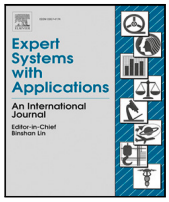
Expert Systems With Applications 249 (2024) 123557

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Breast cancer classification through multivariate radiomic time series analysis in DCE-MRI sequences

Francesco Prinzi [a,b], Alessia Orlando [c], Salvatore Gaglio [d,e], Salvatore Vitabile [a,*]

[a] Department of Biomedicine, Neuroscience and Advanced Diagnostics (BiND), University of Palermo, Palermo, Italy
[b] Department of Computer Science and Technology, University of Cambridge, Cambridge CB2 1TN, UK
[c] Section of Radiology - Department of Biomedicine, Neuroscience and Advanced Diagnostics (BiND), University Hospital "Paolo Giaccone", Palermo, Italy
[d] Department of Engineering, University of Palermo, Palermo, Italy
[e] Institute for High-Performance Computing and Networking, National Research Council (ICAR-CNR), Palermo, Italy

ARTICLE INFO

Keywords:
Radiomics
Machine learning
Time-series analysis
Explainable AI

ABSTRACT

Breast cancer is the most prevalent disease that poses a significant threat to women's health. Despite the Dynamic Contrast-Enhanced MRI (DCE-MRI) has been widely used for breast cancer classification, its diagnostic performance is still suboptimal. In this work, the Radiomic workflow was implemented to classify the whole DCE-MRI sequence based on the distinction in contrast agent uptake between benign and malignant lesions. The radiomic features extracted from each of the seven time instants within the DCE-MRI sequence were fed into a multi-instant features selection strategy to select the discriminative features for time series classification. Several time series classification algorithms including Rocket, MultiRocket, K-Nearest Neighbor, Time Series Forest, and Supervised Time Series Forest were compared. Firstly, a univariate classification was performed to find the five most informative radiomic series, and then, a multivariate time series classification was implemented via a voting mechanism. The Multivariate Rocket model was the most accurate (Accuracy = 0.852, AUC-ROC = 0.852, Specificity = 0.823, Sensitivity = 0.882). The intelligible radiomic features enabled model findings explanations and clinical validation. In particular, the Energy and TotalEnergy were among the most important features, and the most descriptive for the change in signal intensity, which is the main effect of the contrast agent.

1. Introduction

Breast cancer represents the main disease threatens women's health (Sung et al., 2021). Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI) plays a pivotal role in the diagnosis of breast lesions by offering both morphological and hemodynamic information. This imaging technique evaluates the vasculature at multiple time points following the administration of a contrast agent intravenously, enabling the quantitative analysis of signal variations through enhancement kinetic features (Xiao et al., 2021). By examining variations in the contrast agent's absorption, including factors such as the initial peak enhancement and the presence of delayed phase washout, specificity for malignancy identification can be improved. It is established that malignant lesions exhibit an immediate increase in signal intensity, while benign lesions show a slower increase (Padhani, 2002). Hence, the DCE-MRI is a sequence of MRI, whose signal intensity varies due to the contrast agent. The physician aims to evaluate the signal variation captured by the whole sequence to diagnose the disease. Despite the DCE-MRI has been widely used to improve MRI in characterizing breast lesions (Khouli et al., 2009), its specificity is still suboptimal (Kuhl et al., 2005; Orlando, Dimarco, Cannella, & Bartolotta, 2020; Zhang, Ren, et al., 2017).

Artificial Intelligence models have shown impressive results for medical image analysis. Deep Learning architectures were employed for breast cancer classification, segmentation, detection, etc., considering Convolutional Neural Networks, Autoencoder, Generative models, etc (Mridha et al., 2021; Prinzi, Insalaco, Orlando, Gaglio, & Vitabile, 2024; Rautela, Kumar, & Kumar, 2022). These methodologies are focused on the extraction of highly informative features, exploiting the abstraction mechanism of deep neural networks. Many Deep Learning architectures were proposed also for Time Series Analysis (Ismail Fawaz, Forestier, Weber, Idoumghar, & Muller, 2019). For example, the Long Short-term Memory (LSTM) showed outstanding results for learning temporal relationships in time series, as well as many

* Corresponding author.
E-mail addresses: francesco.prinzi@unipa.it (F. Prinzi), orlandoalessiamed@hotmail.it (A. Orlando), salvatore.gaglio@unipa.it (S. Gaglio), salvatore.vitabile@unipa.it (S. Vitabile).

https://doi.org/10.1016/j.eswa.2024.123557
Received 9 October 2023; Received in revised form 10 February 2024; Accepted 20 February 2024
Available online 26 February 2024

convolutional architectures. Nonetheless, these architectures require the training of millions of parameters, imposing a substantial need for a vast quantity of well-annotated data. More importantly, the features extracted via deep architectures are not intelligible, making it difficult to provide a medical interpretation of the trained models. Deep features are commonly interpreted through saliency maps, showing the most important pixels for the predictions. Saliency maps produce only local explanations (i.e., for each instance of the dataset), while they fail to provide a global model explanation. However, to clinically validate the systems and compare them with the medical literature, the global explanation is mandatory.

In recent years, feature extraction from radiological images has been commonly addressed by Radiomics. The radiomic workflow aims to train data-driven models, through several image analysis steps, including data acquisition, identification and segmentation of the Regions of Interest (ROIs), features extraction, and features selection (Gillies, Kinahan, & Hricak, 2016; Lambin et al., 2012). In particular, the ROIs are converted into high-informative quantitative radiomic features, used to correlate a clinical outcome. The radiomic features quantify the ROIs shape, texture, and gray level intensity. Furthermore, more advanced radiomic features can be derived by incorporating wavelet transforms, Laplacian-of-Gaussian filtering, and other preprocessing techniques (Biondi et al., 2023). The radiomic feature extraction process has two main advantages over deep feature extraction. Achieving precise feature extraction does not necessitate a vast dataset as in the case of deep learning, in which extensive data is crucial for model training (Wei, 2021). Deep architectures trained with very small datasets are strongly exposed to the overfitting risk. More importantly, the radiomic features are intelligible, e.g. it is well-known the meaning each radiomic feature expresses, making it possible to explain the models and interpret the most important features. The latter requirement is essential to integrate the models into real clinical practice. In fact, through model interpretation, it is possible to clinically validate the results and compare the model findings with the medical literature. Indeed, several radiomics studies have been introduced for breast cancer classification in DCE-MRI. Four distinct heuristic parameter maps were used by Gibbs et al. (2019) to train a support vector machine. They focused on sub 1-cm Breast Lesion BI-RADS 4 or 5 classification on a dataset composed of 165 lesions. Also in Zhou, Zhang et al. (2020) the parametric maps were exploited, and a 133 court was used to train a logistic regression model. In their study, Parekh and Jacobs (2017) analyzied radiomic features from multiparametric breast MR imaging, which included DCE-MRI. They then aggregated and classified these features using the isoSVM algorithm, with a dataset comprising 124 patients. Zhang et al. (2020) exploited five imaging modalities for feature extraction, including DCE-MRI, and a support vector machine was trained for the classification task. Nagarajan et al. (2013) tackled the extraction of features specifically from the five post-contrast images. Subsequently, they assessed the performance of both a support vector regressor and a fuzzy k-nearest neighbor classifier for the classification of small lesions. Militello et al. (2022) exploited several feature selection algorithms on a court of 111 patients, and a support vector machine was trained for classification. Following this line of research, we also proposed a radiomic model for breast cancer classification (Prinzi, Orlando, Gaglio, Midiri, & Vitabile, 2022). In particular, our evaluation focused on assessing the predictive capabilities of each time instance within the DCE-MRI sequence. Our findings highlighted that the Random Forest classifier, when applied to the third post-contrast time instant, accentuated the distinctions between malignant and benign lesions. Also several deep learning applications were proposed in the literature for DCE-MRI analysis. Zhao et al. (2023) proposed a local–global cross attention fusion network (LG-CAFN) for DCE-MRI breast segmentation and classification. They explored several 2D and 3D deep architectures trained considering three types of sequences: pre-contrast, post-contrast, and subtraction sequences. Ru et al. (2023) proposed an Att-U-Node which uses attention modules to guide a neural ordinary differential equation (ODE) based framework. The proposed model presents also some advantages in terms of interpretability. Park et al. (2023) proposed a 3D model for segmentation. They used 3D U-Net transformer on a proprietary dataset. Chen et al. (2022) propose a Faster-RCNN-based model that first localizes lesions and then a CNN performs classification.

However, the diagnostic efficacy of the DCE-MRI sequence lies in its ability to assess the contrast agent uptake dissimilarity between malignant and benign lesions. From an imaging point of view, this is reflected in the variation in signal intensity between the sequence instants. From a quantitative modeling point of view, the variation uptake can be analyzed through time series classification algorithms.

To the best of our knowledge, no work has been proposed for breast cancer classification in DCE-MRI through time series analysis algorithms. In this work, a time series-based model was proposed for breast cancer classification. More specifically, all seven time instants of the DCE-MRI series were employed simultaneously and a comparative analysis between several time series analysis classifiers was performed. Fig. 1 shows the general workflow. In particular, after feature extraction and harmonization, the most predictive features were selected via a multi-instant feature selection and a univariate time series classification. Then, five predictive features were aggregated via a voting mechanism to implement a multivariate time series classifier. The goal of this modeling is to mimic the physician's diagnostic assessment of the DCE-MRI sequence by evaluating the variation and trend of radiomic features generated by the contrast agent administration. In addition, through the use of intelligible radiomic features, model findings interpretation and clinical validation was performed. This approach addresses a gap in the existing literature, as most previous work tends to either overlook or underutilize time series classifiers in the context of DCE-MRI analysis.

The main contributions of this paper include the following:

- A time series analysis for breast cancer classification by mimicking the DCE-MRI physician's diagnostic process.
- A comparison of several time series classifiers to capture the differences in contrast agent uptake between benign and malignant lesions in DCE-MRI.
- A deep discussion of the model findings obtained from time series analysis as a result of the inherent explainability of radiomic features (Arrieta et al., 2020; Montavon, Samek, & Müller, 2018).

The article is organized as follows: Section 2 Materials and Methods describes the used multi-protocol dataset, its preparation for radiomic features extraction, and the time series classification algorithms. Section 3 Results exposes the obtained results in terms of selected radiomic features, univariate and multivariate time series classification. Section 4 Discussion, discusses the results compared with our previous work (Prinzi et al., 2022), the clinical findings achieved through radiomic features interpretation, and a literature comparison with other works in DCE-MRI. Section 5 remarks on the main conclusions.

