Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multimodal Fusion of Radiomics, Two-Dimensional Deep Learning, and Clinical Risk Factors for Preoperative Prediction of Pathological Complete Response to Neoadjuvant Chemotherapy in Breast Cancer: A Multicenter External Validation Study

Hongen Li
Guangdong Women and Children Hospital
Yihui Zeng
Guangdong Women and Children Hospital
Wen Wang
Shandong Medical and Pharmaceutical University Hospital
Jianbin Tan
Guangdong Women and Children Hospital
Yan Zhang
doctorzhangyan@vip.163.com

Guangdong Women and Children Hospital

Research Article

Keywords: breast neoplasms, neoadjuvant chemotherapy, pathological complete response, radiomics, deep learning, multimodal fusion, magnetic resonance imaging, nomogram

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11382617/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multimodal Fusion of Radiomics, Two-Dimensional Deep Learning, and Clinical Risk Factors for Preoperative Prediction of Pathological Complete Response to Neoadjuvant Chemotherapy in Breast Cancer: A Multicenter External Validation Study

Hongen Li[*1] • Yihui Zeng[*1] • Wen Wang[2] • Jianbin Tan[1] • Yan Zhang[1+]

*H.L. and Y.Z.(Yihui Zeng) contributed equally to this work

Author affiliations:

[1]Department of Radiology, Guangdong Women and Children Hospital, No. 521 Xingnan Avenue, Panyu District, Guangzhou 511442, China

[2]Shandong Medical and Pharmaceutical University Hospital, Shandong, China

[+]Address correspondence to: Y.Z. (Yan Zhang) (email: doctorzhangyan@vip.163.com)

Abstract

Purpose: Accurate preoperative prediction of pathological complete response (pCR) following neoadjuvant chemotherapy (NAC) is essential for individualized treatment decision-making in patients with breast cancer. This study aimed to develop a multimodal fusion model integrating conventional radiomics, two-dimensional (2D) deep learning features, and clinical risk factors based on the pre-treatment third-phase images of dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI), and to validate its efficacy for predicting pCR in an independent institutional cohort.

Methods: This retrospective multicenter study enrolled 372 women with breast cancer who received NAC across two centers. The 245 patients from Center 1 constituted the model development cohort and were stratified by pCR status and randomly split at a 7:3 ratio into a training set (n = 171) and an internal validation set (n = 74), whereas the 127 patients from Center 2 served as an independent external test set. A radiologist manually delineated the tumor region of interest on the third contrast-enhanced phase of DCE-MRI, and all delineations were reviewed by a senior radiologist. A total of 1,197 radiomic features were extracted using PyRadiomics; after Spearman correlation-based redundancy removal and dimensionality reduction with least absolute shrinkage and selection operator (LASSO) regression, 10 features were retained to construct the radiomics model. Six ImageNet-

pretrained convolutional neural networks were compared, and MobileNet V2 was selected to build the 2D deep learning model. After dimensionality reduction by principal component analysis, the deep features were early-fused with the radiomic features to develop a deep learning radiomics (DLR) model. Independent clinical predictors of pCR were identified using univariate and multivariate logistic regression and used to construct a clinical model. The predicted probability of the DLR model was further fused with the independent clinical factors using logistic regression to build the combined model, which was presented as a nomogram. Model performance was evaluated using the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, calibration curves, decision curve analysis, the DeLong test, the integrated discrimination improvement (IDI), and the net reclassification improvement (NRI). Interpretability analyses were conducted using SHAP and Grad-CAM.

Results: The pCR rate of the entire cohort was 14.2% (53/372), with rates of 17.0%, 21.6%, and 6.3% in the training, internal validation, and external test sets, respectively. Multivariate logistic regression identified ER status (adjusted OR = 0.119, P < 0.001), HER-2 status (adjusted OR = 4.870, P = 0.003), and maximum tumor diameter (adjusted OR = 0.742, P = 0.041) as independent predictors of pCR. In the independent external test set, the combined model achieved the best overall discriminative performance, with an AUC of 0.933 (95% CI: 0.879–0.987), accuracy of 0.843, sensitivity of 1.000, specificity of 0.832, and negative predictive value of 1.000, outperforming the clinical model (AUC = 0.925), the deep learning model (0.825), the DLR model (0.745), and the radiomics model (0.559); its AUC was significantly higher than those of the radiomics model (P = 0.0007) and the DLR model (P = 0.014). The deep learning model showed the smallest cross-center performance degradation (ΔAUC = 0.033), markedly smaller than that of the conventional radiomics model (ΔAUC = 0.278). Relative to the four single-modality models, the combined model yielded positive IDI (0.208–0.482, all P < 0.05) and NRI (0.216–0.683) values, good calibration (Hosmer–Lemeshow test, P = 0.069), and a higher net clinical benefit on decision curve analysis. SHAP and Grad-CAM analyses consistently demonstrated that model decisions focused on intratumoral enhancement heterogeneity and peritumoral regions, with the most contributory features derived from both radiomics and deep learning, indicating complementarity between the two sources of information.

Conclusions: The multimodal fusion nomogram integrating independent clinical predictors, conventional radiomic features, and 2D deep learning features can accurately and robustly predict pCR after NAC in breast cancer before surgery, with favorable cross-center

generalizability and interpretability. It may serve as a noninvasive quantitative tool for evaluating NAC response and supporting individualized treatment decision-making.

Keywords: breast neoplasms; neoadjuvant chemotherapy; pathological complete response; radiomics; deep learning; multimodal fusion; magnetic resonance imaging; nomogram

Introduction

Breast cancer is currently the most frequently diagnosed malignancy among women worldwide and poses a serious threat to women's health. According to the GLOBOCAN 2022 estimates of the International Agency for Research on Cancer (IARC), approximately 2.3 million new cases of female breast cancer occurred worldwide in 2022, ranking first among female malignancies [1]. The disease burden of breast cancer in China is likewise increasing continuously, with approximately 357,000 new cases in 2022, and its incidence is rising annually and affecting women at progressively younger ages [2]. Neoadjuvant chemotherapy (NAC) has become one of the standard treatment strategies for locally advanced breast cancer and for selected patients with early-stage disease who have a strong desire for breast conservation. By shrinking the tumor and lowering the clinical stage, NAC not only increases the rate of breast-conserving surgery but also allows in vivo assessment of tumor chemosensitivity, providing an important basis for adjustment of postoperative adjuvant regimens [3-5].

Pathological complete response (pCR), defined as the absence of residual invasive carcinoma in both the breast primary tumor and the axillary lymph nodes in surgical specimens after completion of NAC (ypT0/is ypN0), is currently recognized as the key pathological endpoint for evaluating the efficacy of NAC. The large CTNeoBC pooled analysis showed that patients achieving pCR had significantly longer event-free survival and overall survival than those without pCR, and this association was particularly prominent in aggressive molecular subtypes such as triple-negative and HER-2-positive breast cancer [6,7]. In clinical practice, pCR status directly influences treatment decisions after NAC: treatment de-escalation or a reduced extent of surgery may be considered for good responders, whereas timely treatment intensification or regimen change is required for poor responders [8]. However, pCR can be confirmed only by pathological examination after completion of NAC and surgery, and reliable methods for predicting treatment response before or during therapy are still lacking. Accurate preoperative identification of patients with a high probability of pCR would

facilitate individualized treatment planning, optimization of surgical timing, and doctor–patient communication.

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) offers high soft-tissue resolution, no ionizing radiation, and multiphasic dynamic evaluation of tumor hemodynamics, and it is currently the most sensitive imaging modality for assessing response to NAC in breast cancer [9,10]. The contrast-enhanced phases reflect tumor microvessel density and vascular permeability, and the third enhanced phase can adequately depict peak tumor enhancement and global perfusion characteristics [11]. Nevertheless, conventional imaging assessment relies mainly on qualitative or semi-quantitative indicators such as tumor size, morphology, and degree of enhancement (e.g., the RECIST criteria); it has limited ability to distinguish residual tumor after NAC from post-treatment fibrotic and necrotic tissue and is susceptible to the subjective experience of the reader. Consequently, its sensitivity and stability for pCR prediction remain unsatisfactory [12,13], and the information on tumor microenvironmental heterogeneity embedded in the images cannot be fully exploited.

Radiomics converts conventional visual images into mineable high-dimensional data through automated extraction of high-throughput quantitative features from medical images. The extracted features encompass three-dimensional tumor shape, first-order gray-level statistics, and higher-order textural features based on the gray level co-occurrence matrix, gray level run length matrix, wavelet transforms, and Laplacian of Gaussian filtering, enabling objective and quantitative characterization of intratumoral heterogeneity [14,15]. Previous studies have demonstrated that radiomic models based on pre-treatment MRI have certain predictive value for pCR after NAC in breast cancer [16,17]. However, conventional radiomic features rely on manual predefinition and are sensitive to differences in ROI segmentation, scanner models, and imaging protocols, which limits their cross-center generalizability [18]; moreover, handcrafted features cannot adequately represent the complex high-order spatial patterns contained in images.

In recent years, deep learning techniques represented by convolutional neural networks (CNNs) have achieved breakthroughs in medical image analysis. CNNs can automatically learn hierarchical feature representations from low-level to high-level directly from raw images, avoiding the limitations of handcrafted feature design [19]. Transfer learning, which transfers network parameters pretrained on large-scale natural image datasets such as ImageNet to medical imaging tasks, effectively mitigates the overfitting caused by the limited availability of medical samples [20]. Compared with three-dimensional (3D) models, 2D deep

learning models based on the maximal tumor cross-section require fewer samples and can be trained more efficiently, and they have shown favorable discriminative performance and cross-center stability in several studies of tumor treatment response prediction [21,22]. Nevertheless, the “black-box” nature of deep learning limits its clinical trust to some extent; meanwhile, deep features and handcrafted radiomic features characterize tumor phenotypes from different dimensions and offer potential complementary value.

At the same time, clinicopathological factors such as ER status, HER-2 status, and tumor size have been established as independent predictors of pCR. Multimodal modeling strategies that fuse clinicopathological risk factors with multi-source imaging features can achieve complementarity among information from different dimensions and may further improve predictive performance [23]. Interpretability methods such as SHapley Additive exPlanations (SHAP) and gradient-weighted class activation mapping (Grad-CAM) can clarify model decision mechanisms at the feature level and the pixel level, respectively, whereas nomograms can further promote clinical translation. However, most existing studies are single-center designs that generally lack independent external validation, and few studies have systematically compared the cross-center predictive performance of conventional radiomics, 2D deep learning, clinical models, and their fusion models within a unified framework. Accordingly, this study included 372 patients with breast cancer receiving NAC from two centers and, based on the third contrast-enhanced phase of pre-treatment DCE-MRI, separately constructed a conventional radiomics model, a 2D deep learning model, a deep learning radiomics model, and a clinical model; we further integrated the above information to build a multimodal combined model and a visual nomogram, performed external validation with data from an independent center, and conducted interpretability analyses combining SHAP and Grad-CAM. The aim was to provide an accurate, robust, and interpretable quantitative tool for individualized preoperative prediction of pCR after NAC in breast cancer.

1 Materials and Methods

1.1 Clinical data

This was a retrospective, multicenter study conducted in accordance with the principles of the Declaration of Helsinki. The study protocol was approved by the institutional medical ethics review committees of both participating centers. Owing to the use of anonymized data, the requirement for informed consent was waived (ethics approval numbers: Guangdong Women

and Children Hospital, 20261015; The First Affiliated Hospital of Jinan University, KY-2025-180).

Clinical and imaging data of female patients with breast cancer who received neoadjuvant chemotherapy (NAC) at two independent centers between January 2025 and September 2026 were retrospectively collected. The inclusion criteria were: ① invasive breast carcinoma pathologically confirmed by core needle biopsy; ② surgery after completion of standardized NAC, with postoperative pathological assessment allowing definitive determination of pathological complete response (pCR) status; ③ breast dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) performed within 2 weeks before initiation of NAC; and ④ complete clinicopathological data. The exclusion criteria were: ① poor DCE-MRI image quality or missing sequences before NAC; ② prior radiotherapy, endocrine therapy, or other antitumor treatment before NAC; ③ concurrent other primary malignancies; or ④ incomplete clinicopathological data.

pCR was defined as no residual invasive carcinoma in either the breast primary tumor or the axillary lymph nodes on postoperative pathological examination (ypT0/is ypN0) [24]. Specifically, ypT0/is denotes the absence of invasive carcinoma in the breast (ductal carcinoma in situ may be present), and ypN0 denotes no axillary lymph node metastasis.

A total of 372 patients were included. To evaluate model generalizability, a multicenter external validation strategy was adopted: data from Center 1 (Guangdong Women and Children Hospital; n = 245) constituted the model development cohort and, after stratification by pCR status, were randomly split at a 7:3 ratio into a training set (n = 171) and an internal validation set (n = 74); data from Center 2 (The First Affiliated Hospital of Jinan University; n = 127) served as the independent external test set. The patient selection process is shown in Figure 1.

