



Four-Dimensional Machine Learning Radiomics for the Pretreatment Assessment of Breast Cancer Pathologic Complete Response to Neoadjuvant Chemotherapy in Dynamic Contrast-Enhanced MRI

Marco Caballo, PhD,^{1*}  Wendelien B. G. Sanderink, PhD,¹  Luyi Han, MSc,^{1,2}
Yuan Gao, MSc,^{2,3} Alexandra Athanasiou, MD, MSc,⁴ and Ritse M. Mann, MD, PhD^{1,2}

Background: Breast cancer response to neoadjuvant chemotherapy (NAC) is typically evaluated through the assessment of tumor size reduction after a few cycles of NAC. In case of treatment ineffectiveness, this results in the patient suffering potentially severe secondary effects without achieving any actual benefit.

Purpose: To identify patients achieving pathologic complete response (pCR) after NAC by spatio-temporal radiomic analysis of dynamic contrast-enhanced (DCE) MRI images acquired before treatment.

Study type: Single-center, retrospective.

Population: A total of 251 DCE-MRI pretreatment images of breast cancer patients.

Field strength/sequence: 1.5 T/3 T, T1-weighted DCE-MRI.

Assessment: Tumor and peritumoral regions were segmented, and 348 radiomic features that quantify texture temporal variation, enhancement kinetics heterogeneity, and morphology were extracted. Based on subsets of features identified through forward selection, machine learning (ML) logistic regression models were trained separately with all images and stratifying on cancer molecular subtype and validated with leave-one-out cross-validation.

Statistical tests: Feature significance was assessed using the Mann–Whitney U-test. Significance of the area under the receiver operating characteristics (ROC) curve (AUC) of the ML models was assessed using the associated 95% confidence interval (CI). Significance threshold was set to 0.05, adjusted with Bonferroni correction.

Results: Nine features related to texture temporal variation and enhancement kinetics heterogeneity were significant in the discrimination of cases achieving pCR vs. non-pCR. The ML models achieved significant AUC of 0.707 (all cancers, $n = 251$, 59 pCR), 0.824 (luminal A, $n = 107$, 14 pCR), 0.823 (luminal B, $n = 47$, 15 pCR), 0.844 (HER2 enriched, $n = 25$, 11 pCR), 0.803 (triple negative, $n = 72$, 19 pCR).

Data Conclusions: Differences in imaging phenotypes were found between complete and noncomplete responders. Furthermore, ML models trained per cancer subtype achieved high performance in classifying pCR vs. non-pCR cases. They may, therefore, have potential to help stratify patients according to the level of response predicted before treatment, pending further validation with larger prospective cohorts.

Evidence Level: 4

Technical Efficacy: Stage 4

J. MAGN. RESON. IMAGING 2023;57:97–110.

View this article online at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1002/jmri.28273). DOI: 10.1002/jmri.28273

Received Apr 11, 2022, Accepted for publication May 13, 2022.

*Address reprint requests to: M.C., Advanced X-Ray Tomographic Imaging (AXTI) Lab, Department of Medical Imaging, Radboudumc, Nijmegen, The Netherlands. E-mail: marco.caballo@radboudumc.nl

From the ¹Department of Medical Imaging, Radboud University Medical Center, Nijmegen, The Netherlands; ²Department of Radiology, The Netherlands Cancer Institute, Amsterdam, The Netherlands; ³GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre, Maastricht, the Netherlands; and ⁴Breast Imaging Department, MITERA Hospital, Marousi, Athens, Greece

Additional supporting information may be found in the online version of this article

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Neoadjuvant chemotherapy (NAC) has gained notable acceptance as first-line breast cancer treatment for larger cancers and those with specific immunohistochemical characteristics.^{1,2} However, response to NAC varies considerably. It is estimated that 2%–30% of patients do not benefit from this treatment, and only up to 30%–50% achieve pathologic complete response (pCR).³ Achieving pCR is for most cancers a good prognostic indicator for disease free survival and is therefore commonly used as a surrogate endpoint for treatment success.³

Response to NAC is currently assessed during and after treatment with imaging and subsequently confirmed after surgery via histopathologic analysis.⁴ For imaging evaluation, dynamic contrast-enhanced MRI (DCE-MRI) is recommended. DCE-MRI consists of the acquisition of T1-weighted acquisitions prior to and several times after contrast administration, enabling both morphological and kinetic analysis of enhancing lesions. DCE-MRI enables the evaluation of the level of response of the tumor to NAC by assessing changes in the enhancement pattern over time using standardized methods that mainly use tumor size reduction as a prognostic indicator (e.g. the Response Evaluation Criterion in Solid Tumors [RECIST]).⁴ However, the reduction of the tumor size, if present, typically occurs only after at least two NAC cycles.⁴ This results, in case of treatment ineffectiveness, in waste of medical resources and, importantly, in the patient suffering secondary effects that may be severe, without achieving any actual benefit.⁴ Therefore, the ability to predict the likelihood of response to NAC before any treatment takes place is crucial.

In recent years, the development of methods for pretreatment NAC response prediction has been an active area of research. In the imaging domain, radiomics methodologies have been devised for the automated and quantitative analysis of pretreatment DCE-MRI images, with the aim of quantifying cancer-related imaging biomarkers able to predict the response to NAC.^{5,6} In radiomics, dedicated image analysis algorithms are used to extract and quantify features related to tumor anatomy and function, which can then be used to develop computer models for clinical decision support.⁷ This approach has been demonstrated to outperform visual perception in characterizing cancer biology through imaging.^{8–16}

For prediction of the response to NAC, most studies involved the automatic extraction, through engineered algorithms, of quantitative information related to texture (i.e. spatial enhancement) or kinetics (i.e. temporal enhancement) from DCE-MRI images.^{17–28} This information was then used to develop machine learning (ML) models for NAC response prediction. Most often, tumor (and, sometimes, peritumoral) texture are quantified from the first post-contrast image, or, in a few cases, independently for different temporal phases.^{17–24} Kinetics are characterized, mainly, globally within the whole tumoral environment or within

the tumor region with the highest signal enhancement intensity.^{17–24} To better account for tumor heterogeneity, a recognized challenge that may hinder treatment effectiveness,⁹ a few investigators localized the extraction of kinetic features in different tumor subregions, to quantify differences in enhancement dynamics throughout the tumor environment.^{25–28} Several authors performed classification experiments by stratifying per breast cancer molecular subtype.^{17–21}

Those studies thus uniformly show that both texture and dynamic features could be useful for NAC response prediction. However, the integration of texture and kinetic characteristics remains limited in these prediction models. It has been suggested that spatio-temporal information in tumor enhancement may reflect important phenotypical aspects of the underlying intratumoral heterogeneity that could correlate with cancer outcomes.⁹ This implies that the optimal use of integrated spatio-temporal information, likely in combination with molecular subtypes, could be used to improve the prediction of response to NAC for breast cancer.

Consequently, this study aims to explore the potential of four-dimensional (4D, 3D plus time) ML radiomics for the pretreatment classification of pCR and non-pCR patients with various molecular subtypes of breast cancer treated with NAC, to leverage the complete spatio-temporal information included in DCE-MRI data.

Materials and Methods

Study Dataset and Data Acquisition

This study was approved by the institutional review board of our institution. Consent or waiver was not required, as all data were obtained deidentified from a public dataset (Duke-Breast-Cancer-MRI) stored in the National Cancer Institute's Cancer Imaging Archive.^{29–31} The dataset consisted of 922 pretreatment DCE-MRI images of invasive breast cancer patients, complete with demographic, clinical, pathologic (including ethnicity, age, histologic type, and hormone receptor status), and treatment outcome information. The location of the tumors in the images was provided as a three-dimensional (3D) bounding box annotated by one of multiple breast radiologists. At the moment of MRI examination, eligible patients did not have any prior history of breast cancer, breast surgery, or neoadjuvant therapy.

Patients were recruited at the Duke University School of Medicine (Durham, USA) from January 2000 to March 2014.³⁰ pCR to NAC was defined as the absence of any disease (invasive or in situ) in the breast or in axillary lymph nodes, as specified in the pathology reports resulting from the surgical resection specimen following the first surgical intervention. Non-pCR cases were defined as achieving only a partial response (or none), or with remaining ductal or lobular carcinoma in situ. For our study, we selected the patients receiving NAC, and we excluded those with missing ground truth on pCR (i.e. treatment response assessment unavailable), those with motion artifacts in the MRI scan, and those where image segmentation failed (see the next section for details on image segmentation).

