, Löck S (2020) Image biomarker standardisation initiative. 295:328–338. https://doi.org/arXiv:1612.07003v11

50. Kocak B, Akinci D'Antonoli T, Mercaldo N, et al (2024) METhodological RadiomICs Score (METRICS): a quality scoring tool for radiomics research endorsed by EuSoMII. Insights Imaging 15:8. https://doi.org/10.1186/s13244-023-01572-w

Table title and legends

Table 1: MRI acquisitions parameters.

Bold and italic characters indicate changes compared with the reference acquisition protocol

Table 2: Overall results: Number of reproducible and repeatable features by reconstruction.

The features from all sequences were evaluated. Reproducible features are defined as ICC2-main>0.9 evaluated on all 16 acquisitions (4 repetitions times 4 different acquisition parameters). Repeatable features are defined as OCCC-main>0.9 and were evaluated on the 4 repetitions (scan-rescan-reposition-scan-rescan) with the same acquisition parameters, resulting in 4 times more evaluated features than with reproducibility. The data are presented as %(n). The difference in the proportion of reproducible (and repeatable) features between the two reconstruction methods was assessed using an exact McNemar's test.

Table 3: Number of reproducible features by feature category, sequence and reconstruction method.

Reproducible features are defined as ICC2-main>0.9.

Table 4: Number of reproducible features by feature category, feature class and reconstruction method.

Reproducible features are defined as ICC2-main>0.9.

Table 5: Number of repeatable features by feature category, sequence and reconstruction method, evaluated on 4 acquisitions.

Repeatable features are defined as OCCC-main>0.9.

Table 6: Performance of the fruit classification task depending on the reconstruction type.

Accuracy and p-values were computed at the fold level (per-fruit). Balanced Accuracy and Macro-F1 were calculated on pooled predictions across all folds to account for the single-class nature of individual test folds.

Figure legends

Figure 1: Study design.

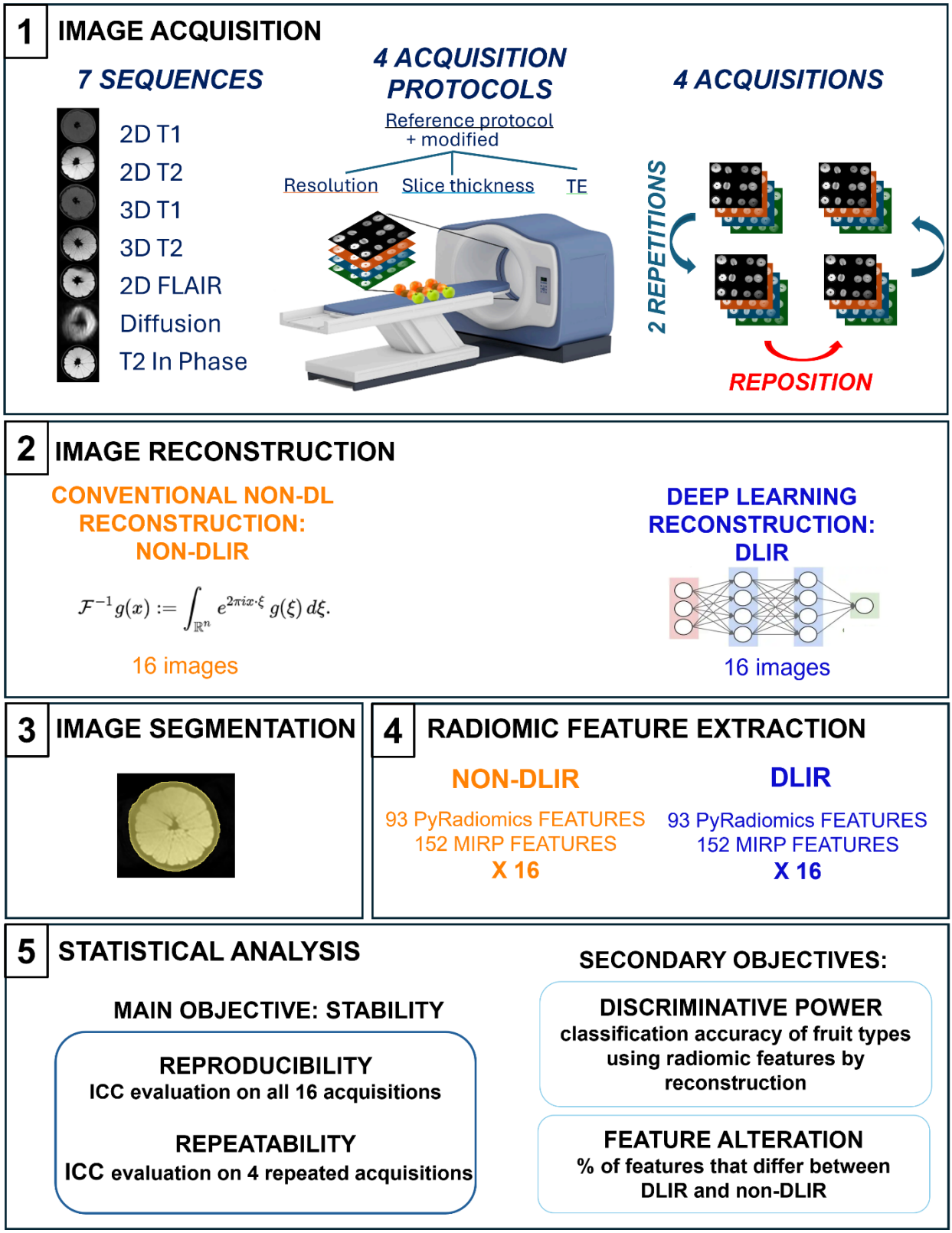


Figure 2: Representative images of fruit MRI and their segmentations. The upper row displays the MRI images; the lower row shows the overlay of the semi-automatic segmentation masks on the corresponding images. 3D T2 MRI of an orange and a kiwi with DLIR (A) and non-DLIR (B). 2D FLAIR MRI of an onion and an apple with DLIR (C) and non-DLIR (D).

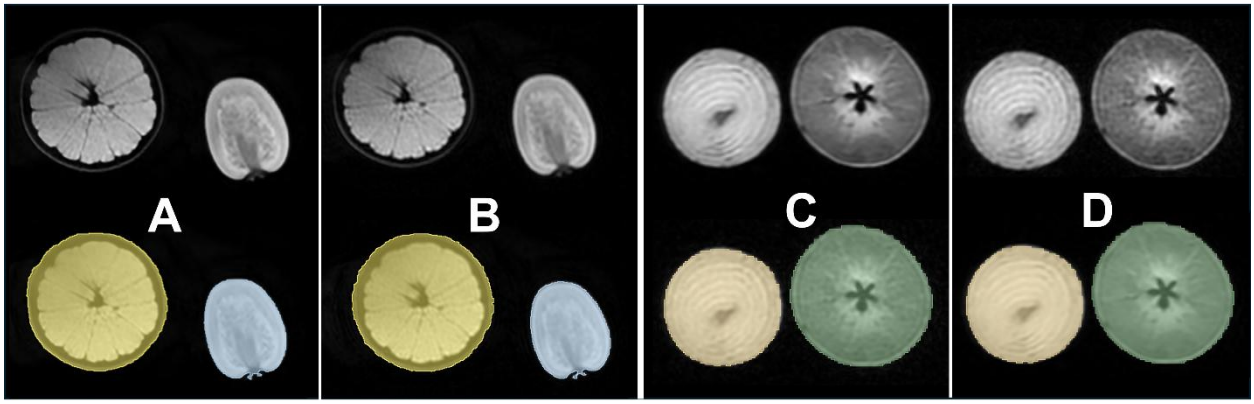


Figure 3: Distribution of 4 radiomic features: scatter plot and boxplot are represented, with corresponding ICC-main. A) PyRadiomics T2 IP GLDM Dependence Non Uniformity Normalized, reproducible with DL reconstruction, not with non-DL; B) PyRadiomic 2D FLAIR GLDM Large Dependence Low Gray Level Emphasis, not reproducible with both reconstructions; C) MIRP 2D T1 NGLM Dependence Count Non-uniformity, reproducible with DL reconstruction, not with non-DL; D) 2D T2 GLDZM Small Distance Low Gray Level emphasis, not reproducible with both reconstructions. For visualization purposes, all features have been Z-normalized.

