Xu *et al. Breast Cancer Research* (2025) 27:80
https://doi.org/10.1186/s13058-025-02033-6

Breast Cancer Research

RESEARCH Open Access

Optimizing breast lesions diagnosis and decision-making with a deep learning fusion model integrating ultrasound and mammography: a dual-center retrospective study

Ziting Xu[1†], Shengzhou Zhong[2†], Yang Gao[1†], Jiekun Huo[3], Weimin Xu[4], Weijun Huang[5], Xiaomei Huang[6], Chifa Zhang[1], Jianqiao Zhou[7], Qing Dan[8], Lian Li[1], Zhouyue Jiang[1], Ting Lang[1], Shuying Xu[1], Jiayin Lu[1], Ge Wen[6*], Yu Zhang[2*] and Yingjia Li[1*]

Abstract

Background This study aimed to develop a BI-RADS network (DL-UM) via integrating ultrasound (US) and mammography (MG) images and explore its performance in improving breast lesion diagnosis and management when collaborating with radiologists, particularly in cases with discordant US and MG Breast Imaging Reporting and Data System (BI-RADS) classifications.

Methods We retrospectively collected image data from 1283 women with breast lesions who underwent both US and MG within one month at two medical centres and categorised them into concordant and discordant BI-RADS classification subgroups. We developed a DL-UM network via integrating US and MG images, and DL networks using US (DL-U) or MG (DL-M) alone, respectively. The performance of DL-UM network for breast lesion diagnosis was evaluated using ROC curves and compared to DL-U and DL-M networks in the external testing dataset. The diagnostic performance of radiologists with different levels of experience under the assistance of DL-UM network was also evaluated.

Results In the external testing dataset, DL-UM outperformed DL-M in sensitivity (0.962 vs. 0.833, $P=0.016$) and DL-U in specificity (0.667 vs. 0.526, $P=0.030$), respectively. In the discordant BI-RADS classification subgroup, DL-UM

†Ziting Xu, Shengzhou Zhong and Yang Gao contributed equally to this work.

*Correspondence:
Ge Wen
m13360022166@163.com
Yu Zhang
yuzhang@smu.edu.cn
Yingjia Li
lyjia@smu.edu.cn

Full list of author information is available at the end of the article

achieved an AUC of 0.910. The diagnostic performance of four radiologists improved when collaborating with the DL-UM network, with AUCs increased from 0.674–0.772 to 0.889–0.910, specificities from 52.1%–75.0 to 81.3–87.5% and reducing unnecessary biopsies by 16.1%–24.6%, particularly for junior radiologists. Meanwhile, DL-UM outputs and heatmaps enhanced radiologists' trust and improved interobserver agreement between US and MG, with weighted kappa increased from 0.048 to 0.713 ($P < 0.05$).

Conclusions The DL-UM network, integrating complementary US and MG features, assisted radiologists in improving breast lesion diagnosis and management, potentially reducing unnecessary biopsies.

Keywords Neural networks, Ultrasonography, Digital mammography, Clinical decision-making, Breast tumours

Background

Mammography (MG) is recommended for breast cancer screening, but its sensitivity is limited in women with dense breasts [1]. Ultrasound (US), as a supplementary screening tool for dense breasts, however, falls short in detecting microcalcifications, a crucial indicator of early breast cancer [2]. MRI, although effective in detecting early breast cancer [3], is currently only recommended for high-risk women due to its high cost and lengthy scans [4]. Therefore, combining US and MG could potentially mitigate these limitations and improve cancer detection, particularly for women with dense breasts [5].

However, discordances of Breast Imaging Reporting and Data System (BI-RADS) classifications between US and MG are inevitable, potentially causing unnecessary anxiety and biopsies [6]. Previous attempts using shear wave elastography and contrast-enhanced US to improve diagnosis in discordant BI-RADS cases [7, 8] have encountered controversy due to high operator variability and discrepancies in diagnostic criteria. A recent study [6] proposed a nomogram integrating visual analysis of US and MG, but it relies on subjective radiological observation, posing challenges for less experienced radiologists. Therefore, to develop a robust and objective method for optimizing diagnosis and management of breast lesions with discordant US and MG classifications is imperative.

Deep learning (DL), allowing automatic analyzing medical images, has shown promise in breast cancer detection and management [9, 10]. Emerging evidence [11–14] suggests that extracting multimodal radiomics features through DL approaches could overcome unimodal imaging limitations, offering comprehensive and complementary diagnostic insights. Numerous studies [15, 16] indicated that artificial intelligence (AI) integration with radiologists could improve diagnostic accuracy, especially for junior radiologists, bolstering the role of artificial intelligence in clinical decision-making. However, how to effectively integrate DL models and radiologists of varying experience levels in cases of discordant MG and US BI-RADS classifications remains unclear. Moreover, radiologists' perceptions of DL outputs may raise uncertainty regarding its clinical applicability, warranting further investigation.

Hence, we aimed to develop a DL network via integrating US and MG images (DL-UM) and investigate its performance in improving breast lesion diagnosis and management when collaborating with radiologists, particularly in cases of discordant US and MG BI-RADS classifications. Additionally, we explored the potential of DL-UM outputs and heatmaps to foster trust of radiologists with various experience in stimulated clinical workflow.

Methods

Statement of ethics

This study was approved by the Institutional Ethics Committee of the hospital (NFEC-202012-K8) and the requirement for informed consent was waived owing to the retrospective design and use of anonymized data.

Study population

Women undergoing breast imaging were consecutively and retrospectively collected from the Medical Centre 1 between June 2019 and June 2021 to form the development dataset for establishing DL network. An external testing dataset was consecutively collected from Medical Centre 2 between January 2021 and June 2021. Figure 1 shows the patient selection flowchart.

Inclusion criteria were as follows: (1) women aged > 18 years with paired MG and US imaging conducted within 1 month; (2) consistent targeted lesion findings on both MG and US; and (3) collected the most suspicious lesion in cases of multiple lesions in the imaging data. Exclusion criteria were as follows: (1) no pathological results; (2) incomplete or low-quality imaging data; (3) radiotherapy or neoadjuvant chemotherapy before examination; (4) patients with incomplete BI-RADS assessment (to exclude BI-RADS 0 cases); and (5) previous biopsy or surgery before examination (to exclude BI-RADS 6 cases).

After selection, 1126 patients from Medical Centre 1 were included for analysis and randomised into the training, validation, and internal testing cohorts in a 7:1:2

