Automated and reusable deep learning (AutoRDL) framework for predicting response to neoadjuvant chemotherapy and axillary lymph node metastasis in breast cancer using ultrasound images: a retrospective, multicentre study

Jingjing You,[a,f] Yue Huang,[b,f] Lizhu Ouyang,[c,f] Xiao Zhang,[d] Pei Chen,[a] Xuewei Wu,[a] Zhe Jin,[a] Hui Shen,[a] Lu Zhang,[a] Qiuying Chen,[a] Shufang Pei,[e,*] Bin Zhang,[a,**] and Shuixing Zhang[a,***]

[a]Department of Radiology, The First Affiliated Hospital of Jinan University, Guangzhou, Guangdong, China
[b]Department of Ultrasound, The First Affiliated Hospital of Kunming Medical University, Kunming, Yunnan, China
[c]Department of Ultrasound, Shunde Hospital of Southern Medical University, Foshan, Guangdong, China
[d]School of Information Science and Technology, Northwest University, Xi'an, China
[e]Department of Ultrasound, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangdong, China

eClinicalMedicine
2024;69: 102499

Published Online xxx
https://doi.org/10.1016/j.eclinm.2024.102499

Summary

Background Previous deep learning models have been proposed to predict the pathological complete response (pCR) and axillary lymph node metastasis (ALNM) in breast cancer. Yet, the models often leveraged multiple frameworks, required manual annotation, and discarded low-quality images. We aimed to develop an automated and reusable deep learning (AutoRDL) framework for tumor detection and prediction of pCR and ALNM using ultrasound images with diverse qualities.

Methods The AutoRDL framework includes a You Only Look Once version 5 (YOLOv5) network for tumor detection and a progressive multi-granularity (PMG) network for pCR and ALNM prediction. The training cohort and the internal validation cohort were recruited from Guangdong Provincial People's Hospital (GPPH) between November 2012 and May 2021. The two external validation cohorts were recruited from the First Affiliated Hospital of Kunming Medical University (KMUH), between January 2016 and December 2019, and Shunde Hospital of Southern Medical University (SHSMU) between January 2014 and July 2015. Prior to model training, super-resolution via iterative refinement (SR3) was employed to improve the spatial resolution of low-quality images from the KMUH. We developed three models for predicting pCR and ALNM: a clinical model using multivariable logistic regression analysis, an image model utilizing the PMG network, and a combined model that integrates both clinical and image data using the PMG network.

Findings The YOLOv5 network demonstrated excellent accuracy in tumor detection, achieving average precisions of 0.880–0.921 during validation. In terms of pCR prediction, the combined model$_{\text{post-SR3}}$ outperformed the combined model$_{\text{pre-SR3}}$, image model$_{\text{post-SR3}}$, image model$_{\text{pre-SR3}}$, and clinical model (AUC: 0.833 vs 0.822 vs 0.806 vs 0.790 vs 0.712, all $p < 0.05$) in the external validation cohort (KMUH). Consistently, the combined model$_{\text{post-SR3}}$ exhibited the highest accuracy in ALNM prediction, surpassing the combined model$_{\text{pre-SR3}}$, image model$_{\text{post-SR3}}$, image model$_{\text{pre-SR3}}$, and clinical model (AUC: 0.825 vs 0.806 vs 0.802 vs 0.787 vs 0.703, all $p < 0.05$) in the external validation cohort 1 (KMUH). In the external validation cohort 2 (SHSMU), the combined model also showed superiority over the clinical and image models (0.819 vs 0.712 vs 0.806, both $p < 0.05$).

Interpretation Our proposed AutoRDL framework is feasible in automatically predicting pCR and ALNM in real-world settings, which has the potential to assist clinicians in optimizing individualized treatment options for patients.

Funding National Key Research and Development Program of China (2023YFF1204600); National Natural Science Foundation of China (82227802, 82302306); Clinical Frontier Technology Program of the First Affiliated Hospital of Jinan University, China (JNU1AF-CFTP-2022-a01201); Science and Technology Projects in Guangzhou

*Corresponding author. Department of Ultrasound, 106 Zhongshan 2nd Road, Yuexiu District, Guangzhou 510180, China.
**Corresponding author. Department of Radiology, 613 Huangpu West Road, Tianhe District, Guangzhou 510630, China.
***Corresponding author. Department of Radiology, 613 Huangpu West Road, Tianhe District, Guangzhou 510630, China.
E-mail addresses: peishufang@gdph.org.cn (S. Pei), xld_Jane_Eyre@126.com (B. Zhang), shui7515@126.com (S. Zhang).
[f]The authors contributed equally to this work.

(202201020022, 2023A03J1036, 2023A03J1038); Science and Technology Youth Talent Nurturing Program of Jinan University (21623209); and Postdoctoral Science Foundation of China (2022M721349).

Keywords: Breast cancer; Ultrasound; Deep learning; Neoadjuvant chemotherapy; Lymph node metastasis

Research in context

Evidence before this study

We systematically searched PubMed for articles published in English with the following terms: ("radiomics" OR "deep learning" OR "artificial intelligence" OR "AI") AND ("breast cancer") AND ("axillary lymph node" OR "ALN" OR ["response" AND "neoadjuvant chemotherapy" OR "NAC"]) from the inception of the database to August 8, 2023. Our findings indicated that previous studies primarily focused on the analysis of MRI data, often developed task-specific deep learning models, heavily relied on manual segmentation, and excluded low-quality images. These factors could potentially limit the reproductivity and clinical applicability of their findings in real-world settings.

Added value of this study

This study proposed an automated and reusable deep learning (AutoRDL) framework that allows for the autonomous detection of tumors and prediction of pCR and ALNM. Additionally, it introduces a super-resolution reconstruction scheme to enhance the spatial resolution of images with low-resolution, thereby improving the predictive performance of our image-based deep learning model. Also, gradient-weighted class activation mapping saliency was employed to quantify pathways contribution to individual AutoRDL decisions.

Implications of all the available evidence

Our findings provide compelling evidence that the ultrasound-based AutoRDL framework, as an automated, reusable, and interpretable approach, has the potential to improve tumor detection and different classification tasks. It exhibits robust performance that can significantly improve clinical decision-making and operative planning. Confirmation of these findings through prospective validation cohort studies will further strengthen the evidence base for the predictive performance of our fully automated AI tool in clinical practice.

Introduction

Predicting the response to neoadjuvant chemotherapy (NAC) and determining the axillary lymph node (ALN) status before surgery are two critical elements in the realm of precision medicine for breast cancer.[1,2] Predicting pathological complete response (pCR) is significant for identifying candidates suitable for more limited operations such as breast-conserving surgery even eliminating invasive surgery after NAC.[3,4] Preoperative identification of ALN status in breast cancer also holds great importance in disease management, surgical decision-making, and prognosis evaluation.[5] Histopathological examination is still the gold standard for assessing pCR and ALN status in breast cancer. Despite this, the quest for a non-invasive method to accurately predict these outcomes presents a formidable challenge.

Deep learning approaches demonstrate the ability to decode underlying information contained within images noninvasively and have been extensively leveraged in different tasks such as image segmentation and classification.[6–8] In recent years, several studies have explored the feasibility of image-based deep learning models for assessing axillary lymph node metastasis (ALNM) and pCR.[9,10] However, most of these studies have predominantly focused on MRI data, with limited emphasis on ultrasound images. Moreover, several significant hurdles hinder the clinical translation of deep learning models into clinically useful tests for predicting pCR and ALNM. First, these models are largely dependent on manual segmentation, which is a time-consuming process and may introduce observer variability.[11–13] Second, the image quality can vary across different centers due to variations in equipment, parameter settings, and operators involved in image acquisition.[14] To attain high model performance, previous studies often excluded images with inadequate quality and concentrated solely on using high-quality images for training their models.[11–13,15] Third, previous studies have developed a wide range of neural networks for the prediction of pCR and ALNM.[16–18] Designing a reusable deep learning architecture for similar tasks in a specific disease has the potential to reduce computational resources and redundant work required in developing new models. Additionally, it promotes efficiency, collaboration, and knowledge sharing within the deep learning community.

In this work, we aimed to propose a fully automated and reusable deep learning (AutoRDL) framework for end-to-end tumor detection and prediction of ALNM and pCR using a multicentre ultrasound dataset with various image qualities. This noninvasive approach may facilitate a decision regarding breast-conserving surgery

or even omitting surgery for patients who achieve a pCR and spare patients with negative ALNM from invasive surgical procedures.

Methods

Ethics

The study protocol was approved by the ethics committee of Guangdong Provincial People's Hospital (approval number: KY2023-318-02) and was endorsed by the participating hospitals, with a waiver for informed consent from patients due to the retrospective nature of the work. We conducted the study in compliance with STARD-2015 guidelines (equator-network.org).