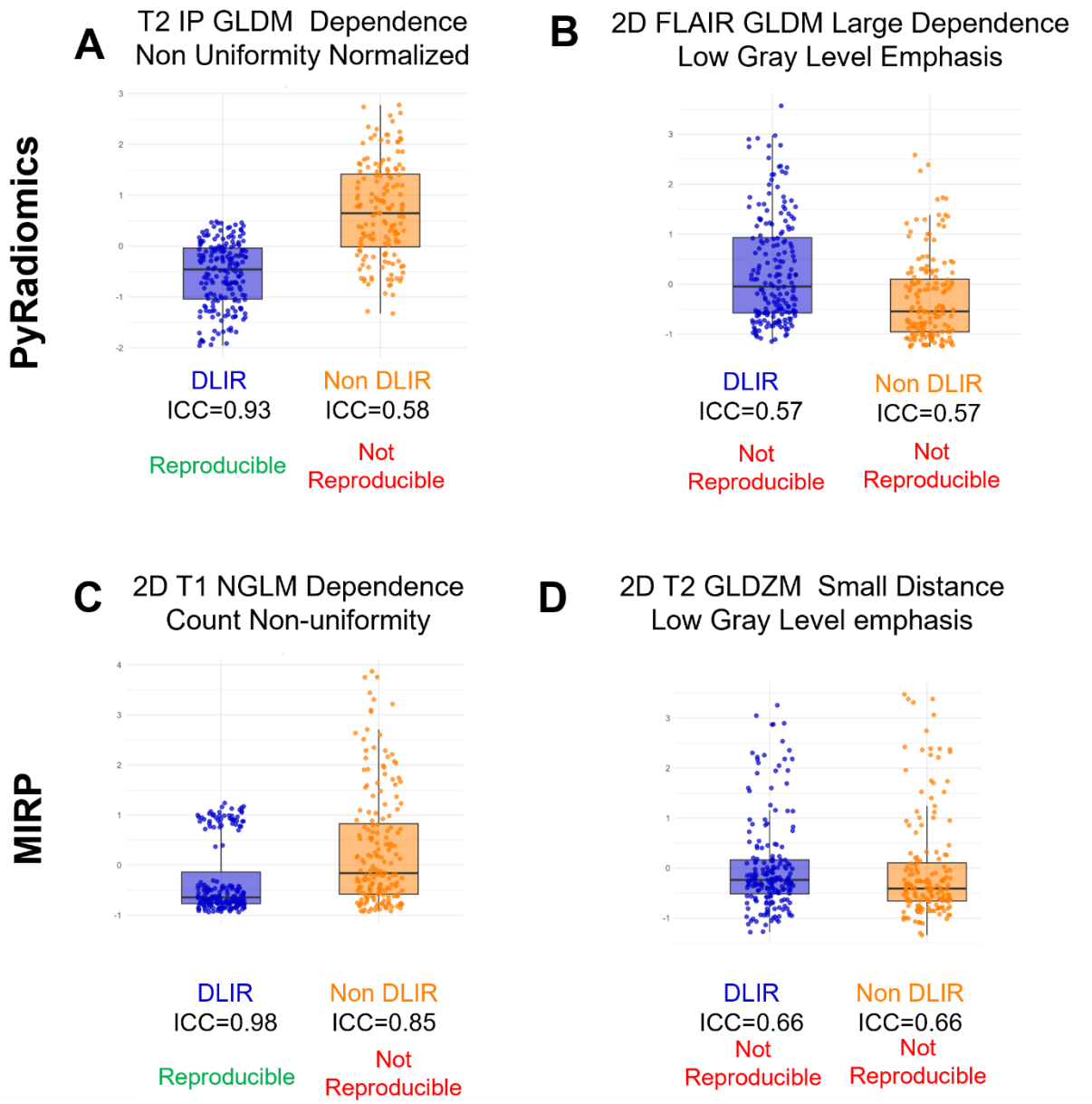


Figure 4: Boxplots of feature reproducibility evaluation by ICC2-main by reconstruction and by radiomic platform.

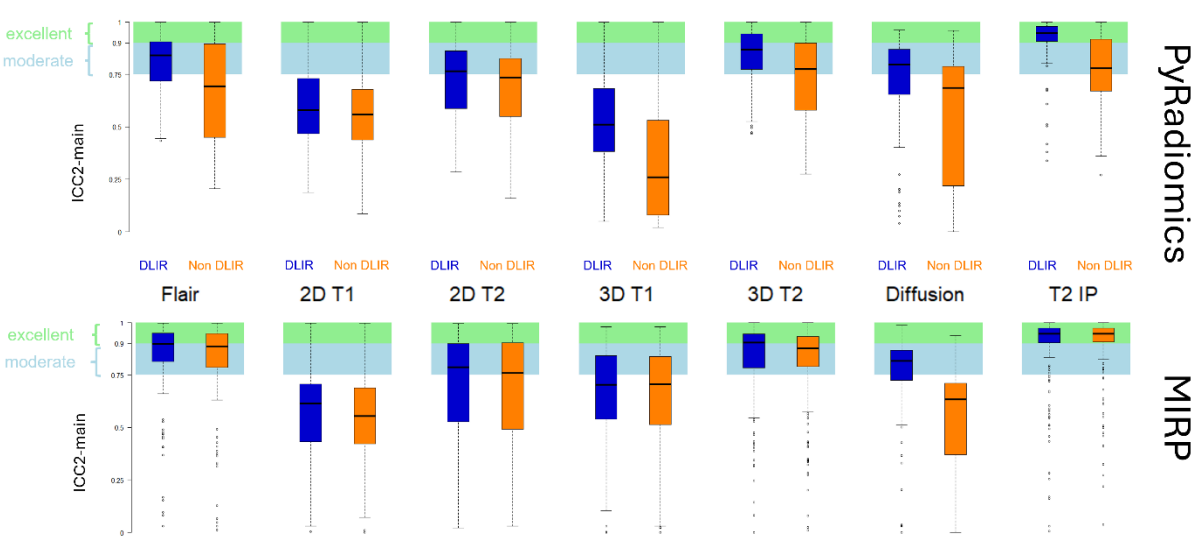


Table 1: MRI acquisitions parameters.