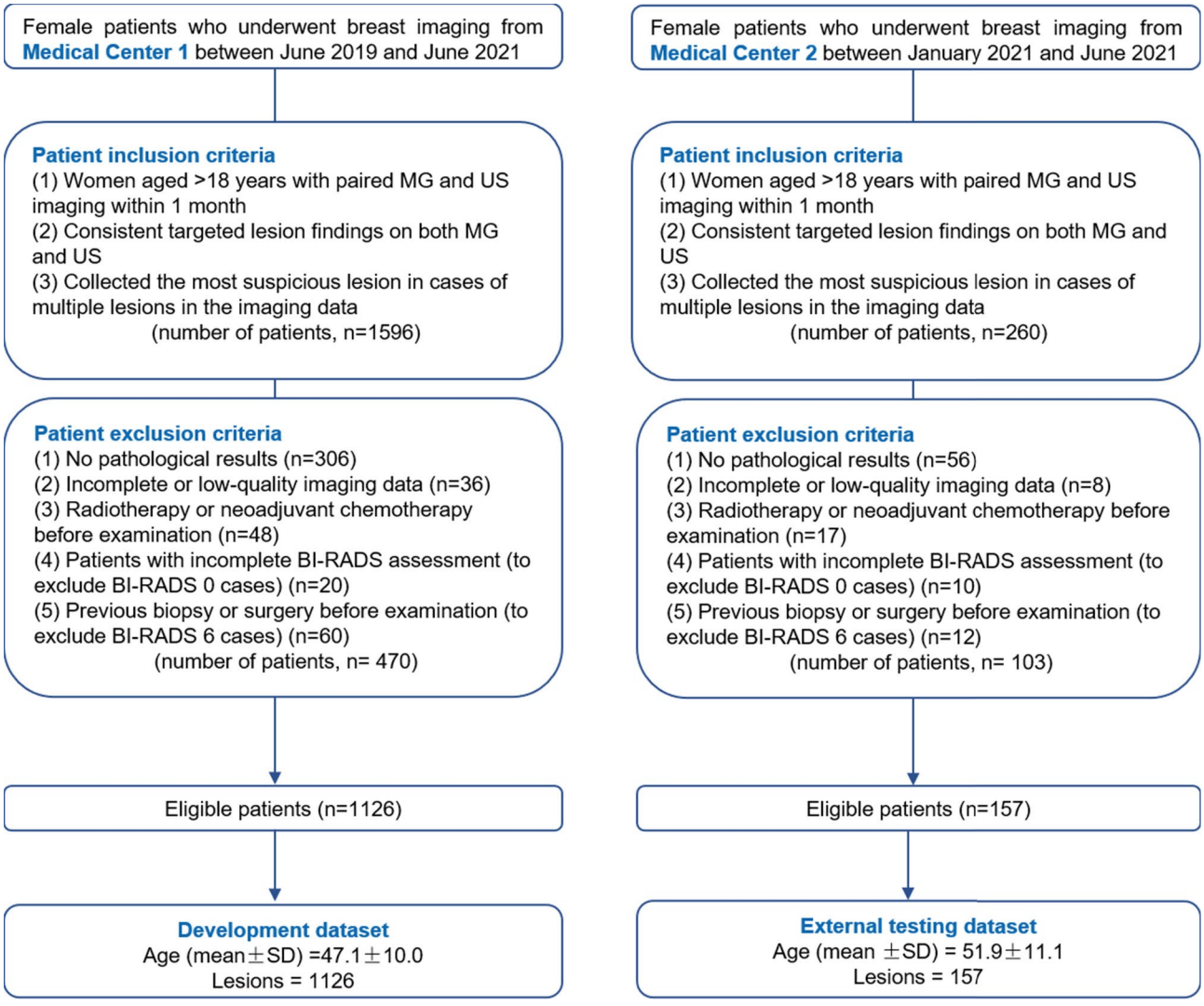

Fig. 1 Flowchart of patient selection. US, ultrasound; MG, mammography

allocation ratio. The external testing cohort comprised 157 patients from Medical Centre 2.

Imaging acquisition and interpretation

MG images were acquired using Mammomat Novation DR (Siemens AG Medical Solutions, Erlangen, Germany) and Selenia Dimensions (Hologic, Bedford, Mass, USA) digital systems, encompassing craniocaudal and medio lateral-oblique views. US images were obtained using different devices, including Aixplorer (SuperSonic Imagine, Aix-en-Provence, France), Logiq E9 (GE Healthcare, Wauwatosa, WI, USA) systems, and other systems with 7.5–15 MHz linear high-frequency transducers. Two-directional (transverse and longitudinal) static images were recorded, focusing on the region of interest in each patient's image data that exhibited the most suspicious lesion. For reliable and reproducible BI-RADS classifications, six senior radiologists (R1–R3 with ≥5 years of experience in breast US, R4–R6 with ≥8 years of experience in MG,) independently reviewed all images. If results diverged, the radiologists resolved discrepancies through discussion to reach a consensus for the final diagnosis. The radiologists were blind to the pathological results but had access to clinical information and prior imaging.

According to the 2013 American College of Radiology BI-RADS criteria, lesions rated as 2 or 3 were considered benign or probably benign, while those classified as 4 or 5 were considered suspicious, warranting tissue diagnosis. A discordant BI-RADS classification between US and MG was defined when a lesion was classified as 4 or 5 on one modality but as 2 or 3 on the other. Based on these standards, all lesions were categorised into subgroups with discordant or concordant BI-RADS classifications.

Data pre-processing

Data pre-processing involved cropping irrelevant regions in US and MG images to minimise their negative effects on network performance, conducted by experienced radiologists (R1–R6) using ITK-SNAP software (http://www.radiantviewer.com). To account for diagnostic significance of adjacent tissue, regions of interest (ROIs) in both MG and US images were expanded by 30% of their shortest lengths. All images were then resized to a standard size (224 pixels × 224 pixels), and intensity values were normalised to the minimum-maximum intensity range (0–1).

Model architecture

Figure 2 depicts the study design and the architecture of the DL-UM network. The network included two feature extraction branches, one each for US and MG images, with a shared feature identification developed using VGG19 [17]. Inputs for each feature extraction branch were 224×224-pixel paired patches from US and MG images after lesion segmentation and image preprocessing. Each branch comprised 5 convolution blocks with convolution layers of 2, 2, 4, 4, and 4, respectively, and of 64, 128, 256, 512, and 512 filters with a kernel size of 3×3, for efficiently extracting and propagating the coarse-to-fine representations. To prevent from gradient vanishing and enhance network sparsity, the last convolutional layers of all convolution blocks were followed by a batch normalization and a nonlinear Rectified Linear Unit activation operator. For the first four blocks, the activated features would pass through a maximum pooling layer with the pooling window size of 2×2 to perform feature dimension reduction for relieving overfitting issue. Meanwhile, they were also input to an additional classification head, including a global average pooling (GAP) layer, two fully connected layers with neuron numbers of 64 and 1, and a sigmoid function, for encouraging the network to capture more discriminative information via deep supervision strategy [18]. In the final convolution block, the maximum pooling layer was removed and features from GAP layer were used for subsequent supervision and final classification. Given the output features from final convolution block, a concatenation operation was embedded into the end of network to receive and integrate the final representations derived from US and MG branches. Finally, the integrated feature representations were used to perform breast tumour classification via the same classification head. A focal loss [19] was used as supervision function to focus network's attention on the samples difficult to classify during training. Meanwhile, the mean absolute error (MAE) to force the final predictions of US and MG branches to be consistent. During the network training, the Adam optimizer was used with the global learning rate of $1 \times 10-4$. Meanwhile, the momentum was 0.9 and the batch size was 16. After every epoch, model parameters were saved and the model with the lowest average loss on the validation set was chosen for evaluation on the test data set. The DL-UM models, developed in Python (3.6.13; Python Software Foundation, Wilmington, DE, USA), were trained and evaluated using five-fold cross-validation, with samples from different classes randomly partitioned by patient. The DL models were trained on a server with eight 12GB NVIDIA GeForce RTX 2080 Ti GPUs, an Intel Xeon E5-2650 v4 CPU @ 2.20 GHz, and 192GB RAM. Key Python packages and versions include TensorFlow (v2.1.0), Keras (v2.3.1), NumPy (v1.19.2), Pandas (v1.1.5), scikit-learn (v0.20.3), Matplotlib (v3.3.4), and Pillow (v8.4.0).

Network interpretability

The Class Activation Map (CAM) [20] technique visualizes the image regions a DL model focused on during classification, pinpointing areas that significantly influence decision-making. CAMs help illustrate which image areas contribute most to the model's decisions, supporting radiologists in evaluating and interpreting the model's performance. In our model, CAMs were generated from the last convolutional layer in each block to highlight areas of network focus. The GAP layer produced an eigenvector representing average feature significance, which was weighted and applied to feature maps for visualization. The resulting heatmap effectively emphasized the critical tumor regions identified by the network.

DL-UM-assisted radiologist

Breast images from Medical Centre 2 were independently evaluated by two junior radiologists (3 years of experience in breast US [R7] and MG [R8]) and two senior radiologists (8 years of experience in breast US [R9] and 10 years of experience in MG [R10]). After analysing US and MG images independently, four radiologists dichotomised all lesions as "possibly benign" and "possibly malignant" respectively. After two months, radiologists (R7–R10) re-analysed the US and MG images in random orders and referred to DL-UM outputs (based on both US and MG images) for a dichotomous classification diagnosis. Meanwhile, radiologists could accept or reject DL-UM suggestions or request AI explanations with heatmaps (for both US and MG images) (see Fig. 2E). Additionally, before and after reviewing the DL-UM outputs, each radiologist had access to clinical information, both US and MG images for each patient but remaining blind to pathologic results. To evaluate DL-UM-assisted radiologists contributing to decision-making at a simulated clinical setting, we quantified the potential reduction in recommended biopsies and unnecessary biopsies based on DL-UM outputs and radiologist interpretations.

