Medical & Biological Engineering & Computing
https://doi.org/10.1007/s11517-024-03071-6

ORIGINAL ARTICLE

Impact of harmonization on the reproducibility of MRI radiomic features when using different scanners, acquisition parameters, and image pre-processing techniques: a phantom study

Ghasem Hajianfar[1] · Seyyed Ali Hosseini[2,3] · Sara Bagherieh[4] · Mehrdad Oveisi[5] · Isaac Shiri[1,6] · Habib Zaidi[1,7,8,9]

Received: 7 October 2023 / Accepted: 5 March 2024

Abstract

This study investigated the impact of ComBat harmonization on the reproducibility of radiomic features extracted from magnetic resonance images (MRI) acquired on different scanners, using various data acquisition parameters and multiple image pre-processing techniques using a dedicated MRI phantom. Four scanners were used to acquire an MRI of a nonanatomic phantom as part of the TCIA RIDER database. In fast spin-echo inversion recovery (IR) sequences, several inversion durations were employed, including 50, 100, 250, 500, 750, 1000, 1500, 2000, 2500, and 3000 ms. In addition, a 3D fast spoiled gradient recalled echo (FSPGR) sequence was used to investigate several flip angles (FA): 2, 5, 10, 15, 20, 25, and 30 degrees. Nineteen phantom compartments were manually segmented. Different approaches were used to pre-process each image: Bin discretization, Wavelet filter, Laplacian of Gaussian, logarithm, square, square root, and gradient. Overall, 92 first-, second-, and higher-order statistical radiomic features were extracted. ComBat harmonization was also applied to the extracted radiomic features. Finally, the Intraclass Correlation Coefficient (ICC) and Kruskal-Wallis's (KW) tests were implemented to assess the robustness of radiomic features. The number of non-significant features in the KW test ranged between 0–5 and 29–74 for various scanners, 31–91 and 37–92 for three times tests, 0–33 to 34–90 for FAs, and 3–68 to 65–89 for IRs before and after ComBat harmonization, with different image pre-processing techniques, respectively. The number of features with ICC over 90% ranged between 0–8 and 6–60 for various scanners, 11–75 and 17–80 for three times tests, 3–83 to 9–84 for FAs, and 3–49 to 3–63 for IRs before and after ComBat harmonization, with different image pre-processing techniques, respectively. The use of various scanners, IRs, and FAs has a great impact on radiomic features. However, the majority of scanner-robust features is also robust to IR and FA. Among the effective parameters in MR images, several tests in one scanner have a negligible impact on radiomic features. Different scanners and acquisition parameters using various image pre-processing might affect radiomic features to a large extent. ComBat harmonization might significantly impact the reproducibility of MRI radiomic features.

Keywords MRI · Radiomics · Robustness · Pre-processing · Harmonization · Scanner effect

✉ Habib Zaidi
habib.zaidi@hcuge.ch

[1] Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, CH-1211 Geneva, Switzerland

[2] Translational Neuroimaging Laboratory, McGill University Research Centre for Studies in Aging, Douglas Hospital, McGill University, Montréal, Québec, Canada

[3] Department of Neurology and Neurosurgery, Faculty of Medicine, McGill University, Montréal, Québec, Canada

[4] School of Medicine, Isfahan University of Medical Sciences, Isfahan, Iran

[5] Department of Computer Science, University of British Columbia, Vancouver, BC, Canada

[6] Department of Cardiology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland

[7] Department of Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, Groningen, Netherlands

[8] Department of Nuclear Medicine, University of Southern Denmark, Odense, Denmark

[9] University Research and Innovation Center, Óbuda University, Budapest, Hungary

Published online: 27 March 2024

1 Introduction

Magnetic resonance imaging (MRI) provides detailed, clinically relevant images of soft tissue structures, which are impossible to attain via any other non-invasive medical imaging modality [1–4]. Therefore, it is frequently utilized for cancer diagnosis, staging, and follow-up. However, contrary to positron emission tomography (PET) and computed tomography (CT), which provide images coded in Hounsfield and kBq/mL units, MRI intensity gray levels do not use a particular standard unit owing to the lack of intensity for each specific tissue [1–4]**.** Consequently, the intensity varies for the same MRI scanner, imaging protocol, and biological tissue, hence the necessity of intensity normalization [1–4]. Moreover, the radiomic feature reproducibility extracted from MR images is affected by numerous parameters, including but not limited to magnetic field strength, gradient strength, MR sequence, image acquisition protocol, and reconstruction algorithm [1–4].

In combination with machine learning, extracting high-throughput quantitative measures from medical images, referred to as radiomics, is used to create models for prediction, screening, diagnosis, response to treatment, and prognosis using medical images and clinical data [1, 3–5]. Hence, it is essential that radiomic features from various modalities, such as PET [6], CT [7], and MRI [8], be reproducible. In other words, it is crucial to obtain features that can be verified by subsequent research with an identical technique, dataset, and/or patient cohort to confirm that the analysis has been conducted without errors.

Previous studies demonstrated that various factors might impact radiomic features in MR images to a large extent, including image pre-, post-processing [9, 10], test-retest [11], and multi-center [12, 13]. To overcome the low reproducibility of radiomic features, several methods have been proposed, among which selecting reproducible features and ComBat harmonization against influential factors seem to be plausible solutions [14].

Harmonization approaches were developed to improve the repeatability of research on radiomic features using medical imaging by removing undesired impacts of vendor-dependent features or resolving inconsistencies across medical images [15]. Harmonizing MR images is feasible using two distinct approaches, namely prior to and following feature extraction [16, 17]. The present study focuses on the second approach, i.e., using harmonized radiomic features once they have been extracted [16, 17].

ComBat harmonization has been widely used for different imaging modalities in a variety of scenarios, thus demonstrating its ability to decrease radiomic feature variability in CT [18], PET [6], and MRI [19]. This popular method was introduced by Johnson et al. [20] to remove batch effects impacts in microarray expression and then applied to PET, CT, and MR images [6, 18, 21, 22]. In addition, Orlhac et al. [1] used ComBat to eliminate the variability of MRI radiomic features in a multi-center study. Moreover, in another study, Li et al. [3] used this method for harmonized MRI radiomic features extracted from 3 and 1.5 Tesla magnetic field strength scanners.

The variability of radiomic features might be caused by several factors, such as varying flip angles (FAs) and inversion recovery (IR) in the same MR scanner and separate scanners with almost the same protocols and situations (in a phantom study) [23–25]. In MRI, the FA affects signal intensity and contrast in various tissues [23]. While a smaller flip angle speeds up scanning and improves the signal-to-noise ratio, it might impair T1 contrast and saturation recovery [23]. IR determines which tissue will be without signal or nulled according to the selection of inversion time (TI) [24, 25]. Since only one texture is used in the phantom, using different TIs affects signal intensity. The signal from the phantom will be nullified if TI is equal to T1 [24, 25]. The signal from the phantom will be positive or negative depending on whether the TI is shorter or longer than the T1 of the phantom, respectively [24, 25]. A variety of image pre-processing techniques had an additional effect on this variability [9, 10]. The current study aims to investigate the effect of the ComBat harmonization method on the reproducibility of MRI radiomic features for different scanners and acquisition parameters with different image pre-processing techniques using a dedicated MRI phantom study.