	In plane resolution (mm x mm)	Matrix	Slice thickness (mm)	TR (ms)	TE (ms)	Flip angle (°)	NEX	b
Reference acquisition								
FLAIR	0.7 x 0.9	352 x 256	3	12000	120	160	1	
2D T1	0.6 x 0.8	384 x 320	3	337	Min Full	160	2	
2D T2	0.8 x 1.2	288 x 192	3	Min	Min Full		1	
3D T1	0.8 x 0.8	300 x 300	1.2	Min	Min Full	15	1	
3D T2	0.8 x 0.8	320 x 320	0.8	1800	140		2	
Diffusion	1.9 x 1.2	128 x 192	3	600	80			1000
T2 IP	0.8 x 0.8	288 x 288	3	5759	108	140	2	
Modified in-plane resolution								
FLAIR	***0.6 x 0.8***	***400 x 300***	3	12000	120	160	1	
2D T1	***0.6 x 0.6***	***416 x 416***	3	337	Min Full	160	2	
2D T2	***0.7 x 1***	***352 x 240***	3	Min	Min Full		1	
3D T1	***0.6 x 0.6***	***400 x 400***	1.2	Min	Min Full	15	1	
3D T2	***0.6 x 0.6***	***380 x 380***	0.8	1800	140		2	
Diffusion	***1.5 x 1.0***	***160 x 240***	3	600	80			1000
T2 IP	***0.8 x 0.8***	***320 x 320***	3	6536	108	140	2	
Modified TE								
FLAIR	0.7 x 0.9	352 x 256	3	12000	***145***	160	1	
2D T1	0.6 x 0.8	384 x 320	3	337	***18***	160	2	
2D T2	0.8 x 1.2	288 x 192	3	Min	***170***		1	
3D T1	0.8 x 0.8	300 x 300	1.2	Min	***Out of Phase***	15	1	
3D T2	0.8 x 0.8	320 x 320	0.8	1800	***170***		2	
Diffusion	1.9 x 1.2	128 x 192	3	8500	***Min***			**2000**
T2 IP	0.8 x 0.8	288 x 288	3	5759	***130***	140	2	
Modified slice thickness								
FLAIR	0.7 x 0.9	352 x 256	***5***	12000	120	160	1	
2D T1	0.6 x 0.8	384 x 320	***5***	337	Min Full	160	2	
2D T2	0.8 x 1.2	288 x 192	***5***	Min	Min Full		1	
3D T1	0.8 x 0.8	300 x 300	***2***	Min	Min Full	15	1	
3D T2	0.8 x 0.8	320 x 320	***1.6***	1800	140		2	
Diffusion	1.9 x 1.2	128 x 192	***5***	600	80			1000
T2 IP	0.8 x 0.8	288 x 288	***5***	5759	108	140	2	

Bold and italic characters indicate changes compared with the reference acquisition protocol

Table 2: Overall results: Number of reproducible and repeatable features by reconstruction.

		PyRadiomics							MIRP						
		DLIR		Non-DLIR		Difference (DLIR - non DLIR)			DLIR		Non-DLIR		Difference (DLIR - non DLIR)		
		%	n	%	n	value	95% CI	p-value	%	n	%	n	value	95% CI	p-value
REPRODUCIBILITY	**All Features**	19.51%	127	9.06%	59	+10.45%	[7.77% - 13.12%]	<0.01	34.10%	355	30.74%	320	+3.36%	[1.66% - 5.07%]	<0.01
	First-Order Features	27.78%	35	17.46%	22	+10.32%	[4.57% - 16.07%]	<0.01	39.95%	153	38.38%	147	+1.57%	[-1.32% - 4.46%]	0.38
	Higher-Order Features	17.52%	92	7.05%	37	+10.48%	[7.46% - 13.49%]	<0.01	30.70%	202	26.29%	173	+4.41%	[2.31% - 6.51%]	<0.01
REPEATABILITY	**All Features**	47.47%	1236	39.02%	1016	+8.45%	[7.05% - 9.85%]	<0.01	46.57%	1723	43.76%	1619	+2.81%	[1.88% - 3.75%]	<0.01
	First-Order Features	54.17%	273	46.43%	234	+7.74%	[4.78% - 10.70%]	<0.01	50.67%	684	47.63%	643	+3.04%	[1.50% - 4.57%]	<0.01
	Higher-Order Features	45.86%	963	37.24%	782	+8.62%	[7.04% - 10.20%]	<0.01	44.21%	1039	41.53%	976	+2.68%	[1.50% - 3.86%]	<0.01

The features from all sequences were evaluated. Reproducible features are defined as ICC2-main>0.9 evaluated on all 16 acquisitions (4 repetitions times 4 different acquisition parameters). Repeatable features are defined as OCCC-main>0.9 and were evaluated on the 4 repetitions (scan-rescan-reposition-scan-rescan) with the same acquisition parameters, resulting in 4 times more evaluated features than with reproducibility. The data are presented as %(n). The difference in the proportion of reproducible (and repeatable) features between the two reconstruction methods was assessed using an exact McNemar's test.

Table 3: Number of reproducible features by feature category, sequence and reconstruction method.

		PyRadiomics						MIRP					
		DLIR		Non-DLIR		Difference (DLIR - non DLIR)		DLIR		Non-DLIR		Difference (DLIR - non DLIR)	
		%	n	%	n	Value	95% CI	%	n	%	n	Value	95% CI
ALL SEQUENCES	**All Features**	19.51%	127	9.06%	59	+10.45%	[7.77% - 13.12%]	34.10%	355	30.74%	320	+3.36%	[1.66% - 5.07%]
	First-Order Features	27.78%	35	17.46%	22	+10.32%	[4.57% - 16.07%]	39.95%	153	38.38%	147	+1.57%	[-1.32% - 4.46%]
	Higher-Order Features	17.52%	92	7.05%	37	+10.48%	[7.46% - 13.49%]	30.70%	202	26.29%	173	+4.41%	[2.31% - 6.51%]
2D FLAIR	**All Features**	20.43%	19	17.20%	16	+3.23%	[-2.31% - 8.76%]	52.35%	78	48.32%	72	+4.03%	[-0.85% - 8.91%]
	First-Order Features	27.78%	5	27.78%	5	+0.00%	[0.00% - 0.00%]	54.55%	30	58.18%	32	-3.64%	[-12.31% - 5.04%]
	Higher-Order Features	18.67%	14	14.67%	11	+4.00%	[-2.85% - 10.85%]	51.06%	48	42.55%	40	+8.51%	[2.87% - 14.15%]
2D T1	**All Features**	0.00%	0	0.00%	0	+0.00%	[0.00% - 0.00%]	1.34%	2	0.00%	0	+1.34%	[-0.51% - 3.19%]
	First-Order Features	0.00%	0	0.00%	0	+0.00%	[0.00% - 0.00%]	0.00%	0	0.00%	0	+0.00%	[0.00% - 0.00%]
	Higher-Order Features	0.00%	0	0.00%	0	+0.00%	[0.00% - 0.00%]	2.13%	2	0.00%	0	+2.13%	[-0.79% - 5.04%]
2D T2	**All Features**	8.60%	8	11.83%	11	-3.23%	[-8.76% - 2.31%]	26.17%	39	27.52%	41	-1.34%	[-3.96% - 1.28%]
	First-Order Features	27.78%	5	33.33%	6	-5.56%	[-16.14% - 5.03%]	50.91%	28	52.73%	29	-1.82%	[-7.97% - 4.34%]
	Higher-Order Features	4.00%	3	6.67%	5	-2.67%	[-9.04% - 3.71%]	11.70%	11	12.77%	12	-1.06%	[-3.14% - 1.01%]
3D T1	**All Features**	1.08%	1	0.00%	0	+1.08%	[-1.02% - 3.17%]	8.05%	12	9.40%	14	-1.34%	[-6.26% - 3.57%]
	First-Order Features	0.00%	0	0.00%	0	+0.00%	[0.00% - 0.00%]	7.27%	4	16.36%	9	-9.09%	[-16.69% - -1.49%]
	Higher-Order Features	1.33%	1	0.00%	0	+1.33%	[-1.26% - 3.93%]	8.51%	8	5.32%	5	+3.19%	[-3.03% - 9.41%]
3D T2	**All Features**	30.11%	28	15.05%	14	+15.05%	[7.20% - 22.91%]	57.43%	85	46.62%	69	+10.81%	[4.02% - 17.60%]

