applied sciences

MDPI

Article

Robustness Analysis of DCE-MRI-Derived Radiomic Features in Breast Masses: Assessing Quantization Levels and Segmentation Agreement

Carmelo Militello [1,*], Leonardo Rundo [2], Mariangela Dimarco [3,4], Alessia Orlando [3], Ildebrando D'Angelo [4,5], Vincenzo Conti [6] and Tommaso Vincenzo Bartolotta [3,5]

1 Institute of Molecular Bioimaging and Physiology, Italian National Research Council (IBFM-CNR), 90015 Cefalù, PA, Italy
2 Department of Information and Electrical Engineering and Applied Mathematics (DIEM), University of Salerno, 84084 Fisciano, SA, Italy; lrundo@unisa.it
3 Section of Radiology—Department of Biomedicine, Neuroscience and Advanced Diagnostics (BiND), University Hospital "Paolo Giaccone", 90127 Palermo, PA, Italy; maridimarco33@gmail.com (M.D.); orlandoalessiamed@gmail.com (A.O.); tommasovincenzo.bartolotta@unipa.it (T.V.B.)
4 Breast Unit, Fondazione Istituto "G. Giglio", 90015 Cefalù, PA, Italy; ildebrando.dangelo@hsrgiglio.it
5 Department of Radiology, Fondazione Istituto "G. Giglio", 90015 Cefalù, PA, Italy
6 Faculty of Engineering and Architecture, University of Enna KORE, 94100 Enna, EN, Italy; vincenzo.conti@unikore.it
* Correspondence: carmelo.militello@cnr.it

Featured Application: The use of highly robust radiomic features is fundamental to reduce intrinsic dependencies and to provide reliable predictive models. This work presents a study on breast tumor DCE-MRI considering the radiomic feature robustness against the quantization settings and segmentation methods.

Abstract: Machine learning models based on radiomic features allow us to obtain biomarkers that are capable of modeling the disease and that are able to support the clinical routine. Recent studies have shown that it is fundamental that the computed features are robust and reproducible. Although several initiatives to standardize the definition and extraction process of biomarkers are ongoing, there is a lack of comprehensive guidelines. Therefore, no standardized procedures are available for ROI selection, feature extraction, and processing, with the risk of undermining the effective use of radiomic models in clinical routine. In this study, we aim to assess the impact that the different segmentation methods and the quantization level (defined by means of the number of bins used in the feature-extraction phase) may have on the robustness of the radiomic features. In particular, the robustness of texture features extracted by PyRadiomics, and belonging to five categories—GLCM, GLRLM, GLSZM, GLDM, and NGTDM—was evaluated using the intra-class correlation coefficient (ICC) and mean differences between segmentation raters. In addition to the robustness of each single feature, an overall index for each feature category was quantified. The analysis showed that the level of quantization (i.e., the 'bincount' parameter) plays a key role in defining robust features: in fact, in our study focused on a dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) dataset of 111 breast masses, sets with cardinality varying between 34 and 43 robust features were obtained with 'binCount' values equal to 256 and 32, respectively. Moreover, both manual segmentation methods demonstrated good reliability and agreement, while automated segmentation achieved lower ICC values. Considering the dependence on the quantization level, taking into account only the *intersection subset* among all the values of 'binCount' could be the best selection strategy. Among radiomic feature categories, GLCM, GLRLM, and GLDM showed the best overall robustness with varying segmentation methods.

Keywords: robustness analysis; radiomic features; quantization levels; segmentation method agreement; DCE-MRI; breast tumors

check for updates

Citation: Militello, C.; Rundo, L.; Dimarco, M.; Orlando, A.; D'Angelo, I.; Conti, V.; Bartolotta, T.V. Robustness Analysis of DCE-MRI-Derived Radiomic Features in Breast Masses: Assessing Quantization Levels and Segmentation Agreement. *Appl. Sci.* **2022**, *12*, 5512. https://doi.org/10.3390/app12115512

Academic Editor: Fabio La Foresta

Received: 26 April 2022
Accepted: 27 May 2022
Published: 29 May 2022

1. Introduction

Radiomics techniques, aimed at analyzing a large amount of minable features extracted from medical images, have shown great potential in different clinical areas [1]. Such a large amount of quantitative radiomic features conveys more information than the qualitative patterns observed by the radiologists' naked eye on the images. In particular, radiomic image-based signatures can be associated with clinical outcomes (i.e., biomarkers) [2,3], thereby improving clinical decision-making tasks [4]. Therefore, the development of robust biomarkers is an essential requirement which could accelerate their usefulness and, consequently, their incorporation into clinical practice [5], providing reliable diagnostic and prognostic biomarkers for precision medicine [6].

It is well known that radiomic features might be affected by several mathematical definitions, and the proliferation of toolboxes does not help this aspect. Thus, standardization initiatives have been carried out by the scientific community in order to deal with the lack of reproducibility and validation of radiomics studies. In particular, the Image Biomarker Standardization Initiative (IBSI) [7] attempts to provide guidelines about the biomarkers' definition. Unfortunately, it is a very complicated pipeline, where every single step (e.g., image acquisition, reconstruction, segmentation, features extraction) must be tackled with caution to ensure reproducibility.

Nevertheless, there are still no comprehensive and clear guidelines to obtain radiomic features that are not only reproducible but also robust. For that reason, an accurate and careful analysis of the robustness of the radiomic features is mandatory to define robust and clinically relevant biomarkers. Due to the wider diffusion and use of radiomics, in recent years, the study of methodologies aimed at improving the reproducibility and robustness of these tools has become a research topic faced, by the scientific community, from different points of view and in several clinical application domains.

There are many *'sources of variability'* that should be considered: many literature works have analyzed and assessed the impact on the robustness of radiomic features related to *intrinsic factors*, such as (*i*) imaging protocol [8,9] and (*ii*) magnetic field strength [10], or due to *extrinsic factors*, such as (*iii*) pre-processing and enhancement techniques [11,12], (*iv*) perturbation on the region of interest (ROI) due to the inter-operator dependence [13,14], and segmentation methods [15,16], but also to (*v*) post-processing techniques, which are certainly modified in original data (e.g., harmonization procedures) [17].