The novelty and main contribution of the current study can be summarized in the following items:

- Exploring the effect of ComBat harmonization on the reproducibility of MRI radiomic features;
- Investigating a broad range of inversion recovery (IR) sequences and flip angles (FA) in a nonanatomic phantom from the TCIA RIDER for MRI scans;
- Employing various pre-processing techniques, such as bin discretization, wavelet filters, and Laplacian of Gaussian, to comprehensively evaluate the impact of these methods on the robustness and consistency of MRI radiomic features;
- Investigating the impact of ComBat harmonization on the reproducibility of MRI radiomic features for different scanners and acquisition parameters with different image pre-processing techniques using a dedicated MRI phantom.

2 Materials and methods

Different steps involved in the implementation of the current study are shown in Fig. 1.

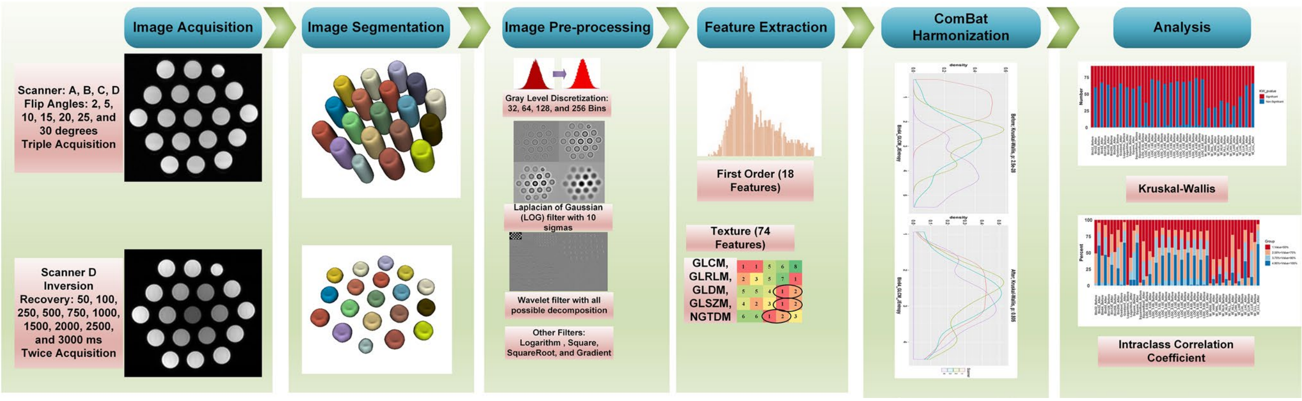


Image Acquisition
Image Segmentation
Image Pre-processing
Feature Extraction
ComBat Harmonization
Analysis
Scanner: A, B, C, D
Flip Angles: 2, 5, 10, 15, 20, 25, and 30 degrees
Triple Acquisition
Scanner D
Inversion Recovery: 50, 100, 250, 500, 750, 1000, 1500, 2000, 2500, and 3000 ms
Twice Acquisition
Gray Level Discretization: 32, 64, 128, and 256 Bins
Laplacian of Gaussian (LOG) filter with 10 sigmas
Wavelet filter with all possible decomposition
Other Filters: Logarithm , Square, SquareRoot, and Gradient
First Order (18 Features)
Texture (74 Features)
GLCM,
GLRLM,
GLDM,
GLSZM,
NGTDM
density
Kruskal-Wallis
Intraclass Correlation Coefficient

Fig. 1 Workflow summarizing the different steps involved in the current study

Table 1 Description of MRI scanners and protocols used in the current study

Name	Scanner	MRI coils	Gradient specifications	Field-of-view
Scanner A	GE 1.5T	8 Channel HD	BRM gradient subsystem (33 mT/m amplitude; 120 T/m-s)	24 × 19 cm
Scanner B	GE 1.5T	8 Channel HD	CRM gradient subsystem (50 mT/m amplitude; 150 T/m-s)	24 × 19 cm
Scanner C	Siemens 1.5T	8 Channel HD	Espree (VB13) 33 mT/m amplitude, 100 T/m-s gradient subsystem	24 × 19 cm
Scanner D	GE 3.0T	8 Channel HD	TwinSpeed gradients (40 mT/m; 150 T/m-s in zoom mode)	24 × 19 cm

2.1 Phantom design

The nonanatomic MRI phantom from TCIA (RIDER database), containing 18 gel-filled tubes (25 mm) and one 20-mm tube filled with 0.25 mM GdDTPA, was used in this study [26–28].

2.2 Evaluated scanners

MR images of the above phantom were acquired on 4 scanners by RIDER database. The description of scanners and protocols are summarized in Table 1. Scanner A was chosen for multiple FAs and three times tests, whereas scanner D was selected for multiple IRs to assess the impact of various FAs, tests, and IRs in one scanner owing to the availability of a large number of images on this scanner.

Table 2 shows variables and constant acquisition parameters for 4 analytical approaches 2 [26, 27]. The first analysis was performed on multiple scanners. Here, 4 scanners were used. For each scanner, 3 images with 1-h and 1-week (scanner D 2-week) intervals were acquired. Other constant acquisition parameters are presented in Table 2. Multiple tests were performed on scanner A with triplet tests (1-h and 1-week intervals). Constant acquisition parameters are shown in Table 2. An investigation of multiple FAs with 2, 5, 10, 15, 20, 25, and 30 degrees was performed by a 3D fast spoiled gradient recalled echo (FSPGR) sequence. For each FA, 3 images with 1-h and 1-week intervals were acquired. Other constant values for acquisition parameters are presented in Table 2. Different inversion times in fast spin-echo inversion recovery sequences included 50, 100, 250, 500, 750, 1000, 1500, 2000, 2500, and 3000 ms. For each IR, 2 images with 2-week intervals were acquired. Constant acquisition parameters are shown in Table 2 [26, 27].

2.3 Image segmentation

Manual segmentation of 19 phantom compartments was performed using 3D Slicer version 4.11 [29]. The FAs series contained 12 slices, where the first and last slice was excluded during manual segmentation owing to the change in intensity in this section. IR images were acquired in a single slice, and each slice was separated and segmented.

2.4 Image pre-processing

Before feature extraction, each image was pre-processed using three methods: (i) Bin discretization (32, 64, 128, and 256 bins), (ii) Laplacian of Gaussian (LOG) filter with 10 sigma's (0.5 to 5 mm in 0.5-mm increment), (iii) wavelet filter with a combination of low- and high-pass filters in 3-dimensions, and (iv) other filters, including logarithm, square, square root, and gradient. It was not possible to use the LOG filter on IR image series since these series

Table 2 List of variables and constant acquisition parameters for 4 different methods

Evaluation	Variable	Constant
Scanners	Scanner A, Scanner B, Scanner C, Scanner D	T1 Measurements: 3D fast spoiled gradient recalled echo (FSPGR) FA: 20 TE: 1.22 ms TR: 6.38 ms Matrix size: 512 × 512 Slice number: 10 Thickness: 5-mm sections Acquisition time per FA: 0:58 sec
Multiple test	Triplet with 1 hour and 1-week interval	Scanner A T1 Measurements: 3D FSPGR FA: 20 TE: 1.22 ms TR: 6.38 ms Matrix size: 512 × 512 Slice number: 10 Thickness: 5-mm sections Acquisition time per FA: 0:58 sec
Flip angles	2, 5, 10, 15, 20, 25, and 30 degrees	Scanner A T1 Measurements: 3D FSPGR TE: 1.22 ms TR: 6.38 ms Matrix size: 512 × 512 Slice number: 10 Thickness: 5-mm sections Acquisition time per FA: 0:58 sec
Inversion recovery	50, 100, 250, 500, 750, 1000, 1500, 2000, 2500, and 3000 ms.	T1 measurements: fast spin-echo inversion recovery sequence TE: 8.7 ms TR: 5000 ms Matrix size: 256 × 256 Slice number: 1 Thick: 10-mm sections Acquisition time per inversion time: 4 min and 25 sec

consisted of a single slice. Sixty-four bin discretization has been adopted for features extracted from LOG and wavelet-filtered images. These methods were implemented using PyRadiomics [30], which is compliant with image biomarker standardization initiative (IBSI) guidelines for radiomic analysis [31, 32]. In this study, we used fixed bin numbers for Bin discretization based on our previous study [8]. LOG filter was used for edge detection and extraction of key points on the image [30]. Low sigma refers to a fine filter, whereas higher sigma makes the filter coarser [30]. For the wavelet filter, we used Coiflets 1 from PyWavelet library [33] with 8 decompositions, including LLL, LLH, LHL, LHH, HLL, HLH, HHL, and HHH [30]. Logarithm, square, and square root filters were applied to the image and logarithm, square, and square root were calculated from image intensities [30]. Gradient calculated the gradient magnitude of an image. Further details about the use of filters can be found in [30–32].

