International Journal of Computer Assisted Radiology and Surgery (2020) 15:921–930
https://doi.org/10.1007/s11548-020-02177-0

ORIGINAL ARTICLE

# Breast DCE-MRI radiomics: a robust computer-aided system based on reproducible BI-RADS features across the influence of datasets bias and segmentation methods

**Mengyun Qiao[1] · Chengkang Li[1] · Shiteng Suo[2] · Fang Cheng[2] · Jia Hua[2] · Dan Xue[3] · Yi Guo[1] · Jianrong Xu[2] · Yuanyuan Wang[1]**

Received: 8 January 2020 / Accepted: 21 April 2020 / Published online: 9 May 2020


## Abstract

**Purpose** A highly accurate and robust computer-aided system based on quantitative high-throughput Breast Imaging Reporting and Data System (BI-RADS) features from dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) can drive the success of radiomic applications in breast cancer diagnosis. We aim to build a stable system with highly reproducible radiomics features, which can make diagnostic performance independent of datasets bias and segmentation methods.

**Method** We applied a dataset of 267 patients including 136 malignant and 131 benign tumors from two MRI manufacturers, where 211 cases from a Philips system and 55 cases from a GE system. First, manual annotations, 3D-Unet and 2D-Unet were applied as different segmentation methods. Second, we designed and extracted 3172 features from six modalities of DCE-MRI based on BI-RADS. Third, the feature selection was conducted. Between-class distance was utilized to eliminate the effect of dataset bias caused by two machines. Concordance correlation coefficient, intraclass correlation coefficient and deviation were employed to evaluate the influence of three segmentation methods. We further eliminated features redundancy using genetic algorithm. Finally, three classifiers including support vector machine (SVM), the bagged trees and *K*-Nearest Neighbor were evaluated by their performance for diagnosing malignant and benign tumors.

**Results** A total of 246 features were preserved to have high stability and reproducibility. The final feature set showed the robust performance under these factors and achieved the area under curve of 0.88, the accuracy of 0.824, the sensitivity of 0.844, the specificity of 0.807 in differentiating benign and malignant tumors with the SVM classifier using manually segmentation results.

**Conclusion** The final selected 246 features are reproducible and show little dependence on segmentation methods and data perturbation. The high stability and effectiveness of diagnosis across these factors illustrate that the preserved features can be used for prognostic analysis and help radiologists in the diagnosis of breast cancer.

**Keywords** Radiomics · Feature reproducibility · Breast tumor · DCE-MRI

✉ Yi Guo
guoyi@fudan.edu.cn

✉ Jianrong Xu
xujianrong_renji@163.com

✉ Yuanyuan Wang
yywang@fudan.edu.cn

[1] Department of Electronic Engineering, Fudan University, Shanghai 200433, China

[2] Department of Radiology, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

[3] Shanghai Cognate Artificial Intelligence Co., Ltd., Shanghai, China

## Introduction

Breast cancer is the most frequent cancer for women and the main cause of cancer-related death all over the world. In 2018, about 268,670 new breast cancer cases are expected to occur globally, accounting for 15.4% of all new cancer cases (1,735,350) and 41,400 breast cancer deaths, accounting for 6.6% of all cancer deaths (609,640) [1]. Early detection and diagnosis of cancer are critical so that treatment and prognosis can be implemented to reduce the breast cancer death rate. An advanced medical imaging technology is a powerful evaluation tool in this field [2]. Many clinical studies have shown that the dynamic contrast-enhanced magnetic

resonance imaging (DCE-MRI) is very important for the diagnosis of benign and malignant breast cancer due to the dynamic enhancement of breast lesions [3].

The Breast Imaging Reporting and Data System (BI-RADS) published by the American College of Radiology is to give a comprehensive and standardized description of breast tumors and offers the guidance for radiologists to categorize breast lesions [4]. For the DCE-MRI, BI-RADS provides seven assessment categories and includes four descriptors in masses, which are the shape, margin, internal enhancement characteristics and time-signal intensity curve (TIC) description [5]. The diagnoses highly rely on radiologists since they make subjective predictable assessments based on BI-RADS characteristics. Thus, a robust computer-aided system based on BI-RADS is extremely important for widespread use to help improve radiologists' diagnosis performance.

However, the stability of a computer-aided system has been influenced by several factors including the dataset bias from machines, segmentation methods, feature sets and classifiers. First, machines from different manufacturers use their own image acquisition and reconstruction schemes, causing various gray distributions of DCE-MR images. Second, since the breast tumor occurs in multiple layers of MRI and borders are blurred, different radiologists outline incompletely coincident boundaries. Besides, automatic methods may neither achieve predictable tumor contours. The variability in data perturbations and segmentation methods may lead to large uncertainty in BI-RADS features. Thus, the stability and reproducibility of radiomics features across different scans and segmentation methods should be investigated before these features are used in a computer-aided diagnosis system.

Our study aims to build a robust computer-aided system based on radiomics features with high reproducibility, stability and classification ability across dataset bias, segmentation methods and classifiers. The experiments were developed on 3172 high-throughput DCE-MRI features extracted from 267 breast cases from two machines (Philips and GE) based on tumor regions segmented by three different methods (manual annotations, 3D U-Net, 2D U-Net). The diagnostic performances of distinguishing malignant and benign tumors were evaluated on the final selected stable feature set by three classifiers [the support vector machine (SVM), the bagged trees and $K$-Nearest Neighbor (KNN)].

The innovations of our proposed methods mainly include four aspects. (1) We use breast MRI BI-RADS to standardize high-throughput features for a detailed and comprehensive description of breast tumors in the DCE-MRI. (2) A stable and reproducible feature set is selected against the dataset bias from machines and segmentation methods. (3) Four metrics including concordance correlation coefficient (CCC), intraclass correlation coefficient (ICC), deviation (Dev) and between-class distance (BD) are applied for features reproducibility and stability evaluation. (4) Experimental results demonstrate that the selected feature set shows stable and great performance of distinguishing benign and malignant tumors by different classifiers.