The main goal of this work is to analyze the dependence of robustness on the quantization level used during feature extraction and the reliability of the segmentation of breast masses on dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) by relying upon two segmentation methods (both automated and manual) obtained by two raters. The proposed analysis exploits hand-crafted features and traditional machine learning approaches, which are still popular in radiomics studies today, aimed at establishing predictive models that are reliable for clinical applications.

The main contributions of this study are:

- an evaluation of the robustness of radiomic features, in terms of ICC, as a function of quantization level in PyRadiomics;
- an assessment of segmentation by comparing three different raters: a automated approach and the manual segmentation of two human operators, to evaluate the features reproducibility;
- a definition of a robustness scale for individual features as a function of ICC and mean standard deviation of the feature differences;
- a definition and quantification of an overall robustness metric for each radiomic category (i.e., GLCM, GLRLM, GLSZM, GLDM, and NGTDM).

The remainder of this work is organized as follows. Section 2 describes the conducted study aimed at dealing with the robustness evaluation. Section 3 illustrates the experimental results in terms of dependence on the quantization level, as well as the segmentation assessment. Finally, the discussion and conclusions are provided in Section 4.

2. Materials and Methods

This section describes the characteristics of the DCE-MRI dataset used and the analysis carried out to assess the robustness of the features, relative to their dependence on the quantization level, as well as on their reproducibility concerning the segmentation method.

The radiomic feature robustness analysis and quantification method were implemented entirely in the MATLAB R2019b (64-bit version) environment MathWorks, Natick, MA, USA). PyRadiomics was used for the extraction of radiomic features, an open-source Python package developed for the standardization of radiomic feature extraction [18]. In particular, we used PyRadiomics version 2.0 and Python 3.7.5.

2.1. Dataset Description

A total of 111 breast masses from DCE-MRI exams were considered in this study, for a total of 1231 MR slices. Table 1 provides some relevant characteristics concerning the MR imaging protocol.

Table 1. Some relevant characteristics concerning the DCE-MRI imaging protocol.

Protocol Characteristic	Value
series	Ax VIBRANT mphase
MR scanner	GE Signa HDxt
magnetic field	1.5 Tesla
repetition time	(37.72–56.92) ms
echo time	(17.64–26.80) ms
flip angle	10°
matrix size	512 × 512 pixels
slice thickness	(2–3) mm
spacing between slices	(1–1.5) mm
pixel spacing	(0.6875–0.7422) mm

2.2. Breast Tumor Segmentation

ROIs containing breast masses were segmented using two different segmentation methods:

- *Automated delineation*. This is a computer-assisted method based on the spatial fuzzy c-means (sFCM) clustering algorithm [19,20]. The sFCM algorithm, compared to the traditional FCM, takes into account the spatial relationship among neighboring pixels, making it less sensitive to noise and other imaging artifacts. This approach has been previously implemented and validated in [21,22];
- *Manual free-hand delineation*. In order to quantify the inter-operator dependence of the robustness of the features, two delineations were performed by two radiologists with more than 5 years of experience with breast MRI, in consensus with a consultant breast radiologist (with more than 30 years of experience with breast imaging).

Segmentations—in terms of original MR slices containing the breast mass and the corresponding masks—were converted from DICOM to the Neuroimaging Informatics Technology Initiative (NIfTI) format [7], to be used successively as input to PyRadiomics [18] for feature extraction. Both automated and manual free-hand delineations were performed by using MATLAB-coded custom segmentation tools. The size of the tumor is included among the exclusion criteria [21]: the masses with maximum diameters lower than 5 mm were excluded. Considering that the images have a pixel spacing varying between 0.6875 and 0.7422 mm, this means that in the limiting case, the ROI has at least 7–8 voxels (along one dimension), which is sufficient to calculate the features. Figure 1 shows two segmentation results obtained through two different segmentation methods.