2.5 Radiomic features extraction

Ninety-two features were extracted within each ROI of phantom images using the IBSI-compatible [34] PyRadiomics package [30] in Python for each pre-processing method, including two feature sets: first-order (FO, 18 features) and textures which also included (i) gray level co-occurrence matrix (GLCM, 23 features), (ii) gray level run length matrix (GLRLM, 16 features), (iii) gray level dependence matrix (GLDM, 14 features), (iv) gray level size zone matrix (GLSZM, 16 features), and (v) neighboring gray tone difference matrix (NGTDM, 5 features).

2.6 ComBat harmonization

The ComBat harmonization method, which Johnson et al. first proposed, assumes that the feature value of y calculated in VOI j and batch (scanner, test, FA, or IR) i is calculated using Eq. (1) [20]:

$$y_{ij} = \alpha + X_{ij}\beta + \gamma_i + \delta_i\varepsilon_{ij} \quad (1)$$

Accordingly, X indicates a design vector (matrix) for biological covariate(s) of interest, whereas α and β stipulate standard linear regression coefficients [20]. In addition, γi captures the additive batch effect on features (normal distribution assumption), while δ_i captures the multiplicative batch effect (inverse gamma distribution assumption) and ε_{ij} represents an error part (assumed to have zero-mean normal distribution) [20].

The method below was developed by Fortin et al. [21, 22] using an empirical Bayes model which estimates γ_i and δ_i parameters (denoted as γ_i* and δ_i*), with the normalized feature value of y for VOIj and batch i as follows [20–22]:

$$y_{ij}^{Combat} = \frac{y_{ij} - \hat{\alpha} - X_{ij}\hat{\beta} - \gamma_i^*}{\delta_i^*} + \hat{\alpha} + X_{ij}\hat{\beta} \quad (2)$$

wherein α and β parameters were estimated and noted as $\hat{\alpha}$ and $\hat{\beta}$ in Eq. (2), respectively [21, 22]. It is worth mentioning that ComBat harmonization employs a transformation method for each feature, which is separately governed by the batch effect observed on feature values [21, 22]. As such, a non-parametric model with an empirical Bayes estimation of the ComBat method was applied, revealing no biological covariates and no assumptions for γ_i, δ_i, and ε_{ij}. The ComBat R function[1] used in this study is publicly available [21, 22].

2.7 Data analysis

The investigation of the ComBat harmonization effect on feature values was performed by Kruskal-Wallis's one-way test. This test was applied to features before and after harmonization among multiple batches. The batches here defined multiple scanners, tests, FAs, and IRs. A *p*-value less than 0.05 was considered statistically significant. Features with a significant *p*-value indicate a significant difference between batches, whereas a non-significant *p*-value indicates no significant difference between batches. The intraclass correlation coefficient (ICC) calculation was performed for each individual radiomic feature over a varoius batch with a random two-way effects model. ICC values were categorized into 4 groups, (i) 90% ≤ ICC ≤ 100%, (ii) 75% ≤ ICC < 90%, (iii) 50% ≤ ICC < 75%, and (iv) ICC < 50% [35]. Radiomic features showing ICC > 90% were selected as robust features against each effective factor. We used irr package version 0.84.1 for ICC and stats package for the KW test in R version 4.0.4 (The R Foundation, Vienna, Austria) [36].

3 Results

Figure 2a depicts the KW test among four scanners before and after ComBat harmonization when using various image pre-processing techniques. The number of features with non-significant *p*-values (lower variability) ranged between 0–5 and 29–74 before and after ComBat harmonization along with various image pre-processing techniques, respectively. The LOG filter with 4.5-mm sigma had the highest non-significant *p*-values (2 before, 74 after). Other LOG filters performed better than other methods (65–72 features after).

Figure 2b shows the KW test among three times tests before and after ComBat harmonization using various image pre-processing steps. The number of non-significant features set is considerably higher than the results of the KW test evaluating other factors, where W_HHH_Before, W_LHH_Before, and Bin256_After had the highest number of significant *p*-value features with 61, 60, and 55 features, respectively. Besides, LOG_2.5S, LOG_3.0S, LOG_4.0S, W_HHL, and W_HLH, after ComBat harmonization, had non-significant *p*-value features, thus demonstrating the constructive impact of ComBat harmonization on features variability. These feature sets had 84, 83, 90, 57, and 64 non-significant features before ComBat harmonization.

Figure 2c shows the KW test among multiple FA before and after ComBat harmonization when using various image pre-processing techniques. The number of features with non-significant *p*-values ranged from 0–33 to 34–90 before and after ComBat harmonization along with different image pre-processing steps, respectively. The wavelet filter with HHL decomposition had the highest number of non-significant *p*-values (12 before and 90 after), followed by the gradient filter, LOG filter with 5.0-mm sigma, and wavelet filter with HLH decomposition, which had 85, 84, and 84 non-significant features after ComBat harmonization. Other wavelet filters performed better than other methods (62–84 features after).

Figure 2d depicts the KW test among multiple IR before and after ComBat harmonization when using various image pre-processing techniques. The number of features with non-significant *p*-values ranged from 3–68 to 65–89 before and after ComBat harmonization along with other image pre-processing techniques, respectively. The wavelet filter with HLH, LHH, and HHH decompositions had the highest non-significant *p*-values before (68, 34, and 45 features) and after (89, 89, and 86 features) ComBat harmonization using various image pre-processing techniques, respectively. The square root filter had 82 non-significant features after ComBat harmonization.