The proposed method included four parts: image segmentation, radiomics features extraction, features selection and tumor diagnosis. Figure 1 shows the flow chart of the entire method.

## Materials and methods

### Dataset

A total of 267 patients with breast tumors include 211 cases from a Philips 3-T system (Achieva or Ingenia; Philips Medical Systems, Best, the Netherlands) and 55 cases from a GE 3-T system (Signa HDxt; GE Medical Systems, Milwaukee, WI, USA) were acquired in Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China from Jan. 2018 to June 2019. The datasets summary is present in Table 1. The dynamic series consisted of five individual dynamic phases: one was obtained before and four after the rapid bolus intravenous injection of 0.1 mmol of gadopentetate dimeglumine per kilogram of body weight and a 10-mL saline solution flush. Each phase was imaged by a 1-min interval. The image size range is from $384 \times 384 \times 150$ to $672 \times 672 \times 150$ mm. The voxel spacing range is from $0.5652 \times 0.5652 \times 0.99$ to $0.8102 \times 0.8102 \times 1$ mm. All images are collected with the institutional review board approval, including a waiver of informed consent.

### Tumor segmentation

Segmentation is an important part of feature extraction since many features are based on the mass region. The slight difference of tumor boundaries among segmentation methods may lead to feature values' huge varieties. To eliminate the influence of segmentation methods, we compare the manual result annotated by the high-experienced radiologist and automatic results by a state-of-the-art deep learning-based method named no-new-Net (nnU-Net) [6].

For the manual segmentation method, a high-experienced radiologist annotated the contour of each tumor by the software Ziosoft (Ziosoft, Inc., Tokyo, Japan), which is considered as the ground truth.

For automatic segmentation methods, the nnU-Net is applied including two steps: preprocessing and deep learning model with the first post-contrast phase of breast DCE-MRI series as input. Firstly, the preprocessing includes cropping original images to the region of nonzero values, resampling to the same voxel spacing in three dimensions and zero-mean (z-score) normalization [7]. Then, to access dif-

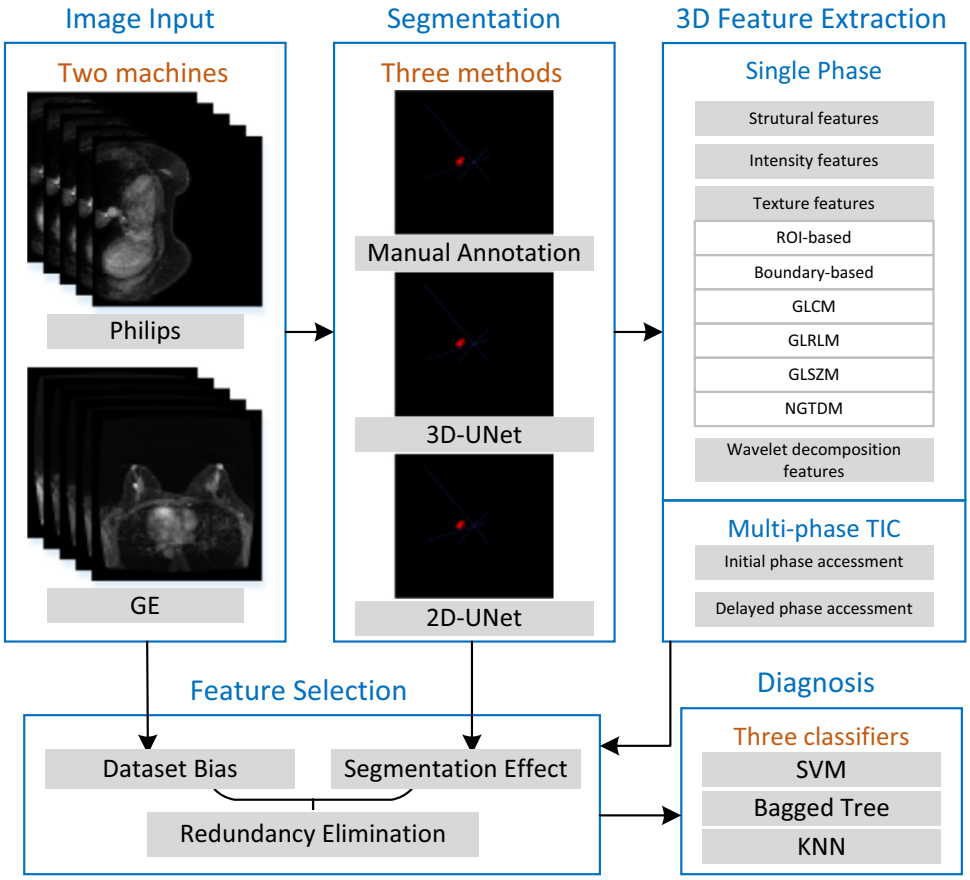


**Fig. 1** The workflow of the proposed radiomics method, including data input, feature selection, 3D feature extraction and diagnosis. The data input consist of images from two machines and segmentation from three methods

**Table 1** Dataset summary

| Machine | Benign cases | Malignant cases | Total |
|---|---|---|---|
| Philips | 99 | 112 | 211 |
| GE | 32 | 23 | 55 |
| All | 131 | 136 | 267 |

ferent segmentation results, we adopt two deep learning models that include 3D U-Net and 2D U-Net [8] for the comparison. These two networks apply the same convolutional encoder–decoder architecture with 3D convolutions and 2D convolutions, respectively. The shrink path which is also considered as an encoder captures the image context which comprises six convolution layers. The first convolution layer uses two convolutional blocks consisting of convolution, instance normalization and Leaky rectified linear units (ReLU). The strided convolutional block is utilized to replace max pooling in the next four layers. The extended path which is considered as a decoder is employed for precise positioning with the transposed convolution and constructed by five layers. The first four layers consist of the convolutional block and the strided convolutional block. The last softmax layer to classify each pixel includes two convolutional blocks and one convolution ($1 \times 1 \times 1$ for 3D U-Net and $1 \times 1$ kernel size for 2D U-Net). The other convolution is with the kernel size of $3 \times 3 \times 3$ for 3D U-Net and $3 \times 3$ for 2D U-Net). To transfer information that may be lost in the encoder path, the skip connections are utilized in the concatenation of outputs from the encoder path and the output of subsequent layers.

