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.

Development and External Validation of an Interpretable Nomogram Integrating Deep Learning and Clinicopathological Signatures on DCE-MRI for Predicting Pathological Complete Response to Neoadjuvant Therapy in Breast Cancer

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 cancer, Neoadjuvant therapy, Pathological complete response, Dynamic contrast-enhanced magnetic resonance imaging, Deep learning, 2.5D input, Multimodal fusion, Nomogram

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11474277/v1

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

Additional Declarations: No competing interests reported.

Development and External Validation of an Interpretable Nomogram Integrating Deep Learning and Clinicopathological Signatures on DCE-MRI for Predicting Pathological Complete Response to Neoadjuvant Therapy in Breast Cancer

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 **Objective:** To systematically compare the predictive performance of two-dimensional (2D) and 2.5-dimensional (2.5D) deep learning input strategies based on dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) for predicting pathological complete response (pCR) after neoadjuvant therapy (NAT) in breast cancer under a fixed backbone network, and to evaluate the cross-center generalizability of a multimodal model integrating clinicopathological information with deep learning signatures. **Methods:** A total of 524 patients with breast cancer who received NAT at two centers were retrospectively enrolled. The 345 patients from center one were divided into a training set (n = 241) and an internal validation set (n = 104) at a 7:3 ratio, and 179 patients from center two served as an independent external test set. Pre-treatment DCE-MRI was used to construct 2D (largest cross-section) and 2.5D (central slice ±2, five slices as multi-channel input) deep learning signatures. Nine convolutional neural network architectures were systematically compared, and the optimal backbone (DenseNet161) was selected. Univariable and multivariable logistic regression analyses of 11 clinicopathological and MRI semantic features were performed to identify independent predictors of pCR, and a clinical signature was constructed using extreme gradient boosting (XGBoost). The independent predictors were then fused with the 2D and 2.5D deep learning signatures via logistic regression to build a combined signature, which was visualized as a nomogram. The receiver operating characteristic (ROC) curve, DeLong test, calibration curve, Hosmer-Lemeshow test, and decision curve analysis were used to evaluate the discrimination, calibration, and clinical net benefit of each signature. Gradient-weighted class activation mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP) were applied to interpret the deep learning and clinical signatures at the pixel and feature levels, respectively. **Results:** Multivariable analysis identified axillary lymph node metastasis status (OR = 0.250, 95% CI: 0.127–0.492, P < 0.001),

human epidermal growth factor receptor 2 (HER-2) status (OR = 3.516, 95% CI: 1.590–7.777, P = 0.002), and Ki-67 expression (OR = 4.035, 95% CI: 1.019–15.979, P = 0.047) as independent predictors of pCR. The 2.5D signature achieved higher areas under the curve (AUCs) than the 2D signature in the training, internal validation, and external test sets (0.758 vs 0.701, 0.551 vs 0.534, and 0.597 vs 0.526, respectively), although the differences were not statistically significant (all P > 0.05). The combined signature yielded the highest discrimination in the training set (AUC = 0.865, 95% CI: 0.819–0.910), significantly outperformed either single deep learning signature across all three cohorts (all P < 0.01), and achieved discrimination comparable to the clinical signature in the external test set (0.726 vs 0.711, P = 0.665). Decision curve analysis showed that both the combined and clinical signatures provided clinical net benefit over a wide range of threshold probabilities. Stratified analysis indicated that HER-2 positivity was the dominant driver of pCR, whereas the incremental predictive value of Ki-67 was mainly observed in the HER-2-negative population (pCR rate, 18.3% vs 7.3%). **Conclusion:** With a fixed backbone network, the 2.5D multi-channel input strategy yields consistent but limited performance gains over the 2D input strategy. The combined signature integrating three independent clinical predictors with deep learning signatures demonstrates robust cross-center discrimination and a wide range of clinical net benefit, and may serve as an auxiliary tool for individualized treatment decision-making before NAT in breast cancer.

Key words Breast cancer; Neoadjuvant therapy; Pathological complete response; Dynamic contrast-enhanced magnetic resonance imaging; Deep learning; 2.5D input; Multimodal fusion; Nomogram

Introduction

Breast cancer is the most common malignancy in women. According to GLOBOCAN 2022, approximately 2.3 million new cases and 670,000 deaths from breast cancer occur worldwide each year, ranking among the leading causes of cancer incidence and death in women[1]. For locally advanced breast cancer or early-stage disease with high-risk features, neoadjuvant therapy (NAT) has become one of the standard treatment strategies; it not only increases the success rates of radical surgery and breast-conserving surgery through downstaging, but also serves as an "in vivo drug sensitivity test" that provides a basis for subsequent adjustment of adjuvant regimens[2,3]. Pathological complete response (pCR) refers to the absence of residual invasive cancer in the breast primary site and of regional lymph node metastasis after NAT (ypT0/is ypN0, with residual ductal carcinoma in situ permitted) and is now widely used as a surrogate endpoint of NAT efficacy[4]. Large pooled analyses have confirmed that patients who achieve pCR have significantly improved event-free survival and overall survival[5,6]. However, pCR rates vary dramatically across molecular subtypes: only about 10%–15% in the hormone receptor–positive/HER-2–negative (HR+/HER-2−)

subtype, whereas triple-negative breast cancer (TNBC) and HER-2–positive subtypes can reach 30%–60%[7,8]. Therefore, if patients more likely to achieve pCR could be accurately identified before treatment begins, this would provide an objective basis for treatment escalation or de-escalation decisions.

At present, the clinical anticipation of pCR relies mainly on pre-treatment clinicopathological indicators, such as axillary lymph node status, Ki-67 expression level, maximum tumor diameter, and molecular subtype[9,10]. These indicators are readily accessible and cost-effective, yet they have two unavoidable shortcomings. First, Ki-67 interpretation is susceptible to inter-observer variability and to the spatial and temporal heterogeneity of tumors; an international multicenter reproducibility study showed that, even after standardized training, observers from different centers may still reach markedly discordant interpretations of the same specimen[11]. Second, core needle biopsy is an invasive sampling procedure, and the tissue obtained represents only a localized region of the tumor, offering quite limited coverage of intratumoral heterogeneity[12]. A noninvasive, reproducible assessment tool that can reflect the biological behavior of the tumor as a whole is therefore urgently needed in clinical practice.

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) can reflect tumor angiogenesis and perfusion characteristics, providing an important window for the noninvasive assessment of tumor biology[13,14]. Studies on pCR prediction based on DCE-MRI radiomics and deep learning have emerged continuously in recent years. Braman et al. were among the first to propose a DCE-MRI radiomics model combining intratumoral and peritumoral regions, confirming the independent predictive value of peritumoral information for treatment response[15]; Cain et al. built multivariable machine learning models on this basis and tested them in an independent validation set[16]. With the introduction of deep learning (DL), models can automatically mine high-dimensional image features that are difficult for the human eye to perceive: Du et al. used convolutional neural networks to predict tumor response from breast MRI, demonstrating potential superiority over conventional methods[17]; Joo et al. fused high-dimensional MRI features with clinical information to construct a multimodal deep learning model in 536 patients[18]. Over the past two years, related research has advanced further in sample size and validation design: based on the multicenter cohort of the I-SPY2 trial (n=929), Hong et al. integrated intratumoral radiomics, peritumoral (9 mm expansion) features, and deep learning features, and the final model achieved areas under the curve (AUCs) of 0.888 and 0.890 in internal validation and an independent external cohort, respectively[19]. Taken together, these studies indicate that fusion of multimodal imaging information is an effective route to improving pCR prediction.

Nevertheless, existing studies share several common problems. First, a considerable proportion of them were modeled at a single center, or—although "external validation" was nominally

incorporated—the samples in fact came from different time periods of the same institution, so true cross-center generalizability remains untested[20,21]. Second, many studies directly apply three-dimensional convolutional neural networks to voxel data; when the sample size is limited, 3D networks have large parameter counts and are prone to overfitting, typically manifesting as excellent training-set performance that drops sharply in the external validation set—a study in the field of hepatocellular carcinoma observed that the AUC of a 3D model fell from 0.751 in the training set and 0.666 in the internal validation set to 0.567 in the external validation set[22]. Third, no unified paradigm has yet been established for the organic integration of imaging and clinicopathological information, and most studies remain at the level of feature concatenation[23]. How to balance the exploitation of three-dimensional spatial information against model generalizability under limited sample conditions therefore remains an urgent problem to be solved.

In recent years, an input strategy termed "2.5D" has attracted increasing attention in oncologic imaging. Its core idea is to take the slice with the largest tumor cross-section as the center, sample several adjacent slices toward the head and foot sides, and feed these slices as multi-channel input into a two-dimensional convolutional network[24]. The advantage of this strategy is that it preserves three-dimensional spatial context while allowing direct reuse of two-dimensional network weights pretrained on natural images, thereby markedly reducing the risk of overfitting when the sample size is limited[25,26]. The strategy has been validated in multiple tumor types: in pathological grading of clear cell renal cell carcinoma[27], differential diagnosis of parotid tumors[28], and prediction of response to neoadjuvant chemoimmunotherapy in esophageal squamous cell carcinoma[29], comparisons of 2D versus 2.5D have all suggested a certain gain from multi-channel input. In breast MRI, Xiang et al. have also attempted to use a multi-branch 2.5D encoder to extract sequence-specific features for pCR prediction[30]. To our knowledge, however, in the specific setting of pCR prediction after NAT for breast cancer, no study has yet performed a head-to-head comparison of 2D and 2.5D input strategies under the same backbone network and the same data split, accompanied by independent multicenter external validation.

Against this background, we retrospectively enrolled 524 patients with breast cancer who received NAT at two centers, with the following objectives: (1) under a fixed backbone network (DenseNet161), to systematically compare the performance of 2D and 2.5D input strategies for predicting pCR and to clarify whether multi-channel fusion yields a tangible gain; (2) to construct and validate a clinical signature based on independent clinical predictors, as well as 2D and 2.5D deep learning signatures; (3) to fuse clinical and imaging information via logistic regression to build a combined signature and to depict it as a nomogram, thereby enabling preoperative individualized prediction; and (4) in the internal validation set and an independent external test set, to comprehensively evaluate the discrimination, calibration, and clinical utility of each signature using

ROC curves, calibration curves, the DeLong test, and decision curve analysis, and to improve model interpretability at both the pixel and feature levels by means of gradient-weighted class activation mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP).

1 Materials and Methods

1.1 Study Population

This was a retrospective, multicenter study conducted in accordance with the principles of the Declaration of Helsinki and approved by the ethics committees of Guangdong Women and Children Hospital (approval no. 20261015) and Binzhou Medical University Hospital (approval no. KYLL-277). Because anonymized data were used, the requirement for informed consent was waived with the approval of the ethics committees.

We retrospectively collected data from patients with breast cancer who underwent breast DCE-MRI examination before neoadjuvant therapy (NAT) at two independent centers between February 2025 and September 2026. The inclusion criteria were: (1) invasive breast cancer confirmed by core needle biopsy, receipt of standard neoadjuvant therapy, and postoperative pathology allowing definitive assessment of pathological complete response (pCR) status; (2) breast DCE-MRI completed within 2 weeks before the start of neoadjuvant therapy; and (3) complete clinical, pathological, and imaging data. The exclusion criteria were: (1) any other antitumor treatment—such as surgery, radiotherapy, endocrine therapy, or targeted therapy—received before neoadjuvant therapy; (2) poor DCE-MRI image quality with obvious artifacts that precluded accurate lesion delineation; (3) concomitant other primary malignancies; and (4) absence of evaluable preoperative MRI sequences.

A total of 524 female patients with breast cancer were finally enrolled, aged 27–77 years, with a mean age of 50.26±9.74 years. A multicenter external validation strategy was adopted to assess the generalizability of the model: center one (n=345) constituted the model development set, which was stratified by pCR status and randomly divided at a 7:3 ratio into a training set (n=241) and an internal validation set (n=104); center two (n=179) served as an independent external test set used only for the final independent evaluation of the model. The patient selection flowchart is shown in Fig. 1.

The pathological definition of pCR followed the 8th edition of the American Joint Committee on Cancer (AJCC) breast cancer staging system[31]: no residual invasive cancer in either the breast primary site or the regional lymph nodes in the surgical specimen after neoadjuvant therapy (ypT0/is ypN0), with residual ductal carcinoma in situ (DCIS) permitted; any case with a residual invasive cancer component was classified as non-pCR.

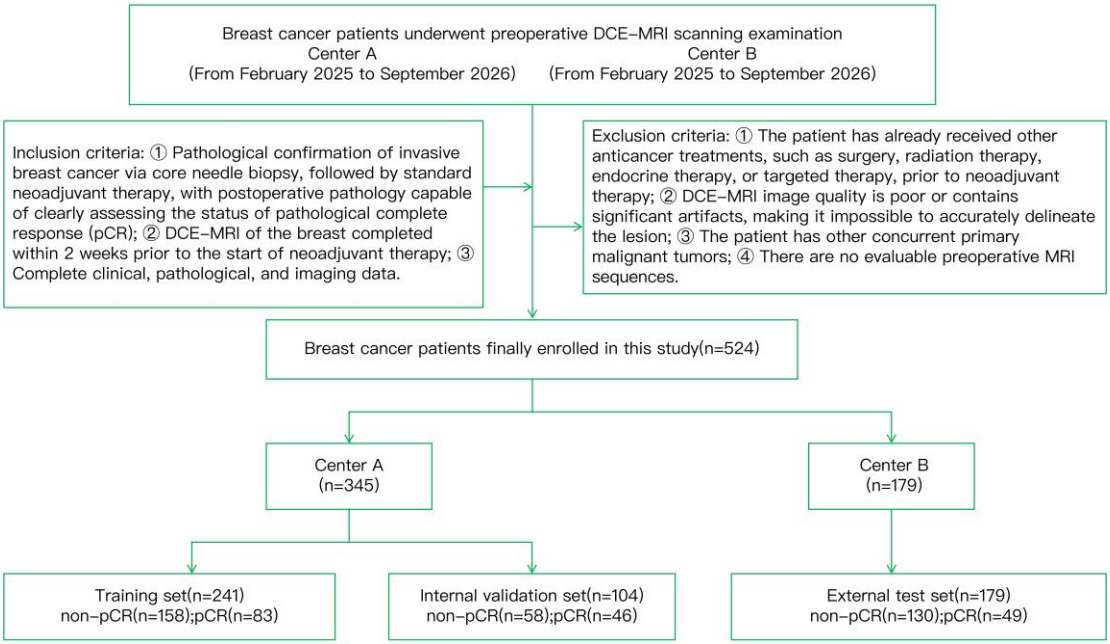


Fig. 1 Flowchart of the study subjects based on inclusion and exclusion criteria.