Supplemental Figures 1-4 depict the KW test results for each radiomic feature before and after ComBat harmonization using various image pre-processing techniques. Supplemental Table 1 shows which radiomic feature had over 20 non-significant (over

[1] https://github.com/Jfortin1/ComBatHarmonization

Fig. 2 Results of Kruskal-Wallis's (KW) test before and after ComBat harmonization, over various image pre-processing techniques for MR images acquired on: **a** four scanners, **b** three times test, **c** various flip angles, and **d** various inversion recoveries. LOG, Laplacian of Gaussian (LOG); S, sigma; W, wavelet; L, low-pass filter; H, high-pass filter

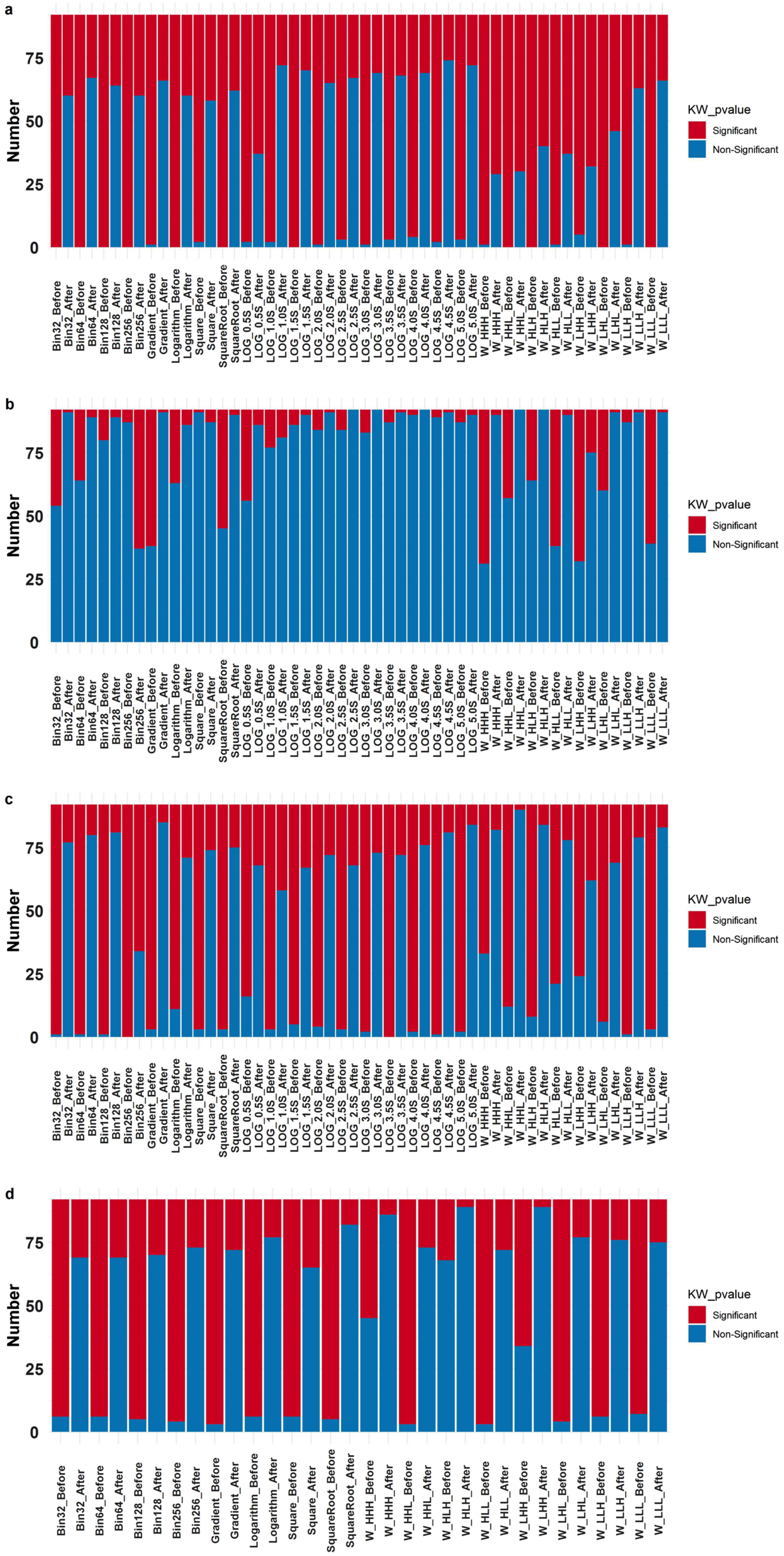


Table 3 Comparison of the number of non-significant features before and after ComBat harmonization when using different image pre-processing techniques

Feature set	Scanner		Three times test		FA		IR	
	Before	After	Before	After	Before	After	Before	After
Bin32	0	60	54	91	1	77	6	69
Bin64	0	67	64	89	1	80	6	69
Bin128	0	64	80	89	1	81	5	70
Bin256	0	60	87	37	0	34	4	73
Gradient	1	66	38	91	3	85	3	72
Logarithm	0	60	63	86	11	71	6	77
Square	2	58	91	87	3	74	6	65
SquareRoot	0	62	45	90	3	75	5	82
LOG_0.5S	2	37	56	86	16	68		
LOG_1.0S	2	72	77	81	3	58		
LOG_1.5S	0	70	86	90	5	67		
LOG_2.0S	1	65	84	91	4	72		
LOG_2.5S	3	67	84	92	3	68		
LOG_3.0S	1	69	83	92	2	73		
LOG_3.5S	3	68	87	91	0	72		
LOG_4.0S	4	69	90	92	2	76		
LOG_4.5S	2	74	89	91	1	81		
LOG_5.0S	3	72	87	90	2	84		
W_HHH	1	29	31	90	33	82	45	86
W_HHL	0	30	57	92	12	90	3	73
W_HLH	0	40	64	92	8	84	68	89
W_HLL	1	37	38	90	21	78	3	72
W_LHH	5	32	32	75	24	62	34	89
W_LHL	0	46	60	91	6	69	4	77
W_LLH	1	63	87	91	1	79	6	76
W_LLL	0	66	39	91	3	83	7	75

*FA filp angle, IR inverstion recovery, LOG Laplacian of Gaussian (LOG), S sigma, W wavelet, L low-pass filter, H high-pass filter

15 for IR) in scanners, three times repeated the test, various flip angles, and inversion recovery across various image pre-processing techniques after ComBat harmonization. Supplemental Figures 5-8 indicate that instance radiomic features were common in 4 analyses. Table 3 shows the numbers of non-significant features before and after ComBat harmoniztion, along with various image pre-processing techniques.

In Fig. 3, we illustrate the ICC percentage of various radiomic feature sets before and after ComBat harmonization. We also show the number of robust features with ICC ≥ 90% in different feature sets before and after harmonization in Table 4. It is evident from Fig. 3a that the impact of different scanners on the LOG feature set and different wavelet feature sets are the least and the most, respectively. Furthermore, ComBat harmonization affects the reproducibility of the radiomic feature set against all parameters, where 80% of feature sets had no robust features (ICC ≥ 90%) before ComBat harmonization. Furthermore, Logarithm_Before, SquareRoot_Before, W_HHH_Before, W_HLH_Before, W_HLL_Before, W_LHH_Before, and W_LLH_Before features sets had no features in the third group (90% < ICC ≥ 75%) either. All of the feature sets showed robust features after ComBat harmonization. After ComBat harmonization, W_LHH_After showed the least reproducibility with 6 robust features, whereas Square_After and Gradient_After led to the highest number of robust features with 60 features. As observed in Fig. 3b, three times, the test showed the most negligible impact on the radiomic features set, and Square feature sets showed the most robustness with 75 and 80 features before and after ComBat harmonization. Different wavelet feature sets are the most robust features against various FA (Fig. 3c). In Fig. 3d, the Logarithm_After features set with 63 robust features showed the most reproducibility, and W-LLH_After showed the least reproducibility with 3 robust features over various IR after ComBat harmonization.

Supplemental Figures 9-12 show the ICC heat map of radiomic features with different image pre-processing techniques over various scanners, three times repeated tests, various flip angles, and inversion recovery before and after ComBat harmonization, respectively.

