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 MRI-Based Multimodal Fusion Model Incorporating Deep Learning and Radiomics for Non-Invasive Prediction of 21-Gene Recurrence Score in Breast Cancer

Lan Wen
The First People's Hospital of Changde
Liwen Peng
Guangdong Province Women and Children Hospital
Zhe Zhang
Peking University Shenzhen Hospital
Zhiqiu Ye
Guangdong Province Women and Children Hospital
Yu Li
Liyu3550@163.com

The First People's Hospital of Changde

Research Article

Keywords: Breast cancer, Dynamic contrast-enhanced MRI, Oncotype DX, Multimodal fusion, Radiogenomics

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11748607/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 MRI-Based Multimodal Fusion Model Incorporating Deep Learning and Radiomics for Non-Invasive Prediction of 21-Gene Recurrence Score in Breast Cancer

Lan Wen[1], Liwen Peng[3], Zhe Zhang[4], Zhiqiu Ye[2*], Yu Li[1*]

1. Department of Radiology, The First People's Hospital of Changde City, No. 818 Renmin Road, 415000, Changde City, Hunan Province, China.

2. Department of Radiology, Guangdong Women and Children Hospital (Women and Children's Hospital, Southern University of Science and Technology), 521 Xingnan Avenue, Panyu District, 511400, Guangzhou, China.

3. Department of Obstetrics, Guangdong Women and Children Hospital (Women and Children's Hospital, Southern University of Science and Technology), 521 Xingnan Avenue, Panyu District, 511400, Guangzhou, China.

4. Department of Radiation Oncology, Peking University Shenzhen Hospital, Shenzhen, China.

*Corresponding authors:

Zhiqiu Ye, Department of Radiology, Guangdong Women and Children Hospital (Women and Children's Hospital, Southern University of Science and Technology), 521 Xingnan Avenue, Panyu District, 511400, Guangzhou, China. E-mail: yezhiqiu0542@126.com

Yu Li, Department of Radiology, The First People's Hospital of Changde City, No. 818 Renmin Road, 415000, Changde City, Hunan Province, China. E-mail: Liyu3550@163.com

E-mail addresses of all authors:

Lan Wen: wenlan7579@163.com; Liwen Peng: 284837423@qq.com; Zhe Zhang: zhangzhejeff@outlook.com; Zhiqiu Ye: yezhiqiu0542@126.com; Yu Li: Liyu3550@163.com

Abstract

Background:

The 21-gene recurrence score (RS) informs adjuvant therapy decisions in selected patients with hormone receptor-positive, HER2-negative early breast cancer. Tissue requirements, costs, and availability can limit testing. We evaluated an MRI-based multimodal fusion model combining clinical variables, handcrafted radiomics, and tri-planar deep-learning representations to predict RS ≥ 18.

Methods:

This retrospective multicohort study evaluated 447 patients with preoperative dynamic contrast-enhanced breast MRI and either measured or transcriptomic-derived 21-gene RS labels. The development cohort combined Peking University Shenzhen Hospital and Duke-Breast-Cancer-MRI cases (n = 324). Development-cohort performance was evaluated using stratified 5-fold cross-validation out-of-fold predictions, with all data-dependent transformations, feature selection, and model fitting restricted strictly within each training fold. TCGA-BRCA cases with RNA-seq-derived RS labels (n = 123) formed the independent external-validation cohort. Four models were evaluated: clinical, radiomics, deep learning, and feature-level fusion using logistic regression. Performance was assessed using AUC with 5,000-replicate bootstrap 95% CIs, threshold-based classification metrics, average precision, Brier score, calibration, and decision-curve analysis. SHAP explained the feature-level fusion model, and Grad-CAM was applied to the deep-learning model.

Results:

The proportion of patients with RS ≥ 18 was 153/324 (47.2%) in the development cohort and 58/123 (47.2%) in the external-validation cohort. External-validation AUCs were 0.669 (95% CI, 0.568–0.763) for the clinical model, 0.616 (0.513–0.711) for radiomics, 0.686 (0.590–0.774) for deep learning, and 0.800 (0.722–0.872) for multimodal fusion. The corresponding average precision values were 0.687, 0.552, 0.686, and 0.777. At the fixed probability threshold of 0.472, fusion achieved 69.1% accuracy, 65.5% sensitivity, 72.3% specificity, 67.9% PPV, and 70.1% NPV. Its Brier score was 0.182. Decision-curve analysis suggested potential net benefit across part of the evaluated threshold range. In the SHAP analysis, a tri-planar deep feature had the highest global importance, with contributions from PR status and radiomic features.

Conclusion:

Multimodal fusion achieved higher observed discrimination than the individual models in an independent cohort with transcriptomic-derived RS labels. This MRI-based framework warrants further evaluation as an adjunct to genomic testing in preoperative workflows.

Keywords: Breast cancer; Dynamic contrast-enhanced MRI; Oncotype DX; Multimodal fusion; Radiogenomics.

Introduction

The 21-gene recurrence-score (RS) assay provides prognostic information and informs adjuvant chemotherapy decisions in selected patients with hormone receptor-positive, human epidermal growth factor receptor 2 (HER2)-negative early breast cancer [1,2]. Clinical evidence from TAILORx supports the use of RS in treatment planning [1]. Tissue requirements, costs, and availability can limit testing in some settings. Imaging-based and clinicopathologic estimates may provide complementary information for patient assessment [3], but require validation before clinical use and should not replace validated genomic assays.

Dynamic contrast-enhanced MRI (DCE-MRI) characterizes breast lesion morphology and contrast enhancement [4,5]. Previous imaging and clinicopathologic studies have examined associations between tumor characteristics and multigene recurrence scores [6-11]. Reproducibility across acquisition settings remains an important challenge for radiomics [12,13]. Handcrafted radiomics quantifies predefined intensity, shape, and texture descriptors using standardized formulations [14,15], whereas convolutional neural networks learn higher-level image representations [16]. Multimodal and MRI-based deep-learning approaches provide additional ways to combine image-derived information [17,18].

Combining clinical variables, radiomic descriptors, and multiplanar deep-learning representations provides a framework for integrating different aspects of tumor characterization [17,19-22]. Recent multiparametric MRI studies have integrated deep learning with habitat-based radiomics to predict Ki-67 expression [23] and combined deep-learning, radiomic, and clinicopathological features to predict recurrence risk [24]. The incremental predictive value and transportability of such multimodal approaches require empirical evaluation [17,21].

In this study, we developed and evaluated a multimodal framework combining clinical variables, handcrafted radiomics, and tri-planar deep-learning representations to predict RS ≥ 18. We compared discrimination, calibration, and decision-curve results with those of the individual models and assessed model explanations.

Materials and Methods

Study Design and Reporting

This retrospective multicohort model-development study followed the principles of the TRIPOD+AI and CLAIM reporting guidelines [25,26]. The study protocol adhered to the Declaration of Helsinki and was approved by the Institutional Review Board of Peking University Shenzhen Hospital (Approval No. PUSH-IRB-20250123). The requirement for written informed consent was waived due to the retrospective nature of the anonymized analytical dataset.