	First-Order Features	50.00%	9	27.78%	5	+22.22%	[3.02% - 41.43%]	55.56%	30	53.70%	29	+1.85%	[-4.42% - 8.12%]
	Higher-Order Features	25.33%	19	12.00%	9	+13.33%	[4.80% - 21.87%]	58.51%	55	42.55%	40	+15.96%	[6.04% - 25.87%]
Diffusion	**All Features**	3.23%	3	0.00%	0	+3.23%	[-0.37% - 6.82%]	11.41%	17	2.68%	4	+8.72%	[4.19% - 13.26%]
	First-Order Features	0.00%	0	0.00%	0	+0.00%	[0.00% - 0.00%]	30.91%	17	7.27%	4	+23.64%	[12.41% - 34.86%]
	Higher-Order Features	4.00%	3	0.00%	0	+4.00%	[-0.43% - 8.43%]	0.00%	0	0.00%	0	+0.00%	[0.00% - 0.00%]
T2 In Phase	**All Features**	73.12%	68	19.35%	18	+53.76%	[43.20% - 64.33%]	82.43%	122	81.08%	120	+1.35%	[-2.39% - 5.09%]
	First-Order Features	88.89%	16	33.33%	6	+55.56%	[32.60% - 78.51%]	81.48%	44	81.48%	44	+0.00%	[-5.13% - 5.13%]
	Higher-Order Features	69.33%	52	16.00%	12	+53.33%	[41.45% - 65.21%]	82.98%	78	80.85%	76	+2.13%	[-2.96% - 7.22%]

Reproducible features are defined as ICC2-main>0.9.

Table 4: Number of reproducible features by feature category, feature class and reconstruction method.

	PyRadiomics						MIRP						
	DLIR		Non-DLIR		Difference (DLIR - non DLIR)		DLIR		Non-DLIR		Difference (DLIR - non DLIR)		
	%	n	%	n	value	95% CI	%	n	%	n	value	95% CI	
All	**19.51%**	**127**	**9.06%**	**59**	**+10.45%**	**[7.77% - 13.12%]**	**34.10%**	**355**	**30.74%**	**320**	**+3.36%**	**[1.66% - 5.07%]**	**All**
First order	27.78%	35	17.46%	22	+10.32%	[4.57% - 16.07%]	39.95%	153	38.38%	147	+1.57%	[-1.32% - 4.46%]	**First order**
	-	-	-	-	-	-	28.57%	4	28.57%	4	+0.00%	[0.00% - 0.00%]	Loc
	-	-	-	-	-	-	44.90%	44	44.90%	44	+0.00%	[-4.90% - 4.90%]	IVH
	-	-	-	-	-	-	39.68%	50	34.13%	43	+5.56%	[0.03% - 11.08%]	Stat
	-	-	-	-	-	-	37.93%	55	38.62%	56	-0.69%	[-5.56% - 4.18%]	IH
Higher order	17.52%	92	7.05%	37	+10.48%	[7.46% - 13.49%]	30.70%	202	26.29%	173	+4.41%	[2.31% - 6.51%]	**Higher order**
GLCM	19.64%	33	6.55%	11	+13.10%	[7.02% - 19.17%]	41.14%	72	36.00%	63	+5.14%	[1.18% - 9.11%]	GLCM
GLRLM	15.18%	17	6.25%	7	+8.93%	[2.59% - 15.26%]	37.50%	42	29.46%	33	+8.04%	[1.90% - 14.17%]	GLRLM
GLSZM	14.29%	16	4.46%	5	+9.82%	[4.31% - 15.33%]	20.54%	23	16.96%	19	+3.57%	[-1.33% - 8.48%]	GLSZM
GLDM	18.37%	18	7.14%	7	+11.22%	[3.80% - 18.64%]	22.32%	25	18.75%	21	+3.57%	[0.13% - 7.01%]	GLDZM
NGTDM	22.86%	8	20.00%	7	+2.86%	[-2.66% - 8.38%]	20.00%	7	14.29%	5	+5.71%	[-1.98% - 13.40%]	NGTDM
							29.46%	33	28.57%	32	+0.89%	[-4.91% - 6.69%]	NGLDM

Reproducible features are defined as ICC2-main>0.9.

Table 5: Number of repeatable features by feature category, sequence and reconstruction method, evaluated on 4 acquisitions.

		PyRadiomics						MIRP					
		DLIR		Non-DLIR		Difference (DLIR - non DLIR)		DLIR		Non-DLIR		Difference (DLIR - non DLIR)	
		%	n	%	n	Value	95% CI	%	n	%	n	Value	95% CI
ALL SEQUENCES	**All Features**	47.47%	1236	39.02%	1016	+8.45%	[7.05% - 9.85%]	46.57%	1723	43.76%	1619	+2.81%	[1.88% - 3.75%]
	First-Order Features	54.17%	273	46.43%	234	+7.74%	[4.78% - 10.70%]	50.67%	684	47.63%	643	+3.04%	[1.50% - 4.57%]
	Higher-Order Features	45.86%	963	37.24%	782	+8.62%	[7.04% - 10.20%]	44.21%	1039	41.53%	976	+2.68%	[1.50% - 3.86%]
2D FLAIR	**All Features**	31.99%	119	25.54%	95	+6.45%	[2.86% - 10.04%]	56.93%	337	54.56%	323	+2.36%	[0.23% - 4.50%]
	First-Order Features	36.11%	26	33.33%	24	+2.78%	[-4.89% - 10.45%]	54.17%	117	55.56%	120	-1.39%	[-4.10% - 1.33%]
	Higher-Order Features	31.00%	93	23.67%	71	+7.33%	[3.29% - 11.38%]	58.51%	220	53.99%	203	+4.52%	[1.56% - 7.48%]
2D T1	**All Features**	13.44%	50	3.76%	14	+9.68%	[6.01% - 13.35%]	28.04%	166	25.51%	151	+2.53%	[-0.13% - 5.20%]
	First-Order Features	22.22%	16	11.11%	8	+11.11%	[1.25% - 20.97%]	30.56%	66	27.78%	60	+2.78%	[-1.65% - 7.21%]
	Higher-Order Features	11.33%	34	2.00%	6	+9.33%	[5.45% - 13.22%]	26.60%	100	24.20%	91	+2.39%	[-0.94% - 5.72%]
2D T2	**All Features**	50.81%	189	32.26%	120	+18.55%	[14.13% - 22.96%]	34.97%	207	32.77%	194	+2.20%	[0.48% - 3.91%]
	First-Order Features	62.50%	45	50.00%	36	+12.50%	[3.12% - 21.88%]	59.72%	129	58.33%	126	+1.39%	[-1.33% - 4.10%]
	Higher-Order Features	48.00%	144	28.00%	84	+20.00%	[15.02% - 24.98%]	20.74%	78	18.09%	68	+2.66%	[0.46% - 4.85%]
3D T1	**All Features**	45.70%	170	50.00%	186	-4.30%	[-7.43% - -1.17%]	36.66%	217	36.99%	219	-0.34%	[-2.99% - 2.31%]
	First-Order Features	37.50%	27	40.28%	29	-2.78%	[-6.57% - 1.02%]	22.22%	48	24.07%	52	-1.85%	[-5.24% - 1.53%]
	Higher-Order Features	47.67%	143	52.33%	157	-4.67%	[-8.44% - -0.89%]	44.95%	169	44.41%	167	+0.53%	[-3.15% - 4.22%]
3D T2	**All Features**	52.42%	195	41.13%	153	+11.29%	[7.30% - 15.28%]	81.76%	121	83.78%	124	-2.03%	[-5.52% - 1.46%]
	First-Order Features	62.50%	45	50.00%	36	+12.50%	[4.86% - 20.14%]	70.37%	38	74.07%	40	-3.70%	[-8.74% - 1.33%]
	Higher-Order Features	50.00%	150	39.00%	117	+11.00%	[6.41% - 15.59%]	88.30%	83	89.36%	84	-1.06%	[-5.72% - 3.59%]
Diffusion	**All Features**	52.15%	194	39.25%	146	+12.90%	[9.26% - 16.55%]	27.20%	161	15.88%	94	+11.32%	[8.29% - 14.34%]
	First-Order Features	61.11%	44	44.44%	32	+16.67%	[7.24% - 26.10%]	45.83%	99	27.31%	59	+18.52%	[12.60% - 24.44%]
	Higher-Order Features	50.00%	150	38.00%	114	+12.00%	[8.10% - 15.90%]	16.49%	62	9.31%	35	+7.18%	[3.92% - 10.44%]
T2 In Phase	**All Features**	85.75%	319	81.18%	302	+4.57%	[1.77% - 7.37%]	86.82%	514	86.82%	514	+0.00%	[-1.40% - 1.40%]
	First-Order Features	97.22%	70	95.83%	69	+1.39%	[-1.31% - 4.09%]	86.57%	187	86.11%	186	+0.46%	[-1.57% - 2.49%]
	Higher-Order Features	83.00%	249	77.67%	233	+5.33%	[1.93% - 8.74%]	86.97%	327	87.23%	328	-0.27%	[-2.15% - 1.61%]