Images were acquired with multislice axial MRI in prone position with 1.5 T or 3 T scanners (GE Healthcare, Little Chalfont, UK and

Siemens, Munich, Germany) and consisted of a non-fat saturated T1-weighted sequence, a fat-saturated gradient echo T1-weighted precontrast sequence, and typically four postcontrast T1-weighted sequences (median acquisition time between postcontrast phases of 131 seconds).³⁰ Complete details about the dataset were previously reported.³⁰

Image Segmentation

Tumor and peritumoral regions were segmented using a previously validated fuzzy C-means algorithm³² implemented in MATLAB (r2018a, The MathWorks, Natick, MA, USA). To segment the tumors, the algorithm was applied within the bounding box enclosing the cancer, provided with the image dataset. To segment the peritumoral regions, the algorithm was applied in the image region that extended 20 mm radially outward from the tumor borders. Both segmentations were performed in 3D in the first postcontrast phase and copied to the other phases. Segmentation results were checked by a medical image analysis scientist (M.C., >4.5 years of experience in breast image segmentation) and manually corrected as necessary (due to, e.g. incorrect initial exclusion of necrotic regions, or incorrect initial inclusion of nearby vasculature or skin). Segmentation results were visually assessed by a board-certified breast radiologist (A.A., 20 years of experience) and further refined as needed using the ImageJ (v. 1.8.0, National Institutes of Health, Bethesda, USA) polyline toolbox (example in Fig. 1). Additional information on image segmentation is described in the Supplemental Digital Content S1.

Radiomic Features

In total, 348 radiomic features divided into five groups were extracted: first postcontrast texture (group 1); time-dependent

texture (group 2); pseudo-4D texture (group 3); enhancement kinetics heterogeneity (group 4); tumor morphology (group 5).

Texture was quantified using first- and second-order descriptors (histogram-based, co-occurrence, run length) with previously developed methods.^{33–36} Briefly, histogram-based features assess the distribution of voxel intensities within the image, by extracting first-order statistical descriptors. Co-occurrence features analyze the spatial distribution of voxel intensities, quantifying the frequency information of intensity values within a neighborhood of voxels along specific linear orientations. Run-length features evaluate the texture by measuring strings of consecutive voxels with the same intensity, quantifying the coarseness of the image in multiple linear directions. To quantify the image texture in this study, three different settings were created. First, texture was calculated in 3D from the first postcontrast phase (group 1, $n = 33$ features) (Fig. 2a). Second, texture features were extracted from each MRI phase (considering precontrast and all postcontrast images), and their mean and variance over time were calculated (group 2, $n = 66$ features). This second approach aimed to investigate the potential prognostic value of the variation of image texture over time, by capturing the overall image texture patterns (mean) and the possible textural changes across different temporal phases (variance) (Fig. 2b). Third, temporal information was directly incorporated in the image texture calculation. For this, each 3D MRI temporal phase was reduced to two-dimensions (2D) through maximum intensity projection (MIP). The resulting 2D projections across time were concatenated to form a 3D volume, representing the image (compressed) spatial information along the first two dimensions, and the image temporal information along the third. Texture was quantified from this pseudo-4D image, allowing to evaluate changes in image textural enhancement simultaneously across space and time (group 3, $n = 33$ features) (Fig. 2c).

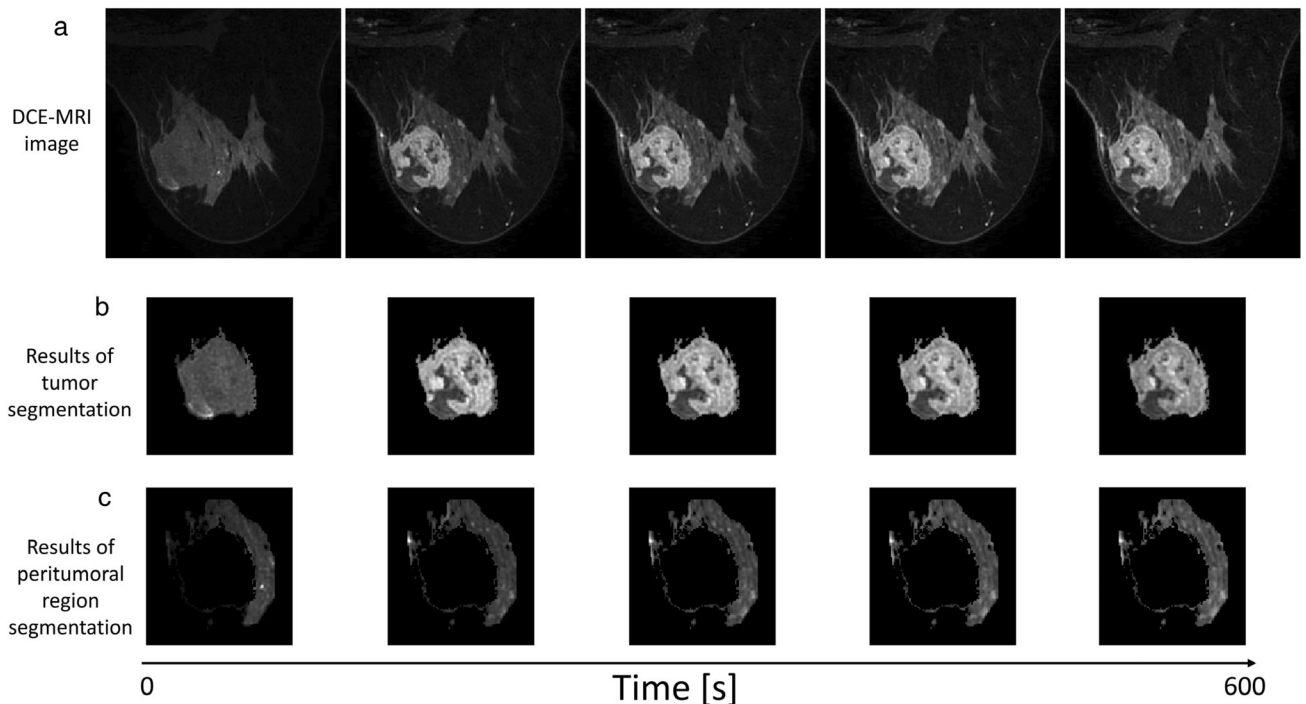


FIGURE 1: Example of (a) DCE-MRI slice (precontrast and four postcontrast phases) and results of (b) tumor and (c) peritumoral region segmentation.

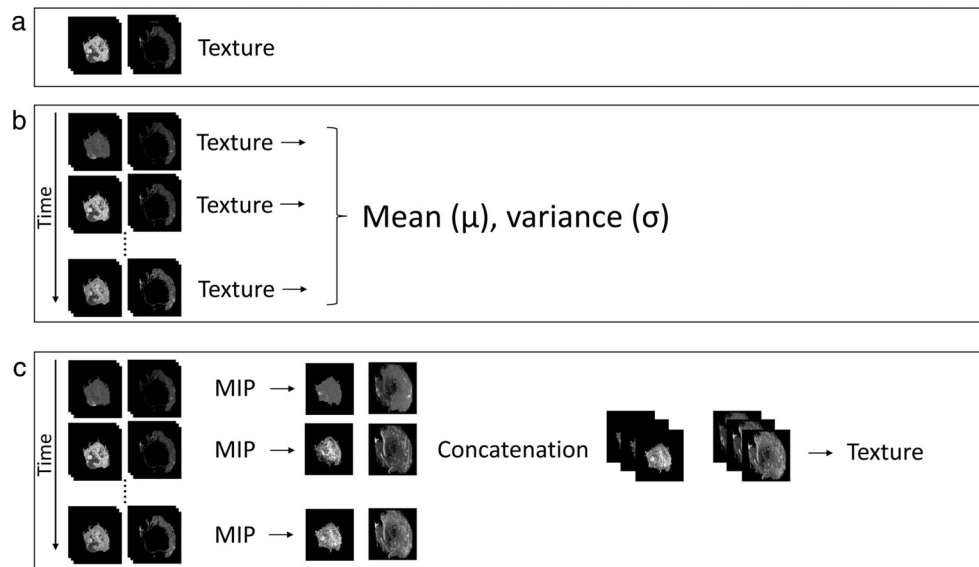


FIGURE 2: Schematics of texture feature calculation: (a) 3D image texture quantified from the first post-contrast phase (group 1); (b) time-dependent texture, quantified as the mean and variance of image texture across temporal phases (group 2); (c) pseudo-4D texture calculation (group 3): 3D phases were first compressed to two dimensions through maximum intensity projection (MIP), concatenated along the temporal dimension, and texture quantified from this resulting 3D volume.