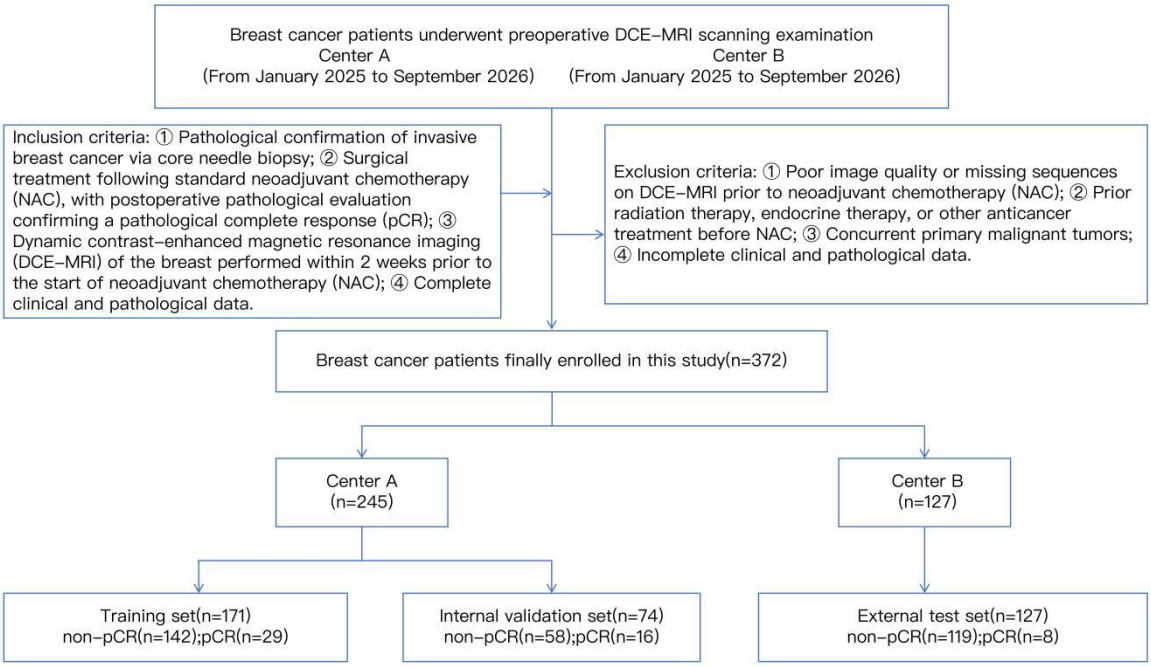


Fig. 1 Flowchart of the study subjects based on inclusion and exclusion criteria.

1.2 Image acquisition

All patients underwent breast DCE-MRI before initiation of NAC. Both scanners operated at a field strength of 3.0 T and were equipped with a 16-channel dedicated breast phased-array surface coil. Center 1 used a Philips Ingenia 3.0 T scanner (Philips Healthcare, the Netherlands) with the following parameters: repetition time/echo time (TR/TE) = 4.4/2.2 ms, slice thickness 1.6 mm, slice gap −0.8 mm, field of view (FOV) 320 × 320 mm, matrix 320 × 320. The DCE-MRI sequence comprised six phases in total, including one pre-contrast phase and five contrast-enhanced phases, with an acquisition time of approximately 60 s per phase. Center 2 used a GE Discovery MR750 3.0 T scanner (GE Healthcare, USA) with TR/TE = 4.1/1.7 ms, slice thickness 1.6 mm, slice gap 0 mm, FOV 300 × 300 mm, and matrix 300 × 300; its DCE-MRI sequence likewise comprised six phases, including one pre-contrast phase and five contrast-enhanced phases, each lasting approximately 60 s.

Gadoteridol was used as the contrast agent at a dose of 0.1 mmol/kg body weight, administered as a bolus through the median cubital vein at a flow rate of 2.0 mL/s, followed by a 15-mL saline flush injected at the same rate. Contrast-enhanced scanning was initiated approximately 20 s after the start of contrast injection. The third phase of DCE-MRI (i.e., the second contrast-enhanced phase, approximately 80 s after injection) was selected for analysis, as this phase can adequately reflect peak tumor enhancement and hemodynamic perfusion characteristics.

1.3 Assessment of clinicopathological characteristics

Clinicopathological characteristics were extracted from the electronic medical record system, including age, menopausal status, age at menarche, family history of breast cancer, maximum tumor diameter, estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER-2), Ki-67 proliferation index, molecular subtype, and axillary lymph node metastasis status. Maximum tumor diameter was measured on preoperative MRI; axillary lymph node status was determined by preoperative ultrasound-guided needle biopsy or sentinel lymph node biopsy.

Immunohistochemical (IHC) markers were evaluated according to the following criteria [25,26]: ER and PR positivity was defined as nuclear staining in ≥1% of tumor cells. HER-2 scoring followed the American Society of Clinical Oncology/College of American Pathologists (ASCO/CAP) guideline criteria: a score of 0 or 1+ was considered negative and 3+ positive; cases scored 2+ were further examined by fluorescence in situ hybridization (FISH), and the presence of gene amplification was classified as positive. A Ki-67 cutoff of 14% was adopted, with ≥14% defined as high expression and <14% as low expression.

According to the consensus of the 2013 St. Gallen International Breast Cancer Conference [27], breast cancer was classified into four molecular subtypes: Luminal A (ER- and/or PR-positive, HER-2-negative, low Ki-67 expression), Luminal B (ER- and/or PR-positive with HER-2-negative and high Ki-67 expression, or HER-2-positive), HER-2-enriched (ER- and PR-negative, HER-2-positive), and triple-negative (ER-, PR-, and HER-2-negative).

1.4 Radiomics and deep learning workflow

The radiomics and deep learning workflow is shown in Figure 2 and consists of five main steps: image segmentation, feature extraction, feature selection, model construction, and performance evaluation.

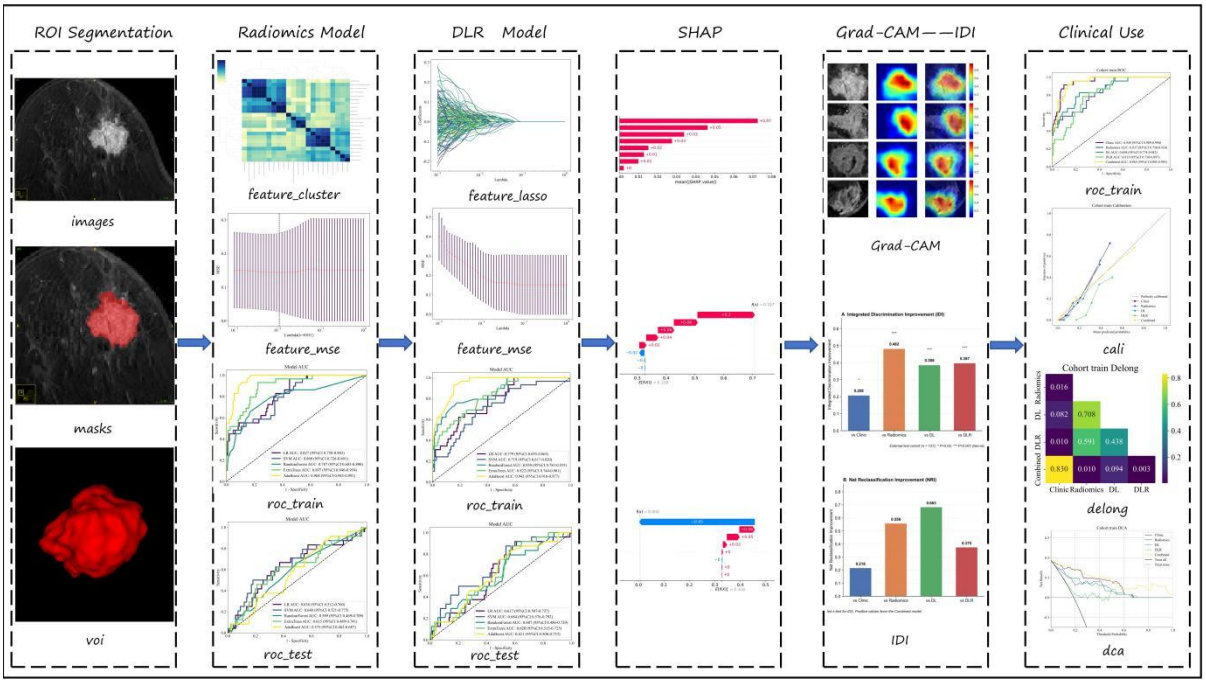


Fig. 2 Workflow of this study.

1.4.1 Image segmentation and definition of the region of interest

Manual segmentation was performed on the third-phase DCE-MRI images using ITK-SNAP software (version 3.8.0; http://www.itksnap.org). A radiologist with 10 years of experience in breast MRI diagnosis manually traced the tumor boundary slice by slice on every section involving the tumor to delineate the region of interest (ROI), taking care to avoid intratumoral necrosis, hemorrhage, and cystic degeneration. All ROIs were reviewed and confirmed by a senior radiologist with more than 20 years of experience in breast imaging diagnosis. The delineating radiologist was blinded to the patients' pathological and clinical information.

1.4.2 Conventional radiomic feature extraction and selection

Radiomic feature extraction was performed using the open-source PyRadiomics Python package (version 3.0.1; http://pyradiomics.readthedocs.io). Before feature extraction, standardized image preprocessing was applied: image intensities were normalized using the Z-score method to eliminate signal intensity differences across scanners and imaging protocols. The normalized value (i.e., the z-score) was calculated as follows:

$$z_i = \frac{x_i - \mu}{\sigma}$$

where x_i denotes the intensity of the i-th voxel, μ the mean image intensity, and σ the standard deviation of image intensity. The normalized images were resampled to an isotropic resolution of 2 × 2 × 2 mm^3 using B-spline interpolation; the gray-level discretization bin

width was set to 5, and a gray-level shift (voxelArrayShift = 300) was applied to ensure that the vast majority of voxel values were positive.

Three image filtering schemes were used for feature extraction: the original image, Laplacian of Gaussian (LoG) filtering (sigma values of 2.0, 3.0, 4.0, and 5.0 mm), and wavelet transformation. The extracted handcrafted features covered six feature classes: ① shape features, describing the three-dimensional spatial morphology of the tumor; ② first-order features, reflecting the statistical distribution of voxel intensities within the ROI; ③ gray level co-occurrence matrix (GLCM) features, characterizing second-order textural information; ④ gray level run length matrix (GLRLM) features; ⑤ gray level size zone matrix (GLSZM) features; and ⑥ gray level dependence matrix (GLDM) and neighborhood gray tone difference matrix (NGTDM) features. A total of 1,197 radiomic features were extracted for each patient.

A two-step feature selection strategy was adopted to reduce data dimensionality and screen for non-redundant, predictive features. The first step was removal of redundant features: pairwise Spearman correlation coefficients were calculated for all radiomic features, and for feature pairs with $|r| \geq 0.90$ only one feature was retained to eliminate highly collinear features. The second step was LASSO feature selection: least absolute shrinkage and selection operator (LASSO) regression was applied to the remaining features for further dimensionality reduction, achieving sparse feature selection through an L1 regularization penalty, and the optimal model was determined using the penalty parameter λ corresponding to the minimum mean squared error of ten-fold cross-validation (lambda.min). Ten radiomic features were ultimately retained for construction of the radiomics model.

Subsequently, multiple machine learning algorithms were used to build classification models, including logistic regression (LR), support vector machine (SVM), random forest (RF), Extra Trees (ET), and AdaBoost. The algorithm achieving the best combined performance on five-fold cross-validation in the training set and on the internal validation set was selected as the final radiomics model.

1.4.3 Deep learning pipeline

(1) Data preparation: Two-dimensional images generated with the transverse crop_tra strategy served as input to the deep learning models. Specifically, for each patient, the transverse image at the maximal tumor cross-section was selected and cropped to the minimum bounding rectangle containing the ROI. Image preprocessing included Z-score normalization

of the RGB channels to standardize the intensity distribution. During training, a real-time data augmentation strategy was used to improve model generalizability, including random cropping, horizontal flipping, and vertical flipping; during testing, only normalization was performed.

(2) Transfer learning and model training: A transfer learning strategy was adopted to overcome the limitations imposed by the small scale of the medical imaging dataset. All deep learning models were initialized with weights pretrained on the large-scale ImageNet natural image dataset and subsequently fine-tuned on the target dataset. To systematically identify the optimal network architecture, six mainstream convolutional neural network models were compared: MobileNet V2, MNASNet 1.0, SqueezeNet 1.0, DenseNet121, ResNet101, and VGG19.