Repeatable features are defined as OCCC-main>0.9.

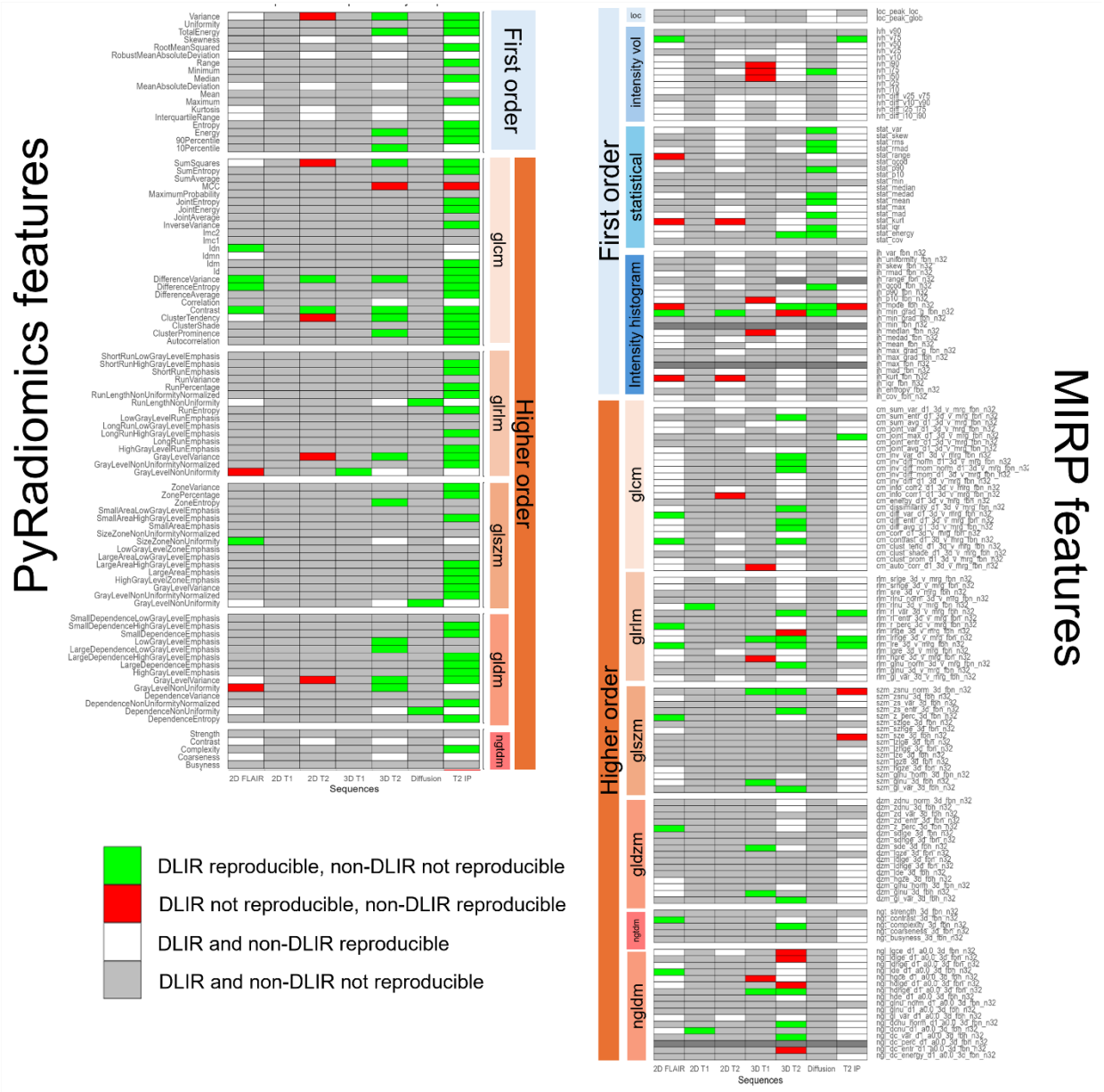
Table 6: Performance of the fruit classification task depending on the reconstruction type.

	PyRadiomics				MIRP			
	Classification **accuracy**				Classification **accuracy**			
	Mean	SD	95%CI	p-value	Mean	SD	95%CI	p-value
DLIR	0.688	0.337	[0.473 - 0.902]	0.567	0.840	0.161	[0.738 - 0.942]	**0.015**
non-DLIR	0.667	0.317	[0.465 - 0.868]		0.797	0.170	[0.689 - 0.905]	
	Balanced accuracy		**Macro F1**		**Balanced accuracy**		**Macro F1**	
DLIR	0.688		0.661		0.840		0.840	
non-DLIR	0.667		0.653		0.797		0.797	

Accuracy and p-values were computed at the fold level (per-fruit). Balanced Accuracy and Macro-F1 were calculated on pooled predictions across all folds to account for the single-class nature of individual test folds.