2. Materials and methods

2.1. Dataset description

The analyzed dataset included 226 breast DCE-MRI. Forty-eight samples were excluded due to a lack of follow-up or absence of a pathological diagnosis, and 12 due to technical or motion artifacts. Eventually, a cohort of 166 breast mass enhancements was included, with a mean size of 15.3 mm (±10.5, size range: 3–75), detected through DCE-MRI in 104 patients. These patients comprised 103 women and 1 man, with a mean age of 51 years (±11, age range: 31–79), who underwent the examinations at University Hospital "Paolo Giaccone" (Palermo, Italy) between April 2018 and March 2020. A consensus classification was carried out by two experienced breast radiologists for the 166 breast mass enhancements. To give a detailed description of

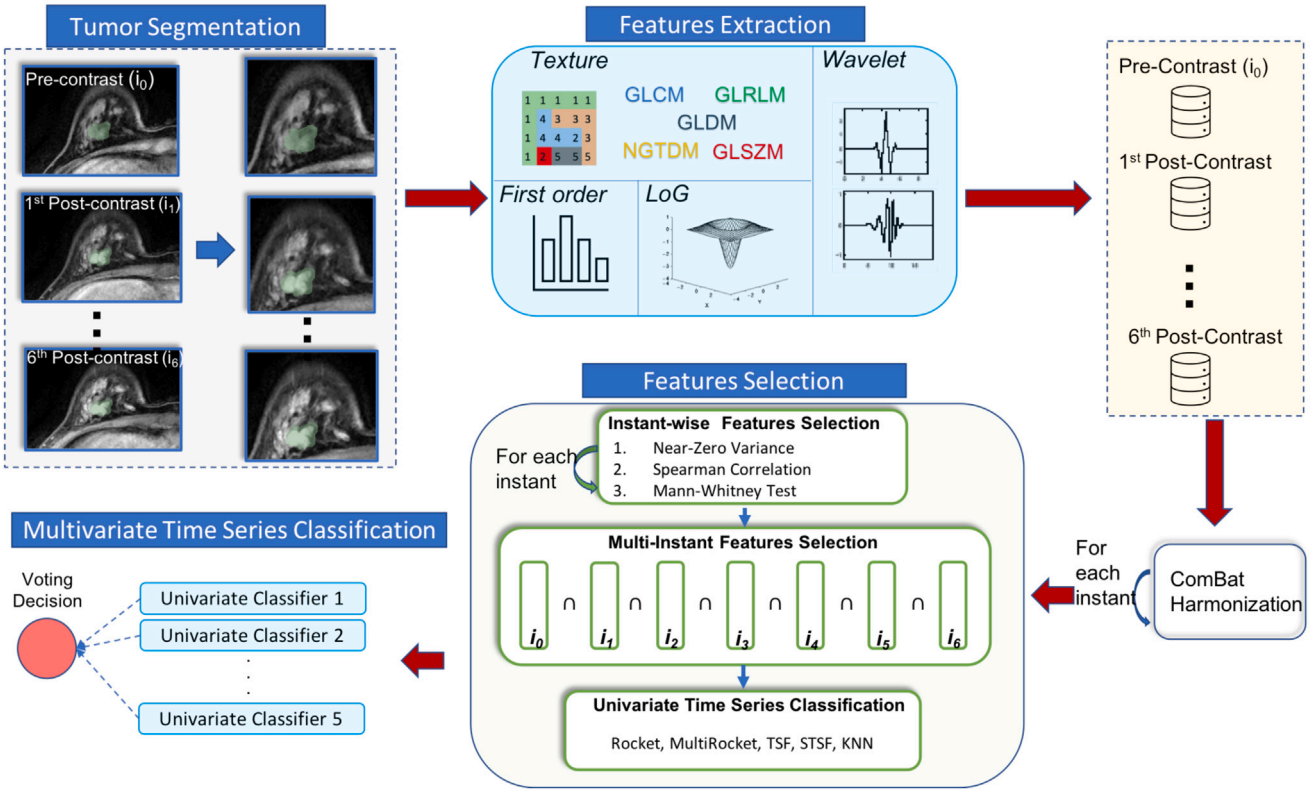


Fig. 1. General workflow. After the manual segmentation step, original, LoG-derived, and Wavelet-derived radiomic features were extracted for each sequence time-instant. The multi-protocol dataset was harmonized using the Combat method. An instant-wise features selection was implemented to discard low variance and correlated features, and to select only statistically significant features. Then, the multi-instant features selection allowed to select the discriminative features for time series classification, selecting only the features statistically significant in each instant of the sequence simultaneously. Finally, an univariate time series classification was performed to select the most predictive features, using several time series classifiers, including Rocket, MultiRocket, Time Series Forest (TSF), Supervised TSF (STSF), and K-nearest Neighbor. The five best features were used to implement a multivariate classifier via a voting mechanism.

the dataset, the BI-RADS lexicon was employed (D'Orsi, Bassett, Feig, et al., 2018), categorizing 73 of these enhancements as BI-RADS 2–3, while 93 were assigned to BI-RADS 4–5. Only benignity/malignancy were used as outcome variables. Breast non-mass enhancements were not considered in our study. This single-center dataset was acquired with a 1.5T MR scanner (GE Signa HDxt, General Electric Healthcare), and a 3D Gradient-echo (GRE) T1-weighted with fat saturation sequence (called Vibrant) was considered. The DCE-MRI sequence allows the evaluation of the tissue/parenchyma enhancement triggered by the introduction of a paramagnetic contrast agent into the vascular system. It consists of the acquisition of a pre-contrast image (without the contrast agent), followed by a series of images (six in our case) acquired after the endovenous introduction of the contrast agent. As a consequence, radiologists study the so-called time-intensity curve for classification. The temporal resolution for these acquisitions was approximately 70 to 90 s. Two different MRI protocols were used (Prinzi et al., 2022), detailed in Table 1. In-plane resolution and slice thickness affect the spatial resolution: the higher the image resolution, the more effective the diagnosis of small pathologies becomes. Increasing the spatial resolution will decrease the pixel size, consequently reducing the signal-to-noise ratio (SNR) of the image. In this perspective, the first protocol shows a higher spatial resolution. It was shown that the effect of the scanner manufacturer is more prevalent in the features compared to slice thickness (Saha, Yu, Sahoo, & Mazurowski, 2017). TR and TE affect the image contrast and the "weighting" (T1 or T2) of the MR image. In the case of T1-weighted (our case) the choice of flip angle is critical for determining signal intensity, image weighting as well as image contrast. It was shown in a 3T case study that increasing TE showed no effect on the T1-weighted images, while increasing TR leads to a slight decrease in blurring (Kim, Lee, Kim, Cho, & Lee, 2013). Protocol 1 exhibited a distribution of 43% benign and 57% malignant lesions, whereas Protocol 2 displayed a distribution of 55% benign and 45% malignant lesions. Some examples of DCE-MRI sequences are shown in Fig. 2 for benign and malignant lesions for both protocols.

Table 1
Employed breast MRI protocols.

	1st	2nd
Number of slices	342	402
Matrix size	452 × 452	352 × 352
In-plane resolution (mm)	0.8 × 0.8	1 × 1
Slice thickness (mm)	0.8	1
Bandwidth	62.5	83.33
Field of view	35 × 35	35 × 35
Time repetition	4.7	3.5
Echo time	2.2	1.6
Flip angle	10	15
Number of lesions	81	85

2.1.1. ROI segmentation

In this study, the 166 breast mass enhancements were manually segmented. The segmentation process was performed using the 3D-Slicer software by three distinct operators along with three radiology residents, each with four years of experience in breast MRI. To ensure precise lesion delineation, each operator selected the post-contrast phase that best highlights the lesion contours. Moreover, to capture the anatomical context more comprehensively, a few millimeters of perilesional fat were included in each segmentation. Subsequently, to

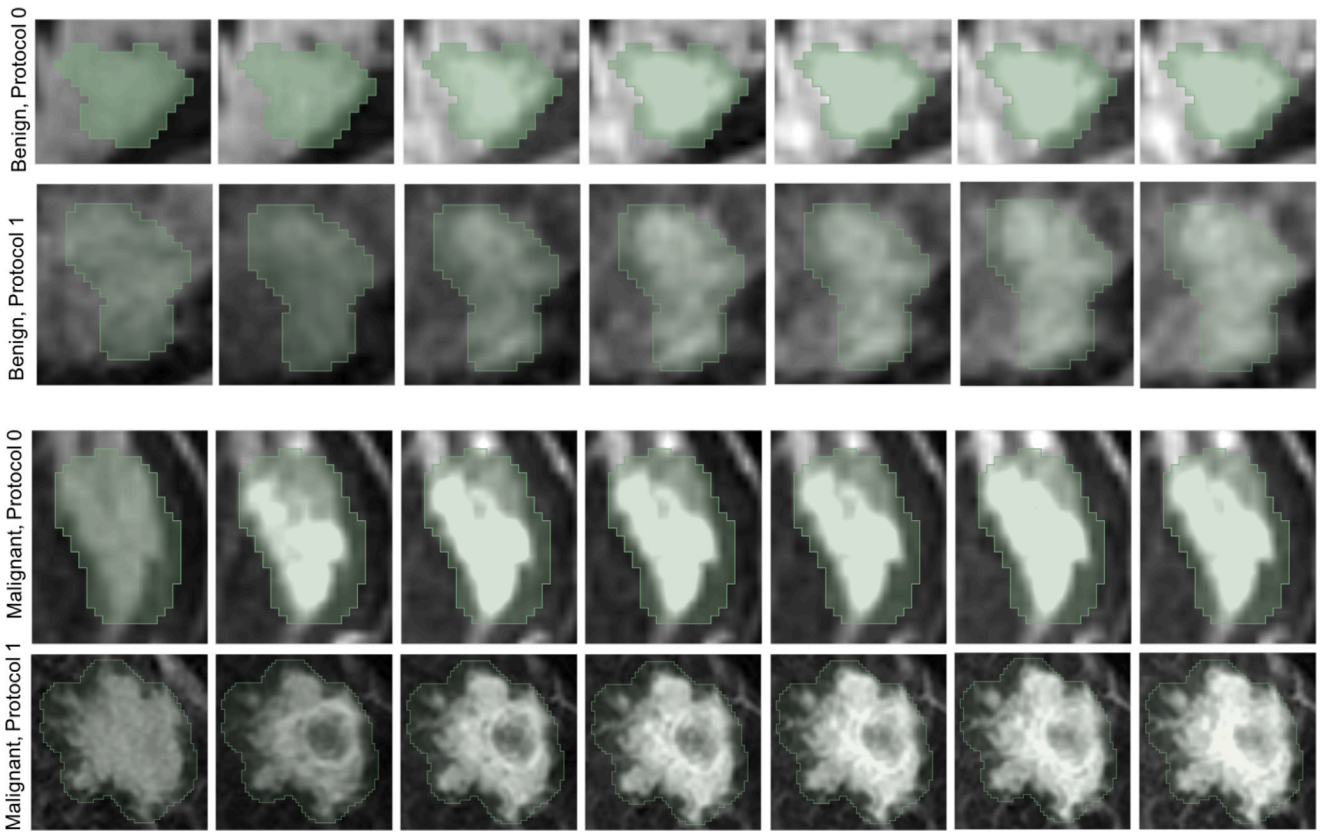


Fig. 2. Four examples of DCE-MRI sequences: two benign cases (above) and two malignant cases (below) for the two protocols. In green are the related manual segmentations.

validate the accuracy and consistency of the segmentations, a third independent breast radiologist with a decade of expertise in breast MR imaging meticulously examined and confirmed the segmentation for all 166 breast masses.