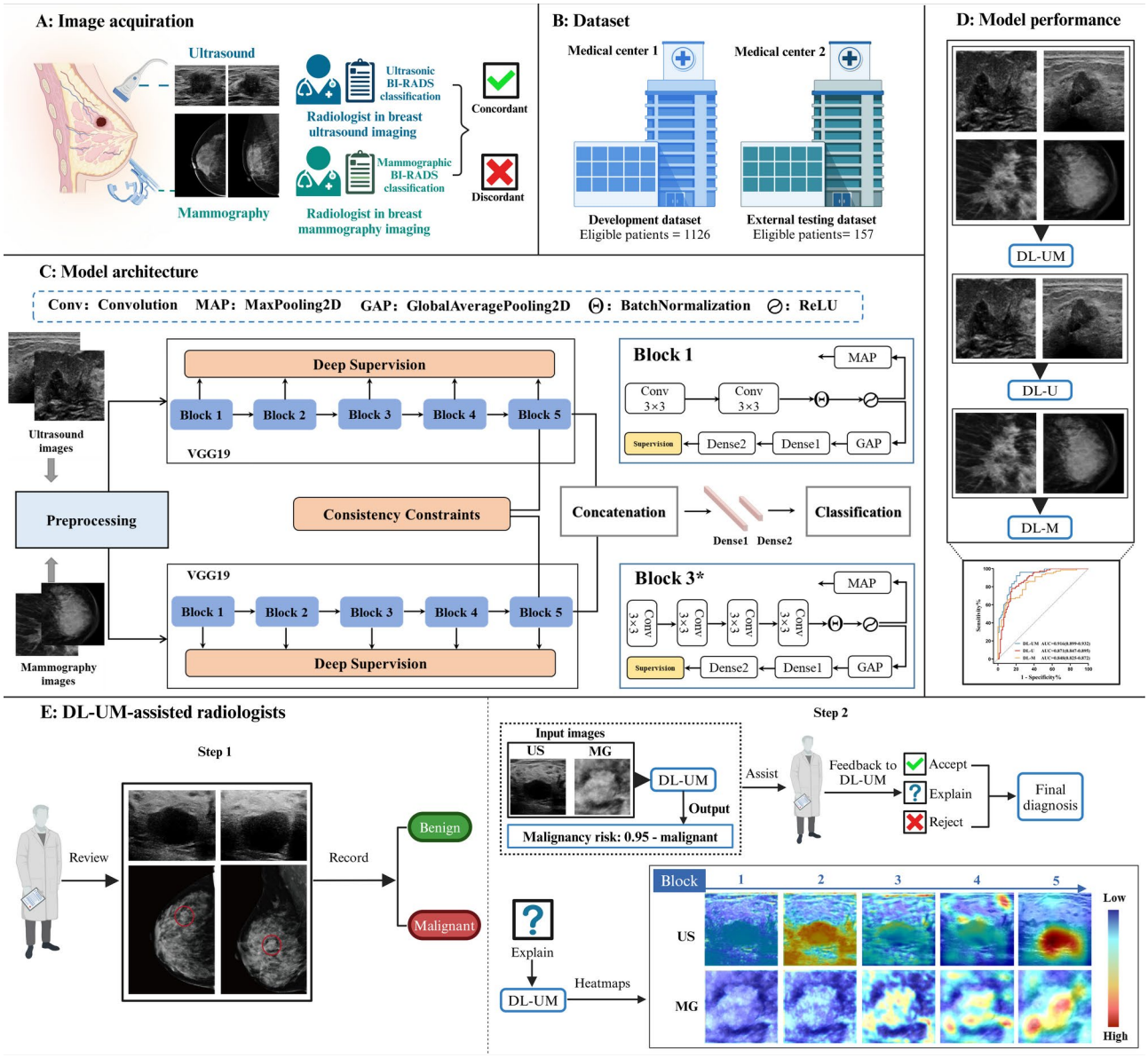

Fig. 2 Design of the study. **A**. Patients underwent paired ultrasonic and mammographic imaging within 1 month and were dichotomised into subgroups with concordant and discordant ultrasonic and mammographic BI-RADS classifications. **B**. Summary of development and external testing datasets from two medical centres. **C.** Architecture of the deep learning (DL-UM) network. The DL-UM network includes two feature extraction branches and a classification head. Each branch comprises five convolution blocks (i.e., $\{CB_i\}_{i=1}^{5}$), with the convolution of layers 2, 2, 4, 4, and 4, and subsequent filtering at 64, 128, 256, 512, and 512 Hz, respectively, using filters with a 3×3 kernel size. A focal loss function is used for deep supervision and classification loss, with the mean absolute error loss function used to force the final output of the two branches to be consistent. * In the VGG19 model used in this study, Block 2, Block 4, and Block 5 share the same architectural structure as Block 3, while Block 1 and Block 3 represent two distinct structural designs. **D.** Diagnostic performances of three DL models were compared with ROC curves. **E.** Overview of DL-UM-assisted radiologist workflow. First step, four radiologists independently reviewed and analysed US and MG images, dichotomising the lesion as "possibly benign" and "possibly malignant". After two months, radiologists respectively re-analysed the US and MG images in random orders and referred to DL-UM outputs ("possibly benign" and "possibly malignant"). Meanwhile, radiologists could accept or reject DL-UM suggestions or request AI explanations with heatmaps, and made the final diagnosis

Following previous studies [21, 22], recommended biopsies included all cases predicted as malignant, while unnecessary biopsies were defined as cases predicted as malignant but pathologically confirmed as benign. Meanwhile missed malignancies were defined as cases predicted as benign but pathologically confirmed as malignant.

Statistical analysis

Continuous variables were analysed using independent t-tests, while categorical variables were compared using

chi-squared (χ^2) or Fisher's exact tests. Predictive performances of the three DL networks and radiologists were assessed with receiver operating characteristic (ROC) curve analyses. Diagnostic accuracy was compared using Delong's test, and sensitivity and specificity were compared with the chi-square test. Inter-observer agreement was analysed using weighted Kappa values. Statistical analyses were performed using SPSS Statistics (version 24.0) and R software (version 3.3.0), with a two-sided *P*-value of 0.05 for significance.

Results

Patient demographics

Overall, we included 1126 patients from Medical Centre 1 (mean age, 47.1 ± 10.0 years; 573 benign and 553 malignant lesions) and 157 patients from Medical Centre 2 (mean age, 51.9 ± 11.1 years; 79 benign and 78 malignant lesions). Demographics are summarised in Table 1. Original US and MG BI-RADS categories with histopathologic results are presented in Supplementary Table 1.

Table 1 Patient demographics

Characteristics	Medical Centre 1 (*n* = 1126)	Medical Centre 2 (*n* = 157)	*P*
BI-RADS classification (%)			0.584
Concordant	600 (53.3%)	80 (51.0%)	
Discordant	526 (46.7%)	77 (49.0%)	
Pathology (%)			0.894
Benign	573 (50.9%)	79 (50.3%)	
Malignant	553 (49.1%)	78 (49.7%)	
Age	47.1 ± 10.0	51.9 ± 11.1	< 0.001
BMI	23.46 ± 3.12	24.10 ± 2.86	0.015
Menopausal			< 0.001
Premenopausal	692 (61.5%)	50 (31.8%)	
Postmenopausal	434 (38.5%)	107 (68.2%)	
Parity			0.868
Nulliparous	47 (4.2%)	7 (4.5%)	
Parous	1079 (95.8%)	150 (95.5%)	
Family history			0.434
No	1105 (98.1%)	156 (99.4%)	
Yes	21 (1.9%)	1 (0.6%)	
Nipple discharge			0.865
No	1086 (96.4%)	151 (96.2%)	
Yes	40 (3.6%)	6 (3.8%)	
Mammography-breast composition			0.700
category A + B	104 (9.2%)	16 (10.2%)	
category C + D	1022 (90.8%)	141 (89.8%)	

Data in parentheses are percentages, except for age and BMI (mean ± SD). The *P*-values represent the comparisons between two medical centers for each characteristic. Abbreviations: BI-RADS: Breast Imaging Reporting and Data System; BMI: Body Mass Index; Mammography-breast composition: almost entirely fatty (category A); scattered areas of fibroglandular tissue (category B); heterogeneously dense (category C); extremely dense (category D)

Diagnostic performance of individual and fusion models

In the external testing dataset (Fig. 3), DL-UM exhibited significant superiority over DL-U in specificity (0.667 [95% CI, 0.624–0.709] vs. 0.526 [95% CI, 0.487–0.564], $P = 0.030$) and over DL-M in sensitivity (0.962 [95% CI, 0.945–0.978] vs. 0.833 [95% CI, 0.802–0.865], $P = 0.016$). This difference was particularly notable in the discordant classification subgroup (Fig. 3 and Supplementary Table 2). Results for development and external test datasets are found in Supplementary Tables 3–8.