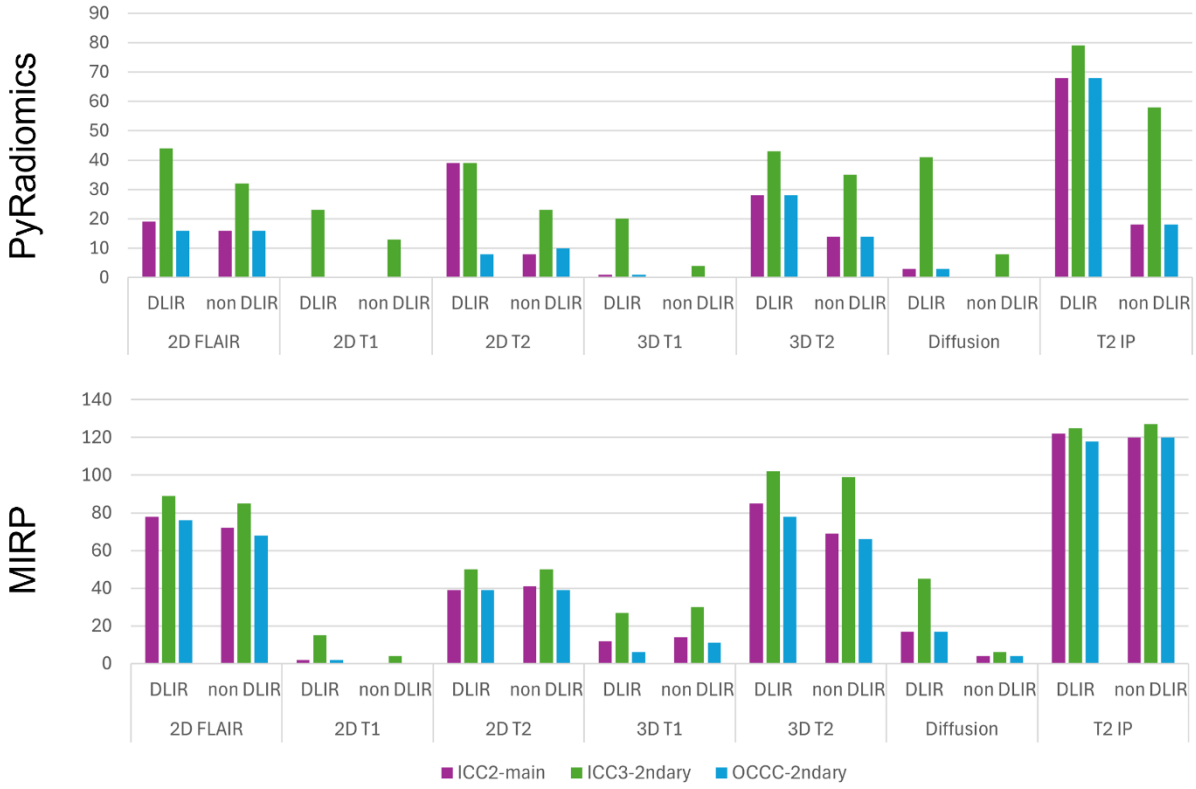
Additional File 1

Heatmap comparing the reproducibility (defined as ICC2-main)>0.9) of each feature between the two reconstructions.


PyRadiomics features
First order
Higher order
glcm
glrlm
glszm
gldm
ngtdm
2D FLAIR
2D T1
2D T2
3D T1
3D T2
Diffusion
T2 IP
Sequences
DLIR reproducible, non-DLIR not reproducible
DLIR not reproducible, non-DLIR reproducible
DLIR and non-DLIR reproducible
DLIR and non-DLIR not reproducible
MIRP features
First order
Higher order
loc
intensity vol
statistical
Intensity histogram
glcm
glrlm
glszm
gldzm
ngtdm
ngldm
2D FLAIR
2D T1
2D T2
3D T1
3D T2
Diffusion
T2 IP
Sequences

Additional File 2

Comparison of the number of reproducible features by sequence, depending on the statistical test used (threshold =0.9).



Additional File 3

Comparison of 3 methods and 2 thresholds in radiomic feature **reproducibility,** evaluated on all sequences.

A. Comparison of feature reproducibility using ICC(2,1), ICC(3,1) and OCCC with a threshold of 0.9

MAIN CRITERIA ICC(2,1)>0.9

	PyRadiomics		MIRP	
	DLIR	**Non-DLIR**	**DLIR**	**Non-DLIR**
All features	19.51% (127)	9.06% (59)	34.07% (355)	30.71% (320)
First order features	27.78% (35)	17.46% (22)	39.84% (153)	38.28% (147)
Higher order features	17.52% (92)	7.05% (37)	30.7% (202)	26.29% (173)

ICC(3,1)- 2ndary >0.9

	PyRadiomics		MIRP	
	DLIR	**Non-DLIR**	**DLIR**	**Non-DLIR**
All features	44.39% (289)	26.57% (173)	43.47% (453)	38.48% (401)
First order features	61.11% (77)	39.68% (50)	47.66% (183)	41.93% (161)
Higher order features	40.38% (212)	23.43% (123)	41.03% (270)	36.47% (240)

OCCC-2ndary >0.9

	PyRadiomics		MIRP	
	DLIR	**Non-DLIR**	**DLIR**	**Non-DLIR**
All features	19.05% (124)	8.91% (58)	32.43% (336)	29.73% (308)
First order features	26.98% (34)	17.46% (22)	39.68% (150)	36.77% (139)
Higher order features	17.14% (90)	6.86% (36)	28.27% (186)	25.68% (169)

B. Comparison of feature reproducibility using ICC(3,1), ICC(2,1) and OCCC with a threshold of 0.75

ICC(2,1)>0.75

	PyRadiomics		MIRP	
	DLIR	**Non-DLIR**	**DLIR**	**Non-DLIR**
All features	53.46% (348)	31.03% (202)	65.26% (680)	55.66% (580)
First order features	57.14% (72)	44.44% (56)	64.06% (246)	57.03% (219)
Higher order features	52.57% (276)	27.81% (146)	65.96% (434)	54.86% (361)

ICC(3,1)-2ndary>0.75

	PyRadiomics		MIRP	
	DLIR	**Non-DLIR**	**DLIR**	**Non-DLIR**
All features	77.11% (502)	64.06% (417)	76.39% (796)	72.84% (759)
First order features	89.68% (113)	80.95% (102)	75.52% (290)	73.18% (281)
Higher order features	74.1% (389)	60% (315)	76.9% (506)	72.64% (478)

OCCC - 2ndary >0.75

	PyRadiomics		MIRP	
	DLIR	**Non-DLIR**	**DLIR**	**Non-DLIR**
All features	52.38% (341)	29.8% (194)	64.77% (671)	54.83% (568)
First order features	55.56% (70)	42.86% (54)	64.02% (242)	57.67% (218)
Higher order features	51.62% (271)	26.67% (140)	65.2% (429)	53.19% (350)

Additional File 4

Number of reproducible features by feature category, sequence and reconstruction method without and with Laplacian of Gaussian (LoG) filtering (sigma =1)

ICC2-main was used at a threshold of 0.9 for reproducibility.

		PyRadiomics							MIRP						
		DLIR		Non-DLIR		Difference (DLIR - non DLIR)			DLIR		Non-DLIR		Difference (DLIR - non DLIR)		
		%	n	%	n	Difference value	95% CI	p-value	%	n	%	n	Difference value	95% CI	p-value
NO FILTERING	**All Features**	19.51%	127	9.06%	59	10.45%	[7.77% - 13.12%]	<0.01	34.10%	355	30.74%	320	3.36%	[1.66% - 5.07%]	<0.01
	First-Order Features	27.78%	35	17.46%	22	10.32%	[4.57% - 16.07%]	<0.01	39.95%	153	38.38%	147	1.57%	[-1.32% - 4.46%]	0.38
	Higher-Order Features	17.52%	92	7.05%	37	10.48%	[7.46% - 13.49%]	<0.01	30.70%	202	26.29%	173	4.41%	[2.31% - 6.51%]	<0.01
LoG FILTERING	**All Features**	26.88%	175	6.30%	41	+20.58%	[17.42% - 23.75%]	<0.01	2.88%	30	13.81%	144	-10.93%	[-13.09% - -8.77%]	<0.01
	First-Order Features	44.44%	56	11.11%	14	+33.33%	[24.81% - 41.85%]	<0.01	6.23%	24	13.25%	51	-7.01%	[-10.58% - -3.45%]	<0.01
	Higher-Order Features	22.67%	119	5.14%	27	+17.52%	[14.23% - 20.82%]	<0.01	0.91%	6	14.13%	93	-13.22%	[-15.91% - -10.53%]	<0.01

Additional File 5

Comparison of 3 methods and 2 thresholds in radiomic feature **repeatability**.