Patients were screened across three cohorts, yielding a final analytical dataset of 447 unique patients. The study population comprised:

- **The In-House Primary Cohort (*n* = 102):** Consecutive patients with invasive breast cancer who underwent preoperative breast MRI and the 21-gene Oncotype DX assay at Peking University Shenzhen Hospital between April 2020 and March 2025.
- **The Duke Public Cohort (*n* = 222):** De-identified imaging and clinical data retrieved from the public Duke-Breast-Cancer-MRI repository on The Cancer Imaging Archive (TCIA).
- **The TCGA-BRCA Transcriptomic Cohort (*n* = 123):** Matched clinical and preoperative MRI datasets from The Cancer Genome Atlas, for which continuous 21-gene Recurrence Scores (RS) were computationally derived using normalized whole-genome RNA-sequencing transcriptomic profiles.

An overview of the multi-cohort integration, three-stream feature extraction architecture, and multimodal feature-fusion workflow is depicted in Figure 1.

Inclusion and Exclusion Criteria

Patients were included if they met the following criteria:

- Histologically confirmed invasive breast carcinoma;
- Documented hormone receptor (ER and/or PR) status and HER2 status confirmed by immunohistochemistry (IHC) or fluorescence in situ hybridization (FISH);
- Preoperative baseline dynamic contrast-enhanced MRI (DCE-MRI) obtained prior to diagnostic core biopsy, surgical intervention, or neoadjuvant therapy; and

- Documented continuous 21-gene Recurrence Score (RS) (via standard Oncotype DX assay or transcriptomic calculation).

Patients were excluded based on the following criteria:

- Missing essential clinicopathological variables (e.g., ER, PR, HER2 status) or non-evaluable genomic/transcriptomic profiles;
- History of minimally invasive procedures or tumor interventions prior to MRI acquisition;
- Inadequate image quality, including severe motion artifacts, missing essential sequences, or acquisition errors;
- Poor lesion depictability, poorly defined margins, or incomplete volumetric coverage extending beyond the field of view.

Outcome Definition and Data Partitioning

The primary objective was to estimate the probability of RS ≥ 18 versus RS < 18. These are study-defined non-low and low RS categories, respectively; RS ≥ 18 should not be interpreted as a universal indication for chemotherapy. RNA-seq-derived RS calculation was used only for TCGA-BRCA. Its weighted calculation combined HER2, estrogen, proliferation, and invasion group scores with CD68, GSTM1, and BAG1 expression; the group-score formulas are provided in the Supplementary Methods. These labels were treated as transcriptomic-derived scores rather than direct commercial Oncotype DX measurements. Label files and analytic records were reconciled before modeling.

The analytic cohort contained 447 unique patients. The development cohort (n = 324) combined the in-house (n = 102) and Duke-Breast-Cancer-MRI (n = 222) cases, while the TCGA-BRCA cohort (n = 123) was held out entirely for independent external validation and was never accessed during model fitting, feature selection, or hyperparameter tuning. Within the development cohort, a stratified 5-fold cross-validation scheme was implemented to generate out-of-fold predictions for all 324 patients, ensuring that each patient was predicted by models fitted exclusively on the remaining folds without data leakage.

Clinical Data and MRI Acquisition

The clinical model incorporated age, estrogen receptor (ER) status, progesterone receptor (PR) status,

HER2 status, and molecular subtype. Molecular subtypes were categorized into luminal A-like, luminal B-like, HER2-enriched, and triple-negative based on immunohistochemical profiles of ER, PR, HER2, and Ki-67.

Preoperative breast MRI was performed in the prone position using dedicated 16-channel phased-array breast coils on 1.5-T and 3.0-T scanners. The imaging protocol included axial T2-weighted imaging (serving as an anatomical reference for lesion boundaries) and dynamic contrast-enhanced (DCE) T1-weighted 3D gradient-echo sequences. DCE-MRI was acquired before and after the intravenous administration of gadolinium (0.1 mmol/kg at 1.5–2.0 mL/s), capturing early, intermediate, and late phases. The DCE-MRI phase exhibiting peak tumor enhancement was selected for radiomic extraction and deep-learning prediction.

Image Preprocessing and Feature Extraction

Signal intensity on peak-phase DCE-MRI was normalized per sequence using the formula $I_{normalized}=[(I-\mu)/\sigma]\times100$. Images were resampled via cubic-spline interpolation to a voxel spacing of $0.9 \times 0.9 \times 2.2$ mm³. Volumetric intratumoral volumes of interest (VOIs) were manually delineated on the late DCE phase using ITK-SNAP (v3.8.0) by a radiologist with 6 years of experience. These VOIs were propagated to the peak-enhancement phase and verified by senior radiologists (with 11 and 21 years of experience) who were blinded to the recurrence score (RS) labels.

For deep-learning feature extraction, bounding-box crops centered on the tumor VOI were extracted from the slice displaying the largest tumor cross-sectional area across axial, coronal, and sagittal planes. Each single-channel grayscale slice was resized to 224×224 pixels and replicated across three channels to match the input architecture of ImageNet-pretrained backbones. Finally, images were normalized using standard ImageNet channel mean and standard deviation statistics.

Radiomics and Deep-Learning Representations

A total of 1,967 handcrafted radiomic features per VOI were extracted using PyRadiomics (v3.0.1) under IBSI-compliant settings across 21 image types: original images (107 features, comprising shape, first-order, and texture descriptors), Laplacian-of-Gaussian (LoG) filtered images (σ = 1.0, 3.0, and 5.0 mm; 93 features each), eight 3D wavelet decomposition subbands (93 features each), 2D local binary patterns

(LBP-2D; 93 features), 3D local binary patterns (LBP-3D-m1, LBP-3D-m2, and LBP-3D-k; 93 features each), and five nonlinear intensity transformations (square, square root, logarithm, exponential, and gradient; 93 features each), using a fixed bin width of 5. Features with pairwise correlation $r > 0.90$ were removed, and LASSO-guided recursive feature elimination retained a subset no larger than 10% of the relevant sample size. SMOTE was applied only within the training partition.

The deep-learning stream used separate 2D ResNet-50 backbones pretrained on ImageNet for axial, coronal, and sagittal tumor views. Training augmentations included random horizontal and vertical flips, affine rotations of ±15°, and intensity jitter. Networks were optimized with AdamW (initial learning rate 1×10^{-4}, weight decay 1×10^{-2}, batch size 32) for up to 50 epochs, with early stopping based on validation loss. Rather than relying solely on the final 2,048-dimensional global average-pooling vector, intermediate convolutional feature maps were flattened and concatenated via multi-scale pooling to yield 16,382 deep features per anatomical plane, resulting in a 49,146-dimensional (16,382 × 3) tri-planar deep representation.