Diagnostic performance and management improvement through DL-UM-assisted radiologists

In the external testing dataset, the diagnostic performance of radiologists with DL-UM assistance significantly improved compared to radiologists alone, with area under the ROC curve (AUC) values increased from 0.734–0.835 to 0.898–0.918 (all $P < 0.05$) and specificities from 57.0%–76.0 to 84.8–86.1% (all $P < 0.05$) (Table 2; Fig. 4). Meanwhile DL-UM-assisted radiologists achieved a significant reduction of 8.3–18.7% in unnecessary biopsies, regardless of radiologists' experience (all $P < 0.05$) (Table 2). Similarly, such improvements were more pronounced in the subgroup of discordant classification cases, with increased AUC from 0.674–0.772 to 0.889–0.910 (all $P < 0.05$) and specificities from 52.1%–75.0 to 81.3–87.5% (all $P < 0.05$), and 16.1%–24.6% cases could avoid unnecessary biopsies (all $P < 0.001$).

Clinical implications of radiologists' trust for DL-UM diagnostic support

During the DL-UM-assisted workflows, 73.1% of radiologists showed positive acceptance of DL-UM outputs (Supplementary Tables 9 and Fig. 5). Within the discordant subgroup, however, there was a notable increase in the demand for explanations (26.0% of cases) (Table 3). Meanwhile DL-UM explanation resulted in a significant reduction in unnecessary biopsies by 19.2%, and also decreasing the rate of missed malignancies by 10.8%. Additionally, compared to senior radiologists, the utilization of heatmaps allowed junior radiologists to achieve a greater reduction in unnecessary biopsies by 19.6% and in missed diagnoses by 11.5% (Table 3).

Interobserver agreement of radiologists with and without the assistance of DL-UM

Without DL-UM, agreement between US and MG was significantly lower in the discordant subgroup (kappa = 0.048) than that in the concordant subgroup (kappa = 0.618, $P < 0.05$). With DL-UM assistance, interobserver agreement significantly enhanced with increased weighted kappa from 0.048 to 0.713 ($P < 0.05$) in the discordant classification subgroup. Moreover, observer agreement between R7 and R9 or between

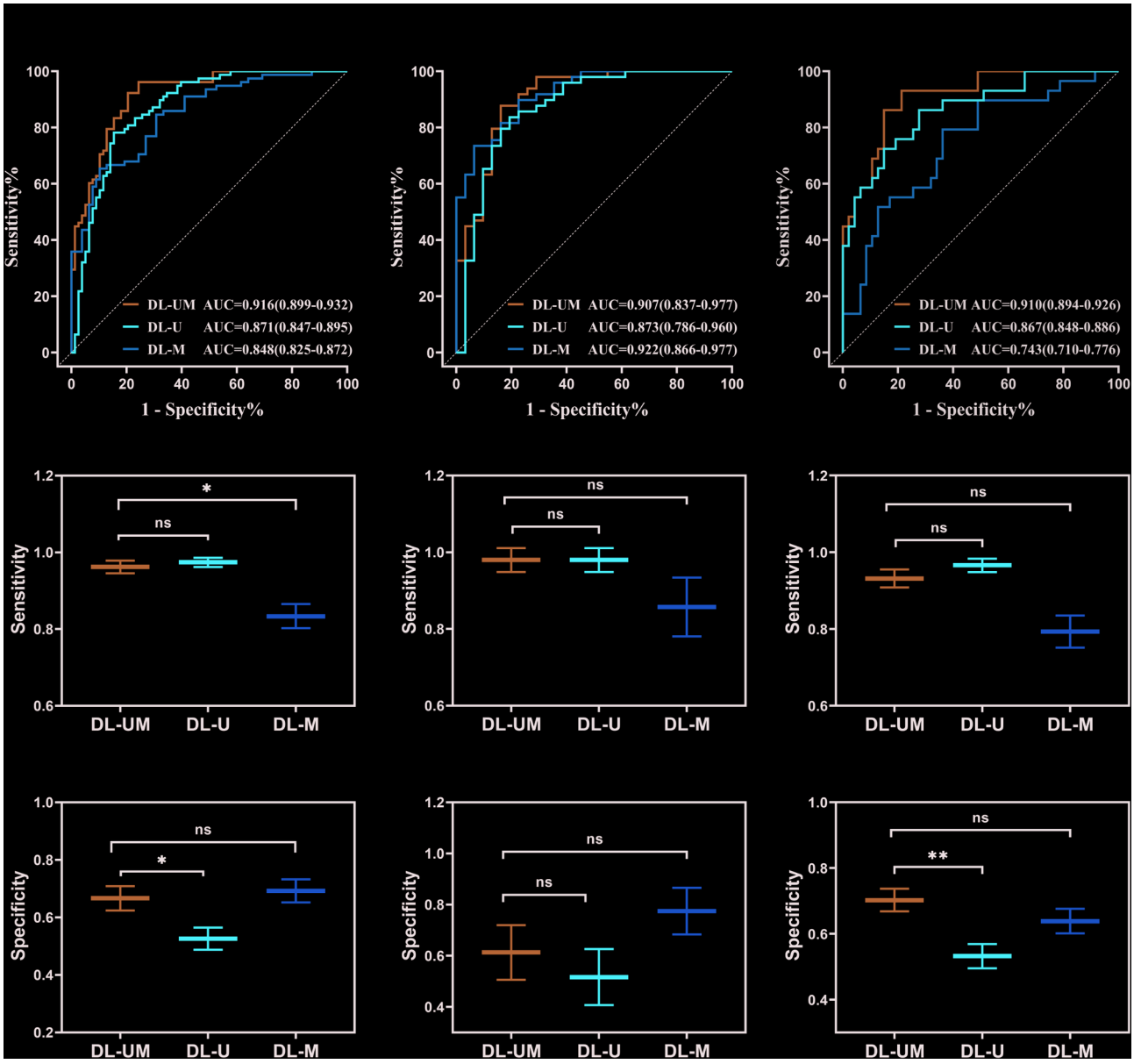

Fig. 3 Comparison of performance among the three DL models. a/d/g: all cases combined; b/e/h: concordant cases; c/f/i: discordant cases; AUC, area under the receiver operating characteristic curve

R8 and R10 significantly increased to 0.739 and 0.687, respectively (both $P < 0.05$) (Fig. 6; Table 4).

Discussion

In this study, the bimodal DL-UM network, integrating US and MG complementary features, significantly improved specificity compared to DL-U and sensitivity compared to DL-M, particularly in the discordant classification subgroup. With the aid of DL-UM, radiologists' diagnostic accuracy and specificity were significantly enhanced, resulting in a notable reduction in unnecessary biopsies, especially for junior radiologists, and improving consistency between US and MG. The potential of DL-UM in building radiologists' trust in AI was further emphasised, with heatmaps aiding in preventing unnecessary biopsies and missed malignancies. These findings highlight the value of DL-UM as a complementary tool to assist radiologists in optimizing breast lesion diagnosis and management.