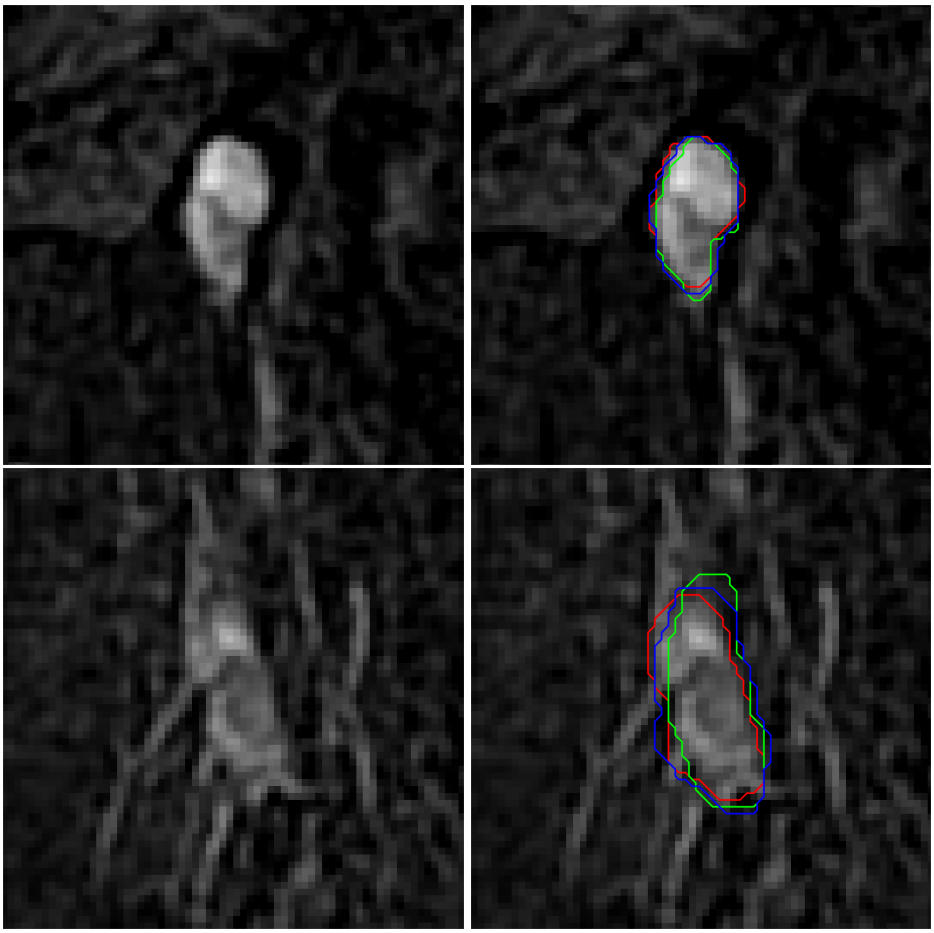

Figure 1. Examples related to a benign (on the top) and a malignant (on the bottom) breast mass. For each example, the DCE-MRI image area is shown without (on the left) and with (on the right) delineation contours. Segmentation results yielded by the automated delineation (red contour) and manual delineations—performed by the first (green contour) and the second (blue contour) radiologist—are compared. All figures are shown with a 2.5× magnification factor.

2.3. Radiomic Feature Extraction

The DCE-MRI images analyzed represent a homogeneous dataset in terms of spatial resolution along the (x, y) plane and slice thickness along the z axis. For this reason, the extraction of the features was performed without any resampling to avoid interpolation artifacts. Radiomic features were extracted from the 3D ROIs delineated in the previous step. Five texture-feature categories were extracted and considered in this study:

- Gray-level co-occurrence matrix (GLCM) [23,24]: spatial relationship between pixels in a specific direction, highlighting the properties of uniformity, homogeneity, randomness, and linear dependencies;
- Gray-level run length matrix (GLRLM) [25]: texture in specific direction, where fine texture has shorter runs while coarse texture presents more long runs with different intensity values;
- Gray-level size zone matrix (GLSZM) [26]: regional intensity variations or the distribution of homogeneity regions;
- Gray-level dependence matrix (GLDM) [27]: quantifies gray-level dependencies;
- Neighboring gray tone difference matrix (NGTDM) [28]: spatial relationship among three or more pixels, closely approaching the human perception of the image.

Conversely, shape-based and first-order features were not considered in the study because they are independent of the quantization level.

It is worth noting that texture features, such as the GLCM features (also known as Haralick's features [23,24]), are calculated from the co-occurrence matrix that is based only on the gray-level values. Therefore, the ROI shape does not play any role, because co-occurrence matrices consider only the pairs of gray-level values composed of two adjacent or neighboring voxels [29].

2.4. Statistical Analysis

Radiomic features were extracted considering different quantization levels (i.e., number of bins equal to 8, 16, 32, 64, 128, 256). The ICC analysis takes into account the extracted features and allows us to determine which are more robust as the number of bins varies. ICC analysis was applied to establish the minimum number of 'bins' that maximizes the set of robust features. In this study, we considered the two-way random-effects model (or mixed-effects), consistency, and the single rater/measurement ICC, named ICC(3,1) [30]. Let k be the number of raters/measurements; the two-way random-effects model (or mixed-effects), consistency, and the single rater/measurement, ICC(3,1)—defined in Equation (1)—was used.

$$ICC(3,1) = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E}, \quad (1)$$

where MS_R and MS_E are the mean squares for rows and for errors, respectively.

2.4.1. Quantization-Level Analysis

The quantization of images, in terms of the rebinning of the gray levels prior to feature computation, has a two-fold goal: (*i*) noise reduction and (*ii*) avoidance of sparse matrices (possibly resulting in unsuitable and poorly robust features for predictive modeling). The Image Biomarker Standardization Initiative (IBSI) [7] explicitly suggests quantization to optimize and improve the development of radiomics models.

The 'binCount' parameter is used by PyRadiomics to determine the image quantization settings (i.e. the number of bins) in the radiomic feature-extraction phase. The optimal value for 'binCount' was chosen so as to maximize the number of robust features (in terms of ICC). As a matter of fact, this choice allowed us to carefully assess the quantization settings, thereby avoiding an arbitrary selection of the number of bins. With more details, to evaluate the robustness as the quantization level changes, the $ICC_{quantLevel}$ was calculated for each feature, considering the set composed of all the segmentations (i.e., $Segm_{automated}$, $Segm_{man1}$, $Segm_{man2}$) and a specific number of bins, as indicated in Equation (2).

$$ICC_{quantLevel} = ICC(Segm_{automated}, Segm_{man1}, Segm_{man2}), \quad (2)$$

2.4.2. Agreement of Segmentation Methods

Each breast mass was segmented *via* two different delineation approaches: (*i*) manual segmentation, performed by two distinct radiologists with 5 years of experience; (*ii*) an automated method based on spatial FCM clustering that was already tested and validated in [21]. The inter-observer agreement of the two segmentation methods (automated and manual) was quantified by calculating the ICC using a two-way random-effects model and the mean differences of the two observers. Regarding the manual segmentation, two distinct series of delineations were analyzed and independently performed by two different raters [31]. For this reason, three ICC series were computed, according to Equations (3)–(5):

$$ICC_{automated-manual1} = ICC(Segm_{automated}, Segm_{man1}), \quad (3)$$

$$ICC_{automated-manual2} = ICC(Segm_{automated}, Segm_{man1}), \quad (4)$$

$$ICC_{manual1-manual2} = ICC(Segm_{man1}, Segm_{man2}), \quad (5)$$

After determining the features showing excellent robustness, we aimed to identify the most relevant features for the analysis at hand by evaluating the dependence on the segmentation method. To this aim, we considered $binCount = 32$, which guarantees the greatest number of features (43) with excellent robustness.

The Shapiro–Wilk test [32] was used to assess the normality of the 129 distributions of the differences (obtained considering 43 features and three segmentations) among the segmentation raters, considering a 0.05 confidence interval. All distributions passed the normality test, obtaining p-values $\ll 0.0001$.

