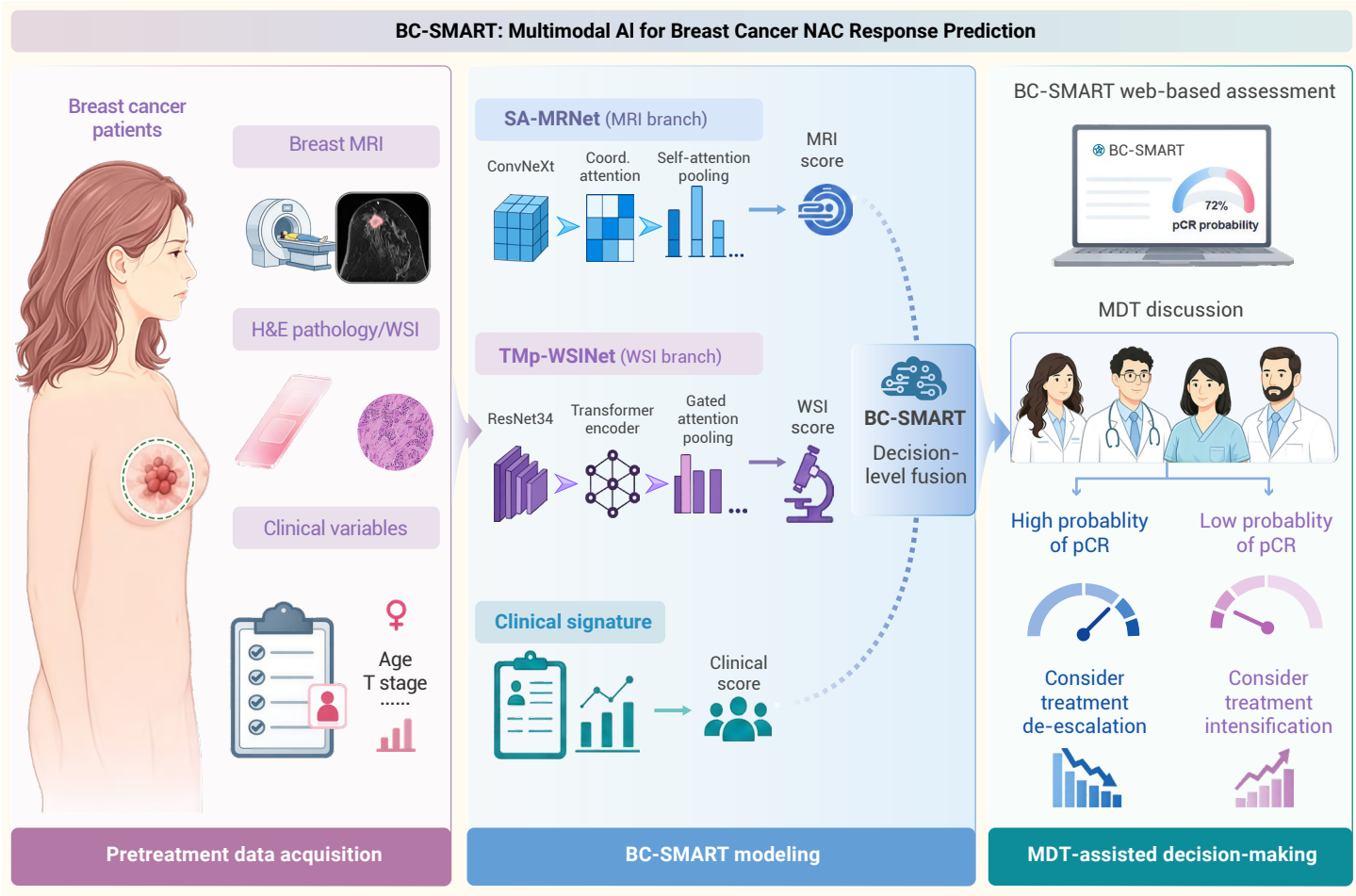
BC-SMART: Multimodal AI framework for breast cancer neoadjuvant chemotherapy response prediction and multidisciplinary evaluation

Wenjie Tang,[1,9] Zijin Weng,[2,9] Yongxin Chen,[1,9] Jinhao Li,[3,9] Jiaxin Chen,[1] Shuang Liu,[1] Yunshi Liang,[4] Siyi Chen,[1] Ruqiong Li,[5] Zhidan Zhong,[1] Liwen Pan,[1] Lei Zhang,[6] Weifeng Liu,[1] Xinhua Wei,[1] Xinqing Jiang,[1] Dilinuer Aishanjiang,[7] Qingcong Kong,[8,*] Yi Dai,[7,*] Enming Cui,[5,*] Jing Yang,[4,*] and Yuan Guo[1,*]
*Correspondence: eyguoyuan@scut.edu.cn (Y.G.); eyjingyang@scut.edu.cn (J.Y.); cem2008@163.com (E.C.); dalydp@126.com (Y.D.); antony_kqc@163.com (Q.K.)
Received: May 8, 2026; Accepted: October 8, 2026; Published Online: October 8, 2026; https://doi.org/10.59717/j.xinn-oncol.2026.100042
© 2026 The Author(s). This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- We developed BC-SMART to predict breast cancer response before neoadjuvant chemotherapy.
- BC-SMART integrates pretreatment magnetic resonance imaging, biopsy whole-slide images, and clinical data.
- The model performed well in both internal and external test cohorts.
- Heatmaps and feature attribution improved interpretation of multimodal predictions.
- BC-SMART may support pretreatment risk assessment and individualized treatment planning.

ARTICLE Cancer Big Data and Artificial Intelligence

OPEN ACCESS

www.the-innovation.org/the-innovation-oncology

BC-SMART: Multimodal AI framework for breast cancer neoadjuvant chemotherapy response prediction and multidisciplinary evaluation

Wenjie Tang,[1,9] Zijin Weng,[2,9] Yongxin Chen,[1,9] Jinhao Li,[3,9] Jiaxin Chen,[1] Shuang Liu,[1] Yunshi Liang,[4] Siyi Chen,[1] Ruqiong Li,[5] Zhidan Zhong,[1] Liwen Pan,[1] Lei Zhang,[6] Weifeng Liu,[1] Xinhua Wei,[1] Xinqing Jiang,[1] Dilinuer Aishanjiang,[7] Qingcong Kong,[8,*] Yi Dai,[7,*] Enming Cui,[5,*] Jing Yang,[4,*] and Yuan Guo[1,*]

[1]Department of Radiology, Guangzhou First People's Hospital, School of Medicine, South China University of Technology, Guangzhou 510180, China
[2]Department of Pathology, The Third Affiliated Hospital, Sun Yat-Sen University, Guangzhou 510630, China
[3]School of Basic Medical Sciences, Tsinghua University, Beijing 100084, China
[4]Department of Pathology, Guangzhou First People's Hospital, School of Medicine, South China University of Technology, Guangzhou 510180, China
[5]Department of Radiology, Jiangmen Central Hospital, Jiangmen 529030, China
[6]Department of Computer & Information Sciences, Towson University, Towson MD 21252, USA
[7]Department of Medical Imaging, Peking University Shenzhen Hospital, Shenzhen 518036, China
[8]Department of Radiology, The Third Affiliated Hospital, Sun Yat-Sen University, Guangzhou 510630, China
[9]These authors contributed equally
*Correspondence: eyguoyuan@scut.edu.cn (Y.G.); eyjingyang@scut.edu.cn (J.Y.); cem2008@163.com (E.C.); dalydp@126.com (Y.D.); antony_kqc@163.com (Q.K.)

Received: May 8, 2026; Accepted: October 8, 2026; Published Online: October 8, 2026; https://doi.org/10.59717/j.xinn-oncol.2026.100042

© 2026 The Author(s). This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

Citation: Tang W., Weng Z., Chen Y., et al. (2026). BC-SMART: Multimodal AI framework for breast cancer neoadjuvant chemotherapy response prediction and multidisciplinary evaluation. *The Innovation Oncology* 1:100042.

Accurate pretreatment prediction of pathologic complete response (pCR) after neoadjuvant chemotherapy (NAC) remains challenging because MRI, biopsy pathology, and clinicopathologic factors capture complementary but incomplete aspects of treatment sensitivity. We developed BC-SMART, a lightweight and interpretable multimodal artificial intelligence framework integrating pretreatment dynamic contrast-enhanced MRI, hematoxylin-eosin-stained biopsy whole-slide images, and routine clinicopathologic variables for pCR prediction in breast cancer. This multicenter retrospective study included 705 women from four institutions, divided into training, internal test, and external test cohorts. The MRI branch used a slice-wise attention network, the pathology branch used an enhanced multiple-instance learning model, and modality-specific scores were combined with clinical variables through transparent decision-level logistic-regression fusion. BC-SMART achieved areas under the receiver operating characteristic curve of 0.903 and 0.901, with accuracies of 0.872 and 0.829, in the internal and external test cohorts, respectively, outperforming clinical, MRI-only, and pathology-only models. Heatmap visualization and Shapley value analysis showed clinically plausible tumor, peritumoral, and tumor-stroma interface regions and quantified the relative contributions of each modality. In a multidisciplinary reader study supported by a prototype web-based interface, BC-SMART improved the performance and confidence of both junior and senior teams. These findings suggest that BC-SMART may provide a practical and interpretable tool for pretreatment pCR prediction and AI-assisted multidisciplinary decision-making in breast cancer.

INTRODUCTION

Breast cancer is the most frequently diagnosed cancer in women worldwide and remains a major contributor to cancer-related mortality.[1] For patients with locally advanced disease or biologically high-risk subtypes, neoadjuvant chemotherapy (NAC) is now widely used to reduce tumor burden, increase the feasibility of breast-conserving surgery, and enable axillary downstaging.[2,3] Pathologic complete response (pCR) after NAC is widely used as a surrogate endpoint for long-term outcomes and is particularly associated with improved event-free and overall survival.[4-6] However, responses to NAC are highly heterogeneous: only a subset of patients achieve pCR, whereas many are exposed to toxicity, financial burden, and delays in optimal local or systemic therapy without deriving meaningful benefit.[7,8] Consequently, there is an urgent clinical need for robust tools that can accurately predict pCR prior to or early during NAC, so that treatment strategies can be tailored to avoid overtreatment in likely good responders and to identify patients who may benefit from treatment escalation or alternative regimens.

Over the past decade, increasing research has explored the utility of AI for predicting NAC response using either single-modality or multimodal data.[9-11] Among available modalities, multiparametric MRI is widely regarded as one of the most sensitive tools for response assessment and is recommended by major guidelines because of its superior soft-tissue contrast and whole-breast coverage.[12,13] Several MRI-based radiomics and deep learning models have shown encouraging performance, with reported areas under the receiver operating characteristic curve (AUCs) ranging from 0.80 to 0.90.[14-16] Nevertheless, while MRI is well suited for characterizing macroscopic tumor burden and perfusion-related features, it cannot fully capture the microscopic heterogeneity of the tumor microenvironment, particularly the complex interactions between cancer cells and the surrounding stroma.[17] In parallel, weakly supervised learning on whole-slide images (WSIs) from pre-treatment biopsies has demonstrated that tissue morphology and stromal patterns hold independent predictive value.[18-20] However, these biopsy-based pathological models are spatially constrained and lack the organ-level context provided by imaging.

To leverage these complementary strengths, recent radiopathomics frameworks have begun to integrate information across scales, theoretically enabling a more comprehensive characterization of treatment sensitivity than any single modality alone.[21,22] However, many existing multimodal approaches rely on relatively complex workflows and deep fusion architectures, which impose substantial computational demands and may hinder implementation in routine clinical practice.[23-25] In addition, the "black-box" nature of these models often obscures the relative contribution of each modality, limiting clinical interpretability and physician trust.[26] Beyond predictive accuracy, evidence remains limited regarding their practical value in real-world clinical scenarios--particularly whether AI assistance can meaningfully improve consensus decision-making within multidisciplinary teams (MDTs). These unmet needs highlight the importance of developing a multimodal model that is not only accurate, but also interpretable, practical, and clinically actionable.

In this study, we developed and validated breast cancer slice-wise multimodal assessment for response to therapy (BC-SMART), a clinically practical multimodal approach for pretreatment prediction of pCR after NAC. BC-SMART integrates baseline MRI, biopsy WSIs, and routine clinicopathologic factors within a streamlined framework designed for both interpretability and potential clinical deployment. Rather than relying on highly complex and computationally intensive architectures, we adopted compact modality-specific models together with a robust decision-level fusion strategy to preserve predictive performance while improving transparency and efficiency. By outputting both modality-specific risk scores and an overall pCR probability, the model provides a clearer view of the relative contributions of MRI, pathology, and clinicopathologic information to the final prediction. In addition, we evaluated its potential role in AI-assisted multidisciplinary assessment. Our overarching aim was to develop an effective and clinically deployable tool for NAC response prediction that could support individualized treatment planning in breast cancer (Figure 1).