Stochastic gradient descent (SGD) was used as the optimizer, with softmax cross-entropy as the loss function. The learning rate was scheduled using a cosine annealing decay strategy, defined as follows:

$$\eta_{\mathrm{t}} = \eta_{\min}^{i} + \frac{1}{2}\left(\eta_{\max}^{i} - \eta_{\min}^{i}\right)\left(1 + \cos\left(\frac{\mathrm{T_{cur}}}{\mathrm{T}_{i}} \cdot \pi\right)\right)$$

where η_min denotes the minimum learning rate (set to 0), η_max the maximum learning rate (set to 0.01), and T_i the total number of iterations (set to 32). This strategy keeps the learning rate relatively high early in training to accelerate convergence and gradually decays it later for fine parameter adjustment. All models were trained for 32 epochs.

(3) Model selection: The model and the optimal epoch were determined solely on the basis of the five-fold cross-validation classification performance in the training set and the performance on the internal validation set; the external test set did not participate in any model selection or hyperparameter tuning and was used only for a one-time independent evaluation of the final model. Considering the combined discriminative performance of training cross-validation and internal validation, the MobileNet V2 model at epoch 22 was ultimately selected as the optimal deep learning model (deep learning signature, DL signature), and its output probability served directly as the deep learning prediction.

1.4.4 Deep learning radiomics features (DLR)

To integrate the interpretability of conventional radiomic features with the ability of deep features to capture high-order spatial patterns, deep learning radiomics (DLR) features were constructed as follows. First, deep features were extracted from the penultimate layer of the

optimal deep learning model (MobileNet V2), i.e., the fully connected layer before the global average pooling layer. Given the high dimensionality of the deep features, principal component analysis (PCA) was applied to compress them into eight principal component features (DL_0 to DL_7), retaining more than 95% of the variance information. An early fusion strategy was then used to concatenate the eight deep learning features with the conventional radiomic features to form a combined feature vector. The radiomics feature selection procedure (Spearman correlation-based redundancy removal → LASSO sparse selection) was repeated on the fused feature set, and eight features (one deep learning feature and seven radiomic features) were ultimately selected for construction of the DLR model. The algorithms LR, SVM, RF, ET, and AdaBoost were likewise compared, and SVM was ultimately selected to build the DLR model.

1.4.5 Construction of the clinical and combined models

(1) Clinical signature: First, univariate logistic regression was performed for all 11 clinicopathological characteristics in the training set, using $P < 0.10$ as the threshold for screening candidate variables; ER, PR, HER-2, molecular subtype, axillary lymph node metastasis, and maximum tumor diameter were selected. Given that PR was highly correlated with ER (Spearman $r = 0.80$) and molecular subtype is derived from ER, PR, HER-2, and Ki-67 status (correlation coefficient $r = -0.87$ with ER status), ER status, HER-2 status, maximum tumor diameter, and axillary lymph node metastasis were ultimately entered into the multivariate logistic regression model to avoid multicollinearity. Multivariate analysis confirmed ER status, HER-2 status, and maximum tumor diameter as independent predictors of pCR ($P < 0.05$), whereas axillary lymph node metastasis did not reach statistical significance ($P = 0.075$). The clinical model was constructed based on these three independent predictors; after comparison of algorithms including LR, SVM, RF, ET, and AdaBoost, ExtraTrees was ultimately selected to build the clinical model.

(2) Combined signature: To integrate the complementary information of multimodal data, the prediction output of the DLR model (the pCR probability) was fused with the independent clinical predictors identified by multivariate analysis, and a linear combination model was constructed using logistic regression. The combined model was visualized as a nomogram to facilitate clinical application.

1.5 Statistical analysis

The Shapiro–Wilk test was used to assess the normality of continuous variables. Continuous variables conforming to a normal distribution are expressed as mean ± standard deviation and were compared between groups using the independent-samples t test; non-normally distributed continuous variables are expressed as median (interquartile range) and were compared using the Mann–Whitney U test. Categorical variables are expressed as frequency (percentage) and were compared using the chi-square test or Fisher's exact test. Baseline characteristics across the cohorts are compared in Table 1.

Univariate and multivariate analyses of clinical characteristics were performed using logistic regression models, with results expressed as odds ratios (ORs) and their 95% confidence intervals (CIs). Radiomic feature selection was performed using Spearman correlation analysis combined with LASSO regression.

Model performance evaluation included the following aspects: ① Discrimination: the receiver operating characteristic (ROC) curve and area under the curve (AUC) were used to evaluate the classification and discrimination ability of each model, and the 95% CI, accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated. ② Calibration: calibration curves were used to assess the agreement between model-predicted probabilities and actual observed probabilities, with quantitative verification by the Hosmer–Lemeshow goodness-of-fit test. ③ Clinical utility: decision curve analysis (DCA) was used to evaluate net benefit across different threshold probabilities. ④ Between-model comparisons: the DeLong test was used to determine whether differences in AUC between models were statistically significant.

All statistical analyses were performed using Python 3.7.12 and Statsmodels 0.13.2, and radiomic features were extracted using PyRadiomics 3.0.1. Machine learning modeling (SVM, RF, etc.) was implemented with Scikit-learn 1.0.2, and the deep learning framework was built on PyTorch 1.11.0 running on CUDA 11.3.1 with cuDNN 8.2.1. The design and reporting of this study followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement and the Checklist for Artificial Intelligence in Medical Imaging (CLAIM). A two-sided $P < 0.05$ was considered to indicate a statistically significant difference.

2 Results

2.1 Baseline clinical characteristics of patients

A total of 372 patients with breast cancer receiving NAC were included, among whom 53 (14.2%) were pathologically confirmed to achieve pCR after surgery. The pCR rates in the training, internal validation, and external test sets were 17.0% (29/171), 21.6% (16/74), and 6.3% (8/127), respectively, with a statistically significant difference among cohorts (χ^2 = 10.89, P = 0.004); the pCR rate of the external test set was significantly lower than those of the development cohorts, reflecting differences in the real-world population composition between the two centers.

The baseline clinicopathological characteristics of the three cohorts are compared in Table 1. Age (P = 0.012), maximum tumor diameter (P < 0.001), menopausal status (P < 0.001), ER status (P = 0.003), HER-2 status (P < 0.001), molecular subtype (P = 0.007), and axillary lymph node metastasis status (P < 0.001) were distributed significantly differently across cohorts, whereas age at menarche (P = 0.116), family history of breast cancer (P = 0.372), PR status (P = 0.227), and Ki-67 expression level (P = 0.431) showed no significant differences. The baseline differences across cohorts objectively reflect discrepancies in population composition and clinical practice among centers in the real world: patients in the external test set were older, had smaller tumors, included a higher proportion of postmenopausal women, and had higher proportions of ER positivity and lymph node positivity. This baseline imbalance made the independent external validation a more stringent test of model generalizability and further underscores the necessity of evaluating cross-center stability in this cohort.

Table 1. Baseline characteristics of patients in the training, internal validation, and external test cohorts.

Characteristic	Overall (n=372)	External test (n=127)	Training (n=171)	Internal validation (n=74)	P value
Age (years)	50.08±9.74	52.09±9.26	48.73±9.87	49.73±9.81	0.012
Age at menarche (years)	12.33±1.16	12.19±0.93	12.46±1.37	12.27±0.93	0.116
Tumor maximum diameter (cm)	3.91±1.84	3.29±1.44	4.18±1.84	4.36±2.16	<0.001
Menopausal status					<0.001
Premenopausal	191 (51.3%)	48 (37.8%)	100 (58.5%)	43 (58.1%)	
Postmenopausal	181 (48.7%)	79 (62.2%)	71 (41.5%)	31 (41.9%)	
Family history of breast cancer					0.372
No	345 (92.7%)	120 (94.5%)	159 (93.0%)	66 (89.2%)	
Yes	27 (7.3%)	7 (5.5%)	12 (7.0%)	8 (10.8%)	
ER status					0.003
Negative	129 (34.7%)	30 (23.6%)	65 (38.0%)	34 (45.9%)	
Positive	243 (65.3%)	97 (76.4%)	106 (62.0%)	40 (54.1%)	

Characteristic	Overall (n=372)	External test (n=127)	Training (n=171)	Internal validation (n=74)	P value
PR status					0.227
Negative	168 (45.2%)	54 (42.5%)	74 (43.3%)	40 (54.1%)	
Positive	204 (54.8%)	73 (57.5%)	97 (56.7%)	34 (45.9%)	
HER-2 status					<0.001
Negative	247 (66.4%)	103 (81.1%)	100 (58.5%)	44 (59.5%)	
Positive	125 (33.6%)	24 (18.9%)	71 (41.5%)	30 (40.5%)	
Ki-67 expression					0.431
Low expression	46 (12.4%)	16 (12.6%)	24 (14.0%)	6 (8.1%)	
High expression	326 (87.6%)	111 (87.4%)	147 (86.0%)	68 (91.9%)	
Molecular subtype					0.007
Luminal A	32 (8.6%)	8 (6.3%)	19 (11.1%)	5 (6.8%)	
Luminal B	221 (59.4%)	91 (71.7%)	91 (53.2%)	39 (52.7%)	
HER-2-enriched	62 (16.7%)	10 (7.9%)	37 (21.6%)	15 (20.3%)	
Triple-negative	57 (15.3%)	18 (14.2%)	24 (14.0%)	15 (20.3%)	
Axillary lymph node metastasis					<0.001
Negative	147 (39.5%)	33 (26.0%)	79 (46.2%)	35 (47.3%)	
Positive	225 (60.5%)	94 (74.0%)	92 (53.8%)	39 (52.7%)	

Note: Continuous variables are expressed as mean ± standard deviation; categorical variables are expressed as number (percentage). ER, estrogen receptor; PR, progesterone receptor; HER-2, human epidermal growth factor receptor 2; pCR, pathological complete response.

2.2 Univariate and multivariate logistic regression analyses of clinical characteristics

To identify clinical factors associated with pCR, we performed univariate and multivariate logistic regression analyses of 11 clinicopathological characteristics in the training set (n = 171) (Table 2). Univariate analysis showed that ER status, PR status, HER-2 status, molecular subtype, and axillary lymph node metastasis were significantly associated with pCR ($P < 0.05$), whereas maximum tumor diameter showed a marginal association (OR = 0.774, 95% CI: 0.596–1.005, P = 0.055); the remaining characteristics showed no statistically significant differences. Using $P < 0.10$ as the candidate threshold and excluding PR (Spearman r = 0.80 with ER), which was highly correlated with ER, and molecular subtype ($r = -0.87$), which is derived from ER/PR/HER-2/Ki-67 status, to avoid multicollinearity before modeling, ER status, HER-2 status, maximum tumor diameter, and axillary lymph node metastasis were entered into the multivariate logistic regression. The results showed that ER status (adjusted OR = 0.119, 95% CI: 0.040–0.355, $P < 0.001$), HER-2 status (adjusted OR = 4.870, 95% CI: 1.686–14.063, P = 0.003), and maximum tumor diameter (adjusted OR = 0.742, 95% CI: 0.558–0.988, P = 0.041) were independent predictors of pCR, whereas axillary lymph node metastasis did not reach statistical significance (adjusted OR = 0.397, 95% CI: 0.144–1.097, P

= 0.075). Specifically, ER-positive patients were significantly less likely to achieve pCR than ER-negative patients; HER-2-positive patients were approximately 4.87 times as likely to achieve pCR as HER-2-negative patients; and for each 1-cm increase in maximum tumor diameter, the likelihood of achieving pCR decreased by approximately 25.8%. These findings are consistent with the biological patterns of breast cancer response to NAC: ER-positive/Luminal tumors are relatively insensitive to cytotoxic chemotherapy, HER-2-positive tumors derive marked benefit from concurrent anti-HER-2 targeted therapy, and a greater tumor burden generally implies a higher probability of residual lesions. Based on these three independent predictors, the clinical model was constructed using the ExtraTrees algorithm. A similar study has shown that, in breast cancer-related genomic feature prediction tasks, the ExtraTrees algorithm achieved the highest predictive performance among models, with an accuracy of 0.999 and a 95% confidence interval of (0.998–1.000), making it well suited to clinical prediction needs.

Table 2. Univariate and multivariate logistic regression analyses of clinical characteristics in relation to pCR.

Characteristic	Univariate analysis		Multivariate analysis	
	OR (95% CI)	P value	OR (95% CI)	P value
ER	0.085 (0.030–0.237)	<0.001	0.119 (0.040–0.355)	<0.001
PR	0.146 (0.056–0.382)	<0.001	—	—
Family history	0.425 (0.053–3.430)	0.422	—	—
Axillary lymph node metastasis	0.263 (0.109–0.634)	0.003	0.397 (0.144–1.097)	0.075
Menopausal status	0.581 (0.247–1.364)	0.212	—	—
Ki-67	5.412 (0.701–41.788)	0.105	—	—
HER-2	7.507 (2.865–19.673)	<0.001	4.870 (1.686–14.063)	0.003
Molecular subtype	2.373 (1.478–3.810)	<0.001	—	—
Tumor diameter	0.774 (0.596–1.005)	0.055	0.742 (0.558–0.988)	0.041
Age at menarche	0.968 (0.710–1.318)	0.835	—	—
Age	0.967 (0.927–1.008)	0.113	—	—

Note: Analyses were conducted on the training set (n = 171). OR, odds ratio; aOR, adjusted odds ratio from the multivariable logistic model; CI, confidence interval. Variables with P < 0.10 in univariate analysis were candidates; PR and molecular subtype were excluded before multivariable modeling owing to strong collinearity (PR vs ER: Spearman r = 0.80; molecular subtype is defined by ER/PR/HER-2/Ki-67 status, r = −0.87 with ER). The multivariable model comprised ER, HER-2, maximum tumor diameter, and axillary lymph node metastasis; "—" indicates that the variable was not entered into the multivariable model. ER, estrogen receptor; PR, progesterone receptor; HER-2, human epidermal growth factor receptor 2.