Fig. 3 ICC percentage value of different radiomic features set over: **a** various scanners, **b** three times repeated tests, **c** various flip angles, and **d** inversion recovery before and after ComBat harmonization

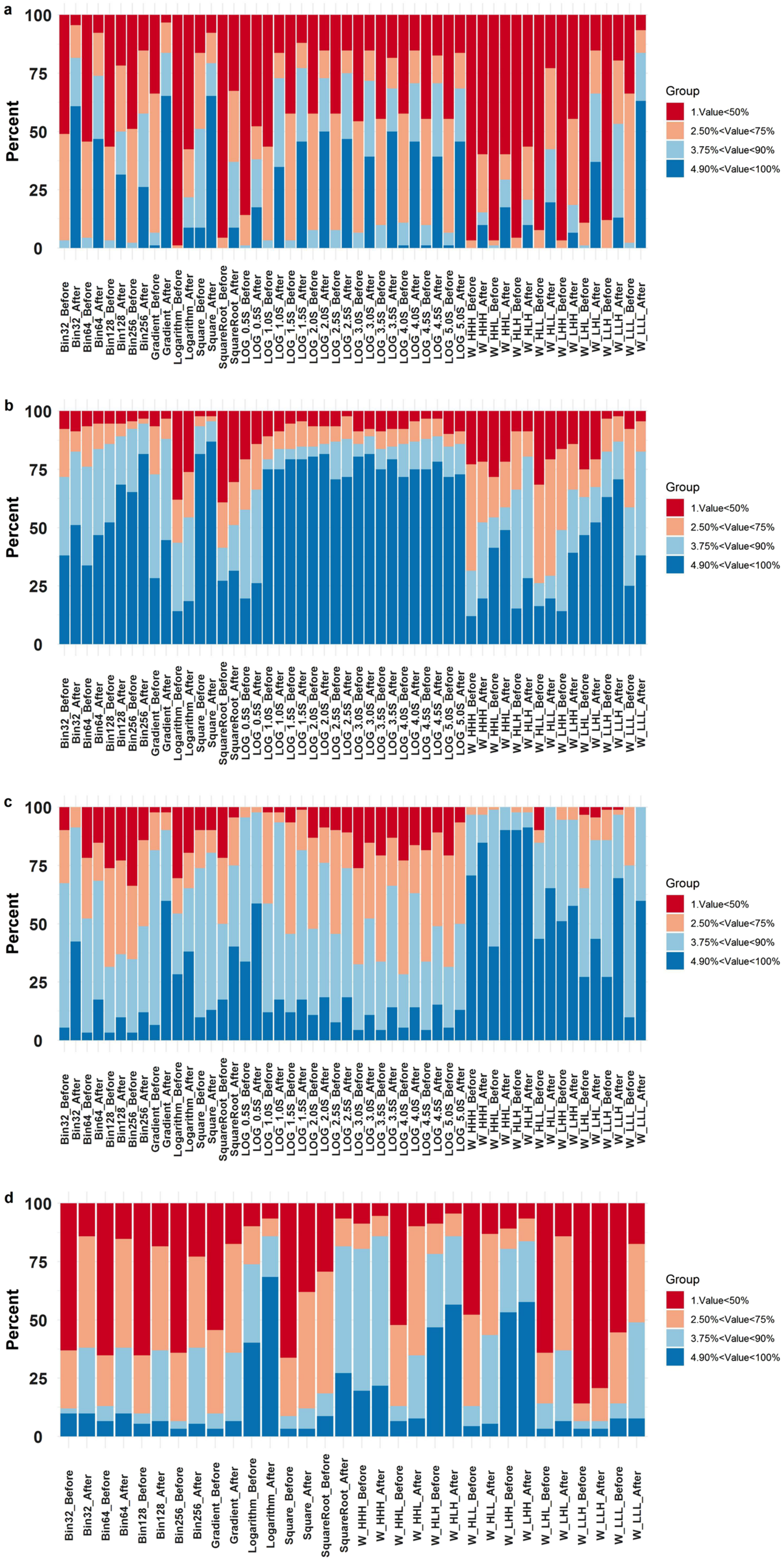


Table 4 Comparison of the number of rubust features (ICC > 90%) before and after ComBat harmonization when using different image pre-processing techniques

Feature set	Scanner		Three times test		FA		IR	
Set	Before	After	Before	After	Before	After	Before	After
Bin32	0	56	35	47	5	39	9	9
Bin64	0	43	31	43	3	16	6	9
Bin128	0	29	48	63	3	9	5	6
Bin256	0	24	60	75	3	11	3	5
Gradient	1	60	26	41	6	55	3	6
Logarithm	0	8	13	17	26	35	37	63
Square	8	60	75	80	9	12	3	3
Square root	0	8	25	29	16	37	8	25
LOG_0.5S	0	16	18	24	31	54		
LOG_1.0S	0	32	69	69	11	16		
LOG_1.5S	0	42	73	73	11	16		
LOG_2.0S	0	46	74	75	10	17		
LOG_2.5S	0	43	65	66	7	17		
LOG_3.0S	0	36	74	75	4	10		
LOG_3.5S	0	46	69	73	4	13		
LOG_4.0S	1	42	66	69	5	13		
LOG_4.5S	1	36	69	72	4	14		
LOG_5.0S	1	42	66	67	5	12		
W_HHH	0	9	11	18	65	78	18	20
W_HHL	0	16	38	45	37	83	6	7
W_HLH	0	9	14	26	83	84	43	52
W_HLL	0	18	15	18	40	60	4	5
W_LHH	0	6	13	36	47	53	49	53
W_LHL	0	34	43	48	25	40	3	6
W_LLH	0	12	58	65	25	64	3	3
W_LLL	0	58	23	35	9	55	7	7

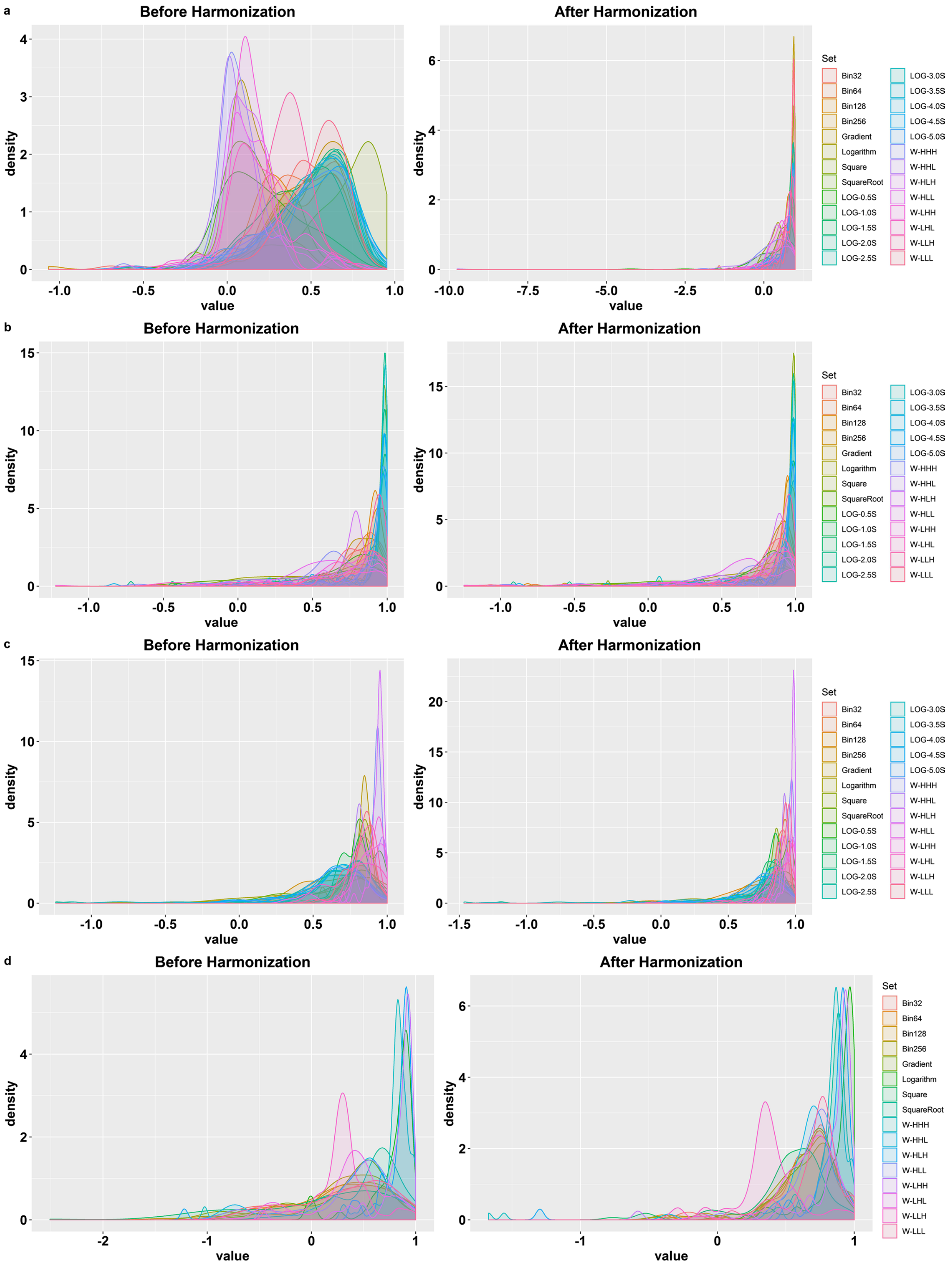
**FA* filp angle, *IR* inverstion recovery, *LOG* Laplacian of Gaussian (LOG), *S* sigma, *W* wavelet, *L* low-pass filter, *H* high-pass filter