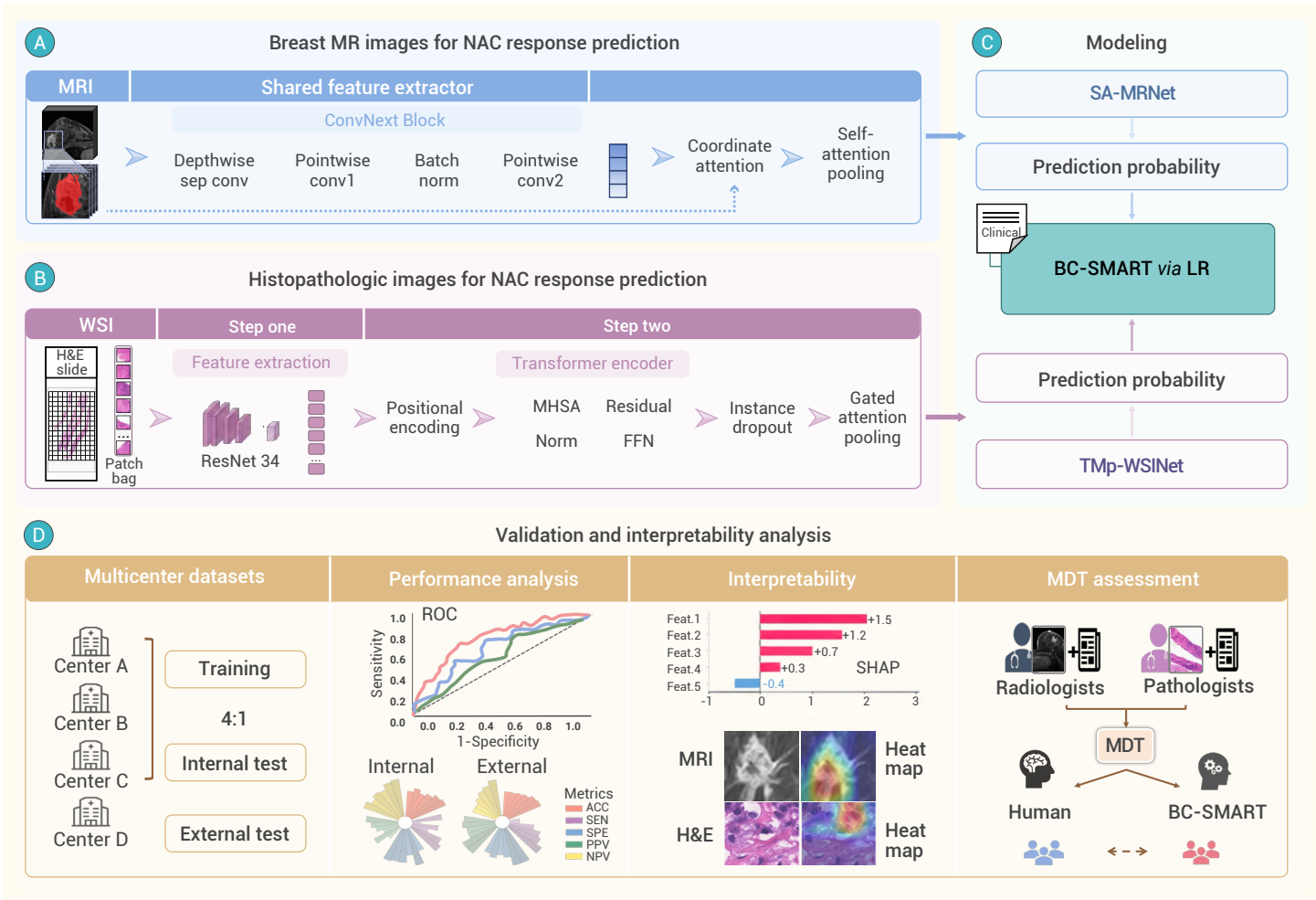


Figure 1. Overview of the multimodal framework and study workflow (A) SA-MRNet for MRI-based prediction of NAC response, integrating slice-wise feature extraction, coordinate attention, and self-attention pooling. (B) TMp-WSINet for WSI-based prediction, combining ResNet-34 feature extraction, Transformer encoding, and gated-attention pooling. (C) BC-SMART integrates MRI, WSI, and clinical signatures through logistic-regression–based decision-level fusion to generate patient-level pCR probabilities. (D) Multicenter validation, model performance assessment, interpretability analyses, and multidisciplinary team evaluation.

MATERIALS AND METHODS

Study participants

This multicenter retrospective study included 705 women with histologically confirmed invasive breast cancer who received NAC followed by surgery at four tertiary hospitals (Centers A–D). Patients were consecutively identified from institutional electronic medical records and pathology databases. The patient selection process and cohort assignment are summarized in Figure 2. Patients from centers A–C were combined and randomly split at the patient level into a training cohort (n = 465) and an internal test cohort (n = 117) at a 4:1 ratio. Patients from center D formed an independent external test cohort (n = 123). For model development, the training cohort was further divided into a model-development subset and a validation subset using stratified sampling according to pCR status. The model-development subset included 372 patients (109 pCR and 263 non-pCR), and the validation subset included 93 patients (27 pCR and 66 non-pCR). This validation subset was used only for model configuration, hyperparameter tuning, early stopping, and backbone/model selection, whereas the internal and external test cohorts were reserved exclusively for locked-model evaluation.

Baseline clinicopathologic variables and treatment-related characteristics were extracted from medical records before NAC, including age, menopausal status, clinical T stage, clinical N stage, MRI-derived tumor size, hormone receptor (HR) status, human epidermal growth factor receptor 2 (HER2) status, Ki-67 index, molecular subtype, and neoadjuvant treatment regimen categories. Detailed definitions of these variables and treatment-related characteristics are provided in Supplementary Appendix 1.

Eligible patients met all of the following criteria: (1) primary invasive breast carcinoma confirmed by biopsy; (2) completion of standard NAC followed by definitive surgery at the participating center;[12,27] (3) availability of pretreatment dynamic contrast-enhanced breast MRI performed before NAC initiation; (4) availability of corresponding pretreatment hematoxylin and eosin (H&E)-stained biopsy WSIs; and (5) complete clinicopathologic data. Exclusion criteria were: (1) history of other malignancies or distant metastasis at diagnosis; (2) prior systemic therapy, radiotherapy, or breast surgery before baseline MRI and/or biopsy; and (3) poor-quality MRI or WSI precluding image analysis.

Within one month after completion of NAC, treatment response was determined on the basis of surgical pathology. pCR was defined as the absence of residual invasive carcinoma in both the breast and sampled axillary lymph nodes (ypT0/is ypN0) and was used as the reference standard for model development.

MRI acquisition and WSI preparation

All patients underwent pretreatment breast MRI at one of the four participating centers using 1.5-T or 3.0-T scanners according to routine institutional protocols. Dynamic contrast-enhanced T1-weighted sequences were acquired with dedicated breast coils, and only the baseline pre-NAC MRI examination was used for model development. Detailed scanner models and acquisition parameters for each center are summarized in Table S1 and Supplementary Appendix 2.

For pathology, ultrasound-guided core-needle biopsy specimens were formalin-fixed, paraffin-embedded, sectioned, and stained with H&E according to standardized laboratory procedures. Slides were digitized at 20× magnification and converted to a unified format before preprocessing; detailed WSI digitization procedures and scanner information are provided in

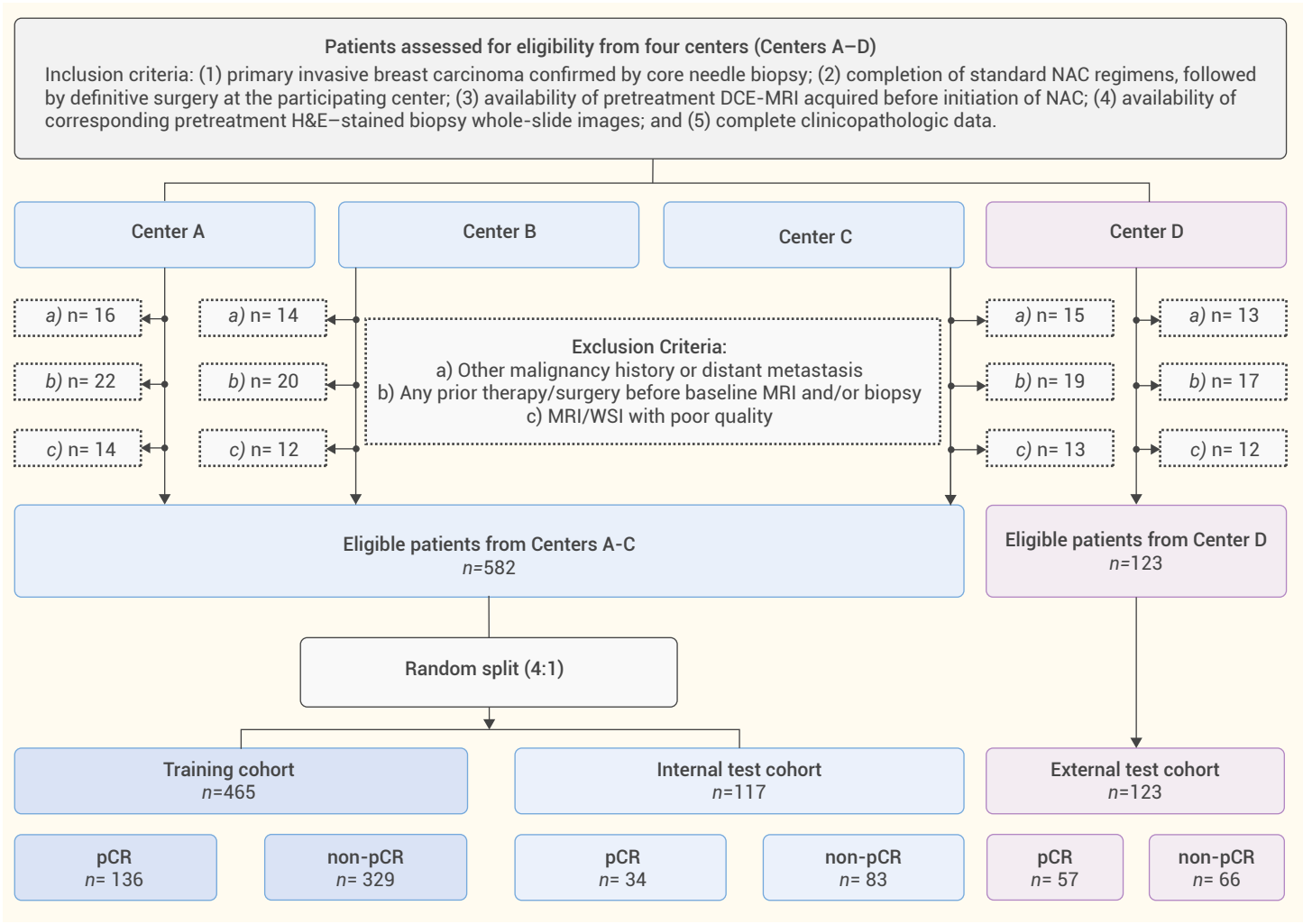


Figure 2. Patient selection flowchart.

Supplementary Appendix 2. For each case, at least one WSI with adequate invasive tumor content and satisfactory image quality was confirmed by experienced breast pathologists and used for downstream analysis.

MRI preprocessing and model development

Tumor regions of interest (ROIs) on baseline MRI were manually delineated by two breast radiologists who were blinded to pathologic outcomes, and any disagreement was resolved by consensus review. To standardize inputs across centers and scanners, baseline MRI volumes were cropped around the radiologist-defined ROI with a small surrounding tissue margin, resampled to a fixed spatial size, and intensity-normalized using a unified preprocessing pipeline. Data augmentation was applied only to the training cohort, whereas deterministic preprocessing was used for the internal and external test cohorts. Detailed ROI annotation and preprocessing procedures are provided in Supplementary Appendix 3.

In this study, we developed a lightweight slice-wise deep learning architecture, the Slice-wise Attention MR Network (SA-MRNet) for pretreatment pCR prediction, as illustrated in Figure S1. Briefly, each preprocessed 3D MRI volume was decomposed into axial 2D slices and processed by a shared convolutional encoder, and slice-level representations were aggregated using attention-based pooling to generate a patient-level prediction. Coordinate attention modules were incorporated to enhance sensitivity to lesion boundaries and peritumoral regions. In addition to SA-MRNet, four commonly used MRI backbones were implemented as reference models under the same data split and preprocessing framework: (1) a ResNet-based slice-wise network, (2) a DenseNet-based slice-wise network, (3) a vision transformer (ViT), and (4) a generic transformer model. Model configuration, including hyperparameter tuning, early stopping, and backbone selection, was determined using the validation subset within the training cohort. After configuration selection, SA-MRNet was retrained on the full training cohort and locked for evaluation in the internal and external test cohorts without further modification, refitting, or recalibration. Detailed network architecture is described in Supplementary Appendix 4.

WSI preprocessing and model development

Pretreatment H&E-stained biopsy WSIs were preprocessed using an automated tissue extraction and quality-control pipeline. Briefly, non-tissue regions and low-information areas were excluded, after which each WSI was tiled into non-overlapping image patches from tissue-containing regions. For each patient, approximately 800 representative tissue patches were sampled to construct a patient-level bag. For downstream analysis, these patches were analyzed within a weakly supervised multiple-instance learning (MIL) framework, in which supervision was provided only at the patient level according to pCR status rather than at the level of individual patches. Data augmentation was applied only to patches in the training cohort, whereas the internal and external test cohorts underwent deterministic preprocessing. Detailed WSI preprocessing procedures are provided in Supplementary Appendix 5.