Enhancement kinetics heterogeneity was characterized through quantification of four kinetic parameters (time to peak, peak enhancement, wash-in rate, wash-out rate), calculated in three different subregions (representing voxel clusters of high, moderate, and low enhancement) of the tumor and surrounding peritumoral region (Figs. 3 and 4a,b).

The choice of evaluating the kinetic parameters in three different subregions (and not in the whole tumor and peritumoral region segmentations) was performed to incorporate heterogeneity in the functional characterization of the tumor (and peritumoral region). This allows to characterize the enhancement kinetics (i.e. temporal information) accounting for spatial information, by first dividing the voxels into subregions of homogeneity, and then by quantifying patterns in kinetic feature statistics within every subregion.²⁵ To identify these subregions, unsupervised k-means clustering was applied to partition the voxels into three clusters of homogeneity, based on temporal enhancement characteristics in each voxel.³³ This algorithm works by iteratively assigning the voxels to a cluster, starting from a random initialization, until the within-cluster variances are minimized. To leverage the temporal enhancement information, the algorithm was applied in 3D on the first postcontrast image, but variances were calculated including the fourth dimension, that is, using the whole perfusion curve associated with each voxel. To allow for result reproducibility, the clustering algorithm was repeated 10 times, with different random cluster initialization sampled from a normal distribution, and the solution with the highest intercluster variance was selected.

Once identified, the spatial extent of each cluster was quantified, and the mean, variance, and maximum value of the four kinetic parameters were calculated in each cluster (group 4, $n = 39$ features) (Fig. 4c). Both texture (in all three settings) and enhancement kinetics heterogeneity features were extracted, separately, from the tumor and the peritumoral region, resulting in a total of 342 features.

Finally, tumor morphology was quantified through the extraction of morphological descriptors: volume, surface area, solidity, equivalent diameter, sphericity, surface area to volume ratio (group 5, $n = 6$ features).

Complete description of all radiomic features, the clustering approach for subregion detection, and calculation of the four kinetics parameters have been previously reported.³³ All radiomic features were extracted in MATLAB (r2018a, The MathWorks).

Feature Analysis

Features were first harmonized to limit the potential bias introduced by the differences in signal-to-noise ratio caused by magnet field strength (1.5 T and 3 T) and associated scan protocol variation. Harmonization was performed using the parametric version of the ComBat method, which yields a transformation of the feature distributions according to the variable being tested (magnet field strength) using additive and multiplicative batch effects (additional details in the Supplemental Digital Content S1).³⁷

Subsequently, radiomic features biased by differences in image resolution were identified, as described in the following lines. To quantify image resolution, slice thickness was used, as it was previously identified as being among the main acquisition-related parameters that could influence the values of radiomic features.³⁸ Features were divided in two groups based on the slice thickness parameter (slice thickness < 2 mm and ≥ 2 mm), and features showing significantly different median values between the two groups were discarded (see Statistical Analysis section for details). All analyses were performed in MATLAB (r2018a, The MathWorks).

Machine Learning Models

ML models were developed to classify pCR vs. non-pCR cases based on the extracted radiomic features. Both univariate and multivariate models were developed to evaluate, respectively, the individual

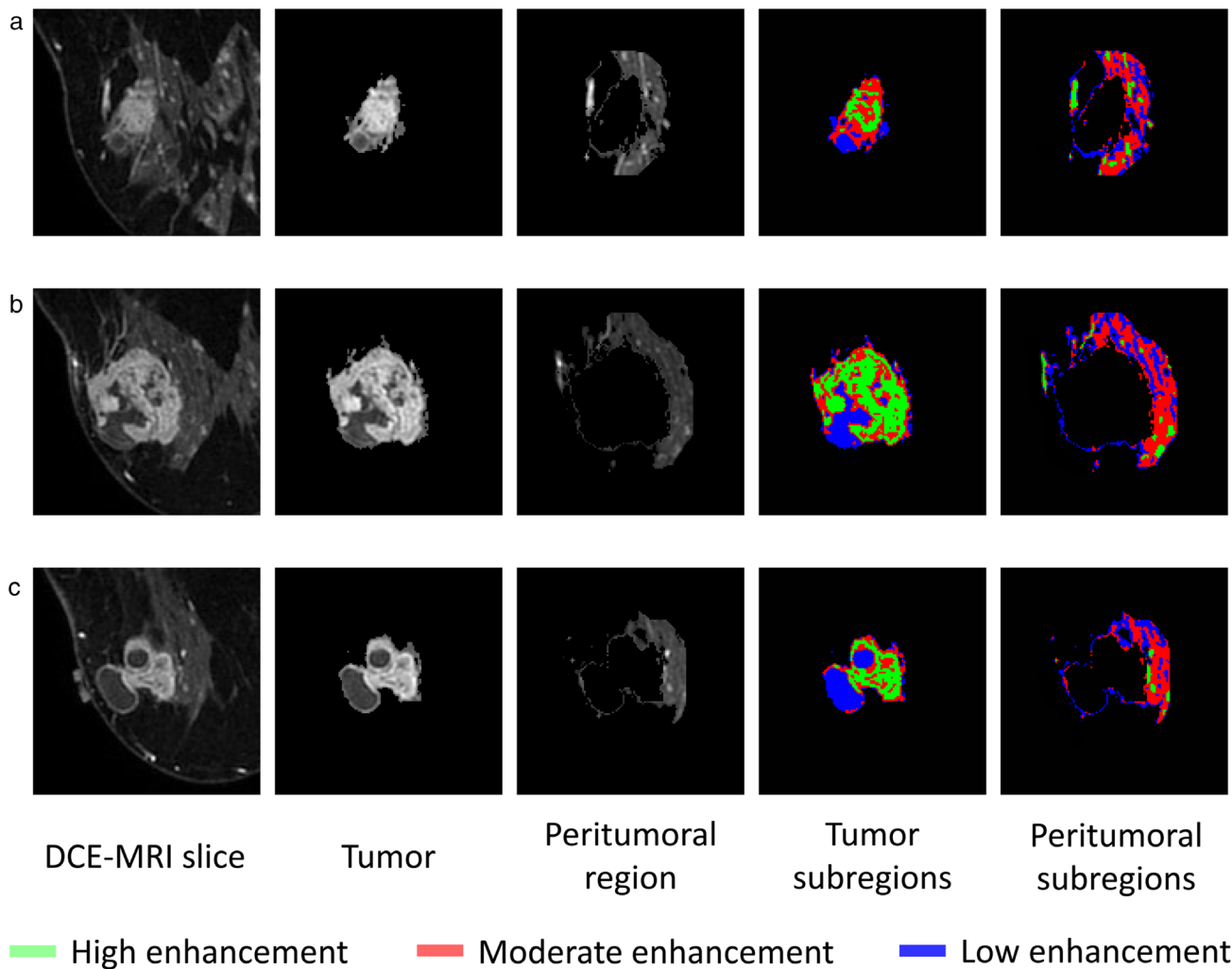


FIGURE 3: Three axial MRI slices of the same patient image and of the same time point (first postcontrast), respective tumor and peritumoral region segmentation, and respective results of the subregion identification algorithm used to quantify enhancement kinetics heterogeneity. (a) most cranial slice containing the tumor; (b) central slice; (c) most caudal slice containing the tumor.

predictive power of each feature, and the collective predictive power of a multifeature radiomic signature. In the univariate analysis, features were fed, one by one, to a ML logistic regression model trained and validated in a leave-one-out cross validation fashion. In the multivariate analysis, another ML logistic regression model was trained with multiple features, and tested in a leave-one-out cross validation fashion to classify pCR vs. non-pCR cases. To prevent overfitting, forward feature selection was performed in each leave-one-out loop using the training examples of the current loop. The number of selected features in each loop was kept lower than 1 for each 10 minority cases. Classification experiments were performed using features obtained from the entire image dataset, and repeated, separately, by stratifying on cancer molecular subtypes (luminal A, luminal B, HER2 enriched, triple negative, obtained through results of immunohistochemical analysis).