Patients and data collection

A retrospective screening was conducted on the clinicopathological and ultrasound data of consecutive female patients diagnosed with breast cancer at three tertiary hospitals, including Guangdong Provincial People's Hospital (GPPH), the First Affiliated Hospital of Kunming Medical University (KMUH), and Shunde Hospital of Southern Medical University (SHSMU) between November 2012 and May 2021. Patients who had baseline two-dimensional ultrasound images were eligible for inclusion in the tumor detection task; however, patients with non-mass, invisible lesions, and lesions with unmatched location and size between the ultrasound image and pathological examination were excluded. A full list of the inclusion and exclusion criteria for the prediction of pCR and ALNM is provided in the Supplementary file. Fig. 1 depicts the patient recruitment pathway and the division of the cohort. Finally, a total of 2556 patients (2632 lesions) were included in the tumor detection task, with 1409 patients (1409 lesions) and 792 patients (792 lesions) included in the prediction tasks of pCR and ALNM, respectively. Patients recruited from GPPH were used as a development cohort, while those recruited from KMUH and SHSMU were assigned to two external validation cohorts. All images downloaded from the workstations were converted into a JPEG format. The images were generated from 10 different ultrasound scanners (Supplementary Table S1), with linear transducers of a frequency range of 5–14 MHz.

After consulting with physicians who specialize in breast cancer and reviewing recent literature on the risk factors relevant to pCR and ALNM, clinicopathological data were collected from medical records, including age, clinical T stage, tumor laterality, tumor location, pathological type, estrogen receptor (ER) status, progesterone receptor (PR) status, human epidermal growth factor receptor 2 (HER-2) status, and Ki-67 proliferation index. Clinical T stage, tumor laterality, and tumor location were obtained from the imaging report. Nuclear staining of ER/PR by immunohistochemistry (IHC) with ≥1% positive tumor cells was defined as ER/PR positive, while staining with <1% positive tumor cells was defined as ER/PR negative.[19] HER-2 was considered positive if IHC resulted in a score of 3+ or 2+ with amplification confirmed by fluorescence in situ hybridization (FISH). HER-2 negative was defined as an IHC score of 0 or 1+ or 2+ with non-amplified FISH.[20] Depending on the receptor status, all the patients were categorized into three subtypes depending on receptor status as follows: (i) TN (triple-negative); (ii) HER2+; (iii) HR+/HER2–. The Ki-67 proliferation index was considered high if it was ≥20%.

Evaluation of NAC response and ALN status

Pathological assessments of NAC response and ALN status were conducted through standard histopathological examinations. pCR was defined as the absence of residual invasive tumor cells in both breast specimens and ipsilateral axillary lymph nodes, regardless of the presence of residual ductal carcinoma in situ (ypT0/is ypN0).[21] ALN status was determined based on the histopathological reports of SLND and ALND, in accordance with the American Joint Committee on Cancer's Staging System for Breast Cancer.[22]

Image preprocessing

The image quality of cases in the KMUH cohort was significantly poorer than that of cases in the GPPH and SHSMU cohorts (average image resolution: 262 × 370, 625 × 847, 571 × 763 pixels, respectively). Therefore, a state-of-the-art image super-resolution via iterative refinement (SR3) technique was trained on the images from the GPPH and SHSMU cohorts to enhance the resolution of images from the KMUH cohort to 512 × 512 pixels.[23] The backbone network used in this study was the U-Net encoder-decoder architecture, in which the original convolution operations were replaced with residual blocks (Supplementary Fig. S1). Further details are described in the Supplementary file.

To obtain the ground truth, a graphical image annotation tool (Python labelImg 1.8.6, https://pypi.org/project/labelImg/) was used to manually label the rectangular region of interest (ROI) on the ultrasound images of the maximum cross-section. The labeling was carried out by a trained radiologist with 5 years of experience. The labeled images were then reviewed and confirmed by a senior radiologist with 20 years of experience.

To overcome the need for a large amount of training data, data augmentation was performed using the NVIDIA MONAI framework. This augmentation process effectively increased the size of the dataset, providing more diverse samples for training the deep learning model. By applying a range of transformations such as flipping, rotation, scaling, cropping, translation, Gaussian noise, and color jitter, the amount of data in the training cohort was in a seven-fold increase. The generalization ability of the deep learning model could be improved, thereby alleviating the risk of overfitting.

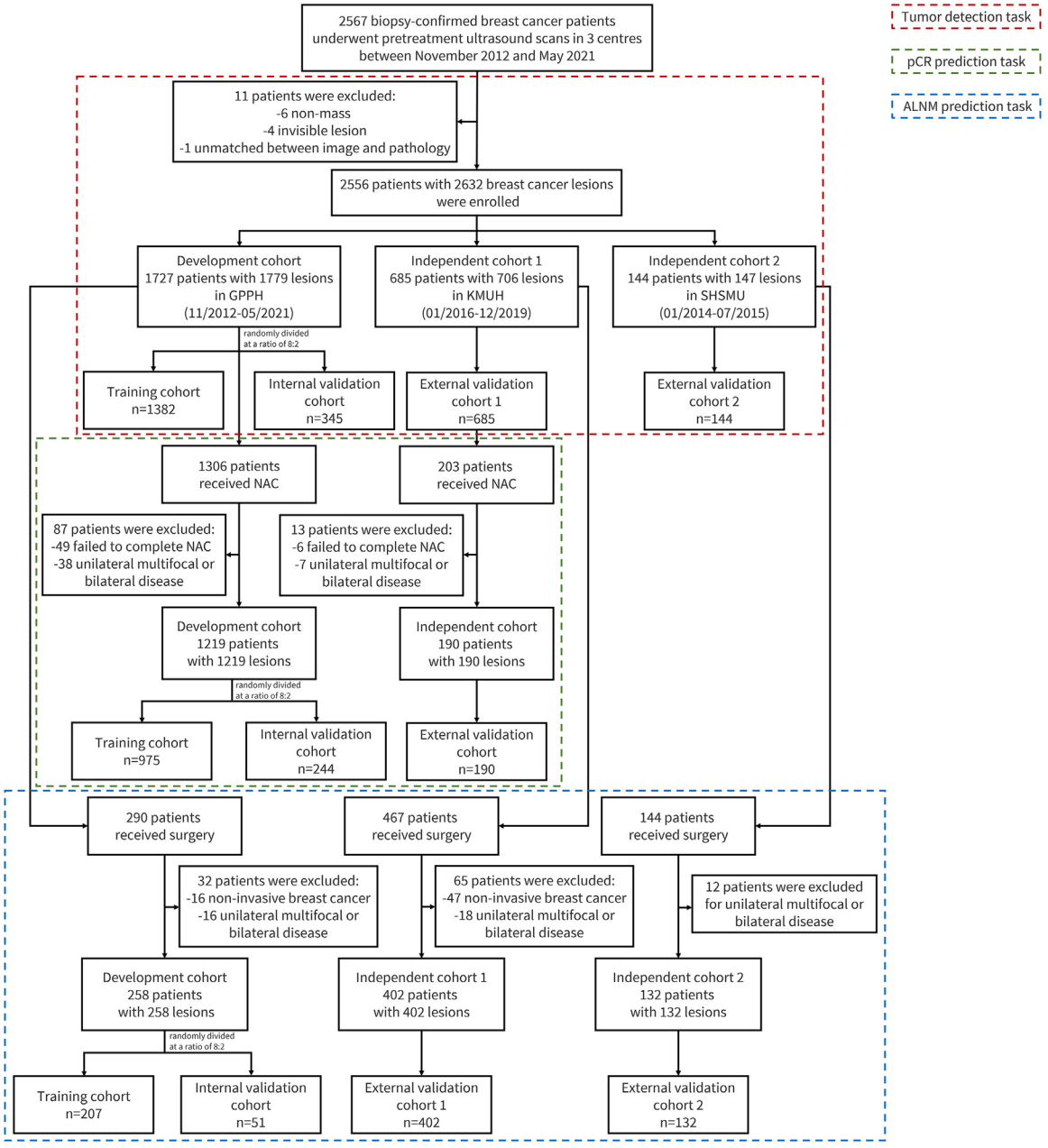

Fig. 1: **Flowchart of study enrollment.** In this multicentre, retrospective study, patients were consecutively enrolled from three independent institutions (GPPH, Guangdong Provincial People's Hospital; KMUH, The First Affiliated Hospital of Kunming Medical University; SHSMU, Shunde Hospital of Southern Medical University) and assigned to the training and two external validation cohorts, respectively. Red, green, and blue frames indicate the enrollment process for tumor detection, pCR prediction, and ALNM prediction tasks, respectively. Abbreviations: NAC, neoadjuvant chemotherapy; pCR, pathological complete response; ALNM, axillary lymph node metastasis.