For WSI-based prediction, we developed an enhanced TransMIL-based architecture, termed TMp-WSINet, to estimate the probability of pCR from each patient's bag of biopsy patches, as illustrated in Figure S1. Briefly, patch-level features were first extracted by an ImageNet-pretrained ResNet-34 encoder and projected into a shared embedding space. To preserve coarse spatial information within each biopsy WSI, normalized two-dimensional patch coordinates were transformed through a learnable coordinate embedding and added to the patch features before transformer-based contextual modeling. During training, instance-level patch dropping was applied to

expose the model to perturbed bag compositions and reduce over-reliance on a small number of dominant patches. Encoded patch tokens were then aggregated using gated-attention pooling to adaptively weight informative regions and generate a global slide-level representation, which was subsequently used to predict the patient-level probability of pCR.

Three additional WSI models were implemented as reference models under the same data split and preprocessing framework: (1) a ResNet-34 model with slide-level pooling, (2) a CLAM-based MIL model, and (3) a baseline TransMIL model. Using the same training–validation workflow, the WSI model configuration was determined using the validation subset within the training cohort. After the configuration was selected, TMp-WSINet was retrained on the full training cohort and locked for evaluation in the internal and external test cohorts without further modification, refitting, or recalibration. Detailed architecture design and training procedures are described in Supplementary Appendix 6.

Clinical and multimodal model development

In the training cohort, candidate clinicopathologic variables included age, menopausal status, baseline tumor size, clinical T stage, clinical N stage, HR status, HER2 status, and Ki-67 index. Univariable logistic regression analyses were first performed with pCR as the dependent variable. Variables with $P < 0.05$ in the univariable analyses were subsequently entered into a multivariable logistic regression model. A backward stepwise selection procedure based on the Akaike information criterion (AIC) was applied and stopped when the minimum AIC was reached. The resulting multivariable logistic regression model, incorporating the selected independent risk factors, was defined as the statistically selected clinical model (Clinical model A) and was used as the original clinical signature to predict the probability of pCR to NAC. In this model, HR status and HER2 status were retained as the final clinicopathologic predictors. To provide a more clinically grounded comparison baseline, we additionally constructed a clinically informed clinical model (Clinical model B) that incorporated HR status, HER2 status, clinical T stage, and clinical N stage, with clinical T and N stage retained regardless of their univariable significance. Clinical model B was fitted in the training cohort and evaluated in the internal and external test cohorts without refitting.

Based on the clinical signature and the two modal-based deep learning signatures, we further constructed decision-level fusion models using logistic regression. After the selected MRI and WSI branches were retrained on the full training cohort and locked, their patient-level prediction scores in the training cohort were used, together with the clinical signature, to fit the fusion models. First, an MRI+WSI fusion model (ST-Fusion model) was developed by jointly modeling the MRI-based SA-MRNet score and the WSI-based TMp-WSINet score; this model represented the combined MRI and WSI probability of achieving pCR. Next, a final multimodal fusion model (BC-SMART) was built by integrating the MRI- and WSI-based deep learning signatures with the original clinical signature derived from Clinical model A, which consisted of HR and HER2 status. The final model was defined as:

$$\log it(p) = \beta_0 + \beta_1 z_{MRI} + \beta_2 z_{WSI} + \beta_3 z_{HR} + \beta_4 z_{HER2}$$

where p denotes the predicted probability of pCR, and z denotes predictors standardized using parameters estimated from the training cohort. The fitted coefficients of the final model are provided in Table S2. The outputs of these models were patient-level probabilities of pCR versus non-pCR. The decision-level fusion models were fitted in the training cohort only. Their performance was then evaluated in the internal and external test cohorts without refitting, recalibration, or further model selection.

Model visualization and interpretability

To improve interpretability, we performed visualization analyses for the MRI- and WSI-based deep learning models. For the MRI model, gradient-weighted class activation mapping (Grad-CAM) was applied to the last convolutional layer of SA-MRNet to generate heatmaps highlighting regions that contributed most to the prediction of pCR versus non-pCR. For each representative case, Grad-CAM maps were overlaid on high-attention slices identified by the self-attention pooling module to illustrate the spatial focus of the network around the primary tumor and peritumoral tissue.

For the WSI model, we exploited the attention weights learned by the TMp-WSINet architecture to construct slide-level attention heatmaps. Patch-level attention scores were mapped back to their spatial locations on the original WSI to visualize highly attended regions that were most informative for predicting pCR. Representative high-attention patches were displayed alongside the corresponding WSIs to facilitate histopathologic interpretation of the deep learning model's focus.The HoverNet model further aided the WSI interpretation by segmenting and classifying cell types from highly attended tiles to quantify the tumor microenvironment. These cell types included tumors, lymphocytes, connective, necrosis, or epithelial cells. For each type, the frequency was calculated for high-attention tiles to analyze fractions in patients with low/high predicted pCR scores.

For the clinical and fusion logistic regression models, feature importance was further explored using coefficient estimates and Shapley value analysis to quantify the relative contribution of HR, HER2, and imaging-based scores to the predicted probability of pCR.

Subgroup analysis

To examine the robustness of the final Multimodal fusion model (BC-SMART) across clinically relevant strata, we conducted predefined subgroup analyses in the internal and external test cohorts based on baseline clinicopathologic characteristics. Patients were stratified by molecular subtype (HR+/HER2−, HER2+, and HR−/HER2− [triple-negative breast cancer, TNBC]), clinical T stage (T1–2 vs T3–4), clinical N stage (N0–1 vs N2–3), and menopausal status (premenopausal *vs* postmenopausal).

Subgroup analyses were performed by applying the final BC-SMART model to the internal and external test cohorts and calculating performance metrics within each predefined subgroup. No subgroup-specific model retraining was performed. For each subgroup, we recalculated discrimination metrics, including AUC, accuracy, sensitivity, specificity, positive predictive value, and negative predictive value, to assess the stability of BC-SMART across different clinical scenarios.

Multidisciplinary reader study and AI assistance

To assess the impact of BC-SMART on human decision-making, we conducted a multidisciplinary reader study on a subset of the test cohorts. To approximate the overall pCR prevalence of the study population (32.2%), 60 patients were selected using stratified random sampling according to pathological response status, including 20 patients with pCR and 40 patients with non-pCR. Six radiologists (three junior and three senior) and six pathologists (three junior and three senior) participated. Each radiologist–pathologist pair formed one MDT, resulting in six MDTs (MDT-1 to MDT-6), including three junior teams (MDT-1 to MDT-3) and three senior teams (MDT-4 to MDT-6).

For each case, MDT members jointly reviewed the pretreatment breast MRI, biopsy H&E WSIs, and clinical information while being blinded to pCR status and model outputs. In the first session (MDT without AI), each MDT assigned a consensus probability of pCR on an 11-point discrete scale from 0.0 to 1.0 in 0.1 increments. This coarse, fixed-step scale was chosen to standardize grading across readers and avoid spurious precision from arbitrary decimal probabilities.

In the second session (MDT with AI), the same MDTs re-evaluated the identical set of 60 cases with assistance from BC-SMART. To facilitate AI-assisted review and explore potential clinical deployment, we developed a prototype web-based interface that integrates MRI visualization, WSI browsing, model inference, and interpretability outputs (MRI Grad-CAM and WSI attention heatmaps), enabling interactive case-level assessment (Figure 3). Through this interface, readers were shown the modality-specific probabilities from the MRI (SA-MRNet), WSI (TMp-WSINet), and clinical branches, the final fused BC-SMART pCR probability, and the corresponding MRI Grad-CAM and WSI attention heatmaps. MDTs then provided an updated consensus pCR probability using the same 0.0–1.0 (step 0.1) scale. All readings were still performed blinded to true pCR status. In the AI-assisted session only, MDTs additionally rated their confidence in the BC-SMART–assisted assessment for each case using a 5-point Likert scale (1 = very uncertain, 5 = very confident).

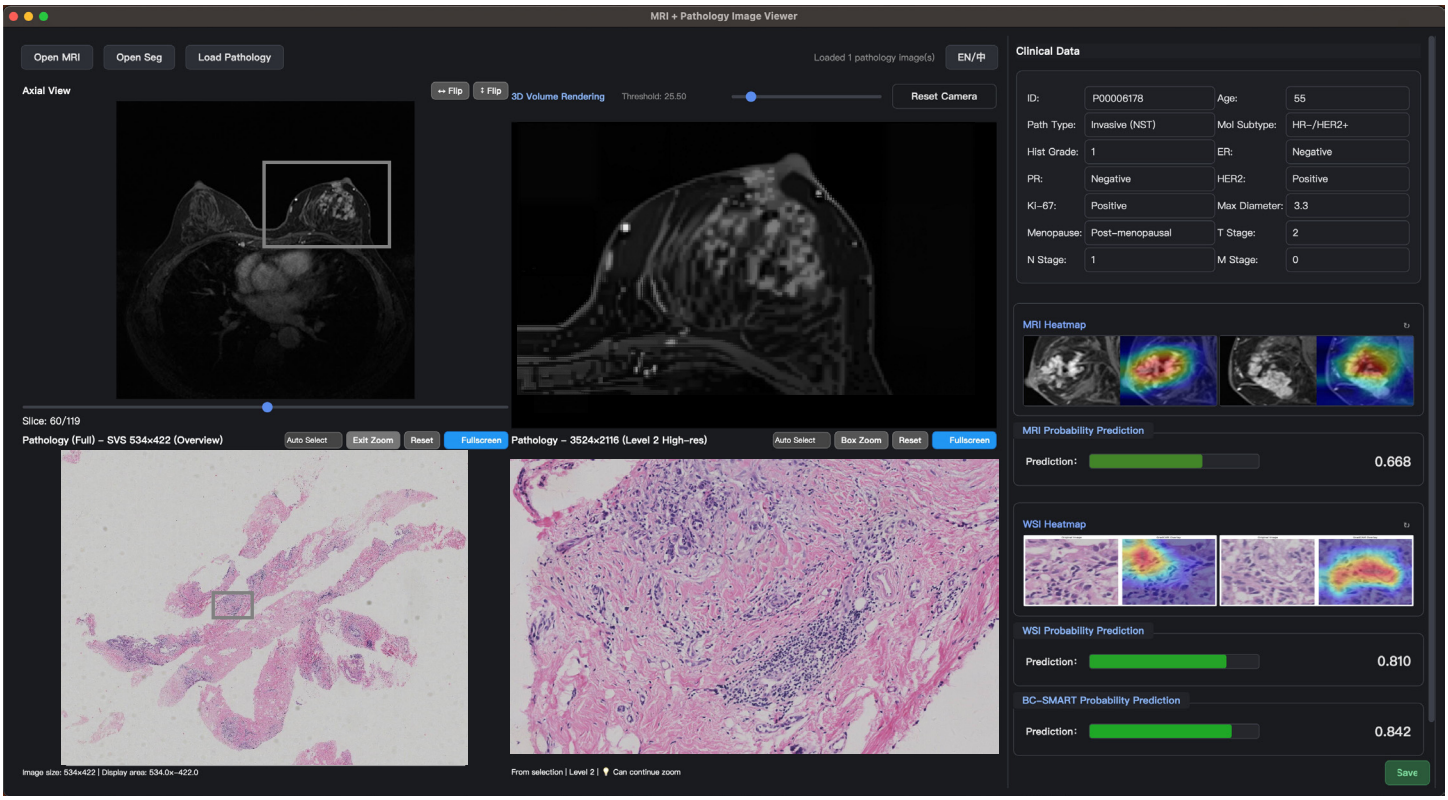


Figure 3. Web-based BC-SMART interface for interactive case review and AI-assisted multidisciplinary assessment Prototype interface displaying pretreatment DCE-MRI, biopsy hematoxylin-eosin WSIs, clinicopathologic information, modality-specific attention heatmaps, and probability predictions generated by the MRI branch, the WSI branch, and the fused BC-SMART model. This interface enables synchronized multimodal review and supports AI-assisted multidisciplinary evaluation at the case level.

Statistical analysis and model evaluation

Statistical analyses were performed using R software (version 3.6.2; www.r-project.org). Categorical variables, including menopausal status, clinical T stage, clinical N stage, HR status, HER2 status, Ki-67 status, and molecular subtype, were compared using the chi-square test or Fisher's exact test, as appropriate. Continuous variables, including age and baseline tumor size, were first assessed for normality. Variables that approximately followed a normal distribution were summarized as mean ± standard deviation (SD) and compared using two-sample t tests, whereas non-normally distributed variables were summarized as median and interquartile range (IQR) and compared using the Mann–Whitney U test.

For model evaluation, the AUC was used as the primary measure of discrimination. Receiver operating characteristic (ROC) curves were generated to visualize the trade-off between sensitivity and specificity across thresholds, and precision–recall (PR) curves with area under the PR curve (AUPRC) were also reported. Accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated using a fixed probability threshold of 0.5. Confusion matrices were generated to provide a more detailed view of classification performance. Ninety-five percent confidence intervals (CIs) for AUC and other performance metrics were estimated using 1000 bootstrap resamples. Comparisons of AUCs between models were performed using the DeLong test.

For the multidisciplinary reader study, MDT-assigned pCR probabilities were treated as continuous prediction scores to construct ROC curves and calculate AUCs for each MDT, both without and with BC-SMART assistance. Within each MDT, AUCs before and after AI assistance, as well as AUCs for MDTs versus the BC-SMART model, were compared using the DeLong test. Likert confidence scores (1–5) from the AI-assisted session were summarized as median (IQR). All statistical tests were two-sided, and $P < 0.05$ was considered statistically significant.