Figure 4a depicts the impact of multiple scanners on radiomic features where the distribution of ICC values is scaled between − 1 and + 1. When the density graph is left uneven, the mean is less than the median in the density plot (DP). Conversely, the plot concentration on the right panel (+ 1) illustrates the robustness. The left panel in Fig. 4a shows the massive distribution of ICC values representing the low reproducibility of radiomic features set over the different scanners before ComBat harmonization. The right panel of the same figure (after ComBat harmonization) indicates the beneficial effect of ComBat harmonization on the radiomic features set robustness. Figure 4b shows that most of the feature sets had high reproducibility before and after harmonization. Figure 4c and d illustrate the ICC values concentration before and after ComBat harmonization over various FAs and IR. Although the impact of FA/IR is less than when using multiple scanners, the DP of radiomic features turns left after ComBat harmonization, which proves the constructive effect of ComBat harmonization on the reproducibility of the radiomic features set.

4 Discussion

We investigated the effect of ComBat harmonization on radiomic features extracted from MRI phantom when applying different image pre-processing techniques on images acquired using various imaging protocols (multiple FAs and IRs) and scanners. The results indicated that ComBat harmonization decreased the variability of radiomic features across various multi-center images and imaging protocols acquired on different scanners. In addition, using different image pre-processing techniques reduced the radiomic features' variability.

The results of the KW test indicate that the number of non-significant radiomic feature sets will rise after applying ComBat harmonization regardless of image pre-processing techniques applied in the current study. Following ComBat harmonization, LOG_4.5S, LOG_1.0S, and LOG_5.0S had the highest number of non-significant radiomic feature sets over multiple scanners in the KW test (Fig. 2) with 74, 72, and 72 non-significant features, respectively. Besides, prior


a
Before Harmonization
After Harmonization
b
Before Harmonization
After Harmonization
c
Before Harmonization
After Harmonization
d
Before Harmonization
After Harmonization
density
value
Set
Bin32
Bin64
Bin128
Bin256
Gradient
Logarithm
Square
SquareRoot
LOG-0.5S
LOG-1.0S
LOG-1.5S
LOG-2.0S
LOG-2.5S
LOG-3.0S
LOG-3.5S
LOG-4.0S
LOG-4.5S
LOG-5.0S
W-HHH
W-HHL
W-HLH
W-HLL
W-LHH
W-LHL
W-LLH
W-LLL
-1.0
-0.5
0.0
0.5
1.0
-10.0
-7.5
-5.0
-2.5
-1.5
-2
-1
0
1
2
3
4
5
6
10
15
20

◂**Fig. 4** ICC value density plots (DPs) of the different radiomic feature sets over various scanners (**a**), three LOG: Laplacian of Gaussian (LOG), S, sigma; W, wavelet; L, low-pass filter; H, high-pass filter, times repeated test (**b**), flip angles (**c**), and inversion recovery (**d**), before (left panel) and after (right panel) ComBat harmonization. LOG, Laplacian of Gaussian (LOG); S, sigma; W, wavelet; L, low-pass filter; H, high-pass filter

to ComBat harmonization, Bin32, Bin64, Bin128, Bin256, logarithm, square root, LOG_1.5S, W_HHL, W_HLH, W_LHL, and W_LLL had no significant radiomic features. MRI scanners have the largest impact on radiomic features among the parameters investigated in the current study. The use of various MRI scanners showed the largest effect on radiomic features before and after ComBat, even more than IR and FA combined. Ninety-four percent of radiomic features that are robust against FA and IR simultaneously are also robust against different scanners. Three times test had the most negligible impact on radiomic features variability.

Concerning the effects of image pre-processing on the variability of radiomic features, the study by Demircioğlu et al. [37] used public radiomic datasets to investigate the effect of various pre-processing filters on the predictive performance of radiomic models. They found that adding features pre-processed with various filters improved the predictive performance, although using pre-processing filters in some datasets showed the opposite [37]. Tuning the filters further improved the results, indicating that pre-processing filters should be used in radiomic studies to improve the predictive performance [37]. Moradmand et al. [38] investigated the impact of pre-processing techniques on MRI radiomic features and reported that 23% of radiomic features after bias field correction were robust (ICC > 90%). Yet, overlooking inter-scanner, inter-vendor, and inter-protocol variations in radiomics research can not only adversely affect the results but may also lead to failure in the process of finding uncertainties in radiomics research. In spite of the significance of such deviations, only few studies have investigated this sphere and identified precautionary measures.

In a study conducted by Orlhac et al. [1], the RIDER MRI phantom scanned on 1.5 T and 3 T scanners were used to extract 42 radiomic features, 40 of which had significant differences prior to ComBat harmonization. Following ComBat harmonization, this number was reduced to 0 features. The same phantom data was used in the present work, but contrary to the above reference, we used all available scanners, including three 1.5 T scanners and one 3 T scanner, besides extracting all 3D IBSI radiomic features and implementing various image pre-processing techniques, including different discretization of bins (32, 64, 128, and 256 bins), logarithm, square, square root, and gradient filters, LOG filter with 10 sigmas, and wavelet filter with 8 decompositions. Furthermore, all data used in the current study were acquired three times using multiple flip angles and inversion recovery settings. Our findings confirm the results of this study, i.e., all features with different bin discretization (32, 64, 128, and 256) had significant differences before scanner harmonization. However, this number varied in other pre-processing methods (0–5 features had non-significant differences).

Furthermore, our results demonstrated that the best feature set was the LOG filter (4.5-mm sigma), with 74 features with non-significant differences in different scanners. In a recent study, Li et al. [3] investigated how pre-processing steps and harmonization procedures (such as the ComBat method for radiomic features) may reduce scanner effects and enhance radiomic features' repeatability in brain MRI radiomics. Their findings are entirely in line with ours in the sense that ComBat harmonization might increase radiomic features reproducibility to a large extent over various image pre-processing steps.