Ablation Study

To investigate the contribution in performance of different radiomic descriptors, multivariate ML classification was repeated per feature group, using the entire image dataset, and compared with the results

of the multivariate model based on features from all groups. A schematic of the whole methodology implemented in this study, including all validation steps, is shown in Fig. 5. All ML models were developed in MATLAB (r2018a, The MathWorks).

Statistical Analysis

Stability of radiomic features to the slice thickness parameter (dichotomized in <2 mm, and ≥ 2 mm) was assessed using the Mann–Whitney U-test. Features yielding a statistically significant difference were deemed biased by image resolution and discarded. To be more conservative on the number of features to retain, correction for multiple comparison was not applied.

After signal-to-noise ratio and resolution bias correction, univariate statistical analysis was performed to evaluate the potential significance of individual descriptors in discriminating pCR vs. non-pCR cases. Since no feature showed normal distribution (Shapiro–Wilk test), the Mann–Whitney U-test was used, with threshold on statistical significance adjusted for multiple comparisons (Bonferroni correction).

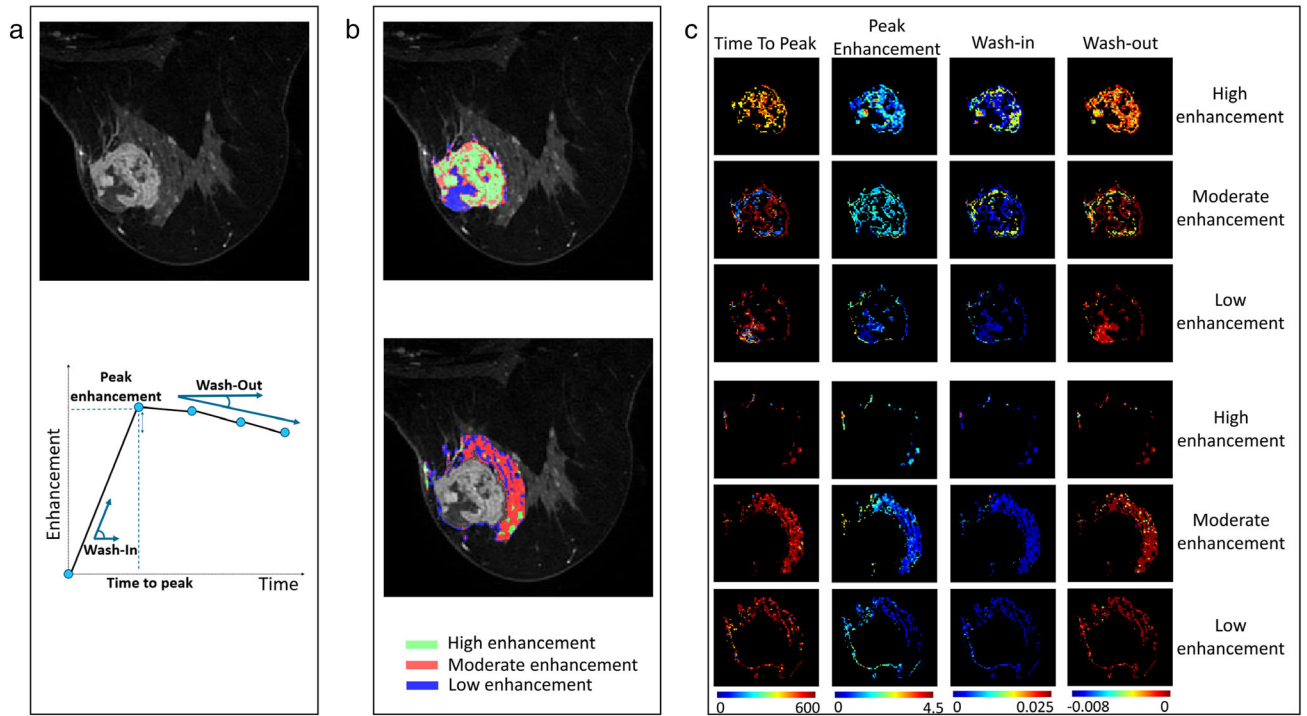


FIGURE 4: Schematics of the enhancement kinetics heterogeneity features (group 4): (a) example of MRI image (top row) and schematics of the four kinetic parameters (time to peak, peak enhancement, wash-in rate, wash-out rate) (bottom row); (b) example result of the clustering algorithm for subregion detection (tumor, top rows, and peritumoral region, bottom row); (c) respective voxel-wise maps of time to peak, peak enhancement, wash-in rate and wash-out rate for the three subregions identified.

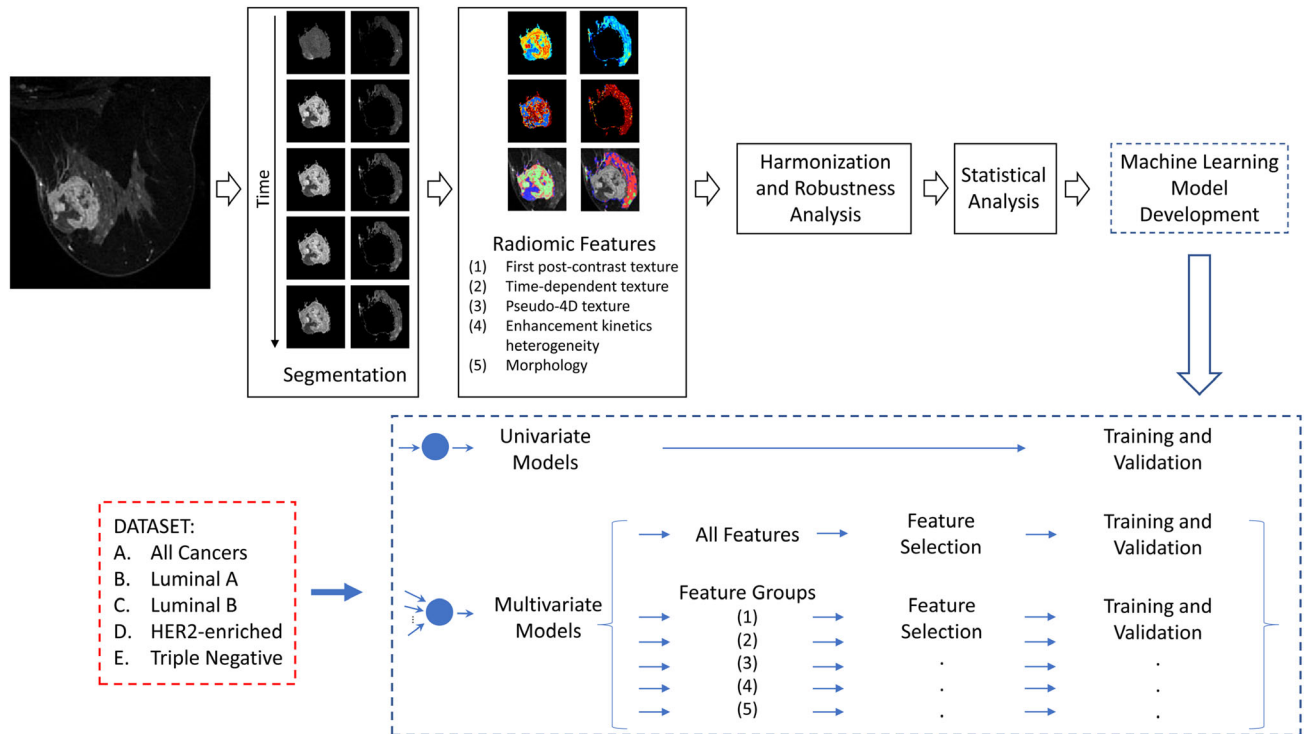


FIGURE 5: Schematics of the whole methodology developed and implemented in this study for radiomic feature extraction, analysis, and machine learning model development.

Univariate statistical analysis was also performed on demographic and clinical characteristics of the included patients, to evaluate any significance between the pCR and the non-pCR group. Continuous variables were analyzed with the Mann–Whitney

U-test, while discrete variables with the Pearson's chi-squared test.