Model Construction and Explainability

Four prediction models were constructed and evaluated: clinical-only, radiomics-only (linear support vector machine), deep-learning-only (tri-planar ResNet-50), and multimodal feature-level fusion. Prior to fusion, concatenated tri-planar deep representations (49,146 dimensions) underwent variance thresholding and LASSO-regularized feature selection strictly within each training fold alongside the selected radiomic subset. The retained clinical variables, selected handcrafted radiomic features, and selected deep-learning features were concatenated and entered into an L2-regularized logistic regression classifier. Hyperparameters were tuned via inner 5-fold cross-validation within the training set. SHAP (SHapley Additive exPlanations) analysis was performed on the external-validation cohort to quantify global and local feature contributions to the fusion model.

The operating probability threshold was fixed at 0.472, corresponding to the positive-class prevalence in the development cohort. Predictions were classified as $RS < 18$ or $RS \geq 18$ using this threshold. SHAP values were used to assess contributions to model predictions.

Statistical Analysis

Continuous variables were summarized as mean ± standard deviation and median (interquartile range); categorical variables were summarized as counts and percentages. Discrimination was evaluated using the area under the receiver operating characteristic curve (AUC), with 95% CIs obtained from 5,000 patient-level bootstrap resamples. AUC, average precision, and Brier score were calculated from continuous prediction scores or probabilities. Accuracy, sensitivity, specificity, positive predictive value, and negative predictive value were calculated at the fixed probability threshold of 0.472. Development-cohort estimates used out-of-fold predictions, and external-validation estimates used predictions for the independently held-out TCGA-BRCA cases.

Calibration was evaluated using grouped calibration curves and Brier scores. Decision-curve analysis compared net benefit across threshold probabilities from 0.0 to 1.0 with treat-all and treat-none reference strategies. These analyses concern classification of the study-defined RS endpoint and do not establish treatment benefit. Baseline characteristics between the development and external-validation cohorts were compared using the independent-samples t-test or Mann–Whitney U test for continuous variables and the χ^2 test or Fisher's exact test for categorical variables. Pairwise differences between model AUCs were statistically compared using DeLong's test. Two-sided $P < 0.05$ was considered statistically significant. Analyses and visualizations were performed using Python v3.8, GraphPad Prism v9.5.0, and SPSS v25.0.

Results

Cohort Characteristics

The analytic cohort comprised 447 patients, of whom 211 (47.2%) had RS ≥ 18. The development cohort contained 153/324 cases with RS ≥ 18 (47.2%), and the external-validation cohort contained 58/123 (47.2%). Baseline characteristics are summarized in Table 1.

Mean age was 53.2 ± 10.5 years in the development cohort and 54.3 ± 10.7 years in the external validation cohort. ER positivity was present in 279/324 (86.1%) and 111/123 (90.2%), PR positivity in

262/324 (80.9%) and 101/123 (82.1%), and HER2 positivity in 91/324 (28.1%) and 26/123 (21.1%), respectively.

Table 1. Baseline Characteristics of Development and Independent External Validation Cohorts

Cohort	n	Age, mean ± SD	Age, median (IQR)	ER+	PR+	HER2+	RS ≥ 18
Development cohort	324	53.2 ± 10.5	52.4 (45.4–60.8)	279 (86.1)	262 (80.9)	91 (28.1)	153 (47.2)
External validation	123	54.3 ± 10.7	53.0 (47.0–62.5)	111 (90.2)	101 (82.1)	26 (21.1)	58 (47.2)

Values are n (%) unless indicated otherwise. ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; IQR, interquartile range; PR, progesterone receptor; RS, recurrence score.

Model Discrimination and Predictive Performance

The multimodal feature-fusion model had the highest observed AUC in both cohorts (Table 2 and Figure 2). In the development cohort, pairwise DeLong's tests demonstrated that the multimodal fusion model achieved a significantly higher AUC than the clinical model (0.841 vs. 0.620; ΔAUC = 0.221 [95% CI, 0.157–0.286]; P < 0.001), the radiomics model (0.841 vs. 0.651; ΔAUC = 0.190 [95% CI, 0.127–0.253]; P < 0.001), and the deep-learning model (0.841 vs. 0.794; ΔAUC = 0.047 [95% CI, 0.019–0.075]; P < 0.001). These findings were corroborated in the independent external-validation cohort, where the multimodal fusion model significantly outperformed the clinical model (0.800 vs. 0.669; ΔAUC = 0.132 [95% CI, 0.029–0.234]; P = 0.012), the radiomics model (0.800 vs. 0.616; ΔAUC = 0.184 [95% CI, 0.070–0.298]; P = 0.002), and the deep-learning model (0.800 vs. 0.686; ΔAUC = 0.115 [95% CI, 0.058–0.171]; P < 0.001).

Table 2. Discrimination and Probabilistic Error Measures

Cohort	Model	AUC (95% bootstrap CI)	Average precision	Brier score
Development	Clinical	0.620 (0.558–0.679)	0.631	0.227
Development	Radiomics	0.651 (0.590–0.711)	0.646	0.229
Development	Deep learning	0.794 (0.745–0.840)	0.766	0.185
Development	Fusion	0.841 (0.797–0.883)	0.822	0.163
External validation	Clinical	0.669 (0.568–0.763)	0.687	0.214
External validation	Radiomics	0.616 (0.513–0.711)	0.552	0.242
External validation	Deep learning	0.686 (0.590–0.774)	0.686	0.222
External validation	Fusion	0.800 (0.722–0.872)	0.777	0.182

Development-cohort metrics were calculated from out-of-fold predictions.

At the fixed probability threshold of 0.472, fusion correctly classified 85/123 external-validation cases, achieving 69.1% accuracy, 65.5% sensitivity, 72.3% specificity, 67.9% PPV, and 70.1% NPV. The clinical model had high specificity (98.5%) but low sensitivity (32.8%). Cohort-specific classification results are reported in Table 3, and external-validation confusion matrices are shown in Figure 3.

Table 3. Classification Performance at the Fixed Probability Threshold of 0.472

Cohort	Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	TN/FP/FN/TP
Development	Clinical	62.7	26.8	94.7	82.0	59.1	162/9/112/41
Development	Radiomics	59.9	46.4	71.9	59.7	60.0	123/48/82/71
Development	Deep learning	71.3	66.7	75.4	70.8	71.7	129/42/51/102
Development	Fusion	76.9	72.5	80.7	77.1	76.7	138/33/42/111
External validation	Clinical	67.5	32.8	98.5	95.0	62.1	64/1/39/19
External validation	Radiomics	59.3	44.8	72.3	59.1	59.5	47/18/32/26
External validation	Deep learning	58.5	55.2	61.5	56.1	60.6	40/25/26/32
External validation	Fusion	69.1	65.5	72.3	67.9	70.1	47/18/20/38

Development-cohort metrics were calculated from out-of-fold predictions.

Calibration Decision Curves and Model Explainability

Fusion had the lowest Brier score in external validation (0.182), compared with 0.214 for the clinical model, 0.242 for radiomics, and 0.222 for deep learning. Calibration plots showed deviations from the ideal 45° line, indicating that calibration requires further assessment before clinical implementation. Decision-curve analysis suggested higher net benefit for fusion across part of the evaluated threshold range (Figure 2); its clinical interpretation depends on a clearly defined decision pathway.