To evaluate the robustness of the MRI radiomic features, a six-level scale (ranging from 0 to 5) was defined, based on a combination of the ICCs coefficient and the standard deviation (SD) of the mean percentage differences between the three raters (i.e., auto, man1, man2), according the conditions in Table 2. Percentage differences were evaluated according to Equation (6).

$$\text{PercentageDifference}(Segm_i, Segm_j) = \frac{abs(Segm_i - Segm_j)}{mean(Segm_i, Segm_j)} \times 100 \quad (6)$$

$\forall i, j \in \{auto, man1, man2\}$.

Table 2. Conditions to evaluate the robustness of each radiomic feature category.

Score (Robustness)	ICC Condition	SD Condition
5 (very high)	≥90%	≤10%
4 (high)	≥85%	≤20%
3 (medium)	≥80%	≤30%
2 (limited)	≥75%	≤40%
1 (low)	≥70%	≤100%
0 (very low)	<70%	>100%

3. Results

3.1. Quantization Setting Dependence Results

Quantization setting dependence was evaluated by means of ICC, as indicated in Section 2.4.1. In order to consider only radiomic features with excellent robustness, the cut-off value was set to 0.9. Figures 2–6 show all the features—divided by category (i.e., GLCM, GLRLM, GLSZM, GLDM, NGTDM)—that obtained $ICC \geq 0.9$ in at least one quantization level.

	Quantization Level '*binCount*' (ICC>0.9)					
feature Name (GLCM)	8	16	32	64	128	256
Autocorrelation					X	
Contrast	X	X	X	X	X	X
DifferenceAverage	X	X	X	X	X	X
DifferenceEntropy	X	X	X	X	X	
DifferenceVariance	X		X			
Id	X	X	X	X	X	X
Idm	X	X	X	X	X	X
Idmn	X	X	X	X	X	X
Idn	X	X	X	X	X	X
Imc1	X		X	X	X	X
Imc2			X	X	X	X
InverseVariance	X	X	X	X	X	X
JointEnergy	X	X				
JointEntropy				X	X	X
MCC						X
MaximumProbability	X	X	X			
SumEntropy						X

Figure 2. Dependence of the quantization level of GLCM radiomic features, considering different values (i.e., 8, 16, 32, 64, 128, and 256) of the 'bincount' PyRadiomics parameter. The features *ClusterProminence*, *ClusterShade*, *ClusterTendency*, *Correlation*, *JointAverage*, and *SumAverage* were discarded, as the ICC did not overcome the cutoff for any 'binCount' setting.

	Quantization Level '*binCount*' (ICC>0.9)					
feature Name (GLRLM)	**8**	**16**	**32**	**64**	**128**	**256**
GrayLevelNonUniformity	X	X	X	X	X	X
LongRunEmphasis	X	X	X	X	X	X
LongRunHighGrayLevelEmphasis	X	X	X		X	
RunEntropy	X					
RunLengthNonUniformity	X	X	X	X	X	X
RunLengthNonUniformityNormalized	X	X	X	X	X	X
RunPercentage	X	X	X	X	X	X
RunVariance	X	X	X	X	X	
ShortRunEmphasis	X	X	X	X	X	X

Figure 3. Dependence of the quantization level of GLRLM radiomic features, considering different values (i.e., 8, 16, 32, 64, 128, and 256) of the 'binCount' PyRadiomics parameter. The features *SumSquares*, *GrayLevelNonUniformityNormalized*, *GrayLevelVariance*, *HighGrayLevelRunEmphasis*, *LongRunLowGrayLevelEmphasis*, *LowGrayLevelRunEmphasis*, *ShortRunHighGrayLevelEmphasis*, and *ShortRunLowGrayLevelEmphasis* were discarded, as the ICC did not overcome the cutoff for any 'binCount' setting.

	Quantization Level '*binCount*' (ICC>0.9)					
feature Name (GLSZM)	**8**	**16**	**32**	**64**	**128**	**256**
GrayLevelNonUniformity	X	X	X	X	X	X
LargeAreaEmphasis	X	X	X	X		
LargeAreaHighGrayLevelEmphasis	X	X	X	X	X	X
SizeZoneNonUniformity	X	X	X	X	X	X
SizeZoneNonUniformityNormalized			X	X	X	X
SmallAreaEmphasis			X	X	X	X
ZoneEntropy	X	X	X	X	X	X
ZonePercentage	X	X	X	X	X	X
ZoneVariance	X	X	X	X		

Figure 4. Dependence of the quantization level of GLSZM radiomic features, considering different values (i.e., 8, 16, 32, 64, 128, and 256) of the 'binCount' PyRadiomics parameter. The features *GrayLevelNonUniformityNormalized*, *GrayLevelVariance*, *HighGrayLevelZoneEmphasis*, *LargeAreaLowGrayLevelEmphasis*, *LowGrayLevelZoneEmphasis*, *SmallAreaHighGrayLevelEmphasis*, and *SmallAreaLowGrayLevelEmphasis* were discarded, as the ICC did not overcome the cutoff for any 'binCount' setting.

	Quantization Level '*binCount*' (ICC>0.9)					
feature Name (GLDM)	**8**	**16**	**32**	**64**	**128**	**256**
DependenceEntropy	X	X	X	X	X	X
DependenceNonUniformity	X	X	X	X	X	X
DependenceNonUniformityNormalized	X	X	X	X	X	X
DependenceVariance	X	X	X	X		
GrayLevelNonUniformity	X	X	X	X	X	X
LargeDependenceEmphasis	X	X	X	X	X	
LargeDependenceHighGrayLevelEmphasis	X	X	X	X	X	X
SmallDependenceEmphasis	X	X	X	X	X	X
SmallDependenceHighGrayLevelEmphasis	X	X	X			

Figure 5. Dependence of the quantization level of GLDM radiomic features, considering different values (i.e., 8, 16, 32, 64, 128, and 256) of the 'binCount' PyRadiomics parameter. The features *GrayLevelVariance*, *HighGrayLevelEmphasis*, *LargeDependenceLowGrayLevelEmphasis*, *LowGrayLevelEmphasis*, and *SmallDependenceLowGrayLevelEmphasis* were discarded, as the ICC did not overcome the cutoff for any 'binCount' setting.