## High-throughput feature extraction

*Image phases* We extract features from six DCE-MRI phases including one non-contrast phase, four post-contrast phases and one designed time-intensity signal map. An example of six phases is shown in Fig. 2.

The designed new MRI modality time-intensity signal map is inspired by TIC-related characteristic to comprehensively reflect the tumor enhancement. The calculation process is similar to the TIC-related feature but based on each pixel value rather than only the mass region from five phases. The specific method is as follows. The enhancement rate (percentage of signal intensity increase) is qualified by the formula:

$$S_{\text{tic}} = \frac{S_{\text{post}} - S_{\text{non}}}{S_{\text{non}}} \times 100 \quad (1)$$

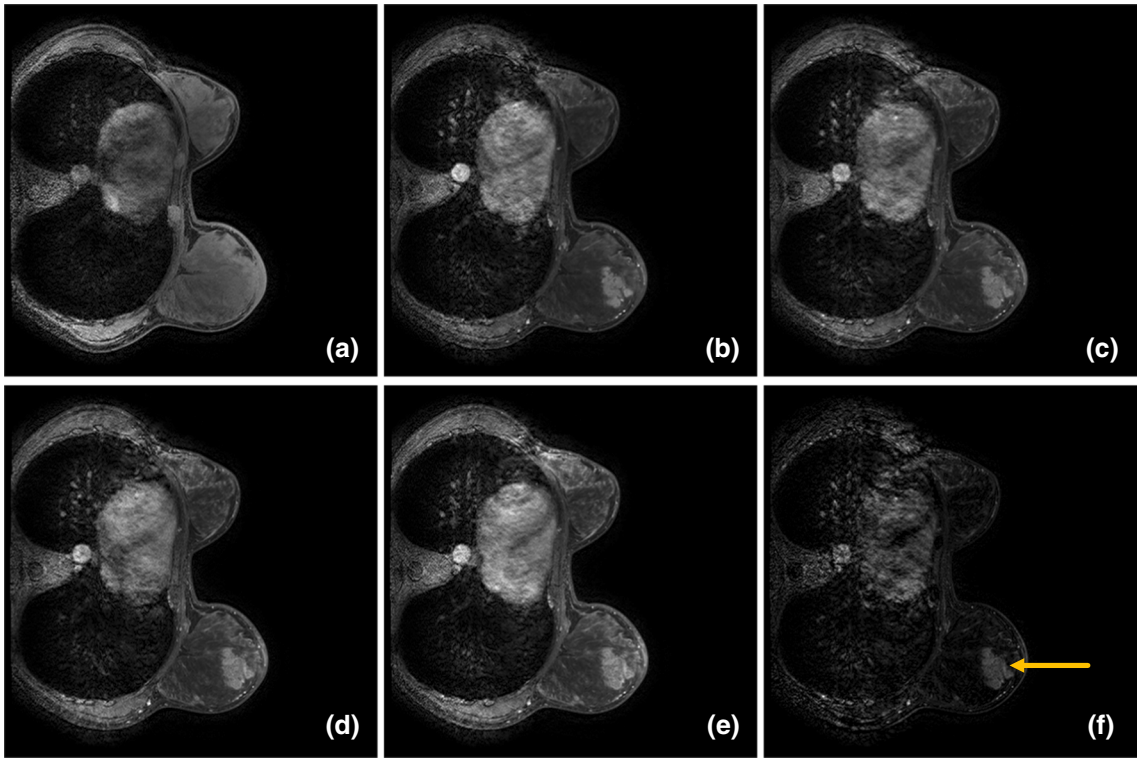


**Fig. 2** An example of breast DCE-MRI showing a strongly enhancing lesion in the under outer quadrant (yellow arrow). **a** Non-contrast phase from DCE-MRI; **b**–**e** post-contrast phases after enhancement; **f** the designed Time-signal Intensity Map

where $S_{post}$ and $S_{non}$ are the post-contrast and non-contrast signal intensities, respectively. The estimation applies the linear regression, where the mean value of the mass region in $S_{post}$ is the response data calculated by Eq. (1) and the image acquisition time is the predictor data. The coefficient estimation from multiple linear regression is the value of each pixel on the time-signal intensity map.

*Feature extraction* The extracted high-throughput 3D features are designed based on breast MRI BI-RADS description, as illustrated in Table 2. According to MRI breast BI-RADS descriptors, 528 features are extracted in each phase separately and four features are calculated by non-contrast and post-contrast phases. In total, we extracted 3172 radiomic features that are grouped into four main categories including 18 intensity features, 39 texture features, 15 structural features, 456 wavelet feature and 4 TIC features. Table 3 illustrates all breast MRI BI-RADS high-throughput features.

Intensity features reflect histogram distributions and gray levels. The total 18 features are acquired in the mass region from the original MRI phases. Structural features are employed to reflect the shape, margin and internal enhancement characteristics as introduced in BI-RADS as shown in Table 2.