In alignment with prior DL studies on MG [23–25], this study found that DL-M demonstrated high specificity but reduced sensitivity, which attributed to the obscuring effect of dense breast parenchyma in two-dimensional imaging of MG [1]. Conversely, as noted in previous AI studies [11, 26, 27], DL-U showed high sensitivity but with low specificity, which may result from overlapping

Table 2 Diagnostic performance and management recommendation improvement through DL-UM-radiologists collaboration

			Recommended biopsy (%)	Unnecessary biopsies (%) ^	AUC	Sensitivity (%)	Specificity (%)
All cases combined (*n* = 157)	US	R7	66.2 (104/157)	32.7 (34/104)	0.734 (0.654–0.814)	89.7 (85.0-94.5)	57.0 (49.2–64.7)
		R7 + DL-UM	54.8 (86/157)	14.0 (12/86)	0.898 (0.844–0.953)	94.9 (91.4–98.3)	84.8 (79.2–90.4)
		R9	59.9 (94/157)	24.5 (23/94)	0.810 (0.739–0.881)	91.0 (86.6–95.5)	70.9 (63.8–78.0)
		R9 + DL-UM	55.4 (87/157)	12.6 (11/87)	0.918 (0.868–0.967)	97.4 (95.0-99.9)	86.1 (80.7–91.5)
	MG	R8	64.3 (101/157)	30.7 (31/101)	0.753 (0.674–0.831)	89.7 (85-9.49.0)	60.8 (53.1–68.4)
		R8 + DL-UM	56.1 (88/157)	13.6 (12/88)	0.911 (0.860–0.963)	97.4 (95.0-99.9)	84.8 (79.2–90.4)
		R10	57.3 (90/157)	21.1 (19/90)	0.835 (0.768–0.902)	91.0 (86.6–95.5)	76.0 (69.3–82.6)
		R10 + DL-UM	54.8 (86/157)	12.8 (11/86)	0.911 (0.860–0.963)	96.2 (93.2–99.2)	86.1 (80.7–91.5)
concordant cases (*n* = 80)	US	R7	71.3 (57/80)	19.3 (11/57)	0.792 (0.680–0.904)	93.9 (88.6–99.1)	64.5 (54.0–75.0)
		R7 + DL-UM	61.3 (49/80)	6.1 (3/49)	0.921 (0.849–0.993)	93.9 (88.6–99.1)	90.3 (83.8–96.8)
		R9	66.3 (53/80)	11.3 (6/53)	0.883 (0.794–0.972)	95.9 (91.6-100.2)	80.7 (72.0-89.3)
		R9 + DL-UM	65.0 (52/80)	7.7 (4/52)	0.925 (0.852–0.999)	98.0 (94.9-101.1)	87.1 (79.8–94.4)
	MG	R8	71.3 (57/80)	21.1 (12/57)	0.766 (0.650–0.882)	91.8 (85.8–97.8)	61.3 (50.6–72.0)
		R8 + DL-UM	63.8 (51/80)	5.9 (3/51)	0.941 (0.876-1.000)	98.0 (94.9-101.1)	90.3 (83.8–96.8)
		R10	68.8 (55/80)	12.7 (7/55)	0.877 (0.784–0.969)	98.0 (94.9-101.1)	77.4 (68.3–86.6)
		R10 + DL-UM	66.3 (53/80)	9.4 (5/53)	0.909 (0.828–0.990)	98.0 (94.9-101.1)	83.9 (75.8–91.9)
discordant cases (*n* = 77)	US	R7	61.0 (47/77)	48.9 (23/47)	0.674 (0.552–0.796)	82.8 (74.3–91.2)	52.1 (40.9–63.2)
		R7 + DL-UM	48.1 (37/77)	24.3 (9/37)	0.889 (0.811–0.967)	96.6 (92.5-100.6)	81.3 (72.5–90.0)
		R9	53.2 (41/77)	41.5 (17/41)	0.737 (0.622–0.851)	82.8 (74.3–91.2)	64.6 (53.9–75.3)
		R9 + DL-UM	45.5 (35/77)	20.0 (7/35)	0.910 (0.839–0.981)	96.6 (92.5-100.6)	85.4 (77.5–93.3)
	MG	R8	57.1 (44/77)	43.1 (19/44)	0.733 (0.619–0.847)	86.2 (78.5–93.9)	60.4 (49.5–71.3)
		R8 + DL-UM	48.1 (37/77)	24.3 (9/37)	0.889 (0.811–0.967)	96.6 (92.5-100.6)	81.3 (72.5–90.0)
		R10	45.5 (35/77)	34.3 (12/35)	0.772 (0.660–0.883)	79.3 (70.3–88.4)	75.0 (65.3–84.7)
		R10 + DL-UM	42.9 (33/77)	18.2 (6/33)	0.903 (0.826–0.980)	93.1 (87.5–98.8)	87.5 (80.1–94.9)

Data are expressed as means and 95% confidence intervals in parentheses or as a count in parentheses. R7: junior radiologist in breast ultrasound; R8: junior radiologist in breast mammography; R9: senior radiologist in breast ultrasound; R10: senior radiologist in breast mammography. ^: Biopsies in benign lesions. Abbreviations: US, ultrasound; MG, mammography; BI-RADS, Breast Imaging Reporting and Data System; AUC, area under the receiver operating characteristic curve

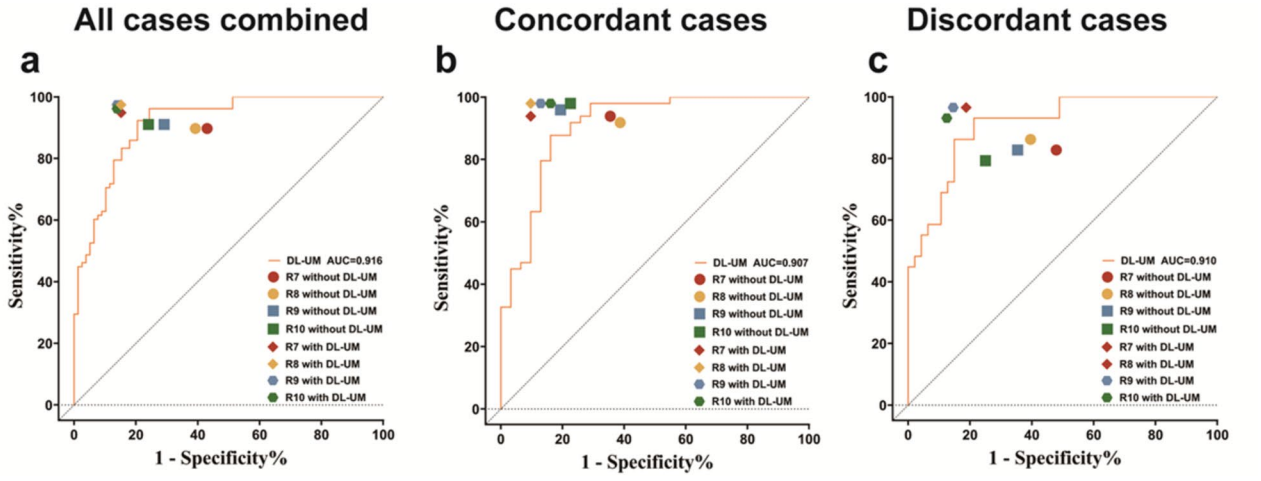

Fig. 4 The diagnostic performance of radiologists with and without the assistance of DL-UM. **a**: all cases combined; **b:** subgroup of concordant cases; **c:** subgroup of discordant cases. R7: junior radiologist in breast ultrasound; R8: junior radiologist in breast mammography; R9: senior radiologist in breast ultrasound; R10: senior radiologist in breast mammography

ultrasonic features in benign and malignant lesions, leading to potential misdiagnoses [28]. While DL-UM did not significantly exceed DL-M in specificity or DL-U in sensitivity, it maintained a sensitivity level comparable to DL-U as well as a specificity level comparable to DL-M simultaneously. This result underscored DL-UM's capability to effectively integrate the complementary diagnostic information of two modalities, mitigating their individual limitations. Such integration enhanced the performance of DL-UM with higher AUC in breast