	Quantization Level '*binCount*' (ICC>0.9)					
feature Name (NGTDM)	**8**	**16**	**32**	**64**	**128**	**256**
Busyness	X	X	X	X	X	X
Coarseness	X	X	X	X	X	X
Complexity	X	X	X	X		
Contrast	X		X			
Strength	X	X	X	X	X	X

Figure 6. Dependence of the quantization level of NGTDM radiomic features, considering different values (i.e., 8, 16, 32, 64, 128, and 256) of the 'binCount' PyRadiomics parameter.

Table 3 summarizes the total of the robust features obtained for each quantization level.

Table 3. Summary of robust features (ICC > 0.9) considering different quantization levels. Bold values represent the setting guaranteeing the maximum number of robust features.

Quantization Level ('binCount')	Robust Features Number	Robust Features Percentage (%)
8	42	85.7
16	38	77.6
32	**43**	**87.8**
64	39	79.6
128	37	75.5
256	34	69.4

3.2. Segmentation Method Dependence Results

The robustness of the MRI radiomic features was evaluated using a six-level scale (going from 0 to 5) and based on a combination of ICCs and the SD of the mean percentage differences between the three raters (i.e., auto, man1, man2). Percentage differences were evaluated according to Equation (6). Each feature was evaluated considering the ICC and SD. The score—from 0 (very high robustness) to 5 (very low robustness)—was assigned according to the conditions reported in Table 2.

The values obtained are illustrated, for each of the five feature categories, in the following figures (Figures 7–11).

	robusteness Index		
GLCM	Auto vs Man1	Auto vs Man2	Man1 vs Man2
Contrast	2	1	3
DifferenceAverage	2	2	4
DifferenceEntropy	3	3	5
DifferenceVariance	1	1	3
Id	3	3	4
Idm	3	3	4
Idmn	4	4	5
Idn	4	4	5
Imc1	2	1	3
Imc2	2	2	5
InverseVariance	3	3	4
MaximumProbability	0	0	3

Figure 7. Robustness of the GLCM radiomic features obtained on DCE-MRI images using automated segmentation against two manual segmentations from two independent readers.

	robusteness Index		
GLRLM	Auto vs Man1	Auto vs Man2	Man1 vs Man2
GrayLevelNonUniformity	2	2	3
LongRunEmphasis	4	4	5
LongRunHighGrayLevelEmphasis	0	0	3
RunLengthNonUniformity	2	2	3
RunLengthNonUniformityNormalized	5	5	5
RunPercentage	5	5	5
RunVariance	2	2	3
ShortRunEmphasis	4	5	5

Figure 8. Robustness of the GLRLM radiomic features obtained on DCE-MRI images using automated segmentation against two manual segmentations from two independent readers.

	robusteness Index		
GLSZM	**Auto vs Man1**	**Auto vs Man2**	**Man1 vs Man2**
GrayLevelNonUniformity	3	2	2
LargeAreaEmphasis	2	2	3
LargeAreaHighGrayLevelEmphasis	2	2	2
SizeZoneNonUniformity	2	2	2
SizeZoneNonUniformityNormalized	1	1	2
SmallAreaEmphasis	1	1	3
ZoneEntropy	3	2	5
ZonePercentage	3	2	3
ZoneVariance	2	2	2

Figure 9. Robustness of the GLSZM radiomic features obtained on DCE-MRI images using automated segmentation against two manual segmentations from two independent readers.

	robusteness Index		
GLDM	**Auto vs Man1**	**Auto vs Man2**	**Man1 vs Man2**
DependenceEntropy	3	3	5
DependenceNonUniformity	2	2	3
DependenceNonUniformityNormalized	2	2	4
DependenceVariance	2	2	3
GrayLevelNonUniformity	2	2	3
LargeDependenceEmphasis	3	3	3
LargeDependenceHighGrayLevelEmphasis	1	2	3
SmallDependenceEmphasis	3	3	3
SmallDependenceHighGrayLevelEmphasis	1	0	3

Figure 10. Robustness of the GLDM radiomic features obtained on DCE-MRI images using automated segmentation against two manual segmentations from two independent readers.

	robusteness Index		
NGTDM	**Auto vs Man1**	**Auto vs Man2**	**Man1 vs Man2**
Busyness	2	2	2
Coarseness	2	0	2
Complexity	2	1	2
Contrast	1	0	2
Strength	2	1	2

Figure 11. Robustness of the NGTDM radiomic features obtained on DCE-MRI images using automated segmentation against two manual segmentation.

Starting from the previously defined scale (see Table 2), an overall robustness value—defined according to Equation (7)—was computed for each radiomic category Φ to summarize the robustness by feature category across the segmentation methods. This formula gives values between 0 (lower robustness) and 1 (higher robustness). Results are shown in Table 4. Instead, for each category of features, Table 5 displays the percentage of robust features (which have obtained a roughness index $\geq$ 3) as the segmentation methodology varies.

$$Robustness(\Phi) = \frac{\sum_{i \in \Phi} S_{\phi_i}}{maxValue \times |\Phi|} \quad (7)$$

where:

- Φ is the radiomic category ($\Phi \in \{GLCM, GLRLM, GLSZM, GLDM, NGTDM\}$);
- S_{ϕ_i} represents the score on the scale described above (see Table 2) for the robustness of each feature i;
- $maxValue$ represents the maximum possible value (i.e., 5);
- $|\cdot|$ represents the cardinality of a given radiomic category Φ.

Table 4. Overall feature category robustness across the three segmentation raters.

Category	Auto vs. Man1	Auto vs. Man2	Man1 vs. Man2
GLCM	0.48	0.45	0.8
GLRLM	0.6	0.63	0.8
GLSZM	0.42	0.36	0.53
GLDM	0.42	0.42	0.67
NGTDM	0.36	0.16	0.4

Table 5. Percentage of features with robustness index ≥ 3.