2.2. Radiomic feature extraction

One thousand twenty-three radiomic features were extracted, following the Imaging Biomarker Standardization Initiative (IBSI) (Zwanenburg et al., 2020) and using the pyradiomics toolkit (van Griethuysen et al., 2017). Radiomic feature extraction has shown several advantages over deep feature extraction. It is possible to perform an accurate feature extraction also on moderate sample sizes, while deep feature extraction requires a large database to avoid the overfitting issue (Wei, 2021). Features can also be extracted at the original spatial resolution of the image, avoiding resizing that can cause information loss. However, the main peculiarity lies in the intelligibility of radiomic features, e.g. it is well-known the meaning each feature expresses. This last aspect allows the clinical validation of the model findings. In this perspective, radiomic workflow can satisfy the need for explainability that the medical context requires (Combi et al., 2022) while enabling the training of accurate models. The following feature categories were extracted:

- First Order (FO) intensity histogram statistics;
- Gray Level Co-occurrence Matrix (GLCM) (Haralick, 1979; Haralick, Shanmugam, & Dinstein, 1973);
- Gray Level Run Length Matrix (GLRLM) (Galloway, 1975);
- Gray Level Size Zone Matrix (GLSZM) (Thibault, Angulo, & Meyer, 2014);
- Gray Level Dependence Matrix (GLDM) (Sun & Wee, 1983);
- Neighboring Gray Tone Difference Matrix (NGTDM) (Amadasun & King, 1989).

First-order features are defined as statistical features that provide information about the overall intensity distribution in the ROI. Textural features are typically associated with the discussed matrices GLCM, GLRLM, GLSZM, GLDM, and NGTDM. The above features were extracted from the original images, as well as Laplacian of Gaussian filtered images (LoG-derived features) and Wavelet Transformed images (Wavelet-derived features). The Discrete-Wavelet-Transforms (DWT) has shown promising results in numerous image processing applications (Al Jumah, 2013; Boix & Canto, 2010; Carlini et al., 2023; Dautov & Özerdem, 2018) due to its multi-resolution analysis (Mallat, 1989; Prinzi, Militello, Conti & Vitabile, 2023a; Ravichandran, Nimmatoori, & Gulam Ahamad, 2016). Many works have demonstrated the high predictive power of wavelet-derived features (Chitalia & Kontos, 2019). Zhou, Lu et al. (2020) demonstrated improved performance using wavelet-derived texture features extracted from MRI in predicting breast cancer response to neoadjuvant chemotherapy. Also Peng et al. (2021) showed that wavelet-derived features may have a high correlation with pathologic complete response to neoadjuvant therapy. Exploiting wavelet features, Mahrooghy et al. (2013) have achieved higher performance for breast cancer recurrence risk prediction. In this work, the Haar kernel was used.

The Laplacian of Gaussian filter is commonly used to highlight areas where rapid changes in intensity occur. Despite the LoG-derived features showing benefits in some context (Li et al., 2022; Pereira, Leite Duarte, Ribeiro Damasceno, de Oliveira Moura Santos, & Nogueira-Barbosa, 2021), the predictive role in DCE-MRI is still unclear. In this work, the LoG filter was used considering 2 and 3 as σ values.

Features were extracted independently from each of the seven DCE-MRI time instants. Consequently, our dataset consisted of 166 samples, encompassing 1037 distinct features across the seven time instants (1 pre-contrast and 6 post-contrast). Table 2 summarizes the extracted features.

Table 2
A comprehensive overview of the extracted features for each instant.

Features	N	Description
Original	93	
Wavelet	744	93 × 8, where 93 are the Original features and 8 are the wavelet decompositions.
LoG	186	93 × 2, where 93 are the Original features and 3 are $\sigma = [2, 3]$.
Total	**1023**	This amount was extracted for each of the seven instants

2.3. Radiomic features preprocessing and selection

Considering the variability introduced by the two protocols and the risk of harming the resulting radiomic analysis (Saha et al., 2018), the ComBat algorithm was employed to harmonize the multi-protocol dataset (Fortin, Cullen, Sheline, Taylor, Aselcioglu, Cook, Adams, Cooper, Fava, McGrath, McInnis, Phillips, Trivedi, Weissman, & Shinohara, 2018; Johnson, Li, & Rabinovic, 2006). ComBat is used for the alignment of feature distributions extracted from images acquired at multiple centers/protocols. The method was applied directly to the radiomic features of each instant within the sequence. The authors (Orlhac et al., 2022) state that, in many examples, each of these distributions is itself composed of 2 or more distributions (e.g. patients with different tumor stages). In the last case, the basic formula for ComBat harmonization can be expanded by introducing covariate terms to consider intra-protocol variability. In our work, no additional information regarding tumor advancement or similar variables was included. Therefore, no correction term was used to consider additional intra-protocol variability. For each sequence instant, low variance features (<0.01), and highly correlated features (Spearman's rank correlation $|\rho| > 0.9$) were discarded (Militello, Prinzi, Sollami, Rundo, La Grutta, & Vitabile, 2023). To compare the malignant and benign distributions, the Mann–Whitney U test was employed. A significance level of $p < 0.05$ was considered as the threshold for statistical significance.

Subsequently, the features selected in each time instant were fed into the multi-instant feature selection process. In particular, the features simultaneously statistically significant in all the seven instants were selected. With this procedure, selected features were relevant to discriminate sequences rather than individual instants.

2.4. Time series analysis

In our previous research (Prinzi et al., 2022), classification was conducted to evaluate the predictivity of each instant separately. The dataset was acquired as discussed in Section 2.1. The features were extracted and preprocessed as discussed in Sections 2.2 and 2.3. For each instant, low variance and correlated features were discarded. However, only the statistically significant features in at least 5 instants of 7 were selected according to the Mann–Whitney U test. Support Vector Machine, XGBoost, and Random Forest were trained considering the selected features according to the Sequential Forward Floating Selection (SFFS) (Raschka, 2018). Therefore, each algorithm at each instant was trained with a subset of features that maximized accuracy, computed via SFFS.

In this work, through time series analysis, a completely different methodology for breast cancer classification was implemented. Considering the large number of features selected after the preprocessing and selection stage, an univariate time series analysis was initially performed: *(1)* for feature selection to identify the most discriminative features for time series classification, and *(2)* to evaluate the optimal time series classifier. The univariate step for feature selection represents a wrapper method because is based on a specific machine learning algorithm. Then, a multivariate time series classification was implemented via a voting mechanism, considering the best classifiers and features found in the univariate step. Model and features assessment were performed considering the accuracy computed during a 20-repeated Stratified 10-fold Cross-Validation (CV) procedure in the training set. Then, the multivariate classification model was evaluated on the independent test set.

Rocket, MultiRocket, Time Series Forest, Supervised Time Series Forest, and K-Nearest Neighbors with several distance metrics were compared as time series classifiers.

2.4.1. Rocket and MultiRocket

The RandOm Convolutional KErnel Transform (ROCKET) (Dempster, Petitjean, & Webb, 2020) algorithm is a kernel-based classifier, in which convolutional filters are applied to extract features from time series data. In our case, the convolutions kernel extracts features from the radiomic time series. Then a RidgeClassifierCV is trained using these features.

For each kernel, Rocket generates two distinct features:

- Maximum (Max): The value represents the highest value within a dataset, equivalent to the global max pooling operation.
- Portion of positive values (ppv): $ppv(Z) = \frac{1}{n}\sum_{i=0}^{n}[z_i > 0]$, where Z represents the output of the convolution operation between the series and the kernels.

This means that using 500 kernels, 1000 features are extracted for each time series. Despite the high dimensionality and the small size of the dataset, it has been demonstrated that Rocket yields a remarkably high classification accuracy when used as input for a linear classifier, such as ridge regression (Dempster et al., 2020).

The Rocket classifier has been used as a benchmark algorithm for classification and showed impressive performance in several datasets (Ruiz, Flynn, Large, Middlehurst, & Bagnall, 2021). However, also the MultiRocket algorithm showed promising performance compared with its predecessor Rocket in terms of accuracy (Dempster, Schmidt, & Webb, 2021; Tan, Dempster, Bergmeir, & Webb, 2022). MultiRocket employs a defined set of kernels characterized by a fixed length and weight configuration, as well as the same set of dilatations. In addition, it uses 4 pooling operators on the convolution output:

- The Proportion of Positive Values (ppv), the same as described in Rocket.
- The Mean of Positive Values (mpv) is a statistical measure used to capture the magnitude of positive values within a series. It is defined as $mpv(Z) = \frac{1}{m}\sum_{i=1}^{m} z_i^+$ where z^+ represents a vector of positive value of length m.
- The Mean of Indices of Positive Values (mipv) is a statistical metric used to capture information about the relative position of positive values within the series. It is defined as:

$$mipv = \begin{cases} \frac{1}{m}\sum_{j=1}^{m} i_j^+ & if \quad m > 0 \\ -1 & otherwise \end{cases} \tag{1}$$

- The Longest Stretch of Positive Values (lspv) is a metric used to ascertain the greatest length of any positive value subsequence within a given series, calculated as $LSPV(Z) = max[j - i | \forall_{i \le k \le j} z_k > 0]$

The main novelty of MultiRocket draws inspiration from the Diverse Representation Canonical Interval Forest Classifier (Middlehurst et al., 2021), where the initial time series undergoes a transformation into its first-order difference, which is subsequently used for feature extraction. For this reason, considering the fixed 6216 kernels, 2 series representations, and 4 pooling operators, MultiRocket extracts about 50,000 features (Tan et al., 2022). However, the applicability of these models for very short time series is unclear. In fact, these algorithms are typically used on more than 60-length time series (UCR Dataset Dau et al., 2019).

2.4.2. Time Series Forest and Supervised Time Series Forest

The Time Series Forest (TSF) algorithm (Deng, Runger, Tuv, & Vladimir, 2013) is an interval-based classifier. TSF trains a random forest with features extracted by the series divided into $\sqrt{m}$ intervals, where m represent the series length, in this case 7. The features extracted are mean, standard deviation, and slope. In addition, the Supervised TSF (STSF) (Cabello, Naghizade, Qi, & Kulik, 2020) was implemented. STFS aims to improve efficiency by exploiting a supervised process to select only the discriminatory intervals from the series and introducing median, interquartile range, minimum, and maximum as features. It is proved that for several datasets, STSF obtained comparable accuracy to state-of-the-art time series classification methods while being significantly more efficient.