RESULTS

Clinicopathological characteristics and treatment regimen

Baseline clinicopathologic characteristics are summarized in Table 1. Among the 705 women with invasive breast cancer, 227 (32.2%) achieved pCR after NAC. The mean age was 49.7 ± 10.7 years, and 373 patients (373/705, 52.9%) were premenopausal. The median baseline tumor diameter was 3.4 cm (IQR, 2.6 - 4.7 cm). Most patients presented with clinical T1–2 (521/705, 73.9%) and N0–1 disease (456/705, 64.7%). HR–positive and HER2-negative status were observed in 472 (472/705, 67.0%) and 412 (412/705, 58.4%) patients, respectively, and 605 (605/705, 85.8%) had a high Ki-67 index (≥20%). With respect to molecular subtype, 306 patients (306/705, 43.4%) had HR+/HER2– tumors, 293 (293/705, 41.6%) had HER2+ tumors, and 106 (106/705, 15.0%) had TNBC. There were no significant differences in baseline characteristics among the training, internal test, and external test cohorts, except for Ki-67 status ($P = 0.015$).

Given the potential biological and therapeutic heterogeneity associated with molecular subtype, we further summarized subtype-specific pCR rates and NAC regimen distributions across cohorts (Table S3). HER2-positive tumors accounted for 41.5%, 41.9%, and 41.5% of patients in the training, internal test, and external test cohorts, respectively. The pCR rates varied across molecular subtypes, with HER2-positive tumors showing higher pCR rates than HR+/HER2– tumors (48.2% vs 7.7% in the training cohort; 51.0% vs 14.0% in the internal test cohort; and 66.7% vs 24.5% in the external test cohort). Because HER2-positive tumors receive distinct systemic treatment strategies, we further stratified HER2-positive patients into HR+/HER2+ and HR–/HER2+ subgroups and summarized subgroup-specific pCR rates and anti-HER2 treatment patterns (Table S4). HR–/HER2+ tumors demonstrated consistently higher pCR rates than HR+/HER2+ tumors across cohorts (56.6% vs 41.8% in the training cohort, 72.2% vs 38.7% in the internal test cohort, and 88.5% vs 44.0% in the external test cohort). Among HER2-positive tumors, dual HER2 blockade plus chemotherapy was the predominant regimen, accounting for 79.8%, 67.4%, and 96.1% of HER2-positive cases in the training, internal test, and external test cohorts, respectively. To further evaluate potential center-level treatment heterogeneity, center-wise distributions of clinicopathological characteristics, molecular subtypes, NAC regimen categories, and pCR status were summarized (Table S5).

Performance of single-modality MRI and WSI models

To compare MRI-only backbones, we evaluated five architectures (ResNet,

Table 1. Baseline clinicopathological characteristics of patients with breast cancer in different cohorts.

Characteristics	Training cohort	Internal test cohort	External test cohort	*p* value
No. Patients	465	117	123	
Age (y)	49.9±10.7	49.5±10.6	49.0±10.7	0.684
Tumor size at baseline (cm)	3.4 (2.5, 4.5)	3.4 (2.6, 4.8)	3.5 (2.7, 4.9)	0.272
Menopausal status				0.608
Premenopausal	243 (52.26)	60 (51.28)	70 (56.91)	
Postmenopausal	222 (47.74)	57 (48.72)	53 (43.09)	
Clinical T stage				0.115
T1-2	355 (76.34)	82 (70.09)	84 (68.29)	
T3-4	110 (23.66)	35 (29.91)	39 (31.71)	
Clinical N stage				0.180
N0-1	310 (66.67)	75 (64.10)	71 (57.72)	
N2-3	155 (33.33)	42 (35.90)	52 (42.28)	
HR status				0.207
Positive	317 (68.17)	81 (69.23)	74 (60.16)	
Negative	148 (31.83)	36 (30.77)	49 (39.84)	
HER2 status				0.997
Positive	193 (41.51)	49 (41.88)	51 (41.46)	
Negative	272 (58.49)	68 (58.12)	72 (58.54)	
Ki-67 status				0.015
≥20%	410 (88.17)	99 (84.62)	96 (78.05)	
<20%	55 (11.83)	18 (15.38)	27 (21.95)	
Molecular subtypes				0.743
HR+/HER2-	207 (44.52)	50 (42.74)	49 (39.84)	
HER2+	193 (41.51)	49 (41.88)	51 (41.46)	
HR-/HER2- (TN)	65 (13.98)	18 (15.38)	23 (18.70)	

Data are presented as n (%), except for age at baseline (mean ± SD) and baseline tumor size (median [interquartile range]). Abbreviations: pCR, pathologic complete response; HR, hormone receptor; HER2, human epidermal growth factor receptor 2; SD, standard deviation.

ViT, Transformer, DenseNet, and SA-MRNet) for pCR prediction. As shown in Table S6 and Figures 4A, C & E, among the locked candidate MRI models, the slice-wise attention MR network (SA-MRNet) showed the highest overall discrimination in both test cohorts. In the internal test cohort, SA-MRNet yielded the highest AUC (0.797, 95% CI 0.707–0.886), outperforming ResNet, ViT, Transformer and DenseNet (AUC range, 0.642–0.721). In the external test cohort, SA-MRNet maintained the best or near-best performance (AUC 0.787, 95% CI 0.708–0.866). SA-MRNet also reached the highest average precision (AP) in the internal test cohort and a competitive AP in the external test cohort, with a favorable combination of accuracy, sensitivity, specificity, PPV and NPV. These findings suggest that SA-MRNet generalized well across the internal and external test cohorts and was suitable as the MRI backbone for the downstream multimodal framework.

For WSI-only models, four MIL-based networks (ResNet, CLAM, TransMIL and TMp-WSINet) were compared (Table S7 & Figures 4B, D & F). Among these locked candidate WSI models, the TransMILplus-based WSI network (TMp-WSINet) showed the strongest and most consistent overall performance. In the internal test cohort, TMp-WSINet achieved an AUC of 0.778 (95% CI 0.683–0.872), clearly exceeding ResNet, CLAM and baseline TransMIL (AUC range, 0.566–0.681). In the external test cohort, TMp-WSINet further improved to an AUC of 0.820 (95% CI 0.745–0.896), again outperforming the other WSI backbones (AUC range, 0.561–0.696) and yielding the most favorable accuracy and precision–recall profiles.

Taken together, these findings support SA-MRNet and TMp-WSINet as the selected MRI and WSI backbones, respectively, for the subsequent multimodal fusion analysis.

Performance of clinical and multimodal models

In the training cohort, univariable logistic regression identified HR status and HER2 status as significantly associated with pCR (both P < 0.001; Table S8), whereas age, menopausal status, tumor size, clinical T stage, clinical N stage, and Ki-67 index were not. In multivariable analysis, compared with HR-negative disease, HR-positive disease was associated with a lower likelihood of achieving pCR (OR, 0.289; 95% CI, 0.184–0.453; P < 0.001); compared with HER2-negative disease, HER2-positive disease was associated with a higher likelihood of achieving pCR (OR, 4.309; 95% CI, 2.757–6.734; P < 0.001). These two variables were retained in the statistically selected clinical model (Clinical model A). In addition, to provide a more clinically grounded baseline comparator, we constructed a clinically informed clinical model (Clinical model B) incorporating HR status, HER2 status, clinical T stage, and clinical N stage.

We then compared the finalized Clinical model A, Clinical model B, the MRI-based SA-MRNet model, the WSI-based TMp-WSINet model, the MRI–WSI fusion model (ST-Fusion), and the final multimodal fusion model, BC-SMART model (Table 2, Figure 5). In the internal test cohort, the AUCs of Clinical model A, Clinical model B, SA-MRNet, TMp-WSINet, and ST-Fusion models

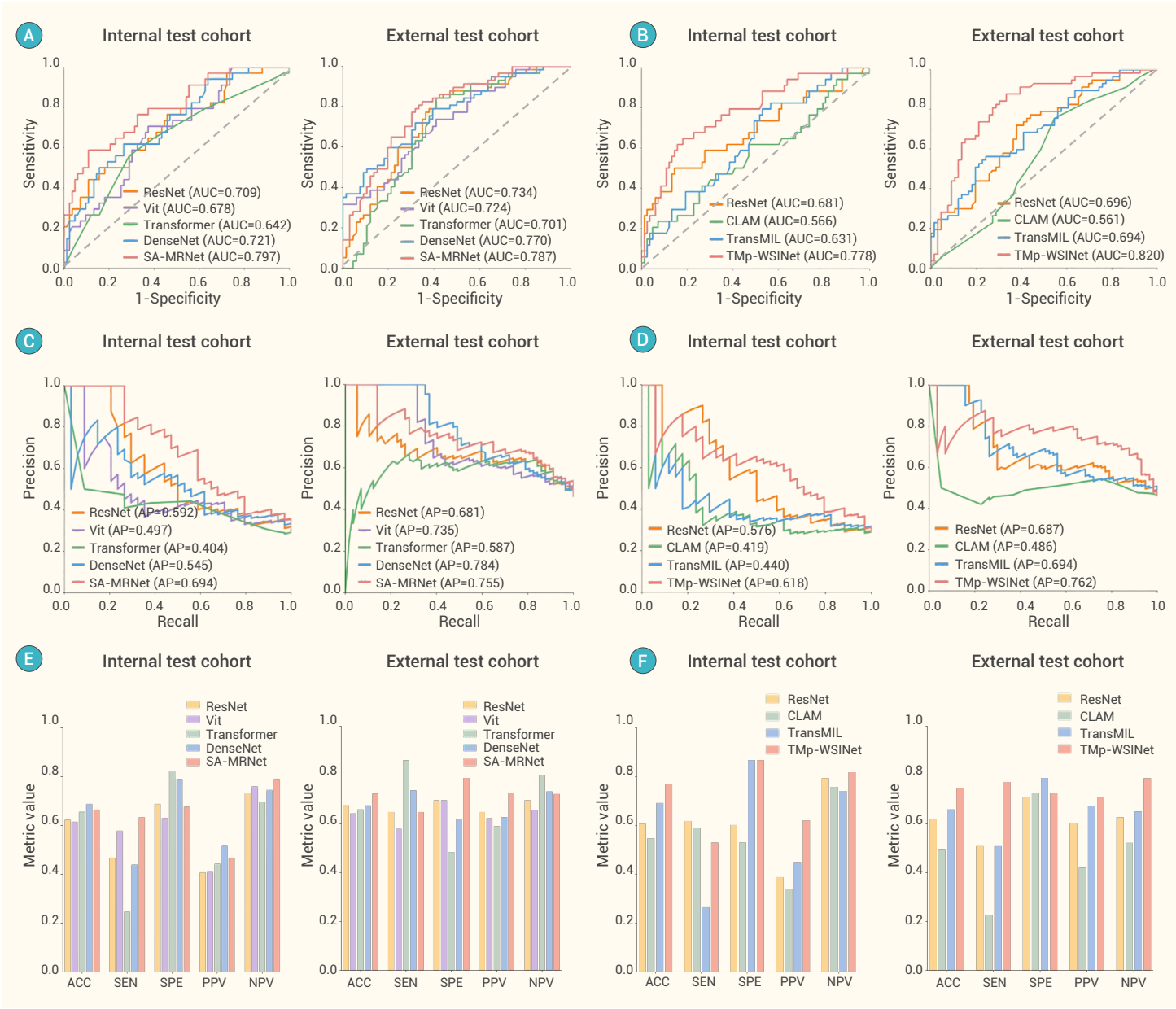


Figure 4. Comparison of MRI- and WSI-based single-modality models for pCR prediction after NAC (A–B) ROC curves for MRI and WSI models in the internal and external test cohorts. (C–D) Corresponding precision–recall curves. (E–F) Bar plots summarizing accuracy, sensitivity, specificity, positive predictive value, and negative predictive value for MRI and WSI models, respectively.

were 0.748, 0.777, 0.797, 0.778, and 0.869, respectively, whereas BC-SMART achieved the highest AUC of 0.903 (95% CI 0.836–0.970). A similar pattern was observed in the external test cohort, where BC-SMART reached an AUC of 0.901 (95% CI 0.847–0.956), outperforming Clinical model A (AUC, 0.743), Clinical model B (AUC, 0.757), all MRI-only or WSI-only models (AUC range, 0.787–0.820), and slightly exceeding the MRI–WSI fusion model (AUC, 0.855; Figure 5A). Compared with Clinical model A, Clinical model B showed modestly improved discrimination across cohorts, with AUCs of 0.768, 0.777, and 0.757 in the training, internal test, and external test cohorts, respectively; however, BC-SMART remained superior to both clinical baselines. DeLong tests showed that BC-SMART significantly outperformed Clinical model A in both test cohorts (all $P < 0.001$), as well as Clinical model B ($P = 0.008$ in the internal test cohort and $P < 0.001$ in the external test cohort). BC-SMART also outperformed SA-MRNet ($P = 0.028$ in the internal test cohort and $P < 0.001$ in the external test cohort) and TMp-WSINet ($P < 0.001$ in the internal test cohort and $P = 0.026$ in the external test cohort). The AUC difference between BC-SMART and ST-Fusion did not reach statistical significance ($P = 0.149$ in the internal test cohort and $P = 0.053$ in the external test cohort), although BC-SMART showed a consistent numerical advantage.