When evaluating the predictive power of the univariate ML models developed upon individual features, the area under the

receiver operating characteristics (ROC) curve (AUC) was calculated, and the associated 95% confidence interval (CI) obtained with bootstrapping (1000 bootstraps). A feature was deemed individually informative if the lower limit of the 95% CI did not overlap with random chance (AUC = 0.5).

In the evaluation of the multivariate ML models (based on all data, and stratified per molecular subtype), performance was quantified through the AUC. Statistical significance was assessed using the associated 95% CI,³⁹ with the threshold corrected for multiple testing (Bonferroni correction, five tests).

In the ablation study, AUCs of multivariate ML models based on individual feature groups were compared using the DeLong method for AUC comparison.

For all analyses performed, significance threshold was set to 0.05.

Results

Results of patient selection are shown in Figure 6. Briefly, of the 292 eligible patients receiving NAC, 41 were excluded due to missing ground truth on NAC response outcome or failure of image segmentation. Therefore, 251 patients were selected (59 achieving pCR). Demographic, clinical, and treatment outcome characteristics are reported in Table 1. Progesterone receptor (PR) status, human epidermal growth factor receptor 2 (HER2) status, molecular subtype, and recurrence status were significantly different for the pCR and the non-pCR group (Table 1). DCE-MRI acquisition parameters are reported in Table 2.

In terms of field strength, 115 (45.8%) patients were acquired at 1.5 T, and 136 (54.2%) at 3 T. pCR prevalence in the two groups (1.5 T and 3 T) was similar: $n = 31$ (27%) and $n = 28$ (21%) for 1.5 T and 3 T, respectively. Median slice thickness was 2 mm (range: 1.04 mm - 2.5 mm). In total, 87 images had slice thickness <2 mm, and 164 ≥ 2 mm, with similar pCR prevalence in the two groups: $n = 18$ (21%) and $n = 41$ (25%) for slice thickness < 2 mm

and ≥ 2 mm, respectively. A fraction of the features (26/348) was found to be dependent on slice thickness and therefore deemed not robust and discarded.

Nine features (five extracted from the tumor, four from the peritumoral region, marked with an asterisk in Table 3) achieved statistical significance in discriminating pCR vs. non-pCR cases. Seven of these features belong to the group of enhancement kinetics heterogeneity descriptors (group 4), and two to the time-dependent texture (group 2).

Twelve features were found to be individually informative (i.e. with an associated AUC 95% CI not overlapping with random chance) of pCR based on the univariate ML analysis performed with the entire dataset ($n = 251$, 59 pCR). When stratifying on molecular subtype, 18 features were deemed individually predictive for luminal A ($n = 107$, 14 pCR), 8 for luminal B ($n = 47$, 15 pCR), 4 for HER2 enriched ($n = 25$, 11 pCR), and 7 for triple negative ($n = 72$, 19 pCR). Overall, most features were from group 2 (time-dependent texture) and group 4 (enhancement kinetics heterogeneity). Only two texture features extracted from the first postcontrast phase (group 1), and three tumor morphological features, were found to be predictive in triple negative and luminal B cancers, respectively. In all, 65% of all predictive features were extracted from the tumor and 35% from the peritumoral region (Table 3).

The multivariate ML model resulted in an AUC = 0.707 in pCR classification for the entire dataset. When training the model on specific molecular subtypes, performance increased, with an AUC = 0.824 (luminal A), AUC = 0.823 (luminal B), AUC = 0.844 (HER2 enriched), AUC = 0.803 (triple negative). ROC curves, 95% CI, and P values are reported in Figure 7 and Table 4.

Results of the multivariate ML model trained on a feature group level (ablation study) are reported in Table 5. Performance based on the entire initial feature space (before feature analysis and selection) was higher than that obtained per single feature group (AUC = 0.707 vs. AUC ranging between 0.473 and 0.617). Statistical significance was achieved against texture features extracted from the first post-contrast phase (group 1), pseudo-4D texture (group 3), and tumor morphological features (group 5).

Discussion

We developed and evaluated a ML 4D radiomics approach for the spatio-temporal analysis of pretreatment DCE-MRI images for the classification of patients achieving pCR at the end of the treatment, vs. partial or non-responders. We performed both univariate and multivariate analyses to investigate the performance of the implemented radiomic descriptors, using the entire dataset available and stratifying on molecular subtype.

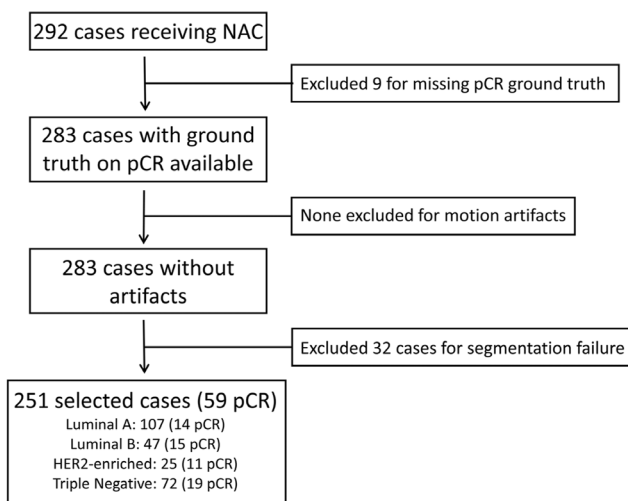


FIGURE 6: Flowchart of the patient selection process.

TABLE 1. Characteristics of the Patients Included in This Study

Demographic and histopathologic characteristics	Age (years)	Mean (SD)	49.2 (10.9)
		Range	25.0–76.6
	Race	White or Hispanic	162 (64.5%)
		Other	89 (35.5%)
	Metastatic at presentation	Yes	251 (100%)
		No	0 (0%)
	Laterality	Left	79 (31.5%)
		Right	91 (36%)
		Not available	81 (32.5%)
	Bilateral cancer	Yes	163 (64.9%)
		No	6 (2.4%)
		Not available	82 (32.7%)
	Estrogen receptor (ER) status	Negative	103 (41.0%)
		Positive	148 (59.0%)
	Progesterone receptor (PR) status*	Negative	136 (54.2%)
		Positive	115 (45.8%)
	Human epidermal growth factor receptor 2 (HER2) status*	Negative	179 (71.3%)
		Positive	72 (28.7%)
	Molecular subtype*	Luminal A (ER+ and/or PR+, HER2–)	107 (42.6%)
		Luminal B (ER+ and/or PR+, HER2+)	47 (18.7%)
		HER2-enriched (HER2+, ER–, PR–)	25 (10.0%)
		Triple Negative (ER–, PR–, HER2–)	72 (28.7%)
	Histologic type	Ductal	158 (63.0%)
		Lobular	10 (4.0%)
		Not Available	83 (33.0%)
	Recurrence*	Yes	33 (13.1%)
		No	218 (86.9%)
Treatment outcome characteristics	pCR—Luminal A	Yes	14 (13%)
		No	93 (87%)
	pCR—Luminal B	Yes	15 (32%)
		No	32 (68%)
	pCR—HER2 enriched	Yes	11 (44%)
		No	14 (56%)
	pCR—Triple Negative	Yes	19 (26%)
		No	53 (74%)

*Statistically significant based on nonparametric test (Mann–Whitney U-test for continuous variables, Pearson's chi-squared test for discrete variables) with Bonferroni correction for multiple comparison.

Features resulting in statistically significant difference between pCR and non-pCR cases are marked with an asterisk. Unless indicated otherwise in the table, numbers between brackets indicate the percentage of patients falling in the described category.

TABLE 2. Acquisition Parameters of the MRI Images Used in This Study

MRI acquisition parameters	Magnet Field Strength	1.5 T	115 patients (45.8%)
		3 T	136 patients (54.2%)
	Slice thickness (mm)	Median (range)	2.0 (1.04–2.5)
	Repetition time (msec)	Median (range)	4.98 (3.54–7.39)
	Echo time (msec)	Median (range)	2.35 (1.29–2.70)
	Acquisition matrix (array size)	Minimum–Maximum	340 × 340–448 × 448
	Flip angle (degrees)	Median (range)	10 (7–12)
	Field of view (cm)	Median (range)	350 (270–440)

Unless indicated otherwise in the table, numbers between brackets indicate the percentage of patients falling in the described category.