2.4.3. K-nearest neighbors

The K-Nearest Neighbors classifier for time series is a distance-based algorithm in which specific metrics are used to determine the samples' distances. It represents a benchmark for time series classification because it is simple and does not require tuning of numerous hyperparameters. A comprehensive comparison was made considering several metrics. In particular: Dynamic Time Warping (DTW) (Sakoe & Chiba, 1978), Weighted DTW (WDTW) (Jeong, Jeong, & Omitaomu, 2011) and Derivative DTW (DDTW) (Keogh & Pazzani) were used, as well as edit distances such as Longest Common SubSequence distance (LCSS) (Vlachos, Kollios, & Gunopulos, 2002), Edit distance for Real Penalty (ERP) (Chen & Ng, 2004), Edit Distance for Real sequences (EDR) (Chen, Özsu, & Oria, 2005), Time Warp Edit distance (TWE) (Marteau, 2008). The K value was set to 3 for the experiments.

3. Experimental results

The dataset was divided into two sets: a training/validation set (80%) and a test set (20%). The division of the dataset was performed randomly and in a stratified manner to ensure an equal representation of malignant and benign lesions in both the training/validation and test groups. Ultimately, the training/validation dataset contained a total of 132 lesions, with 67 classified as malignant and 65 as benign. The test dataset consisted of 34 examples, equally distributed between the two classes. To assess the model performance, various metrics were taken into account, including Accuracy, Sensitivity, Specificity, Negative Predictive Value (NPV), Positive Predictive Value (PPV), and Area Under the Receiver Operating Characteristic (AUC-ROC). To ensure a fair and accurate comparison between different algorithms, the same seed was set for all probabilistic terms within the algorithms and for generating the stratified cross-validation folds.

3.1. Workflow reproducibility

Reproducibility is one of the key issues in radiomic works. Many initiatives have been proposed to increase their reproducibility, and this work is particularly consistent with the CLEAR guidelines (Kocak et al., 2023). Following are the main aspects.

(i) The former is related to the image scanner and the acquisition protocol. The variability in image quality depends on hospital resources (and thus the available scanners) and the progress of healthcare within a given geographical area. Several scanners are currently used to acquire MRIs and can result in very different images in terms of resolution and noise. For example, when compared to 1.5T, a 3T scanner results in better contrast resolution of the enhanced lesions (Menezes et al., 2016). This discrepancy significantly influences various aspects of a radiologist's reporting and mainly the radiomic feature extraction, where meticulous relationships between the pixels are considered (Ziayee et al., 2022). In addition, for the same scanner, image acquisition can differ significantly with the use of multiple protocols. In this case, there are several examples in which the ComBat method was effective (Li, Ammari, Balleyguier, Lassau, & Chouzenoux, 2021) and for this reason, we employed the ComBat method as reference harmonization method (Nan et al., 2022).

(ii) Segmentation is one of the key steps to increase system reproducibility. In our case, segmentations were performed by three distinct operators along with three senior radiologist residents and then validated by another expert radiologist. The procedure was described in detail, indicating the software, the segmentation mode, and the protocol for validating segmentations.

(iii) The feature extraction process can be performed through several tools and different settings. In this case, pyradiomics was employed as an IBSI-compliant toolkit (Zwanenburg et al., 2020). All parameters not explicitly stated for extraction should be considered as default settings (Kocak et al., 2023).

(iv) Finally, for model training and test, the randomness of all models was set with the same seed, as well as the split to generate the training and test subsets and to generate the folds of the cross-validation procedures. In addition, a model interpretation was performed comparing the model findings with the clinical literature, showing interesting overlapping with previous radiomic works as well as clinical findings on breast cancer using DCE-MRI.

Regarding the software and libraries used for the entire workflow, the following versions were employed: for Segmentation 3D Slicer Version 4.11.20200930; for feature extraction pyradiomics 3.0.1; for feature harmonization NeuroCombat-sklearn 0.1.3; for model training and evaluation scikit-learn v1.1.2 and sktime v0.13.4. in python 3.7 environment.

Table 3
Feature variability before and after harmonization.

	t_0	t_1	t_2	t_3	t_4	t_5	t_6
Before	878	727	909	913	918	922	934
After	176	180	178	195	189	193	211

3.2. Features harmonization impact

To demonstrate how the variability introduced by the two protocols was mitigated through data harmonization, a statistical analysis was performed to compare the harmonization process effects. Specifically, the Mann–Whitney U test was applied to compare the feature distribution of the two protocols before and after feature harmonization. Table 3 shows the number of features with a statistically significant difference, before and after harmonization, for the 1023 features of the seven instants. Before harmonization, on average, 86% of the features result in statistically different distributions between the two protocols, potentially greatly affecting the radiomic analysis without any harmonization step. However, after the harmonization process, only 18% of features resulted in a different distribution between the two protocols. This significant decrease reduces the variability introduced by the two protocols, making the radiomic analysis performed more robust.

3.3. Features selected

Five original features, 4 LoG-derived, and 15 Wavelet-derived were selected after the preprocessing and multi-instant selection phases. These features resulted uncorrelated and statistically significant in each instant of the sequence simultaneously. The number of wavelet-derived features exceeded both the original and LoG-derived features significantly. This disparity stems from the initial count of Wavelet features, which was 744 per instant, four times greater than the LoG-derived and eight times greater than the original features. Considering this proportion, the original features were the most frequently selected, followed by LoG-derived, and lastly, wavelet-derived features. Then, a total of 24 features were used for the univariate time series analysis.

Table 4
Rocket validation accuracy of the five most accurate features.

Category	Feature	Class	Accuracy
FO	Energy	Original	0.717
FO	TotalEnergy	Original	0.737
NGTDM	Strength	Original	0.696
GLCM	Imc2	Wavelet HLH	0.729
FO	90Percentile	LoG $\sigma = 2$	0.671
			0.710

Table 5
MultiRocket validation accuracy of the five most accurate features.

Category	Feature	Class	Accuracy
FO	Energy	Original	0.694
FO	TotalEnergy	Original	0.687
NGTDM	Busyness	LoG $\sigma = 2$	0.679
GLCM	Imc2	Wavelet HLH	0.676
GLDM	DependenceEntropy	Original	0.673
			0.681

Table 6
TSF validation accuracy of the five most accurate features (LDHGLE is for LargeDependenceHighGrayLevelEmphasis).

Category	Feature	Class	Accuracy
FO	Energy	Original	0.632
FO	TotalEnergy	Original	0.636
GLDM	LDHGLE	Wavelet LLH	0.635
GLCM	Imc2	Wavelet HLH	0.630
GLSZM	SmallAreaEmphasis	Wavelet HHH	0.641
			0.635

Table 7
STSF validation accuracy of the five most accurate features (LDE is for LargeDependenceEmphasis).

Category	Feature	Class	Accuracy
FO	TotalEnergy	Original	0.651
NGTDM	Strength	Original	0.648
GLCM	Imc2	Wavelet HLH	0.677
GLSZM	SmallAreaEmphasis	Wavelet HHH	0.652
GLDM	LDE	Wavelet HHH	0.648
			0.654

3.4. Univariate analysis

For each of the 24 features, Rocket, MultiRocket, TSF, STSF and, KNN with all the discussed distances, were trained. In general, Rocket-based algorithms outperform TSF, STSF, and KNN. In particular, considering the 24 selected radiomic features, MultiRocket was slightly better than Rocket, with an average accuracy of 0.628 vs. 0.617 (see Supplemental Material, Table 1). Tables 4 and 5 show the top-five accurate features for Rocket and MultiRocket models, respectively. Considering the top-five accurate features, Rocket performs much better, achieving an average accuracy of 0.710, compared with 0.681 of MultiRocket. The STSF models outperform the TSF ones, showing an average accuracy of 0.602 vs. 0.594 over the 24 radiomic features (see Supplemental Material, Table 1). The difference between STSF and TSF is higher when the top-five features are considered, as shown in Tables 6 and 7. In fact, 0.654 and 0.635 were computed for STSF and TSF, respectively. The performance computed for KNN, considering each discussed distance were significantly lower than Rocket-based and TSF-based algorithms (see Supplemental Material, Table 2). Considering the best models (e.g., Rocket-based and TSF-based), the Total Energy and Icm2 features resulted in the most predictive for each time series classifier employed. The Energy features in three of the four best models.

3.5. Multivariate analysis

The top-five features discussed for the Rocket models were used for multivariate analysis using a voting mechanism. The meaning of the five features is crucial in drawing important clinical conclusions, as outlined in the discussions. In particular:

- The Original First Order Energy is a metric that quantifies the magnitude of voxel values within an image. A higher value indicates a larger sum of the squares of these voxel values. Visually, the lesion with high Energy should exhibit a very bright appearance characterized by very high intensities.
- Original First Order TotalEnergy: is a metric derived by scaling the Energy feature with respect to the volume of the voxel in cubic millimeters.
- Original ngtdm Strength: exhibits a high value when an image demonstrates a slow transition in intensity, accompanied by more pronounced variations in gray-level intensities.
- Wavelet — HLH glcm Imc2: measures the complexity of the texture.
- LoG — First Order 90Percentile.

Table 8 shows all the validation metrics for the top-five best features resulting from the univariate analysis, reported considering the mean and standard deviation of the 20-repeated 10-fold CV. Table 9 shows the validation performance for the multivariate time series analysis and the instant-wise analysis (Prinzi et al., 2022). Table 10 displays the performance metrics calculated for both the multivariate time series analysis and the instant-wise analysis (Prinzi et al., 2022) on the test set. In the test phase, the Multivariate Rocket model outperforms the instant-wise model in terms of accuracy, sensitivity, and NPV. On the other hand, the instant-wise model (Prinzi et al., 2022) resulted in higher AUC-ROC, specificity, and PPV. However, the difference between specificity and sensitivity is significantly smaller for the Multivariate Rocket, making the model more balanced compared with the instant-wise model. It is possible to observe the substantial improvement of the time series approach over the instant-wise one, in the 20-repeated 10-fold CV performance (Table 9). In fact, considering the small size of the dataset, the CV procedure enables a more precise performance evaluation. Accuracy and AUC-ROC resulted higher for the Multivariate Rocket model. Specificity was slightly lower with a significantly higher sensitivity. PPV value was similar for both models and higher PPV for Multivariate Rocket. In addition, a lower standard deviation was computed for Multivariate Rocket in each metric.