Development of an AutoRDL framework

To improve efficiency, minimize manual input, and seamlessly integrate into the routine clinical workflow, an AutoRDL framework was developed for tumor detection and prediction of pCR and ALNM. First, a tumor detection network was established to automatically localize tumor regions. Upon successful detection, specific target areas were cropped from the images and used as input for the prediction network. Notably, we pertained the prediction network on the pCR task and then fine-tuned it on the ALNM task. Supplementary Fig. S2 illustrates an overall design of

the proposed AutoRDL framework. The implementation of the deep learning framework is available at: https://github.com/code202308/ALNM_pCR. Convergence curves during the training process are shown in Supplementary Fig. S3.

Automatic tumor detection

A state-of-the-art deep learning network called You Only Look Once version 5 (YOLOv5) was employed as the backbone to detect tumors in a dataset consisting of 2556 patients with 2632 lesions. After obtaining the detection results, the ultrasound images were subjected to a cropping process. This process aimed to eliminate irrelevant data and retain only the essential information for further analysis. Manual delineation was considered as the ground truth for the model.

pCR prediction

The size of the target regions, referring to the ROIs where tumors were present, varied significantly. Thus, a substantial number of tumor sample blocks with diverse sizes were generated. Meanwhile, the implicitly contained information within these tumor sample blocks could be varied. To fully exploit the crucial information within these sample blocks and prevent the loss of image details, a progressive multi-granularity (PMG) classification network was employed to develop the prediction model, i.e., image model.[24] The details about the model development are provided in the Supplementary file.

ALNM prediction

Since the prediction of pCR and ALNM shared similar profiles as classification tasks, we decided to utilize the same network architecture that was used in the pCR prediction task to develop the prediction model for the ALNM task.

Integration with clinicopathologic data

Clinicopathologic data including age, tumor size, pathological type, ER, PR, HER-2, and Ki-67, were used to construct the clinical model for predicting pCR and ALNM via multivariable logistic regression analysis. Subsequently, these clinical variables were incorporated into a fully connected layer of the image model to establish the combined model.

Benchmarking against previous AI models

To illustrate the advantage of our AutoRDL framework, we conducted elaborately comparative studies by benchmarking it against other previously published AI methods, including U-Net with DenseNet-121,[16] ResNet-50,[25] and DenseNet-201[13] for the pCR prediction task, as well as Mask R–CNN with DenseNet-121,[26] ResNet-50[18] and Inception V3[12] for the ALNM prediction task.

Statistics

The statistical analyses were performed with SPSS (version 23.0) and Python (version 3.7). The peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM) were used to evaluate the quality of the reconstructed images. The mean average precision (mAP) was used to assess the tumor detection performance. We calculated the area under the curve (AUC), sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), confusion matrix, calibration curve, and decision curve analysis (DCA) to comprehensively evaluate the model performance from multiple dimensions of discrimination, calibration, and clinical usefulness. Differences in AUCs between the models were compared using the DeLong test. A two-tailed $p < 0.05$ was considered statistically significant.

We utilized local interpretable model-agnostic explanations (LIME)[27] and gradient-weighted class activation mapping (Grad-CAM)[28] techniques to interpret the predictions generated by AutoRDL. LIME generates explanations that highlight the significant features and their contributions to the model's prediction at a local level. To facilitate visualization and analysis, we firstly conducted dimensionality reduction on the 64-dimensional deep features obtained from the neural networks for both the pCR and ALNM tasks. We utilized the Principal Component Analysis algorithm to further reduce the dimensionality of these deep features to eight dimensions for each network. Subsequently, we concatenated these reduced features with their corresponding clinical features and conducted feature importance analysis using LIME. Grad-CAM leverages the gradients flowing into the last convolutional layer of the network. These gradients indicate the importance of each feature map in the layer concerning the final prediction. By calculating the weighted sum of these gradients, Grad-CAM generates a heatmap that highlights the image regions most relevant to the predicted class. The heatmap produced by Grad-CAM can be superimposed on the original image, enabling us to visually identify the discriminative regions that the network focuses on when making predictions. This provides valuable insights into the decision-making process of the network and aids in interpreting and explaining its predictions.

Role of the funding source

The funders had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Study design and patient characteristics

The overall study design is shown in Fig. 2. We trained and independently validated an AutoRDL framework that used ultrasound images to detect tumors and

independently predict pCR and ALNM. A total of 2556 patients were enrolled in the automated tumor detection task. Among them, there were 1382 patients from GPPH, 345 patients from GPPH, 685 patients from KMUH, and 144 patients from SHSMU, distributed across the training, internal validation, and two external validation cohorts, respectively. The pCR prediction task included 1409 patients (mean age, 49.2 ± 10.3 years), of which, 975 and 244 patients from GPPH, and 190 patients from KMUH were allocated to the training, internal validation, and external validation cohorts, respectively. The ALNM prediction task consisted of 792 patients (mean age, 51.9 ± 12.0 years), with 207 and 51 patients from GPPH for the training and internal validation cohorts while 402 and 132 patients from KMUH and SHSMU for two external validation cohorts. Clinicopathological data of the pCR and ALNM prediction tasks are summarized in Supplementary Tables S2 and S3. Baseline characteristics are well balanced between the training and internal validation cohort except for the clinical T stage in the pCR prediction task.

Performance of image super-resolution reconstruction

Image super-resolution via iterative refinement achieved a PSNR of 28.11 (95% confidence interval [CI]: 27.28–28.99) and an SSIM of 0.93 (95%CI: 0.91–0.94) on average, indicting significant improvement of image quality. Supplementary Fig. S4 shows the visual results of the image super-resolution reconstruction for the randomly selected cases.

Accuracy of the tumor detection

The YOLOv5 network demonstrated excellent detection performance among all cohorts, with mAP values of 0.978 (95%CI: 0.965–0.988), 0.921 (95%CI: 0.904–0.936), 0.894 (95%CI: 0.879–0.908), and 0.880 (95%CI: 0.863–0.892) in the training, internal validation, and two external validation cohorts, respectively. The results suggested that the YOLOv5 network had an effective tumor detection ability, even under the variations in tumor sizes and locations. Supplementary Fig. S5 displays representative examples of automatic tumor detection.

Performance of the pCR and ALNM prediction

Tables 1 and 2 provide the performances of the clinical model, image model, and combined model for the prediction of pCR and ALNM in the training and validation cohorts. For the prediction of pCR, the combined model achieved a training AUC of 0.996 (95%CI: 0.984–0.998), while the clinical and image models yielded training AUCs of 0.882 (95%CI: 0.869–0.894) and 0.978 (95%CI: 0.966–0.989), respectively. The combined model outperformed both the clinical model (0.851 vs 0.738, $p < 0.0005$) and image model (0.851 vs 0.811, $p = 0.001$) in the internal validation cohort. For the external validation cohort (KMUH cohort), the predictive accuracy of the image $model_{post\text{-}SR3}$ was statistically higher than that of the image $model_{pre\text{-}SR3}$ (AUC: 0.806 vs 0.790, $p = 0.0027$). Furthermore, the combined $model_{post\text{-}SR3}$ (AUC = 0.833) was superior to the combined $model_{pre\text{-}SR3}$ (0.822, $p = 0.0042$), image $model_{post\text{-}SR3}$ (0.806, $p = 0.0055$), image $model_{pre\text{-}SR3}$ (0.790, $p = 0.0014$), as well as clinical model (0.712, $p = 0.00029$). In addition, the combined $model_{post\text{-}SR3}$ demonstrated overwhelmingly higher metrics such as sensitivity, specificity, PPV, and NPV compared to other models. Fig. 3 displays the ROC curves, AUC values, confusion matrices, calibration curves, and DCA curves of the clinical, image, and combined models. The comparative results of the models are shown in Supplementary Tables S4 and S5.

In the ALNM prediction task, likewise, the optimal predictive model achieved the highest accuracy across the training, internal validation, and two external validation cohorts, with AUCs of 0.960 (95%CI: 0.948–0.971), 0.856 (95%CI: 0.842–0.868), 0.825 (95% CI: 0.814–0.842), and 0.819 (95%CI: 0.806–0.835), respectively. The accuracy of the image model in the external validation cohort 1 (KMUH cohort) could be improved by the super-resolution reconstruction scheme (AUC: 0.802 vs 0.787, $p = 0.0054$). The AUC of the combined $model_{post\text{-}SR3}$ was higher than that of the combined $model_{pre\text{-}SR3}$, image $model_{post\text{-}SR3}$, and clinical model (0.825 vs 0.806 vs 0.802 vs 0.703, all $p < 0.05$). In the external validation cohort 2 (SHSMU cohort), the combined model was superior to those of the clinical model (AUC: 0.819 vs 0.712, $p < 0.0005$) and the image model (AUC: 0.819 vs 0.806, $p < 0.005$). Furthermore, the image model also demonstrated higher accuracy compared to the clinical model in validation cohorts (all $p < 0.005$). The ROC curves, AUC values, confusion matrices, calibration curves, and DCA curves of the predictive models are illustrated in Fig. 4, and the performance comparison of different models are summarized in Supplementary Tables S6–S8.