Category	Auto vs. Man1	Auto vs. Man2	Man1 vs. Man2
GLCM	50%	50%	100%
GLRLM	50%	50%	100%
GLSZM	33.3%	0%	44.4%
GLDM	33.3%	33.3%	100%
NGTDM	0%	0%	0%

4. Discussion and Conclusions

The aim of this work was to acquire further insights into the dependence of radiomic features' robustness on the quantization level used during feature extraction, and the reproducibility of features in relation to segmented ROIs of breast masses. In particular, first of all, in order to quantify the robustness and to provide a score—for each radiomic feature and for each category (i.e., GLCM, GLRLM, GLSZM, GLDM, NGTDM)—a correlation analysis based on the ICC was carried out to identify the features that are the least dependent on the level of quantization (i.e., the 'binCount' parameter). After determining the best value for 'binCount' and considering the subset containing only the robust features, an additional analysis was carried out in order to assess the reproducibility by comparing, relying upon the ICC, three different segmentations: two that were independently obtained through manual delineation by two radiologists, and one that was obtained using a validated automated segmentation approach [21,33] based on spatial FCM clustering. This second analysis led to the quantification of not only a robustness index for each feature, but also to the definition and quantification of an overall robustness index for each category of features.

Among the GLCM features, *Contrast*, *DifferenceAverage*, *DifferenceVariance*, *Imc1*, *Imc2*, and *MaximumProbability* seem to have a poor robustness, related to the segmentation method. Even if, in the comparison of manual1 vs. manual2, their robustness index is good (surely due to a higher accordance between the two manual segmentations), in the comparison with the automated method, the *Imc1* is lower. This denotes a high dependence on the ROI. The category has a medium-to-high overall robustness index (0.48–0.8). Among the GLRLM features, *LongRunHighGrayLevelEmphasis* obtains the worst score in the comparison between automated and manual segmentations, followed by *GrayLevelNonUniformity*, *RunLengthNonUniformity*, and *RunVariance*. On the other hand, the other features present a very good score in all comparisons, and the overall category score is high (0.6–0.8). Nearly all GLSZM features have a medium/low index that brings the category a very low overall index, varying in the range 0.36–0.53, depending on the segmentation. GLDM almost all have a medium other score, except for *LargeDependenceHighGrayLevelEmphasis* and *SmallDependenceHighGrayLevelEmphasis*, showing an overall score in the range 0.42–0.67. Finally, all features belonging to the NGTDM category show very low robustness indexes and, consequently, the category obtains an overall score in the range 0.16–0.4.

Considering the dependence on the quantization level, taking into account only the *intersection subset* among all the values of 'binCount' could be the best selection strategy. Among radiomic categories, GLCM, GLRLM, and GLDM showed the best overall robustness against variations due to the segmentation method.

A fair comparison against analogous literature works is not possible, because each one does not necessarily refer to the same disease and, consequently, does not use the same

data. As a matter of fact, rather than comparing results, we believe it is more appropriate to summarize the results obtained (see Table 6), in order to provide an overview of the literature and the parameters for analyzing the robustness of the characteristics.

The main focus of our work was on the development of a reliable system in terms of robust radiomic features according to the quantization settings and segmentation approaches. The adopted methods are not computationally expensive since traditional statistical analyses (mostly based on the ICC) are applied after the radiomic feature extraction. This assessment allowed us to propose a lightweight system based on classic machine learning techniques. Moreover, the type of analysis performed disregards variations in the data that could affect the stability of the system. In fact, the analysis performed on the features is computed from the segmented ROI and by quantifying the correlation (*via* ICC) as the quantization level and segmentation approach change.

Presently, clinical decision support systems (CDSSs) are becoming increasingly prevalent in clinical routines, and are able to assist the work of clinicians [34,35]. In the near future, radiomics will certainly represent a tool that clinicians will rely on. An analysis and quantification approach, such as the one proposed here, inserted upstream in the analysis and modeling pipeline will enable even more reliable radiomics tools. Our study can represent an improvement over a traditional radiomic analysis pipeline that might be affected by parameter choices. In fact, the used techniques, inserted upstream of the modeling pipeline, allows for the definition of predictive models based only on features (or feature categories) that are robust against levels of quantization and segmentation methods. Certainly, considering that clinical scenarios are among the most critical, any tool that takes advantage of ICT must be carefully validated before being integrated into clinical practice.

Furthermore, as future developments, it would be very interesting to validate the insights gained from this study on a more extensive dataset in order to assess the repeatability of the analysis made and its ability to generalize. A pre-analysis and feature calibration phase—such as the one proposed in this study—is absolutely essential to have an initial set of non-redundant and robust features. In fact, every radiomic study consists of several phases [3]: reducing the uncertainty on the input features allows to improve the repeatability and robustness of the study.

Table 6. Overview of the state-of-the-art results, along with the considered settings, to assess the robustness of radiomic features.

Related Work	Imaging, Disease (# Samples)	Dependence	Extraction Tool (# Features)	Main Findings
Shafiq-Ul-Hassan et al. [8]	CT, phantom (8 by 8 CT scanners)	voxel size; gray levels	in-house tool (213)	In total, 150 features are reproducible across voxel sizes
Escudero Sanchez et al. [9]	CT, liver cancer (43)	slice thickness	PyRadiomics (107)	In total, 75–90% of features are highly robust
Whitney et al. [10]	MRI (DCE-MRI), breast cancer (612)	magnetic field strength	PyRadiomics (38)	In total, 5 features are robust across field strength
Scalco et al. [11]	MR (T2w)/prostate cancer/14	image signal normalization	PyRadiomics (91)	In total, 60% of features have a poor reproducibility
De Farias et al. [12]	CT, various lesion types (10,000 slices) + validation on NSCLC (17,938 slices)	super-resolution	PyRadiomics (75)	In total, 10 texture features have excellent robustness
Zwanenburg et al. [13]	CT, NSCLC (31) + HNSCC (19)	image perturbation	N.A. (4032)	In total, 2310 (57.3%) NSCLC features are robust; 582 (14.4%) HNSCC features are robust; 454 (11.3%) features are robust in both cohorts
Mottola et al. [14]	CT, RCC (98) + CK (93)	image resampling and perturbation	in-house tool (32)	In total, 94.6% and 87.7% of features achieve the best reproducibility in RCC and CK
Tixier et al. [15]	MRI (FLAIR, T1w), glioblastoma (98)	segmentation method	PyRadiomics (108)	IH and GLCM features are the most robust; GLSZM features have a mixed robustness
Granzier et al. [16]	MR (T1w), breast cancer (102)	inter-observer segmentation variability	RadiomiX (1328) + PyRadiomics (833)	In total, 41.6% (552/1328) and 32.8% (273/833) of all RadiomiX and Pyradiomics features, respectively, are robust
Le et al. [17]	CTA, culprit lesions in carotid arteries (41)	inter-observer segmentation variability; image configurations	PyRadiomics (93)	In total, 55.9% (52/93) of features have excellent robustness; 33.3% (31/93) of features have moderate robustness; 10.8% (10/93) of features have poor robustness
Proposed Work	MRI (DCE-MRI), breast cancer (111)	quantization level; inter-observer segmentation variability	PyRadiomics (49)	In total, 87.8% (43/49) of features were robust (ICC $\geq$ 0.9) with binCount = 32; GLCM and GLRLM features have high robustness; GLDM features have moderate robustness