Precision–recall analysis further highlighted the benefit of multimodal fusion: BC-SMART yielded the highest average precision in both the internal (0.818) and external (0.890) test cohorts, indicating superior identification of pCR cases compared with the clinical and single-modality models (Figure 5B). BC-SMART also provided the most favorable overall metric profile (Figure 5C), with accuracies of 0.872 and 0.829 in the internal and external test cohorts, respectively, together with enhanced sensitivity and NPV (sensitivities of 0.853 and 0.930; NPVs of 0.936 and 0.925 in the internal and external cohorts, respectively).

Confusion matrices showed that BC-SMART correctly identified most pCR cases while maintaining an acceptable false-positive rate in both cohorts (Figure 5D). Calibration curves demonstrated good agreement between predicted and observed probabilities for all models, with BC-SMART lying closest to the ideal diagonal (Figure S2). Decision-curve analysis indicated that BC-SMART delivered the greatest net clinical benefit across a broad range of decision thresholds compared with the clinical model and all unimodal and bimodal models (Figure S2). Collectively, these results support BC-SMART as the best-performing model among those evaluated and suggest its potential clinical utility for pretreatment pCR prediction in this multicenter setting.

Table 2. Performance of the clinical and multimodal models for predicting pCR to NAC in different cohorts.

Cohorts	AUC	ACC	SEN	SPE	PPV	NPV	*p* value
Training cohort (n=465)							
Clinical model A	0.756 (0.711 - 0.800)	0.669 (0.668 - 0.670)	0.882 (0.828 - 0.937)	0.581 (0.527 - 0.634)	0.465 (0.404 - 0.526)	0.923 (0.886 - 0.959)	**<0.001**
Clinical model B	0.768 (0.721 - 0.813)	0.731 (0.697 - 0.766)	0.346 (0.265 - 0.426)	0.891 (0.854 - 0.924)	0.566 (0.471 - 0.667)	0.767 (0.745 - 0.791)	**<0.001**
SA-MRNet model	0.965 (0.944 - 0.984)	0.961 (0.944 - 0.978)	0.956 (0.919 - 0.985)	0.964 (0.942 - 0.982)	0.915 (0.872 - 0.957)	0.981 (0.966 - 0.994)	**<0.001**
TMp-WSINet model	0.983 (0.966 - 0.996)	0.978 (0.966 - 0.991)	0.971 (0.941 - 0.993)	0.982 (0.967 - 0.994)	0.957 (0.922 - 0.985)	0.988 (0.976 - 0.997)	**0.038**
ST-Fusion model	0.999 (0.997 - 1.000)	0.978 (0.966 - 0.991)	0.971 (0.941 - 0.993)	0.982 (0.967 - 0.994)	0.957 (0.922 - 0.985)	0.988 (0.976 - 0.997)	0.072
BC-SMART model	0.999 (0.999 - 1.000)	0.989 (0.978 - 0.998)	0.978 (0.949 - 1.000)	0.994 (0.985 - 1.000)	0.985 (0.964 - 1.000)	0.991 (0.979 - 1.000)	ref
Internal test cohort (n=117)							
Clinical model A	0.748 (0.646 - 0.849)	0.709 (0.706 - 0.713)	0.735 (0.587 - 0.884)	0.699 (0.600 - 0.797)	0.500 (0.361 - 0.639)	0.866 (0.784 - 0.947)	**<0.001**
Clinical model B	0.777 (0.670 - 0.869)	0.778 (0.718 - 0.838)	0.382 (0.206 - 0.559)	0.940 (0.880 - 0.988)	0.722 (0.529 - 0.929)	0.788 (0.745 - 0.835)	**0.008**
SA-MRNet model	0.797 (0.707 - 0.886)	0.718 (0.630 - 0.792)	0.676 (0.508 - 0.809)	0.735 (0.631 - 0.818)	0.511 (0.370 - 0.650)	0.847 (0.747 - 0.912)	**0.028**
TMp-WSINet model	0.778 (0.683 - 0.872)	0.769 (0.685 - 0.836)	0.529 (0.367 - 0.685)	0.867 (0.778 - 0.924)	0.621 (0.440 - 0.773)	0.818 (0.725 - 0.885)	**<0.001**
ST-Fusion model	0.869 (0.797 - 0.942)	0.838 (0.835 - 0.840)	0.794 (0.658 - 0.930)	0.855 (0.780 - 0.931)	0.692 (0.547 - 0.837)	0.910 (0.847 - 0.974)	0.149
BC-SMART model	0.903 (0.836 - 0.970)	0.872 (0.870 - 0.874)	0.853 (0.734 - 0.972)	0.880 (0.809 - 0.950)	0.744 (0.607 - 0.881)	0.936 (0.882 - 0.990)	ref
External test cohort (n=123)							
Clinical model A	0.743 (0.658 - 0.828)	0.699 (0.696 - 0.703)	0.404 (0.276 - 0.531)	0.955 (0.904 - 1.000)	0.885 (0.762 - 1.000)	0.649 (0.555 - 0.744)	**<0.001**
Clinical model B	0.757 (0.667 - 0.838)	0.699 (0.634 - 0.764)	0.404 (0.281 - 0.526)	0.955 (0.894 - 1.000)	0.885 (0.750 - 1.000)	0.649 (0.602 - 0.702)	**<0.001**
SA-MRNet model	0.787 (0.708 - 0.866)	0.724 (0.639 - 0.795)	0.649 (0.519 - 0.760)	0.788 (0.675 - 0.869)	0.725 (0.591 - 0.829)	0.722 (0.610 - 0.812)	**<0.001**
TMp-WSINet model	0.820 (0.745 - 0.896)	0.756 (0.673 - 0.823)	0.772 (0.648 - 0.862)	0.742 (0.626 - 0.833)	0.721 (0.598 - 0.818)	0.790 (0.674 - 0.873)	**0.026**
ST-Fusion model	0.855 (0.789 - 0.921)	0.813 (0.811 - 0.815)	0.825 (0.726 - 0.923)	0.803 (0.707 - 0.899)	0.783 (0.679 - 0.888)	0.841 (0.751 - 0.932)	0.053
BC-SMART model	0.901 (0.847 - 0.956)	0.829 (0.827 - 0.832)	0.930 (0.864 - 0.996)	0.742 (0.637 - 0.848)	0.757 (0.657 - 0.858)	0.925 (0.853 - 0.996)	ref

Data are presented as metric values with 95% confidence intervals. Clinical model A refers to the statistically selected clinical model, whereas Clinical model B refers to the clinically informed clinical model. P values are derived from DeLong tests comparing the AUC of each model with that of the BC-SMART model in the corresponding cohort.

Feature- and image-level interpretability of the BC-SMART multimodal framework

To better understand how different modalities and clinicopathologic factors contribute to the final prediction, we first performed shapley additive explanations (SHAP) analysis on the logistic-regression fusion model (Figure S3). In both the internal and external test cohorts, the WSI score and MRI score showed the largest mean absolute SHAP values, indicating that histopathology and MRI were the primary drivers of the predicted pCR probability, whereas HR and HER2 status provided smaller but additional contributions. SHAP summary plots further demonstrated that higher WSI and MRI scores were consistently associated with positive SHAP values, corresponding to an increased likelihood of pCR. The ranking and direction of these effects were highly consistent across the two independent test cohorts, supporting the stability of modality-level interpretability.

To illustrate case-level interpretability, we visualized attention maps on whole-slide images together with Grad-CAM heatmaps on MRI and SHAP-based decompositions of the fusion model for two representative patients (Figure 6). For the MRI model, Grad-CAM in correctly classified cases predominantly focused on the enhancing tumor core and peritumoral regions, rather than distant breast parenchyma, in line with radiologists' visual assessment of treatment-sensitive areas. For the WSI model, attention maps in both pCR and non-pCR cases predominantly focused on invasive tumor nests and the surrounding stromal components, with varying degrees of lymphocytic infiltration across cases.

Taken together, these feature- and image-level analyses suggest that the multimodal BC-SMART framework integrates complementary MRI and WSI information with established clinicopathologic factors while attending to anatomically and histologically plausible regions, thereby achieving accurate prediction with improved transparency and clinical interpretability.

Subgroup analysis of the BC-SMART model

We further evaluated the robustness of BC-SMART across clinically relevant subgroups defined by molecular subtype, clinical T and N stage, and menopausal status (Table S9). Across all predefined subgroups in both the internal and external test cohorts, BC-SMART maintained high discriminative ability, with AUCs generally ranging from 0.787 to 1.000 and accuracies ranging from 0.776 to 0.923.

By molecular subtype, AUCs in the internal test cohort were 0.841 for HR+/HER2−tumors, 0.928 for HER2+ tumors, and 1.000 for TNBC, with corresponding values of 0.802, 0.908, and 0.909 in the external test cohort. Performance was particularly strong in HER2+ and TNBC subgroups. For all subtypes, accuracies and negative predictive values (NPVs) were consistently high (≥0.75), supporting the identification of patients unlikely to achieve pCR.

Stratification by clinical T stage showed similarly robust results. In the internal test cohort, AUCs were 0.906 for T1–2 and 0.960 for T3–4 disease; in the external test cohort, the corresponding AUCs were 0.864 and 0.976.

For clinical N stage, the internal test cohort yielded AUCs of 0.932 for N0–1 and 0.787 for N2–3 disease, while in the external cohort the AUCs were 0.893

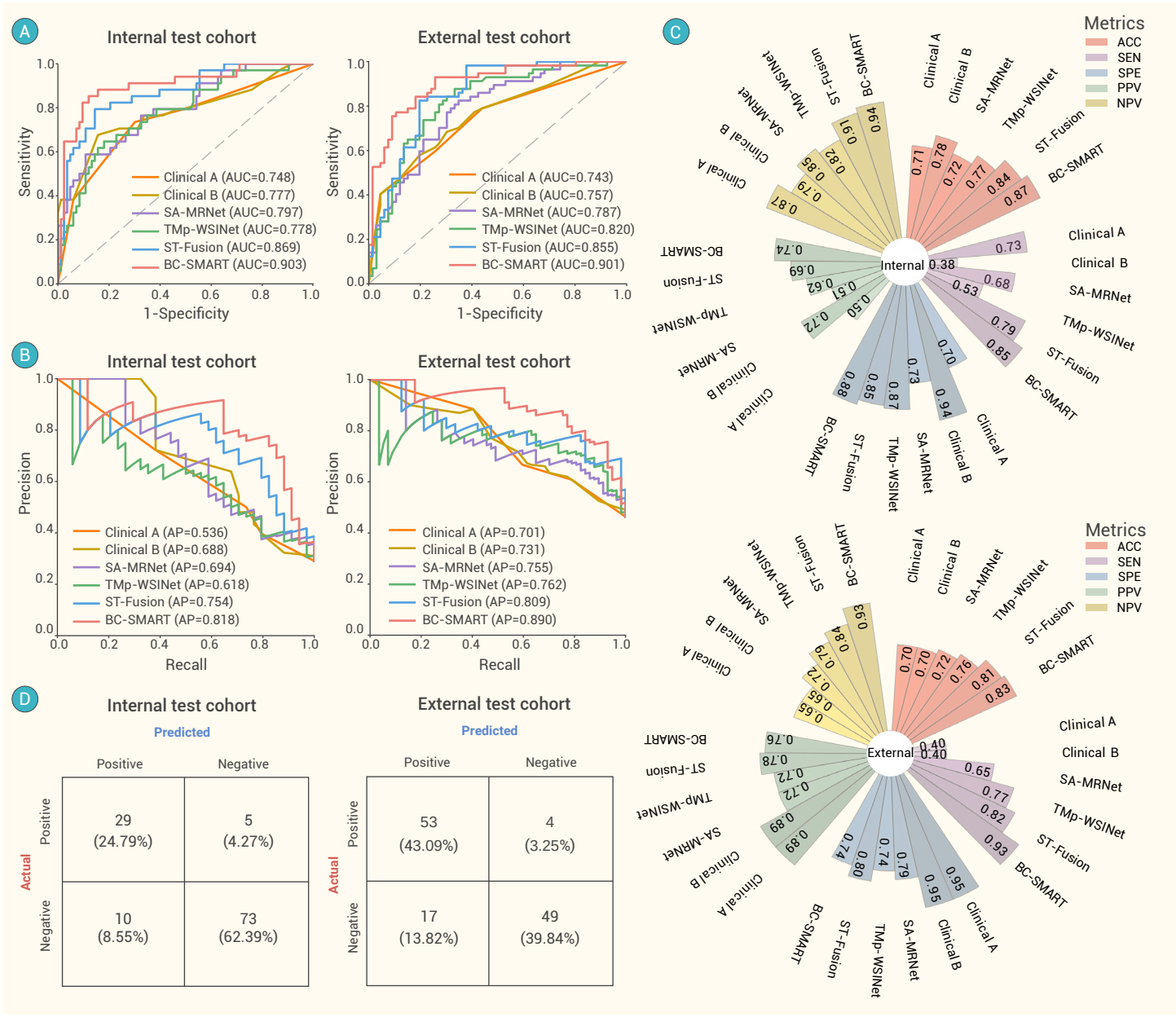


Figure 5. Performance of clinical, single-modality, and multimodal models for predicting pCR after NAC in the internal and external test cohorts (A) ROC curves comparing the clinical model A, clinical model B, MRI model (SA-MRNet), WSI model (TMp-WSINet), MRI–WSI fusion model (ST-Fusion), and full multimodal model (BC-SMART) in the internal (left) and external (right) test cohorts. (B) PR curves of the same six models in the internal and external test cohorts. (C) Radar plots summarizing key performance metrics (ACC, SEN, SPE, PPV, and NPV) for all models in the internal (top) and external (bottom) test cohorts. (D) Confusion matrices of the BC-SMART model in the internal and external test cohorts, showing the distribution of true/false positives and true/false negatives.

and 0.904, respectively, again accompanied by accuracies and NPVs of at least 0.80 in both cohorts.