We conducted a visualization using the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm to compare the classification effects of clinical, image, clinical-image combined features. The results suggested that the combined features exhibited superior characterization ability (Supplementary Fig. S6).

Interpretation of the AutoRDL framework

To enhance the interpretability of the AutoRDL, the LIME explainer was applied, with the features for each case presented in Supplementary Fig. S7. We also generated activation maps and selected a total of 12 random example explanations for different prediction results in each task individually. The heatmaps highlighted the central region of the tumor in the positive cases (i.e., pCR and ALNM) while the tumor boundary was highlighted in the negative cases (i.e., non-pCR and non-ALNM) (Fig. 5).

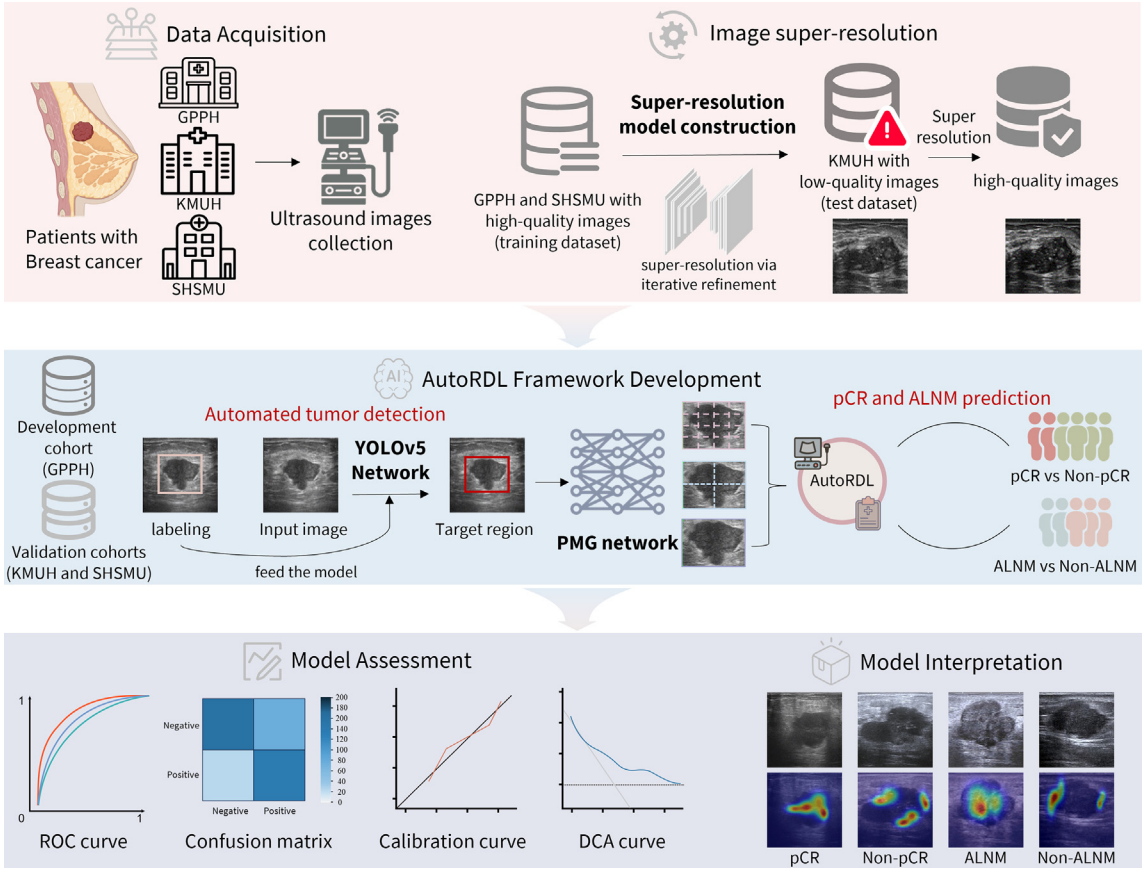

***Fig. 2:* Overview of the study design.** The clinicopathologic and ultrasound data of patients were collected from the GPPH, KMUH, and SHSM. Thereafter, an image super-resolution via iterative refinement (SR3) model was trained on the GPPH and SHSMU with high-quality images to enhance the resolution of images from KMUH with low image quality. Our proposed AutoRDL framework contained two subnetworks: YOLOv5 network for automated tumor detection and PMG network for the prediction of pCR and ALNM. Model performance was assessed using the ROC curve, confusion matrix, calibration curve, and DCA curve. Finally, individual decisions made by the AutoRDL framework were visualized and interpreted by Grad-CAM. Abbreviations: GPPH, Guangdong Provincial People's Hospital; KMUH, First Affiliated Hospital of Kunming Medical University; SHSM, Shunde Hospital of Southern Medical University; AutoRDL, automated and reusable deep learning; YOLOv5, You Only Look Once version 5; PMG, progressive multi-granularity; pCR, pathological complete response; ALNM, axillary lymph node metastasis; ROC, receiver operator characteristic; DCA, decision curve analysis; Grad-CAM, gradient-weighted class activation mapping.

Performance comparison with previous AI models

Compared with the segmentation networks Mask R–CNN and U-Net, our YOLOv5 showed comparable or slightly higher mAP values, particularly in the KMUH cohort (0.894 vs 0.869 vs 0.854). Detailed performance comparisons of the tumor detection networks are presented in Supplementary Table S9.

As for the pCR prediction task using the combined model, our AutoRDL framework outperformed the U-Net with DenseNet-121, ResNet-50, and DenseNet-201 in all the validation cohorts (all $p < 0.05$). For the ALNM prediction task, our framework also showed superior prediction performance compared to ResNet-50, Mask R–CNN with DenseNet-121, and Inception V3 in the validation cohorts (all $p < 0.05$). Detailed comparative results are shown in Supplementary Tables S10–S13.

Discussion

In this multicentre study, for the first time, we developed an AutoRDL framework that allows the automated tumor detection as well as prediction of pCR and ALNM from various image quality ultrasound images in patients with breast cancer. The AutoRDL framework showed excellent performance in three tasks, which consist of tumor detection (mAPs: 0.921, 0.894, and 0.880) as well as prediction of pCR (AUCs: 0.851 and 0.833) and ALNM (AUCs: 0.856, 0.825, and 0.819) in both the internal and external validation cohorts. The comparative analysis demonstrated that our AutoRDL framework outperformed other state-of-the-art deep learning models that have already been published. Notably, the predictive accuracy of the image models based on low-quality images could be improved by

Models	AUC	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Training cohort					
Clinical model	0.882 (0.869–0.894)	86.0 (84.2–88.3)	81.8 (80.5–84.7)	77.0 (75.4–78.8)	88.1 (86.2–89.4)
Image model	0.978 (0.966–0.989)	94.8 (93.3–96.2)	92.1 (90.7–93.9)	91.7 (90.0–93.1)	96.5 (94.0–97.8)
Combined model	0.996 (0.984–0.998)	98.0 (96.5–99.0)	91.8 (90.4–93.5)	92.5 (91.3–94.0)	97.2 (95.8–98.4)
Internal validation cohort					
Clinical model	0.738 (0.719–0.752)	69.2 (67.3–72.6)	71.0 (69.7–73.2)	65.3 (63.9–67.0)	74.9 (73.3–76.7)
Image model	0.811 (0.797–0.835)	78.5 (77.1–80.0)	76.6 (74.9–78.4)	71.6 (70.2–73.0)	84.0 (82.7–85.9)
Combined model	0.851 (0.828–0.865)	84.0 (81.6–85.1)	77.1 (74.9–78.3)	73.5 (72.1–75.4)	97.2 (95.8–98.4)
External validation cohort					
Clinical model	0.712 (0.701–0.729)	66.0 (64.1–67.7)	71.8 (68.6–73.1)	63.3 (61.8–64.6)	72.5 (70.6–74.3)
Image model$_{pre\text{-}SR3}$	0.790 (0.778–0.805)	75.4 (73.6–76.9)	72.3 (70.7–74.0)	68.1 (66.4–69.8)	81.1 (79.8–82.7)
Image model$_{post\text{-}SR3}$	0.806 (0.798–0.824)	77.0 (75.2–78.9)	73.8 (72.2–75.1)	69.7 (68.0–71.5)	82.9 (81.8–84.3)
Combined model$_{pre\text{-}SR3}$	0.822 (0.805–0.831)	79.0 (77.8–80.3)	73.2 (71.9–74.6)	71.3 (69.8–72.9)	83.0 (81.8–84.4)
Combined model$_{post\text{-}SR3}$	0.833 (0.820–0.847)	81.2 (79.5–82.4)	75.8 (74.0–77.1)	72.0 (70.5–74.1)	87.3 (85.4–88.9)

Note: Data in parentheses are the 95% confidence interval. pCR, pathological complete response; AUC, area under the curve; PPV, positive predictive value; NPV, negative predictive value; SR3, image super-resolution via iterative refinement.