A. Comparison of feature repeatability using OCCC, ICC(1,1) and ICC(2,1) with a threshold of 0.9

MAIN CRITERIA: OCCC>0.9

	PyRadiomics		MIRP	
	DLIR	**Non-DLIR**	**DLIR**	**Non-DLIR**
All features	47.68% (1215)	39.21% (999)	46.57% (1723)	46.97% (1877)
First order features	54.17% (273)	46.43% (234)	50.67% (684)	50.41% (735)
Higher order features	46.09% (942)	37.43% (765)	44.21% (1039)	45% (1142)

ICC(1,1) - 2ndary >0.9

	PyRadiomics		MIRP	
	DLIR	**Non-DLIR**	**DLIR**	**Non-DLIR**
All features	49.14% (1252)	40.54% (1033)	48.27% (1798)	48.25% (1941)
First order features	55.56% (280)	47.42% (239)	51.05% (702)	51.04% (758)
Higher order features	47.55% (972)	38.85% (794)	46.64% (1096)	46.61% (1183)

ICC(2,1) - 2ndary >0.9

	PyRadiomics		MIRP	
	DLIR	**Non-DLIR**	**DLIR**	**Non-DLIR**
All features	49.25% (1255)	40.58% (1034)	48.32% (1800)	48.27% (1942)
First order features	55.75% (281)	47.42% (239)	51.13% (703)	51.04% (758)
Higher order features	47.65% (974)	38.89% (795)	46.68% (1097)	46.65% (1184)

B. Comparison of feature repeatability using OCCC, ICC(1,1) and ICC(2,1) with a threshold of 0.75

MAIN CRITERIA: OCCC>0.75				
	PyRadiomics		MIRP	
	DLIR	Non-DLIR	DLIR	Non-DLIR
All features	75.12% (1914)	69.74% (1777)	76.08% (2815)	72.75% (2907)
First order features	82.74% (417)	81.15% (409)	76% (1026)	73.05% (1065)
Higher order features	73.24% (1497)	66.93% (1368)	76.13% (1789)	72.58% (1842)

ICC(1,1) - 2ndary >0.75				
	PyRadiomics		MIRP	
	DLIR	Non-DLIR	DLIR	Non-DLIR
All features	75.67% (1928)	71.19% (1814)	76.75% (2859)	73.55% (2959)
First order features	82.94% (418)	81.75% (412)	75.35% (1036)	72.73% (1080)
Higher order features	73.87% (1510)	68.59% (1402)	77.57% (1823)	74.03% (1879)

ICC(2,1) - 2ndary >0.75				
	PyRadiomics		MIRP	
	DLIR	Non-DLIR	DLIR	Non-DLIR
All features	76.14% (1940)	71.59% (1824)	77.18% (2875)	73.95% (2975)
First order features	83.93% (423)	82.74% (417)	75.78% (1042)	73.2% (1087)
Higher order features	74.22% (1517)	68.84% (1407)	78% (1833)	74.39% (1888)

Additional File 6

Number (and %) of different features between DL and non-DL reconstructions with or without pre- and post-processing steps (Kruskal Wallis test with Benjamini–Hochberg correction for multiple comparisons)

A. Differences between PyRadiomics features

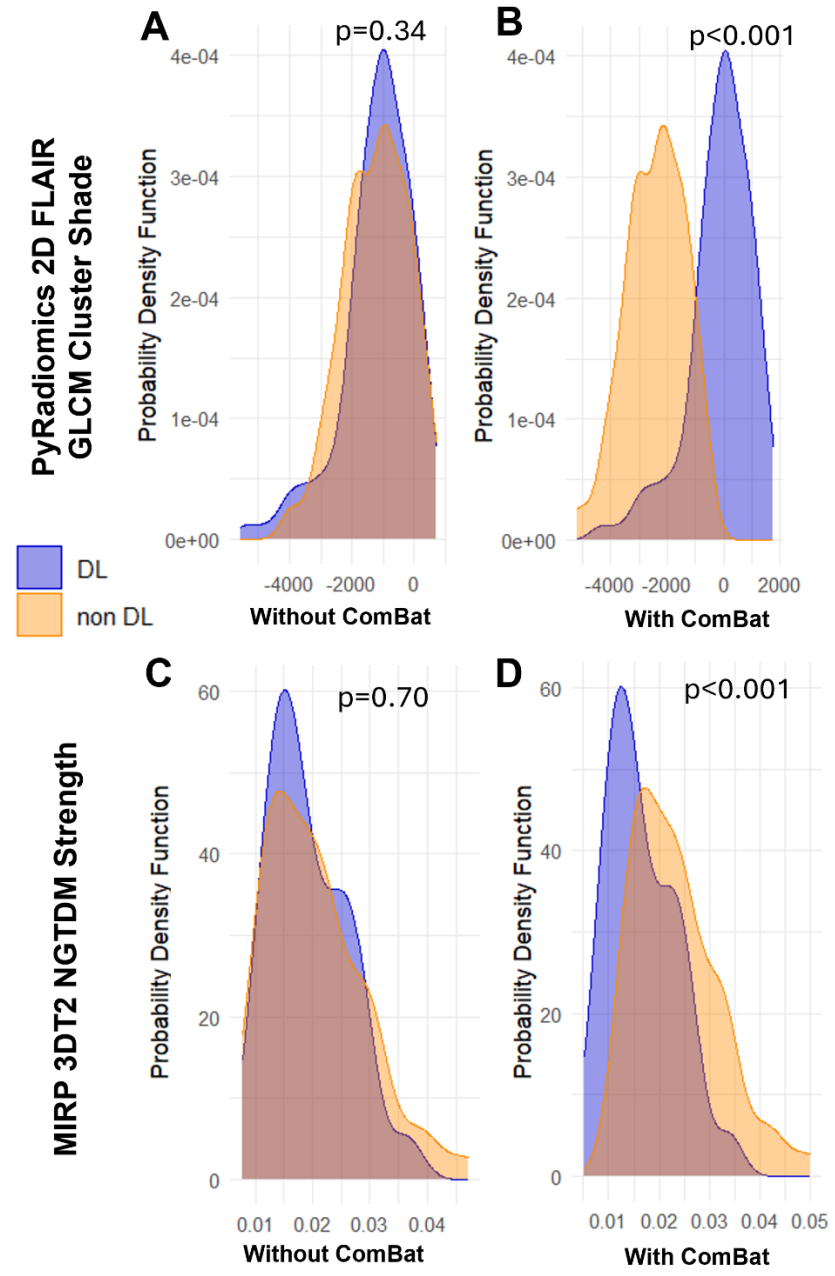
	N4 bias correction Image intensity Normalisation		N4 bias correction Image intensity Normalisation and ComBat	
Category	n	%	n	%
Overall	502	77.11	505	77.57
First Order	89	70.63	**80**	63.49
Higher Order	413	78.67	425	80.95
GLCM	126	75	132	78.57
GLRLM	91	81.25	94	83.93
GLSZM	93	83.04	89	79.46
GLDM	81	82.65	87	88.78
NGTDM	22	62.86		