Author Contributions: Conceptualization, C.M. and L.R.; methodology, C.M. and L.R.; software, C.M.; validation, C.M. and L.R.; formal analysis, C.M. and L.R.; investigation, C.M.; resources, M.D., A.O., and I.D.; data curation, M.D., A.O., and I.D.; writing—original draft preparation, C.M. and L.R.; writing—review and editing, M.D., A.O., V.C., and T.V.B.; visualization, C.M. and L.R.; supervision, V.C. and T.V.B.; project administration, T.V.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of "Azienda Ospedaliera Universitaria Policlinico P.Giaccone" of Palermo, Italy (protocol code n.1/2020-15/01/2020).

Informed Consent Statement: Retrospective data collection was approved by the Ethics Committee. The requirement for evidence of informed consent was waived because of the retrospective nature of our study.

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

Abbreviations

The following abbreviations are used in this manuscript:

CK	Contralateral normal kidney
CT	Computed tomography
CTA	Computed tomography angiography
DCE-MRI	Dynamic contrast-enhanced magnetic resonance imaging
EVK	Enhancement variance kinetics
FCM	Fuzzy c-means
FO	First-order
GLCM	Gray-level co-occurrence matrix
GLDM	Gray-level dependence matrix
GLRLM	Gray-level run length matrix
GLSZM	Gray-level size zone matrix
HNSCC	Head and neck squamous cell carcinoma
IBSI	Image Biomarker Standardization Initiative
ICC	Intra-class correlation coefficient
IH	Intensity histogram
IVH	Intensity–volume histogram
KCA	Kinetic curve assessment
MR	Magnetic resonance
NGTDM	Neighboring gray tone difference matrix
NIfTI	Neuroimaging Informatics Technology Initiative
NSCLC	Non-small cell lung cancer
RCC	Renal cell carcinoma
ROI	Region of interest
sFCM	Spatial fuzzy c-means
T1w	T1 weighed
T2w	T2 weighed

References

1. Gillies, R.; Kinahan, P.; Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* **2016**, *278*, 563–577. [CrossRef] [PubMed]
2. Rundo, L.; Militello, C.; Vitabile, S.; Russo, G.; Sala, E.; Gilardi, M.C. A survey on nature-inspired medical image analysis: A step further in biomedical data integration. *Fundam. Inform.* **2020**, *171*, 345–365. [CrossRef]
3. Militello, C.; Rundo, L.; Dimarco, M.; Orlando, A.; Woitek, R.; D'Angelo, I.; Russo, G.; Bartolotta, T.V. 3D DCE-MRI Radiomic Analysis for Malignant Lesion Prediction in Breast Cancer Patients. *Acad. Radiol.* **2022**, *29*, 830–840. [CrossRef] [PubMed]
4. Rundo, L.; Pirrone, R.; Vitabile, S.; Sala, E.; Gambino, O. Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine. *J. Biomed. Inform.* **2020**, *108*, 103479. [CrossRef] [PubMed]