1.2 Image Acquisition

DCE-MRI examinations were performed on two 3.0T MR scanners, both equipped with a dedicated 16-channel breast phased-array coil. Center one used a Philips Ingenia 3.0T scanner (Philips Healthcare, the Netherlands) with repetition time/echo time (TR/TE) = 4.4/2.2 ms, a slice thickness of 1.6 mm, a slice gap of −0.8 mm, and six dynamic phases of 60 s each; center two used a GE Discovery MR750 3.0T scanner (GE Healthcare, USA) with TR/TE = 4.1/1.7 ms, a slice thickness of 1.6 mm, and a slice gap of 0 mm. The contrast agent gadoterate meglumine was injected through the median cubital vein at a dose of 0.2 mmol/kg body weight and a flow rate of 2.0 ml/s, followed by a 15 ml saline flush after the injection.

1.3 Clinicopathological and MRI Semantic Feature Assessment

All MRI images were independently evaluated by two radiologists with more than 10 years of experience in breast MRI diagnosis, both of whom were unaware of the pathological results; in case of disagreement, a senior radiologist with more than 20 years of experience made the final judgment. The evaluation followed the 5th edition of the Breast Imaging Reporting and Data System (BI-RADS) of the American College of Radiology (ACR)[32]. The recorded MRI semantic features included MRI-assessed axillary lymph node metastasis status (positive/negative) and maximum tumor diameter.

Clinicopathological features were extracted from the electronic medical record system, including age, menopausal status, age at menarche, family history of breast cancer, and immunohistochemical markers. The immunohistochemical markers included estrogen receptor (ER), progesterone receptor

(PR), human epidermal growth factor receptor-2 (HER-2), and the proliferation index Ki-67, all based on the pathology report of the pre-NAT core needle biopsy. The criteria for ER and PR were as follows[33]: positive nuclear staining in <1% of tumor cells was considered negative, and ≥1% was considered positive. The criteria for HER-2 were as follows[34]: an immunohistochemistry score of "−" or "+" was judged as negative and "+++" as positive; cases scored "++" underwent further fluorescence in situ hybridization (FISH) testing, with gene amplification judged as positive and no amplification as negative. The criterion for Ki-67 was as follows[35]: ≥14% was considered high expression and <14% low expression. According to the consensus of the 2013 St. Gallen International Breast Cancer Conference[35], breast cancers were classified into four molecular subtypes: Luminal A, Luminal B, HER-2-enriched, and triple-negative. A total of 11 clinicopathological and MRI semantic features were ultimately included in the subsequent analyses.

1.4 Image Segmentation and Region of Interest Definition

Manual segmentation was performed on the third phase of DCE-MRI images using ITK-SNAP software (version 3.8.0). Two radiologists independently delineated the tumor region of interest (ROI) slice by slice, avoiding areas of necrosis, hemorrhage, and cystic change; in case of disagreement, a senior radiologist made the final judgment. The ROI required for 2D deep learning was obtained by cropping the minimum bounding rectangle of the largest tumor slice.

1.5 Deep Learning Workflow

The deep learning workflow is shown in Fig. 2 and comprises five steps: image segmentation, construction of 2D and 2.5D data, deep learning model training, signature construction, and model evaluation.

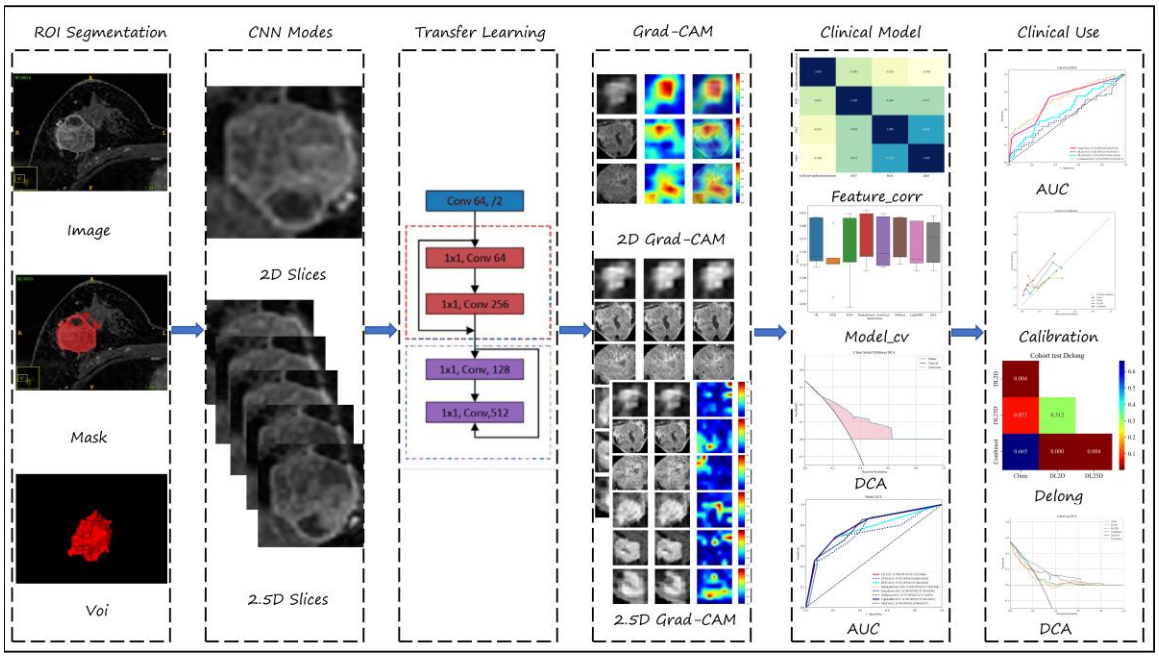


Fig. 2 Flowchart of the deep learning workflow.

1.5.1 Construction of 2D and 2.5D Data

2D data construction: for each patient, the slice with the largest ROI area was selected as the representative image; only the minimum bounding rectangle enclosing the ROI was retained, and it was expanded outward by 10 pixels on all four sides to incorporate peritumoral information, thereby preserving key contextual information of the peritumoral region while reducing background noise and algorithmic complexity.

2.5D data construction (multi-channel image fusion): taking the slice with the largest ROI area as the central slice, adjacent slices at ±1 and ±2 toward the head and foot sides were selected, yielding five two-dimensional images per patient that were stacked as a five-channel input to simulate three-dimensional structural information, i.e., 2.5-dimensional (2.5D) imaging. This method can simultaneously capture the cross-section of the primary tumor and features of the peritumoral microenvironment, compensating for the drawback that using only the largest cross-section may miss key contextual information, and thus achieves more comprehensive spatial information coverage.

1.5.2 Deep Learning Model Training

A transfer learning strategy was adopted in this study, and all models were initialized with weights pretrained on the ImageNet dataset to enhance their adaptability to heterogeneous multicenter data. To identify the network architecture best suited to this study, nine mainstream convolutional neural networks (CNNs) were systematically compared, including ResNet50, ResNet101, ResNet152,

DenseNet121, DenseNet161, DenseNet201, VGG11, VGG16, and VGG19, each trained and compared under both 2D and 2.5D inputs.

The deep learning training pipeline mainly comprised three components: data preprocessing, data augmentation, and training parameter settings. Z-score normalization was applied to standardize the intensity distribution of image channels, and the gray values of image slices were normalized to the interval [−1, 1] through min–max transformation; all cropped subregion images were resized to 224×224 pixels using nearest-neighbor interpolation to meet the input requirements of the deep learning models; during training, online data augmentation techniques (random cropping and horizontal/vertical flipping) were applied to improve model robustness, whereas test images were only standardized without any augmentation. The learning rate was scheduled with a cosine decay strategy, defined as follows:

$$\eta_t = \eta_{min} + \tfrac{1}{2}(\eta_{max} - \eta_{min})\left(1 + \cos\left(\pi \cdot T_{cur} / T_i\right)\right) \quad (1)$$

where η_min was set to 0, η_max was set to 0.01, T_i denotes the total number of training epochs (set to 64 in this study), and T_cur is the current epoch. The optimizer was stochastic gradient descent (SGD), and the loss function was softmax cross entropy.

1.5.3 Signature Construction

Deep learning signatures: the pCR probability computed by the output layer of the CNN was defined as the deep learning signature; accordingly, a 2D deep learning signature (DL2D) and a 2.5D deep learning signature (DL25D) were constructed.

Clinical signature: univariable logistic regression analysis was performed on all 11 clinicopathological and MRI semantic features, and variables with $P<0.05$ were further entered into multivariable logistic regression analysis to identify independent predictors of pCR (axillary lymph node metastasis status, HER-2, and Ki-67). On this basis, the performance of multiple machine learning classifiers—logistic regression, support vector machine, k-nearest neighbors, random forest, extremely randomized trees, XGBoost, LightGBM, and multilayer perceptron—was compared; with the discrimination of the internal validation set as the selection criterion and both discrimination and stability taken into account, XGBoost was selected to construct the clinical signature.

Combined signature: the independent clinical predictors identified by multivariable logistic regression and the output probabilities of the DL2D and DL25D signatures were fused at the feature level via logistic regression to construct a combined signature (Combined), which was visualized as a nomogram, so as to integrate imaging and clinical information to improve predictive performance.

1.5.4 Model Evaluation Metrics

The receiver operating characteristic (ROC) curve was used to evaluate the discriminative ability of each signature, with the area under the curve (AUC) as the primary metric. The AUC represents the probability that, when a positive sample and a negative sample are drawn at random, the classifier will rank the positive sample ahead of the negative sample according to the computed score; it ranges from 0 to 1, with an AUC of 0.5 indicating no discriminative ability (equivalent to random guessing) and an AUC of 1.0 indicating perfect discrimination. The AUC can be expressed as:

$$AUC = \int_0^1 TPR(x)\, dx \qquad (2)$$

where x is the false positive rate (FPR) and TPR(x) is the true positive rate (TPR) corresponding to a given false positive rate x.

Calibration curves were used to assess the agreement between the predicted probabilities of the model and the actual observations; the Hosmer–Lemeshow (HL) goodness-of-fit test was used to quantify the degree of deviation between the two, with a smaller HL statistic indicating better calibration. The HL test statistic H is calculated as follows:

$$H = \Sigma_{g=1}^{G} (O_g - E_g)^2 / [\, n_g P_g (1 - P_g) \,] \qquad (3)$$

where G is the number of groups into which the predicted probabilities are divided (usually 10 groups by deciles of predicted probability), O_g is the observed number of events in group g, and E_g is the expected number of events in group g, calculated as E_g = n_g × P_g; P_g is the mean predicted probability of group g, and n_g is the number of observations in group g.

Decision curve analysis (DCA) was used to evaluate the clinical utility of the models. Based on the concept of net benefit, DCA compares the clinical benefit of different prediction models at various treatment decision thresholds to select the optimal model, and it can incorporate the preferences of patients and decision-makers. Net benefit is defined as the difference between the number of correctly treated cases (true positives, TP) and the weighted number of incorrectly treated cases (false positives, FP), calculated as follows:

$$Net\ benefit = TP/n - (FP/n) \times [\, p_t / (1 - p_t) \,] \qquad (4)$$

where n is the total sample size, TP and FP are the numbers of true-positive and false-positive cases at the corresponding threshold, p_t is the decision threshold probability, and the weighting term p_t/(1−p_t) corrects for the clinical cost incurred by incorrect treatment.

ROC curves were plotted in the training, internal validation, and external test sets to evaluate the discriminative ability of each signature: the ROC curve takes FPR as the horizontal axis and TPR as the vertical axis, and diagnostic performance is judged by the position of the curve and the AUC, with an AUC closer to 1 indicating better discrimination; differences in AUC among signatures were compared using the DeLong test. In addition, model reliability was evaluated with calibration curves combined with the HL goodness-of-fit test, and the clinical utility of the prediction models and their

potential benefit in clinical decision-making scenarios were assessed with DCA. To further improve model interpretability, gradient-weighted class activation mapping (Grad-CAM) was used to visualize the regions attended by the deep learning models, so as to identify image regions that contributed significantly to model decisions.

1.6 Statistical Analysis

The Shapiro–Wilk test was used to assess the normality of continuous variables. Normally distributed continuous variables with homogeneous variance are expressed as mean ± standard deviation and were compared between groups using the independent-samples t test; continuous variables not following a normal distribution are expressed as median (interquartile range) and were compared between groups using the Mann–Whitney U test. Categorical variables are expressed as frequency (percentage) and were compared between groups using the χ^2 test or Fisher's exact test. Univariable and multivariable logistic regression analyses were used to identify independent predictors of pCR, and differences in AUC among models were compared using the DeLong test. All tests were two-sided, and $P<0.05$ was considered statistically significant. SHapley Additive exPlanations (SHAP) was used to perform feature-level interpretability analysis of the decision mechanism of the clinical signature (XGBoost): SHAP values of each feature were computed across all 524 patients, and the contribution of each feature to the model output was quantified and ranked by the mean absolute SHAP value (mean |SHAP|); a beeswarm plot was used to display the directional relationship between feature values and SHAP values, and waterfall plots of representative cases were drawn to illustrate the individualized decision process, forming a complementary verification with the Grad-CAM pixel-level visualization.