Texture features describe the gray-level variations of images in detail. Four types of texture features are categorized to highlight different tumor features that may not be visible in the original image [9]. Gray-level co-occurrence matrix (GLCM) [10] texture features reflect the specified

**Table 2** A brief BI-RADS description of breast tumor in MRI

| Focus | Category | Description |
|---|---|---|
| Masses | Shape | Oval |
| | | Round |
| | | Irregular |
| | Margin | Circumscribed |
| | | Not circumscribed -Irregular - Spiculated |
| | Internal enhancement characteristics | Homogeneous |
| | | Heterogeneous |
| | | Rim enhancement |
| | | Dark internal septations |
| Kinetic curve assessment | Initial phase | Slow |
| | | Medium |
| | | Fast |
| | Delayed phase | Persistent |
| | | Plateau |
| | | Washout |

**Table 3** Summary of the quantitative high-throughput features

| Type | Name | | Number |
|---|---|---|---|
| Intensity | Energy, h-entropy, kurtosis, max, mean absolute deviation, mean, media, min, range, root mean square, skewness, standard deviation, h-uniformity, variance, h-mean, h-variance, h-skewness, h-kurtosis | | 18 |
| Texture | GLCM(8) | Energy, contrast, correlation, homogeneity, variance, sum average, entropy, dissimilarity | 39 |
| | GLRLM(13) | Short run emphasis, long run emphasis, gray-level nonuniformity, run-length nonuniformity, run percentage, low gray-level run emphasis, high gray-level run emphasis, short run low gray-level emphasis, short run high gray-level emphasis, long run low gray-level emphasis, long run high gray-level emphasis, gray-level variance, run-length variance | |
| | GLSZM(13) | Small zone emphasis, large zone emphasis, gray-level nonuniformity, zone-size nonuniformity, zone percentage, low gray-level zone emphasis, high gray-level zone emphasis, small zone low gray-level emphasis, small zone high gray-level emphasis, large zone low gray-level emphasis, large zone high gray-level emphasis, gray-level variance, zone-size variance | |
| | NGTDM(5) | coarseness, contrast, busyness, complexity, strength | |
| Structure | Compactness, compactness-square, max-length, spherical disproportion, sphericity, superficial-area, surface to volume ratio, volume, region to bounding-box ratio, max major-length, min -length, eccentricity, orientation, solidity, Fourier-descriptors | | 15 |
| Wavelet | LLL HLL LHL HHL LLH HLH LHH HHH decomposition | | 456 |
| TIC | Initial phase regression, delayed phase regression | | 4 |

spatial linear relationships between the frequencies of two gray levels within a certain range. Gray-level run-length matrix (GLRLM) [11, 12] checks the runs of a set of consecutive collinear image points with the same gray value in a given direction, which describes the coarseness of the texture. Gray-level size zone matrix (GLSZM) [13] provides a statistical representation by estimating a binary conditional probability density function of image distribution values. Neighborhood gray-tone difference matrix (NGTDM) measures the gradation of each pixel to its grayscale difference between adjacent pixels to describe the spatial changes in the intensity or dynamic range of intensity [14].

To reflect more detailed information of images, wavelet decomposition features are introduced to decompose a two-dimensional image into four components which are Low pass/Low pass (LL), Low pass/High pass (LH), High pass/Low pass (HL) and High pass/High pass (HH). Each component containing 114 features, and there are 456 wavelet features in total [9].

In addition, four features are designed based on the TIC description in BI-RADS to evaluate the relative enhancement before and after the injection of gadopentetate dimeglumine. The kinetic curve assessment includes the initial phase and delayed phase, which are calculated in the mass region. The enhancement of the former two post-contrast phases and the former three post-contrast phases are assessed as the initial phase enhancement description. The enhancement of the last three post-contrast phases and all post-contrast phases are assessed as the delayed phase enhancement description.

## Feature selection

### Feature selection from dataset bias

The radiomic-based approach ought to be robust against various machines. However, different machines may cause the intensity and noise discrepancy. The distribution of each feature is utilized to eliminate the influence of two frequently used machines: Philips and GE. The between-class distance (BD) is the normalized distance between two feature sets and employed to measure the distribution differences:

$$\mathrm{BD}_i = \frac{\left|\mu_{F_{\mathrm{manu}_i}} - \mu_{F_{\mathrm{auto}_i}}\right|}{\sqrt{\sigma^2_{F_{\mathrm{manu}_i}} + \sigma^2_{F_{\mathrm{auto}_i}}}} \tag{2}$$

where $\mu_{F_{\mathrm{manu}_i}}$ 和 $\sigma^2_{F_{\mathrm{manu}_i}}$, $\mu_{F_{\mathrm{auto}_i}}$ 和 $\sigma^2_{F_{\mathrm{auto}_i}}$ are the mean and standard deviation of the $i$th feature for Philips and GE images. In this study, a relatively high BD value means that features calculated from these two machines are dissimilar and non-repeatable. Elements with a BD value of less than 0.2 are defined as reproducible features.

### Feature selection from segmentation methods

The automatic results acquired by nnU-Net and the manual segmentation by radiologists are applied to assess the feature stability. This step is to eliminate the effect of segmentation methods. Three metrics are calculated to evaluate the features' similarities.

The first metric is the concordance correlation coefficient (CCC) [15] to measure the agreement between two variables, defined as:

$$\mathrm{CCC} = \frac{2S_{ab}}{S_a^2 + S_b^2 + (\bar{a} - \bar{b})^2} \tag{3}$$

where $\bar{a}$ and $\bar{b}$ are the mean values of variables $a$ and $b$. $S_a^2$ and $S_b^2$ are the corresponding variances. $S_{ab}$ is a correlation coefficient between $a$ and $b$. Elements with a CCC value of more than 0.9 are defined as a high agreement between features.