Another noteworthy finding that is highlighted in our results is that several times testing turned out to have the most negligible impact on radiomic features. In the context of radiomics, imaging at multiple time points enables researchers to analyze features' robustness to temporal variabilities, e.g., organ expansion, shrinkage, and motion [39]. However, the significance of temporal variations fades in relation to inter-scanner variations, which are capable of causing much more fundamental inconsistencies between samples [40]. Lee et al. [41] investigated the robustness of radiomic features in an MRI phantom. The ICC for test-retest analysis in phantoms with different materials was reported to be high (average ICC = 0.96 for T1-w images). While our study employed a three-time test, the ICC was also high for the majority of pre-processing methods, such as the LOG filter with different sigma.

Few previous studies explored the impact of image pre-processing techniques and ComBat harmonization at the same time and managed to follow concise protocols [3, 10, 42]. Nevertheless, there is a key point to keep in mind when interpreting the present findings since Baeßler et al. [4] showed in a phantom study that the number of robust features in a FLAIR MR image is higher than in T1- and T2-weighted images. Consequently, it is expected that different MR sequences might affect radiomic features, which was not explored in the current study.

Alternative methods for feature extraction, such as deep learning-based feature [43], Bag of Features (BoF) [44] and Local Binary Patterns (LBPs) [45] were also reported in the literature. These techniques were not used in the current study. Further analysis is required to explore the effect of various feature extraction methods. Our findings underscore the importance of harmonization in MRI radiomics, potentially enhancing diagnostic accuracy and reliability in multi-center studies. By reducing variability across different scanners and protocols, ComBat harmonization could lead to more consistent radiomic features, improving patient care through better-informed diagnostic and prognostic models. This advancement holds promise for standardized imaging biomarkers in clinical practice, offering a path toward more personalized and precise medical interventions.

5 Conclusion

ComBat harmonization appears to be a decent solution to enhance MRI radiomic features reproducibility. The use of multiple scanners had the highest impact on radiomic features variability, followed by IR and FA. Most of the robust features against scanners are robust against IR and FA. However, acquiring several test images on a single scanner had the lowest impact on radiomic features among the remaining parameters. The main contribution of the current study is the consideration of various image pre-processing and data acquisition protocols using different scanners and 3 times repeated scanning to avoid any errors. However, our study inherently bears some limitations, the main one being that the effect of different MR imaging protocols was overlooked. Future studies will evaluate the effect of other MRI scanning protocols on the reproducibility of radiomic features to tackle this limitation. Another limitation is the use of only phantom images. Clinical studies are required to validate these results.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11517-024-03071-6.

Funding Open access funding provided by University of Geneva This work was supported by the Swiss National Science Foundation under grant SNSF 320030_176052.

Data availability The data used in this study are publically available from the TCIA archive. (https://doi.org/10.7937/K9/TCIA.2015.MI4QDDHU).

Code availability Open-source libraries used in this study. PyRadiomics (https://github.com/AIM-Harvard/pyradiomics) for image preprocessing and feature extraction in Python, ComBat harmonization (https://github.com/Jfortin1/ComBatHarmonization) in R and irr and stats packages for statistical analysis.

Declarations

Conflict of interest The authors declare no competing interests.

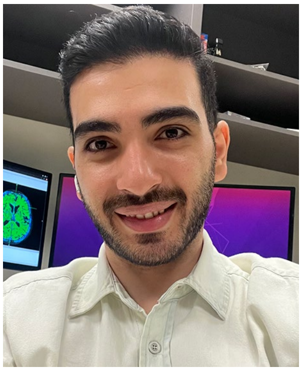
References

1. Orlhac F, Lecler A, Savatovski J, Goya-Outi J, Nioche C, Charbonneau F et al (2021) How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. Eur Radiol 31(4):2272–2280
2. Karayumak SC, Bouix S, Ning L, James A, Crow T, Shenton M et al (2019) Retrospective harmonization of multi-site diffusion MRI data acquired with different acquisition parameters. Neuroimage 184:180–200
3. Li Y, Ammari S, Balleyguier C, Lassau N, Chouzenoux E (2021) Impact of preprocessing and harmonization methods on the removal of scanner effects in brain mri radiomic features. Cancers 13(12):3000
4. Baeßler B, Weiss K, Dos Santos DP (2019) Robustness and reproducibility of radiomics in magnetic resonance imaging: a phantom study. Investig Radiol 54(4):221–228
5. Mahon RN, Ghita M, Hugo GD, Weiss E (2020) ComBat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets. Phys Med Biol 65(1):015010
6. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L et al (2018) A postreconstruction harmonization method for multicenter radiomic studies in PET. J Nucl Med 59(8):1321–1328
7. Meyer M, Ronald J, Nelson RC, Ramirez-Giraldo JC, Solomon J, Patel BN et al (2019) Reproducibility of CT radiomic features within the same patient: influence of radiation dose and CT reconstruction settings. Radiology. 293(3):583–591
8. Shiri I, Hajianfar G, Sohrabi A, Abdollahi H, Shayesteh PS, Geramifar P et al (2020) Repeatability of radiomic features in magnetic resonance imaging of glioblastoma: test–retest and image registration analyses. Med Phys 47(9):4265–4280
9. Hosseini SA, Shiri I, Hajianfar G, Ghafarian P, Karam MB, Ay MR (2021) The impact of preprocessing on the PET-CT radiomics features in non-small cell lung cancer. Front Biomed Technol 8(4):261–272
10. Khodabakhshi Z, Gabrys H, Wallimann P, Guckenberger M, Andratschke N, Tanadini-Lang S Magnetic resonance imaging radiomic features stability in brain metastases: impact of image preprocessing, image-, and feature-level harmonization. https://doi.org/10.2139/ssrn.4671310
11. Hosseini SA, Shiri I, Hajianfar G, Bahadorzadeh B, Ghafarian P, Zaidi H, Ay MR (2022) Synergistic impact of motion and acquisition/reconstruction parameters on 18F-FDG PET radiomic features in non-small cell lung cancer: phantom and clinical studies. Med Phys 49(6):3783–3796
12. Chirra P, Leo P, Yim M, Bloch BN, Rastinehad AR, Purysko A et al (2019) Multisite evaluation of radiomic feature reproducibility and discriminability for identifying peripheral zone prostate tumors on MRI. J Med Imaging 6(2):024502
13. Hajianfar G, Hosseini SA, Amini M, Shiri I, Zaidi H (2022) MRI radiomic features harmonization: a multi-center phantom study. In: 2022 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC). https://doi.org/10.1109/NSS/MIC44845.2022.10399264
14. Li Z-C, Chen Y, Li Q, Sun Q, Luo R (2017) Automatic extraction of MRI radiomics features in glioblastoma multiforme: a

reproducibility evaluation. In: 2017 3rd IEEE International Conference on Cybernetics (CYBCONF). IEEE