When stratified by menopausal status, BC-SMART also performed well. In the internal test cohort, AUCs were 0.867 in premenopausal and 0.924 in postmenopausal women; in the external test cohort, the corresponding AUCs were 0.902 and 0.893, with accuracies and NPVs all exceeding 0.75.

Taken together, these subgroup analyses indicate that BC-SMART preserves strong predictive performance across diverse biological and clinical strata, supporting its robustness in heterogeneous patient subsets.

Human-machine collaboration based on BC-SMART

The overall workflow of the human–AI collaboration study is illustrated in Figure 7A. In the reader study comprising 60 cases selected using stratified sampling according to pathological response status, the standalone BC-SMART model achieved an AUC of 0.894 (95% CI, 0.792–0.974), with an accuracy of 0.850, sensitivity of 0.800, and specificity of 0.875 (Table 3). Without AI assistance, MDT performance varied according to clinical experience. Junior MDTs (MDT-1 to MDT-3) demonstrated limited discrimination, with AUCs ranging from 0.561 to 0.573, whereas senior MDTs (MDT-4 to MDT-6) achieved moderately higher AUCs ranging from 0.688 to 0.746. However, all MDTs showed significantly lower AUCs than the standalone BC-SMART model (all DeLong test $P < 0.05$; Table 3, Figure 7B).

With the assistance of BC-SMART, the diagnostic performance of all MDTs improved substantially. Among junior MDTs, AUCs increased to 0.756–0.808, with corresponding accuracies of 0.750–0.817. Among senior MDTs, AUCs increased to 0.836–0.879, with accuracies of 0.800–0.867 (Table 3, Figure 7B). These improvements were accompanied by higher specificity and positive predictive values, especially for the advanced multidisciplinary teams, while the sensitivity and negative predictive values remained at clinically acceptable levels (Figure 7C). Pairwise DeLong comparisons between each MDT before and after AI assistance confirmed significant increases in AUC for all six teams ($P = 0.005$, 0.003, 0.009, 0.012, 0.046, and 0.035 for MDT-1 to MDT-6, respectively; Figure 7D). After AI assistance, the performance of all MDTs approached that of the standalone BC-SMART model, with no statisti-

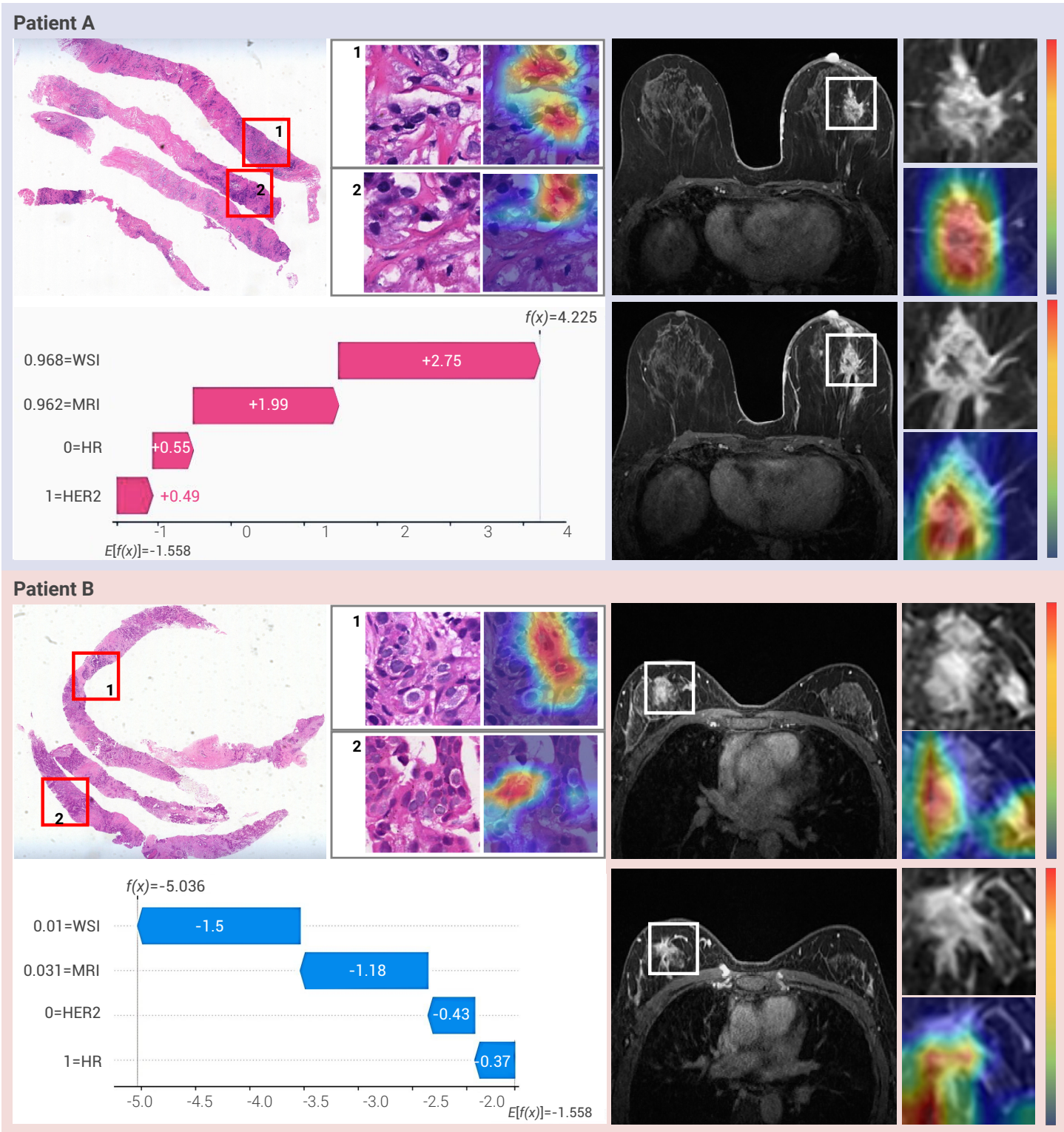


Figure 6. Case-level multimodal interpretability of BC-SMART Representative high-pCR-probability (Patient A, top) and low-pCR-probability (Patient B, bottom) cases. WSI panels show high-attention regions and heatmaps highlighting influential tumor and stromal areas, while MRI panels show tumor ROIs and Grad-CAM maps focused on tumor and peritumoral regions. Shapley value plots depict the contribution of each modality to the final prediction.

cally significant differences observed between AI-assisted MDTs and BC-SMART (all P > 0.05).

Regarding confidence in AI-assisted assessment, MDTs generally reported high confidence in BC-SMART–assisted evaluations. The median confidence score was 4 (IQR, 3–4 or 4–5) for MDT-1, MDT-2, and MDT-4–MDT-6, whereas MDT-3 showed a median score of 3 (IQR, 3–4). Across MDTs, the proportion of high-confidence ratings (Likert scores 4–5) ranged from 55.0% (33/60) to 75.0% (45/60), while MDT-3 showed a lower proportion of high-confidence ratings (36.7%, 22/60) (Figure 7E).

To further illustrate the interpretability of BC-SMART and its potential role in MDT decision-making, representative cases were visualized (Figures S4–S7). These cases illustrated the complementary information provided by radiologic and pathologic features for pCR prediction and showed how BC-SMART integrates macro-scale MRI characteristics and

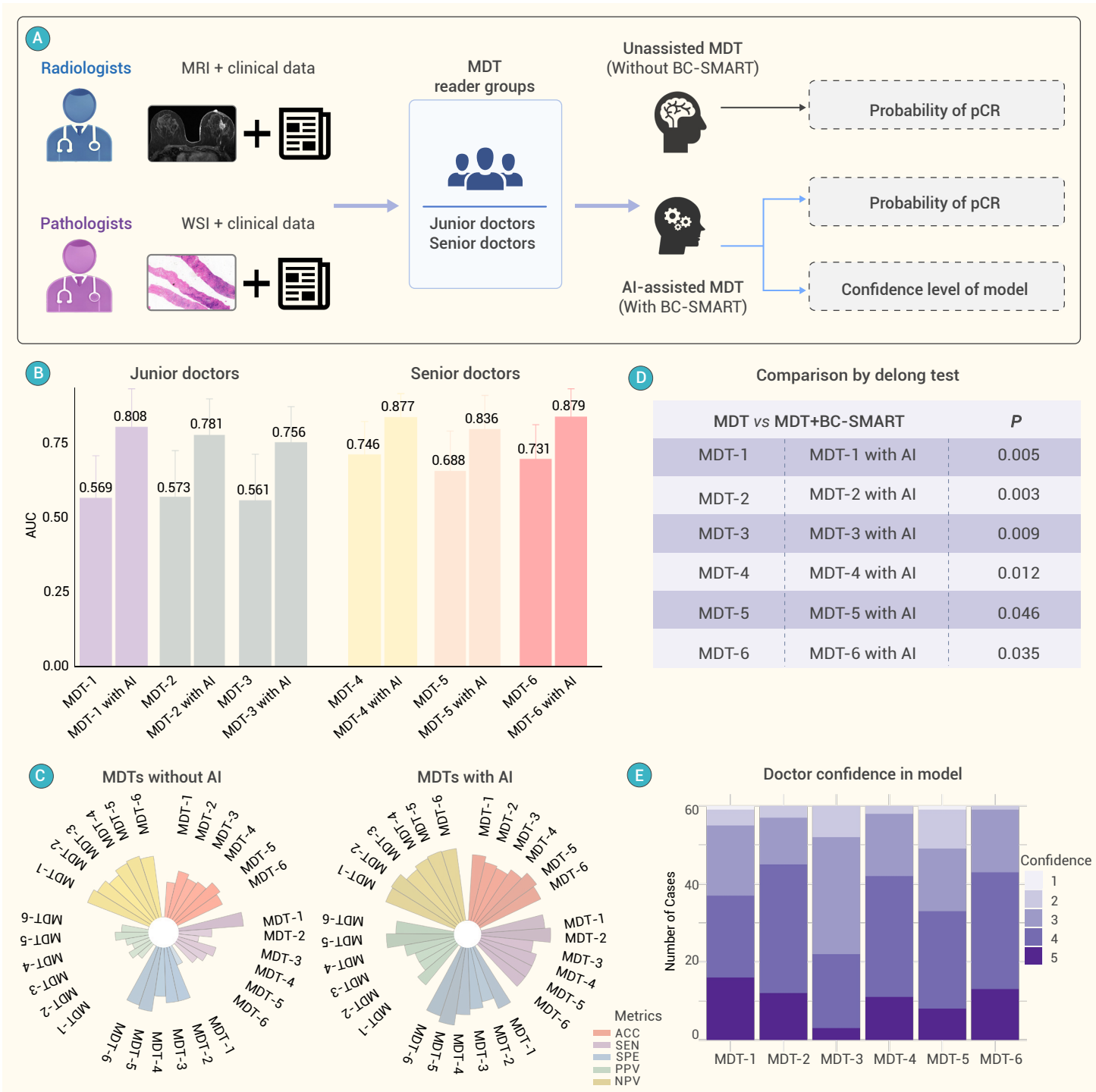


Figure 7. Human–AI collaboration analysis of MDT performance with and without BC-SMART assistance (A) Schematic of the reading workflow. (B) AUCs with 95% CIs before and after AI assistance. (C) Radar plots of diagnostic performance. (D) DeLong test comparison of AUCs between MDTs with and without BC-SMART. (E) Physician confidence after BC-SMART-assisted assessment.

micro-scale histopathologic patterns to generate interpretable predictions. The web-based interface displayed MRI, WSI, clinical information, modality-specific attention heatmaps, and final BC-SMART predictions, facilitating review of model outputs during AI-assisted MDT assessment.

DISCUSSION

In this multicenter retrospective study, we developed and validated BC-SMART, a lightweight multimodal framework integrating pretreatment MRI features, biopsy WSI characteristics, and routine clinicopathologic factors to predict pCR to NAC in invasive breast cancer. BC-SMART demonstrated favorable predictive performance and overall accuracy, outperforming conventional clinical models and single-modality models in both test cohorts. Beyond accurate risk stratification, the framework offers transparent feature interpretability by effectively bridging macroscopic radiological patterns with microscopic histopathological traits. Furthermore, BC-SMART was proven to tangibly enhance the confidence and precision of MDT decision-making, providing a transparent, efficient, and clinically deployable solution for precision oncology.

Recent radiopathomics frameworks have begun to integrate cross-scale information, leveraging the complementary strengths of macroscopic MRI and microscopic WSI to better characterize tumor heterogeneity. While previous studies, such as Guo et al.'s ultrasound-pathology model[28] and Mao et al.'s MRI-WSI-based MIFAPS system,[24] have demonstrated the strong predictive potential of this approach, they often entail complex workflows, serial

Table 3. Diagnostic performance of MDTs for predicting pCR to NAC with and without BC-SMART assistance.