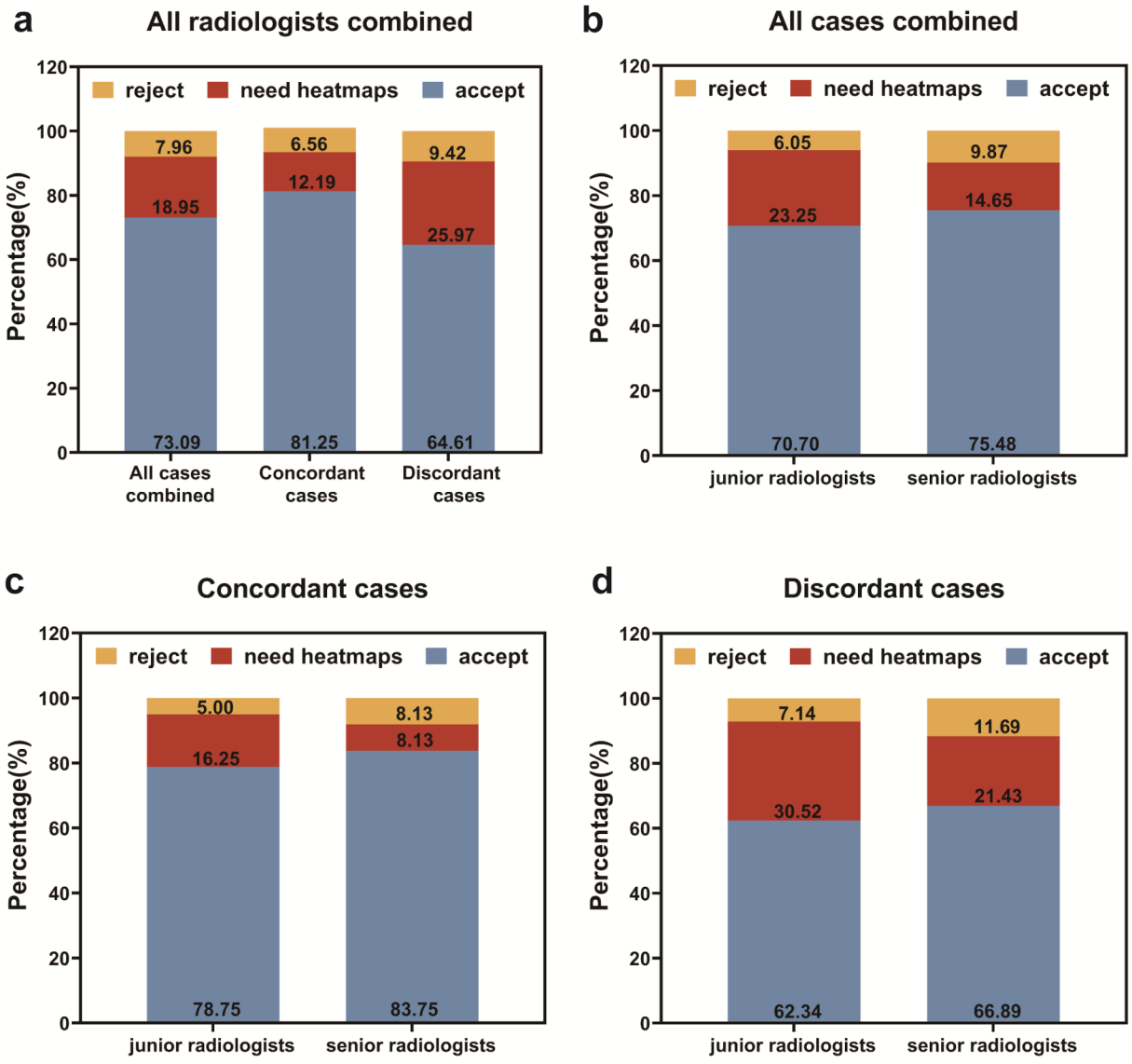

Fig. 5 Radiologists' attitude to the outputs of DL-UM. **A**: all radiologists; **B**: all cases combined; **C**: subgroup of concordant cases; **D**: subgroup of discordant cases

cancer detection and improved adaptability of DL-UM across complex clinical scenarios.

While most previously reported multimodal radiomics studies [29, 30] have shown promise in breast cancer diagnosis, but often stopped at merging outputs from individual modalities [25, 31–33], neglecting valuable complementary diagnostic information between multimodalities. By contrast, the DL-UM comprehensively extracted feature across multiple modalities, incorporating both intramodal features from identical imaging data and intermodal features from diverse imaging types. This approach focused on integrating diagnostic features from both US and MG, recognizing their complementary and correlated diagnostic role [34]. When applied to complex scenarios like discordant MG and US BI-RADS classification, DL-UM was optimised through focal loss supervision for challenging and difficult samples and minimising feature disparities between US and MG classifiers with MAE. Our results underscored the effectiveness of the DL-UM, particularly in cases with discordant MG and US classifications.

In stimulated clinical workflows without the involvement of DL-UM, even experienced radiologists exhibited reduced diagnostic accuracy in cases of discordant BI-RADS classification. However, the incorporation of DL-UM significantly enhanced diagnostic performance, particularly benefiting less experienced radiologists who achieved comparable results to their senior

Table 3 Clinical impact of radiologists' requirement for explanations using heatmaps in DL-UM-radiologists collaboration data are expressed as a count of cases in parentheses. R7: junior radiologist in breast ultrasound; R8: junior radiologist in breast mammography; R9: senior radiologist in breast ultrasound; R10: senior radiologist in breast mammography

		Percentage (%)	Unnecessary biopsies without DL-UM	Unnecessary biopsies with DL-UM	Decreased biopsies (%)	Missed malignancies without DL-UM	Missed malignancies with DL-UM	Avoided missed malignancies (%)
All cases combined	R7-R10	18.95 (119/628)	15.94 (62/389)	2.31 (8/347)	13.63	10.04 (24/239)	1.78 (5/281)	8.26
	Junior radiologists	23.25 (73/314)	20.49 (42/205)	3.45 (6/174)	17.04	11.93 (13/109)	2.14 (3/140)	9.78
	Senior radiologists	14.65 (46/314)	10.87 (20/184)	1.16 (2/173)	9.71	8.46 (11/130)	1.42 (2/141)	7.04
Concordant cases	R7-R10	12.19 (39/320)	10.81 (24/222)	1.46 (3/205)	9.35	8.16 (8/98)	3.48 (4/115)	4.69
	Junior radiologists	16.25 (26/160)	15.79 (18/114)	1.00 (1/100)	14.79	10.87 (5/46)	3.33 (2/60)	7.54
	Senior radiologists	8.13 (13/160)	5.56 (6/108)	1.90 (2/105)	3.65	5.77 (3/52)	3.64 (2/55)	2.13
Discordant cases	R7-R10	25.97 (80/308)	22.75 (38/167)	3.52 (5/142)	19.23	11.35 (16/141)	0.60 (1/166)	10.75
	Junior radiologists	30.52 (47/154)	26.37 (24/91)	6.76 (5/74)	19.62	12.70 (8/63)	1.25 (1/80)	11.45
	Senior radiologists	21.43 (33/154)	18.42 (14/76)	0 (0/68)	18.42	10.26 (8/78)	0 (0/86)	10.26

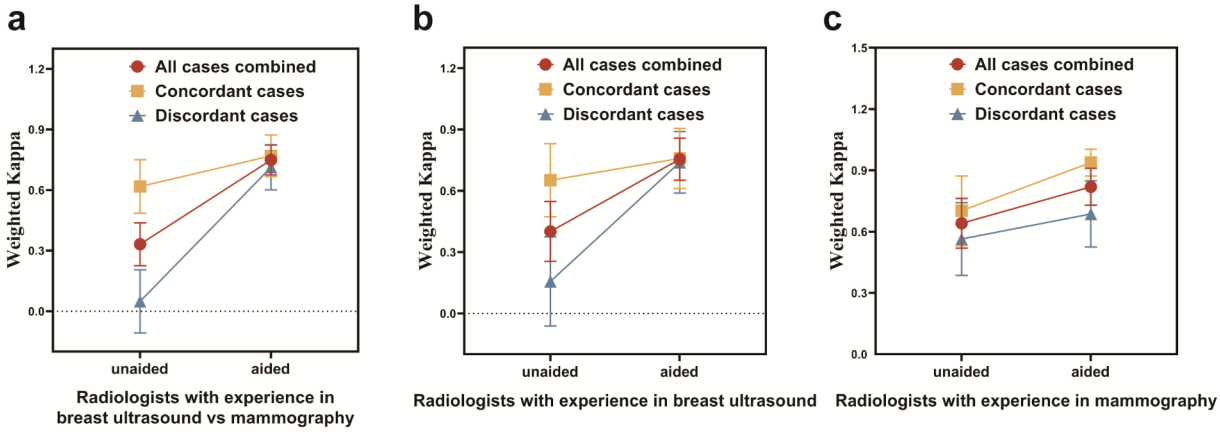

Fig. 6 Interobserver agreement between radiologists with and without the aid of DL-UM. R7: junior radiologist in breast ultrasound; R8: junior radiologist in breast mammography; R9: senior radiologist in breast ultrasound; R10: senior radiologist in breast mammography