The second metric is deviation (Dev), which demonstrates the relative differences between manual segmentation features and automatic segmentation features. The average deviation of the $i$th feature $\mathrm{Dev}_i$ is defined as follows:

$$\mathrm{Dev}_i = \frac{1}{N}\sum_{n=1}^{N}\frac{\left|F_{\mathrm{manu}_{n,i}} - F_{\mathrm{auto}_{n,i}}\right|}{\left|F_{\mathrm{manu}_{n,i}}\right|}, \quad i = 1, 2, \ldots, I \tag{4}$$

where $F_{\mathrm{manu}_{n,i}}$ and $F_{\mathrm{auto}_{n,i}}$ are the $i$th feature extracted by manual segmentation and automatic segmentation for the $n$th patient, respectively. $N$ is the number of all cases, which is 267 in our work. $I$ is the feature number, which is 3172. Elements with a Dev value less than 0.1 are considered as the low differences between features.

The third metric is the intraclass correlation coefficient (ICC) [16], which describes how strongly units in the same group resemble each other in statistics. The features consist of $N$ paired data values for manual segmentation and automatic segmentation. ICC is defined as:

$$\mathrm{ICC}_i = \frac{1}{Ns_i^2}\sum_{n=1}^{N}\left(F_{\mathrm{manu}_{n,i}} - \bar{F}_i\right)\left(F_{\mathrm{auto}_{n,i}} - \bar{F}_i\right) \tag{5}$$

where

$$\bar{F}_i = \frac{1}{2N}\sum_{n=1}^{N}\left(F_{\mathrm{manu}_{n,i}} + F_{\mathrm{auto}_{n,i}}\right) \tag{6}$$

$$s_i^2 = \frac{1}{2N}\left\{\sum_{n=1}^{N}\left(F_{\mathrm{manu}_{n,i}} - \bar{F}_i\right)^2 + \sum_{n=1}^{N}\left(F_{\mathrm{auto}_{n,i}} - \bar{F}_i\right)^2\right\} \tag{7}$$

Elements with an ICC value of more than 0.9 are considered as a high similarity between features.

### Redundancy elimination

It is also necessary to eliminate redundancies of high-throughput features. The genetic algorithm (GA) is a stochastic optimization process of natural selection and genetic variation during simulating biological evolution. The mRMR algorithm is an approximation of the theoretically best dependent feature selection algorithm. It maximizes the mutual information between the selected feature's joint distribution and categorical variables, therefore enabling the genetic algorithm to operate this function at a very low cost. Here, we combine the genetic algorithm (GA) method and the minimal-redundancy-maximal-relevance (mRMR) to reduce feature redundancy and select a feature set that maximizes the relevance and minimizes the redundancy [17]. Finally, a stable and representative feature set for the DCE-MRI is preserved.

### Breast tumor diagnosis

The final selected features were then fed into classifiers to verify the efficiency of distinguishing benign and malignant tumors in breast MR images. Features are extracted on three segmentation results of the same dataset, respectively, to evaluate the influence of segmentation methods on the final stable feature set. Also, cases from two machines are compared to access the reproducibility of the final feature set for the dataset bias. Three classifiers including the support vector machine (SVM) classifier [18], the bagged trees [19] and $K$-Nearest Neighbor (KNN) [20] are employed to eliminate the influence of classifiers and verify the robust performance for different situations, as presented in Fig. 1.

## Experiments and results

### Evaluation metrics

The Dice score metric is used to evaluate the accuracy of segmentation results, as defined in Eq. (1).

$$\mathrm{Dice}\left(bw_{\mathrm{gt}}, bw_{\mathrm{seg}}\right) = \frac{2\left|bw_{\mathrm{gt}} \cap bw_{\mathrm{seg}}\right|}{\left|bw_{\mathrm{gt}}\right| + \left|bw_{\mathrm{seg}}\right|} \tag{8}$$

Five metrics are used to evaluate the overall performance for these classifiers including the area under the ROC curve (AUC), accuracy (ACC), sensitivity (SENS), specificity (SPEC) and precision (PREC):

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}, \quad \mathrm{SENS} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$
$$\mathrm{SPEC} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \quad \mathrm{PREC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

where $TP$ and $FN$ represent the number of correctly and incorrectly classified malignant tumors, $TN$ and $FP$ refer to the number of correctly and incorrectly classified benign tumors, respectively.

All images processing was performed on MATLAB R2018b (MathWorks, Inc., Natick, MA, USA).

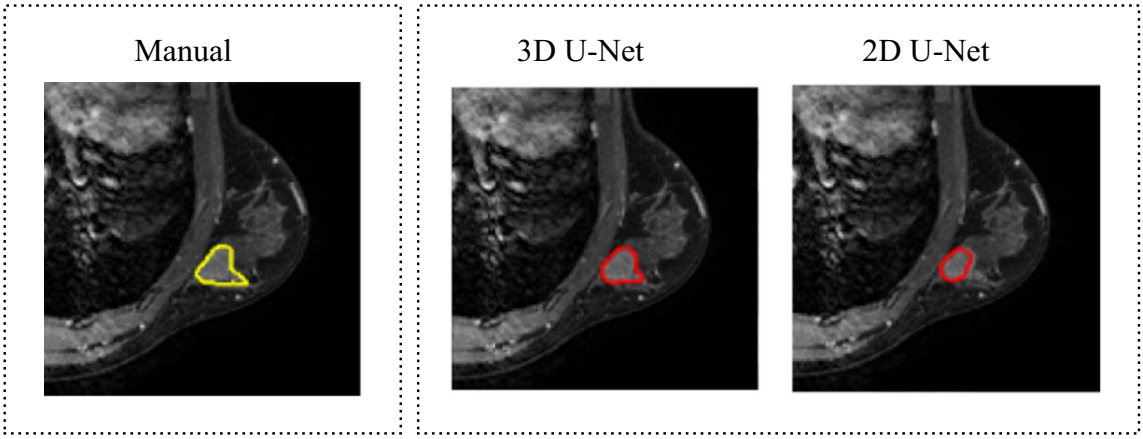


**Fig. 3** An example of breast DCE-MRI segmentation by three methods: manual annotation by the radiologist, nnU-Net based on 3D U-Net and 2D U-Net models