SHAP analysis described contributions from clinical, radiomic, and deep-learning inputs to model predictions (Figures 4 and 5). DL_Feat_4498 showed the highest global importance, followed by PR status and additional deep-learning and radiomic features (Figure 5a–b). Feature correlation clustering described relationships among the representative features (Figure 4c). Grad-CAM overlays were generated for representative cases to visualize regions contributing to the deep-learning predictions (Figure 6). These explanations do not establish biological causality.

Discussion

In this multicohort study of 447 patients across three datasets, fusion of clinical variables, radiomics, and tri-planar deep features had the highest observed AUC in the development and external-validation cohorts. The external-validation AUC was 0.800 (95% CI, 0.722–0.872), and the Brier score was 0.182. Because the external target was transcriptomic-derived RS, this result supports prediction of an independent molecularly derived label rather than equivalence to a commercial Oncotype DX measurement.

Previous imaging and clinicopathologic investigations have associated tumor characteristics with recurrence-score categories [6-11]. Differences in study populations and inputs limit direct performance comparisons [8,11]. Deep-learning models provide nonlinear image representations [16], while scanner-related variability remains a concern for multicenter applications [17,27]. In this study, fusion had the highest observed AUC and lowest Brier score in both the development and external-validation cohorts. The tri-planar architecture provided axial, coronal, and sagittal context without isotropic 3D convolutions. Validation against measured Oncotype DX results in the intended clinical population remains necessary.

SHAP analysis indicated contributions from multiple input types, but does not establish cross-modality interactions or biological causality. DL_Feat_4498 had the highest global importance, followed by PR

status. The 21-gene assay includes genes related to estrogen signaling, HER2, proliferation, and invasion [1,2]. Handcrafted radiomics uses standardized quantitative descriptors [15] to characterize image intensity and texture [28], while deep networks provide learned image representations. The contribution of PR status is consistent with relationships between hormone-receptor biology and recurrence scores [1,2].

Decision-curve results suggest a potential role for further evaluation in preoperative assessment. The framework should be considered an investigational adjunct to validated genomic testing; prospective studies are needed to define the intended action and its clinical consequences.

Potential clinical applications require prospective evaluation:

Preoperative assessment: the fixed probability threshold of 0.472 provides an operating point for this study, but does not establish an optimal threshold for referral, genomic testing, or treatment decisions.

Insufficient tissue for genomic testing: the potential role of MRI-based estimates in this setting requires dedicated validation; the present results do not support substituting imaging predictions for a validated genomic assay.

Resource use: clinical utility, accessibility, and cost effectiveness should be assessed before implementation, including in settings where genomic testing is limited.

Multicenter translation also requires attention to differences in scanner field strength, acquisition, and reconstruction [29]. Radiomic feature values can vary across acquisition settings and software implementations [14,30]. This study used standardized feature extraction [15], intensity normalization, and spatial resampling. Future deployment will require a frozen preprocessing pipeline and evaluation of harmonization approaches, such as ComBat for extracted features or image-level domain adaptation [12,27,29], with data-dependent transformations fitted using training data and assessed in held-out cohorts.

This study has several limitations that warrant consideration:

1) Transcriptomic-derived target: external labels were calculated from RNA-seq rather than measured with the commercial Oncotype DX assay. External performance therefore concerns a transcriptomic-derived RS target, and agreement with measured assay results requires further validation.

2) Clinical cutoff: RS ≥ 18 represents a historical non-low-score category and should not be interpreted as a universal indication for chemotherapy. Future work should assess contemporary decision thresholds and incorporate age, nodal status, and other relevant clinical factors.

3) Interpretability: SHAP and Grad-CAM describe fitted-model behavior and do not establish causal biological mechanisms or prove cross-modality synergy.

4) Clinical utility: retrospective prediction of RS categories and decision-curve results do not establish treatment benefit or cost effectiveness. Prospective evaluation in a defined clinical pathway is required.

Conclusion

The multimodal MRI framework achieved higher observed discrimination than the individual models for predicting RS ≥ 18, including in an independent cohort with transcriptomic-derived labels. Further validation against measured Oncotype DX results in the intended clinical population is needed before clinical implementation.

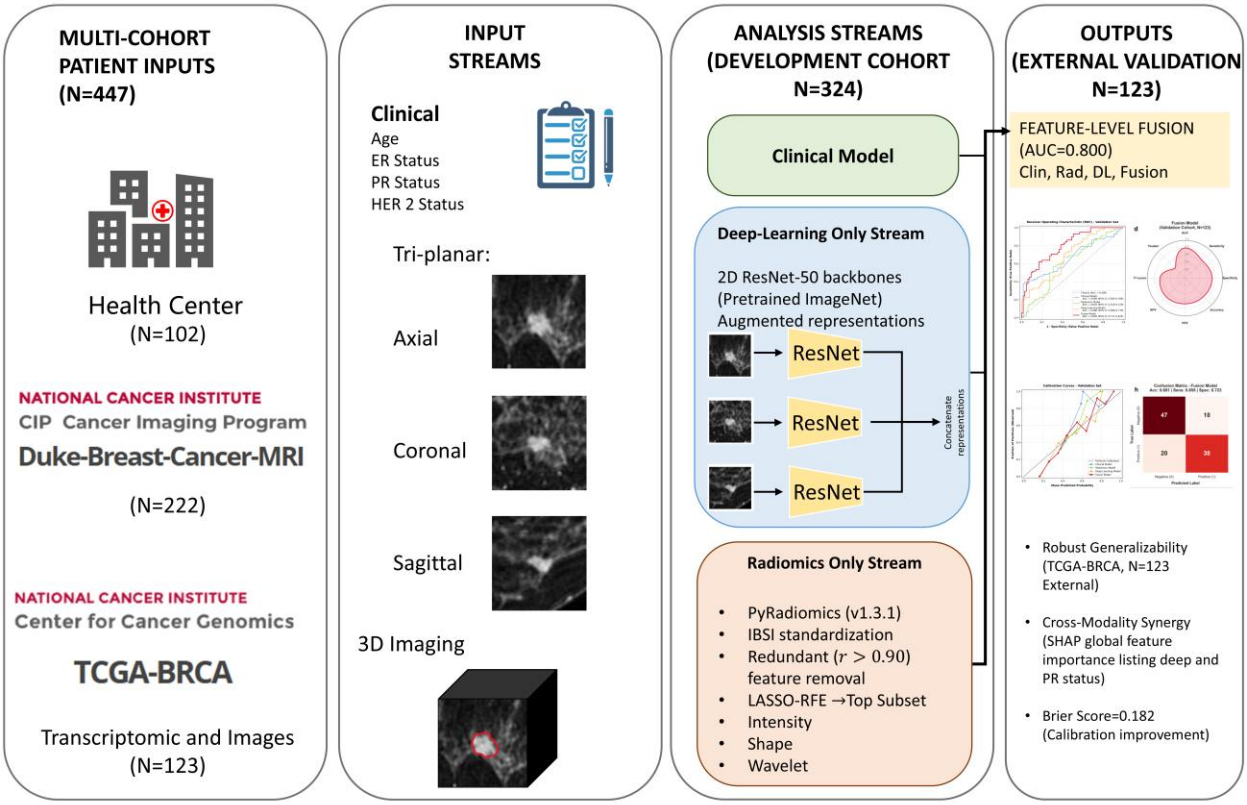

Figure 1. Multimodal radiogenomic workflow and cohort integration.