	AUC (95% CI)	ACC	SEN	SPE	PPV	NPV	P value
BC-SMART	0.894 (0.792-0.974)	0.850 (0.750-0.933)	0.800 (0.600-0.950)	0.875 (0.750-0.975)	0.762 (0.609-0.938)	0.897 (0.816-0.973)	ref
Junior Doctors							
MDT-1	0.569 (0.422-0.711)	0.517 (0.400-0.633)	0.750 (0.550-0.950)	0.400 (0.250-0.550)	0.385 (0.300-0.472)	0.762 (0.600-0.917)	**<0.001**
MDT-2	0.573 (0.420-0.728)	0.617 (0.500-0.733)	0.500 (0.300-0.700)	0.675 (0.525-0.800)	0.435 (0.280-0.593)	0.730 (0.636-0.825)	**0.001**
MDT-3	0.561 (0.407-0.716)	0.583 (0.467-0.700)	0.400 (0.200-0.600)	0.675 (0.525-0.800)	0.381 (0.222-0.550)	0.692 (0.606-0.784)	**<0.001**
Senior Doctors							
MDT-4	0.746 (0.612-0.862)	0.600 (0.483-0.717)	0.550 (0.350-0.750)	0.625 (0.475-0.750)	0.423 (0.292-0.560)	0.735 (0.636-0.842)	**0.025**
MDT-5	0.688 (0.532-0.829)	0.667 (0.550-0.783)	0.500 (0.300-0.700)	0.750 (0.600-0.875)	0.500 (0.320-0.700)	0.750 (0.660-0.848)	**0.019**
MDT-6	0.731 (0.596-0.851)	0.633 (0.517-0.750)	0.450 (0.250-0.700)	0.725 (0.575-0.850)	0.450 (0.278-0.636)	0.725 (0.639-0.818)	**0.010**
Junior Doctors with BC-SMART							
MDT-1	0.808 (0.642-0.936)	0.817 (0.700-0.900)	0.800 (0.600-0.950)	0.825 (0.700-0.925)	0.696 (0.545-0.857)	0.892 (0.805-0.972)	0.179
MDT-2	0.781 (0.649-0.902)	0.767 (0.667-0.867)	0.850 (0.700-1.000)	0.725 (0.575-0.850)	0.607 (0.500-0.750)	0.906 (0.818-1.000)	0.152
MDT-3	0.756 (0.617-0.876)	0.750 (0.633-0.850)	0.700 (0.500-0.900)	0.775 (0.650-0.900)	0.609 (0.464-0.778)	0.838 (0.750-0.938)	0.084
Senior Doctors with BC-SMART							
MDT-4	0.877 (0.767-0.960)	0.800 (0.683-0.900)	0.750 (0.550-0.950)	0.825 (0.700-0.925)	0.682 (0.526-0.850)	0.868 (0.778-0.966)	0.595
MDT-5	0.836 (0.700-0.954)	0.867 (0.783-0.950)	0.750 (0.550-0.950)	0.925 (0.825-1.000)	0.833 (0.682-1.000)	0.881 (0.805-0.971)	0.443
MDT-6	0.879 (0.750-0.977)	0.833 (0.733-0.917)	0.750 (0.550-0.950)	0.875 (0.750-0.975)	0.750 (0.600-0.929)	0.875 (0.789-0.971)	0.630

Data are presented as metric values with 95% confidence intervals. MDT-1~6 indicate multidisciplinary diagnostic teams, each consisting of one radiologist and one pathologist; MDT-1–MDT-3 correspond to junior teams and MDT-4–MDT-6 to senior teams. P values are derived from DeLong tests comparing the AUC of each MDT (with or without BC-SMART assistance) against the AUC of the BC-SMART model in the corresponding cohort.

imaging requirements, or substantial computational demands. To address these barriers, BC-SMART adopts lightweight unimodal backbones and a transparent decision-level fusion strategy, aiming to preserve the benefits of multimodal integration while improving computational efficiency and clinical practicality. Consistent with prior radiopathomics findings,[24,29,30] BC-SMART delivered a clear and clinically meaningful performance gain over both single-modality backbones and the parsimonious clinical model. In the internal and external test cohorts, BC-SMART achieved AUCs of 0.903 (95% CI, 0.836–0.970) and 0.901 (95% CI, 0.847–0.956), respectively, and consistently ranked first across the overall metric profile. Importantly, the advantage was not confined to discrimination: BC-SMART showed a favorable "rule-out" profile with high NPVs (0.936 in the internal test cohort and 0.925 in the external test cohort), supporting its utility for identifying patients unlikely to achieve pCR who may benefit from closer monitoring or consideration of treatment adaptation. Compared with MRI-only and WSI-only models, the fused predictor also mitigated the sensitivity–specificity trade-off, suggesting that complementary macro- and micro-scale cues, combined with subtype-level clinicopathologic information, jointly enable more reliable pretreatment risk stratification in multicenter settings. Subgroup analyses further supported the robustness of BC-SMART, with generally maintained discrimination across molecular subtypes and key clinical strata.

To clarify where these gains originate, the performance and design rationale of the two unimodal backbones were explored, which constitute the foundation of the downstream decision-level fusion. From the perspective of MRI single-modality modeling, our results support SA-MRNet as a practical backbone for pretreatment pCR prediction. Its advantage was empirically confirmed in a comparison with representative CNN and transformer baselines (ResNet, DenseNet, ViT, and a vanilla Transformer). In the test cohorts, SA-MRNet consistently outperformed these alternatives, achieving the highest AUC in the internal test cohort (0.797) and again leading in the external test cohort (0.787), while also attaining the highest accuracy. We attribute this robustness to two design choices: slice-wise encoding that treats each volume as an ordered stack, improving tolerance to cross-center variation in slice thickness and coverage, and coordinate attention that emphasizes the enhancing tumor core and peritumoral regions while suppressing background noise. Conceptually, this design aligns with emerging "2.5D-style" strategies that add limited through-plane context on top of 2D backbones to approximate 3D capacity at substantially lower computational cost.[31-33]

On the pathology side, a similar pattern was observed for TMp-WSINet, an enhanced yet parameter-efficient TransMIL variant for biopsy WSIs, which consistently outperformed ResNet-, CLAM-, and baseline TransMIL-based models in both test cohorts. TMp-WSINet achieved the highest discrimination in both test cohorts (AUC=0.778 internally; AUC=0.820 externally), outperforming the competing backbones. In line with recent transformer-based multiple-instance learning approaches for WSIs, such as CLAM and TransMIL,[34,35] TMp-WSINet is designed to capture slide-level context under weak supervision while keeping the architecture compact. Building on the original TransMIL framework,[35] TMp-WSINet introduces learnable positional encoding to inject coarse spatial configuration, instance-level dropout to reduce over-reliance on a small subset of highly activated patches, and a gated-attention aggregator to recalibrate patch contributions at the slide level.[36] While these modifications may improve robustness to weak-label noise within sampled biopsy slides, they cannot fully compensate for biopsy-level sampling bias when the sampled tissue is not representative of the whole tumor. This limitation also highlights the value of MRI, which provides broader lesion-level and whole-tumor context beyond the sampled biopsy tissue. Taken together with the MRI results, these findings support SA-MRNet and TMp-WSINet as compact yet effective unimodal backbones that provide a solid foundation for subsequent multimodal fusion in BC-SMART.

In addition, in the multivariable clinical analysis, only HR and HER2 status remained independent predictors, consistent with prior evidence that subtype-level differences are strongly associated with chemosensitivity.[37,38] Accordingly, these two variables were retained to form a parsimonious, statistically selected clinical model. To provide a more clinically grounded comparison, we additionally constructed a clinically informed clinical model that retained clinical T and N stage together with HR and HER2 status, regardless of univariable significance. This clinically informed model showed modestly improved discrimination over the parsimonious clinical model, but both clinical

baselines remained clearly inferior to BC-SMART in the test cohorts. These findings suggest that clinicopathologic variables contribute useful subtype-level and anatomical tumor-burden information, but are insufficient on their own to capture the macro heterogeneity and microenvironmental complexity underlying pCR. When the HR/HER2-based clinical signature was further incorporated into the MRI–WSI decision-level fusion (ST-Fusion), multimodal performance showed a further upward trend, yielding BC-SMART, although the incremental gain over ST-Fusion did not reach statistical significance. Notably, although HR and HER2 status provided important subtype-related information associated with treatment response, substantial heterogeneity remained across molecular subgroups and neoadjuvant treatment patterns. In our cohorts, treatment regimens generally reflected routine subtype-guided clinical practice, with most HER2-positive patients receiving anti-HER2-based therapy. However, treatment response was not fully explained by molecular subtype or treatment strategy alone, as pCR rates varied across cohorts and subgroups. Although treatment regimens were not explicitly incorporated into the model, the consistent performance of BC-SMART across independent test cohorts suggests that MRI and WSI features provide complementary phenotypic information beyond conventional clinicopathologic factors. Collectively, these results indicate that clinical variables provide complementary prognostic context, whereas the dominant discriminative signal is captured by imaging and histopathology.

At the decision-level fusion stage, we employed a sparse logistic regression head based on only three scalar inputs (the MRI risk score, the WSI risk score, and a clinicopathologic score derived from HR/HER2 status), thereby avoiding reliance on high-capacity "black-box" fusion networks.[39] This design allows the relative contribution of each modality to the final pCR probability to be directly quantified through regression coefficients and SHAP decompositions, and confirms that the primary performance gains of the model are driven by WSI and MRI features. Within each imaging branch, Grad-CAM for MRI and attention heatmaps for WSIs provided intuitive spatial visualizations, showing that the networks predominantly focus on enhancing tumor cores, peritumoral regions, and histologic tumor–stroma interfaces that are clinically plausible correlates of treatment sensitivity. Together, these analyses establish a multi-level interpretability framework spanning modality-level contribution, spatially localized attention, and case-level attribution. Although such approaches do not provide mechanistic causality in a strict sense, they offer transparent and clinically meaningful attribution that can enhance clinician trust, facilitate error analysis and model refinement, and generate testable hypotheses for future studies linking radiopathologic phenotypes to underlying molecular pathways.[40,41]

We also investigated how BC-SMART functions as an assistive tool within multidisciplinary workflows. Our reader study demonstrated that BC-SMART alone achieved higher discrimination than any MDT, particularly the junior teams. Importantly, after MDTs were provided with BC-SMART probabilities and visual explanations, all teams demonstrated significant improvements in AUC, accompanied by favorable changes in accuracy and predictive values. These findings suggest that AI assistance provided complementary quantitative information that improved overall risk stratification while maintaining clinically relevant sensitivity. The confidence ratings further indicate that doctors generally perceived BC-SMART as a trustworthy aid, although the degree of trust appeared to depend on team experience, calibration, and familiarity with AI outputs.[42] To facilitate potential clinical translation, we implemented a prototype web-based interface integrating MRI and WSI visualization, model inference, and interpretability outputs, enabling interactive review of model predictions during MDT assessment. Collectively, these findings support BC-SMART as a complementary "second-reader" tool that may reduce performance variability across MDTs, particularly among less experienced clinicians, while preserving human oversight and allowing clinicians to interpret model outputs through attention maps and calibrated probabilities.[43,44]

This study has several limitations. First, our framework focuses on pretreatment prediction and does not incorporate on-treatment imaging or longitudinal clinical data, so its value for guiding dynamic treatment adaptation remains to be determined. Second, the WSI branch was developed using core-needle biopsy slides, which provide only partial sampling of the tumor and may not fully capture whole-tumor heterogeneity. Although TMp-WSINet includes mechanisms to improve robustness to weak-label noise within sampled slides, it cannot recover biologically relevant tumor or microenvironmental regions that were not captured by the biopsy specimen. Third, although attention maps and SHAP analysis suggested plausible radiologic-pathologic patterns, we did not perform molecular or immune profiling (e.g., transcriptomic or spatial omics analyses) to biologically validate the learned features. Fourth, although the reader-study cohort was designed to approximate the observed pCR prevalence of the overall cohort, the evaluation was still conducted in a selected retrospective subset rather than a consecutively enrolled prospective population. Fifth, although molecular subtype-specific analyses and treatment regimen distributions were summarized, treatment heterogeneity was not explicitly modeled, and residual confounding related to NAC regimens cannot be completely excluded. Furthermore, although center-wise clinical characteristics and treatment patterns were summarized, more extensive center-rotation validation strategies, such as leave-one-center-out validation, were not performed and should be explored in future multicenter studies. Finally, the utility of BC-SMART for prognostic stratification, long-term risk assessment, and prospective clinical deployment remains to be established. Future studies will focus on prospective validation in real-world clinical workflows, evaluation of prognostic implications, extension to cohorts receiving NAC combined with targeted or immunotherapy, and development of modality-incomplete learning strategies to better handle missing data in clinical practice.

CONCLUSION

In this study, BC-SMART provides a lightweight yet accurate multimodal solution for pretreatment pCR prediction by integrating routinely acquired MRI, biopsy WSIs, and minimal clinicopathologic variables. By pairing compact unimodal backbones with a transparent decision-level fusion, the framework showed consistent performance across the evaluated cohorts while offering interpretable, cross-scale evidence through modality attribution and spatial heatmaps. Importantly, BC-SMART can meaningfully augment multidisciplinary decision-making by improving discrimination and perceived confidence. Future work will focus on prospective, workflow-integrated validation and streamlined deployment across heterogeneous clinical sites to establish clinical utility and operational robustness in real-world practice.