In the univariate ML analysis, features from both the tumor and the peritumoral region were found to be predictive of pCR. Most of these features were derived from the time-dependent texture analysis and from the cluster-based enhancement kinetics heterogeneity descriptors. Some features overlapped among the analyses performed (i.e. with the entire dataset and stratifying by subtype), while others were predictive only for specific subtypes. When using the entire dataset, nine of these features showed a significant difference between pCR and non-pCR groups. This suggests that intrinsic phenotypical characteristics, both at spatial and temporal level, may be present between pCR and non-pCR groups. These features show that responders have, overall, a higher peak enhancement, wash-in and wash-out rates in subregions with high and moderate enhancement, and a higher textural variation across temporal phases. This indicates that pronounced enhancement seems to be associated with a higher likelihood to achieve pCR. These findings agree with those of a previous independent study conducted on 15 patients²⁵ and with oncologic literature.⁴⁰

Texture extracted from the first postcontrast phase and tumor morphology resulted, instead, in only a few features with predictive power, and only for triple negative and luminal B cancers. This suggests that the proposed spatio-temporal radiomic analysis is superior to the morphological features analysis, and at least complementary, if not superior, to the quantification of 3D image texture from only the first postcontrast image (a common approach in DCE-MRI radiomics¹⁰). On the other hand, pseudo-4D texture did not seem particularly informative, potentially due to the image compression through MIP that could result in loss of spatial enhancement information. These findings were confirmed in the ablation study, where classification performance achieved with spatio-temporal radiomic features was significantly higher than that achieved using only 3D texture and tumor morphology.

In the multivariate analysis, prediction performance increased when stratifying by cancer subtype. This is likely due to the ML model being tailored to a specific phenotype, thanks to the selection of the features that are most predictive of pCR for a given molecular subtype. This also suggests that the power of different radiomic features in pCR prediction can change across subtypes, a fact that can be explained by the already demonstrated phenotypical differences of molecular subtypes of breast cancer.⁴⁰ These findings are concordant with the univariate analysis, where only a partial overlap of individually predictive features was found among molecular subtypes.

In this study, we focused on the classification of patients achieving pCR at the end of the treatment, vs. all the others. While this is, typically, the most evaluated clinical outcome,^{17–28} due to pCR being a strong indicator of favorable prognosis,³ future studies should also investigate the classification of partial responders, nonresponders, and patients presenting disease progression. Although this would require larger and more detailed datasets, it would allow for a more detailed prediction of the level of response that can be achieved in a specific patient before start of treatment, potentially maximizing treatment effectiveness through the selection of the most appropriate therapeutic strategy for each patient.

Limitations

The texture features extracted could be supplemented with additional descriptors, especially to better characterize the distribution of image texture over time, and the clustering algorithm for subregion identification could be improved to allow for the tailoring of the selection of the number of subregions to each single tumor (i.e. not selected a priori). Furthermore, the potential variability of radiomic features due to image segmentation was not investigated. However, this effect should be mitigated by using an automatic algorithm as baseline segmentation and therefore is not expected to impact our results greatly.³⁴ Finally, the dataset

TABLE 3. Radiomic Features Individually Predictive of pCR (i.e. With a 95% CI on the AUC Obtained With the Univariate ML Analysis Not Overlapping With Random Chance)

Molecular Subtype	Category	Feature Name	Group	Image Region	AUC
Luminal A	2	Skewness (temporal variance)	First order	Tumor	0.69
	2	Entropy (temporal variance)	First order	Tumor	0.73
	2	Max probability (temporal variance)	Co-occurrence	Tumor	0.75
	2	Maximum (temporal variance)	First order	Peritumoral	0.63
	2	Run percentage (temporal variance)	Run length	Peritumoral	0.72
	3	Fifth Percentile	First order	Peritumoral	0.69
	4	Average wash-In	High enhancement cluster	Tumor	0.69
	4	Wash-in variance	High enhancement cluster	Tumor	0.81
	4	Average peak enhancement	Moderate enhancement cluster	Tumor	0.71
	4	Peak enhancement variance	Moderate enhancement cluster	Tumor	0.81
	4	Average wash-in	Moderate enhancement cluster	Tumor	0.72
	4	Wash-out variance	Moderate enhancement cluster	Tumor	0.81
	4	Peak enhancement variance	High enhancement cluster	Peritumoral	0.81
	4	Wash-in variance	High enhancement cluster	Peritumoral	0.81
	4	Peak enhancement variance	Moderate enhancement cluster	Peritumoral	0.81
	4	Wash-in variance	Moderate enhancement cluster	Peritumoral	0.80
	4	Wash-out variance	Moderate enhancement cluster	Peritumoral	0.80
	4	Peak enhancement variance	Low enhancement cluster	Peritumoral	0.73
Luminal B	4	Average peak enhancement	High enhancement cluster	Tumor	0.72
	4	Average wash-in	High enhancement cluster	Tumor	0.72
	4	Wash-in variance	High enhancement cluster	Tumor	0.70
	4	Maximum wash-in	High enhancement cluster	Tumor	0.69
	4	Volume fraction	Moderate enhancement cluster	Peritumoral	0.69
	5	Surface area	N.A.	Tumor	0.72
	5	Equivalent diameter	N.A.	Tumor	0.71
	5	Sphericity	N.A.	Tumor	0.69
HER2-enriched	2	High gray-level run emphasis (temporal variance)	Run length	Tumor	0.82
	2	Short run high gray-level emphasis (temporal variance)	Run length	Tumor	0.83
	4	Wash-out variance	Moderate enhancement cluster	Tumor	0.81
	4	Wash-out variance	Moderate enhancement cluster	Peritumoral	0.82
Triple negative	1	Low gray-level run emphasis	Run length	Tumor	0.70
	1	Short run emphasis	Run length	Tumor	0.69
	2	Short run emphasis (temporal mean)	Run length	Tumor	0.69
	2	Low gray-level run emphasis (temporal mean)	Run length	Tumor	0.70

TABLE 3. Continued

Molecular Subtype	Category	Feature Name	Group	Image Region	AUC
All	2	Correlation (temporal variance)	Co-occurrence	Peritumoral	
	3	Short run emphasis	Run length	Tumor	0.69
	3	Low gray-level run emphasis	Run length	Tumor	0.70
	2	High gray-level run emphasis (temporal variance)* Feature value pCR = 0.33 (0.03) Feature value non-pCR = 0.18 (0.009)	Run length	Tumor	0.62
	2	Short run high gray-level emphasis (temporal variance)* Feature value pCR = 0.32 (0.03) Feature value non-pCR = 0.17 (0.009)	Run length	Tumor	0.63
	2	Run Percentage (Temporal Variance)	Run Length	Peritumoral	0.61
	4	Peak enhancement variance* Feature value pCR = 3.2 (0.08) Feature value non-pCR = 3.8 (0.14)	High enhancement cluster	Tumor	0.60
	4	Average wash-in	High enhancement cluster	Tumor	0.60
	4	Wash-in variance* Feature value pCR = 1.2e-4 (2.6e-6) Feature value non-pCR = 8.7e-5 (3.5e-6)	High enhancement cluster	Tumor	0.68
	4	Wash-out variance* Feature value pCR = 1.1e-4 (3.0e-6) Feature value non-pCR = 6.3e-5 (5.1e-6)	Moderate enhancement cluster	Tumor	0.70
	4	Wash-in variance	Low enhancement cluster	Tumor	0.63
	4	Peak enhancement Variance* Feature value pCR = 25.0 (0.5) Feature value non-pCR = 11.8 (0.4)	High enhancement cluster	Peritumoral	0.70
	4	Wash-in variance* Feature value pCR = 1.6e-4 (4.3e-5) Feature value non-pCR = 2.4e-4 (5.9e-6)	High enhancement cluster	Peritumoral	0.70
	4	Wash-out variance* Feature value pCR = 3.9e-5 (1.1e-6) Feature value non-pCR = 9.7e-6 (5.3e-7)	Moderate enhancement cluster	Peritumoral	0.68
	4	Peak enhancement variance* Feature value pCR = 4.7 (1.2) Feature value non-pCR = 2.8 (1.3)	Low enhancement cluster	Peritumoral	0.65

*Also, statistically significant based on nonparametric Mann–Whitney U-test test with Bonferroni correction for multiple comparison. Legend of feature category: (1) first post-contrast texture; (2) time-dependent texture; (3) pseudo-4D texture; (4) enhancement kinetics heterogeneity; (5) morphology. Features resulting in statistically significant difference between pCR and non-pCR cases are marked with an asterisk. The mean values and standard errors for these features, for pCR and non-pCR cases, are reported. HER2 = Human Epidermal Growth Factor Receptor 2; AUC = area under the receiver operating characteristics curve; pCR = pathologic complete response; N.A. = not applicable.