All data analyses were completed on the OnekeyAI platform (version 4.9.1) using Python 3.7.12; statistical analyses were performed with Statsmodels 0.13.2; machine learning (including support vector machine and multilayer perceptron) was implemented with scikit-learn 1.0.2; the deep learning framework was built on PyTorch 1.11.0, with performance optimization via CUDA 11.3.1 and cuDNN 8.2.1; deep learning training used an NVIDIA 4090 GPU and the MONAI 0.8.1 framework. The SHAP interpretability analysis was implemented with SHAP 0.42.1.

2 Results

2.1 Clinicopathological and MRI Semantic Characteristics

A total of 524 female patients with breast cancer were included in this study, aged 27–77 years, with a mean age of 50.26±9.74 years. Among them, 345 were from center one (the model development set, stratified and randomly divided at a 7:3 ratio into a training set of 241 patients and an internal validation set of 104 patients) and 179 were from center two (the independent external test

set). The pCR rates of the three cohorts were 34.4% (83/241) in the training set, 44.2% (46/104) in the internal validation set, and 27.4% (49/179) in the external test set.

The clinicopathological and MRI semantic characteristics of the patients in the three cohorts are shown in Table 1. For continuous variables, the mean age of patients in the external test set was 52.17±9.75 years, higher than 49.71±9.70 years in the training set and 48.23±9.34 years in the internal validation set (P=0.002); age at menarche in the external test, training, and internal validation sets was 12.16±0.85, 12.49±1.36, and 12.27±0.98 years, respectively (P=0.010); and maximum tumor diameter on MRI was 3.33±1.44, 4.21±1.88, and 4.18±1.88 cm, respectively (P<0.001). All of these differences were statistically significant.

For categorical variables, the distributions of menopausal status (P<0.001), ER status (P=0.005), HER-2 status (P<0.001), molecular subtype (P=0.007), and MRI-assessed axillary lymph node metastasis status (P<0.001) differed significantly among the three cohorts. Specifically, the proportion of postmenopausal patients in the external test set was 64.80% (116/179), higher than 46.06% (111/241) in the training set and 40.38% (42/104) in the internal validation set; the ER-positive proportion was 69.27% (124/179), higher than 54.77% (132/241) and 53.85% (56/104), respectively; the HER-2-positive proportion was 29.05% (52/179), lower than 47.72% (115/241) and 51.92% (54/104), respectively; and the proportion of MRI-assessed axillary lymph node metastasis positivity was 72.07% (129/179), significantly higher than 47.72% (115/241) and 34.62% (36/104), respectively. In contrast, the distributions of PR status (P=0.896), family history of breast cancer (P=0.291), and Ki-67 expression (P=0.853) did not differ significantly among the three cohorts.

Table 1. Baseline characteristics of our cohorts.

Feature_name	ALL	Test	Train	Val	P-value
Age	50.26±9.74	52.17±9.75	49.71±9.70	48.23±9.34	0.002
Ageatmenarche	12.34±1.14	12.16±0.85	12.49±1.36	12.27±0.98	0.01
Tumor_size	3.90±1.79	3.33±1.44	4.21±1.88	4.18±1.88	<0.001
Menopausal					<0.001
Premenopausal	255(48.66)	63(35.20)	130(53.94)	62(59.62)	
Post-menopausal	269(51.34)	116(64.80)	111(46.06)	42(40.38)	
Family history					0.291
No	492(93.89)	171(95.53)	222(92.12)	99(95.19)	
Yes	32(6.11)	8(4.47)	19(7.88)	5(4.81)	
ER					0.005
Negative	212(40.46)	55(30.73)	109(45.23)	48(46.15)	
Positive	312(59.54)	124(69.27)	132(54.77)	56(53.85)	
PR					0.896
Negative	277(52.86)	93(51.96)	127(52.70)	57(54.81)	
Positive	247(47.14)	86(48.04)	114(47.30)	47(45.19)	
HER2					<0.001
Negative	303(57.82)	127(70.95)	126(52.28)	50(48.08)	
Positive	221(42.18)	52(29.05)	115(47.72)	54(51.92)	
Ki67					0.853
Low expression	61(11.64)	22(12.29)	26(10.79)	13(12.50)	
High expression	463(88.36)	157(87.71)	215(89.21)	91(87.50)	
Molecular typing					0.007
Luminal A	33(6.30)	8(4.47)	18(7.47)	7(6.73)	
Luminal B	294(56.11)	119(66.48)	122(50.62)	53(50.96)	
Her-2 overexpression	106(20.23)	20(11.17)	60(24.90)	26(25.00)	

Feature_name	ALL	Test	Train	Val	P-value
Triple negative	91(17.37)	32(17.88)	41(17.01)	18(17.31)	
Axillary lymph node metastasis					<0.001
Negative	244(46.56)	50(27.93)	126(52.28)	68(65.38)	
Positive	280(53.44)	129(72.07)	115(47.72)	36(34.62)	

Note: MRI Magnetic Resonance Imaging; ER estrogen receptor; PR progesterone receptor; HER-2 human epidermal growth factor receptor-2.

2.2 Univariable and Multivariable Analyses of Clinical Features

With pCR as the outcome variable, univariable logistic regression analysis was performed on the 11 clinicopathological and MRI semantic features (Table 2). The results showed that six features were significantly associated with pCR (all P<0.05): MRI-assessed axillary lymph node metastasis status (OR=0.174, 95% CI: 0.094–0.321, P<0.001), PR status (OR=0.237, 95% CI: 0.132–0.426, P<0.001), and ER status (OR=0.243, 95% CI: 0.138–0.428, P<0.001) were all significantly negatively associated with pCR, whereas HER-2 status (OR=5.849, 95% CI: 3.226–10.604, P<0.001), Ki-67 expression (OR=3.195, 95% CI: 1.063–9.606, P=0.039), and molecular subtype (OR=1.706, 95% CI: 1.242–2.344, P=0.001) were significantly positively associated with pCR. Age (P=0.567), menopausal status (P=0.545), age at menarche (P=0.691), family history of breast cancer (P=0.209), and maximum tumor diameter (P=0.120) showed no significant association with pCR (all P>0.05).

The six features with P<0.05 in the univariable analysis were entered into multivariable logistic regression analysis (Table 2). The results showed that three features were independent predictors of pCR, namely MRI-assessed axillary lymph node metastasis status (OR=0.250, 95% CI: 0.127–0.492, P<0.001), HER-2 status (OR=3.516, 95% CI: 1.590–7.777, P=0.002), and Ki-67 expression (OR=4.035, 95% CI: 1.019–15.979, P=0.047); ER status (P=0.056), PR status (P=0.187), and molecular subtype (P=0.512) did not retain independent effects in the multivariable analysis (all P>0.05).

Based on these three independent predictors, we further compared the performance of eight machine learning classifiers for constructing the clinical signature. Given the multicenter external validation strategy adopted in this study, model selection was based on the discrimination of the internal validation set (the external test set was used only for the final independent evaluation). The AUCs of the classifiers in the internal validation set were: XGBoost 0.815, random forest 0.809, extremely randomized trees 0.809, logistic regression 0.799, LightGBM 0.798, multilayer perceptron 0.783, k-nearest neighbors 0.764, and support vector machine 0.740. Among them, XGBoost achieved

the highest discrimination and performed robustly across cohorts; it was therefore ultimately selected to construct the clinical signature (Clinic).

Table 2. Univariable and multivariable analysis of clinical and imaging features associated with pathological complete response.

Feature_name	UNI		MULTI	
	OR(95%CI)	P_value	OR(95%CI)	P_value
Axillary lymph node metastasis	0.174(0.094-0.321)	<0.001	0.250(0.127-0.492)	<0.001
PR	0.237(0.132-0.426)	<0.001	0.512(0.189-1.385)	0.187
ER	0.243(0.138-0.428)	<0.001	0.310(0.094-1.030)	0.056
Family history	0.483(0.155-1.504)	0.209		
menopausal	0.848(0.497-1.447)	0.545		
Tumor_size	0.887(0.764-1.031)	0.120		
Age at menarche	0.960(0.784-1.175)	0.691		
age	0.992(0.965-1.020)	0.567		
Molecular typing	1.706(1.242-2.344)	0.001	0.740(0.301-1.821)	0.512
Ki67	3.195(1.063-9.606)	0.039	4.035(1.019-15.979)	0.047
Her2	5.849(3.226-10.604)	<0.001	3.516(1.590-7.777)	0.002

Note: ER estrogen receptor; PR progesterone receptor; HER-2 human epidermal growth factor receptor-2; MRI Magnetic Resonance Imaging; OR odds ratio; CI confidence interval.

2.3 Construction and Performance Comparison of 2D and 2.5D Deep Learning Signatures

To identify the optimal network architecture and input strategy, we constructed deep learning signatures using each of the nine CNN architectures described in Section 1.5.2 under both 2D and 2.5D data construction schemes, and evaluated their performance for predicting pCR in the training, internal validation, and external test sets (Tables 3 and 4).

2D deep learning signatures (Table 3): in the training set, DL2D_resnet152 achieved the highest discrimination (AUC=0.800, 95% CI: 0.7407–0.8595), followed by DL2D_vgg16 (AUC=0.749, 95% CI: 0.6850–0.8134) and DL2D_vgg11 (AUC=0.727, 95% CI: 0.6590–0.7951). In the internal validation set, DL2D_densenet201 achieved the highest discrimination (AUC=0.652, 95% CI: 0.5465–0.7579), with the AUCs of the remaining networks ranging from 0.510 to 0.578. In the external test set, DL2D_vgg11 (AUC=0.576, 95% CI: 0.4779–0.6741) and DL2D_vgg16 (AUC=0.559, 95% CI: 0.4659–0.6528) performed relatively better, whereas DL2D_resnet101,

DL2D_resnet152, and DL2D_densenet201 yielded AUCs of 0.460, 0.479, and 0.465, respectively, all below 0.5.

2.5D deep learning signatures (Table 4): in the training set, DL25D_densenet161 achieved the highest discrimination (AUC=0.758, 95% CI: 0.6936–0.8223), with an accuracy, sensitivity, and specificity of 75.5% (182/241), 60.2% (50/83), and 83.5% (132/158), and positive and negative predictive values of 65.8% and 80.0%, respectively; it was followed by DL25D_densenet201 (AUC=0.638, 95% CI: 0.5647–0.7115) and DL25D_densenet121 (AUC=0.634, 95% CI: 0.5596–0.7087). In the internal validation set, DL25D_resnet50 achieved relatively high discrimination (AUC=0.607, 95% CI: 0.4959–0.7185); in the external test set, DL25D_densenet161 (AUC=0.597, 95% CI: 0.5011–0.6926) and DL25D_resnet101 (AUC=0.566, 95% CI: 0.4685–0.6636) performed better.

Comparison of the AUCs between the training set and the external test set showed that most networks experienced performance degradation of varying degrees in the external test set, suggesting overfitting: DL2D_resnet152 decreased from 0.800 to 0.479 (a difference of 0.321), DL2D_vgg11 from 0.727 to 0.576 (0.151), DL25D_densenet161 from 0.758 to 0.597 (0.161), and DL2D_densenet121 from 0.578 to 0.486 (0.092). In addition, the training-set AUCs of DL2D_resnet50 (0.484), DL2D_vgg19 (0.491), and DL25D_resnet152 (0.483) were all below 0.5, indicating that these networks failed to learn effectively from the training data and are unsuitable as representative deep learning signatures.

To avoid selecting networks with inadequate discrimination and to exclude the confounding influence of network architecture itself on the comparison between 2D and 2.5D, we used the same backbone network, DenseNet161, to construct the representative deep learning signatures under both input strategies, thereby examining the effect of the multi-channel 2.5D input strategy in isolation with the network architecture held fixed. Under 2.5D input, this network ranked first among the nine networks in both the training set and the external test set (AUCs of 0.758 and 0.597); under 2D input, its AUCs in the training, internal validation, and external test sets were 0.701, 0.534, and 0.526, respectively, all above the chance level. Accordingly, a 2D deep learning signature (DL2D) and a 2.5D deep learning signature (DL25D) were constructed; the AUCs of the 2.5D deep learning signature were higher than those of the 2D deep learning signature in all three cohorts (0.758 vs 0.701, 0.551 vs 0.534, and 0.597 vs 0.526).

Table 3. Metrics of 2D deep learning signatures.