Abbreviations

AUC Area under the curve
MDT Multidisciplinary teams
NAC Neoadjuvant chemotherapy
pCR Pathologic complete response
WSI Whole-slide image
HR Hormone receptor
HER2 Human epidermal growth factor receptor 2
H&E Hematoxylin and eosin
ROI Regions of interest
MIL Multiple-instance learning
AIC Akaike information criterion
TNBC Triple-negative breast cancer
SD Standard deviation
IQR Interquartile range
ROC Receiver operating characteristic
PR Precision–recall
PPV Positive predictive value
NPV Negative predictive value
CIs Confidence intervals
SHAP Shapley Additive Explanations

REFERENCES

1. Bray F., Laversanne M., Sung H., et al. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**:229–263. DOI:10.3322/caac.21834
2. Tamirisa N. and Hunt K.K. (2022). Neoadjuvant chemotherapy, endocrine therapy, and targeted therapy for breast cancer: ASCO guideline. *Ann. Surg. Oncol.* **29**:1489–1492. DOI:10.1245/s10434-021-11223-3

3. Derks M.G.M. and Van de Velde C.J.H. (2018). Neoadjuvant chemotherapy in breast cancer: More than just downsizing. *Lancet Oncol.* **19**:2–3. DOI:10.1016/S1470-2045(17)30914-2
4. Cortazar P., Zhang L., Untch M., et al. (2014). Pathological complete response and long-term clinical benefit in breast cancer: The CTNeoBC pooled analysis. *Lancet* **384**:164–172. DOI:10.1016/S0140-6736(13)62422-8
5. Boman C., Tranchell C., Liu X., et al. (2025). Prognosis after pathologic complete response to neoadjuvant therapy in early-stage breast cancer: A population-based study. *J. Natl. Compr. Canc. Netw.* **23**. DOI:10.6004/jnccn.2024.7093
6. Fayanju O.M., Ren Y., Thomas S.M., et al. (2018). The clinical significance of breast-only and node-only pathologic complete response (pCR) after neoadjuvant chemotherapy (NACT): A review of 20,000 breast cancer patients in the National Cancer Data Base (NCDB). *Ann. Surg.* **268**:591–601. DOI:10.1097/SLA.0000000000002953
7. Dong M., Chen J., Lu N., et al. (2025). Unraveling breast cancer response to neoadjuvant chemotherapy through integrated genomic, transcriptomic, and circulating tumor DNA analysis. *Breast Cancer Res.* **27**:64. DOI:10.1186/s13058-025-02026-5
8. Sammut S.J., Crispin-Ortuzar M., Chin S.-F., et al. (2022). Multi-omic machine learning predictor of breast cancer therapy response. *Nature* **601**:623–629. DOI:10.1038/s41586-021-04278-5
9. Kim J., Yu D., Kwon Y., et al. (2020). Genomic characteristics of triple-negative breast cancer nominate molecular subtypes that predict chemotherapy response. *Mol. Cancer Res.* **18**:253–263. DOI:10.1158/1541-7786.MCR-19-0453
10. Valenza C., Saldanha E.F., Gong Y., et al. (2025). Circulating tumor DNA clearance as a predictive biomarker of pathologic complete response in patients with solid tumors treated with neoadjuvant immune checkpoint inhibitors: A systematic review and meta-analysis. *Ann. Oncol.* **36**:726–736. DOI:10.1016/j.annonc.2025.03.019
11. Wei W., Ma M., and Liu Z. (2024). Research progress of artificial intelligence in evaluating the efficacy of neoadjuvant chemotherapy for breast cancer. *EngMedicine* **1**:100024. DOI:10.1016/j.engmed.2024.100024
12. Gradishar W.J., Moran M.S., Abraham J., et al. (2025). NCCN Guidelines® Insights: Breast Cancer, Version 5.2025. *J. Natl. Compr. Canc. Netw.* **23**:426-436. DOI:10.6004/jnccn.2025.0053
13. Mann R.M., Kuhl C.K., Kinkel K., et al. (2008). Breast MRI: Guidelines from the European society of breast imaging. *Eur. Radiol.* **18**:1307–1318. DOI:10.1007/s00330-008-0863-7
14. Huang Y., Wang X., Cao Y., et al. (2025). Nomogram for predicting neoadjuvant chemotherapy response in breast cancer using MRI-based intratumoral heterogeneity quantification. *Radiology* **315**:e241805. DOI:10.1148/radiol.241805
15. Tang W., Jin C., Kong Q., et al. (2025). Development and validation of an MRI spatiotemporal interaction model for early noninvasive prediction of neoadjuvant chemotherapy response in breast cancer: A multicentre study. *EClinicalMedicine* **85**:103298. DOI:10.1016/j.eclinm.2025.103298
16. Liang Y., Tang W., Kong Q., et al. (2025). Localized intra- and inter-tumoral heterogeneity for predicting treatment response to neoadjuvant chemotherapy in breast cancer. *IEEE J. Biomed. Health Inform.* **30**:6174–6185. DOI:10.1109/JBHI.2025.3647872
17. Kashyap A., Rapsomaniki M.A., Barros V., et al. (2022). Quantification of tumor heterogeneity: From data acquisition to metric generation. *Trends Biotechnol.* **40**:647–676. DOI:10.1016/j.tibtech.2021.11.006
18. Ogier du Terrail J., Leopold A., Joly C., et al. (2023). Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. *Nat. Med.* **29**:135–146. DOI:10.1038/s41591-022-02155-w
19. Yu J.-G., Wu Z., Ming Y., et al. (2023). Prototypical multiple instance learning for predicting lymph node metastasis of breast cancer from whole-slide pathological images. *Med. Image Anal.* **85**:102748. DOI:10.1016/j.media.2023.102748
20. Qu H., Zhou M., Yan Z., et al. (2021). Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning. *NPJ Precis. Oncol.* **5**:87. DOI:10.1038/s41698-021-00225-9
21. Feng L., Liu Z., Li C., et al. (2022). Development and validation of a radiopathomics model to predict pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: A multicentre observational study. *Lancet Digit. Health* **4**:e8–e17. DOI:10.1016/S2589-7500(21)00215-6
22. Wang J., Dai J., Cheng Y., et al. (2025). Advances in the use of radiomics and pathomics for predicting the efficacy of neoadjuvant therapy in tumors. *Transl. Oncol.* **58**:102435. DOI:10.1016/j.tranon.2025.102435
23. Zhang W., Zhang S., You J., et al. (2025). Attention-based multimodal fusion transformer for predicting the efficacy of neoadjuvant therapy in breast cancer: a cross-institutional retrospective study. *Breast Cancer Res.* **28**:4. DOI:10.1186/s13058-025-02181-9
24. Mao N., Dai Y., Zhou H., et al. (2025). A multimodal and fully automated system for prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer. *Sci. Adv.* **11**:eadr1576. DOI:10.1126/sciadv.adr1576
25. Zhao Z., Chen D., Wei X., et al. (2025). A multimodal synergistic model for personalized neoadjuvant immunochemotherapy in esophageal cancer. *Cell Rep. Med.* **6**:102479. DOI:10.1016/j.xcrm.2025.102479
26. Ghasemi A., Hashtarkhani S., Schwartz D.L., et al. (2024). Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review. *Cancer Innov.* **3**:e136. DOI:10.1002/cai2.136
27. Li J. and Jiang Z. (2024). Chinese society of clinical oncology breast cancer (CSCO BC) guidelines in 2024: International contributions from China. *Cancer Biol. Med.* **21**:838–843. DOI:10.20892/j.issn.2095-3941.2024.0374
28. Guo J., Chen B., Cao H., et al. (2024). Cross-modal deep learning model for predicting pathologic complete response to neoadjuvant chemotherapy in breast cancer. *NPJ Precis. Oncol.* **8**:189. DOI:10.1038/s41698-024-00678-8
29. Lin H., Hua J., Gong Z., et al. (2025). Multimodal radiopathological integration for prognosis and prediction of adjuvant chemotherapy benefit in resectable lung adenocarcinoma: a multicentre study. *Cancer Lett.* **616**:217557. DOI:10.1016/j.canlet.2025.217557
30. Hu C., Qiao X., Huang R., et al. (2024). Development and validation of a multimodality model based on whole-slide imaging and biparametric MRI for predicting postoperative biochemical recurrence in prostate cancer. *Radiol. Imaging Cancer* **6**:e230143. DOI:10.1148/rycan.230143
31. He J., Xu J., Chen W., et al. (2025). Development of a deep learning model for T1N0 gastric cancer diagnosis using 2.5D radiomic data in preoperative CT images. *NPJ Precis. Oncol.* **9**:249. DOI:10.1038/s41698-025-01055-9
32. Cen Y., Nong H., Du D., et al. (2025). CT-based 2.5D deep learning-multi-instance learning for predicting early recurrence of hepatocellular carcinoma and correlating with recurrence-related pathological indicators. *J. Hepatocell. Carcinoma* **12**:2095-2108. DOI:10.2147/JHC.S541402
33. Mai W., Fan X., Zhang L., et al. (2025). Deep learning for differential diagnosis of parotid tumors based on 2.5D magnetic resonance imaging. *Ann. Med.* **57**:2520401. DOI:10.1080/07853890.2025.2520401
34. Lu M.Y., Williamson D.F.K., Chen T.Y., et al. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**:555–570. DOI:10.1038/s41551-020-00682-w
35. Shao Z., Bian H., Chen Y., et al. (2021). TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. *Adv. Neural Inf. Process. Syst.* **34**:2136–2147. DOI:10.48550/arXiv.2106.00908
36. Ilse M., Tomczak J.M. and Welling M. (2018). Attention-based deep multiple instance learning. *Proceedings of the 35th International Conference on Machine Learning* **80**:2127-2136. https://proceedings.mlr.press/v80/ilse18a.html
37. Houssami N., Macaskill P., von Minckwitz G., et al. (2012). Meta-analysis of the association of breast cancer subtype and pathologic complete response to neoadjuvant chemotherapy. *Eur. J. Cancer* **48**:3342–3354. DOI:10.1016/j.ejca.2012.05.023
38. Asaoka M., Narui K., Suganuma N., et al. (2019). Clinical and pathological predictors of recurrence in breast cancer patients achieving pathological complete response to neoadjuvant chemotherapy. *Eur. J. Surg. Oncol.* **45**:2289–2294. DOI:10.1016/j.ejso.2019.08.001
39. Rudin C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**:206–215. DOI:10.1038/s42256-019-0048-x
40. Eke C. and Shuib L. (2025). The role of explainability and transparency in fostering trust in AI healthcare systems: A systematic literature review, open issues and potential solutions. *Neural Comput. Appl.* **37**:1999–2034. DOI:10.1007/s00521-024-10868-x
41. Tonekaboni S., Joshi S., McCradden M.D., et al. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. *Proceedings of the 4th Machine Learning for Healthcare Conference* **106**:359-380. https://proceedings.mlr.press/v106/tonekaboni19a.html
42. Gaube S., Suresh H., Raue M., et al. (2021). Do as AI say: Susceptibility in deployment of clinical decision-aids. *npj Digit. Med.* **4**:31. DOI:10.1038/s41746-021-00385-9
43. Topol E.J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **25**:44–56. DOI:10.1038/s41591-018-0300-7
44. Jha S. and Topol E.J. (2016). Adapting to artificial intelligence: Radiologists and pathologists as information specialists. *JAMA* **316**:2353–2354. DOI:10.1001/jama.2016.17438

FUNDING AND ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (grant No. 82302314); the Guangdong Special Support Program for Health Talents (grant No. 0920260252); the Basic and Applied Basic Research Foundation of Guangdong Province (grant Nos. 2022A1515110792, 2023A1515220097, and 2024A1515010653); and the Science and Technology Projects of Guangzhou (grant Nos. 2024A03J1030, 2025A03J4162, 2025A03J4163, and 2023A04J1275). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

W.T., Z.W., Y.C., and J.L. contributed equally to this work. W.T., Y.D., E.C., J.Y., and Y.G. conceived and designed the study. W.T., Z.W., Y.C., and J.L. performed data analysis, model development, figure preparation, and manuscript drafting. J.C., S.L., Y.L., S.C., R.L., Z.Z., L.P., and D.A. collected and curated clinical, imaging, and pathology data. L.Z., W.L., X.W., X.J., and Q.K. contributed to methodological discussion, data interpretation, and manuscript revision. W.T., W.L., Q.K., and Y.G. supervised the study and provided funding support. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

ETHICAL STATEMENT AND PATIENT CONSENT

This multicenter retrospective study was conducted in accordance with the Declaration of Helsinki and was approved by the institutional review boards or ethics committees of all participating centers: Guangzhou First People's Hospital (Center A; approval No. S-2023-083-01/-02/-03), Peking University Shenzhen Hospital (Center B; approval No. 2025-126), Jiangmen Central Hospital (Center C; approval No. 2025-292A), and the Third Affiliated Hospital of Sun Yat-sen University (Center D; approval No. A2025-695-01). Given the retrospective nature of the study and the use of de-identified clinical data, the requirement for written informed consent was waived by the ethics committees.

DATA AND CODE AVAILABILITY

De-identified data and analysis code supporting the findings of this study are available from the corresponding author, Yuan Guo, upon reasonable request. Raw imaging and pathology data are not publicly available due to patient privacy protection and institutional regulations.

SUPPLEMENTAL INFORMATION

It can be found online at https://doi.org/10.59717/j.xinn-oncol.2026.100042