3.6. Leave-one-protocol-out evaluation

To validate the promising model performance, a leave-one-protocol-out evaluation was employed. Leveraging the multi-protocol dataset, this methodology involves training the model using samples from one protocol and subsequently evaluating its performance with samples from the opposing protocol. This procedure simulates a kind of external model validation. This evaluation method approximates, even if only composed of two protocols, the leave-one-center-out cross-validation, as performed in Soda et al. (2021). However, in our case, this approach assumes training a machine learning model on 81 samples for the first protocol and 85 for the second one, making model training complicated. Specifically, we employed the optimal configuration used for the whole analysis, e.g. the Multivariate Rocket models with the features outlined in Table 8. The performance achieved in the two protocols exhibits differences and demonstrates a reasonable generalization capability. In both cases accuracy and AUROC overlap. Despite the very-small dataset size, training on Protocol 1 and testing on Protocol 2 yields an accuracy and AUROC of 0.691, whereas training on Protocol 2 and testing on Protocol 1 an accuracy and AUROC of 0.605. While the selected features and model showcase effectiveness even when the two protocols are analyzed separately, it is crucial to

Table 8
Cross-validation performance of the top five features using the Rocket algorithm.

Model	Accuracy	AUC-ROC	Specificity	Sensitivity	PPV	NPV
Original FO Energy	0.717 ± 0.133	0.718 ± 0.133	0.719 ± 0.184	0.717 ± 0.196	0.733 ± 0.159	0.731 ± 0.156
Original FO TotalEnergy	0.736 ± 0.109	0.736 ± 0.110	0.727 ± 0.160	0.745 ± 0.173	0.747 ± 0.126	0.754 ± 0.140
Original NGTDM Strength	0.696 ± 0.111	0.696 ± 0.110	0.610 ± 0.169	0.782 ± 0.150	0.680 ± 0.118	0.747 ± 0.149
Wavelet HLH GLCM Imc2	0.728 ± 0.119	0.727 ± 0.120	0.666 ± 0.192	0.789 ± 0.160	0.722 ± 0.143	0.773 ± 0.149
LoG $\sigma = 2$ FO 90Percentile	0.671 ± 0.120	0.671 ± 0.121	0.681 ± 0.182	0.661 ± 0.172	0.693 ± 0.157	0.670 ± 0.133

Table 9
Corss-validation performance of the multivariate Rocket classification algorithm using a voting mechanism and its comparison against the previous instant-wise analysis (Prinzi et al., 2022).

Model	Accuracy	AUC-ROC	Specificity	Sensitivity	PPV	NPV
Multivariate Rocket	0.742 ± 0.117	0.743 ± 0.118	0.710 ± 0.171	0.775 ± 0.156	0.742 ± 0.131	0.767 ± 0.138
Instant-wise (Prinzi et al., 2022)	0.710 ± 0.130	0.741 ± 0.135	0.738 ± 0.177	0.683 ± 0.178	0.743 ± 0.153	0.703 ± 0.145

Table 10
The overall model performance on the indipendent test set achieved using the Rocket algorithm and its comparison against the previous instant-wise analysis (Prinzi et al., 2022).

Model	Accuracy	AUC-ROC	Specificity	Sensitivity	PPV	NPV
Multivariate Rocket	0.852	0.852	0.823	0.882	0.833	0.875
Instant-wise (Prinzi et al., 2022)	0.823	0.877	0.882	0.764	0.866	0.789

acknowledge the limitations posed by the small dataset. The current results, while promising, warrant caution, and the generalization of findings should be approached with care. Further investigations with a more comprehensive dataset are recommended to validate the observed performance.

4. Discussion and comparison

In our previous research (Prinzi et al., 2022), each instant within the DCE-MRI sequence was evaluated separately. The Random Forest model exhibited promising performance when trained on features from the third post-contrast instant. This result was elucidated by taking into account that, during the third post-contrast instant, the contrast agent is effectively absorbed in both malignant and benign lesions, thereby emphasizing their distinct characteristics. However, the instant-wise analysis does not fully exploit the potential of the DCE-MRI sequence because classification is performed on each instant separately.

In this view, the proposed method introduces several novelties and advantages. Firstly, the series structure of the DCE-MRI acquisition was analyzed through time series classification algorithms and exploiting all the sequence instants simultaneously. The approach assumes the whole series more informative than individual time instants. In addition, the approach results similar to the radiologist's diagnostic process, which considers the whole sequence to determine the lesion benignity or malignancy.

The MRI acquisition protocols were set by medical professionals, including radiologic technologists and radiologists, based on clinical evidence and individual patient needs. As a consequence, our dataset is composed of two protocols reflecting the real clinical scenario, introducing challenges and enhancing the validity of our work. For multi-protocol dataset management, we performed a data harmonization to align distributions from the two protocols. This step was critical because it was proven a multi-protocol dataset can influence radiomic analysis (Saha et al., 2017). In fact, Table 3 showed a significant reduction in inter-protocol variability due to the use of combat. Specifically, on average 86% of features were statistically different between the two protocols before harmonization, while only 18% after harmonization. In addition, the use of multi-instant feature selection enabled the selection of descriptive features for time series classification, independently of the specific instant within the instant under consideration.

Several time series classifiers were compared, proving Rocket and MultiRocket as the best on small datasets and short time series settings. In particular, the top-five radiomic features for Rocket were aggregated via a voting mechanism. The LoG-derived features were not as predictive as the original and Wavelet-derived features.

Compared with our previous work, the proposed model demonstrates superiority in various aspects. Given the limited dataset size, metrics calculated in cross-validation are considered more reliable. In this context, the gaps observed in AUROC, specificity, and sensitivity are minimal (Table 9). Specifically, the average AUROC is higher (0.743 vs. 0.741) with a lower standard deviation (0.118 vs. 0.135). Although specificity is marginally lower (0.710 vs. 0.738), the standard deviation is also lower (0.171 vs. 0.177). Furthermore, the positive predictive value (PPV) is comparable on average (0.742 vs. 0.743) but exhibits a significantly lower standard deviation (0.131 vs. 0.153). Notably, all other metrics show a marked improvement. In the test dataset (Table 10), it is crucial to highlight that the proposed method demonstrates a more balanced sensitivity and specificity than the instance-wise analysis, as well as PPN and NPV. This balance results in a model that is more adept at predicting both benign and malignant classes with greater equality. In addition, it is important to emphasize the notable improvement of the proposed method in terms of sensitivity, a critical metric for breast cancer classification. Sensitivity is particularly crucial as it describes the probability that a lesion is malignant when it is indeed malignant. Recognizing that misclassifying a malignant lesion as benign is a more serious error than the inverse, emphasizing a higher sensitivity is imperative. In this context, the preference is for a model that excels in sensitivity over specificity, as the former plays a pivotal role in minimizing the risk associated with misclassification of a malignant lesion as benign.

4.1. Literature comparison

Despite the use of a two-protocol dataset, which undoubtedly elevates the complexity of the classification task, the achieved performance are either superior or in alignment with the state-of-the-art. In fact, Gibbs et al. (2019) used radiomic features extracted from three parameter maps, obtaining a high specificity (97%–100%) and a low sensitivity (56%–67%) model using a support vector machine. Zhang et al. (2020) achieved the same trend using a support vector machine, achieving a sensitivity of 0.714 and a specificity of 0.800 considering only the pharmacokinetic parameters maps. Also in Militello et al. (2022) a support vector machine was trained to achieve a higher specificity (0.741 ± 0.114) with respect to sensitivity (0.709 ± 0.176). Focusing on radiomics analysis, an opposite trend was found by Zhou,

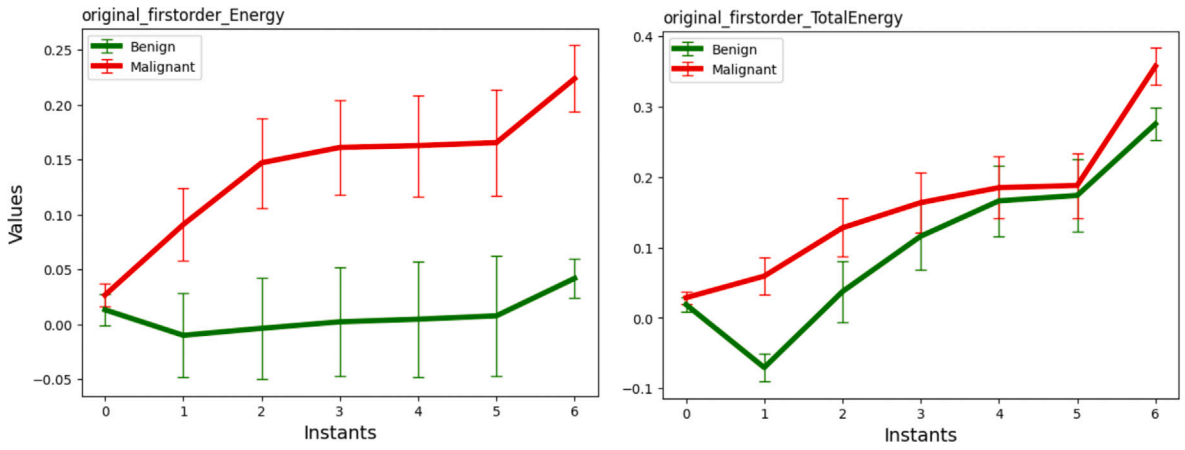


Fig. 3. Average trend of Energy and Total Energy features for the 81 benign lesions (green) and 85 malignant lesions (red).

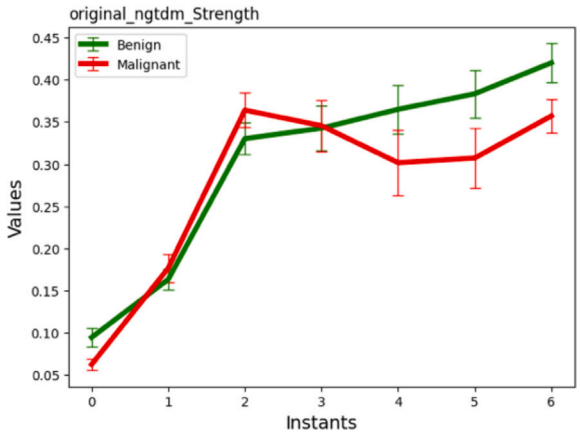


Fig. 4. Average trend of the NGTDM Strength feature for benign lesions (green) and malignant lesions (red).