Signature	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Cohort
DL2D_resnet50	0.676	0.484	0.4074 - 0.5610	0.06	1	1	0.669	train
DL2D_resnet101	0.535	0.53	0.4529 - 0.6074	0.554	0.525	0.38	0.692	train
DL2D_resnet152	0.73	0.8	0.7407 - 0.8595	0.747	0.722	0.585	0.844	train
DL2D_densenet121	0.664	0.578	0.4997 - 0.6566	0.349	0.829	0.518	0.708	train
DL2D_densenet161	0.647	0.701	0.6317 - 0.7712	0.711	0.614	0.492	0.802	train
DL2D_densenet201	0.581	0.514	0.4362 - 0.5928	0.398	0.677	0.393	0.682	train
DL2D_vgg11	0.73	0.727	0.6590 - 0.7951	0.554	0.823	0.622	0.778	train
DL2D_vgg16	0.73	0.749	0.6850 - 0.8134	0.651	0.772	0.6	0.808	train
DL2D_vgg19	0.44	0.491	0.4166 - 0.5648	0.88	0.209	0.369	0.767	train
DL2D_resnet50	0.606	0.575	0.4634 - 0.6872	0.674	0.552	0.544	0.681	val
DL2D_resnet101	0.51	0.51	0.3971 - 0.6232	0.913	0.19	0.472	0.733	val
DL2D_resnet152	0.577	0.578	0.4675 - 0.6884	0.761	0.431	0.515	0.694	val
DL2D_densenet121	0.606	0.565	0.4525 - 0.6780	0.674	0.552	0.544	0.681	val
DL2D_densenet161	0.529	0.534	0.4210 - 0.6480	0.739	0.362	0.479	0.636	val
DL2D_densenet201	0.635	0.652	0.5465 - 0.7579	0.587	0.672	0.587	0.672	val
DL2D_vgg11	0.587	0.535	0.4219 - 0.6486	0.63	0.552	0.527	0.653	val
DL2D_vgg16	0.519	0.54	0.4274 - 0.6521	0.891	0.224	0.477	0.722	val
DL2D_vgg19	0.558	0.541	0.4287 - 0.6530	0.587	0.534	0.5	0.62	val
DL2D_resnet50	0.626	0.503	0.4059 - 0.5997	0.449	0.692	0.355	0.769	test
DL2D_resnet101	0.369	0.46	0.3663 - 0.5539	0.918	0.162	0.292	0.84	test
DL2D_resnet152	0.682	0.479	0.3799 - 0.5787	0.204	0.862	0.357	0.742	test
DL2D_densenet121	0.385	0.486	0.3913 - 0.5810	0.959	0.169	0.303	0.917	test
DL2D_densenet161	0.631	0.526	0.4247 - 0.6271	0.408	0.715	0.351	0.762	test
DL2D_densenet201	0.425	0.465	0.3746 - 0.5544	0.776	0.292	0.292	0.776	test
DL2D_vgg11	0.62	0.576	0.4779 - 0.6741	0.49	0.669	0.358	0.777	test

Signature	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Cohort
DL2D_vgg16	0.458	0.559	0.4659 - 0.6528	0.816	0.323	0.312	0.824	test
DL2D_vgg19	0.676	0.522	0.4207 - 0.6235	0.347	0.8	0.395	0.765	test

Table 4. Metrics of 2.5D deep learning signatures.

Signature	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Cohort
DL25D_resnet50	0.51	0.5	0.4225 - 0.5777	0.663	0.43	0.379	0.708	train
DL25D_resnet101	0.568	0.54	0.4630 - 0.6176	0.542	0.582	0.405	0.708	train
DL25D_resnet152	0.432	0.483	0.4073 - 0.5582	0.904	0.184	0.368	0.784	train
DL25D_densenet121	0.627	0.634	0.5596 - 0.7087	0.578	0.652	0.466	0.746	train
DL25D_densenet161	0.755	0.758	0.6936 - 0.8223	0.602	0.835	0.658	0.8	train
DL25D_densenet201	0.639	0.638	0.5647 - 0.7115	0.627	0.646	0.481	0.767	train
DL25D_vgg11	0.481	0.518	0.4410 - 0.5946	0.759	0.335	0.375	0.726	train
DL25D_vgg16	0.585	0.534	0.4572 - 0.6101	0.446	0.658	0.407	0.693	train
DL25D_vgg19	0.498	0.525	0.4478 - 0.6018	0.687	0.399	0.375	0.708	train
DL25D_resnet50	0.625	0.607	0.4959 - 0.7185	0.696	0.569	0.561	0.702	val
DL25D_resnet101	0.606	0.597	0.4871 - 0.7067	0.652	0.569	0.545	0.673	val
DL25D_resnet152	0.587	0.573	0.4625 - 0.6844	0.739	0.466	0.523	0.692	val
DL25D_densenet121	0.538	0.519	0.4064 - 0.6311	0.739	0.379	0.486	0.647	val
DL25D_densenet161	0.558	0.551	0.4376 - 0.6636	0.63	0.5	0.5	0.63	val
DL25D_densenet201	0.596	0.526	0.4108 - 0.6409	0.5	0.672	0.548	0.629	val
DL25D_vgg11	0.577	0.583	0.4730 - 0.6935	0.739	0.448	0.515	0.684	val
DL25D_vgg16	0.567	0.582	0.4714 - 0.6920	0.761	0.414	0.507	0.686	val
DL25D_vgg19	0.587	0.593	0.4839 - 0.7027	0.717	0.483	0.524	0.683	val
DL25D_resnet50	0.503	0.522	0.4265 - 0.6168	0.633	0.454	0.304	0.766	test
DL25D_resnet101	0.682	0.566	0.4685 - 0.6636	0.429	0.777	0.42	0.783	test
DL25D_resnet152	0.581	0.555	0.4586 - 0.6521	0.551	0.592	0.337	0.778	test
DL25D_densenet121	0.637	0.506	0.4117 - 0.6002	0.367	0.738	0.346	0.756	test
DL25D_densenet161	0.665	0.597	0.5011 - 0.6926	0.469	0.738	0.404	0.787	test
DL25D_densenet201	0.654	0.561	0.4674 - 0.6541	0.388	0.754	0.373	0.766	test
DL25D_vgg11	0.335	0.501	0.4053 - 0.5976	0.98	0.092	0.289	0.923	test

Signature	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Cohort
DL25D_vgg16	0.587	0.505	0.4089 - 0.6002	0.469	0.631	0.324	0.759	test
DL25D_vgg19	0.659	0.514	0.4141 - 0.6132	0.347	0.777	0.37	0.759	test

Note: AUC area under the receiver operating characteristic curve; PPV positive predictive value; NPV negative predictive value.

2.4 Grad-CAM Visualization

To further improve the interpretability of the deep learning models, Grad-CAM was used to visualize the regions attended by the models. Fig. 3 and Fig. 4 show the activation maps of the 2D and 2.5D models in pCR(−) and pCR(+) samples, respectively. Under both input strategies, the high-response regions of the models were mainly distributed over the tumor parenchyma and the immediately adjacent peritumoral region, and the extent of activation differed between pCR(+) and pCR(−) samples, suggesting that the models distinguish pCR status mainly on the basis of imaging features in the intratumoral and peritumoral regions.

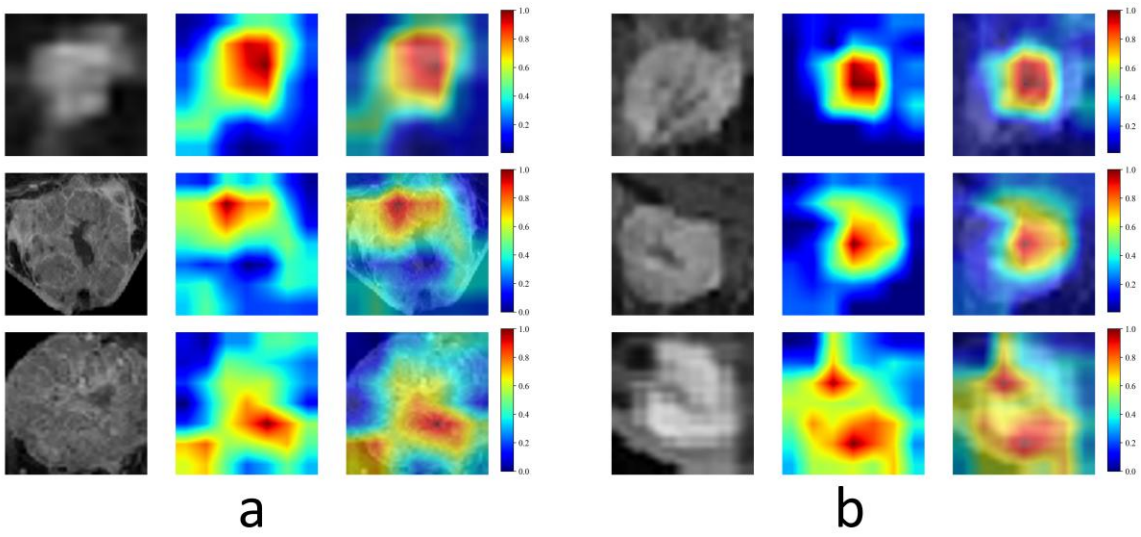


Fig. 3 Grad-CAM visualization for model interpretability (2D). a: activation maps of the DL2D model in pCR (+) samples; b: activation maps in pCR (−) samples.

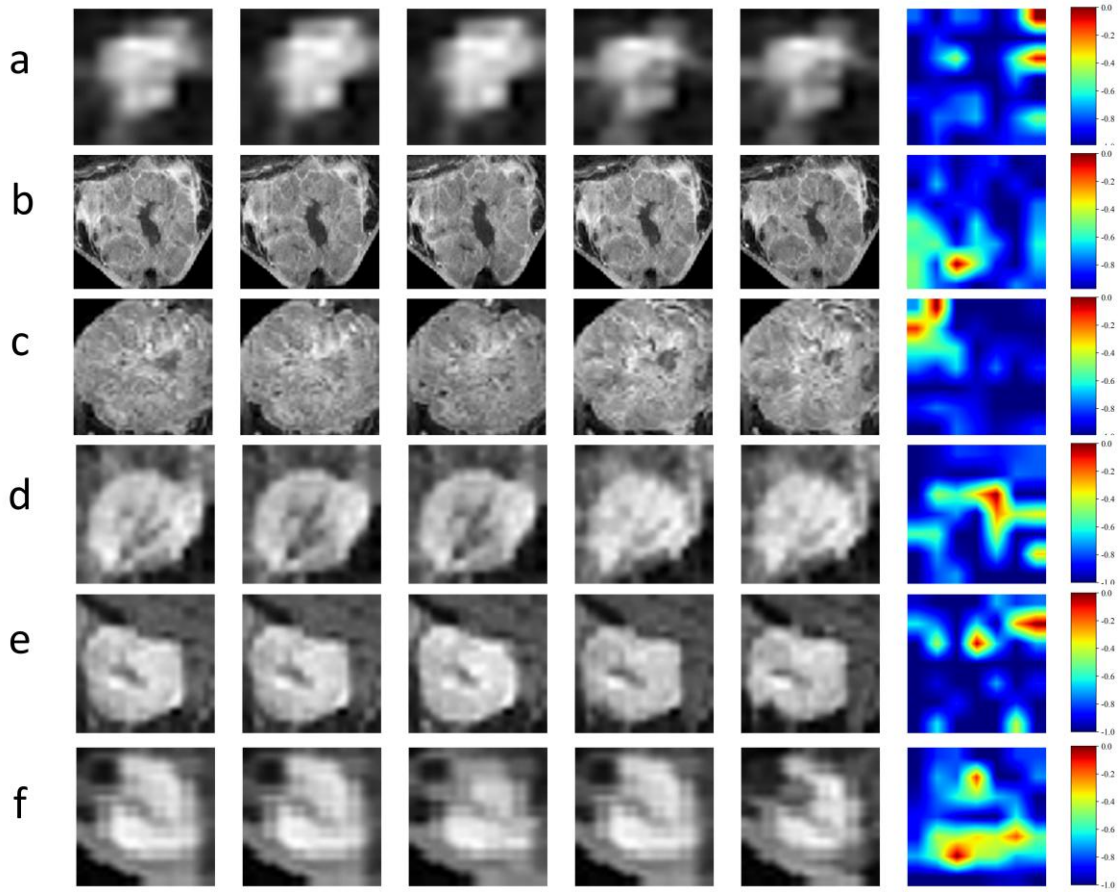


Fig. 4 Grad-CAM visualization for model interpretability (2.5D). a, b, c: activation maps of the DL25D model in pCR (+) samples; d, e, f: activation maps in pCR (−) samples.

2.5 Performance Comparison of Predictive Signatures

To evaluate the gain from fusing multi-source information, we constructed a clinical signature (Clinic), a 2D deep learning signature (DL2D), a 2.5D deep learning signature (DL25D), and a combined signature (Combined), and compared their performance for predicting pCR in the training, internal validation, and external test sets (Table 5 and Fig. 5).

In the training set, the combined signature achieved the highest discrimination (AUC=0.865, 95% CI: 0.8189–0.9104), with an accuracy, sensitivity, and specificity of 73.9% (178/241), 95.2% (79/83), and 62.7% (99/158), and positive and negative predictive values of 57.2% and 96.1%, respectively; it was followed by the clinical signature (AUC=0.785, 95% CI: 0.7294–0.8409), with an accuracy, sensitivity, specificity, positive predictive value, and negative predictive value of 68.9% (166/241), 79.5% (66/83), 63.3% (100/158), 53.2%, and 85.5%, respectively; the 2.5D deep learning signature ranked next (AUC=0.758, 95% CI: 0.6936–0.8223), with corresponding values of 75.5% (182/241), 60.2% (50/83), 83.5% (132/158), 65.8%, and 80.0%; the 2D deep learning signature

showed the lowest discrimination (AUC=0.701, 95% CI: 0.6317–0.7712), with an accuracy of 64.7% (156/241).

In the internal validation set, the clinical signature achieved the highest discrimination (AUC=0.815, 95% CI: 0.7361–0.8940), with an accuracy, sensitivity, specificity, positive predictive value, and negative predictive value of 76.0% (79/104), 67.4% (31/46), 82.8% (48/58), 75.6%, and 76.2%, respectively; the combined signature ranked second (AUC=0.775, 95% CI: 0.6838–0.8672), with corresponding values of 74.0% (77/104), 78.3% (36/46), 70.7% (41/58), 67.9%, and 80.4%; both were markedly higher than the 2.5D deep learning signature (AUC=0.551, 95% CI: 0.4376–0.6636) and the 2D deep learning signature (AUC=0.534, 95% CI: 0.4210–0.6480).

In the external test set, the combined signature achieved the highest discrimination (AUC=0.726, 95% CI: 0.6388–0.8133), with an accuracy, sensitivity, specificity, positive predictive value, and negative predictive value of 67.6% (121/179), 67.3% (33/49), 67.7% (88/130), 44.0%, and 84.6%, respectively; the clinical signature ranked second (AUC=0.711, 95% CI: 0.6261–0.7969), with corresponding values of 67.6% (121/179), 73.5% (36/49), 65.4% (85/130), 44.4%, and 86.7%; the 2.5D deep learning signature (AUC=0.597, 95% CI: 0.5011–0.6926) and the 2D deep learning signature (AUC=0.526, 95% CI: 0.4247–0.6271) both showed lower discrimination than the two signatures above.