2.3 Radiomic feature screening and the radiomics model

Based on the manually segmented ROIs on the third-phase DCE-MRI images, a total of 1,197 radiomic features were extracted. Highly redundant features with $|r| \geq 0.90$ were first removed

by Spearman correlation analysis, and the remaining features were then subjected to LASSO sparse selection using the λ corresponding to the minimum ten-fold cross-validated mean squared error (lambda.min) as the optimal penalty parameter. Ten optimal radiomic features were ultimately selected, including two LoG-filtered first-order features (log-sigma-3-0-mm-3D_firstorder_Kurtosis and log-sigma-4-0-mm-3D_firstorder_Minimum), one shape feature (original_shape_Sphericity), one original-image GLSZM feature (original_glszm_ZoneEntropy), three wavelet first-order features (wavelet-HHL_firstorder_Kurtosis, wavelet-HLH_firstorder_Maximum, and wavelet-LHH_firstorder_Skewness), two wavelet GLCM features (wavelet-HHL_glcm_Idn and wavelet-HLH_glcm_ClusterShade), and one wavelet NGTDM feature (wavelet-HLH_ngtdm_Strength). The LASSO coefficient convergence paths and the weights of each feature are shown in Supplementary Figures S1–S2. Notably, the selected features were dominated by wavelet- and LoG-filtered features (7/10), suggesting that fine spatial texture and multiscale textural information of the tumor are of substantial value for pCR prediction.

Based on these 10 features, after comparison of algorithms including LR, SVM, RF, ET, and AdaBoost by five-fold cross-validation in the training set, SVM was selected to construct the final radiomics model (Radiomics signature). The AUCs of this model in the training, internal validation, and external test sets were 0.837 (95% CI: 0.751–0.924), 0.533 (95% CI: 0.225–0.841), and 0.559 (95% CI: 0.345–0.773), respectively (Table 3). The radiomics model performed well on the training set but showed marked performance degradation on the internal validation and independent external test sets (the AUC decreased by 0.278 from the training set to the external test set), suggesting that handcrafted features based on fixed mathematical definitions are sensitive to cross-center differences in scanners and imaging protocols and have limited cross-center generalizability.

2.4 Construction and performance of the 2D deep learning model

Using the transverse cropped image at the maximal tumor cross-section (crop_tra) as model input, we systematically compared the pCR prediction performance of six ImageNet-pretrained convolutional neural networks (MobileNet V2, MNASNet 1.0, SqueezeNet 1.0, DenseNet121, ResNet101, and VGG19) under identical training configurations. Changes in ROC curves for training cross-validation and internal validation were monitored epoch by epoch, and the optimal network architecture and epoch were determined according to their combined discriminative performance (the external test set did not participate in model

selection). The MobileNet V2 model at epoch 22 was ultimately selected as the optimal deep learning model (DL signature), and its generalizability was evaluated once on the external test set. Its AUCs in the training, internal validation, and external test sets were 0.858 (95% CI: 0.774–0.943), 0.817 (95% CI: 0.587–1.000), and 0.825 (95% CI: 0.677–0.973), respectively (Table 3). Notably, the AUC of the deep learning model decreased by only 0.033 from the training set to the external test set, and its cross-center stability was significantly superior to that of the conventional radiomics model (decrease of 0.278), indicating that deep features automatically learned through transfer learning possess stronger cross-center transferability.

2.5 Construction of deep learning radiomics features (DLR)

To integrate the interpretability of handcrafted features with the high-order representational capacity of deep features, we extracted deep features from the penultimate layer of the optimal deep learning model, MobileNet V2, reduced them by PCA to eight principal components (DL_0–DL_7), and early-fused them with the radiomic features by concatenation. After Spearman redundancy removal and LASSO sparse selection of the fused feature set, eight optimal features were ultimately retained for construction of the DLR model, including one deep learning feature (DL_7) and seven radiomic features (log-sigma-3-0-mm-3D_firstorder_Kurtosis, log-sigma-5-0-mm-3D_gldm_LargeDependenceLowGrayLevelEmphasis, original_shape_Sphericity, wavelet-HHL_firstorder_Kurtosis, wavelet-HLH_glcm_ClusterShade, wavelet-LHH_firstorder_Skewness, and wavelet-LLL_ngtdm_Contrast) (Supplementary Figure S3). Algorithm comparison on the fused feature set showed that SVM performed most robustly, and SVM was therefore used to build the DLR model. The AUCs of this model in the training, internal validation, and external test sets were 0.813 (95% CI: 0.730–0.897), 0.542 (95% CI: 0.256–0.828), and 0.745 (95% CI: 0.587–0.903), respectively (Table 3).

2.6 Comprehensive comparison of the performance of the five prediction models

The detailed classification performance of the five prediction models constructed in this study (Clinic, Radiomics, DL, DLR, and Combined) in the three cohorts is shown in Table 3; ROC curves are presented in Figure 3, and confusion matrices and predicted probability distributions are provided in the Supplementary Materials.

Table 3. Predictive performance of the five models in the training, internal validation, and external test cohorts.

Model	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Cohort
Clinic	0.902	0.949	0.909–0.990	0.913	0.900	0.677	0.978	Training
Radiomics	0.846	0.837	0.751–0.924	0.565	0.910	0.591	0.901	Training
DL	0.780	0.858	0.774–0.943	0.826	0.770	0.452	0.951	Training
DLR	0.748	0.813	0.730–0.897	0.783	0.740	0.409	0.937	Training
Combined	0.854	0.945	0.905–0.985	0.957	0.830	0.564	0.988	Training
Clinic	0.538	0.654	0.436–0.873	1.000	0.400	0.333	1.000	Internal validation
Radiomics	0.769	0.533	0.225–0.841	0.333	0.900	0.500	0.818	Internal validation
DL	0.885	0.817	0.587–1.000	0.833	0.900	0.714	0.947	Internal validation
DLR	0.731	0.542	0.256–0.828	0.500	0.800	0.429	0.842	Internal validation
Combined	0.654	0.658	0.406–0.910	0.833	0.600	0.385	0.923	Internal validation
Clinic	0.898	0.925	0.854–0.995	0.875	0.899	0.368	0.991	External test
Radiomics	0.512	0.559	0.345–0.773	0.750	0.496	0.091	0.967	External test
DL	0.858	0.825	0.677–0.973	0.750	0.866	0.273	0.981	External test
DLR	0.504	0.745	0.587–0.903	1.000	0.471	0.113	1.000	External test
Combined	0.843	0.933	0.879–0.987	1.000	0.832	0.286	1.000	External test

Note: AUC, area under the curve; PPV, positive predictive value; NPV, negative predictive value. Clinic, clinical model; Radiomics, radiomics model; DL, deep learning model; DLR, deep learning radiomics model; Combined, combined model.

In the training set, the clinical model achieved the highest discriminative performance (AUC = 0.949), followed closely by the combined model (AUC = 0.945). In the internal validation set, the deep learning model performed best (AUC = 0.817). The key finding was observed in the independent external test set: the Combined model achieved the best overall discriminative performance of all models, with an AUC of 0.933 (95% CI: 0.879–0.987), accuracy of 0.843, sensitivity of 1.000, specificity of 0.832, and NPV of 1.000; the AUC was 0.925 for the clinical model, 0.825 for the deep learning model, 0.745 for the DLR model, and only 0.559 for the radiomics model. Compared with the clinical model, the combined model improved the external test AUC by 0.008 and increased sensitivity from 0.875 to 1.000, meaning that none of the eight pCR patients in the external test set was missed. In a low-prevalence population with a pCR rate of only 6.3%, the combined model achieved an NPV of 1.000, supporting its high clinical value as a preoperative rule-out tool. It should be noted that, owing to the low prevalence (6.3%) in the external test set, PPVs were generally low across models (0.091–0.368); this is an inevitable consequence of Bayes' theorem and cannot be simply interpreted as inadequate model performance.

From the perspective of cross-center generalization, the five models exhibited distinctly different performance degradation patterns: the conventional radiomics model degraded most severely (training 0.837 → external test 0.559; ΔAUC = 0.278), followed by the DLR model (0.813 → 0.745; ΔAUC = 0.068), whereas the deep learning model (0.858 → 0.825; ΔAUC = 0.033) and the clinical model (0.949 → 0.925; ΔAUC = 0.024) were the most stable. While

maintaining external discriminative performance comparable to that of the clinical model, the combined model integrated imaging information and achieved the highest AUC (0.933) of all models, realizing effective complementarity of multimodal information.

The DeLong test results showed that, in the external test set, the AUC of the combined model was significantly higher than those of the radiomics model (P = 0.0007) and the DLR model (P = 0.014), whereas the differences from the deep learning model (P = 0.162) and the clinical model (P = 0.795) were not statistically significant; the complete pairwise AUC comparisons among models are shown in Figure 5. Calibration curve analysis (Figure 4) showed good agreement between the predicted and observed probabilities of the combined model in the training, internal validation, and external test sets, with calibration curves close to the ideal diagonal; the Hosmer–Lemeshow goodness-of-fit test indicated good calibration of the combined model in the external test set (χ^2 = 14.53, df = 8, P = 0.069). Decision curve analysis (Figure 6) showed that, across a wide range of threshold probabilities, the net benefit of the combined model was higher than those of all single-modality models and of the two extreme strategies of "treat all" or "treat none", confirming its favorable clinical utility.

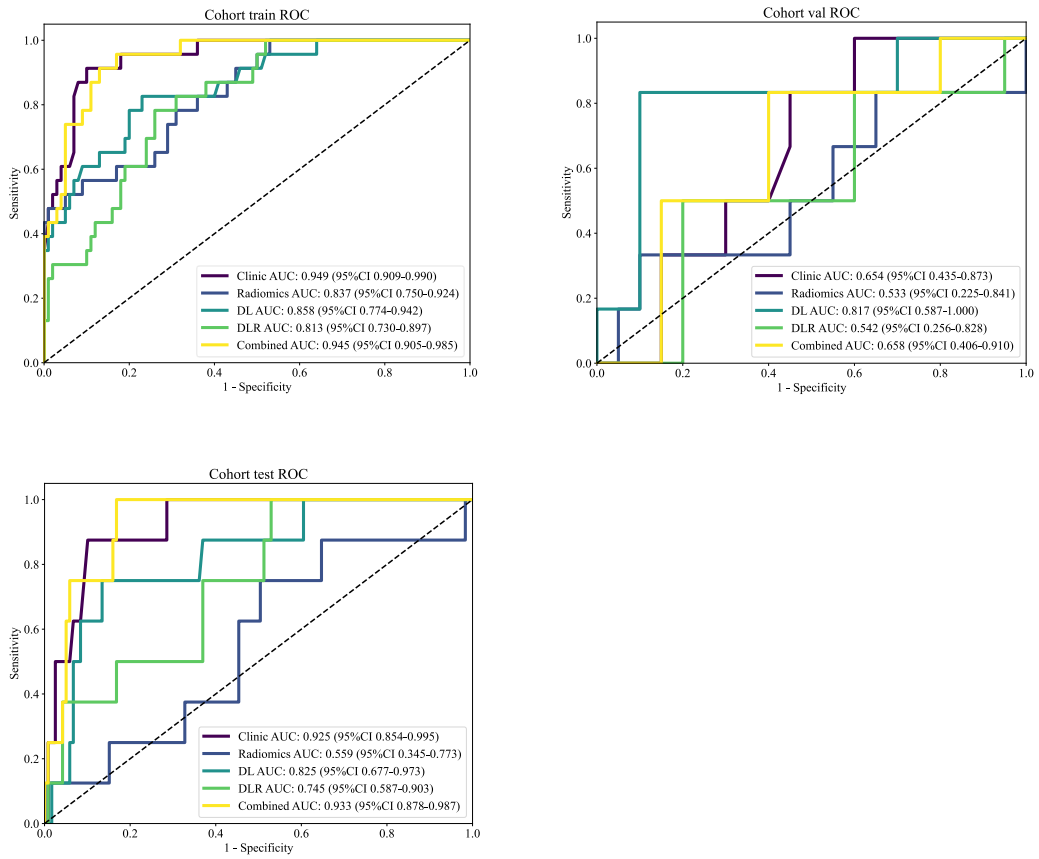


Fig. 3 ROC curves of different signatures across cohorts.