15. Pinto MS, Paolella R, Billiet T, Van Dyck P, Guns P-J, Jeurissen B et al (2020) Harmonization of brain diffusion MRI: concepts and methods. Front Neurosci 14:396
16. Da-Ano R, Masson I, Lucia F, Doré M, Robin P, Alfieri J et al (2020) Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. Sci Rep 10(1):1–12
17. Da-Ano R, Visvikis D, Hatt M (2020) Harmonization strategies for multicenter radiomics investigations. Phys Med Biol 65(24):24TR02
18. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I (2019) Validation of a method to compensate multicenter effects affecting CT radiomics. Radiology 291(1):53–59
19. Saint Martin M-J, Orlhac F, Akl P, Khalid F, Nioche C, Buvat I et al (2021) A radiomics pipeline dedicated to Breast MRI: validation on a multi-scanner phantom study. MAGMA 34(3):355–366
20. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8(1):118–127
21. Fortin JP, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K et al (2017) Harmonization of multi-site diffusion tensor imaging data. Neuroimage 161:149–170
22. Fortin JP, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA et al (2018) Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167:104–120
23. Mills TC, Ortendahl DA, Hylton NM, Crooks LE, Carlson JW, Kaufman L (1987) Partial flip angle MR imaging. Radiology 162(2):531–539
24. Constable RT, Smith RC, Gore JC (1992) Signal-to-noise and contrast in fast spin echo (FSE) and inversion recovery FSE imaging. J Comput Assist Tomogr 16(1):41–47
25. Kellman P, Arai AE, McVeigh ER, Aletras AH (2002) Phase-sensitive inversion recovery for detecting myocardial infarction using gadolinium-delayed hyperenhancement. Magn Reson Med 47(2):372–383
26. Jackson EF (2015) Rider Phantom Mri. The Cancer Imaging Archive
27. Jackson EF, Barboriak DP, Bidau LM, Meyer CR (2009) Magnetic resonance assessment of response to therapy: tumor change measurement, truth data and error sources. Transl Oncol 2(4):211–215
28. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P et al (2013) The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 26(6):1045–1057
29. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-C, Pujol S et al (2012) 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging 30(9):1323–1341
30. Van Griethuysen JJ, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V et al (2017) Computational radiomics system to decode the radiographic phenotype. Cancer Res 77(21):e104–e1e7
31. Depeursinge A, Andrearczyk V, Whybra P, van Griethuysen J, Müller H, Schaer R, et al. Standardised convolutional filtering for radiomics. https://arxiv.org/abs/2006.05470. 2020.
32. Whybra P, Zwanenburg A, Andrearczyk V, Schaer R, Apte AP, Ayotte A et al (2024) The image biomarker standardization initiative: standardized convolutional filters for reproducible radiomics and enhanced clinical insights. Radiology 310(2):e231319
33. Lee G, Gommers R, Waselewski F, Wohlfahrt K, O'Leary A (2019) PyWavelets: a Python package for wavelet analysis. J Open Source Softw 4(36):1237
34. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJ, Andrearczyk V, Apte A et al (2020) The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiol 295(2):328–338
35. Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 15(2):155–163
36. Team RC (2012) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, p 2021
37. Demircioğlu A (2022) The effect of preprocessing filters on predictive performance in radiomics. Eur Radiol Exp 6(1):40
38. Moradmand H, Aghamiri SMR, Ghaderi R (2020) Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. J Appl Clin Med Phys 21(1):179–190
39. Zhao B (2021) Understanding sources of variation to improve the reproducibility of radiomics. Front Oncol 11:826
40. Williams S IV (2009) Using control charts for computer-aided diagnosis of brain images. Mathematics & Statistics UNM
41. Lee J, Steinmann A, Ding Y, Lee H, Owens C, Wang J et al (2021) Radiomics feature robustness as measured using an MRI phantom. Sci Rep 11(1):3973
42. Nan Y, Del Ser J, Walsh S, Schönlieb C, Roberts M, Selby I et al (2022) Data harmonisation for information fusion in digital healthcare: a state-of-the-art systematic review, meta-analysis and future research directions. Inform Fusion 82:99–122
43. Muhammed Sunnetci K, Ulukaya S, Alkan A (2022) Periodontal bone loss detection based on hybrid deep learning and machine learning models with a user-friendly application. Biomed Signal Process Control 77:103844
44. Sunnetci KM, Alkan A (2023) Biphasic majority voting-based comparative COVID-19 diagnosis using chest X-ray images. Expert Syst Appl 216:119430
45. Nanni L, Lumini A, Brahnam S (2010) Local binary patterns variants as texture descriptors for medical image analysis. Artif Intell Med 49(2):117–125

Ghasem Hajianfar is a dedicated PhD candidate at the University of Geneva. He studied the MSc of Medical Physics. His work focuses specifically on data analysis, machine/deep learning and radio(geno)mics analysis in medical images.



Seyyed Ali Hosseini is a highly skilled researcher and scientist with over a decade of experience in medical imaging, artificial intelligence, and data analysis. He is pursuing a Ph.D. in Neuroscience at McGill University, where he conducts cutting-edge research in the field of computational medical imaging and artificial intelligence, with a focus on Radiomics and deep learning. He also holds an M.Sc. in Medical Physics from Tehran University of Medical Science and a B.Sc. in Medical Imaging from Shahid Beheshti University of Medical Science. Seyyed Ali is proficient in computer programming languages such as Python and MATLAB, statistical analysis software like R, and machine learning frameworks including Keras and PyTorch. He has a solid background in data acquisition and pre/post-processing of medical image data, and his research interests include computational and analytical medical imaging, Radiomics, machine learning, deep learning, and neuroscience.

Sara Bagherieh is a motivated medical student and an enthusiastic early-career researcher with a solid background of medical research, as a RSNA & SIIM scholarship awardee. With 30+ published research papers, working as a research assistant at different departments and reviewing 30+ manuscripts for high-impact journals has fine-tuned Sara's research skills and equipped her with numerous tools to improve her understanding of research projects and help administer them.

Mehrdad Oveisi is a Lecturer in computer science at the University of British Columbia (UBC), Vancouver, Canada, where he teaches courses on artificial intelligence and machine learning. He is also an AI Specialist / Data Scientist at King's College London (KCL), UK. He has engaged in research in theoretical artificial intelligence, such as for his Ph.D. dissertation, as well as in the application of AI/ML in biomedical domains. He has also contributed to the development of several scientific applications at UBC, KCL, and Canada's Michael Smith Genome Sciences Centre (GSC).

Isaac Shiri received his Ph.D. in Medical Physics from the University of Geneva, Geneva, Switzerland. He currently leads the Artificial Intelligence in Cardiovascular Imaging group and laboratory at the Bern University Hospital, Bern University, Bern, Switzerland, and is interested in medical image analysis and artificial intelligence.

Habib Zaidi is Chief physicist and head of the PET Instrumentation & Neuroimaging Laboratory at Geneva University Hospital and full Professor at the medical school of the University of Geneva. He is also a Professor at the University of Groningen (Netherlands), the University of Southern Denmark (Denmark) and Óbuda University (Hungary). His research is supported by the Swiss National Foundation, the European Commission, private foundations and industry (Total 10M+ US$) and centres on hybrid imaging instrumentation (PET/CT and PET/MRI), computational modelling and radiation dosimetry and deep learning. He was guest editor for 14 special issues of peer-reviewed journals and serves and serves as founding Editor-in-Chief (scientific) of the *British Journal of Radiology (BJR)|Open*, Deputy Editor for *Medical Physics* and is on the editorial board of leading journals in medical physics and medical imaging. He has been elevated to the grade of fellow of the IEEE, AIMBE, AAPM, IOMP, AAIA and the BIR. His academic accomplishments in the area of quantitative PET imaging have been well recognized by his peers since he is a recipient of many awards and distinctions among which the prestigious (100'000$) *2010 Kuwait Prize of Applied Sciences* (known as the *Middle Eastern Nobel Prize*). Prof. Zaidi has been an invited speaker of over 160 keynote lectures and talks at an International level, has authored over 400+ peer-reviewed articles (h-index=76, >21'500+ citations) in prominent journals and is the editor of four textbooks.