Overall, the combined signature achieved the highest discrimination in the training set and significantly outperformed the clinical signature and each single deep learning signature; in the internal validation set and the external test set, it significantly outperformed the 2D and 2.5D deep learning signatures, whereas its difference from the clinical signature was not statistically significant, indicating that after incorporating imaging information the combined signature can achieve discrimination comparable to that of the clinical signature. In addition, the AUCs of the 2.5D deep learning signature were higher than those of the 2D deep learning signature in the training, internal validation, and external test sets.

Table 5. Metrics on different signatures.

Note: AUC area under the receiver operating characteristic curve; PPV positive predictive value; NPV negative predictive value.

Signature	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Cohort
Clinic	0.689	0.785	0.7294 - 0.8409	0.795	0.633	0.532	0.855	train
DL2D	0.647	0.701	0.6317 - 0.7712	0.711	0.614	0.492	0.802	train
DL25D	0.755	0.758	0.6936 - 0.8223	0.602	0.835	0.658	0.8	train
Combined	0.739	0.865	0.8189 - 0.9104	0.952	0.627	0.572	0.961	train
Clinic	0.76	0.815	0.7361 - 0.8940	0.674	0.828	0.756	0.762	val
DL2D	0.529	0.534	0.4210 - 0.6480	0.739	0.362	0.479	0.636	val
DL25D	0.558	0.551	0.4376 - 0.6636	0.63	0.5	0.5	0.63	val
Combined	0.74	0.775	0.6838 - 0.8672	0.783	0.707	0.679	0.804	val
Clinic	0.676	0.711	0.6261 - 0.7969	0.735	0.654	0.444	0.867	test
DL2D	0.631	0.526	0.4247 - 0.6271	0.408	0.715	0.351	0.762	test
DL25D	0.665	0.597	0.5011 - 0.6926	0.469	0.738	0.404	0.787	test
Combined	0.676	0.726	0.6388 - 0.8133	0.673	0.677	0.44	0.846	test

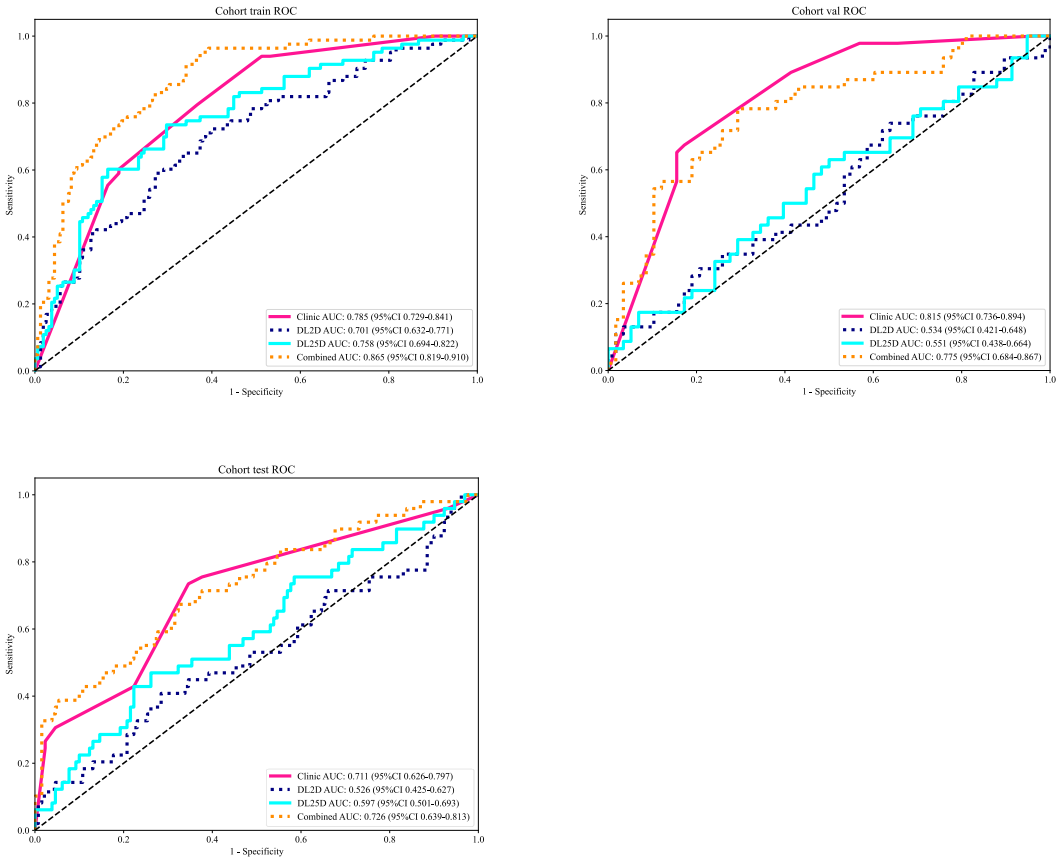


Fig. 5 ROC curves of different signatures in different cohorts.

2.6 Calibration Curve Assessment

Calibration curves combined with the Hosmer–Lemeshow (HL) goodness-of-fit test were used to assess the agreement between the predicted probabilities of each signature and the observed outcomes (Fig. 6). In the training set, the 2D deep learning signature and the clinical signature were well calibrated (HL test P=0.409 and P=0.530, respectively), the 2.5D deep learning signature showed obvious deviation (P=0.001), and the combined signature was at a borderline level (P=0.025). In the internal validation set, the 2D deep learning signature, the 2.5D deep learning signature, and the clinical signature were well calibrated (HL test P=0.641, 0.670, and 0.212, respectively), whereas the combined signature showed relatively obvious deviation (P=0.020). In the external test set, the 2D deep learning signature, the 2.5D deep learning signature, and the clinical signature were all well calibrated (HL test P=0.664, 0.766, and 0.104, respectively), and the combined signature was at a borderline level (P=0.046). These results indicate that the calibration of the clinical signature and the deep learning signatures was generally stable across the three cohorts; although the combined signature had higher discrimination, its probability calibration was relatively poorer, suggesting that

the output probabilities of the linear fusion model should still be interpreted with caution in cross-center application.

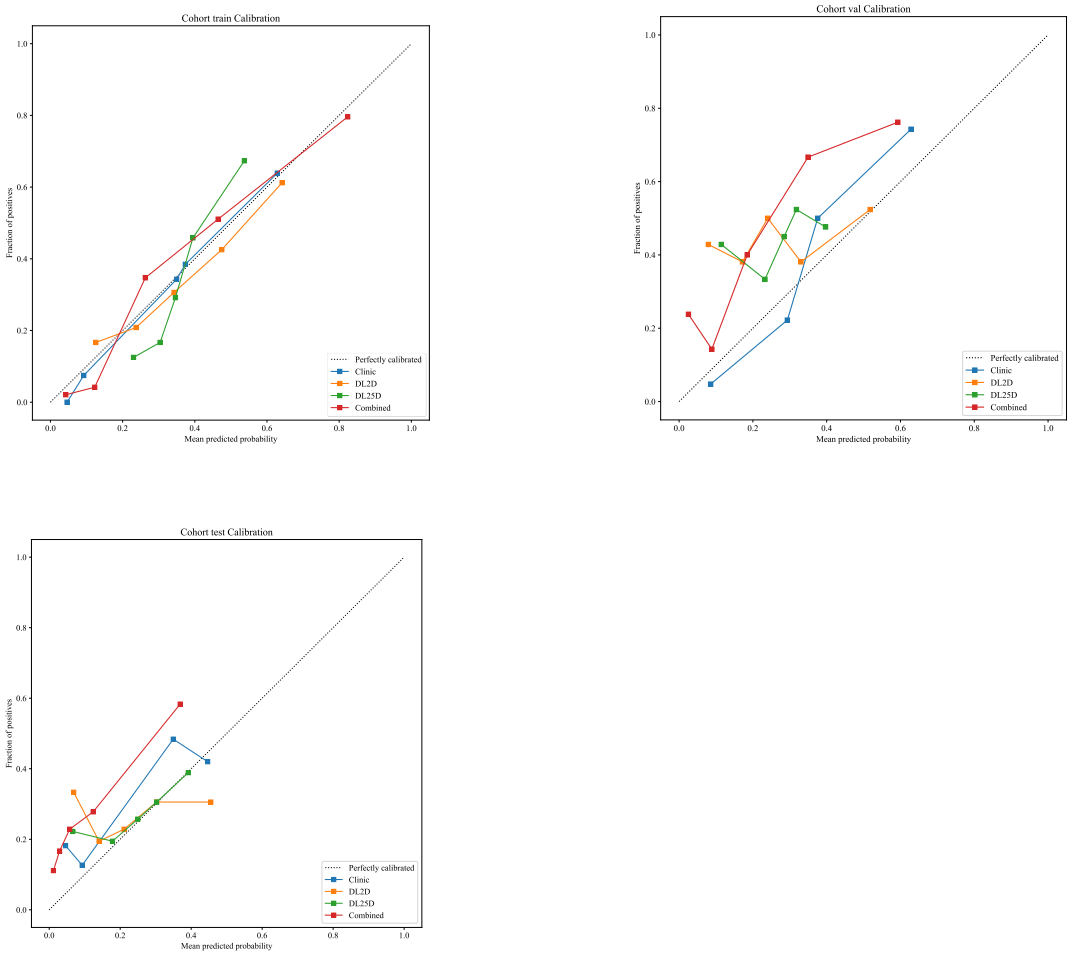


Fig. 6 Calibration curves of different signatures in different cohorts.

2.7 DeLong Test

The DeLong test was used for pairwise comparisons of the AUCs of the signatures (Fig. 7). In the training set, the combined signature significantly outperformed the 2D deep learning signature (ΔAUC=0.164, P<0.001), the 2.5D deep learning signature (ΔAUC=0.107, P<0.001), and the clinical signature (ΔAUC=0.080, P<0.001); the differences between the clinical signature and the 2D and 2.5D deep learning signatures were not statistically significant (ΔAUC=0.084 and 0.027, P=0.066 and P=0.517, respectively), nor was the difference between the 2D and 2.5D deep learning signatures (ΔAUC=0.057, P=0.180).

In the internal validation set, the combined signature significantly outperformed the 2D deep learning signature (ΔAUC=0.241, P<0.001) and the 2.5D deep learning signature (ΔAUC=0.224,

P<0.001); the clinical signature also significantly outperformed the 2D (ΔAUC=0.281, P<0.001) and 2.5D (ΔAUC=0.264, P<0.001) deep learning signatures; the difference between the combined signature and the clinical signature was not statistically significant (ΔAUC=0.040, P=0.320).

In the external test set, the combined signature significantly outperformed the 2D deep learning signature (ΔAUC=0.200, P<0.001) and the 2.5D deep learning signature (ΔAUC=0.129, P=0.004); the clinical signature significantly outperformed the 2D deep learning signature (ΔAUC=0.185, P=0.004), whereas its difference from the 2.5D deep learning signature was not statistically significant (ΔAUC=0.114, P=0.071); the difference between the combined signature and the clinical signature was not statistically significant (ΔAUC=0.015, P=0.665).

In addition, the AUC of the 2.5D deep learning signature was higher than that of the 2D deep learning signature in all three cohorts, but the DeLong test showed that the differences between the two were not statistically significant in any of the three cohorts (P=0.180, 0.806, and 0.312, respectively). These results indicate that the combined signature significantly outperformed each single deep learning signature in all three cohorts and achieved discrimination comparable to that of the clinical signature in the internal validation set and the external test set; although the 2.5D input strategy brought AUC improvements over the 2D input in all three cohorts, none of the differences reached statistical significance.

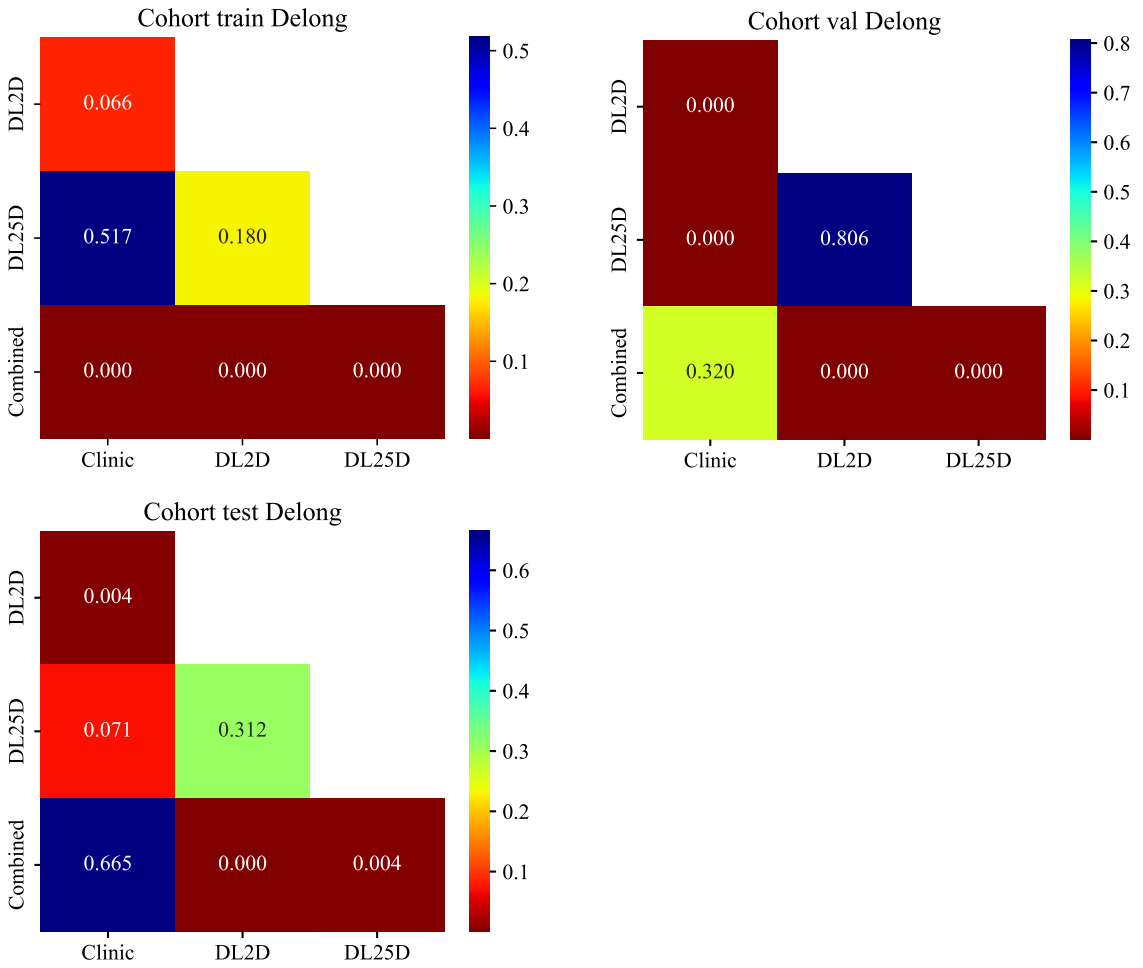


Fig. 7 DeLong test results of pairwise AUC comparisons among signatures.