Table 4 Interobserver agreement with and without the collaboration of DL-UM

	US R7 + R9 vs. MG R8 + R10			MG: R8 vs. R10		
	Without DL-UM	With DL-UM	Without DL-UM	With DL-UM	Without DL-UM	With DL-UM
All cases combined	0.332 (0.225–0.438)	0.749 (0.675–0.823)	0.401 (0.255–0.548)	0.755 (0.652–0.858)	0.641 (0.519–0.763)	0.820 (0.729–0.910)
Concordant cases	0.618 (0.485–0.750)	0.769 (0.666–0.873)	0.652 (0.473–0.831)	0.759 (0.611–0.906)	0.703 (0.532–0.873)	0.945 (0.870-1.000)
Discordant cases	0.048 (-0.108-0.205)	0.713 (0.601–0.824)	0.157 (-0.062-0.375)	0.739 (0.589–0.890)	0.564 (0.386–0.742)	0.687 (0.525–0.849)

US, radiologists with breast US imaging experience; MG, radiologists with breast MG imaging experience

kappa value < 0 indicated poor agreement; <0.20, slight agreement; <0.21 to 0.40, fair agreement; <0.41 to 0.60, moderate agreement;< 0.61 to 0.80, substantial agreement; 0.81 to 1.00, almost perfect agreement

R7: junior radiologist in breast ultrasound; R8: junior radiologist in breast mammography; R9: senior radiologist in breast ultrasound; R10: senior radiologist in breast mammography

counterparts. In this study, 65% (392 out of 603) of discordant cases underwent unnecessary biopsies, a concern often associated with high recall rates and biopsy rates [7, 8]. However, the addition of DL-UM resulted in a notable reduction in unnecessary biopsies. Prior research [35, 36] has reported significant improvements in diagnostic agreement among radiologists with varying levels of experience when AI is introduced, consistent with our findings. Notably, DL-UM assistance improved inter-observer consistency between US and MG, especially in cases with divergent MG and US BI-RADS classifications. Such improvement highlighted AI's potential to support radiologists with diverse backgrounds in breast imaging, reducing subjective bias and addressing uncertainties in areas such as image interpretation, result communication, and treatment decisions.

Understanding radiologists' trust in AI is crucial for its integration into clinical practice [37]. Overall,

radiologists expressed positive feedback regarding DL-UM outputs, which could enhance their confidence in image interpretation and patient management [38], particularly for junior radiologists. However, trust in DL results diminished when radiologists hesitated, particularly when US and MG classification diverge, leading to a surge in demand for AI explanations. Heatmaps play a vital role in gaining radiologists' trust in DL-UM by highlighting lesion boundaries, peri-tumoral areas, and calcifications in both US and MG images, aligning closely with visual diagnoses. Adjusting the initial diagnosis based on heatmaps improved breast cancer detection and reduced unnecessary biopsies. Our findings emphasized the necessity of providing explanations in AI implementation, especially for inconclusive diagnoses or when there is skepticism regarding DL-UM output, particularly among less experienced radiologists.

There are still some limitations in this study. First, excluding patients with follow-ups may introduce selection bias. Second, 6.5% (83/1283) of the patients underwent biopsy despite having BI-RADS classifications of 2 or 3 on both MG and US, influenced by factors beyond BI-RADS, such as palpation findings and patient preferences in routine clinical settings. This could potentially affect the practical utility of DL-UM in clinical decision-making. Third, as this study is retrospective and involves only two medical centres, further prospective studies with larger sample sizes from multiple centres are necessary to improve model performance and generalisability. Finally, in this study, ROIs were manually outlined to ensure the consistent targeting of lesions on both US and MG images. However, there are ongoing efforts to develop automated segmentation software for multimodalities to address the requirements of large-scale datasets and integrate them into clinical workflows in the future.

Conclusion

The DL-UM bimodal fusion network, integrating US and MG complementary features, showed good performance for breast lesion diagnosis, particularly for those cases of discordant US and MG BI-RADS classification. The DL-UM network showed great potential to support radiologists in breast lesion diagnosis and management, reducing unnecessary biopsies. Following prospective multicentre clinical trials, the DL-UM network may be evolved into an advanced software module, seamlessly integrating into clinical practice to aid decision-making and advance precision healthcare.

Abbreviations

US	Ultrasound
MG	Mammography
BI-RADS	Breast Imaging Reporting and Data System
DL	Deep Learning
AI	Artificial Intelligence
GAP	Global Average Pooling
MAE	Mean Absolute Error
CAMs	Class Activation Maps
ROC	Receiver Operating Characteristic

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13058-025-02033-6.

Supplementary Material 1

Acknowledgements

None.

Author contributions

Conceptualization, data curation, formal analysis, methodology, writing– original draft: Z.X., S.Z. and Y.G.; validation: J.H. and W.X.; resources: W.H. and J.Z.; investigation: C.Z., Q.D., L.L., Z.J., T.Z., S.X., J.L.; writing– review & editing: X.H., G.W., Y.Z. and Y.L.; project administration: G.W., Y.Z. and Y.L.; funding acquisition: Y.L. The authors read and approved the final manuscript.

Funding

This research was supported by the National Natural Science Foundation of China (82271998 and 82071949), College Students' Innovative Entrepreneurial Training Plan Program (202212121022), and Guangzhou Municipal Science and Technology Department: 2023 Key Research and Development Plan Projects (2023B03J1350).

Data availability

Due to the privacy of patients, the data and materials related to patients cannot be available for public access but can be obtained from the corresponding authors on reasonable request approved by the institutional review board of all enrolled centers.

Declarations

Ethics approval and consent to participate

This study was approved by the Institutional Ethics Committee of the hospital (NFEC-202012-K8) and the requirement for informed consent was waived owing to the retrospective design and use of anonymized data.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Ultrasound, Nanfang Hospital, Southern Medical University, Guangzhou 510515, China
[2]School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, Guangdong, China
[3]Department of Imaging, Zengcheng Branch of Nanfang Hospital, Southern Medical University, Guangzhou 510515, China
[4]Department of Radiology, Nanfang Hospital, Southern Medical University, Guangzhou 510515, China
[5]Department of Ultrasound, First People's Hospital of Foshan, Foshan 510515, Guangdong, China
[6]Department of Medical Imaging, Nanfang Hospital, Southern Medical University, Guangzhou 510515, Guangdong, China
[7]Department of Ultrasound, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai 200025, China
[8]Shenzhen Key Laboratory for Drug Addiction and Medication Safety, Department of Ultrasound, Institute of Ultrasonic Medicine, Peking University Shenzhen Hospital, Shenzhen Peking University-The Hong Kong University of Science and Technology Medical Center, Shenzhen 518036, China