Table 1: **Performances of the clinical model, image model, and combined model for the prediction of pCR in the training and validation cohorts.**

increasing resolution via super-resolution. Further, the combination of images with existing clinical biomarkers led to a significant improvement in prediction accuracy. AutoRDL represents a crucial step towards integrating artificial intelligence into clinical practice to better inform clinical decision-making and operative planning.

Artificial intelligence techniques have been employed to capture high-dimensional characteristics of tumors, facilitating accurate predictions of the response to NAC.[29] In imaging-based research, previous studies have mainly focused on the analysis of MRI data, with only a few studies utilizing ultrasound images as the basis for analysis.[10,25,30–33] Jiang et al.[13] constructed a deep learning radiomic nomogram to predict pCR in patients with locally advanced breast cancer by integrating pre- and post-treatment ultrasound data. Wu et al.[17] also developed a deep learning model integrating pre-, early-stage, and post-treatment images for the prediction of pCR. However, there was a time lag in acquiring post-treatment images since they were obtained after NAC.

Models	AUC	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Training cohort					
Clinical model	0.813 (0.798–0.832)	81.7 (80.5–82.8)	83.6 (82.4–85.1)	73.0 (71.7–74.8)	84.6 (83.3–86.1)
Image model	0.928 (0.916–0.937)	90.3 (88.9–91.8)	97.4 (95.8–98.6)	86.7 (85.4–88.5)	88.3 (86.4–89.2)
Combined model	0.960 (0.948–0.971)	92.2 (90.5–93.5)	92.3 (91.1–93.4)	88.7 (87.2–90.8)	94.8 (93.7–96.5)
Internal validation cohort					
Clinical model	0.738 (0.726–0.764)	77.0 (75.8–79.5)	67.5 (65.0–70.2)	68.4 (67.5–71.0)	78.5 (76.0–79.6)
Image model	0.810 (0.788–0.819)	73.4 (71.7–75.0)	75.5 (74.0–76.7)	74.2 (72.7–75.6)	80.1 (78.8–81.4)
Combined model	0.856 (0.842–0.868)	81.9 (80.7–83.0)	74.8 (73.2–76.1)	74.3 (72.9–75.6)	81.5 (79.2–83.4)
External validation cohort 1					
Clinical model	0.703 (0.687–0.720)	67.8 (66.0–69.2)	65.7 (64.3–66.8)	66.5 (65.3–67.9)	74.4 (73.1–76.2)
Image model$_{pre\text{-}SR3}$	0.787 (0.771–0.805)	71.2 (69.5–72.8)	67.4 (65.8–69.2)	66.9 (65.4–68.7)	77.3 (75.7–78.8)
Image model$_{post\text{-}SR3}$	0.802 (0.785–0.818)	72.9 (71.5–74.2)	69.1 (67.9–70.5)	68.5 (66.9–70.1)	79.0 (77.5–80.1)
Combined model$_{pre\text{-}SR3}$	0.806 (0.790–0.821)	75.2 (74.0–76.6)	70.5 (69.0–72.1)	68.9 (67.4–70.5)	80.1 (78.8–81.3)
Combined model$_{post\text{-}SR3}$	0.825 (0.814–0.842)	77.1 (75.8–78.2)	72.3 (70.8–73.7)	71.0 (70.1–72.2)	81.4 (79.6–82.9)
External validation cohort 2					
Clinical model	0.712 (0.698–0.720)	69.3 (67.8–71.2)	70.2 (68.8–72.0)	67.6 (65.9–69.8)	76.2 (74.5–78.8)
Image model	0.806 (0.787–0.816)	73.7 (72.1–75.4)	75.6 (74.3–76.7)	71.2 (69.4–73.3)	80.6 (78.0–82.7)
Combined model	0.819 (0.806–0.835)	75.6 (74.0–76.9)	75.5 (74.1–77.6)	73.4 (71.8–74.6)	80.6 (79.8–82.3)

Note: Data in parentheses are the 95% confidence interval. ALNM, axillary lymph node metastasis; AUC, area under the curve; PPV, positive predictive value; NPV, negative predictive value; SR3, image super-resolution via iterative refinement.

Table 2: **Performances of the clinical model, image model, and combined model for the prediction of ALNM in the training and validation cohorts.**

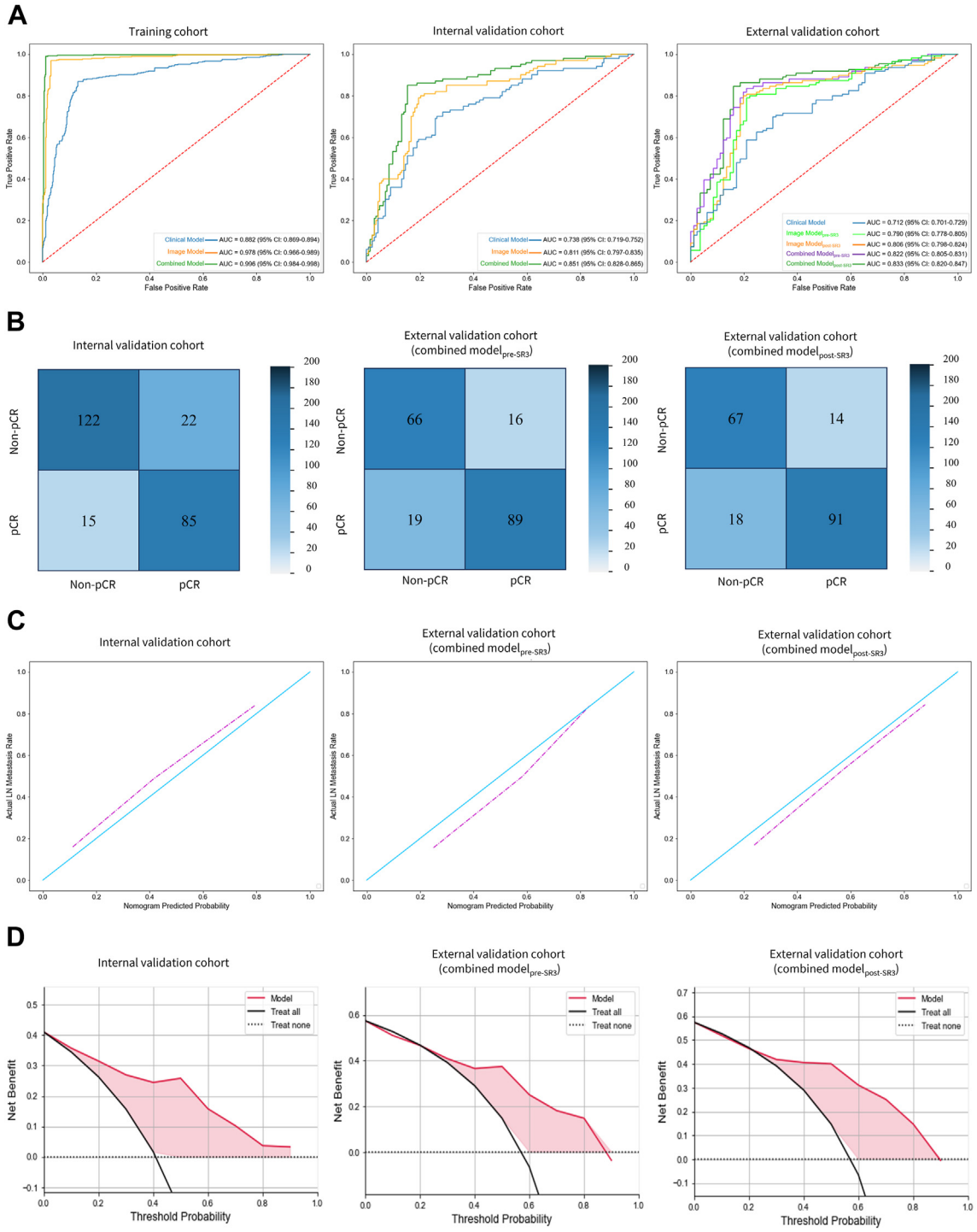

***Fig. 3*: Performance of the clinical, image, and combined models in predicting pCR.** (A): ROC curves of the clinical, image, and combined models for predicting pCR in the training, internal validation, and external validation cohorts. Specially, in the external validation cohort, the image and combined models were divided into the models before and after the SR3 method. (B): Confusion matrices of the combined model without the SR3 method in the internal validation cohort and the combined model$_{pre\text{-}SR3}$ and combined model$_{post\text{-}SR3}$ in the external validation cohort. The confusion matrices show the pair-wise comparison; diagonal: number of cases correctly classified; off-diagonal: number of cases incorrectly classified. (C): Calibration curves of the combined model without SR3 method in the internal validation cohort and the combined model$_{pre\text{-}SR3}$ and combined model$_{post\text{-}SR3}$ in the external validation cohort. Calibration curves show excellent agreement between the model-