2.8 Decision Curve Analysis

DCA was used to evaluate the clinical net benefit of each signature at different decision threshold probabilities (Fig. 8). In the training set, the combined signature provided the highest net benefit across threshold probabilities, followed by the clinical signature; taking a threshold probability of 0.3 as an example, the net benefits of the combined, clinical, 2.5D deep learning, and 2D deep learning signatures were 0.201, 0.180, 0.137, and 0.126, respectively. In the internal validation set, the clinical signature provided the highest net benefit (0.297 at a threshold probability of 0.3), followed by the combined signature (0.205), with the 2.5D and 2D deep learning signatures at 0.115 and 0.095, respectively. In the external test set, at a threshold probability of 0.2, the net benefits of the clinical, 2.5D deep learning, combined, and 2D deep learning signatures were 0.138, 0.099, 0.092, and 0.049, respectively; at a threshold probability of 0.4, the clinical signature and the combined signature still maintained positive net benefit (0.061 and 0.060, respectively), whereas the net benefits of the 2D and 2.5D deep learning signatures were close to zero.

These results indicate that both the combined signature and the clinical signature can provide considerable clinical net benefit over a wide range of threshold probabilities, whereas the clinical utility of the single deep learning signatures is limited.

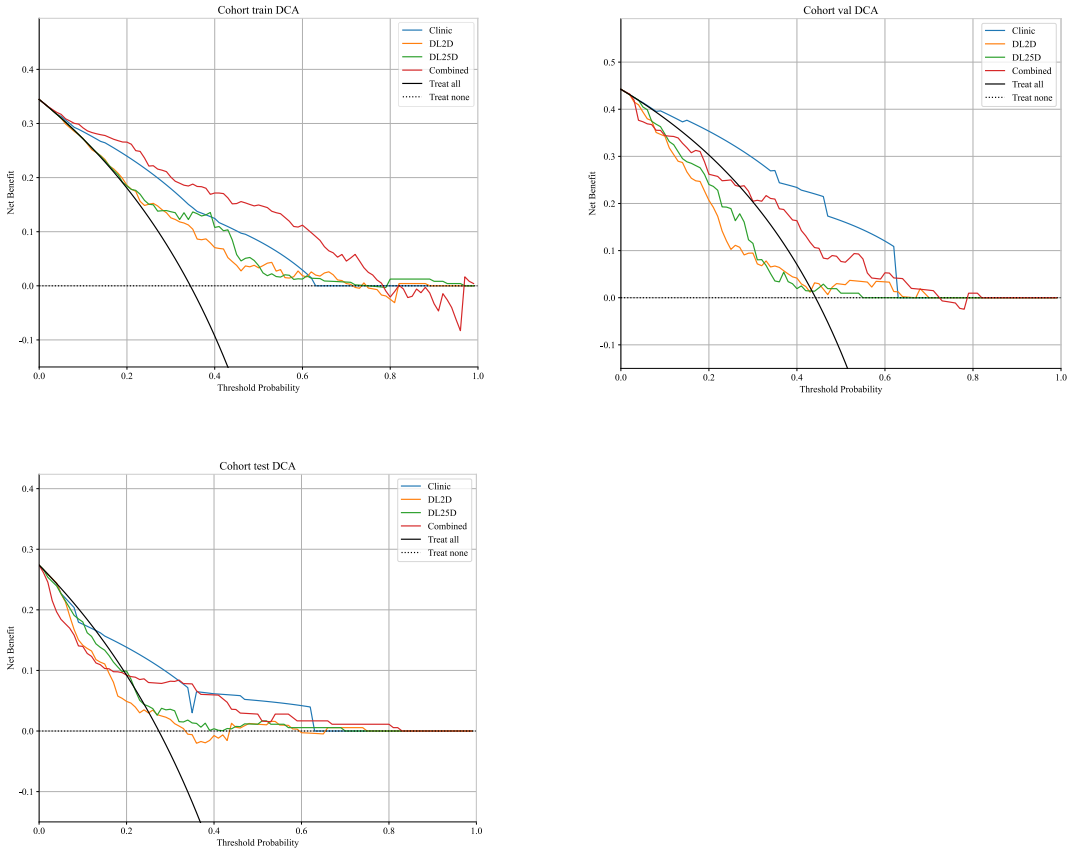


Fig. 8 Decision curve analysis of different signatures in different cohorts.

2.9 Nomogram Construction and Individualized Prediction

To facilitate individualized clinical prediction, we constructed a visualized nomogram for predicting the probability of pCR after neoadjuvant therapy for breast cancer (Fig. 9), based on the three independent pCR predictors identified by multivariable logistic regression (MRI-assessed axillary lymph node metastasis status, HER-2 status, and Ki-67 expression) together with the 2D and 2.5D deep learning signatures. The nomogram integrates all variables into a quantitative scoring system, in which the predicted pCR probability of each patient is converted from the sum of the scores corresponding to each variable, thereby providing an intuitive tool for preoperative individualized risk assessment. The AUCs of the combined signature in the training, internal validation, and external test sets were 0.865, 0.775, and 0.726, respectively, indicating that the nomogram has favorable discrimination and cross-center applicability.

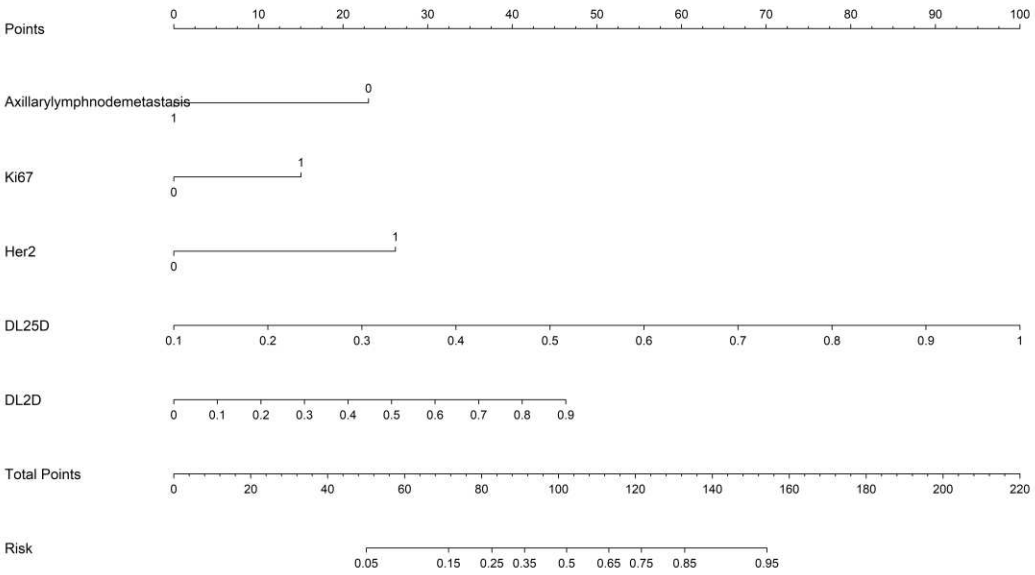


Fig. 9 Nomogram of the integrated model for predicting pathological complete response in breast cancer.

2.10 SHAP-Based Interpretability Analysis of the Clinical Signature

Feature importance ranking based on mean |SHAP| showed that HER-2 status contributed most to the predictions of the clinical signature (mean |SHAP|=0.135), followed by axillary lymph node metastasis status (0.126) and Ki-67 expression (0.020) (Fig. 10A). Feature direction analysis showed that HER-2 positivity, high Ki-67 expression, and negative axillary lymph node metastasis jointly pointed to a higher predicted probability of pCR, consistent with the direction of the ORs obtained from multivariable logistic regression. Waterfall plots of representative cases further illustrated the decision process of the model at the individual level: case one (Fig. 10B) had a predicted pCR probability of 0.629, driven mainly by HER-2 positivity and high Ki-67 expression; case two was a patient from the external test set (Fig. 10C) who, owing to axillary lymph node metastasis, HER-2 negativity, and low Ki-67 expression, had a predicted pCR probability of only 0.046 and was ultimately confirmed as non-pCR on pathology.

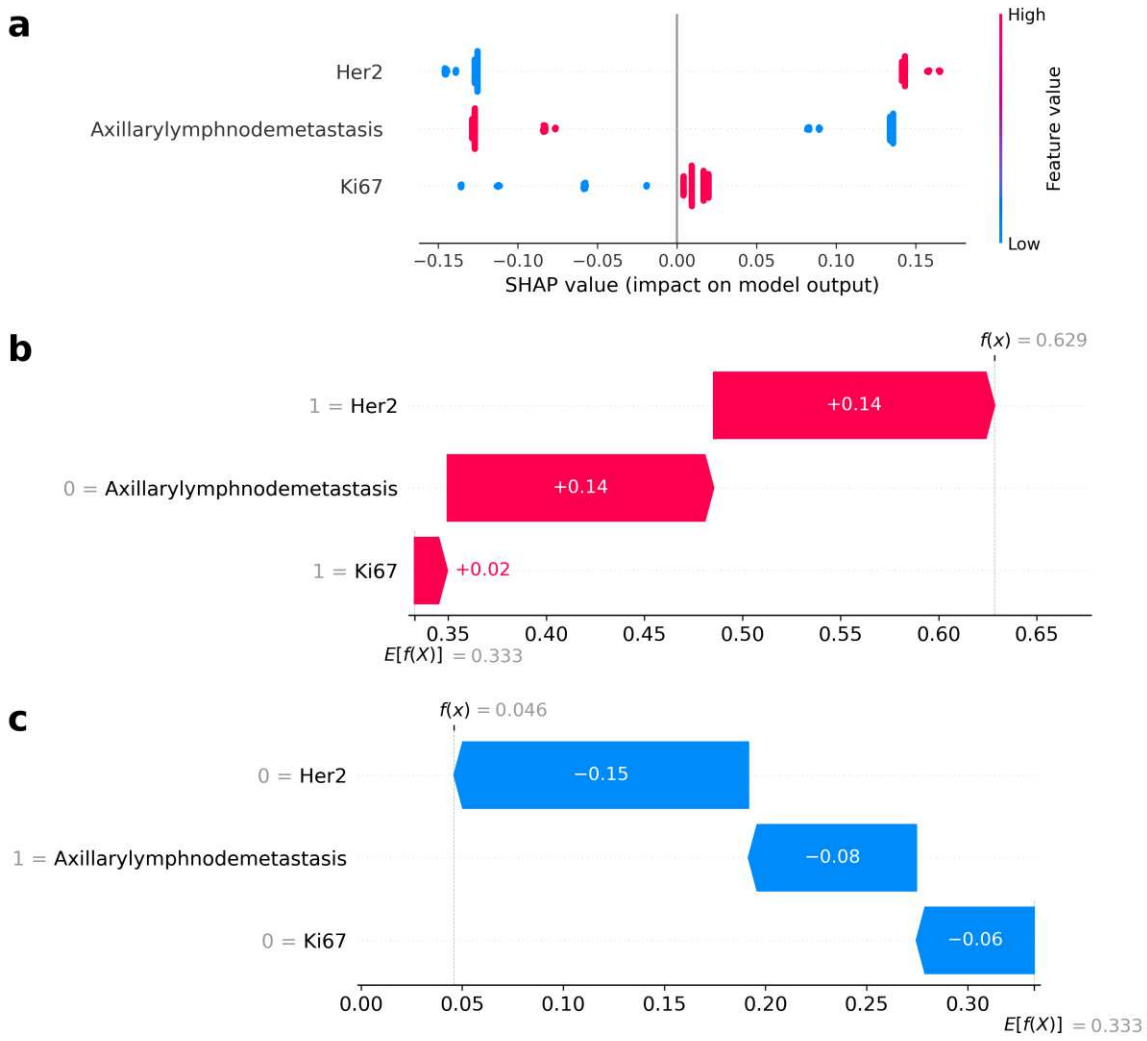


Fig. 10 SHAP-based interpretation of the clinical signature (XGBoost classifier). a: feature importance ranking based on mean absolute SHAP values of the three independent predictors across all 524 patients; b, c: individualized SHAP waterfall plots for representative cases predicted as pCR and non-pCR, respectively. Red and blue bars represent feature contributions pushing the prediction toward and away from pCR.