## Tumor segmentation

We used three segmentation methods including radiologist annotations, nnU-Net with 3D U-Net and 2D U-Net model. Figure 3 shows the result comparison for the same DCE-MRI. The network used the Adam optimizer, a batch size of 2 and 19 for 3D U-Net and 2D U-Net, respectively, and epochs of 250 iterations. In addition, the adaptive adjustment strategy was utilized for the learning rate in the training process, where the initial learning rate was $3 \times 10^{-4}$ and was reduced by a factor of 0.2 if the training loss no longer improved after 30 epochs. The training stopped if there was no improvement in the loss after 50 epochs. During the experiments, we conducted a variety of data augmentation techniques on our limited training data, including mirroring, random scaling, gamma correction augmentation, random rotations and random elastic deformations.

Manual segmentation is regarded as the ground truth, while the automatic segmentation methods nnU-Net with 3D U-Net and 2D U-Net model achieved the dice score of 0.8 and 0.7, respectively. They were applied to extract features and compare their diagnostic performance to verify that the diagnosis performance of the selected feature set is not affected by segmentation results.

## Feature selection

First, to evaluate the influences of different machines, 211 cases from Philips and 55 cases from GE are employed. The lower BDs result in more reproducible features under the effect of machines. A total of 954 features were found to be reproducible (BD < 0.2). The gray-intensity-related features are mostly eliminated such as the intensity, texture and wavelet features. Second, to eliminate the influence of segmentation methods, features whose CCC > 0.9, Dev < 0.1, ICC > 0.9 were preserved as reproducible features, resulting in 967 features remaining. The results show that shape-related and margin-related features were more likely to be affected by segmentation boundaries. Then, the intersection set of above features remained 563 features, which was then eliminated redundancy by the GA. For GA algorithm, the number of individuals is set as 50, the maximum genetic generation is 30, the binary value of variables is 24 and the generation gap is 0.9. Finally, 246 features were preserved as illustrated in Table 4, which can comprehensively describe all BI-RAD categories.

## Diagnostic performance

There are three main experimental factors, which can potentially affect the prediction of radiomic-based tumor classification: segmentation method, dataset bias and classifiers. The training and testing cases were randomly selected from all datasets, accounting for 80% and 20% of all patients, respectively.

### The stability comparison of different feature set

The diagnosis performance experiments to compare the stability of feature set were conducted based on three features including all 3172 features, 563features before redundancy elimination by GA and the final selected 246 features. The classifier applied SVM with Linear kernel. The results are presented in Table 5.

### The influence of segmentation methods and classifiers

The diagnosis performance experiments to classify tumors into benign and malignant ones were conducted based on the final selected 246 features. The evaluation of features reproducibility and stability includes two machines, three segmentation methods and three classifiers. First, manual

**Table 4** The final stable feature set and the links between the remaining features and BI-RADS descriptors

| Descriptors | Features | Numbers |
|---|---|---|
| Shape | max major-length, max-length, min-length, orientation, spherical disproportion, superficial-area, volume | 18 |
| Margin | Variance of annular region; Annular region SNR; Std of annular region; Variance of annular region | 15 |
| Internal enhancement characteristics | Kurtosis, mean, energy, entropy, mean, variance, kurtosis, max, mean absolute deviation, media, range, root mean square, solidity, standard deviation, variance, GLCM-contrast, GLCM-correlation, GLCM-dissimilarity, GLCM-entropy, GLCM-homogeneity, GLCM-sum average, GLCM-variance, GLRLM-GLN, GLRLM-GLV, GLRLM-HGRE, GLRLM-LRE, GLRLM-LRHGE, GLRLM-LRLGE, GLRLM-RLN, GLRLM-RLN, GLRLM-SRHGE, GLRLM-SRLEG, GLSZM-LZE, GLSZM-SZE, GLSZM-SZHGE, GLSZM-ZP, GLSZM-ZSN, GLSZM-ZSV, NGTDM-busyness, NGTDM-coarseness, NGTDM-complexity, NGTDM-strength | 211 |
| Initial phase | Regression of the first three phases | 1 |
| Delayed phase | Regression of the last three phases | 1 |

The same feature names extracted by different phases or wavelets are only listed once
All abbreviations are used by Initials of features in Table 3

**Table 5** The diagnostic performance of different feature set with three segmentation results by the classifier SVM

| Classifier | Ground truth | Segmentation dice = 0.8 | Segmentation dice = 0.7 |
|---|---|---|---|
| 3172 features | | | |
| ACC | 0.775 | 0.749 | 0.726 |
| SPEC | 0.753 | 0.714 | 0.674 |
| PREC | 0.717 | 0.648 | 0.549 |
| 563 features | | | |
| ACC | 0.782 | 0.786 | 0.782 |
| SPEC | 0.763 | 0.768 | 0.774 |
| PREC | 0.732 | 0.74 | 0.755 |
| 246 features | | | |
| ACC | 0.824 | 0.820 | 0.813 |
| SPEC | 0.807 | 0.810 | 0.80 |
| PREC | 0.786 | 0.794 | 0.79 |