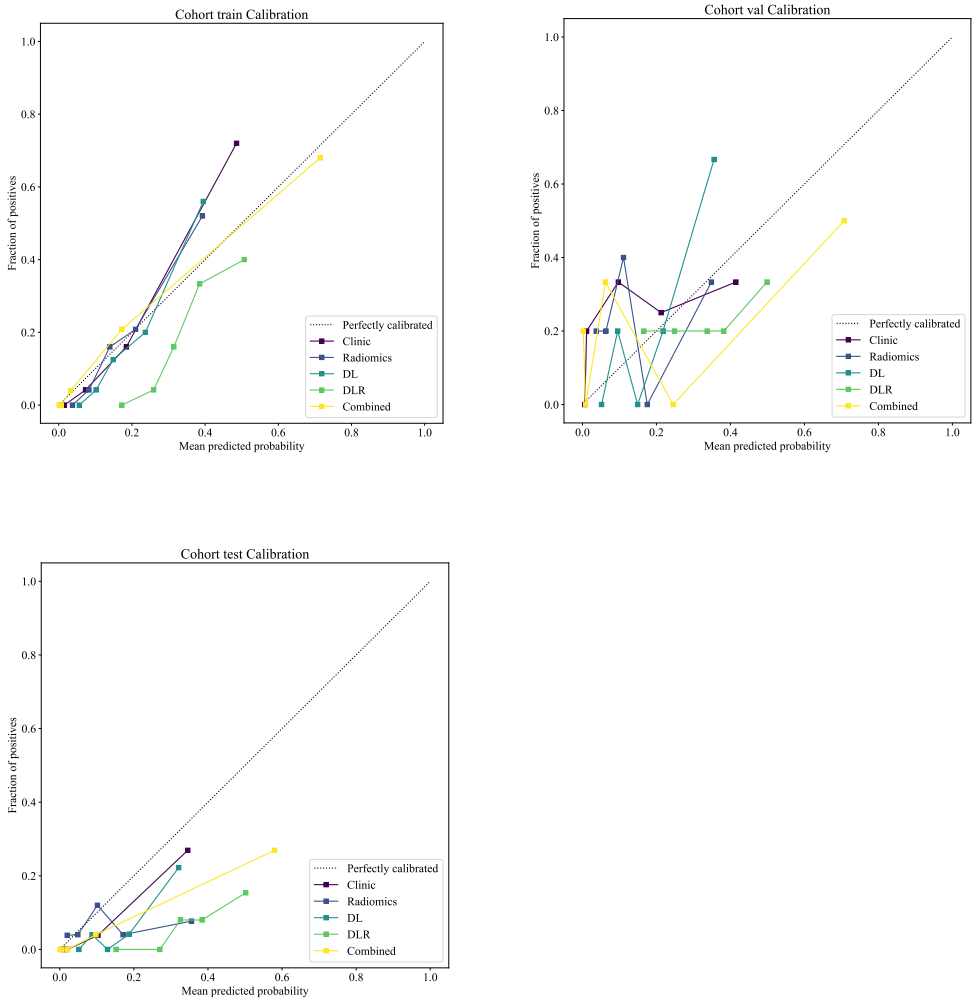
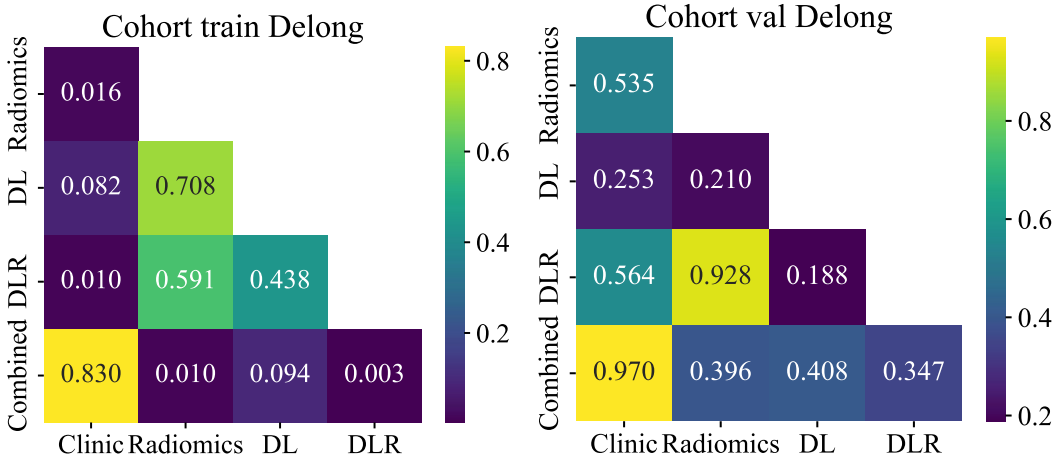


Fig. 4 Calibration curves of different signatures across the three cohorts.



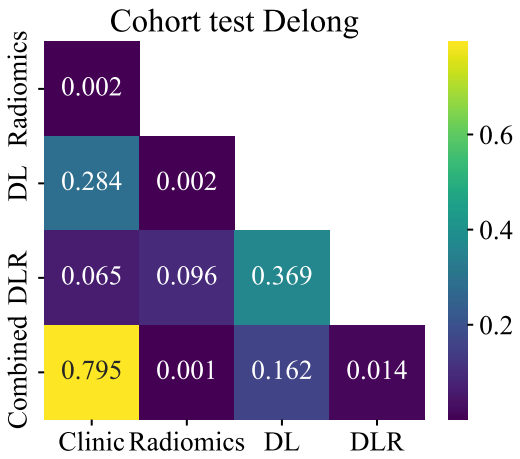


Fig. 5 DeLong test results for different signatures.

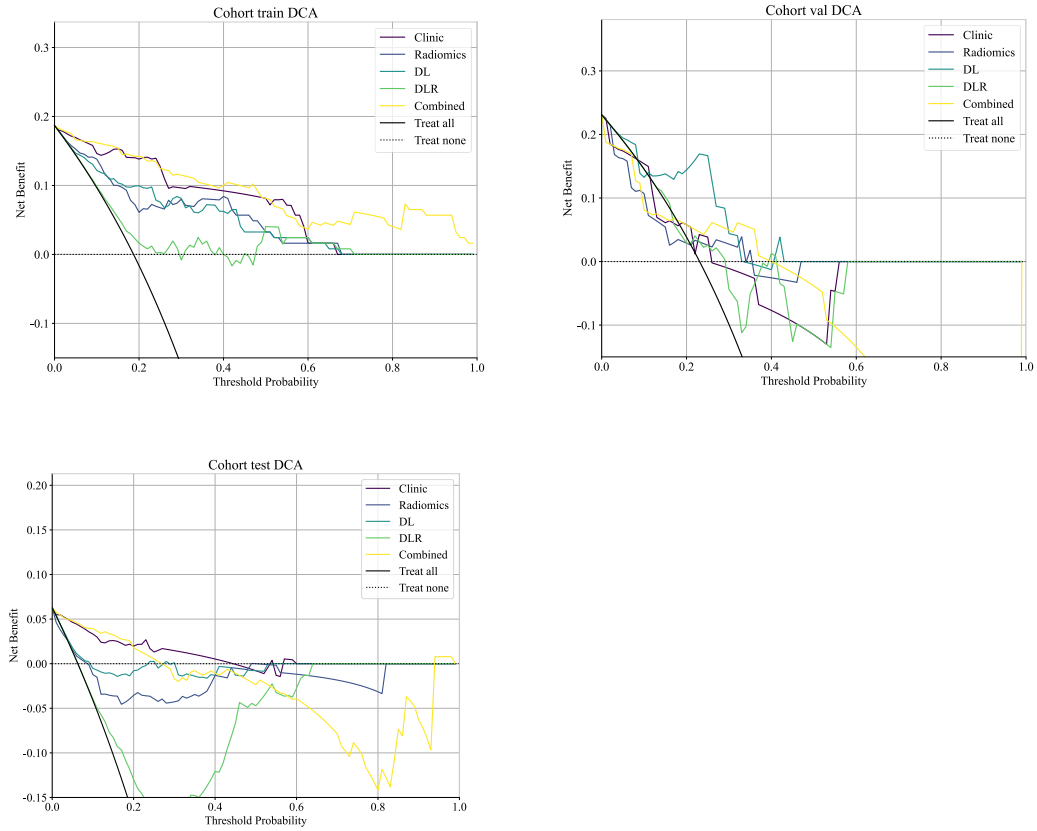


Fig. 6 Decision curves of different signatures for the external test cohort.

2.7 Reclassification analysis of the incremental value of the models

To further quantify the incremental predictive value of the combined model relative to each single-modality model, we calculated the integrated discrimination improvement (IDI) and net reclassification improvement (NRI) in the independent external test set (Figure 7). The results showed that the IDI of the combined model relative to all four single-modality models was

positive and statistically significant: IDI = 0.208 (P = 0.028) relative to the clinical model, 0.482 (P < 0.001) relative to the radiomics model, 0.386 (P = 0.0002) relative to the deep learning model, and 0.397 (P < 0.001) relative to the DLR model, indicating that the combined model significantly improved the overall discrimination between pCR and non-pCR patients compared with each single-modality model. The NRI results were consistent: the NRIs of the combined model relative to the clinical, radiomics, deep learning, and DLR models were 0.216, 0.558, 0.683, and 0.375, respectively, suggesting that the combined model could reclassify a substantial proportion of patients into more accurate risk categories. These findings confirm that the multimodal fusion strategy provided stable and statistically significant incremental predictive value in the independent external cohort.

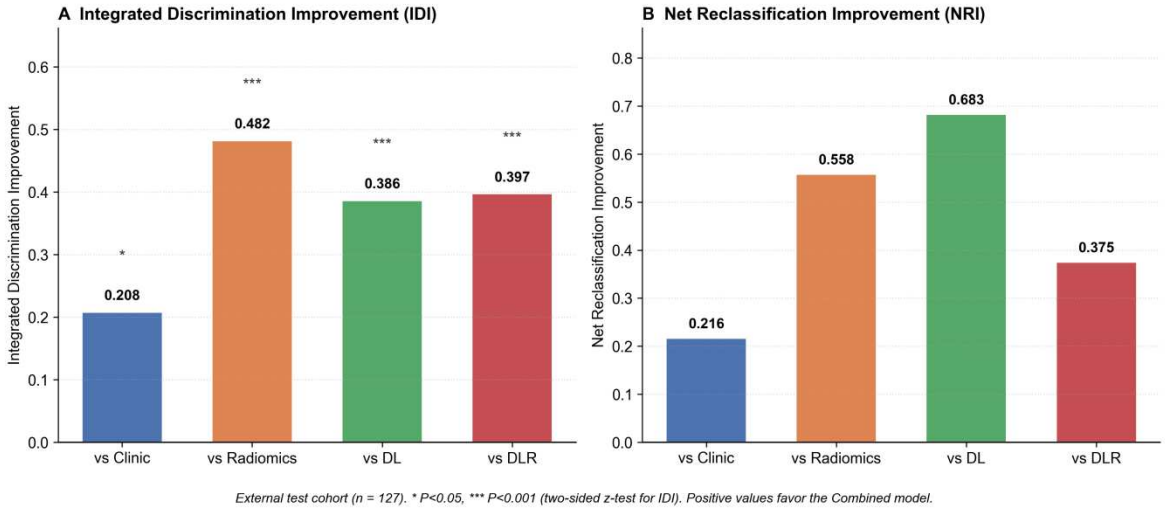


*Figure 7. Incremental value of the combined model over the four single-modality models in the external test cohort (n = 127). (A) Integrated Discrimination Improvement (IDI); (B) Net Reclassification Improvement (NRI). Positive values indicate that the combined model improves discrimination (IDI) and risk reclassification (NRI) relative to the corresponding single-modality model. * P < 0.05, *** P < 0.001 (two-sided z-test for IDI).*

2.8 Model interpretability analysis

To elucidate the decision mechanism of the DLR model, we used SHAP (SHapley Additive exPlanations) to explain the predictions of the final DLR model on all 372 patients (Figure 8A). The feature importance ranking based on mean absolute SHAP values (mean |SHAP|) showed that the three features contributing most to model predictions were, in descending order, log-sigma-3-0-mm-3D_firstorder_Kurtosis (mean |SHAP| = 0.073), the deep learning feature DL_7 (0.046), and wavelet-LLL_ngtdm_Contrast (0.034), followed in order by wavelet-LHH_firstorder_Skewness (0.028), wavelet-HLH_glcm_ClusterShade (0.015),

original_shape_Sphericity (0.013), log-sigma-5-0-mm-3D_gldm_LargeDependenceLowGrayLevelEmphasis (0.010), and wavelet-HHL_firstorder_Kurtosis (0.002). Feature direction analysis showed that lower log-sigma-3-0 first-order kurtosis, lower wavelet-LLL NGTDM contrast, higher DL_7, higher wavelet-LHH first-order skewness, higher wavelet-HLH GLCM cluster shade, and greater tumor sphericity were associated with a higher predicted probability of pCR (all |Spearman r| between SHAP values and feature values > 0.97, $P < 0.001$). Notably, the two features ranking first and second in contribution were derived from multiscale-filtered radiomics (LoG first-order kurtosis) and deep learning (DL_7), respectively, forming a complementary pair and confirming, at the feature level, the effectiveness of the DLR fusion strategy.