3 Discussion

In this study of 524 patients with breast cancer receiving NAT at two centers, we performed a head-to-head comparison of 2D and 2.5D input strategies under a fixed backbone network (DenseNet161) and tested the generalizability of the models in an independent external cohort. The main findings can be summarized in three aspects. First, the AUCs of the 2.5D multi-channel input were higher than those of the 2D input in the training, internal validation, and external test sets (0.758 vs 0.701, 0.551 vs 0.534, and 0.597 vs 0.526), with a completely consistent direction, although the

DeLong test showed that none of the differences reached statistical significance. Second, the combined signature—constructed by fusing the three independent clinical predictors (axillary lymph node metastasis status, HER-2, and Ki-67) with the 2D and 2.5D deep learning signatures via logistic regression—achieved the highest discrimination in the training set (AUC=0.865) and significantly outperformed each single deep learning signature in all three cohorts; in the external test set, the AUC of the combined signature (0.726) was comparable to that of the clinical signature (0.711) (ΔAUC=0.015, P=0.665), whereas the clinical signature was better calibrated (HL test P=0.104 vs 0.046). Third, multivariable analysis indicated that axillary lymph node metastasis status was the strongest negative predictor of pCR (OR=0.250, 95% CI: 0.127–0.492), whereas HER-2 status (OR=3.516, 95% CI: 1.590–7.777, P=0.002) and Ki-67 expression (OR=4.035, 95% CI: 1.019–15.979, P=0.047) were independent positive predictors, with a trend toward a negative interaction between Ki-67 and HER-2 (detailed below).

The interaction pattern between Ki-67 and HER-2 is another finding of note in this study. After the entire cohort was divided into four subgroups according to HER-2 status and Ki-67 expression, the pCR rates showed a clear gradient: the HER-2–negative/low-Ki-67 group was the lowest (7.3%, 3/41), the HER-2–negative/high-Ki-67 group was the next (18.3%, 48/262), and the HER-2–positive/low-Ki-67 group (60.0%, 12/20) and the HER-2–positive/high-Ki-67 group (57.2%, 115/201) were markedly higher and almost identical; this gradient remained consistent in the training set (0.0%, 19.4%, 50.0%, and 54.2%, respectively). Stratified analysis showed that among HER-2–negative patients, the pCR rate of those with high Ki-67 expression was about 2.5 times that of those with low expression (18.3% vs 7.3%, P=0.127); among HER-2–positive patients, the level of Ki-67 expression had almost no effect on the pCR rate (57.2% vs 60.0%, P=0.997). The logistic regression product-term test (entire cohort) yielded an interaction-term OR of 0.314 (95% CI: 0.068–1.457, P=0.139; after adjustment for axillary lymph node status, OR=0.316, P=0.153), which, although not statistically significant, indicated a negative interaction on the multiplicative scale (a sub-additive effect); the relative excess risk due to interaction (RERI) estimated on the additive scale was approximately −3.9, in a consistent direction. These results suggest that HER-2 positivity (in the context of standard anti-HER-2 therapy) is the dominant driver of pCR and that its treatment benefit to some extent "masks" the incremental value of Ki-67; the predictive value of Ki-67 is mainly evident in the HER-2–negative population, in which high proliferative activity indicates relative sensitivity to cytotoxic chemotherapy. This pattern is consistent with previous evidence: recent pooled analyses have likewise shown the highest pCR rates in the HER-2–positive subtype[5,36]; Villacampa et al. confirmed that the prognostic value of pCR is subtype-dependent[7,37]; the predictive significance of Ki-67 as a proliferation marker is clearer in the hormone receptor–positive/HER-2–negative population[38,39]; and biomarker analyses from the I-SPY2 platform also support restratifying treatment benefit according to HER-2 and

proliferative features[40]. It is worth noting that the HER-2–negative/low-Ki-67 subgroup (about 8% of the entire cohort) had a pCR rate of only 7.3%; this population biologically corresponds largely to a Luminal A–like phenotype, derives limited benefit from standard neoadjuvant chemotherapy, and may warrant prioritized consideration for treatment escalation, intensified endocrine therapy, or clinical trials of novel agents. From a model-building perspective, although HER-2 positivity and high Ki-67 expression overlap substantially in biology (about 91% of HER-2–positive tumors in this cohort had high Ki-67 expression), both retained independent effects in the multivariable analysis (HER-2: OR=3.516, P=0.002; Ki-67: OR=4.035, P=0.047), indicating that the two carry complementary rather than redundant predictive information; this also provides biological plausibility for the robust cross-center discrimination (internal validation AUC 0.815) achieved by the clinical signature relying on only three variables—axillary lymph node status, HER-2, and Ki-67.

As for the discrimination of the deep learning signatures themselves, the external test set result of this study (AUC 0.597) is lower than that of several recent multicenter studies. Huang et al. included 1339 patients from multiple centers and analyzed longitudinal MRI using a self-supervised temporal learning framework, achieving relatively stable cross-center performance[41]. We speculate that four main reasons account for this gap. First, this study used only single-timepoint images from a single preoperative DCE-MRI sequence, whereas the above studies mostly fused multi-sequence or multi-timepoint information, which provides richer biological information. Second, the sample size was relatively limited (524 patients), making it difficult for deep learning models to adequately learn the complex mapping between imaging and pCR under limited-sample conditions. Third, the external test set of this study came from another center, whose case mix differed markedly from that of the development set, and cross-center transfer itself can cause performance degradation. Fourth, we deliberately did not adopt 3D convolutional networks in order to avoid overfitting under small-sample conditions. Similar declines in external validation performance have also been reported in the field of radiomics: Xie et al. constructed an intratumoral-plus-peritumoral vascular radiomics model in a TNBC population, with an AUC of 0.82 in the primary cohort that decreased to 0.67 in the external validation set[42], suggesting that this phenomenon is not unique to the present study but is closely related to sample size, the dimensionality of imaging information, and cross-center heterogeneity.

The core methodological question of this study was whether the 2.5D multi-channel input can bring stable gains with the network architecture held fixed. The results showed that the AUCs of the 2.5D signature were higher than those of the 2D signature in all three cohorts, with a consistent direction, which is in line with the performance of the 2.5D strategy in other tumor types: the 2.5D multi-instance learning model built by Sun et al. on the basis of venous-phase CT achieved an AUC of 0.820 for predicting vascular invasion in gastric cancer, higher than that of the 3D model (0.711) and the 2D model (0.786)[43]. However, the magnitude of the gain from 2.5D in this study was small (an

AUC increase of 0.071 in the external test set) and did not reach statistical significance on the DeLong test (P=0.312), falling short of the more pronounced advantages reported in other studies. We consider three possible reasons: first, the sample size was small, making it difficult for the performance difference to reach statistical significance; second, only five slices (the central slice ±1 and ±2) were sampled, covering relatively limited spatial context, and the gain might be more pronounced if more slices or three orthogonal planes were used; third, the predictive signal of the images in this cohort was inherently weak (the AUCs of both the 2D and 2.5D signatures clustered in the range of 0.53–0.60), and under low-signal conditions the room for improvement from the input strategy is correspondingly limited.

It should be emphasized that this study used the same backbone network (DenseNet161) under both the 2D and 2.5D strategies, a design of important methodological value. Most previous studies changed the network architecture at the same time when comparing different input strategies, making it difficult to determine whether the performance differences originated from the input strategy or from the network itself[44,45]. By fixing the architecture, the data split, and the training pipeline, this study reduced the comparison to a single variable, thereby making the conclusion that "2.5D is superior to 2D" more reliable. Moreover, DenseNet161 itself performed best among the nine candidate networks (ranking first in both the training set and the external test set under 2.5D input), indicating that this choice was not arbitrarily designated but an objective result of network screening.

It is also worth noting that the introduction of transfer learning alleviated the small-sample problem to some extent: all models were initialized with ImageNet pretrained weights, allowing the 2D networks to leverage the general feature representations learned from natural images, which is particularly practical in medical imaging tasks with limited sample sizes.

The clinical signature achieved discrimination comparable to that of the combined signature in the external test set (AUC 0.711 vs 0.726, ΔAUC=0.015, P=0.665), suggesting that in cross-center scenarios the robustness of standard clinicopathological information is not inferior to that of imaging information. In fact, the absolute performance of the deep learning signatures in this study was not high—especially the 2D signature (external test set AUC of only 0.526)—indicating that relying solely on preoperative DCE-MRI to predict pCR still has obvious limitations. This does not mean, however, that imaging information is of no value: the combined signature significantly outperformed each single deep learning signature in all three cohorts and achieved the highest discrimination in the training set (AUC=0.865); decision curve analysis also showed that the combined signature and the clinical signature provided considerable net benefit over a wide range of threshold probabilities, whereas the clinical utility of the single deep learning signatures was limited[46,47]. It should be pointed out that the absolute probability calibration of the combined signature was at a borderline level in the external test set (HL test P=0.046), whereas the clinical signature was well calibrated

(P=0.104), suggesting that although linear fusion improved discrimination, its output probabilities should still be interpreted with caution in cross-center application, and recalibration strategies such as Platt scaling or isotonic regression could be introduced if necessary[48,49]. From a clinical perspective, the potential significance of the combined model may lie in its "redundancy" value: when the interpretation of indicators such as Ki-67 is subject to inter-observer variability[50], imaging information can provide a portion of independent and reproducible supplementary evidence, thereby buffering, to some extent, the decision risk arising from fluctuation of a single indicator. Of course, this potential value still needs to be confirmed by prospective studies.

The difference in case mix between the external test set and the development set merits separate discussion. The proportion of postmenopausal patients in the external test set (64.80%) was higher than that in the training set (46.06%), the ER-positive proportion (69.27% vs 54.77%) and the MRI-assessed axillary lymph node metastasis–positive proportion (72.07% vs 47.72%) were both higher, and the HER-2–positive proportion (29.05% vs 47.72%) was lower; the pCR rates of the three cohorts also differed markedly (34.4% in the training set, 44.2% in the internal validation set, and 27.4% in the external test set). Such distributional differences (i.e., dataset shift) directly affect the performance of deep learning models: the feature–outcome mapping learned in the development set is no longer fully applicable once the case mix changes, which is one of the important reasons for the performance decline in the external test set. By contrast, the variables on which the clinical signature relies (axillary lymph node status, HER-2, and Ki-67) are themselves well-recognized pCR-related factors and therefore behave more robustly in new populations. This finding suggests that future multicenter studies should describe the baseline differences among cohorts in detail while reporting overall performance, and consider domain adaptation or recalibration strategies to mitigate dataset shift.

Grad-CAM visualization showed that the high-response regions of both the 2D and 2.5D models were mainly concentrated in the tumor parenchyma and its immediately adjacent peritumoral region, which is consistent with observations from other breast cancer imaging studies[51,52] and also indicates, to some extent, that the models did not make judgments by relying solely on segmentation boundaries or background artifacts. The peritumoral region was repeatedly identified as a high-response area, echoing previous studies emphasizing the independent predictive value of the peritumoral microenvironment for treatment response[53,54]. It should be noted, however, that Grad-CAM is a coarse localization method whose heatmaps reflect regions that contribute substantially to the output rather than causal evidence; therefore, the results of this study should not be used to infer causal relationships between specific imaging signs and pCR. Complementarily, the SHAP feature-level analysis confirmed that the decisions of the clinical signature strictly follow the three well-recognized pCR-related factors—axillary lymph node metastasis status, HER-2 status, and Ki-67 expression—and that the contribution direction of each feature is consistent with the logistic

regression ORs. Together, the pixel-level (Grad-CAM) and feature-level (SHAP) lines of evidence indicate that the model decisions have a definite clinico-biological basis rather than representing incidental fitting of confounders.

This study has several limitations. First, it was a retrospective study and selection bias may exist; although consecutive enrollment was adopted and explicit inclusion and exclusion criteria were defined, patient selection was still constrained by the completeness of historical data. Second, the sample size was limited (524 patients) and came from only two centers, and external validation relied on a single independent center; the cross-center generalizability of the model still needs to be verified in more centers and larger samples. Third, only single-timepoint preoperative DCE-MRI was used, and diffusion-weighted imaging, T2-weighted imaging, and longitudinal imaging during treatment were not included; multiple studies suggest that multi-sequence and longitudinal information can significantly improve pCR prediction[55,56]. Fourth, the 2.5D input sampled only five slices (the central slice ±2), covering limited spatial context, and was not compared with other parameter settings (such as ±3 or ±4 slices, or three orthogonal planes). Fifth, tumor segmentation was performed manually, inter- and intra-observer agreement was not systematically evaluated, and the robustness of the model to segmentation variability was not verified. Sixth, pCR was used as the endpoint, and long-term outcomes such as disease-free survival and overall survival were not evaluated; although pCR is a well-recognized surrogate endpoint, it is ultimately not fully equivalent to survival benefit[57]. Seventh, only nine classic CNN architectures were compared, and 3D convolutional networks, Transformers, and more recent self-supervised or foundation model frameworks were not included, so the performance of the multi-channel strategy combined with these newer architectures could not be evaluated. Eighth, this study did not perform a parallel comparison with radiomics models. Ninth, the analysis of the Ki-67–HER-2 interaction was limited by the small sample size of the low-Ki-67 subgroup (only 61 patients in the entire cohort, about 12%), and no pCR events were observed in the HER-2–negative/low-Ki-67 subgroup of the training set, leaving the interaction test underpowered; the related results should be regarded as exploratory findings. These limitations suggest that the results of this study should still be considered exploratory conclusions and that the model still requires prospective, multicenter external validation before entering clinical decision-making pathways; meanwhile, future studies should be designed and reported with reference to guidelines such as TRIPOD+AI to improve the reproducibility and transparency of model research[48].