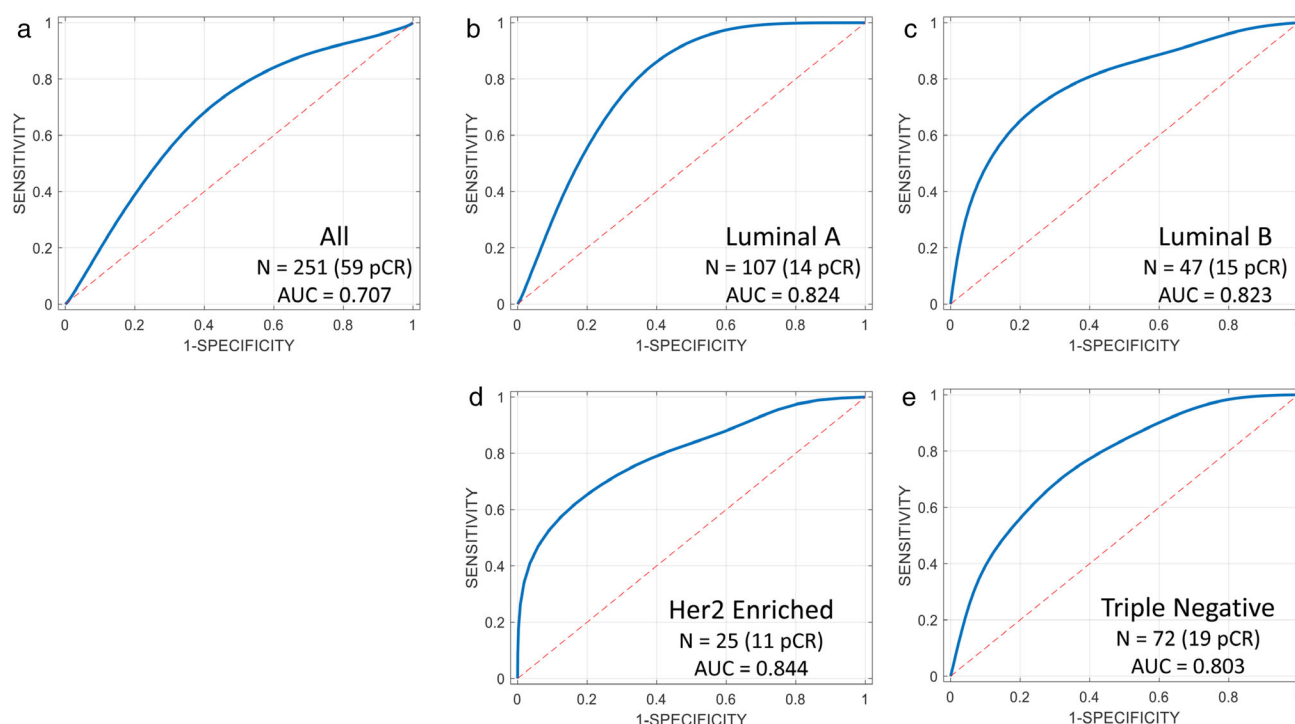


FIGURE 7: Fitted binormal ROC curves for the multivariate machine learning models in the prediction of pathologic complete response (pCR) vs. non-pCR cases: (a) all cancers; (b) only luminal A; (c) only luminal B; (d) only human epidermal growth factor receptor 2 (HER2) enriched; (e) only triple negative.

TABLE 4. Results of the Multivariate Machine Learning Logistic Regression Models for pCR Prediction

	Patients	AUC	C.I.	<i>P</i>
Luminal A	<i>N</i> = 107 (14 pCR)	0.824	[0.718–0.895]	<0.01
Luminal B	<i>N</i> = 47 (15 pCR)	0.823	[0.619–0.933]	<0.02
HER2 enriched	<i>N</i> = 25 (11 pCR)	0.844	[0.611–0.968]	<0.05
Triple negative	<i>N</i> = 72 (19 pCR)	0.803	[0.659–0.897]	<0.01
All	<i>N</i> = 251 (59 pCR)	0.707	[0.619–0.789]	<0.01

HER2 = human epidermal growth factor receptor 2; AUC = area under the receiver operating characteristics curve; pCR = pathologic complete response; C.I. = 95% confidence interval; *P* = *P* value.

TABLE 5. Results of the multivariate machine learning model trained and validated on the entire dataset per feature group

Feature Group	AUC	CI	<i>P</i>
First postcontrast texture	0.562	[0.474–0.612]	0.008
Time-dependent texture	0.617	[0.522–0.697]	0.15
Pseudo-4D texture	0.572	[0.485–0.618]	0.03
Enhancement kinetics heterogeneity	0.601	[0.534–0.659]	0.06
Morphology	0.473	[0.393–0.561]	0.0009
All	0.707	[0.619–0.789]	N.A.

Reported *P* values refer to the AUC comparison with the multivariate model trained with features belonging to different groups. AUC = area under the receiver operating characteristics curve; CI = 95% confidence interval; *P* = *P* value; N.A. = not applicable.

used was from a single institution, and therefore no independent testing was performed. While cross-validation has the advantage of not overfitting results on specific characteristics of a given test set, independent validation of the method with larger datasets from different institutions should be conducted in the future, to evaluate result reproducibility, or the lack thereof.

Future studies may also include the evaluation of the proposed methods in images with higher spatio-temporal resolution, which can provide finer information on tumor enhancement and, therefore, a potentially improved cancer characterization. Deep learning approaches for the spatio-temporal analysis of these images could also be investigated, to evaluate the possible gain in performance prediction when larger datasets become available.

Future work may also include other MRI sequences, for example, diffusion tensor imaging (DTI), stretched-exponential (SEM) and intravoxel incoherent motion (IVIM) models of diffusion-weighted imaging (DWI), to further improve treatment response assessment.^{41–43}

Conclusion

4D ML-based radiomic analysis of pretreatment DCE-MRI images of breast cancer was found promising in the classification of pCR vs. non-pCR cases for patients undergoing NAC. Time-dependent texture and enhancement kinetics heterogeneity features were found especially informative. These features may have a higher clinical value in predicting response to NAC from pretreatment imaging compared to 3D image texture and to tumor morphology.

Conflicts of Interest

The authors declare no potential conflicts of interest associated with this work.