To further illustrate the individualized decision process of the model, SHAP waterfall plots for two representative cases are presented in Figures 8B–C. Case 1 (Figure 8B) had a predicted pCR probability of 0.707 (population baseline expected value E[f(X)] = 0.328). The features driving its prediction toward pCR were, in order: log-sigma-3-0 first-order kurtosis (value −1.66, markedly below the population mean; SHAP value +0.20, the largest contribution), DL_7 (value +1.27; +0.08), wavelet-LHH first-order skewness (value +1.38; +0.06), wavelet-HLH GLCM cluster shade (value +1.52; +0.04), and sphericity (value +1.46; +0.02); only the wavelet-LLL NGTDM contrast made a slight negative contribution (−0.02). The predicted probability of Case 2 (Figure 8C) was only 0.004, and its key determinant was an abnormally elevated wavelet-LLL NGTDM contrast (value +12.70, approximately 13 standard deviations above the population mean); the SHAP value of this single feature reached −0.45, exceeding the sum of the positive contributions of all other features and driving the predicted probability down from the baseline of 0.328 to nearly zero. This contrast suggests that the intratumoral enhancement heterogeneity captured by wavelet-LLL NGTDM contrast may be a key imaging phenotype of non-pCR; meanwhile, the pCR determination of the model was driven jointly by multiscale textural features, morphological features, and deep features rather than relying on any single feature.

SHAP analysis clarified the decision basis of the model at the feature level, complementing and cross-validating the pixel-level attention visualization of Grad-CAM.

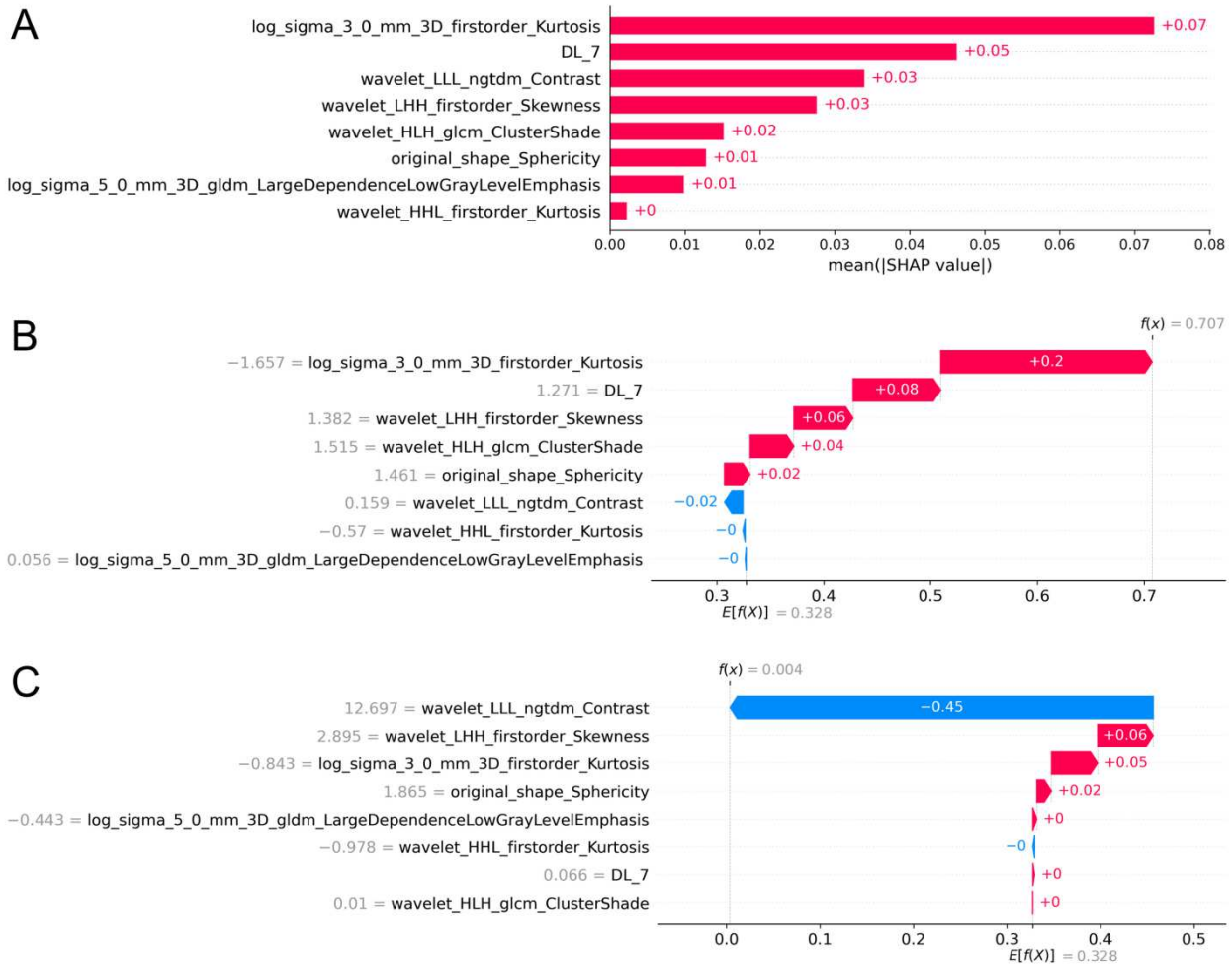


Fig. 8 SHAP-based interpretation of the DLR signature (SVM classifier). (A) Feature importance ranking based on the mean absolute SHAP values of the eight selected features across all 372 patients. (B) Individualized SHAP waterfall plot for a representative case predicted as pCR (f(x) = 0.707; E[f(X)] = 0.328). (C) Individualized SHAP waterfall plot for a representative case predicted as non-pCR (f(x) = 0.004). Red and blue bars represent feature contributions pushing the prediction toward and away from pCR, respectively. f(x), predicted pCR probability for the individual patient; E[f(X)], expected value of the model output over the study population (base value).

Grad-CAM visualization showed (Figure 9) that the high-value areas of the attention heatmap of the deep learning model were mainly concentrated in the regions with the most marked enhancement within the tumor parenchyma and in the peritumoral enhancement rim, largely overlapping with the ROI regions used for radiomic feature extraction. This suggests that the model focused on pCR-related imaging phenotypes, such as tumor enhancement patterns and the peritumoral microenvironment, rather than on background noise or imaging artifacts. This pixel-level visualization corroborated the feature-level explanations of SHAP, jointly supporting the biological plausibility of the model's decisions. Notably, the deep learning attention regions revealed by Grad-CAM were highly consistent with the tumor heterogeneity

regions represented by the key radiomic features screened by SHAP (e.g., LoG first-order kurtosis and NGTDM contrast), indicating biological overlap between the enhancement patterns automatically captured by the deep learning model and the textural information quantified by handcrafted radiomic features. By integrating the above imaging information with independent clinical predictors such as ER, HER-2, and tumor size, the combined model achieved complementarity between imaging phenotype and clinical risk, which also explains, at the mechanistic level, why its discriminative performance in the external test set (AUC = 0.933) was superior to that of each single-modality model.

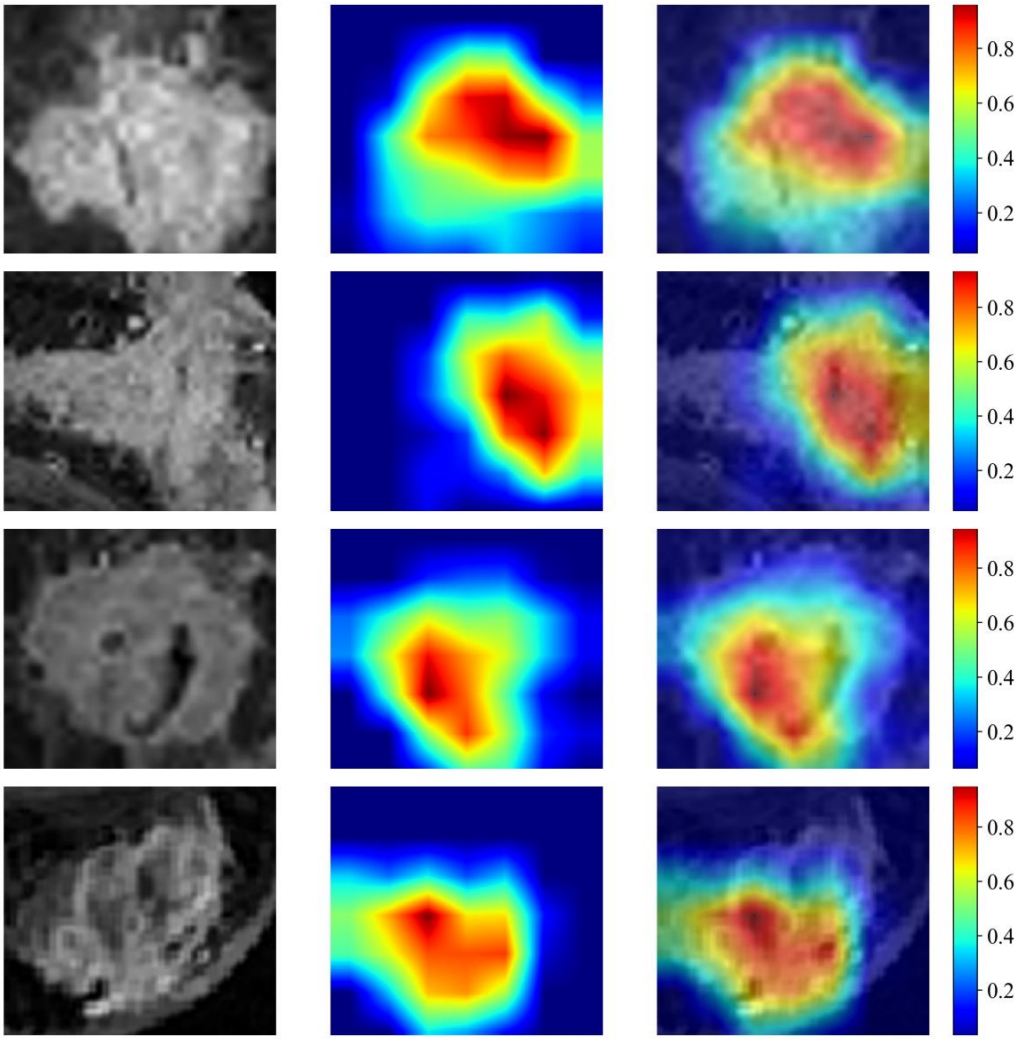


Fig. 9 Grad-CAM visualization of the deep learning model.

2.9 Construction of the nomogram

To facilitate clinical translation, we integrated the independent clinical predictors identified by multivariate analysis (ER status, HER-2 status, and maximum tumor diameter) with the predicted probability of the DLR model, constructed the combined prediction model using

logistic regression, and presented it in the form of a nomogram (Figure 10). Clinicians can sum the scores of each predictor for a patient and directly read the patient's preoperative probability of achieving pCR at the corresponding position on the total-points scale of the nomogram, thereby providing individualized quantitative reference for NAC regimen selection, communication of treatment expectations, and surgical timing decisions.

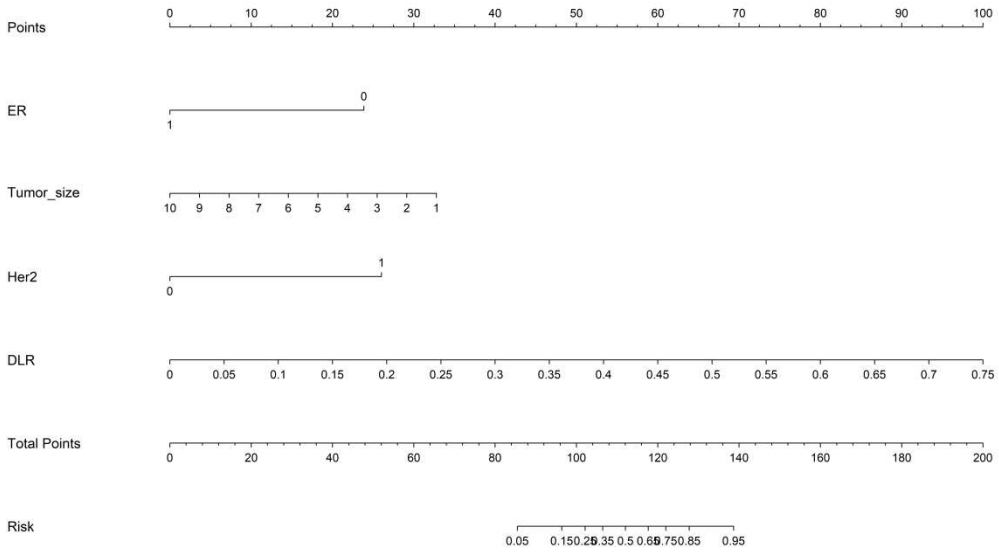


Fig. 10 Nomogram integrating clinical independent predictors and the DLR signature for preoperative prediction of pCR in breast cancer patients.