Zhang et al. (2020), achieving a sensitivity of 85% with respect to a specificity of 65%. Furthermore, in the study conducted by Parekh and Jacobs (2017), which involved the analysis of different image sequences, a higher sensitivity (0.93) was observed in comparison to specificity (0.85). Compared with the validation performance in Prinzi et al. (2022) only specificity resulted slightly lower (0.710 ± 00171 vs. 0.738 ± 0.177). The sensitivity resulted significantly higher (0.775 ± 0.156 vs 0.683 ± 0.178). The same considerations can be extended to the test set. Except compared with the Parekh work (Parekh & Jacobs, 2017), in which features from DCE-MRI, DCE High Spatial Resolution, DWI, ADC map, T1, and T2 were aggregated, our performance results in line or higher. In particular, a more balanced sensitivity and specificity were computed, which means fair benign and malignant classification rates. Also compared with the clinical diagnostic performance, the Multivariate Rocket classifier shows a similar trend in terms of sensitivity and specificity. In particular, was shown that MRI provided an overall sensitivity and specificity of 94.6% and 74.2%, respectively, while for the contrast-enhanced MRI, overall sensitivity and specificity were 91.5% and 64.7% (Aristokli, Polycarpou, Themistocleous, Sophocleous, & Mamais, 2022). Focusing on DCE-MRI, Zhang et al. (2016) computed a sensitivity of 93.2% and a specificity of 71.1%, while 0.87 of sensitivity and 0.74 of specificity were found by Dong, Kang, Cheng, and Zhang (2021). In this view, our model is coherent with the clinical diagnostic performance, showing a slightly lower specificity and higher sensitivity.

4.2. Model interpretation and clinical validation

The main advantage of radiomic features extraction lies in their intelligibility. Radiomic feature extraction allows model introspection while maintaining high classification accuracy. In our case, model introspection and features intelligibility allow comparing the model findings with the real physician diagnosis process and the related clinical literature (Liang et al., 2023; Prinzi, Militello, Scichilone, and Gaglio & Vitabile, 2023). The main results concern features closely related to changes in the intensity of gray levels: Energy and Total Energy. In Fig. 2 is visually evident that the first effect of the contrast agent administration is the increase in gray level (signal) intensities, which is precisely the magnitude that the Energy and Total Energy features describe. Fig. 3 shows the average trend of the two features exhibiting this effect. The malignant lesion shows a more rapid increase in energy, and thus signal intensity, until the last instant of the DCE-MRI sequence. The quantitative explanation of these model findings allows a clinical validation: much faster growth in the first instants (the initial peak enhancement) is evident for malignant rather than benign lesions, indicating that the contrast agent is absorbed faster in malignant lesions (Padhani, 2002). Energy and Total Energy are the statistical features that describe the intensity of the signal and thus explain this clinical aspect. Fig. 3 shows that the greater differences between the two trends lie in the first instants. Peak enhancement typically occurs within the first 2 min after the injection of the contrast agent (Mann, Kuhl, Kinkel, & Boetes, 2008). In our case then occurs in about the first three instants as shown in Fig. 3.

Additionally, elevated NGTDM Strength values indicate that ROIs exhibit a gradual shift in intensity, coupled with a prevalence of substantial and coarse variations in gray-level intensities. In our case, as shown in Fig. 4, the greater difference lies in the latest time instants of the sequence, in which benign lesions show higher Strength values. This indicates that, when both lesions absorb the contrast agent, benign lesions display a more consistent pattern characterized by fewer rapid changes in intensity. High values mean an image with a slow change in intensity but larger coarse differences in gray level intensities. This means that when the contrast agent is absorbed by both lesions, benign lesions show a more regular pattern, with less rapid changes in intensity. This is explained because it was already clinically proved that malignant lesions have typically a heterogeneous internal enhancement in the delayed phase (Agrawal, Su, Nalcioglu, Feig, & Chen, 2009; Tozaki, Igarashi, & Fukuda, 2006).

Features explanation becomes complicated when high-level features such as wavelet-derived are considered. This is because the clinical interpretation of the results is performed by the physician on the original images, which are very different from the transformed images (e.g. Wavelet Transformed and LoG filtered). This makes the association

of radiomic features with clinical findings unfair. However, from a quantitative point of view, the other feature series show a different trend between benign and malignant tumors (See Figure 1 of Supplemental Materials for the average trend of the other selected radiomic sequences).

4.3. Limitations

The use of an external test set acquired at another center/scanner would be the optimal approach to evalue the model performance. We tried to search some DCE-MRI open-source datasets as external test. Unfortunately, datasets in the public domain are frequently scarce, and even when available, they may differ significantly from the specific dataset under analysis. Especially in the case of DCE-MRI, the situation becomes more intricate due to its time series structure. This structure may vary depending on the contrast agent used, resulting in different absorption times, consequently leading to variations in time resolution and number of instants within the sequence. For instance, the Duke-Breast-Cancer-MRI dataset (Saha et al., 2018) is accessible; however, it exclusively comprises invasive breast cancer patients, the related annotations are sequences composed of 4/5 instants and the segmentations are represented in the form of bounding boxes and thus very different from the segmentation prepared for our study. In another available dataset, which is the dataset proposed in Zhao et al. (2023), samples are categorized into malignant and benign, but the entire acquisition process takes less than 3 min consisting of 9 instants (1 pre-contrast and 8 post-contrast) and is acquired using a 3T scanner. In contrast, our dataset involves acquisitions every 70/90 s and 7 instants, rendering the time intervals of this dataset staggered compared to ours. Many other works use similar datasets; unfortunately, they are used for other purposes (Huang et al., 2014) or are not made accessible.

5. Conclusion and future works

The proposed Multivariate Rocket model leverages the entire DCE-MRI sequence to classify breast cancer. More specifically, it employs multi-instant feature selection and univariate time series classification techniques to identify the five most informative features for radiomic time series classification. Then, the top-five features were aggregated via a voting mechanism to implement a Multivariate Rocket model. The use of radiomic features extracted from the whole DCE-MRI sequence mimics the radiologist's diagnostic process, which analyzes the entire DCE-MRI sequence rather than individual instants for the diagnosis. In fact, the implemented Multivariate Rocket model outperformed the instant-wise analysis of our previous work (Prinzi et al., 2022). Furthermore, the use of radiomic features was essential to clinically justify and explain the model findings, showing how the trend of Energy, Total Energy, and NGTDM Strength, can be approximated to some clinical evidence (Agrawal et al., 2009; Mann et al., 2008; Padhani, 2002; Tozaki et al., 2006). The dataset derived from real clinical practice reflects the heterogeneity inherent in actual clinical scenarios. It encompasses images obtained through diverse protocols and a variety of lesion types and sizes, thereby significantly enhancing the validity of the conducted research. To the best of our knowledge, as highlighted in the introduction, this sequence is typically addressed through the analysis of individual time instants or, at most, through subtraction methods. Our work is distinguished by focusing on the analysis of DCE-MRI through time series classifiers, a perspective that, among other advantages, aligns with the current medical diagnostic process for this particular medical imaging modality. This approach addresses a gap in the existing literature, as the majority of previous works tend to overlook or underutilize time-series classifiers in the context of DCE-MRI analysis. One of the main future directions is the information fusion of different temporal instants as well as the extraction of an informative embedding. This salient embedding that encapsulates information from all instants can be the input for several new and recently proposed classifiers (Liu et al., 2022; Xia et al., 2022). Through these approaches, higher accuracy might be achieved even if model explainability can be strongly affected.

CRediT authorship contribution statement

Francesco Prinzi: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Visualization. **Alessia Orlando:** Validation, Investigation, Data curation. **Salvatore Gaglio:** Validation, Investigation, Resources, Writing – review & editing, Supervision. **Salvatore Vitabile:** Validation, Investigation, Resources, Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

Funding

This research was co-funded by the Italian Complementary National Plan PNC-I.1 "Research initiatives for innovative technologies and pathways in the health and welfare sector, D.D. 931 of 06/06/2022, "DARE - DigitAl lifelong pRevEntion" initiative), code PNC0000002, CUP B53C22006460001; and Project INNOVA "Italian network of excellence for advanced diagnosis", CUP B73C22001770006, Italian Complementary National Plan (PNC) to the National Recovery and Resilience Plan (PNRR).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eswa.2024.123557.

References

Agrawal, G., Su, M.-Y., Nalcioglu, O., Feig, S. A., & Chen, J.-H. (2009). Significance of breast lesion descriptors in the ACR BI-RADS mri lexicon. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *115*(7), 1363–1380. http://dx.doi.org/10.1002/cncr.24156.

Al Jumah, A. (2013). Denoising of an image using discrete stationary wavelet transform and various thresholding techniques. *Journal of Signal and Information Processing*, http://dx.doi.org/10.4236/jsip.2013.41004.

Amadasun, M., & King, R. (1989). Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man, and Cybernetics*, *19*(5), 1264–1274. http://dx.doi.org/10.1109/21.44046.

Aristokli, N., Polycarpou, I., Themistocleous, S., Sophocleous, D., & Mamais, I. (2022). Comparison of the diagnostic performance of magnetic resonance imaging (MRI), ultrasound and mammography for detection of breast cancer based on tumor type, breast density and patient's history: A review. *Radiography*, http://dx.doi.org/10.1016/j.radi.2022.01.006.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. http://dx.doi.org/10.1016/j.inffus.2019.12.012.

Biondi, R., Renzulli, M., Golfieri, R., Curti, N., Carlini, G., Sala, C., et al. (2023). Machine learning pipeline for the automated prediction of MicrovascularInvasion in HepatocellularCarcinomas. *Applied Sciences*, *13*(3), http://dx.doi.org/10.3390/app13031371.

Boix, M., & Canto, B. (2010). Wavelet transform application to the compression of images. *Mathematical and Computer Modelling*, *52*(7–8), 1265–1270. http://dx.doi.org/10.1016/j.mcm.2010.02.019.

Cabello, N., Naghizade, E., Qi, J., & Kulik, L. (2020). Fast and accurate time series classification through supervised interval search. In *2020 IEEE international conference on data mining* (pp. 948–953). Los Alamitos, CA, USA: IEEE Computer Society, http://dx.doi.org/10.1109/ICDM50108.2020.00107.

Carlini, G., Gaudiano, C., Golfieri, R., Curti, N., Biondi, R., Bianchi, L., et al. (2023). Effectiveness of radiomic ZOT features in the automated discrimination of oncocytoma from clear cell renal cancer. *Journal of Personalized Medicine*, *13*(3), http://dx.doi.org/10.3390/jpm13030478.

Chen, L., & Ng, R. (2004). On the marriage of lp-norms and edit distance. In *Proceedings of the thirtieth international conference on very large data bases - volume 30* (pp. 792–803). VLDB Endowment, http://dx.doi.org/10.5555/1316689.1316758.

Chen, L., Özsu, M. T., & Oria, V. (2005). *Robust and fast similarity search for moving object trajectories* (pp. 491–502). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/1066157.1066213.