Received: 2 April 2024 / Accepted: 22 April 2025
Published online: 14 May 2025

References

1. Hussein H, Abbas E, Keshavarzi S, Fazelzad R, Bukhanov K, Kulkarni S, Au F, Ghai S, Alabousi A, Freitas V. Supplemental breast cancer screening in women with dense breasts and negative mammography: a systematic review and meta-analysis. Radiology (2023) 306(3).
2. Kunitake J, Sudilovsky D, Johnson LM, Loh HC, Choi S, Morris PG, Jochelson MS, Iyengar NM, Morrow M, Masic A et al. Biomineralogical signatures of breast microcalcifications. Sci Adv (2023) 9(8).
3. Mann RM, Athanasiou A, Baltzer P, Camps-Herrero J, Clauser P, Fallenberg EM, Forrai G, Fuchsjager MH, Helbich TH, Killburn-Toppin F, et al. Breast cancer screening in women with extremely dense breasts: recommendations of the European society of breast imaging (EUSOBI). Eur Radiol. 2022;32(6):4036–45.
4. Barba D, Leon-Sosa A, Lugo P, Suquillo D, Torres F, Surre F, Trojman L, Caicedo A. Breast cancer, screening and diagnostic tools: all you need to know. Crit Rev Oncol Hematol. 2021;157:103174.
5. Glechner A, Wagner G, Mitus JW, Teufer B, Klerings I, Bock N, Grillich L, Berzaczy D, Helbich TH, Gartlehner G. Mammography in combination with breast ultrasonography versus mammography for breast cancer screening in women at average risk. Cochrane Database Syst Rev (2023) 3(3).
6. Xu Z, Lin Y, Huo J, Gao Y, Lu J, Liang Y, Li L, Jiang Z, Du L, Lang T, et al. A bimodal nomogram as an adjunct tool to reduce unnecessary breast biopsy following discordant ultrasonic and mammographic BI-RADS assessment. Eur Radiol. 2024;34(4):2608–18.
7. Pu H, Zhang XL, Xiang LH, Zhang JL, Xu G, Liu H, Tang GY, Zhao BH, Wu R. The efficacy of added shear wave elastography (SWE) in breast screening for women with inconsistent mammography and conventional ultrasounds (US). Clin Hemorheol Microcirc. 2019;71(1):83–94.
8. Shao SH, Li CX, Yao MH, Li G, Li X, Wu R. Incorporation of contrast-enhanced ultrasound in the differential diagnosis for breast lesions with inconsistent results on mammography and conventional ultrasound. Clin Hemorheol Microcirc. 2020;74(4):463–73.
9. Dhar T, Dey N, Borra S, Sherratt RS. Challenges of deep learning in medical image analysis—improving explainability and trust. IEEE Trans Technol Soc. 2023;4(1):68–75.
10. Yang Y, Guan S, Ou Z, Li W, Yan L, Situ B. Advances in AI-based cancer cytopathology. Interdiscip Med (2023) 1(3).
11. Yang Y, Zhong Y, Li J, Feng J, Gong C, Yu Y, Hu Y, Gu R, Wang H, Liu F, et al. Deep learning combining mammography and ultrasound images to predict the malignancy of BI-RADS US 4A lesions in women with dense breasts: a diagnostic study. Int J Surg. 2024;110(5):2604–13.
12. Assari Z, Mahloojifar A, Ahmadinejad N. A bimodal BI-RADS-guided GoogLeNet-based CAD system for solid breast masses discrimination using transfer learning. Comput Biol Med. 2022;142:105160.
13. Antropova N, Huynh BQ, Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Med Phys. 2017;44(10):5162–71.
14. Jiang M, Lei S, Zhang J, Hou L, Zhang M, Luo Y. Multimodal imaging of target detection algorithm under artificial intelligence in the diagnosis of early breast cancer. J Healthc Eng. 2022;2022:9322937.
15. Drozdov I, Dixon R, Szubert B, Dunn J, Green D, Hall N, Shirandami A, Rosas S, Grech R, Puttagunta S et al. An artificial neural network for nasogastric tube position decision support. Radiol Artif Intell (2023) 5(2).
16. Yu Q, Ning Y, Wang A, Li S, Gu J, Li Q, Chen X, Lv F, Zhang X, Yue Q, et al. Deep learning-assisted diagnosis of benign and malignant Parotid tumors based on contrast-enhanced CT: a multicenter study. Eur Radiol. 2023;33(9):6054–65.
17. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In. Ithaca: Cornell University Library, arXiv.org;; 2019.
18. Chen-Yu L, Xie S, Gallagher P, Zhang Z, Tu Z. Deeply-supervised Nets. In. Ithaca: Cornell University Library, arXiv.org;; 2014.
19. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell. 2020;42(2):318–27.
20. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In. IEEE; 2016:2921–2929.
21. Shen Y, Shamout FE, Oliver JR, Witowski J, Kannan K, Park J, Wu N, Huddleston C, Wolfson S, Millet A, et al. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. Nat Commun. 2021;12(1):5645.
22. Wang SJ, Liu HQ, Yang T, Huang MQ, Zheng BW, Wu T, Qiu C, Han LQ, Ren J. Automated breast volume scanner (ABVS)-based radiomic nomogram: a potential tool for reducing unnecessary biopsies of BI-RADS 4 lesions. Diagnostics (2022) 12(1).
23. Romero-Martin S, Elias-Cabot E, Raya-Povedano JL, Gubern-Merida A, Rodriguez-Ruiz A, Alvarez-Benito M. Stand-alone use of artificial intelligence for digital mammography and digital breast tomosynthesis screening: a retrospective evaluation. Radiology. 2022;302(3):535–42.
24. Lauritzen AD, Rodriguez-Ruiz A, von Euler-Chelpin MC, Lynge E, Vejborg I, Nielsen M, Karssemeijer N, Lillholm M. An artificial intelligence-based mammography screening protocol for breast cancer: outcome and radiologist workload. Radiology. 2022;304(1):41–9.
25. Tan T, Rodriguez-Ruiz A, Zhang T, Xu L, Beets-Tan R, Shen Y, Karssemeijer N, Xu J, Mann RM, Bao L. Multi-modal artificial intelligence for the combination of automated 3D breast ultrasound and mammograms in a population of women with predominantly dense breasts. Insights Imaging. 2023;14(1):10.
26. Yi M, Lin Y, Lin Z, Xu Z, Li L, Huang R, Huang W, Wang N, Zuo Y, Li N et al. Biopsy or follow-up: AI improves the clinical strategy of US BI-RADS 4A breast nodules using a convolutional neural network. Clin Breast Cancer (2024) 24(5).
27. Gu Y, Xu W, Liu T, An X, Tian J, Ran H, Ren W, Chang C, Yuan J, Kang C, et al. Ultrasound-based deep learning in the establishment of a breast lesion risk stratification system: a multicenter study. Eur Radiol. 2023;33(4):2954–64.
28. Lang M, Liang P, Shen H, Li H, Yang N, Chen B, Chen Y, Ding H, Yang W, Ji X, et al. Head-to-head comparison of perfluorobutane contrast-enhanced US and multiparametric MRI for breast cancer: a prospective, multicenter study. Breast Cancer Res. 2023;25(1):61.
29. Misra S, Yoon C, Kim KJ, Managuli R, Barr RG, Baek J, Kim C. Deep learning-based multimodal fusion network for segmentation and classification of breast cancers using B-mode and elastography ultrasound images. Bioeng Transl Med (2023) 8(6).
30. Xu Z, Wang Y, Chen M, Zhang Q. Multi-region radiomics for artificially intelligent diagnosis of breast cancer using multimodal ultrasound. Comput Biol Med. 2022;149:105920.
31. Jiang T, Song J, Wang X, Niu S, Zhao N, Dong Y, Wang X, Luo Y, Jiang X. Intratumoral and peritumoral analysis of mammography, tomosynthesis, and multiparametric MRI for predicting Ki-67 level in breast cancer: a radiomics-based study. Mol Imaging Biol. 2022;24(4):550–9.
32. Chen S, Guan X, Shu Z, Li Y, Cao W, Dong F, Zhang M, Shao G, Shao F. A new application of multimodality radiomics improves diagnostic accuracy of nonpalpable breast lesions in patients with microcalcifications-only in mammography. Med Sci Monit. 2019;25:9786–93.
33. Huang W, Tan K, Zhang Z, Hu J, Dong S. A review of fusion methods for omics and imaging data. IEEE/ACM Trans Comput Biol Bioinform. 2023;20(1):74–93.
34. Xi X, Li W, Li B, Li D, Tian C, Zhang G. Modality-correlation embedding model for breast tumor diagnosis with mammography and ultrasound images. Comput Biol Med. 2022;150:106130.
35. Kim HJ, Choi WJ, Gwon HY, Jang SJ, Chae EY, Shin HJ, Cha JH, Kim HH. Improving mammography interpretation for both novice and experienced readers: a comparative study of two commercial artificial intelligence software. Eur Radiol. 2024;34(6):3924–34.
36. Lopez-Almazan H, Javier PF, Larroza A, Perez-Cortes JC, Pollan M, Perez-Gomez B, Salas TD, Casals M, Llobet R. A deep learning framework to classify breast density with noisy labels regularization. Comput Methods Programs Biomed. 2022;221:106885.
37. Ho S, Doig GS, Ly A. Attitudes of optometrists towards artificial intelligence for the diagnosis of retinal disease: a cross-sectional mail-out survey. Ophthalmic Physiol Opt. 2022;42(6):1170–9.
38. Calisto FM, Santiago C, Nunes N, Nascimento JC. BreastScreening-AI: evaluating medical intelligent agents for human-AI interactions. Artif Intell Med. 2022;127:102285.

Publisher's note