**Table 6** The comparison results of different segmentation results on three classifiers

| Classifier | Ground truth | Segmentation dice = 0.8 | Segmentation dice = 0.7 |
|---|---|---|---|
| SVM | | | |
| AUC | 0.88 | 0.88 | 0.87 |
| ACC | 0.824 | 0.820 | 0.813 |
| SENS | 0.844 | 0.832 | 0.830 |
| SPEC | 0.807 | 0.810 | 0.80 |
| PREC | 0.786 | 0.794 | 0.79 |
| Bagged trees | | | |
| AUC | 0.82 | 0.84 | 0.85 |
| ACC | 0.783 | 0.771 | 0.779 |
| SENS | 0.783 | 0.802 | 0.795 |
| SPEC | 0.783 | 0.748 | 0.765 |
| PREC | 0.771 | 0.710 | 0.740 |
| KNN | | | |
| AUC | 0.82 | 0.82 | 0.82 |
| ACC | 0.760 | 0.756 | 0.749 |
| SENS | 0.759 | 0.779 | 0.754 |
| SPEC | 0.761 | 0.738 | 0.744 |
| PREC | 0.748 | 0.702 | 0.725 |

and automatic segmentation results were utilized to extract features and verify that the diagnosis performance of the selected feature set is not affected by segmentation results. Second, three classifiers were utilized to validate the adaptation ability of the final feature set to different classifiers. The kernel function applied in SVM is Linear kernel. For KNN, the number of neighbors is set as 10, and distance metric is cosine where the distance weight is equal. The number of Bagged trees learners is 30. The results are presented in Table 6. Manual segmentation is regarded as the ground truth, while the 3D U-Net and 2D U-Net model achieved the dice score of 0.8 and 0.7. Figure 4 shows the ROC curve of different classifiers on the same segmentation results. When applying the same classifier, the diagnostic performances show slight differences among ground truth-based features and automatic segmentation-based features. The reproducible features show great and stable discriminations between benign and malignant tumors under the circumstances of different segmentation results, indicating that they are effective and robust in breast tumor diagnosis.

### The influence of dataset bias

The performance of DCE-MRI cases from two machines was also compared. For a fair comparison, manual segmentation and SVM classifiers were applied. The experimental results are presented in Table 7. Our method achieved the AUC of 0.88 and 0.87 for Philips and GE cases. The final feature set held a similar classification performance on machines of Philips and GE, which demonstrates that the selected features have great reproducibility and stability even with the influence of dataset perturbations from machines.

**Table 7** The comparison diagnostic performance of the same feature set on different machines applying the manual segmentation and SVM

| Machine | AUC | ACC | SENS | SPEC | PREC |
|---|---|---|---|---|---|
| Philips | 0.88 | 0.829 | 0.835 | 0.825 | 0.802 |
| GE | 0.87 | 0.811 | 0.869 | 0.767 | 0.741 |
| All | 0.88 | 0.824 | 0.844 | 0.807 | 0.786 |

## Discussion

We conducted the feature selection to eliminate the influence of dataset bias and segmentation methods and the diagnosis of malignant and benign tumors to verify the effectiveness of our selection strategy. For different segmentation methods experiments, the texture and ROI-based features are of high repeatability and stability mainly since the calculated area is mostly related to the mass inside the region. Some GLCM, GLRLM and GLSZM functions meet the cutoffs, which can be explained as the gray matrix reflects the intensity change of the entire tumor area. Apart from them, other features are easily affected by the segmentation results. Boundary-related and structural features show low repeatability because these features reflect the relative differences between the outside and inside of tumor boundaries which are sensitive to contours [9]. From Table 5, all extracted features are not stable with different segmentation results as input and achieve unstable diagnostic performance. After feature selection of eliminating the influence of segmentation methods, the 563 features calculated by different segmentation results achieve similar diagnostic performance with ACC of 0.78. However, without redundancy elimination by GA, the performance shows relatively low ACC, SPEC and PREC compared with the final selected feature set as input.

The segmentation method achieved the dice score of 0.8 and 0.7 for uuU-Net with 2D U-Net and 3D U-Net models. It shows that 3D models perform better than 2D models since in the 3D network, one case is processed as a single subject and the continuous information of adjacent slices can be accessed in a 3D convolution of networks. The diagnosis experiments illustrate that even the segmentation shows discriminations which can be seen in Fig. 3, and the final feature set still maintains the great classification performance. Also, the diagnosis performance of different segmentation methods on three classifiers experiments presents similar results which show that the feature set is stable and robust for each classifier. From Table 6, SVM shows better performance under the situation of groudtruth contours with the AUC of 0.88 than Bagged tree and KNN with the same AUC of 0.82. For segmentation results with the dice score of 0.8, the diagnostic performance is increased from 0.82, 0.84 to 0.88 by applying different classifiers of KNN, Bagged Tree and SVM, respectively.

For dataset bias experiments, machines mainly affect gray intensities of the DCE-MRI. Therefore, the intensity, texture and wavelet features are the main concerns for eliminating the influence of machines. The texture features and the intensity-related features such as the grayscale and histogram are easily affected by machines and relatively more unreproducible. Compared with intensity and texture features, the performance of wavelet features is still poor. The diagnosis