Chen, Y., Wang, L., Luo, R., Wang, S., Wang, H., Gao, F., et al. (2022). A deep learning model based on dynamic contrast-enhanced magnetic resonance imaging enables accurate prediction of benign and malignant breast lessons. *Frontiers in Oncology*, *12*, Article 943415. http://dx.doi.org/10.1016/j.bspc.2021.103319.

Chitalia, R. D., & Kontos, D. (2019). Role of texture analysis in breast MRI as a cancer biomarker: a review. *Journal of Magnetic Resonance Imaging*, *49*(4), 927–938. http://dx.doi.org/10.1002/jmri.26556.

Combi, C., Amico, B., Bellazzi, R., Holzinger, A., Moore, J. H., Zitnik, M., et al. (2022). A manifesto on explainability for artificial intelligence in medicine. *Artificial Intelligence in Medicine*, *133*, Article 102423. http://dx.doi.org/10.1016/j.artmed.2022.102423.

Dau, H. A., Bagnall, A., Kamgar, K., Yeh, C.-C. M., Zhu, Y., Gharghabi, S., et al. (2019). The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica*, *6*(6), 1293–1305.

Dautov, Ç. P., & Özerdem, M. S. (2018). Wavelet transform and signal denoising using wavelet method. In *2018 26th signal processing and communications applications conference* (pp. 1–4). IEEE, http://dx.doi.org/10.1109/SIU.2018.8404418.

Dempster, A., Petitjean, F., & Webb, G. I. (2020). ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, *34*(5), 1454–1495. http://dx.doi.org/10.1007/s10618-020-00701-z.

Dempster, A., Schmidt, D. F., & Webb, G. I. (2021). MiniRocket: A very fast (almost) deterministic transform for time series classification. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining* (pp. 248–257). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3447548.3467231.

Deng, H., Runger, G., Tuv, E., & Vladimir, M. (2013). A time series forest for classification and feature extraction. *Information Sciences*, *239*, 142–153. http://dx.doi.org/10.1016/j.ins.2013.02.030.

Dong, H., Kang, L., Cheng, S., & Zhang, R. (2021). Diagnostic performance of dynamic contrast-enhanced magnetic resonance imaging for breast cancer detection: an update meta-analysis. *Thoracic Cancer*, *12*(23), 3201–3207. http://dx.doi.org/10.1111/1759-7714.14187.

D'Orsi, C., Bassett, L., Feig, S., et al. (2018). Breast imaging reporting and data system (BI-RADS). In *Breast imaging atlas* (4th ed). Reston: American College of Radiology.

El Khouli, R. H., Macura, K. J., Jacobs, M. A., Khalil, T. H., Kamel, I. R., Dwyer, A., et al. (2009). Dynamic contrast-enhanced MRI of the breast: quantitative method for kinetic curve type assessment. *AJR. American Journal of Roentgenology*, *193*(4), W295. http://dx.doi.org/10.2214/AJR.09.2483.

Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., et al. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, *167*, 104–120. http://dx.doi.org/10.1016/j.neuroimage.2017.11.024.

Galloway, M. M. (1975). Texture analysis using gray level run lengths. *Computer Graphics and Image Processing*, *4*(2), 172–179. http://dx.doi.org/10.1016/S0146-664X(75)80008-6.

Gibbs, P., Onishi, N., Sadinski, M., Gallagher, K. M., Hughes, M., Martinez, D. F., et al. (2019). Characterization of sub-1 cm breast lesions using radiomics analysis. *Journal of Magnetic Resonance Imaging*, *50*(5), 1468–1477. http://dx.doi.org/10.1002/jmri.26732.

Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: Images are more than pictures, they are data. *Radiology*, *278*(2), 563–577. http://dx.doi.org/10.1148/radiol.2015151169.

Haralick, R. M. (1979). Statistical and structural approaches to texture. *Proceedings of the IEEE*, *67*(5), 786–804. http://dx.doi.org/10.1109/PROC.1979.11328.

Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-3*(6), 610–621. http://dx.doi.org/10.1109/TSMC.1973.4309314.

Huang, W., Li, X., Chen, Y., Li, X., Chang, M.-C., Oborski, M. J., et al. (2014). Variations of dynamic contrast-enhanced magnetic resonance imaging in evaluation of breast cancer therapy response: a multicenter data analysis challenge. *Translational Oncology*, *7*(1), 153–166. http://dx.doi.org/10.1593/tlo.13838.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, *33*(4), 917–963. http://dx.doi.org/10.1007/s10618-019-00619-1.

Jeong, Y.-S., Jeong, M. K., & Omitaomu, O. A. (2011). Weighted dynamic time warping for time series classification. *Pattern Recognition*, *44*(9), 2231–2240. http://dx.doi.org/10.1016/j.patcog.2010.09.022.

Johnson, W. E., Li, C., & Rabinovic, A. (2006). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127. http://dx.doi.org/10.1093/biostatistics/kxj037.

Keogh, E. J., & Pazzani, M. J. Derivative dynamic time warping. In *Proceedings of the 2001 SIAM international conference on data mining* (pp. 1–11). http://dx.doi.org/10.1137/1.9781611972719.1.

Kim, J. W., Lee, S. H., Kim, N., Cho, K.-S., & Lee, J. S. (2013). Impact of the parameter variation on the image blurring in 3 T magnetic resonance imaging: A phantom study. *Journal of the Korean Society of Radiology*, *68*(4), 355–362. http://dx.doi.org/10.3348/jksr.2013.68.4.355.

Kocak, B., Baessler, B., Bakas, S., Cuocolo, R., Fedorov, A., Maier-Hein, L., et al. (2023). CheckList for EvaluAtion of radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and eusomii. *Insights into Imaging*, *14*(1), 1–13. http://dx.doi.org/10.1186/s13244-023-01415-8.

Kuhl, C. K., Schrading, S., Leutner, C. C., Morakkabati-Spitz, N., Wardelmann, E., Fimmers, R., et al. (2005). Mammography, breast ultrasound, and magnetic resonance imaging for surveillance of women at high familial risk for breast cancer. *Journal of Clinical Oncology*, *23*(33), 8469–8476. http://dx.doi.org/10.1200/JCO.2004.00.4960.

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Van Stiphout, R. G., Granton, P., et al. (2012). Radiomics: extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*, *48*(4), 441–446. http://dx.doi.org/10.1016/j.ejca.2011.11.036.

Li, Y., Ammari, S., Balleyguier, C., Lassau, N., & Chouzenoux, E. (2021). Impact of preprocessing and harmonization methods on the removal of scanner effects in brain MRI radiomic features. *Cancers*, *13*(12), 3000. http://dx.doi.org/10.3390/cancers13123000.

Li, Y., Ammari, S., Lawrance, L., Quillent, A., Assi, T., Lassau, N., et al. (2022). Radiomics-based method for predicting the glioma subtype as defined by tumor grade, IDH mutation, and 1p/19q Codeletion. *Cancers*, *14*(7), 1778. http://dx.doi.org/10.3390/cancers14071778.

Liang, P., Yang, J., Wang, W., Yuan, G., Han, M., Zhang, Q., et al. (2023). Deep learning identifies intelligible predictors of poor prognosis in chronic kidney disease. *IEEE Journal of Biomedical and Health Informatics*, http://dx.doi.org/10.1109/JBHI.2023.3266587.

Liu, J., Wei, J., Heidari, A. A., Kuang, F., Zhang, S., Gui, W., et al. (2022). Chaotic simulated annealing multi-verse optimization enhanced kernel extreme learning machine for medical diagnosis. *Computers in Biology and Medicine*, *144*, Article 105356. http://dx.doi.org/10.1016/j.compbiomed.2022.105356.

Mahrooghy, M., Ashraf, A. B., Daye, D., Mies, C., Feldman, M., Rosen, M., et al. (2013). Heterogeneity wavelet kinetics from DCE-MRI for classifying gene expression based breast cancer recurrence risk. In *Medical image computing and computer-assisted intervention–mICCAI 2013: 16th international conference, nagoya, Japan, September 22-26, 2013, proceedings, part II 16* (pp. 295–302). Springer, http://dx.doi.org/10.1007/978-3-642-40763-5_37.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(7), 674–693. http://dx.doi.org/10.1109/34.192463.

Mann, R. M., Kuhl, C. K., Kinkel, K., & Boetes, C. (2008). Breast MRI: guidelines from the European society of breast imaging. *European Radiology*, *18*, 1307–1318. http://dx.doi.org/10.1148/radiol.2018181352.

Marteau, P.-F. (2008). Time warp edit distance with stiffness adjustment for time series matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(2), 306–318. http://dx.doi.org/10.1109/TPAMI.2008.76.

Menezes, G. L., Stehouwer, B. L., Klomp, D. W., van der Velden, T. A., van den Bosch, M. A., Knuttel, F. M., et al. (2016). Dynamic contrast-enhanced breast MRI at 7T and 3T: an intra-individual comparison study. *SpringerPlus*, *5*(1), 1–12. http://dx.doi.org/10.1186/s40064-015-1654-7.

Middlehurst, M., Large, J., Flynn, M., Lines, J., Bostrom, A., & Bagnall, A. (2021). HIVE-COTE 2.0: a new meta ensemble for time series classification. *Machine Learning*, *110*(11), 3211–3243. http://dx.doi.org/10.1007/s10994-021-06057-9.

Militello, C., Prinzi, F., Sollami, G., Rundo, L., La Grutta, L., & Vitabile, S. (2023). CT radiomic features and clinical biomarkers for predicting coronary artery disease. *Cognitive Computation*, *15*(1), 238–253. http://dx.doi.org/10.1007/s12559-023-10118-7.

Militello, C., Rundo, L., Dimarco, M., Orlando, A., Woitek, R., D'Angelo, I., et al. (2022). 3D DCE-MRI radiomic analysis for malignant lesion prediction in breast cancer patients. *Academic Radiology*, *29*(6), 830–840. http://dx.doi.org/10.1016/j.acra.2021.08.024.

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15. http://dx.doi.org/10.1016/j.dsp.2017.10.011.

Mridha, M. F., Hamid, M., Monowar, M. M., Keya, A. J., Ohi, A. Q., Islam, M., et al. (2021). A comprehensive survey on deep-learning-based breast cancer diagnosis. *Cancers*, *13*(23), 6116. http://dx.doi.org/10.3390/cancers13236116.

Nagarajan, M. B., Huber, M. B., Schlossbauer, T., Leinsinger, G., Krol, A., & Wismüller, A. (2013). Classification of small lesions in breast MRI: evaluating the role of dynamically extracted texture features through feature selection. *Journal of Medical and Biological Engineering*, *33*(1), http://dx.doi.org/10.5405/jmbe.1183.

Nan, Y., Del Ser, J., Walsh, S., Schönlieb, C., Roberts, M., Selby, I., et al. (2022). Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions. *Information Fusion*, *82*, 99–122. http://dx.doi.org/10.1016/j.inffus.2022.01.001.