5. Lambin, P.; Leijenaar, R.; Deist, T.M.; Peerlings, J.; de Jong, E.; van Timmeren, J.; Sanduleanu, S.; Larue, R.; Even, A.; Jochems, A.; et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* **2017**, *14*, 749–762. [CrossRef] [PubMed]
6. Fornacon-Wood, I.; Mistry, H.; Ackermann, C.J.; Blackhall, F.; McPartlin, A.; Faivre-Finn, C.; Price, G.J.; O'Connor, J. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. *Eur. Radiol.* **2020**, *30*, 6241–6250. [CrossRef] [PubMed]
7. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.W.L.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* **2020**, *295*, 328–338. [CrossRef]
8. Shafiq-Ul-Hassan, M.; Zhang, G.; Latifi, K.; Ullah, G.; Hunt, D.; Balagurunathan, Y.; Abdalah, M.; Schabath, M.; Goldgof, D.; Mackin, D.; et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med. Phys.* **2017**, *44*, 1050–1062. [CrossRef]
9. Escudero Sanchez, L.; Rundo, L.; Gill, A.; Hoare, M.; Mendes Serrao, E.; Sala, E. Robustness of radiomic features in CT images with different slice thickness, comparing liver tumour and muscle. *Sci. Rep.* **2021**, *11*, 8262. [CrossRef]
10. Whitney, H.; Drukker, K.; Edwards, A.; Papaioannou, J.; Medved, M.; Karczmar, G.; Giger, M. Robustness of radiomic features of benign breast lesions and hormone receptor positive/HER2-negative cancers across DCE-MR magnet strengths. *Magn. Reson. Imaging* **2020**, *82*, 111–121. [CrossRef]
11. Scalco, E.; Belfatto, A.; Mastropietro, A.; Rancati, T.; Avuzzi, B.; Messina, A.; Valdagni, R.; Rizzo, G. T2w-MRI signal normalization affects radiomics features reproducibility. *Med. Phys.* **2020**, *47*, 1680–1691. [CrossRef] [PubMed]
12. de Farias, E.C.; di Noia, C.; Han, C.; Sala, E.; Castelli, M.; Rundo, L. Impact of GAN-based lesion-focused medical image super-resolution on the robustness of radiomic features. *Sci. Rep.* **2021**, *11*, 21361. [CrossRef] [PubMed]
13. Zwanenburg, A.; Leger, S.; Agolli, L.; Pilz, K.; Troost, E.; Richter, C.; Löck, S. Assessing robustness of radiomic features by image perturbation. *Sci. Rep.* **2019**, *9*, 614. [CrossRef] [PubMed]
14. Mottola, M.; Ursprung, S.; Rundo, L.; Sanchez, L.; Klatte, T.; Mendichovszky, I.; Stewart, G.; Sala, E.; Bevilacqua, A. Reproducibility of CT-based radiomic features against image resampling and perturbations for tumour and healthy kidney in renal cancer patients. *Sci. Rep.* **2021**, *11*, 11542. [CrossRef]
15. Tixier, F.; Um, H.; Young, R.; Veeraraghavan, H. Reliability of tumor segmentation in glioblastoma: Impact on the robustness of MRI-radiomic features. *Med. Phys.* **2019**, *46*, 3582–3591. [CrossRef]
16. Granzier, R.; Verbakel, N.; Ibrahim, A.; van Timmeren, J.; van Nijnatten, T.; Leijenaar, R.; Lobbes, M.; Smidt, M.; Woodruff, H. MRI-based radiomics in breast cancer: Feature robustness with respect to inter-observer segmentation variability. *Sci. Rep.* **2020**, *10*, 14163. [CrossRef]
17. Le, E.; Rundo, L.; Tarkin, J.M.; Evans, N.; Chowdhury, M.; Coughlin, P.; Pavey, H.; Wall, C.; Zaccagna, F.; Gallagher, F.; et al. Assessing robustness of carotid artery CT angiography radiomics in the identification of culprit lesions in cerebrovascular events. *Sci. Rep.* **2021**, *11*, 3499. [CrossRef]
18. van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.C.; Pieper, S.; Aerts, H.J.W.L. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* **2017**, *77*, e104–e107. [CrossRef]
19. Chuang, K.S.; Tzeng, H.L.; Chen, S.; Wu, J.; Chen, T.J. Fuzzy c-means clustering with spatial information for image segmentation. *Comput. Med. Imaging Graph.* **2006**, *30*, 9–15. [CrossRef]
20. Li, B.N.; Chui, C.K.; Chang, S.; Ong, S.H. Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation. *Comput. Biol. Med.* **2011**, *41*, 1–10. [CrossRef]
21. Militello, C.; Rundo, L.; Dimarco, M.; Orlando, A.; Conti, V.; Woitek, R.; D'Angelo, I.; Bartolotta, T.V.; Russo, G. Semi-automated and interactive segmentation of contrast-enhancing masses on breast DCE-MRI using spatial fuzzy clustering. *Biomed. Signal Process. Control* **2022**, *71*, 103113. [CrossRef]
22. Militello, C.; Ranieri, A.; Rundo, L.; D'Angelo, I.; Marinozzi, F.; Bartolotta, T.V.; Bini, F.; Russo, G. On Unsupervised Methods for Medical Image Segmentation: Investigating Classic Approaches in Breast Cancer DCE-MRI. *Appl. Sci.* **2021**, *12*, 162. [CrossRef]
23. Haralick, R.M.; Shanmugam, K.; Dinstein, I. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *SMC-3*, 610–621. [CrossRef]
24. Haralick, R.M. Statistical and structural approaches to texture. *Proc. IEEE* **1979**, *67*, 786–804. [CrossRef]
25. Galloway, M.M. Texture analysis using gray level run lengths. *Comput. Graph. Image Process.* **1975**, *4*, 172–179. [CrossRef]
26. Thibault, G.; Angulo, J.; Meyer, F. Advanced Statistical Matrices for Texture Characterization: Application to Cell Classification. *IEEE Trans. Biomed. Eng.* **2014**, *61*, 630–637. [CrossRef] [PubMed]
27. Sun, C.; Wee, W.G. Neighboring gray level dependence matrix for texture classification. *Comput. Vis. Graph. Image Process.* **1983**, *23*, 341–352. [CrossRef]
28. Amadasun, M.; King, R. Textural features corresponding to textural properties. *IEEE Trans. Syst. Man Cybern.* **1989**, *19*, 1264–1274. [CrossRef]

29. Rundo, L.; Tangherloni, A.; Galimberti, S.; Cazzaniga, P.; Woitek, R.; Sala, E.; et al. HaraliCU: GPU-powered Haralick feature extraction on medical images exploiting the full dynamics of gray-scale levels. In Proceedings of the International Conference on Parallel Computing Technologies (PaCT), Almaty, Kazakhstan, 19–23 August 2019; Malyshkin, V., Ed.; Springer International Publishing: Cham, Switzerland, 2019; LNCS 11657, pp. 304–318. [CrossRef]
30. Shrout, P.E.; Fleiss, J.L. Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* **1979**, *86*, 420–428. [CrossRef]
31. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef]
32. Shapiro, S.; Wilk, M. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* **1965**, *52*, 591–611. [CrossRef]
33. Militello, C.; Vitabile, S.; Rundo, L.; Russo, G.; Midiri, M.; Gilardi, M.C. A fully automatic 2D segmentation method for uterine fibroid in MRgFUS treatment evaluation. *Comput. Biol. Med.* **2015**, *62*, 277–292. [CrossRef] [PubMed]
34. Furqan Qadri, S.; Ai, D.; Hu, G.; Ahmad, M.; Huang, Y.; Wang, Y.; Yang, J. Automatic Deep Feature Learning via Patch-Based Deep Belief Network for Vertebrae Segmentation in CT Images. *Appl. Sci.* **2019**, *9*, 69. [CrossRef]
35. Hirra, I.; Ahmad, M.; Hussain, A.; Ashraf, M.U.; Saeed, I.A.; Qadri, S.F.; Alghamdi, A.M.; Alfakeeh, A.S. Breast Cancer Classification From Histopathological Images Using Patch-Based Deep Learning Modeling. *IEEE Access* **2021**, *9*, 24273–24287. [CrossRef]