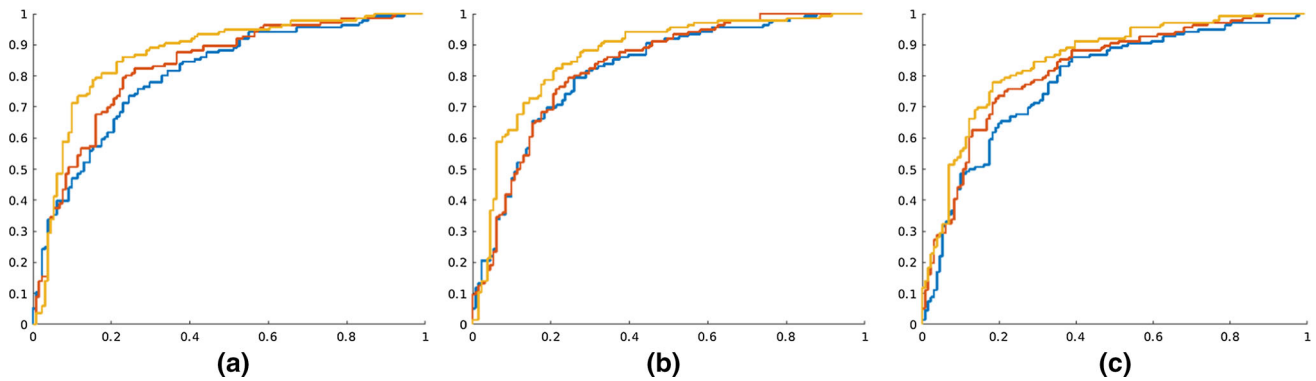


**Fig. 4** ROC curves of different classifiers (yellow: SVM; red: Bagged trees; blue: KNN) by the final feature set extracted on segmentation of **a** ground truth; **b** 3D U-Net, dice = 0.8; **c** 2D U-Net, dice = 0.7

performance of different machine shows slight discrimination, illustrating that our final feature set is robust and solves the problem of dataset bias. The stability of the machine makes the high-throughput BI-RADS function a possibility for future classification or prognosis in multicenter clinical diagnosis.

## Conclusion

In this work, we proposed an effective, robust and stable breast tumor diagnosis system with little dependency on the segmentation methods and dataset perturbations. To assess the uncertainty of quantitative imaging features extracted from the DCE-MRI, we conducted the feature selection across three segmentation methods, two machines and three classifiers. The persevered features can give a comprehensive description of breast MRI BI-RADS. In addition, the classification experiments of malignant and benign tumors demonstrate that our reproducible features have high stability and great diagnostic performance. These BI-RADS features could be used for breast tumor analysis in the future. Our future work will focus on the reproducible features' application and the combination of deep learning and radiomic-based methods for breast tumor analysis in the DCE-MRI. Overall, our variability analysis of reproducible quantitative BI-RADS features is a step forward toward the enhancements of radiomic-based clinical predictions.

**Acknowledgements** This work was supported by the National Natural Science Foundation of China (Grant 61871135, 81627804 and 81830058) and the Science and Technology Commission of Shanghai Municipality (Grant 18511102901, 18511102904, 17411953400).

### Compliance with ethical standards

**Conflict of interest** We have no conflict of interest to declare.

**Ethical approval** All procedures performed in studies involving human participants were following the ethical standards of the institutional and/or national research committee and with the Declaration of Helsinki. Informed consent was obtained from all individual participants included in the study. This study has been approved by the ethics committee of the Renji Hospital.

## References

1. Siegel RL, Miller KD, Jemal A (2018) Cancer statistics, 2018. CA Cancer J Clin 60(5):277–300
2. Milosevic M, Jankovic D, Milenkovic A, Stojanov D (2018) Early diagnosis and detection of breast cancer. Technol Health Care 26(4):729–759
3. Hara N, Okuizumi M, Koike H, Kawaguchi M, Bilim V (2010) Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is a useful modality for the precise detection and staging of early prostate cancer. Prostate 62(2):140–147
4. Morris EA, Comstock CE, Lee CH (2013) ACR BI-RADS® Magnetic resonance imaging. In: ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. American College of Radiology, Reston, VA, pp 127–143
5. Agrawal G, Su MY, Nalcioglu O, Feig SA, Chen JH (2009) Significance of breast lesion descriptors in the ACR BI-RADS MRI lexicon. Cancer 115:1363–1380
6. Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger PF, Kohl S (2018) Nnu-net: self-adapting framework for u-net-based medical image segmentation
7. David PM, Rusty OB (2013) Improving cross-device attacks using zero-mean unit-variance normalization. J Cryptogr Eng 3(2):99–110
8. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, vol 1, pp 234–241
9. Hu Y, Qiao M, Guo Y, Wang Y, Yu J, Li J, Chang C (2017) Reproducibility of quantitative high-throughput BI-RADS features extracted from ultrasound images of breast cancer. Med Phys 44(7):3676–3685
10. Haralick RM, Shanmugam K, Dinstein I (1973) Textural features for image classification. Stud Media Commun (SMC) 3(6):610–621
11. Chu A, Sehgal CM, Greenleaf JF (1990) Use of gray value distribution of run lengths for texture analysis. Pattern Recognit Lett 11(6):415–419
12. Galloway MM (1975) Texture analysis using gray level run lengths. Comput Graph Image Process 4(2):172–179
13. Thibault G, Fertil B, Navarro C, Pereira S, Cau P, Levy N, Sequeira J, Mari J (2009) Texture indexes and gray level size zone matrix application to cell nuclei classification. In: 10th International conference on pattern recognition and information processing (Atlantic City, New Jersey), pp 140–145
14. Amadasun M, King R (1989) Textural features corresponding to textural properties. IEEE Trans Syst Man Cybern 19(5):1264–1274
15. Lin IK (1989) A concordance correlation coefficient to evaluate reproducibility. Biometrics 45(1):255–268
16. McGraw Kenneth O, Wong SP (1996) Forming inferences about some intraclass correlation coefficients. Psychol Methods 1(1):30–46
17. Qiao M, Hu Y, Guo Y, Wang Y, Yu J (2018) Breast tumor classification based on a computerized breast imaging reporting and data system feature system. J Ultrasound Med 37(2):403–415
18. Rebentrost P, Mohseni M, Lloyd S (2013) Quantum support vector machine for big feature and big data classification. Phys Rev Lett 113(13):130503
19. Sexton J, Laake P (2009) Standard errors for bagged and random forest estimators. Comput Stat Data Anal 53(3):801–811
20. Islam MJ, Wu QMJ, Ahmadi M, Sid-Ahmed MA (2010) Investigating the performance of Naive-Bayes classifiers and K-nearest neighbor classifiers. J Converg Inf Technol 5(2):133–137