Orlando, A., Dimarco, M., Cannella, R., & Bartolotta, T. V. (2020). Breast dynamic contrast-enhanced-magnetic resonance imaging and radiomics: state of art. *Artificial Intelligence in Medical Imaging*, http://dx.doi.org/10.35711/aimi.v1.i1.6.

Orlhac, F., Eertink, J. J., Cottereau, A.-S., Zijlstra, J. M., Thieblemont, C., Meignan, M., et al. (2022). A guide to ComBat harmonization of imaging biomarkers in multicenter studies. *Journal of Nuclear Medicine*, *63*(2), 172–179. http://dx.doi.org/10.2967/jnumed.121.262464.

Padhani, A. R. (2002). Dynamic contrast-enhanced MRI in clinical oncology: Current status and future directions. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *16*(4), 407–422. http://dx.doi.org/10.1002/jmri.10176.

Parekh, V. S., & Jacobs, M. A. (2017). Integrated radiomic framework for breast cancer and tumor biology using advanced machine learning and multiparametric MRI. *NPJ Breast Cancer*, *3*(1), 1–9. http://dx.doi.org/10.1038/s41523-017-0045-3.

Park, G. E., Kim, S. H., Nam, Y., Kang, J., Park, M., & Kang, B. J. (2023). 3D breast cancer segmentation in DCE-mri using deep learning with weak annotation. *Journal of Magnetic Resonance Imaging*, http://dx.doi.org/10.1002/jmri.28960.

Peng, S., Chen, L., Tao, J., Liu, J., Zhu, W., Liu, H., et al. (2021). Radiomics analysis of multi-phase DCE-MRI in predicting tumor response to neoadjuvant therapy in breast cancer. *Diagnostics*, *11*(11), 2086. http://dx.doi.org/10.3390/diagnostics11112086.

Pereira, H. M., Leite Duarte, M. E., Ribeiro Damasceno, I., de Oliveira Moura Santos, L. A., & Nogueira-Barbosa, M. H. (2021). Machine learning-based CT radiomics features for the prediction of pulmonary metastasis in osteosarcoma. *The British Journal of Radiology*, *94*(1124), Article 20201391. http://dx.doi.org/10.1259/bjr.20201391.

Prinzi, F., Insalaco, M., Orlando, A., Gaglio, S., & Vitabile, S. (2024). A YOLO-based model for breast cancer detection in mammograms. *Cognitive Computation*, *16*(1), 107–120. http://dx.doi.org/10.1007/s12559-023-10189-6.

Prinzi, F., Militello, C., Conti, V., & Vitabile, S. (2023a). Impact of wavelet kernels on predictive capability of radiomic features: A case study on COVID-19 chest X-ray images. *Journal of Imaging*, *9*(2), 32. http://dx.doi.org/10.3390/jimaging9020032.

Prinzi, F., Militello, C., Scichilone, N., Gaglio, S., & Vitabile, S. (2023). Explainable machine-learning models for COVID-19 prognosis prediction using clinical, laboratory and radiomic features. *IEEE Access*, *11*, 121492–121510. http://dx.doi.org/10.1109/ACCESS.2023.3327808.

Prinzi, F., Orlando, A., Gaglio, S., Midiri, M., & Vitabile, S. (2022). ML-based radiomics analysis for breast cancer classification in DCE-MRI. In M. Mahmud, C. Ieracitano, M. S. Kaiser, N. Mammone, & F. C. Morabito (Eds.), *Applied intelligence and informatics* (pp. 144–158). Cham: Springer Nature Switzerland, http://dx.doi.org/10.1007/978-3-031-24801-6_11.

Raschka, S. (2018). Mlxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *Journal of Open Source Software*, *3*(24), 638. http://dx.doi.org/10.21105/joss.00638.

Rautela, K., Kumar, D., & Kumar, V. (2022). A systematic review on breast cancer detection using deep learning techniques. *Archives of Computational Methods in Engineering*, *29*(7), 4599–4629. http://dx.doi.org/10.1007/s11831-022-09744-5.

Ravichandran, D., Nimmatoori, R., & Gulam Ahamad, M. (2016). Mathematical representations of 1D, 2D and 3D wavelet transform for image coding. *International Journal on Advanced Computer Theory and Engineering*, *5*, 1–8.

Ru, J., Lu, B., Chen, B., Shi, J., Chen, G., Wang, M., et al. (2023). Attention guided neural ODE network for breast tumor segmentation in medical images. *Computers in Biology and Medicine*, *159*, Article 106884. http://dx.doi.org/10.1016/j.compbiomed.2023.106884.

Ruiz, A. P., Flynn, M., Large, J., Middlehurst, M., & Bagnall, A. (2021). The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, *35*(2), 401–449. http://dx.doi.org/10.1007/s10618-020-00727-3.

Saha, A., Harowicz, M. R., Grimm, L. J., Kim, C. E., Ghate, S. V., Walsh, R., et al. (2018). A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. *British Journal of Cancer*, *119*(4), 508–516. http://dx.doi.org/10.1038/s41416-018-0185-8.

Saha, A., Yu, X., Sahoo, D., & Mazurowski, M. A. (2017). Effects of MRI scanner parameters on breast cancer radiomics. *Expert Systems with Applications*, *87*, 384–391. http://dx.doi.org/10.1016/j.eswa.2017.06.029.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*(1), 43–49.

Soda, P., D'Amico, N. C., Tessadori, J., Valbusa, G., Guarrasi, V., Bortolotto, C., et al. (2021). AIforCOVID: Predicting the clinical outcomes in patients with COVID-19 applying AI to chest-X-rays. An Italian multicentre study. *Medical Image Analysis*, *74*, Article 102216. http://dx.doi.org/10.1016/j.media.2021.102216.

Sun, C., & Wee, W. G. (1983). Neighboring gray level dependence matrix for texture classification. *Computer Vision, Graphics, and Image Processing*, *23*(3), 341–352. http://dx.doi.org/10.1016/0734-189X(83)90032-4.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249. http://dx.doi.org/10.3322/caac.21660.

Tan, C. W., Dempster, A., Bergmeir, C., & Webb, G. I. (2022). MultiRocket: multiple pooling operators and transformations for fast and effective time series classification. *Data Mining and Knowledge Discovery*, 1–24. http://dx.doi.org/10.1007/s10618-022-00844-1.

Thibault, G., Angulo, J., & Meyer, F. (2014). Advanced statistical matrices for texture characterization: Application to cell classification. *IEEE Transactions on Biomedical Engineering*, *61*(3), 630–637. http://dx.doi.org/10.1109/TBME.2013.2284600.

Tozaki, M., Igarashi, T., & Fukuda, K. (2006). Positive and negative predictive values of BI-RADS®-MRI descriptors for focal breast masses. *Magnetic Resonance in Medical Sciences*, *5*(1), 7–15. http://dx.doi.org/10.2463/mrms.5.7.

van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., et al. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, *77*(21), e104–e107. http://dx.doi.org/10.1158/0008-5472.CAN-17-0339.

Vlachos, M., Kollios, G., & Gunopulos, D. (2002). Discovering similar multidimensional trajectories. In *Proceedings 18th international conference on data engineering* (pp. 673–684). IEEE, http://dx.doi.org/10.1109/ICDE.2002.994784.

Wei, P. (2021). Radiomics, deep learning and early diagnosis in oncology. *Emerging Topics in Life Sciences*, *5*(6), 829–835. http://dx.doi.org/10.1042/ETLS20210218.

Xia, J., Yang, D., Zhou, H., Chen, Y., Zhang, H., Liu, T., et al. (2022). Evolving kernel extreme learning machine for medical diagnosis via a disperse foraging sine cosine algorithm. *Computers in Biology and Medicine*, *141*, Article 105137. http://dx.doi.org/10.1016/j.compbiomed.2021.105137.

Xiao, J., Rahbar, H., Hippe, D. S., Rendi, M. H., Parker, E. U., Shekar, N., et al. (2021). Dynamic contrast-enhanced breast MRI features correlate with invasive breast cancer angiogenesis. *NPJ Breast Cancer*, *7*(1), 42. http://dx.doi.org/10.1038/s41523-021-00247-3.

Zhang, Q., Peng, Y., Liu, W., Bai, J., Zheng, J., Yang, X., et al. (2020). Radiomics based on multimodal MRI for the differential diagnosis of benign and malignant breast lesions. *Journal of Magnetic Resonance Imaging*, *52*(2), 596–607. http://dx.doi.org/10.1002/jmri.27098.

Zhang, Y., Ren, H., et al. (2017). Meta-analysis of diagnostic accuracy of magnetic resonance imaging and mammography for breast cancer. *Journal of Cancer Research and Therapeutics*, *13*(5), 862. http://dx.doi.org/10.4103/jcrt.JCRT_678_17.

Zhang, L., Tang, M., Min, Z., Lu, J., Lei, X., & Zhang, X. (2016). Accuracy of combined dynamic contrast-enhanced magnetic resonance imaging and diffusion-weighted imaging for breast cancer detection: a meta-analysis. *Acta Radiologica*, *57*(6), 651–660. http://dx.doi.org/10.1177/0284185115597265.

Zhao, X., Liao, Y., Xie, J., He, X., Zhang, S., Wang, G., et al. (2023). BreastDM: A DCE-MRI dataset for breast tumor image segmentation and classification. *Computers in Biology and Medicine*, *164*, Article 107255. http://dx.doi.org/10.1016/j.compbiomed.2023.107255.

Zhou, J., Lu, J., Gao, C., Zeng, J., Zhou, C., Lai, X., et al. (2020). Predicting the response to neoadjuvant chemotherapy for breast cancer: wavelet transforming radiomics in MRI. *BMC Cancer*, *20*, 1–10. http://dx.doi.org/10.1186/s12885-020-6523-2.

Zhou, J., Zhang, Y., Chang, K.-T., Lee, K. E., Wang, O., Li, J., et al. (2020). Diagnosis of benign and malignant breast lesions on DCE-MRI by using radiomics and deep learning with consideration of peritumor tissue. *Journal of Magnetic Resonance Imaging*, *51*(3), 798–809. http://dx.doi.org/10.1002/jmri.26981.

Ziayee, F., Schimmöller, L., Blondin, D., Boschheidgen, M., Wilms, L., Vach, M., et al. (2022). Impact of dynamic contrast-enhanced MRI in 1.5 T versus 3 T MRI for clinically significant prostate cancer detection. *European Journal of Radiology*, *156*, Article 110520. http://dx.doi.org/10.1016/j.ejrad.2022.110520.

Zwanenburg, A., Vallières, M., Abdalah, M. A., Aerts, H. J., Andrearczyk, V., Apte, A., et al. (2020). The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, *295*(2), 328. http://dx.doi.org/10.1148/radiol.2020191145.