In summary, this study compared the performance of 2D and 2.5D input strategies for predicting pCR after NAT in breast cancer under a fixed backbone network. The 2.5D multi-channel input brought AUC improvements over the 2D input with a consistent direction in all three cohorts, but the differences did not reach statistical significance, suggesting that under the current sample size and single-sequence imaging conditions the gain of the multi-channel strategy is limited. The combined

signature—built from three independent clinical predictors (axillary lymph node metastasis status, HER-2, and Ki-67) together with the 2D and 2.5D deep learning signatures—performed best in the training set (AUC=0.865) and achieved discrimination comparable to that of the clinical signature in the external test set (0.726 vs 0.711); the clinical signature was better calibrated, whereas the combined signature had a wider range of clinical net benefit in decision curve analysis. The Ki-67–HER-2 interaction analysis indicated that HER-2 status dominates the pCR benefit and that the incremental value of Ki-67 is mainly evident in the HER-2–negative population. The combined model and the nomogram are expected to assist individualized decision-making before NAT, but their clinical value still requires validation in prospective multicenter studies.

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).

Clinical trial number

not applicable

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

[1] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin . 2024;74(3):229-263.

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

[3] Schmid P, Cortes J, Dent R, et al. Overall Survival with Pembrolizumab in Early-Stage Triple-Negative Breast Cancer. N Engl J Med . 2024;391(21):1981-1991.

[4] Litton JK, Regan MM, Pusztai L, et al. Standardized Definitions for Efficacy End Points in Neoadjuvant Breast Cancer Clinical Trials: NeoSTEEP. J Clin Oncol . 2023;41(27):4435-4447.

[5] 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;384(9938):164-172.

[6] Spring LM, Fell G, Arfe A, et al. Pathological complete response after neoadjuvant chemotherapy and impact on breast cancer recurrence and survival: a comprehensive meta-analysis. Clin Cancer Res . 2020;26(12):2838-2848.

[7] von Minckwitz G, Untch M, Blohmer JU, et al. Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. J Clin Oncol . 2012;30(15):1796-1804.

[8] Multiomic integration reveals subtype-specific predictors of neoadjuvant treatment response in breast cancer. Sci Adv . 2024;10(45):eadu1521.

[9] Helal C, Djerroudi L, Ramtohul T, Laas E, Vincent-Salomon A, Jin M, Seban RD, Bieche I, Bello-Roufai D, Bidard FC, Cottu P, Loirat D, Carton M, Lerebours F, Kiavue N, Romano E, Bonneau C, Cabel L. Clinico-pathological factors predicting pathological response in early triple-negative breast cancer. NPJ Breast Cancer. 2025 Feb 13;11(1):15.

[10] Fang S, et al. Interpretable Machine Learning for Predicting Neoadjuvant Chemotherapy Response in Breast Cancer Using the Baseline Clinical and Pathological Characteristics. Cancer Med . 2025;14(17):e71221.

[11] Wang Y, Zou J, Cao Q, et al. Multi-institutional evaluation comparing guidance from International Ki67 Working Group vs National Health Commission of China on immunohistochemistry-based Ki67 assessment alongside the Quantitative Dot Blot method. Front Oncol . 2025;14:1510273.

[12] Xu M, Yan Y, Chen Y, et al. Expression and Subtype Discordance Between Core Needle Biopsy and Surgical Specimen in Breast Cancer. J Surg Res . 2025;307:42-52.

[13] Kataoka M, Honda M, Iima M, et al. Quantitative MR Imaging of the Breast. Magn Reson Med Sci . 2025;24(3):419-429.

[14] Ma W, Yang L, Zhang Y, Gao Y, Jie H, Huang C. Heterogeneity Assessment of Breast Cancer Tumor Microenvironment: Multiparametric Quantitative Analysis with DCE-MRI and Discovery of Radiomics Biomarkers. Breast Cancer (Dove Med Press). 2025 Jul 8;17:573-581.

[15] Braman NM, Etesami M, Prasanna P, et al. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI. Breast Cancer Res . 2017;19(1):57.

[16] Cain EH, Saha A, Harowicz MR, Marks JR, Marcom PK, Mazurowski MA. Multivariate Machine Learning Models for Prediction of Pathologic Response to Neoadjuvant Therapy in Breast Cancer Using MRI Features: A Study Using an Independent Validation Set. Breast Cancer Res Treat . 2019;173(2):455-463.

[17] Du S, Xie W, Gao S, et al. Early prediction of neoadjuvant therapy response in breast cancer using MRI-based neural networks: data from the ACRIN 6698 trial and a prospective Chinese cohort. Breast Cancer Res . 2025;27(1):52.

[18] Joo S, Ko ES, Kwon S, et al. Multimodal deep learning models for the Prediction of Pathologic Response to Neoadjuvant Chemotherapy in Breast Cancer. Sci Rep . 2021;11(1):18800.

[19] Hong D, Peng J, Xu P, Liu W, Liu Z, Ye Z. A Multimodal Fusion Model of Radiomics and Deep Learning Integrating the Tumor Microenvironment Accurately Predicts Pathological Complete Response in Breast Cancer. Acad Radiol . 2026;33(4):1473-1483.

[20] Suleman MU, Mursaleen M, Khalil U, et al. Assessing the generalizability of artificial intelligence in radiology: a systematic review of performance across different clinical settings. Ann Med Surg (Lond) . 2025;87(12):8803-8811.

[21] Zhang J, Wu Q, Lei P, Zhu X, Li B, et al. MRI-based radiomics models for early predicting pathological response to neoadjuvant chemotherapy in triple-negative breast cancer: A systematic review and meta-analysis. J Appl Clin Med Phys . 2025;26(10):e70296.

[22] Zhang YB, Chen ZQ, Bu Y, Lei P, Yang W, et al. Construction of a 2.5D Deep Learning Model for Predicting Early Postoperative Recurrence of Hepatocellular Carcinoma Using Multi-View and Multi-Phase CT Images. J Hepatocell Carcinoma . 2024;11:2223-2239.

[23] Yang H, Yang M, Chen J, Yao G, Zou Q, Jia L. Multimodal deep learning approaches for precision oncology: a comprehensive review. Brief Bioinform . 2025;26(1):bbae699.

[24] Zhang R, Wang K, Wang S, et al. Multimodal deep learning model for prediction of breast cancer recurrence risk and correlation with oncotype DX. Breast Cancer Res . 2025;27:178.

[25] Niu H, Li L, Wang X, Xu H. The value of predicting breast cancer with a DBT 2.5D deep learning model. Discov Oncol . 2025;16:420.

[26] He J, Xu J, Chen W, et al. Development of a deep learning model for T1N0 gastric cancer diagnosis using 2.5D radiomic data in preoperative CT images. npj Precis Oncol . 2025;9(1):249.

[27] Huang T, Ke M, Liu Q, Ying M, Hu M, Fu X, Hu Y, Xu M. A computed tomography-based deep learning model for non-invasively predicting World Health Organization (WHO)/International Society of Urological Pathology (ISUP) pathological grades of clear cell renal cell carcinoma (ccRCC): a multicenter cohort study. Transl Androl Urol. 2025 Jul 30;14(7):2018-2028.

[28] Mai W, Fan X, Zhang L, Li J, Chen L, et al. Deep learning for differential diagnosis of parotid tumors based on 2.5D magnetic resonance imaging. Ann Med . 2025;57(1):2520401.

[29] Chen D, Wang W, Li Q, Rao X, Shi X, Lu D, Diao D, Zhai J, Cai K. Preoperative CT-based Radiomics for Predicting Response to Neoadjuvant Chemoimmunotherapy in Esophageal Squamous Cell Carcinoma. Radiol Imaging Cancer. 2026 Jan;8(1):e250128.

[30] Xiang X, Yin W, Li J. Radiomics-Guided Multi-Sequence Learning for Pathological Complete Response Prediction from Breast MRI with Missing Auxiliary Sequences. J Imaging . 2026;12(6):271.

[31] Giuliano AE, Connolly JL, Edge SB, et al. Breast cancer—major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. CA Cancer J Clin . 2017;67(4):290-303.

[32] D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA, et al. ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System . 5th ed. Reston, VA: American College of Radiology; 2013.

[33] Allison KH, Hammond MEH, Dowsett M, McKernin SE, Carey LA, Fitzgibbons PL, Hayes DF, et al. Estrogen and Progesterone Receptor Testing in Breast Cancer: ASCO/CAP Guideline Update. J Clin Oncol . 2020;38(12):1346-1366.

[34] Wolff AC, Hammond MEH, Allison KH, et al. Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: ASCO – College of American Pathologists Guideline Update. J Clin Oncol . 2023;41(22):3867-3872.

[35] Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thürlimann B, Senn HJ; Panel members. 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;24(9):2206-2223.

[36] Ye Y, Zhang Z, Zhao H, Zhao B. A system review of neoadjuvant immune checkpoint blockade for breast cancer. Front Immunol . 2025;16:1537926.

[37] Villacampa G, Navarro V, Matikas A, Ribeiro JM, Schettini F, Tolosa P, Martínez-Sáez O, Sánchez-Bayona R, Ferrero-Cafiero JM, Salvador F, Papakonstantinou A, Prat A, Oliveira M, Pascual T. Neoadjuvant Immune Checkpoint Inhibitors Plus Chemotherapy in Early Breast Cancer: A Systematic Review and Meta-Analysis. JAMA Oncol . 2024;10(10):1331-1341.

[38] Gómez-Bravo R, et al. Ki67 dynamic predicts endocrine sensitivity in estrogen receptor-positive/HER2-negative breast cancer patients undergoing preoperative endocrine therapy. ESMO Open . 2025;10(11):105845.

[39] Colomer R, González-Farré B, et al. Biomarkers in breast cancer 2024: an updated consensus statement by the Spanish Society of Medical Oncology and the Spanish Society of Pathological Anatomy. Clin Transl Oncol . 2024;26(12):2935-2951.

[40] Khoury K, Meisel J, ... Esserman LJ. Datopotamab-deruxtecan in early stage breast cancer: the I-SPY2.2 sequential multiple assignment randomized phase 2 trial. Nat Med . 2024;30(12):3728-3736.

[41] Huang X, Xu Z, Zhao Y, et al. Longitudinal MRI-based deep learning model for predicting pathological complete response in breast cancer: a multicenter, retrospective cohort study. npj Precis Oncol . 2026;10(1):57.

[42] Xie T, Gong J, Zhao Q, Wu C, Wu S, Peng W, Gu Y. Development and validation of peritumoral vascular and intratumoral radiomics to predict pathologic complete responses to neoadjuvant chemotherapy in patients with triple-negative breast cancer. BMC Med Imaging . 2024;24(1):136.

[43] Sun X, Wang P, Ding R, Ma L, Zhang H, Zhu L.Preoperative prediction value of 2.5D deep learning model based on contrast-enhanced CT for lymphovascular invasion of gastric cancer. Sci Rep . 2025;15(1):25646.

[44] Yang S, Jeong JS, Song D, Han JY, Lim SH, Kim S, et al. Comparison of 2D, 2.5D, and 3D segmentation networks for mandibular canals in CBCT images: a study on public and external datasets. BMC Oral Health . 2025;25(1):1126.

[45] Zuo XY, Liu HF. Multi-scale deep learning models based on MRI for predicting pathological differentiation and evaluating its association with recurrence-free survival in hepatocellular carcinoma: an explainable machine learning study. J Gastrointest Oncol . 2026;17(2):84.

[46] Gao Y, Ventura-Díaz S, ... Mann RM. An explainable longitudinal multi-modal fusion model for predicting neoadjuvant therapy response in women with breast cancer. Nat Commun . 2024;15(1):9613.

[47] Wu T, Liu X, Si Z, Zhang Y, Ma Z, Zhu Y, Gan L. Combination of ultrasound-based radiomics and deep learning with clinical data to predict response in breast cancer patients treated with neoadjuvant chemotherapy. Front Oncol . 2025;15:1525285.

[48] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ . 2024;385:e078378.

[49] Riley RD, Archer L, Snell KIE, Ensor J, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ . 2024;384:e074820

[50] Boyaci C, Sun W, Hartman J, Ács B, et al. Global scoring method of Ki67 immunohistochemistry in breast cancer demonstrates improved concordance using real-world multi-institutional data. Breast Cancer Res . 2025;27:159.

[51] Kim SY, Lee J, Cho N, Kim YG, et al. Deep-learning based discrimination of pathologic complete response using MRI in HER2-positive and triple-negative breast cancer. Sci Rep . 2024;14:23065.

[52] Feng J, Jiang Z, Zhang J. Bidirectional Perceptual Multimodal Interaction Network Based on Contrastive Learning for Breast Cancer pCR Prediction. Tomography . 2026;12(5):74.

[53] Park I, Shin SJ, Go H, Ko J, Bae SJ, et al. Spatial analysis of tumor immune microenvironment of TNBC with different neoadjuvant chemotherapy outcomes using multiplex Immunofluorescence. Breast Cancer Res . 2025;27:181.

[54] Chen W, et al. Multiregional MRI-based deep learning radiomics to predict axillary response after neoadjuvant chemotherapy in breast cancer patients. Insights Imaging . 2026;17(1):21.

[55] Huang YH, Shi ZY, Zhu T, Zhou TH, et al. Longitudinal MRI-Driven Multi-Modality Approach for Predicting Pathological Complete Response and B Cell Infiltration in Breast Cancer. Adv Sci (Weinh) . 2025;12(12):2413702.

[56] Yang Z, He J, Li T, Liu C, Wang Y, et al. A multicenter study of a predictive model for pathological complete response after neoadjuvant therapy in breast cancer using multimodal digital biomarkers. Chin J Cancer Res . 2025;37(6):984-999.

[57] Conforti F, Nekljudova V, Sala I, Ascari R, Solbach C, Untch M, Denkert C, Bagnardi V, Pala L, Fasching PA, Schneeweiss A, Lück HJ, Pagan E, De Pas T, van Mackelenbergh M, Huober J, et al. Surrogate End Points for Overall Survival in Neoadjuvant Randomized Clinical Trials for Early Breast Cancer. J Clin Oncol . 2025;43(12):1441-1452.