B. Differences between MIRP features

	N4 bias correction Image intensity Normalisation		N4 bias correction Image intensity Normalisation and ComBat	
Category	n	%	n	%
Overall	**340**	**32,82**	**420**	**40,54**
First Order	**71**	**18,78**	**115**	**30,42**
Loc	6	42,86	7	50
IVH	15	15,31	25	25,51
Stat	24	17,14	40	28,57
IH	26	20,63	43	34,13
Higher Order	**269**	**40,88**	**305**	**46,35**
GLCM	62	35,43	65	37,14
GLRLM	48	42,86	49	43,75
GLSZM	56	50	64	57,14
GLDZM	40	35,71	54	48,21
NGTDM	10	28,57	15	42,86
NGLDM	53	47,32	58	51,79

C. Graphical example

Example of two radiomic features (PyRadiomics 2D Flair GLCM Cluster Shade (A and B) and MIRP 3DT2 NGTDM Strength (C and D)) for which the difference between DL and non-DL reconstruction was worsened by ComBat correction. Difference was evaluated with Kruskall -Wallis test with Benjamini–Hochberg correction for multiple comparisons.



Additional File 7

A. SNR and CNR measurements

SNR

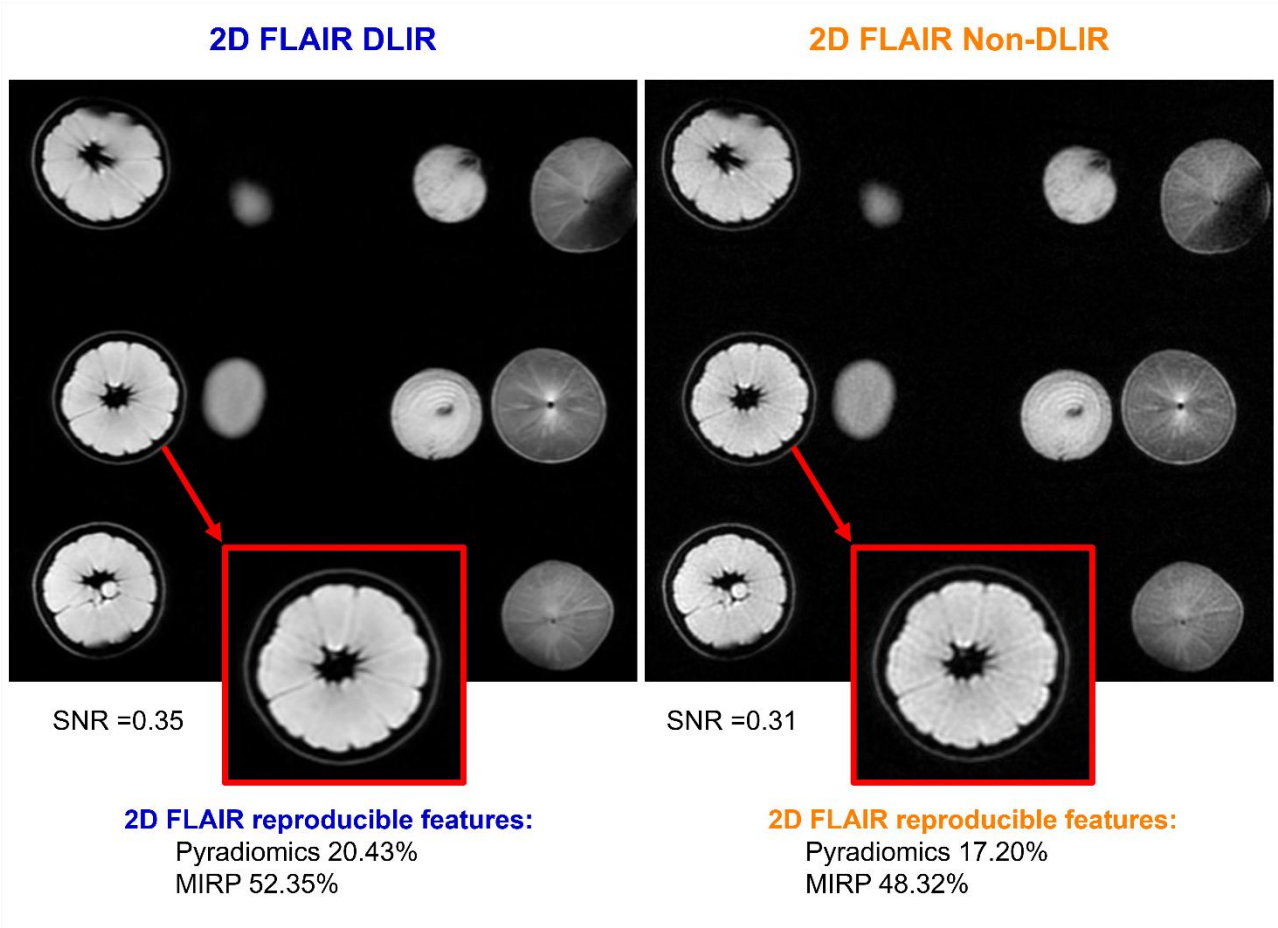
	DLIR		Non-DLIR		
Sequence	**mean**	**SD**	**mean**	**SD**	**p-value**
2DFLAIR	1.91E+08	2.27E+08	1.62E+08	2.19E+08	<0.001
2D T1	5.44E+07	3.27E+07	6.77E+07	3.94E+07	<0.001
2D T2	2.49E+08	2.91E+08	2.23E+08	2.93E+08	<0.001
3D T1	3.79E+07	5.53E+07	1.56E+08	1.33E+08	<0.001
3D T2	5.61E+07	8.70E+07	1.89E+08	1.47E+08	<0.001
Diffusion	5.96E+01	5.16E+01	3.76E+01	3.14E+01	<0.001
T2 IP	3.01E+08	3.47E+08	2.85E+08	3.91E+08	<0.001

	DLIR		Non-DLIR		
Sequence	**mean**	**SD**	**mean**	**SD**	**p-value**
2DFLAIR	1.91E+08	2.27E+08	1.62E+08	2.19E+08	<0.001
2D T1	5.44E+07	3.27E+07	6.77E+07	3.94E+07	<0.001
2D T2	2.49E+08	2.91E+08	2.23E+08	2.93E+08	<0.001
3D T1	3.79E+07	5.53E+07	1.56E+08	1.33E+08	<0.001
3D T2	5.61E+07	8.70E+07	1.89E+08	1.47E+08	<0.001
Diffusion	5.89E+01	5.16E+01	3.69E+01	3.14E+01	<0.001
T2 IP	3.01E+08	3.47E+08	2.85E+08	3.91E+08	<0.001

SNR and CNR measurements by sequences and reconstruction

B. Graphical example

2D slice of a 2DFLAIR acquisition, with DLIR and non-DLIR. An orange is magnified in the red box. The signal to noise ratio (SNR) is higher in the DLIR. The number of reproducible features is higher with DLIR for both radiomic platforms.


2D FLAIR DLIR
2D FLAIR Non-DLIR
SNR =0.35
SNR =0.31
2D FLAIR reproducible features:
Pyradiomics 20.43%
MIRP 52.35%
2D FLAIR reproducible features:
Pyradiomics 17.20%
MIRP 48.32%