It would be more feasible to adjust treatment decision at an early stage if pre-treatment images were significant in predicting pCR, which could maximize the likelihood of achieving pCR and minimize unnecessary adverse effects, expenses and risks of disease progression for patients who do not achieve pCR.[34] Furthermore, there have been additional studies focusing on predicting the response to NAC using ultrasound images.[15,35] Nevertheless, these studies were often limited to single-center investigations with small sample sizes, and their findings were not validated in external cohorts.[15,35] Liu et al. predicted pCR for HER-positive breast cancer patients.[16] In our study, we intentionally relaxed the exclusion criteria to enhance the generalizability of our model. This allowed us to include a diverse range of primary breast cancers, regardless of tumor stage, pathological type, or molecular subtype. The data we utilized, to our knowledge, was the largest to date for pCR assessment. Moreover, to assess the generalizability of our deep learning model in predicting pCR, we conducted tests on an external validation cohort. This validation step was crucial in demonstrating the reliability and effectiveness of our model beyond the initial study population.[36,37]

Breast cancer with a negative axilla can be considered for exemption from unnecessary axillary surgery, leading to a significant reduction in potential risks associated with postoperative complications. The Memorial Sloan Kettering Cancer Center (MSKCC) nomogram is a well-validated tool that incorporates several clinical variables to predict the likelihood of ALNM.[38] However, it should be noted that histopathological data, such as histological tumor size, lymphovascular invasion, and multifocality, may not be available preoperatively. In contrast, our deep learning model utilized only clinicopathological data after a biopsy of breast cancer, a standard procedure preoperatively, which could serve as a non-invasive tool for the prediction of outcomes. Additionally, the proposed nomogram did not incorporate any information derived from the images of the tumor to predict outcomes or guide treatment decisions. Zhou et al.[12] constructed a deep learning model with the specific purpose of predicting ALNM in T1/T2 breast cancer with clinically negative axilla. In order to enhance the applicability of the study findings to real-world clinical scenarios, we further expanded our inclusion criteria to encompass patients with primary breast cancer across all tumor stages. In another study, deep learning radiomics was employed to analyze conventional ultrasound and wave elastography data for the prediction of ALNM in patients with early-stage breast cancer.[18] Although the previous study yielded satisfactory results, it was limited to a single-center setting and lacked external validation.[18] In the current study, we sought to address this limitation by validating our AutoRDL framework for predicting ALNM using two independent external validation cohorts. This rigorous validation approach was employed to assess the generalization performance of our model and confirm its reliability across diverse patient populations.[36,37]

There was variability in the quality of the ultrasound images due to the scans being performed by multiple physicians using different machines.[14] For instance, the images from the KMUH cohort had poorer spatial resolution compared to those from other participating institutions. In previous studies, images with insufficient quality were typically excluded, and the models were trained and evaluated exclusively on high-quality images to obtain higher model performance.[11–13,15] In the present study, for the first time, we applied a novel super-resolution via iterative refinement method to enhance the spatial resolution of the medical images. This approach allowed us to increase the sample size for training and enabled the broader clinical implementation of our deep learning model in real-world settings. By overcoming the resolution limitations of the imaging system, this software technology paved the way for more accurate and reliable diagnoses and treatment decisions, even in primary hospitals or resource-constrained medical environments. Moreover, previous studies predominantly relied on manual delineation of medical images for predicting pCR and ALNM in patients with breast cancer, which could be time-consuming, labor-intensive, and subjective, potentially introducing variability and bias into the predictions.[11–13] To address these challenges, to our knowledge, our fully automated AutoRDL framework used the largest data set for tumor detection and classification in breast cancer, which detected breast tumors prior to the prediction task, and demonstrated accurate detection performance despite the variations in tumor size, shape, and location. Regarding the classification tasks, we first constructed the pCR prediction model using a specific PMG network. Subsequently, we reused the same network architecture to perform ALNM prediction following parameter fine-tuning. The reusable neural network pipeline also demonstrated satisfactory performance in predicting ALNM. The advantage of this approach lies in the similarity of the two tasks, which enables more

predicted and observed pCR probabilities. (D): DCA curves of the combined model without SR3 method in the internal validation cohort and the combined model$_{pre\text{-}SR3}$ and combined model$_{post\text{-}SR3}$ in the external validation cohort. The plot shows the net benefit (y-axis) across a range of risk thresholds (x-axis) of the combined model compared with intervention in all participants (all) or no intervention (none). The decision curves show that the combined model$_{post\text{-}SR3}$ had a higher net benefit than the combined model$_{pre\text{-}SR3}$ in predicting pCR when the threshold probability in the clinical decision was 0.323–0.782. Abbreviations: pCR, pathological complete response; ROC, receiver operator characteristic; SR3, image super-resolution via iterative refinement; DCA, decision curve analysis.