3 Discussion

Based on the third contrast-enhanced phase of pre-treatment DCE-MRI from 372 patients with breast cancer receiving NAC at two independent centers, this study systematically constructed and head-to-head compared the predictive performance of a clinical model, a conventional radiomics model, a 2D deep learning model, a deep learning radiomics (DLR) model, and a multimodal combined model within a unified technical framework, and completed external validation using independent-center data. The principal findings of this study are as follows: ① the combined model achieved the best overall discriminative performance in the independent external test set, with an AUC of 0.933 (95% CI: 0.879–0.987), and none of the eight pCR patients in the external test set was missed; ② models of different modalities exhibited strikingly different cross-center generalization patterns, with the deep learning model showing the smallest performance degradation (ΔAUC = 0.033) and the

conventional radiomics model the most marked degradation (ΔAUC = 0.278); ③ ER status, HER-2 status, and maximum tumor diameter were independent clinical predictors of pCR; ④ SHAP and Grad-CAM demonstrated, at the feature level and pixel level respectively, that model decisions had a clear tumor biological basis; and ⑤ the integrated discrimination improvement (IDI), net reclassification improvement (NRI), and decision curve analysis (DCA) consistently indicated that the combined model had significant incremental predictive value and net clinical benefit. These results suggest that fusing clinical risk factors, handcrafted radiomic features, and 2D deep features is an effective strategy for achieving accurate and robust preoperative prediction of pCR after NAC in breast cancer.

Regarding the clinical model, multivariate logistic regression in this study showed that ER positivity (adjusted OR = 0.119), HER-2 positivity (adjusted OR = 4.870), and smaller maximum tumor diameter (adjusted OR = 0.742 per cm) were independently associated with pCR, which is highly consistent with the conclusions of previous studies [28-30]. From the perspective of tumor biology, ER-positive Luminal tumors are well differentiated with low proliferation rates and are relatively insensitive to cytotoxic chemotherapeutic agents, whereas ER-negative tumors proliferate actively and are more sensitive to DNA-damaging chemotherapy; HER-2-positive tumors derive marked benefit from anti-HER-2 targeted therapy (e.g., trastuzumab and pertuzumab). Studies such as NeoSphere and KRISTINE have demonstrated that dual-target therapy combined with chemotherapy can raise the pCR rate of HER-2-positive breast cancer to above 40% [31-33]. Maximum tumor diameter reflects the overall tumor burden: the larger the lesion, the higher the probability of residual drug-resistant clones and occult micrometastases, and therefore the lower the likelihood of pCR. The clinical model based on these three indicators already achieved a high AUC of 0.925 in the external test set, indicating that clinicopathological factors are an important cornerstone of pCR prediction. However, its sensitivity was 0.875, and one pCR patient was still missed, suggesting that reliance on clinical factors alone is not sufficient to achieve fully accurate individualized prediction, which also provides a rationale for introducing imaging information.

For the conventional radiomics model, it performed well on the training set (AUC = 0.837) but showed marked degradation on the internal validation set (0.533) and the independent external test set (0.559). This phenomenon is consistent with the widely recognized "distribution shift" problem in the radiomics field [34–36]: although Z-score intensity normalization and 2-mm isotropic resampling were performed before feature extraction, the

two centers used scanners from different vendors (Philips Ingenia and GE Discovery MR750), with differences in TR/TE, FOV, matrix, and signal calibration; handcrafted features based on fixed mathematical definitions (especially LoG and wavelet textural features) are sensitive to subtle changes in gray-level distribution and noise characteristics, making it difficult to directly transfer the feature–outcome relationships learned on the development cohort to the external cohort. Fitting of the SVM classifier to the training distribution may also have aggravated this problem. Most previous single-center MRI-based radiomics studies of breast cancer pCR reported AUCs of 0.70–0.90 but generally lacked truly independent external validation [37–39]; their reported performance is close to our training-set results, which indirectly suggests that internal validation performance may overestimate the true generalizability of a model and further underscores the necessity of multicenter external validation.

In sharp contrast, the 2D deep learning model demonstrated significantly superior cross-center stability. The AUC of MobileNet V2 decreased by only 0.033 from the training set to the external test set (0.858 → 0.825), and it performed best in the internal validation set (0.817). The possible reasons are as follows: CNNs automatically learn hierarchical feature representations from low-level edges and textures to high-level semantic patterns in an end-to-end manner, and their discriminative basis does not depend on any fixed mathematical definition of gray levels, conferring stronger intrinsic robustness to scanner parameter differences; the ImageNet-pretrained transfer learning strategy, combined with real-time data augmentation and cosine annealing learning rate scheduling, effectively alleviated overfitting caused by the limited medical sample size [40,41]. Among the six candidate networks, MobileNet V2, with the smallest parameter scale, performed best; its inverted residual structure and linear bottleneck design preserve representational capacity while limiting redundant parameters, which may make it better suited to the moderate sample size of 372 patients in this study. In addition, using the 2D image at the maximal tumor cross-section as input markedly reduced the sample size and memory requirements compared with 3D models; the maximal cross-section typically contains the most representative morphological and enhancement phenotypes of the tumor, and in small-sample scenarios 2D models can achieve performance comparable to or even more stable than that of 3D models, which is consistent with the findings of recent treatment response prediction studies in other tumors [42-44].

In terms of fusion modeling, we constructed the DLR model by early-fusing MobileNet V2 deep features, after PCA dimensionality reduction, with radiomic features, and then built the

final combined model by late-fusing the DLR-predicted probability with three independent clinical factors. The external test AUC of the DLR model (0.745) was 0.186 higher than that of the conventional radiomics model (0.559), suggesting that the introduction of deep features partially rescued the cross-center failure of handcrafted features; the external AUC of the final combined model further increased to 0.933, indicating substantial information complementarity between imaging phenotypes and clinical risk factors rather than simple performance superposition. Notably, the AUC difference between the combined model and the clinical model did not reach statistical significance on the DeLong test (P = 0.795), which is largely attributable to the limited statistical power imposed by only eight pCR events in the external test set. Nevertheless, the IDI of the combined model relative to the clinical model remained statistically significant (IDI = 0.208, P = 0.028), and sensitivity increased from 0.875 to 1.000, meaning that the practical gain of the combined model was mainly reflected in eliminating missed pCR diagnoses and improving overall discrimination rather than in changes in the AUC value itself. Relative to the radiomics and DLR models, the AUC advantage of the combined model was clearly statistically significant (DeLong P = 0.0007 and 0.014; IDI = 0.482 and 0.397, both P < 0.001). The NRI results (0.216–0.683) further confirmed that the combined model could reclassify a substantial proportion of patients into more accurate risk categories; the Hosmer–Lemeshow test (external test P = 0.069) and calibration curves showed good agreement between its predicted probabilities and actual event rates, and DCA indicated that it could provide a higher net benefit across a wide range of thresholds. jointly supporting the comprehensive advantage of the combined model in the three dimensions of discrimination, calibration, and clinical utility [45,46]. It should be objectively noted that the AUC of the combined model in the internal validation set was only 0.658, which may be related to the small sample size of this cohort (n = 74, with only 16 pCR events) and unstable weight estimation of the logistic regression fusion layer when the number of events is insufficient; its return to high performance in the independent external cohort with a larger sample size (n = 127) and markedly different population composition also indirectly suggests that the small-sample internal validation result was subject to chance.

The interpretability analyses enhanced the clinical credibility of the model. Global SHAP ranking showed that the three features with the highest contributions were LoG-3-mm first-order kurtosis (mean |SHAP| = 0.073), the deep feature DL_7 (0.046), and wavelet-LLL-domain NGTDM contrast (0.034). First-order kurtosis describes the peakedness of the gray-level distribution within the ROI; lower kurtosis reflects a relatively homogeneous

intratumoral enhancement signal and may correspond to tumors of uniform composition with chemotherapy-sensitive clones. NGTDM contrast quantifies the intratumoral heterogeneity represented by neighborhood gray-level differences; higher heterogeneity suggests coexistence of multiple subclones with different biological behaviors and a greater probability of drug-resistant cell populations. The representative case analyses intuitively corroborated this: in the non-pCR case (predicted probability 0.004), NGTDM contrast was approximately 13 standard deviations above the population mean, and the negative contribution of this single feature (SHAP = −0.45) overwhelmed all other features; the pCR case (0.707), in contrast, was pushed toward pCR jointly by low kurtosis, high DL_7, and relatively high sphericity, with a regular spherical morphology often suggesting a well-demarcated tumor with less infiltrative growth. Most critically, the two features with the highest contributions came from handcrafted radiomics and deep learning, respectively, directly demonstrating the complementary value of the two types of information at the feature level. Grad-CAM visualization further showed that the high-value attention areas of the model were concentrated in the regions with the most marked enhancement of the tumor parenchyma and in the peritumoral enhancement rim rather than in background or artifact regions; the peritumoral region harbors information on tumor angiogenesis and microenvironmental invasion. and previous studies have likewise confirmed that peritumoral imaging features are associated with treatment response and prognosis [47,48]. The pixel-level attention distribution was highly consistent with the tumor heterogeneity regions indicated by feature-level SHAP, forming a chain of mutually validating evidence and showing that the high performance of the model did not arise from incidental fitting of imaging confounders [49,50].

In terms of clinical translation, we presented the combined model in the form of a nomogram. After obtaining biopsy pathology (ER and HER-2 status) and preoperative MRI (maximum tumor diameter, radiomic features, and deep features), clinicians can noninvasively calculate an individual patient's pCR probability without additional examinations or increased medical burden. Considering that both the sensitivity and NPV of the combined model in the external test set reached 1.000, its most promising application scenario may be as a preoperative "rule-out" tool: for patients predicted as non-pCR, it could prompt clinicians to consider intensifying the NAC regimen, prolonging the treatment course, or switching strategies in a timely manner, and to cautiously evaluate the feasibility of de-escalated surgery; for patients predicted to have a high probability of pCR, it could provide quantitative support for doctor–patient communication regarding breast-conserving surgery and treatment de-escalation.

However, it should also be noted that, affected by the low pCR prevalence (6.3%) in the external cohort, the positive predictive value (PPV) of the model was only 0.286; that is, a substantial proportion of those predicted as pCR ultimately had residual lesions. Its results should therefore serve as auxiliary evidence for multidisciplinary decision-making rather than as a standalone basis for guiding treatment de-escalation.

This study still has several limitations. First, it was a retrospective design; despite consecutive enrollment and strict inclusion and exclusion criteria, selection bias may still exist, and the conclusions require validation by prospective studies. Second, the external test set included only eight pCR events (6.3%); the small number of events led to wide AUC confidence intervals, low PPV, and insufficient power for some statistical tests. The significant differences in pCR prevalence and baseline clinical characteristics between the two cohorts, while making external validation more stringent, also constitute potential population confounding; further validation in larger multicenter cohorts with higher prevalence is needed in the future. Third, this study used only the 2D image at the maximal tumor cross-section from the third contrast-enhanced phase of DCE-MRI and did not fully utilize whole-tumor three-dimensional information, hemodynamic information from multiphasic dynamic enhancement, or other functional imaging sequences such as DWI/ADC; subsequent multiparametric, 3D, or image-sequence-based models may further improve performance [51]. Fourth, the ROI still relied on manual physician delineation, and this study included only two centers and two scanner models; the compatibility of the model with equipment from more vendors remains to be evaluated. Fifth, this study did not incorporate the specific NAC regimen (e.g., whether HER-2-positive patients received dual-target therapy), dose intensity, or dynamic MRI changes during treatment into the model, although these factors may likewise influence pCR. Future studies may combine multiparametric MRI, interim dynamic imaging, and molecular pathological information in multicenter prospective cohorts to construct and validate more generalizable dynamic prediction models.

In conclusion, within a multicenter framework this study demonstrated that the multimodal combined model integrating independent clinical predictors, conventional radiomic features, and 2D deep features has excellent preoperative predictive performance for pCR after NAC in breast cancer and favorable cross-center generalizability, with an external test AUC of 0.933. SHAP and Grad-CAM analyses ensured the interpretability of the model, while the nomogram facilitated its clinical translation. This tool is expected to provide a noninvasive, quantitative reference for individualized preoperative evaluation of NAC response,

optimization of treatment strategies, and surgical decision-making, and to offer a methodological example for the application of multimodal fusion modeling to tumor treatment response prediction.

Acknowledgments

Not applicable.

Authors' contributions

Hongen Li, Yihui Zeng, and Yan Zhang designed the study. Hongen Li performed the majority of the experiments. Yihui Zeng provided assistance with the experiments and data analysis. Yihui Zeng and Wen Wang wrote the draft of the manuscript and revised it. Jianbin Tan confirmed the authenticity of all the raw data. All authors read and approved the final version of the manuscript.