Study design integrating three datasets (n = 447): Peking University Shenzhen Hospital (n = 102), Duke-Breast-Cancer-MRI (n = 222), and TCGA-BRCA (n = 123). The development cohort (n = 324) used clinical variables, tri-planar 2D ResNet-50 representations, and handcrafted radiomic descriptors. These features were concatenated for logistic regression fusion, and development performance was assessed using out-of-fold predictions. Fusion was evaluated in the held-out TCGA-BRCA cohort with transcriptomic-derived labels, with an AUC of 0.800 (95% CI, 0.722–0.872) and Brier score of 0.182.

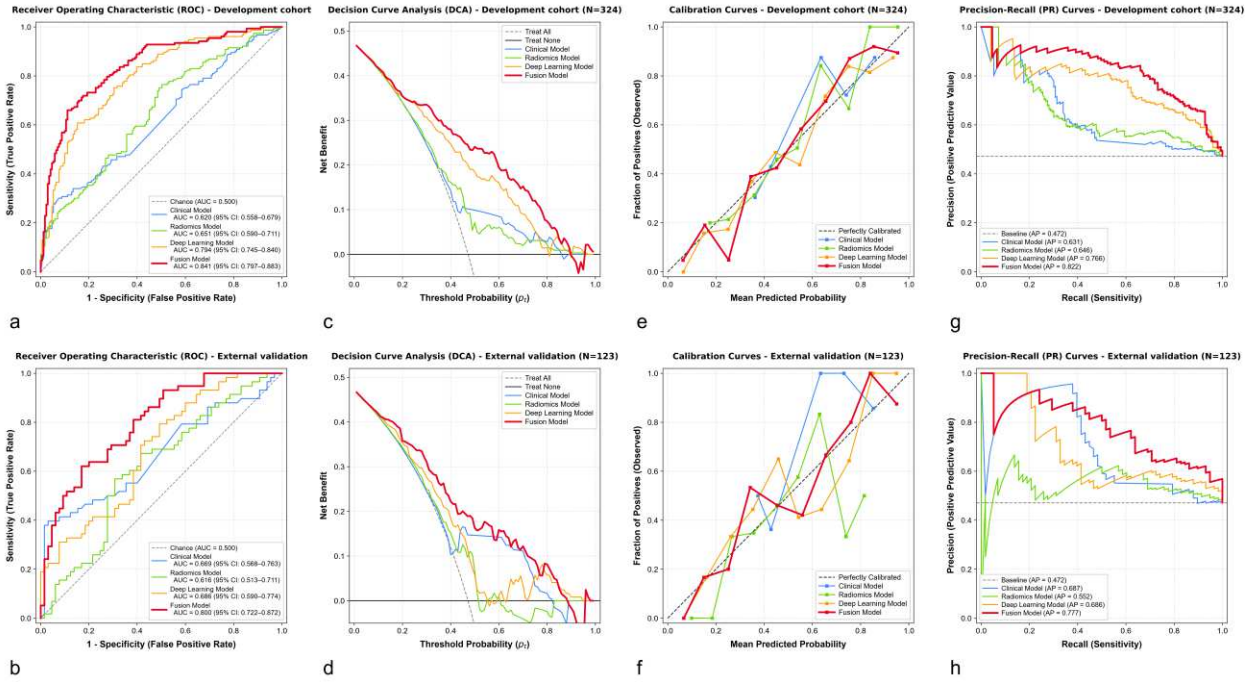

Figure 2. Comprehensive performance metrics across development and external validation cohorts.

Comparative evaluation of clinical, radiomics, deep-learning, and fusion models. Panels a and b show ROC curves for the development cohort (n = 324; out-of-fold predictions) and the independent external-validation cohort (n = 123), respectively. Panels c and d show decision curves; panels e and f show grouped calibration curves; and panels g and h show precision-recall curves. AUC confidence intervals and average precision values are reported in Table 2.

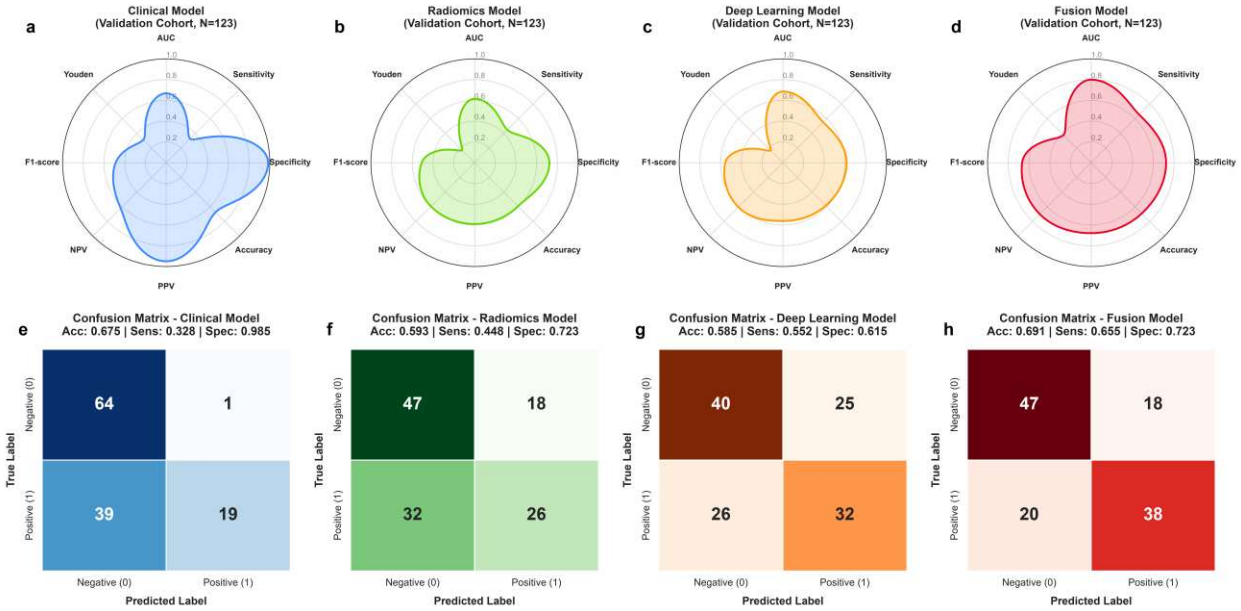

Figure 3. External-validation classification performance and confusion matrices (n = 123). Panels a–d show radar plots for the clinical, radiomics, deep-learning, and fusion models, respectively; panels e–h show their corresponding confusion matrices at the fixed probability threshold of 0.472. Positive and negative labels denote RS ≥ 18 and RS < 18, respectively.