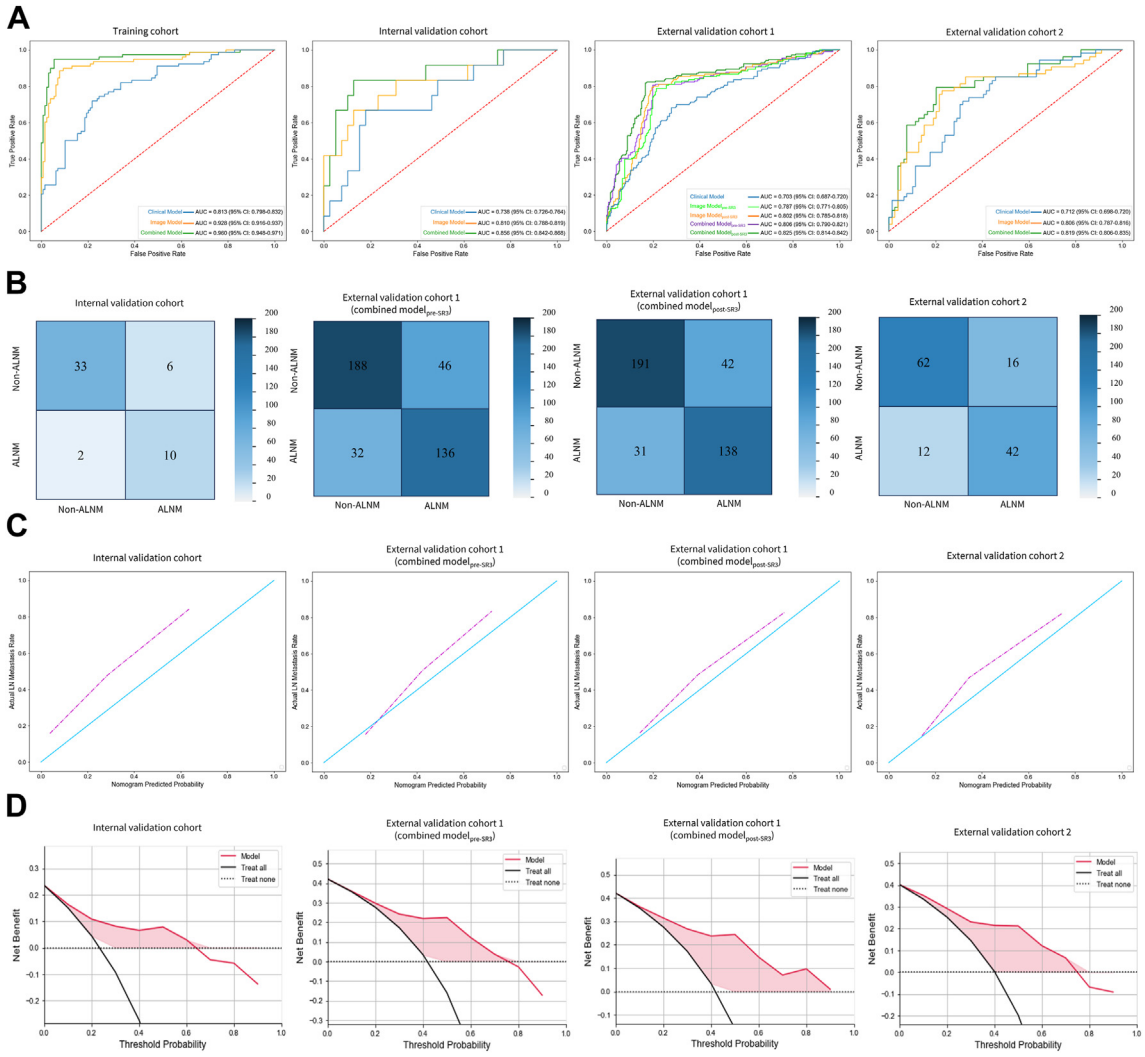

Fig. 4: **Performance of the clinical, image, and combined models in predicting ALNM.** (A): ROC curves of the clinical model, image model, and combined model for predicting ALNM in the training, internal validation, external validation cohort 1, and external validation cohort 2. Among the external validation cohort 1, the image and combined models were divided into models before and after the SR3 approach. (B): Confusion matrices of the combined model without the SR3 approach in the internal validation cohort and external validation cohort 2, as well as the combined model$_{pre\text{-}SR3}$ and combined model$_{post\text{-}SR3}$ in the external validation cohort 1. (C): Calibration curves of the combined model without SR3 in the internal validation cohort and external validation cohort 2, as well as the combined model$_{pre\text{-}SR3}$ and combined model$_{post\text{-}SR3}$ in the external validation cohort 1. Calibration curves display excellent agreement between the model-predicted and observed ALNM probabilities. (D): DCA curves of the combined model without SR3 approach in the internal validation cohort and external validation cohort 2, as well as the combined model$_{pre\text{-}SR3}$ and combined model$_{post\text{-}SR3}$ in the external validation cohort 1. The plot shows the net benefit (y-axis) across a range of risk thresholds (x-axis) of the combined model compared with intervention in all participants (all) or no intervention (none). The decision curves show that the combined model$_{post\text{-}SR3}$ had a higher net benefit than the combined model$_{pre\text{-}SR3}$ in predicting ALNM when the threshold probability in the clinical decision was 0.108–0.755. Abbreviations: ALNM, axillary lymph node metastasis; ROC, receiver operator characteristic; SR3, image super-resolution via iterative refinement; DCA, decision curve analysis.

efficient training and sharing of learned features through the utilization of the same network architecture. This ultimately improves the model's ability to generalize to some extent. Moreover, without the need for developing two separate architectures, our approach greatly reduced computational consumption and saved time in model development, while eliminating the redundancy of results due to multiple networks.

Despite promising findings, our study had some limitations. First, the retrospective design of this study may have inevitably introduced selection bias. Therefore, further validation using prospective cohorts are

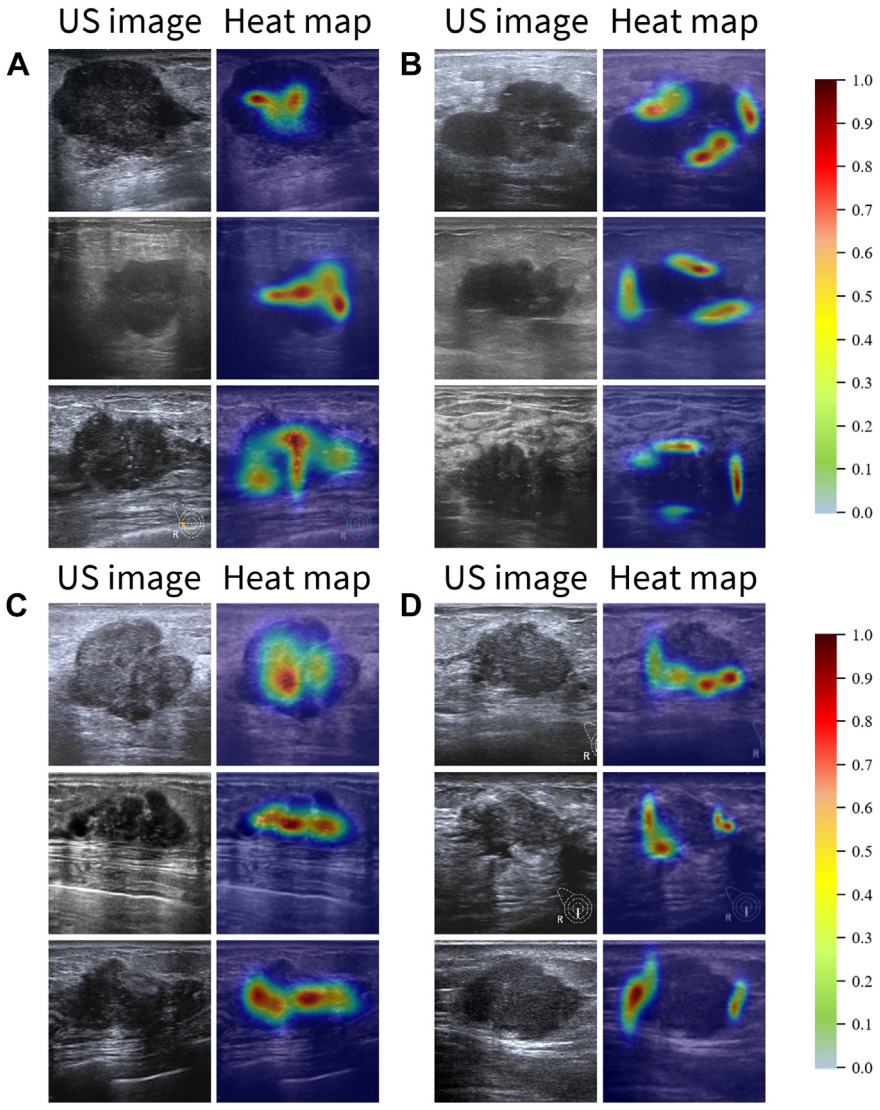

Fig. 5: **Visualization results of the pCR and ALNM prediction.** Color-coded heatmaps were superimposed on the corresponding ultrasound images. The central region of the tumor was significant for the positive cases (pCR [A] and ALNM [C]) while it was the tumor boundary for negative cases (non-pCR [B] and non-ALNM [D]), which could be decoded by the color bar on the right. Abbreviations: AutoRDL, automated and reusable deep learning; pCR, pathological complete response; ALNM, axillary lymph node metastasis.

needed to ensure the reliability and applicability of the developed model in real-world clinical practice. Second, our ultrasound images were acquired from 10 different devices across multiple centers, which could potentially introduce systematic variations in the images. Nevertheless, harmonization algorithms can be designed to minimize the variability of ultrasound images caused by different scanners and protocols, which may potentially enhance the generalizability of deep learning models. In this present study, we mitigated the differences between low- and high-quality images by using image super-resolution techniques, and further reduces the domain gap between different central data by detecting lesion regions. In the future, more effective harmonization or domain adaptive algorithms can be developed to remove the scanner-specific biases while preserving the biological properties of images. Third, we only used pre-treatment images to predict pCR following NAC. Previous studies have indicated that incorporating longitudinal images at multiple time points may provide more comprehensive and informative data for pCR prediction.[15–17] Fourth, the patient population was limited to the Asian population. Therefore, it is necessary to thoroughly evaluate the generalization

performance of our deep learning model in different geographic settings. Fifth, multiple lesions were excluded from the study due to the potential challenge of obtaining one-to-one pathological results for each lesion. Similarly, non-mass and invisible lesions were also excluded, as they were difficult to be segmented or detected on ultrasound images. Finally, although Grad-GAMs were used to identify salient visual features and facilitate the interpretation of the image-based deep learning model, the prediction results of the model should also be interpreted with caution. Hence, there is still a need for comprehensive consideration in clinical decision-making.

In conclusion, our study proposed an AutoRDL framework as a non-invasive and effective tool for automated prediction of pCR and ALNM in patients with breast cancer. This framework holds promise in offering valuable insights for treatment decisions in routine clinical practice. However, further refinement and validation are required prior to integrating this model into routine clinical use. Prospective multicentre validation will be conducted to evaluate its performance across different clinical settings and patient populations. Through this validation process, the model can be further improved to ensure its reliability and accuracy, ultimately becoming a valuable reference for clinical treatment decision-making.