REFERENCES

- Kamangar F, Dores GM, Anderson WF. Patterns of cancer incidence, mortality, and prevalence across five continents: Defining priorities to reduce cancer disparities in different geographic regions of the world. *J Clin Oncol* 2006;24(14):2137-2150.
- Korde LA, Somerfield MR, Carey LA, et al. Neoadjuvant chemotherapy, endocrine therapy, and targeted therapy for breast cancer: ASCO guideline. *J Clin Oncol* 2021;39(13):1485-1505.
- Cortazar P, Geyer CE Jr. Pathologic complete response in neoadjuvant treatment of breast cancer. *Ann Surg Oncol* 2015;22(5):1441-1446.
- Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. *J Natl Cancer Inst* 2000;92(3):205-216.
- Lo Gullo R, Eskreis-Winkler S, Morris EA, Pinker K. Machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy. *Breast* 2020;49:115-122.
- Granzier RWY, van Nijmegen TJA, Woodruff HC, Smidt ML, Lobbes MBI. Exploring breast cancer response prediction to neoadjuvant systemic therapy using MRI-based radiomics: A systematic review. *Eur J Radiol* 2019;121:108736.
- Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: The bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 2017;14(12):749-762.
- Chitalia RD, Rowland J, McDonald ES, et al. Imaging phenotypes of breast cancer heterogeneity in preoperative breast dynamic contrast enhanced magnetic resonance imaging (DCE-MRI) scans predict 10-year recurrence. *Clin Cancer Res* 2020;26(4):862-869.
- Ashraf A, Daye D, Gavenonis S, et al. Identification of intrinsic imaging phenotypes for breast cancer tumors: Preliminary associations with gene expression profiles. *Radiology* 2014;272(2):374-384.
- Chitalia RD, Kontos D. Role of texture analysis in breast MRI as a cancer biomarker: A review. *J Magn Reson Imaging* 2019;49(4):927-938.
- Drukker K, Li H, Antropova N, Edwards A, Papaioannou J, Giger ML. Most-enhancing tumor volume by MRI radiomics predicts recurrence-free survival "early on" in neoadjuvant treatment of breast cancer. *Cancer Imaging* 2018;18(1):12.
- Hu Q, Whitney HM, Li H, Ji Y, Liu P, Giger ML. Improved classification of benign and malignant breast lesions using deep feature maximum intensity projection MRI in breast cancer diagnosis using dynamic contrast-enhanced MRI. *Radiol Artif Intell* 2021;3(3):e200159.
- Antropova N, Huynh B, Li H, Giger ML. Breast lesion classification based on dynamic contrast-enhanced magnetic resonance images sequences with long short-term memory networks. *J Med Imaging (Bellingham)* 2019;6(1):011002.
- Chitalia R, Viswanath V, Pantel AR, et al. Functional 4-D clustering for characterizing intratumor heterogeneity in dynamic imaging: Evaluation in FDG PET as a prognostic biomarker for breast cancer. *Eur J Nucl Med Mol Imaging* 2021;48:3990-4001. <https://doi.org/10.1007/s00259-021-05265-8>.
- Dalmış MU, Gubern-Mérida A, Vreemann S, et al. Artificial intelligence-based classification of breast lesions imaged with a multiparametric breast MRI protocol with ultrafast DCE-MRI, T2, and DWI. *Invest Radiol* 2019;54(6):325-332.
- Pötsch N, Dietzel M, Kapetas P, et al. An a.i. classifier derived from 4D radiomics of dynamic contrast-enhanced breast MRI data: Potential to avoid unnecessary breast biopsies. *Eur Radiol* 2021;31(8):5866-5876.
- Braman NM, Etesami M, Prasanna P, et al. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathologic complete response to neoadjuvant chemotherapy based on breast DCE-MRI. *Breast Cancer Res* 2017;19(1):57.
- Cain EH, Saha A, Harowicz MR, et al. Multivariate machine learning models for prediction of pathologic response to neoadjuvant therapy in breast cancer using MRI features: A study using an independent validation set. *Breast Cancer Res Treat* 2019;173(2):455-463.
- Chamming's F, Ueno Y, Ferré R, et al. Features from computerized texture analysis of breast cancers at pretreatment MR imaging are associated with response to Neoadjuvant chemotherapy. *Radiology* 2018;286(2):412-420.
- Drukker K, Edwards A, Doyle C, Papaioannou J, Kulkarni K, Giger ML. Breast MRI radiomics for the pretreatment prediction of response to neoadjuvant chemotherapy in node-positive breast cancer patients. *J Med Imaging (Bellingham)* 2019;6(3):034502.
- Li W, Newitt DC, Gibbs J, et al. Predicting breast cancer response to neoadjuvant treatment using multi-feature MRI: Results from the I-SPY 2 TRIAL. *NPJ Breast Cancer* 2020;6(1):63.
- Tahmassebi A, Wengert GJ, Helbich TH, et al. Impact of machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy and survival outcomes in breast cancer patients. *Invest Radiol* 2019;54(2):110-117.
- Chen S, Shu Z, Li Y, et al. Machine learning-based radiomics nomogram using magnetic resonance images for prediction of neoadjuvant chemotherapy efficacy in breast cancer patients. *Front Oncol* 2020;10:1410.

24. Eun NL, Kang D, Son EJ, et al. Texture analysis with 3.0-T MRI for association of response to neoadjuvant chemotherapy in breast cancer. *Radiology* 2020;294(1):31-41.
25. Ashraf A, Gaonkar B, Mies C, et al. Breast DCE-MRI kinetic heterogeneity tumor markers: Preliminary associations with Neoadjuvant chemotherapy response. *Transl Oncol* 2015;8(3):154-162.
26. Jahani N, Cohen E, Hsieh M-K, et al. Prediction of treatment response to Neoadjuvant chemotherapy for breast cancer via early changes in tumor heterogeneity captured by DCE-MRI registration. *Sci Rep* 2019;9(1):12114.
27. Wu J, Gong G, Cui Y, Li R. Intratumor partitioning and texture analysis of dynamic contrast-enhanced (DCE)-MRI identifies relevant tumor sub-regions to predict pathologic response of breast cancer to neoadjuvant chemotherapy. *J Magn Reson Imaging* 2016;44(5):1107-1115.
28. Drisis S, El Adoui M, Flamen P, et al. Early prediction of neoadjuvant treatment outcome in locally advanced breast cancer using parametric response mapping and radial heterogeneity from breast MRI. *J Magn Reson Imaging* 2020;51(5):1403-1411.
29. Saha A, Harowicz MR, Grimm LJ, et al. "Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations" [data set]. *Cancer Imag Archiv* 2021. <https://doi.org/10.7937/TCIA.e3sv-re93>.
30. Saha A, Harowicz MR, Grimm LJ, et al. A machine learning approach to radiogenomics of breast cancer: A study of 922 subjects and 529 DCE-MRI features. *Br J Cancer* 2018;119(4):508-516.
31. Clark K, Vendt B, Smith K, et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *J Digit Imaging* 2013;26(6):1045-1057.
32. Chen W, Giger ML, Bick U. A fuzzy c-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images. *Acad Radiol* 2006;13(1):63-72.
33. Caballo M, Sanderink WBG, Han L, Gao Y, Athanasiou A, Mann RM. 4D radiomics in dynamic contrast-enhanced MRI: Prediction of pathologic complete response and systemic recurrence in triple-negative breast cancer. *Proc SPIE Med Imaging* 2022;12033:120331G.
34. Caballo M, Pangallo DR, Mann RM, Sechopoulos I. Deep learning-based segmentation of breast masses in dedicated breast CT imaging: Radiomic feature stability between radiologists and artificial intelligence. *Comput Biol Med* 2020;118:103629.
35. Caballo M, Pangallo DR, Sanderink W, et al. Multi-marker quantitative radiomics for mass characterization in dedicated breast CT imaging. *Med Phys* 2021;48(1):313-328.
36. Caballo M, Hernandez AM, Lyu SH, et al. Computer-aided diagnosis of masses in breast computed tomography imaging: Deep learning model with combined handcrafted and convolutional radiomic features. *J Med Imaging (Bellingham)* 2021;8(2):024501.
37. Whitney HM, Li H, Ji Y, Liu P, Giger ML. Harmonization of radiomic features of breast lesions across international DCE-MRI datasets. *J Med Imaging (Bellingham)* 2020;7(1):012707.
38. Saha A, Yu X, Sahoo D, Mazurowski MA. Effects of MRI scanner parameters on breast cancer radiomics. *Expert Syst Appl* 2017;87:384-391.
39. Altman DG, Bland JM. How to obtain the P value from a confidence interval. *BMJ* 2011;343:d2304. <https://doi.org/10.1136/bmj.d2304>.
40. Esserman LJ, Berry DA, Cheang MCU, et al. Chemotherapy response and recurrence-free survival in neoadjuvant breast cancer depends on biomarker profiles: Results from the I-SPY 1 TRIAL (CALGB 150007/150012; ACRIN 6657). *Breast Cancer Res Treat* 2012;132(3):1049-1062.
41. Almutlaq ZM, Wilson DJ, Bacon SE, et al. Evaluation of mono-exponential, stretched-exponential and intravoxel incoherent motion MRI diffusion models in early response monitoring to neoadjuvant chemotherapy in patients with breast cancer-a preliminary study. *J Magn Reson Imaging* 2022. <https://doi.org/10.1002/jmri.28113>.
42. Musall BC, Abdelhafez AH, Adrada BE, et al. Functional tumor volume by fast dynamic contrast-enhanced MRI for predicting neoadjuvant systemic therapy response in triple-negative breast cancer. *J Magn Reson Imaging* 2021;54(1):251-260.
43. Furman-Haran E, Nissan N, Ricart-Selma V, Martinez-Rubio C, Degani H, Camps-Herrero J. Quantitative evaluation of breast cancer response to neoadjuvant chemotherapy by diffusion tensor imaging: Initial results. *J Magn Reson Imaging* 2018;47(4):1080-1090.