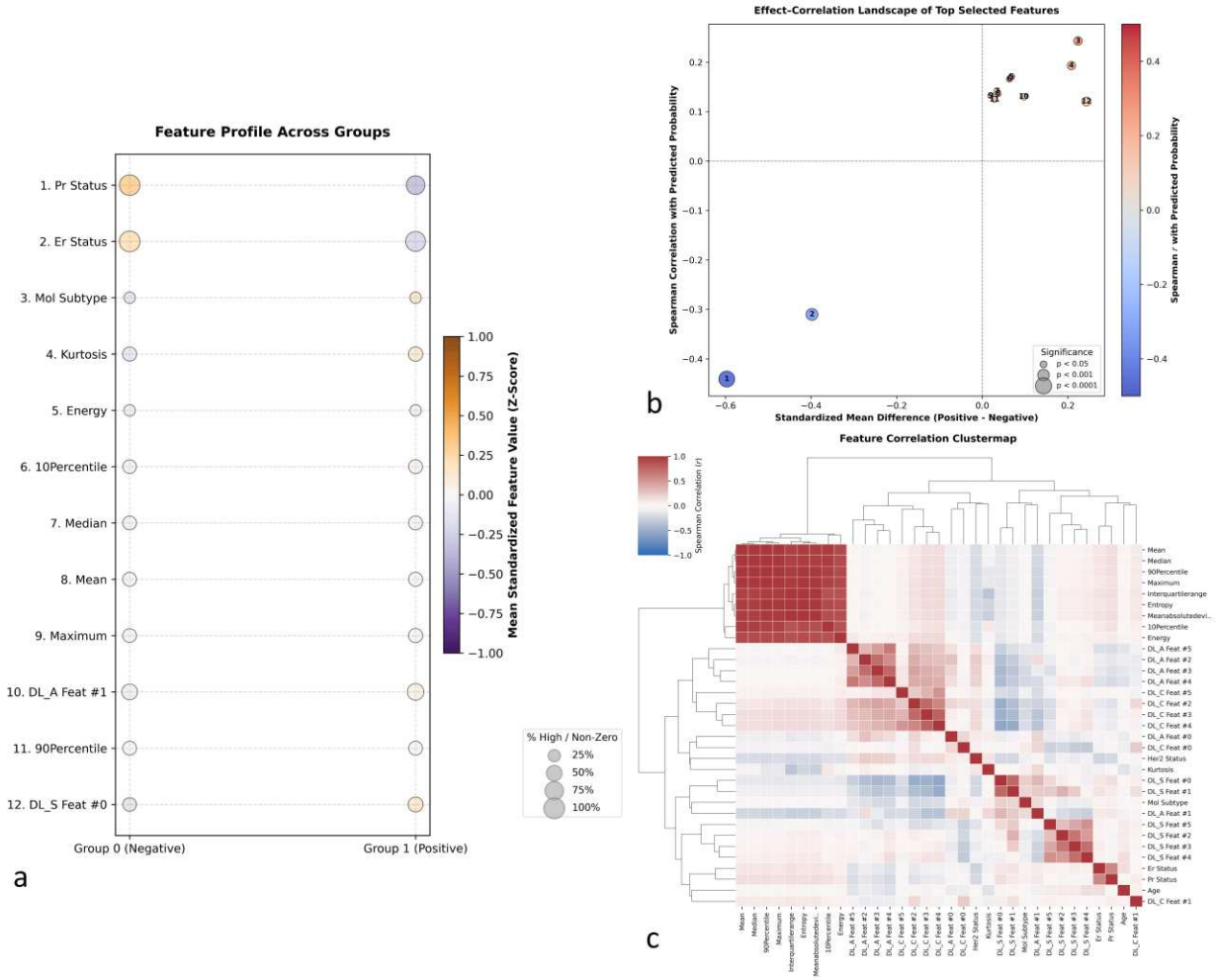

Figure 4. Feature profile, effect-correlation landscape, and cluster map.

a, Group-level feature profiles comparing standardized mean z-scores between the study-defined negative (Group 0) and positive (Group 1) categories. b, Effect-correlation plot of standardized mean difference against Spearman correlation with predicted probability, with bubble size representing statistical significance. c, Hierarchical clustering heatmap of Spearman correlations among representative clinical, radiomic, and deep-learning features.

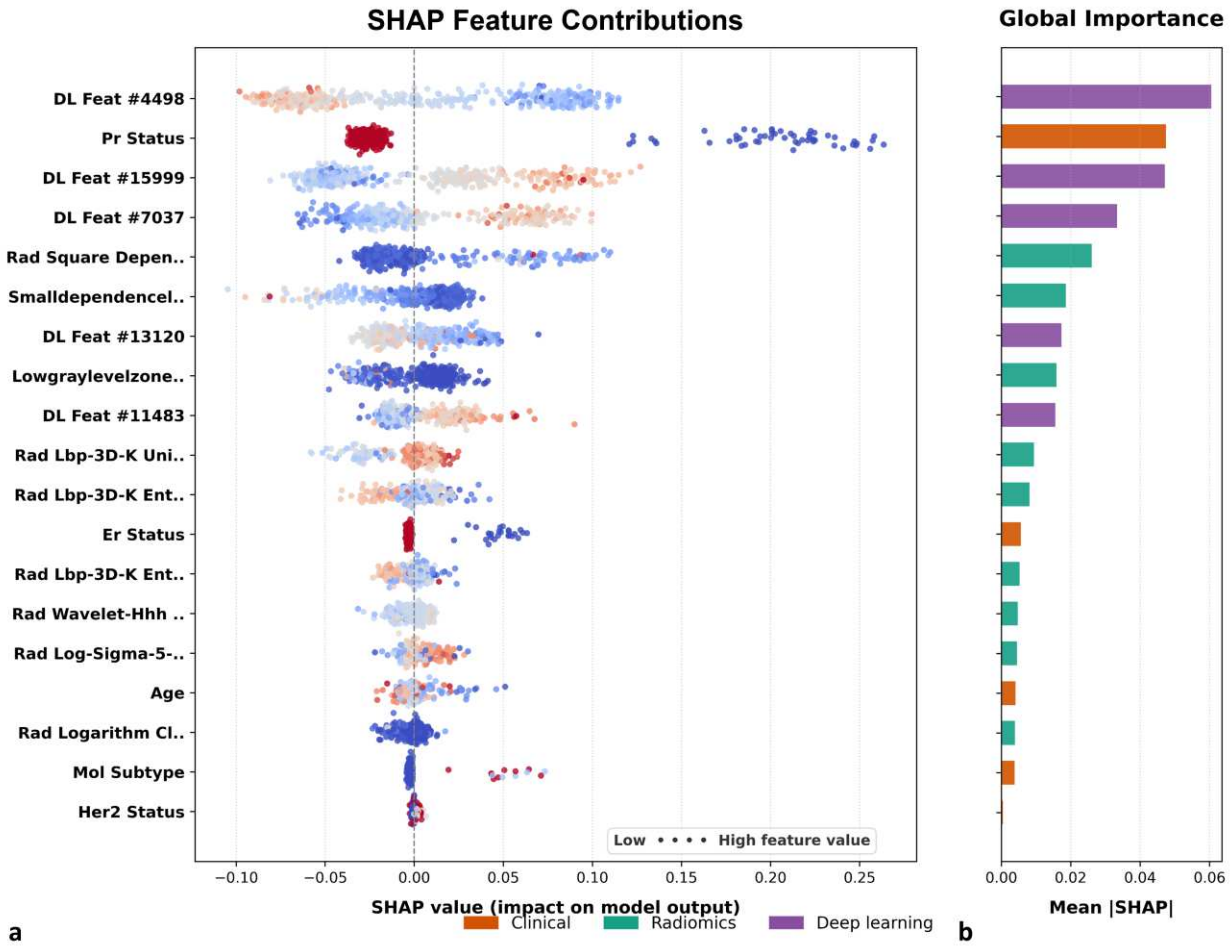

Figure 5. SHAP feature contributions and global importance.