Contributors

Conceptualization: JJY, XZ, BZ, SXZ; Data collection: JJY, YH, LZOY, PC, HS, LZ; Accessing and verifying the underlying data: JJY, YH, LZOY, SFP. Formal analysis: HY, LZOY, XZ, XWW, ZJ, QYC, BZ; Methodology: JJY, XZ, XWW, ZJ, QYC, BZ; Writing: JJY, BZ; Supervision: SFP, BZ, SXZ. All authors reviewed the manuscript, approved the submitted version, and had final responsibility for the decision to submit for publication.

Data sharing statement

Data will be made available to interested research partners upon reasonable request to the corresponding authors SFP, BZ or SXZ. The source code for the AutoRDL is available online (https://github.com/code202308/ALNM_pCR). Moreover, all experimental and implementation details are available in appendix.

Declaration of interests

We declare no competing interests.

Acknowledgements

We acknowledge financial support from the National Key Research and Development Program of China (2023YFF1204600); the National Natural Science Foundation of China (82227802, 82302306); the Clinical Frontier Technology Program of the First Affiliated Hospital of Jinan University, China (JNU1AF-CFTP-2022-a01201); the Science and Technology Projects in Guangzhou (202201020022, 2023A03J1036, 2023A03J1038); the Science and Technology Youth Talent Nurturing Program of Jinan University (21623209); and the Postdoctoral Science Foundation of China (2022M721349). The authors would like to express their gratitude to EditSprings (https://www.editsprings.cn) for the expert linguistic services provided.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2024.102499.

References

1 Imbriaco M, Ponsiglione A. Predicting pathologic complete response after neoadjuvant chemotherapy. *Radiology*. 2021;299(2):301–302.
2 Bae MS. Using deep learning to predict axillary lymph node metastasis from US images of breast cancer. *Radiology*. 2020;294(1):29–30.
3 Pilewskie M, Morrow M. Axillary nodal management following neoadjuvant chemotherapy: a review. *JAMA Oncol*. 2017;3(4):549–555.
4 Kuerer HM, Smith BD, Krishnamurthy S, et al. Eliminating breast surgery for invasive breast cancer in exceptional responders to neoadjuvant systemic therapy: a multicentre, single-arm, phase 2 trial. *Lancet Oncol*. 2022;23(12):1517–1524.
5 Ahmed M, Purushotham AD, Douek M. Novel techniques for sentinel lymph node biopsy in breast cancer: a systematic review. *Lancet Oncol*. 2014;15(8):e351–e362.
6 Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203–211.
7 Qian X, Pei J, Zheng H, et al. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. *Nat Biomed Eng*. 2021;5(6):522–532.
8 Wang Y, Acs B, Robertson S, et al. Improved breast cancer histological grading using deep learning. *Ann Oncol*. 2022;33(1):89–98.
9 Santucci D, Faiella E, Gravina M, et al. CNN-based approaches with different tumor bounding options for lymph node status prediction in breast DCE-MRI. *Cancers*. 2022;14(19):4574.
10 Vulchi M, Adoui ME, Braman N, et al. Development and external validation of a deep learning model for predicting response to HER2-targeted neoadjuvant therapy from pretreatment breast MRI. *J Clin Oncol*. 2019;37(15_suppl):593.
11 Guo X, Liu Z, Sun C, et al. Deep learning radiomics of ultrasonography: identifying the risk of axillary non-sentinel lymph node involvement in primary breast cancer. *EBioMedicine*. 2020;60:103018.
12 Zhou L-Q, Wu X-L, Huang S-Y, et al. Lymph node metastasis prediction from primary breast cancer US images using deep learning. *Radiology*. 2020;294(1):19–28.
13 Jiang M, Li C-L, Luo X-M, et al. Ultrasound-based deep learning radiomics in the assessment of pathological complete response to neoadjuvant chemotherapy in locally advanced breast cancer. *Eur J Cancer*. 2021;147:95–105.
14 Fowler AM, Mankoff DA, Joe BN. Imaging neoadjuvant therapy response in breast cancer. *Radiology*. 2017;285(2):358–375.
15 Gu J, Tong T, He C, et al. Deep learning radiomics of ultrasonography can predict response to neoadjuvant chemotherapy in breast cancer at an early stage of treatment: a prospective study. *Eur Radiol*. 2022;32(3):2099–2109.
16 Liu Y, Wang Y, Wang Y, et al. Early prediction of treatment response to neoadjuvant chemotherapy based on longitudinal ultrasound images of HER2-positive breast cancer patients by Siamese multi-task network: a multicentre, retrospective cohort study. *EClinicalMedicine*. 2022;52:101562.
17 Wu L, Ye W, Liu Y, et al. An integrated deep learning model for the prediction of pathological complete response to neoadjuvant chemotherapy with serial ultrasonography in breast cancer patients: a multicentre, retrospective study. *Breast Cancer Res*. 2022;24(1):81.
18 Zheng X, Yao Z, Huang Y, et al. Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. *Nat Commun*. 2020;11(1):1236.
19 Allison KH, Hammond MEH, Dowsett M, et al. Estrogen and progesterone receptor testing in breast cancer: ASCO/CAP guideline update. *J Clin Oncol*. 2020;38(12):1346–1366.
20 Wolff AC, Hammond MEH, Allison KH, et al. Human epidermal growth factor receptor 2 testing in breast cancer: American society of clinical oncology/college of American pathologists clinical practice guideline focused update. *J Clin Oncol*. 2018;36(20):2105–2122.
21 von Minckwitz G, Untch M, Blohmer J-U, et al. Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. *J Clin Oncol*. 2012;30(15):1796–1804.
22 Giuliano AE, Connolly JL, Edge SB, et al. Breast Cancer-Major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA Cancer J Clin*. 2017;67(4):290–303.
23 Saharia C, Ho J, Chan W, Salimans T, Fleet DJ, Norouzi M. Image super-resolution via iterative refinement. *IEEE Trans Pattern Anal Mach Intell*. 2023;45(4):4713–4726.

24 Fine-grained visual classification via progressive multi-granularity training of igsaw patches. In: Du R, Chang D, Bhunia AK, et al., eds. *European Conference on Computer Vision*. 2020.
25 Huang Y, Zhu T, Zhang X, et al. Longitudinal MRI-based fusion novel model predicts pathological complete response in breast cancer treated with neoadjuvant chemotherapy: a multicenter, retrospective study. *EClinicalMedicine*. 2023;58:101899.
26 Lee YW, Huang CS, Shih CC, Chang RF. Axillary lymph node metastasis status prediction of early-stage breast cancer using convolutional neural networks. *Comput Biol Med*. 2021;130:104206.
27 Tulio Ribeiro M, Singh S, Guestrin C. "Why should I trust You?": Explaining the predictions of any Classifier. *arXiv*. 2016 [arXiv:1602.04938 p.].
28 Zhou B, Khosla A, Lapedriza À, Oliva A, Torralba A. Learning deep features for discriminative localization. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015:2921–2929.
29 Sammut S-J, Crispin-Ortuzar M, Chin S-F, et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*. 2022;601(7894):623–629.
30 Bitencourt AGV, Gibbs P, Rossi Saccarelli C, et al. MRI-based machine learning radiomics can predict HER2 expression level and pathologic response after neoadjuvant therapy in HER2 overexpressing breast cancer. *EBioMedicine*. 2020;61:103042.
31 Braman NM, Etesami M, Prasanna P, et al. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI. *Breast Cancer Res*. 2017;19(1):57.
32 Liu Z, Li Z, Qu J, et al. Radiomics of multiparametric MRI for pretreatment prediction of pathologic complete response to neoadjuvant chemotherapy in breast cancer: a multicenter study. *Clin Cancer Res*. 2019;25(12):3538–3547.
33 Shi Z, Huang X, Cheng Z, et al. MRI-Based quantification of intratumoral heterogeneity for predicting treatment response to neoadjuvant chemotherapy in breast cancer. *Radiology*. 2023;308(1):e222830.
34 Korde LA, Somerfield MR, Carey LA, et al. Neoadjuvant chemotherapy, endocrine therapy, and targeted therapy for breast cancer: ASCO guideline. *J Clin Oncol*. 2021;39(13):1485–1505.
35 Byra M, Dobruch-Sobczak K, Klimonda Z, Piotrzkowska-Wroblewska H, Litniewski J. Early prediction of response to neoadjuvant chemotherapy in breast cancer sonography using siamese convolutional neural networks. *IEEE J Biomed Health Inform*. 2021;25(3):797–805.
36 Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (claim): a guide for authors and reviewers. *Radiol Artif Intell*. 2020;2(2):e200029.
37 Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.
38 Bevilacqua JLB, Kattan MW, Fey JV, Cody HS, Borgen PI, Van Zee KJ. Doctor, what are my chances of having a positive sentinel node? A validated nomogram for risk estimation. *J Clin Oncol*. 2007;25(24):3670–3679.