Funding

No

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to patient privacy and ethical restrictions but are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

The research protocol has been approved by the institutional review boards of both participating centers, and informed consent from participants has been waived in compliance with relevant ethical regulations (Ethics Approval Numbers: 20261015; KYLL-277).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

[1] Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024 May-Jun;74(3):229-263.

[2] Han B, Zheng R, Zeng H, et al. Cancer incidence and mortality in China, 2022. J Natl Cancer Cent. 2024 Feb 2;4(1):47-53. Feng X, Zhang Y, Gao A, et al. Breast cancer trends in China and its comparative analysis with group of 20 countries: Projections for 2022-2040. Cancer. 2026 May 15;132(10):e70379.

[3] Loibl S, André F, Bachelot T, et al. Early breast cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann Oncol. 2024 Feb;35(2):159-182.

[4] Pérez-García JM, Cortés J, Ruiz-Borrego M, et al. 3-year invasive disease-free survival with chemotherapy de-escalation using an 18F-FDG-PET-based, pathological complete response-adapted strategy in HER2-positive early breast cancer (PHERGain): a randomised, open-label, phase 2 trial. Lancet. 2024 Apr 27;403(10437):1649-1659.

[5] Wang F, Wang Y, Xiong B, et al. Neoadjuvant pyrotinib and trastuzumab in HER2-positive breast cancer with no early response (NeoPaTHer): efficacy, safety and biomarker analysis of a prospective, multicentre, response-adapted study. Signal Transduct Target Ther. 2025 Jan 29;10(1):45.

[6] Cortazar P, Zhang L, Untch M, et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. Lancet. 2014 Jul 12;384(9938):164-72.

[7] Pusztai L, Denkert C, O'Shaughnessy J, et al. Event-free survival by residual cancer burden with pembrolizumab in early-stage TNBC: exploratory analysis from KEYNOTE-522. Ann Oncol. 2024 May;35(5):429-436.

[8] Garrigós L, Ruiz-Borrego M, Pérez-García JM, et al. A chemotherapy-free, pathological response-adapted strategy using trastuzumab-pertuzumab and T-DM1 in HER2-positive early breast cancer: the PHERGain-2 study. Ann Oncol. 2026;37(6):837-848.

[9] van der Voort A, Louis FM, van Ramshorst MS, et al. MRI-guided optimisation of neoadjuvant chemotherapy duration in stage II-III HER2-positive breast cancer (TRAIN-3): a multicentre, single-arm, phase 2 study. Lancet Oncol. 2024 May;25(5):603-613.

[10] Conti M, Morciano F, Bufi E, et al. Surgical planning after neoadjuvant treatment in breast cancer: a multimodality imaging-based approach focused on MRI. Cancers (Basel). 2023 Feb 24;15(5):1439.

[11] Ramtohul T, Lepagney V, Bonneau C, et al. Use of pretreatment perfusion MRI-based intratumoral heterogeneity to predict pathologic response of triple-negative breast cancer to neoadjuvant chemoimmunotherapy. Radiology. 2024 Sep;312(3):e240575.

[12] Caracciolo M, Castello A, Urso L, et al. Comparison of MRI vs. [18F]FDG PET/CT for treatment response evaluation of primary breast cancer after neoadjuvant chemotherapy: literature review and future perspectives. J Clin Med. 2023 Aug 17;12(16):5355.

[13] Yılmaz E, Mısır ME, Seker ME, et al. Subtle ground enhancement a new definition for breast magnetic resonance imaging (MRI) after neoadjuvant chemotherapy: is it a marker for microscopic residual disease? Clin Radiol. 2026 Jun;97:107349.

[14] Lambin P, Woodruff HC, Mali SA, et al. Radiomics Quality Score 2.0: towards radiomics readiness levels and clinical translation for personalized medicine. Nat Rev Clin Oncol. 2025 Nov;22(11):831-846.

[15] Whybra P, Zwanenburg A, Andrearczyk V, et al. The Image Biomarker Standardization Initiative: standardized convolutional filters for reproducible radiomics and enhanced clinical insights. Radiology. 2024 Feb;310(2):e231319.

[16] Ye T, Jin S, Xu H, et al. Can radiomics from dynamic contrast-enhanced MRI effectively predict response to neoadjuvant chemotherapy in breast cancer?: A meta-analysis. Clin Radiol. 2026 Jan;92:107085.

[17] Gaudio M, Vatteroni G, De Sanctis R, et al. Incorporating radiomic MRI models for presurgical response assessment in patients with early breast cancer undergoing neoadjuvant systemic therapy: collaborative insights from breast oncologists and radiologists. Crit Rev Oncol Hematol. 2025 Jun;210:104681.

[18] Shao J, Wei M, Li K, et al. Radiomics: current applications and future directions. MedComm (2020). 2026 May 23;7(6):e70773.

[19] Zhao X, Bai JW, Guo Q, et al. Clinical applications of deep learning in breast MRI. Biochim Biophys Acta Rev Cancer. 2023 Mar;1878(2):188864.

[20] Haghighi F, Hosseinzadeh Taher MR, Gotway MB, et al. Self-supervised learning for medical image analysis: discriminative, restorative, or adversarial? Med Image Anal. 2024 May;94:103086.

[21] Wang W, Liang H, Zhang Z, et al. Comparing three-dimensional and two-dimensional deep-learning, radiomics, and fusion models for predicting occult lymph node metastasis in laryngeal squamous cell

carcinoma based on CT imaging: a multicentre, retrospective, diagnostic study. EClinicalMedicine. 2024 Jan 3;67:102385.

[22] Wang T, Li Z, Yu H, et al. Prediction of microvascular invasion in hepatocellular carcinoma based on preoperative Gd-EOB-DTPA-enhanced MRI: comparison of predictive performance among 2D, 2D-expansion and 3D deep learning models. Front Oncol. 2023 Feb 3;13:987781.

[23] Mo Z, Yang M, Zhu Z, et al. Multiomic integration reveals subtype-specific predictors of neoadjuvant treatment response in breast cancer. Sci Adv. 2025 Jul 4;11(27):eadu1521.

[24] Harbeck N, Modi S, Pusztai L, et al. DESTINY-Breast11 Trial Investigators. Neoadjuvant trastuzumab deruxtecan alone or followed by paclitaxel, trastuzumab, and pertuzumab for high-risk HER2-positive early breast cancer (DESTINY-Breast11): a randomised, open-label, multicentre, phase III trial. Ann Oncol. 2026 Feb;37(2):166-179.

[25] Burstein HJ, Curigliano G, Gnant M, et al. Gallen International Breast Cancer Consensus 2025. Electronic address: https://twitter.com/SG_BCC; Panelists of the St. Gallen International Breast Cancer Consensus 2025. Tailoring treatment to cancer risk and patient preference: the 2025 St Gallen International Breast Cancer Consensus Statement on individualizing therapy for patients with early breast cancer. Ann Oncol. 2025 Dec;36(12):1433-1446.

[26] Rakha EA, Tan PH, Van Bockstal MR, et al. International expert consensus recommendations for HER2 reporting in breast cancer: focus on HER2-low and ultralow categories. Mod Pathol. 2026 Jan;39(1):100925.

[27] Goldhirsch A, Winer EP, Coates AS, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann Oncol. 2013 Sep;24(9):2206-23.

[28] Bayram S, Tatli AM, Cakir MO, et al. Predictors of pathologic complete response and its prognostic value in early breast cancer: a real-world cohort study. Curr Oncol. 2025 Oct 28;32(11):603.

[29] Perry LM, Sevilimedu V, Polidorio N, et al. Predictors of pathologic complete response with neoadjuvant chemo-immunotherapy in early-stage triple-negative breast cancer. Ann Surg Oncol. 2025 Jun;32(6):3991-4001.

[30] Imamura S, Mori K, Yasojima H, et al. Genomically determined subtypes and clinicopathological features as predictors of the efficacy of preoperative chemotherapy combined with HER2-targeted therapy for early-stage HER2-positive breast cancer. BMC Cancer. 2025 Nov 4;25(1):1707.

[31] Gianni L, Pienkowski T, Im YH, et al. Efficacy and safety of neoadjuvant pertuzumab and trastuzumab in women with locally advanced, inflammatory, or early HER2-positive breast cancer (NeoSphere): a randomised multicentre, open-label, phase 2 trial. Lancet Oncol. 2012 Jan;13(1):25-32.

[32] Hurvitz SA, Martin M, Symmans WF, et al. Neoadjuvant trastuzumab, pertuzumab, and chemotherapy versus trastuzumab emtansine plus pertuzumab in patients with HER2-positive breast cancer (KRISTINE): a randomised, open-label, multicentre, phase 3 trial. Lancet Oncol. 2018 Jan;19(1):115-126.

[33] Yuan Y, Liu X, Xu G, et al. Comparative efficacy and safety of pyrotinib plus trastuzumab versus trastuzumab plus pertuzumab and trastuzumab monotherapy in neoadjuvant treatment of HER2-positive breast cancer: a systematic review and meta-analysis. Cancer Treat Rev. 2025;134:102901.

[34] Koch LM, Baumgartner CF, Berens P. Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study. NPJ Digit Med. 2024 May 9;7(1):120.

[35] Wennmann M, Rotkopf LT, Bauer F, et al. Reproducible radiomics features from multi-MRI-scanner test-retest-study: influence on performance and generalizability of models. J Magn Reson Imaging. 2025 Feb;61(2):676-686.

[36] Guevorguian P, Chinnery T, Lang P, et al. External validation of a CT-based radiomics signature in oropharyngeal cancer: assessing sources of variation. Radiother Oncol. 2023 Jan;178:109434.

[37] Zeng Q, Xiong F, Liu L, et al. Radiomics based on DCE-MRI for predicting response to neoadjuvant therapy in breast cancer. Acad Radiol. 2023 Sep;30 Suppl 2:S38-S49. doi: 10.1016/j.acra.2023.04.009. Epub 2023 May 9.

[38] Hwang KP, Elshafeey NA, Kotrotsou A, et al. A radiomics model based on synthetic MRI acquisition for predicting neoadjuvant systemic treatment response in triple-negative breast cancer. Radiol Imaging Cancer. 2023 Jul;5(4):e230009.

[39] Bi C, Chen A, Ran F, et al. Pretreatment MRI radiomics for predicting pathological Miller-Payne grading in breast cancer following neoadjuvant chemotherapy. Cancer Imaging. 2026 Jan 16;26(1):25.

[40] Theodoris CV, Xiao L, Chopra A, et al. Transfer learning enables predictions in network biology. Nature. 2023 Jun;618(7965):616-624. doi: 10.1038/s41586-023-06139-9. Epub 2023 May 31.

[41] Zhao Y, Li C, Shu C, et al. Tackling small sample survival analysis via transfer learning: a study of colorectal cancer prognosis. Artif Intell Med. 2026 Aug;178:103426.

[42] Zhu J, Xu B, Fan T, et al. Sub-regional radiomics combining multichannel 2-dimensional or 3-dimensional deep learning for predicting neoadjuvant chemo-immunotherapy response in esophageal squamous cell carcinoma: a multicenter study. NPJ Precis Oncol. 2025 Jul 21;9(1):248.

[43] Shi K, Chen C, Fei Y, et al. Imaging-based transformer model predicts early therapy response in advanced nasopharyngeal carcinoma: a dual-center study. Insights Imaging. 2025 Dec 2;16(1):267.

[44] Yang Y, Mu X, Liu L, et al. Prediction of neoadjuvant chemotherapy efficacy for locally advanced nasopharyngeal carcinoma using MRI-based deep learning features combined with vision transformer. Acad Radiol. 2026 May;33(5):2222-2235.

[45] Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024 Apr 16;385:e078378.

[46] Vickers AJ, Van Claster B, Wynants L, et al. Decision curve analysis: confidence intervals and hypothesis testing for net benefit. Diagn Progn Res. 2023 Jun 6;7(1):11.

[47] Fürböck C, Janickova I, Langs G, et al. Deep-learning with domain-specific pretraining for breast cancer neoadjuvant chemotherapy response prediction from pre-treatment B-mode ultrasound. Cancers (Basel). 2026 Apr 23;18(9):1345.

[48] Wang Y, Zhang Y, Liu Z, et al. Multiparametric MRI-based habitat analysis integrating deep learning and radiomics for predicting preoperative Ki-67 expression level in breast cancer. BMC Med Imaging. 2026 Jan 16;26(1):80.

[49] Yang Y, Zhang H, Gichoya JW, et al. The limits of fair medical imaging AI in real-world generalization. Nat Med. 2024 Oct;30(10):2838-2848.

[50] Dawood M, Branson K, Tejpar S, et al. Confounding factors and biases abound when predicting molecular biomarkers from histological images. Nat Biomed Eng. 2026 Mar 2.

[51] Huang YH, Shi ZY, Zhu T, et al. Longitudinal MRI-driven multi-modality approach for predicting pathological complete response and B cell infiltration in breast cancer. Adv Sci (Weinh). 2025 Mar;12(12):e2413702.