a, SHAP summary plot showing the distribution of feature contributions to the feature-level logistic regression fusion output. b, Global importance ranked by mean absolute SHAP value, with DL_Feat_4498, PR status, and radiomic texture descriptors among the leading contributors. SHAP values describe model behavior and do not establish causal effects.

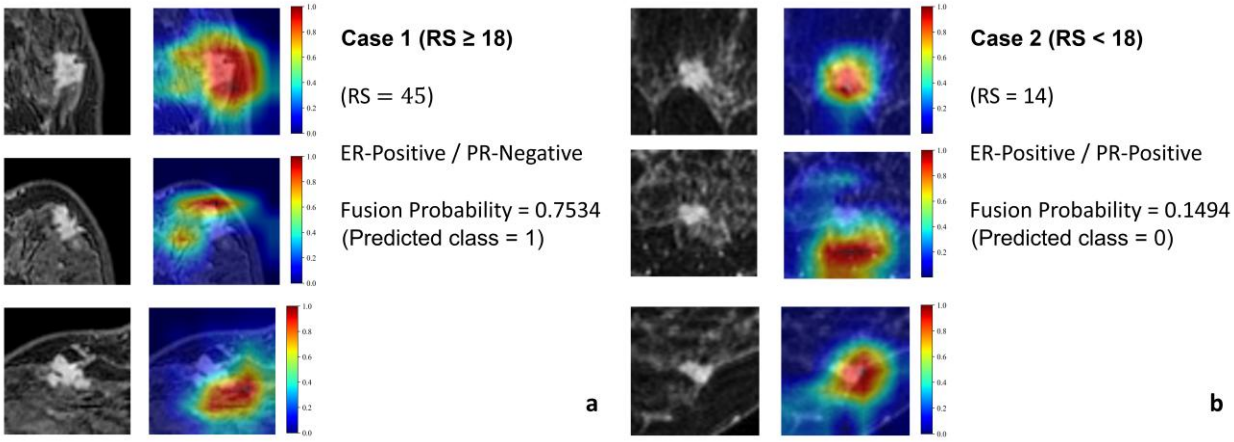

Figure 6. Representative cases and Grad-CAM overlays.

Multiplanar DCE-MRI views and Grad-CAM overlays for two representative cases. a, Case 1 (RS ≥ 18): ER-positive/PR-negative invasive carcinoma with RS = 45; fusion probability 0.7534 and predicted class 1. b, Case 2 (RS < 18): ER-positive/PR-positive invasive carcinoma with RS = 14; fusion probability 0.1494 and predicted class 0. Overlays indicate image regions contributing to deep-learning predictions and do not establish histopathologic mechanisms.

References

1. Sparano JA, Gray RJ, Makower DF, et al. Clinical Outcomes in Early Breast Cancer With a High 21-Gene Recurrence Score of 26 to 100 Assigned to Adjuvant Chemotherapy Plus Endocrine Therapy: A Secondary Analysis of the TAILORx Randomized Clinical Trial. *JAMA Oncol*. Mar 1 2020;6(3):367-374. doi:10.1001/jamaoncol.2019.4794
2. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med*. Dec 30 2004;351(27):2817-26. doi:10.1056/NEJMoa041588
3. Orucevic A, Bell JL, King M, McNabb AP, Heidel RE. Nomogram update based on TAILORx clinical trial results - Oncotype DX breast cancer recurrence score can be predicted using clinicopathologic data. *Breast*. Aug 2019;46:116-125. doi:10.1016/j.breast.2019.05.006
4. Saha A, Harowicz MR, Grimm LJ, et al. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. *Br J Cancer*. Aug 2018;119(4):508-516. doi:10.1038/s41416-018-0185-8
5. Abdullah KA, Marziali S, Nanaa M, Escudero Sánchez L, Payne NR, Gilbert FJ. Deep learning-based breast cancer diagnosis in breast MRI: systematic review and meta-analysis. *Eur Radiol*. Aug 2025;35(8):4474-4489. doi:10.1007/s00330-025-11406-6
6. Chen Y, Xie L, Tang W, et al. MRI Radiomics Signatures of 21-Gene Recurrence Score for Predicting Survival in ER+/HER2- Breast Cancer. *Cancer Med*. Sep 2025;14(17):e71172. doi:10.1002/cam4.71172
7. Park SJ, Kim WH, Kim J, et al. Predictions of Oncotype DX((R)) High-Risk Classification Using Magnetic Resonance Imaging-Based Intratumoral Heterogeneity. *Bioengineering (Basel)*. May 24 2026;13(6)doi:10.3390/bioengineering13060611
8. Kim N, Adam R, Maldjian T, Duong TQ. Radiomics Analysis of Breast MRI to Predict Oncotype Dx Recurrence Score: Systematic Review. *Diagnostics (Basel)*. Apr 22 2025;15(9)doi:10.3390/diagnostics15091054
9. Chiacchiaretta P, Mastrodicasa D, Chiarelli AM, et al. MRI-Based Radiomics Approach Predicts Tumor Recurrence in ER + /HER2 - Early Breast Cancer Patients. *J Digit Imaging*. Jun 2023;36(3):1071-1080. doi:10.1007/s10278-023-00781-5
10. Romeo V, Cuocolo R, Sanduzzi L, et al. MRI Radiomics and Machine Learning for the Prediction of Oncotype Dx Recurrence Score in Invasive Breast Cancer. *Cancers (Basel)*. Mar 18 2023;15(6)doi:10.3390/cancers15061840
11. Alkushi A, Omair A, Arabi H, Masuadi E, Abualkhair O. Predictability of 21-Gene Recurrence Score Assay by Using Pathological and Immunohistochemical Parameters in Breast Cancer. *Breast Cancer (Auckl)*. 2020;14:1178223420977848. doi:10.1177/1178223420977848
12. Leithner D, Nevin RB, Gibbs P, et al. ComBat Harmonization for MRI Radiomics: Impact on Nonbinary Tissue Classification by Machine Learning. *Invest Radiol*. Sep 1 2023;58(9):697-701. doi:10.1097/rli.0000000000000970
13. Orlhac F, Lecler A, Savatovski J, et al. How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. *Eur Radiol*. Apr 2021;31(4):2272-2280. doi:10.1007/s00330-020-07284-9
14. Paquier Z, Chao SL, Acquisto A, et al. Radiomics software comparison using digital phantom and patient data: IBSI-compliance does not guarantee concordance of feature values. *Biomed Phys Eng Express*. Sep 13 2022;8(6)doi:10.1088/2057-1976/ac8e6f

15. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology*. May 2020;295(2):328-338. doi:10.1148/radiol.2020191145
16. Ha R, Chang P, Mutasa S, et al. Convolutional Neural Network Using a Breast MRI Tumor Dataset Can Predict Oncotype Dx Recurrence Score. *J Magn Reson Imaging*. Feb 2019;49(2):518-524. doi:10.1002/jmri.26244
17. Nishizawa T, Maldjian T, Jiao Z, Duong TQ. Attention-based multimodal deep learning for interpretable and generalizable prediction of pathological complete response in breast cancer. *J Transl Med*. Jul 10 2025;23(1):774. doi:10.1186/s12967-025-06617-w
18. Wang X, Hu X, Wang C, et al. Automatic Segmentation and Molecular Subtype Classification of Breast Cancer Using an MRI-based Deep Learning Framework. *Radiol Imaging Cancer*. May 2025;7(3):e240184. doi:10.1148/rycan.240184
19. Boehm KM, El Nahhas OSM, Marra A, et al. Multimodal histopathologic models stratify hormone receptor-positive early breast cancer. *Nat Commun*. Mar 2 2025;16(1):2106. doi:10.1038/s41467-025-57283-x
20. 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*. Oct 14 2025;27(1):178. doi:10.1186/s13058-025-02129-z
21. Shangguan J, Shchukina E, Monov D, Larina S. Integrating deep learning and radiomics for precise identification of luminal A/B breast cancer subtypes on dynamic contrast-enhanced MRI. *Cancer Imaging*. Feb 3 2026;26(1)doi:10.1186/s40644-026-00996-z
22. Zhang Z, Xie H, Liu Y. 23RO Multi-regional fusion framework for predicting neoadjuvant therapy response in HER2-positive breast cancer: A multicenter MRI study. *ESMO Open*. 2026;11doi:10.1016/j.esmoop.2026.107801
23. Wang Y, Zhang Y, Liu Z, Xiong Y, Li M, Zhang L, et al. Multiparametric MRI-based habitat analysis integrating deep learning and radiomics for predicting preoperative Ki-67 expression level in breast cancer. *BMC Med Imaging*. 2026;26(1):80. doi:10.1186/s12880-026-02151-3
24. Gong X, Xu K, Hua M, Zhang Y, Zheng H, Fan S, et al. Deep learning radiomics based on preoperative multiparametric MRI in predicting breast cancer recurrence risk. *BMC Med Imaging*. 2026;26(1):469. doi:10.1186/s12880-026-02774-6
25. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *Bmj*. Apr 16 2024;385:e078378. doi:10.1136/bmj-2023-078378
26. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiol Artif Intell*. Jul 2024;6(4):e240300. doi:10.1148/ryai.240300
27. Jia W, Li H, Ali R, et al. Investigation of ComBat Harmonization on Radiomic and Deep Features from Multi-Center Abdominal MRI Data. *J Imaging Inform Med*. Apr 2025;38(2):1016-1027. doi:10.1007/s10278-024-01253-0
28. Neha F, Shukla DK. Radiomics in Medical Imaging: Methods, Applications, and Challenges. *J Imaging*. May 23 2026;12(6)doi:10.3390/jimaging12060220
29. Pei W, Wei Y, Lai P, et al. Impact of ComBat harmonization on MRI radiomic-analysis of nasopharyngeal carcinoma: a dual-center study. *Front Oncol*. 2026;16:1705567. doi:10.3389/fonc.2026.1705567
30. Fusco R, Festa G, Sansone M, et al. Assessing the multi-software robustness of radiomic biomarkers: a three-tool evaluation. *Front Oncol*. 2026;16:1764691. doi:10.3389/fonc.2026.1764691

List of abbreviations

AUC: Area under the receiver operating characteristic curve; CI: Confidence interval; CLAIM: Checklist for Artificial Intelligence in Medical Imaging; DCE-MRI: Dynamic contrast-enhanced magnetic resonance imaging; ER: Estrogen receptor; FISH: Fluorescence in situ hybridization; Grad-CAM: Gradient-weighted class activation mapping; HER2: Human epidermal growth factor receptor 2; IBSI: Image Biomarker Standardization Initiative; IHC: Immunohistochemistry; IQR: Interquartile range; LASSO: Least absolute shrinkage and selection operator; LBP-2D: Two-dimensional local binary patterns; LBP-3D: Three-dimensional local binary patterns; LoG: Laplacian-of-Gaussian; MRI: Magnetic resonance imaging; NPV: Negative predictive value; PPV: Positive predictive value; PR: Progesterone receptor; RNA-seq: RNA sequencing; ROC: Receiver operating characteristic; RS: Recurrence score; SD: Standard deviation; SHAP: SHapley Additive exPlanations; SMOTE: Synthetic minority over-sampling technique; SVM: Support vector machine; TCGA-BRCA: The Cancer Genome Atlas Breast Invasive Carcinoma; TCIA: The Cancer Imaging Archive; TRIPOD+AI: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence; VOI: Volume of interest.

Declarations

Ethics approval and consent to participate This study was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Review Board of Peking University Shenzhen Hospital (Approval No. PUSH-IRB-20250123). The requirement for written informed consent was waived by the Institutional Review Board due to the retrospective nature of the study and the use of anonymized clinical and imaging data. Publicly available datasets (Duke-Breast-Cancer-MRI and TCGA-BRCA) were de-identified by their respective repositories and required no additional institutional ethical approval.

Consent for publication

All clinical data and medical images presented in this manuscript (including representative MRI scans) have been fully anonymized and contain no identifiable personal information.

Availability of data and materials The public datasets analyzed during the current study are available in The Cancer Imaging Archive (TCIA) repository for the Duke-Breast-Cancer-MRI cohort

(https://www.cancerimagingarchive.net/) and the Genomic Data Commons (GDC) Data Portal / TCIA for the TCGA-BRCA cohort (https://portal.gdc.cancer.gov/).

The in-house clinical and MRI dataset from Peking University Shenzhen Hospital is not publicly available due to patient privacy regulations and institutional data protection policies, but is available from the corresponding authors upon reasonable request and with institutional approval.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

LW and YL conceived and designed the study. LP, ZZ, and ZY collected and curated the clinical, genomic, and imaging data. LW and ZZ performed image segmentation, radiomic feature extraction, deep-learning modeling, and statistical analyses. ZY and YL verified the radiological annotations and supervised the clinical interpretation. LW and ZZ drafted the initial manuscript. ZY and YL critically revised the manuscript for important intellectual content. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx