Original Article

Ultrasound-based deep learning radiomics for multi-stage assisted diagnosis in reducing unnecessary biopsies of BI-RADS 4A lesions

Xiangyu Lu[1#^], Yun Lu[2#], Wuyuan Zhao[1], Yunliang Qi[3], Hongjuan Zhang[1], Wenhao Sun[1], Huaikun Zhang[1], Pei Ma[1], Ling Guan[2], Yide Ma[1^]

[1]School of Information Science and Engineering, Lanzhou University, Lanzhou, China; [2]Department of Ultrasound, Gansu Provincial Cancer Hospital, Lanzhou, China; [3]Zhejiang Laboratory, Hangzhou, China

Contributions: (I) Conception and design: X Lu, W Zhao, Y Ma; (II) Administrative support: Y Ma, L Guan; (III) Provision of study materials or patients: L Guan, Y Lu; (IV) Collection and assembly of data: X Lu, L Guan, Y Lu; (V) Data analysis and interpretation: X Lu, Y Lu, W Zhao; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work as co-first authors.

Correspondence to: Yide Ma, PhD. School of Information Science and Engineering, Lanzhou University, No. 222 South Tianshui Road, Lanzhou 730030, China. Email: ydma@lzu.edu.cn; Ling Guan, MBBS. Department of Ultrasound, Gansu Provincial Cancer Hospital, No. 2 East Street Xiaoxihu, Lanzhou 730050, China. Email: guanling1966@sina.com.

Background: Even with the Breast Imaging Reporting and Data System (BI-RADS) guiding risk stratification on ultrasound (US) images, inconsistencies in diagnostic accuracy still exist, leading patients being subjected to unnecessary biopsies in clinical practice. This study investigated the construction of deep learning radiomics (DLR) models to improve the diagnostic consistency and reduce the unnecessary biopsies for BI-RADS 4A lesions.

Methods: A total of 746 patients with breast lesions were enrolled in this retrospective study. Two DLR models based on US images and clinical variables were developed to conduct breast lesion risk re-stratification as BI-RADS 3 or lower and BI-RADS 4A or higher (DLR_LH), while simultaneously identifying BI-RADS 4A lesions with low malignancy probabilities to avoid unnecessary biopsy (DLR_BM). A three-round reader study with a two-stage artificial intelligence (AI)-assisted diagnosis process was performed to verify the assistive capability and practical benefits of the models in clinical applications.

Results: The DLR_LH model achieved areas under the receiver operating characteristic curve (AUCs) of 0.963 and 0.889 with sensitivities of 92.0% and 83.3%, in the internal and external validation cohorts, respectively. The DLR_BM model exhibited AUCs of 0.977 and 0.942, with sensitivities of 94.1% and 86.4%, respectively. Both models were evaluated using integrated features of US images and clinical variables. Ultimately, 27.7% of BI-RADS 4A lesions avoided unnecessary biopsies. In the three-round reader study, all readers achieved significantly higher diagnostic accuracy and specificity, while maintaining outstanding sensitivity comparable to human experts, both before and after model assistance ($P<0.05$). These findings demonstrate the positive impact of the DLR models in assisting radiologists to enhance their diagnostic capabilities.

Conclusions: The models performed well in breast US imaging interpretation and BI-RADS risk re-stratification, and demonstrated potential in reducing unnecessary biopsies of BI-RADS 4A lesions, indicating the promising applicability of the DLR models in clinical diagnosis.

^ ORCID: Xiangyu Lu, 0000-0002-2095-6830; Yide Ma, 0000-0003-4394-7029.

Keywords: Breast tumor; deep learning radiomics (DLR); ultrasonography; Breast Imaging Reporting and Data System stratification (BI-RADS stratification); diagnosis

Submitted Mar 21, 2024. Accepted for publication Dec 03, 2024. Published online Feb 07, 2025.
doi: 10.21037/qims-24-580
View this article at: https://dx.doi.org/10.21037/qims-24-580

Introduction

Breast cancer (BC) is the most commonly diagnosed cancer, with an incidence rate that continues to increase by about 0.5% per year and a progressively decreasing patient age at diagnosis (1,2). Epidemiological data indicate that several risk factors, including gender, age, family history, menstrual history, and genetic mutations, are related to BC (3). Breast imaging examinations are conducted for early detection, location, and malignancy probability evaluation of breast tumors during clinical diagnosis. Currently, the preferred imaging methods for breast tumor diagnosis are mammography, ultrasound (US), and magnetic resonance imaging (MRI). US is a widely used modality complementary to mammography, especially for detecting small and non-palpable lesions in dense breasts (4), given the limited sensitivity of mammography (5). Moreover, breast US plays an essential role throughout the clinical process of BC diagnosis and treatment owing to its advantages of lower cost, real-time assessment capability, and non-ionizing radiation compared to other imaging modalities (6).

The Breast Imaging Reporting and Data System (BI-RADS) (7) lexicon for US has been developed to standardize breast imaging terminology for describing lesion characteristics, providing risk stratification and management recommendations to ensure uniformity with clinical practice. Radiologists primarily focus on identifying the presence of lesions and suspicious lesions with a likelihood of malignancy, guided by the BI-RADS lexicon. This guideline establishes associations between US findings and the final assessment using categories 1–6 by considering lesion features such as shape, margin, orientation, and micro-calcification, and describes the presence of lesions along with the probability of malignancy. Among these categories, BI-RADS 2 and 3 indicate lesions with no suspicion or a low probability of malignancy (<2%), which for which follow-up or continued surveillance are recommended. BI-RADS 4 and 5 represent suspicious lesions with a likelihood of malignancy ranging from 2% to 95%, necessitating biopsies to confirm the pathological properties (3). Therefore, tumors assessed as BI-RADS 2/3 are considered low risk, whereas those evaluated as BI-RADS 4/5 are considered high risk. Based on the specific imaging characteristics of the lesion, the BI-RADS 4 category is further subdivided into 4A, 4B, and 4C to reflect different levels of malignant risk. Due to the heterogeneity and variety of breast lesions (8,9), there are ambiguous cutoff points for adjacent BI-RADS scores, particularly between BI-RADS 3 and BI-RADS 4A. This categorization is crucial as the score influences the subsequent diagnostic and treatment recommendations regarding whether patients need short-term follow-up imaging or a biopsy (10). Meanwhile, subjective imaging interpretation and inconsistent readings among radiologists have resulted in a high number of false-positive findings and unnecessary biopsies of lesions (11,12), most of which were ultimately benign, particularly in the case of BI-RADS 4A lesions. It has been shown that even after short-term re-evaluation of images, up to 29% of BI-RADS assessments may be reclassified (13). Therefore, improving the assessment of BI-RADS subcategories and further reducing the rate of unnecessary biopsies among BI-RADS 4A breast lesions is essential for making more accurate diagnostic recommendations, thus underscoring the urgent need to establish clinical decision-making support models.

In recent years, artificial intelligence (AI)-assisted systems have become increasingly important in multiple medical fields (14-16), and deep learning (DL) techniques have been proposed to assist radiologists in automatic detection and classification of breast imaging findings. Many studies have built prediction models for the classification of breast lesions as benign or malignant (17-21), and a few studies have focused on BI-RADS subcategory assessment (9,21-23). In fact, it is difficult for radiologists to precisely determine whether a breast lesion is benign or malignant, and such binary classification does not completely align with clinical practice because radiologists primarily evaluate BI-RADS subcategories by assessing the malignancy probabilities of breast lesions. Radiomics approaches have shown

potential for non-invasive assessment of cancer by capturing quantitative features (24), and deep learning radiomics (DLR) has made significant progress in many medical tasks, including BC diagnosis (25-27). To achieve higher diagnostic performance, transfer learning was utilized in this study (28) by employing the publicly available Breast Ultrasound Image (BUSI) (29) dataset.

We committed to replicating and assisting radiologists' decision-making in breast tumor interpretation and risk stratification according to the BI-RADS lexicon. Two DLR models were established to assess BI-RADS risk, categorizing lesions as low-risk (BI-RADS 2/3) or high-risk (BI-RADS 4/5). Additionally, binary classification was implemented for distinguishing between benign and malignant BI-RADS 4A lesions to help reduce false-positive rates (FPRs) and avoid unnecessary biopsy. Both the US representation and clinical variables were used to describe patient characteristics, and the prediction scores from the DLR models were provided to assist radiologists in diagnosis. The ability of the models to assist radiologists was explored to understand the potential value of the models in clinical practice. We present this article in accordance with the TRIPOD+AI reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-24-580/rc).

Methods

Patient population

We conducted a retrospective study from November 2020 to November 2023 from the Gansu Provincial Cancer Hospital. A total of 944 patients with breast tumors were included and confirmed by surgical or needle biopsy at this institution. Some 746 patients were eligible for the final study, comprising 438 patients with benign tumors and 308 patients with BC, including 11 BI-RADS 2, 280 BI-RADS 3, 148 BI-RADS 4A, 73 BI-RADS 4B, 67 BI-RADS 4C, and 167 BI-RADS 5 lesions, categorized by an expert radiologist with over 30 years of experience. These patients were divided into different cohorts for construction of models, and the inclusion and exclusion criteria are shown in *Figure 1* and Figure S1. More details are provided in the supplementary materials.

This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Ethics Committee of the Gansu Provincial Cancer Hospital (No. A202303090008) and the requirement for informed consent for this retrospective analysis was waived.

Data acquisition and image processing

All breast US examinations were performed by radiologists with over 10 years of breast imaging experience, using the EPIQ7 (Philips Healthcare, Erlangen, Netherlands) and Resona R9 (Mindray Medical, Shenzhen, China) equipment. Besides the imaging data (US video, US images, and color Doppler flow images), the ultrasonic diagnosis report, which included variables such as patient's basic information, clinical symptoms, menstruation, marriage and childbearing history, family history, and previous history of BC, was also recorded. All cases (n=746) were grouped into two distinct data sets, namely data_LH and data_BM, based on the BI-RADS risk level and pathology results, respectively. The recently collected patients (n=109) from May 2023 served as an external validation (EV) cohort, and the remaining patients (n=637) were randomly split into a training cohort and an internal validation (IV) cohort at a ratio of 8:2. The IV cohort and EV cohort were used to verify the performance of the final trained model. To prepare for training, we performed image pre-processing by cropping the largest regions of interest (ROIs) surrounding the lesions and resizing them to 224 × 224 pixels.

DLR model development

Both DLR models utilized the ResNet-50 (30) architecture as backbone (*Figure 2*). Compared to the original ResNet-50 framework, we modified it by replacing its fully connected layer with two new fully connected layers. These layers output a feature vector of size 128 as the imaging embedding and two neurons for generating predictive probabilities (Figure S2). To capture more accurate imaging features, the DLR_LH and DLR_BM models were respectively trained to save the best-performing parameters according to the area under the curve (AUC) performance on IV cohort, and the best-performing model was used to learn the image representations. The Adam optimizer with a batch size of 16 images and a maximum of 100 epochs was used for models training with a five-fold cross-validation strategy on an NVIDIA GeForce RTX 3080Ti (NVIDIA, Santa Clara, CA, USA) under Python 3.8 (Python Software Foundation, Wilmington, DE, USA) and Pytorch 2.1.2 (Meta AI, New York, NY, USA).

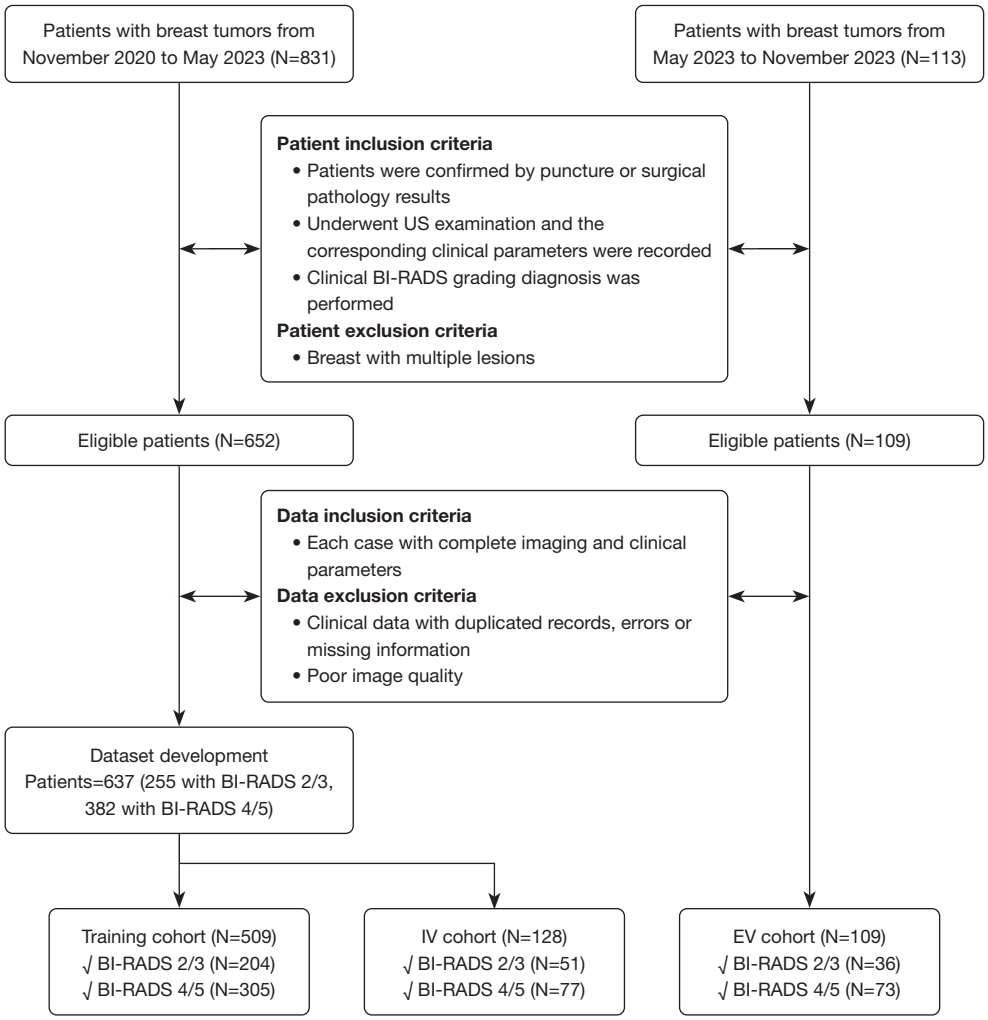

Figure 1 Flowchart of patient recruitment for DLR_LH model construction. US, ultrasound; BI-RADS, Breast Imaging Reporting and Data System; IV, internal validation; EV, external validation.

The main tasks of this study were as follows: (I) to conduct risk stratification of breast tumors as low-risk probability (BI-RADS 2/3) or high-risk probability (BI-RADS 4/5), and (II) to predict the malignancy rate among breast lesions diagnosed in the BI-RADS 4A subcategory. The construction procedure for both DLR models involved ResNet-50-based model training, image representation learning, and result predictions based on the integrated multilayer perceptron (MLP) classifier (Figure S3). The MLP-based classifier was developed through the integration of image embedding and clinical variables. For BI-RADS subcategories, BI-RADS 4 or higher represented potentially malignant lesions, and biopsy was recommended to confirm the pathological properties (31). However, a previous study revealed that approximately 78% of BI-RADS 4 lesions were confirmed as benign but still underwent unnecessary biopsies. Notably, in their subdivided BI-RADS 4A cases, the FPR reached as high as 92% (32). To reduce the FPR among breast lesions diagnosed as category BI-RADS 4A, we established an additional DLR_BM model utilizing the data_BM dataset, employing the same model structure and training strategy as DLR_LH. To better interpret the model diagnosis process, both DLR models generated heatmaps for visualization (33), which could highlight the regions of model interest.

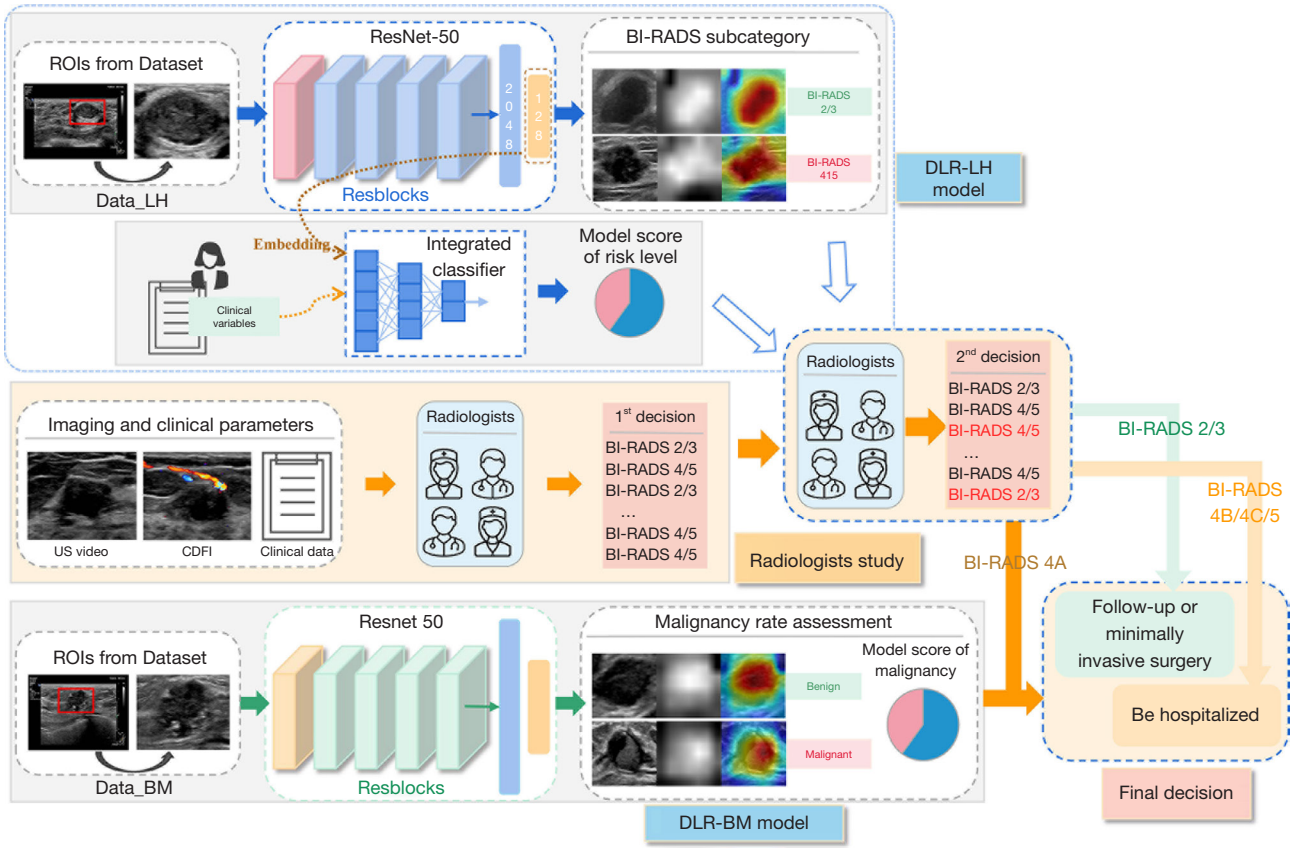

Figure 2 Workflow of the DLR models for BI-RADS risk re-stratification assessment. The imaging representations learned from the DL models were integrated with the clinical variable for model construction and then providing two-stage assistance with the heatmaps and AI scores. The radiologists gave an initial decision on each lesion and refined their decisions, if uncertain, based on the information provided by the DLR_LH and DLR_BM models. ROI, region of interest; DLR, deep learning radiomics; DL, deep learning; US, ultrasound; CDFI, color Doppler flow imaging; BI-RADS, Breast Imaging Reporting and Data System; AI, artificial intelligence.

Three-round reading with two-stage assistance

A three-round radiologists study was conducted to compare the diagnostic performance and further explore the capabilities of assistance (*Figure 2*). Three radiologists, comprising one senior radiologist with more than 10 years of experience and two junior radiologists with more than seven years of experience, participated in this study. Each radiologist was asked to evaluate the lesions on the validation cohort independently, while blinded to the US reports and pathology results.

To align with real clinical scenarios, US images, color Doppler flow imaging (CDFI) images, and clinical data for each patient were available for radiologists to make an initial diagnosis in the first round. In the second round, the radiologists were asked to refine their initial diagnostic decisions with first-stage assistance by referring to the heatmaps and scores from the DLR_LH model. After this round, radiologists could more efficiently categorize the lesions as BI-RADS 3 or lower, and BI-RADS 4A or higher. For lesions categorized as BI-RADS 4A, we conducted a third round of radiologist study with second-stage assistance using the heatmaps and scores of malignancy probability from the DLR_BM model for a final decision. The assisting procedure provided additional evidence for subsequent clinical treatment decisions, and the performance of the radiologists' decision-making process was compared before and after receiving help of DLR models.

Statistical analysis

For statistical analysis, continuous variables such as age,

Table 1 Baseline characteristics of patients with breast tumors

Characteristic	Patient cohort with breast tumors (N=746)	
	Benign	Malignant
Number of patients	438	308
Age (years)	38.77±11.02	52.00±10.66
Acquisition equipment		
EPIQ7/Resona R9	323/115	211/97
Age		
<40 years	221	34
≥40 years	217	274
Lesion size		
≤2 cm	227	121
2 cm, ≤5 cm	205	179
>5 cm	6	8
Subcategory of BI-RADS		
2	10	1
3	278	2
4A	124	24
4B	22	51
4C	3	64
5	1	166
Family history of BC		
Yes	1	1
No	437	307
Histologic type (benign)		
Fibroadenoma	268	–
Intraductal papilloma	14	–
Adenopathy	38	–
Benign phyllodes tumor	13	–
Other	105	–
Histologic type (malignant)		
Invasive ductal carcinoma	–	269
Ductal carcinoma *in situ*	–	18
Invasive lobular carcinoma	–	5
Other	–	16

Data are expressed as the mean ± standard deviation or number. Family history of BC refers to breast cancer in first-degree relatives. BI-RADS, Breast Imaging Reporting and Data System; BC, breast cancer.

were described using mean and standard deviation, and comparisons between two groups were made using the Mann-Whitney *U* test and Student's *t*-test. Categorical variables were described as numbers and percentages, and the chi-square test was used for between-group comparisons. Receiver operating characteristic (ROC) analysis was used to assess the diagnostic performance of the DLR models with a 95% confidence interval (CI) derived from Monte Carlo simulation. The evaluation metrics, including accuracy, sensitivity, and specificity were derived from the confusion matrices. Accuracy represents the percentage of patients correctly diagnosed. Sensitivity refers to the proportion of true positive predictions among all positive cases, whereas specificity refers to the proportion of true negative predictions among all negative cases. Radiologists' performance before and after DLR models' assistance was evaluated using McNemar's test. Statistical significance was set at $P<0.05$. To assess the value of DLR models in reducing unnecessary biopsy rates, the original BI-RADS 4A breast lesions could be downgraded to BI-RADS 3 if the DLR_LH model predicted a low risk level and the DLR_BM model simultaneously predicted benign findings.

Results

Patient characteristics

A total of 746 patients with breast lesions, with an average age of 44.23±12.67 years (range, 13–79 years), were enrolled for further analysis. Two clinically applicable DLR models were constructed to perform AI-assisted diagnosis. Among them, 438 patients (average age 38.77±11.02 years) were confirmed to have benign lesions, including fibroadenoma, intraductal papilloma, adenopathy, benign phyllodes tumor, and some other histologic types. Meanwhile, 308 patients (average age 52±10.66 years) were confirmed to have malignant lesions, such as invasive ductal carcinoma, ductal carcinoma *in situ*, invasive lobular carcinoma, and some other histologic types. The baseline characteristics and some typical cases are listed in *Table 1* and Figure S4.

Associations of US features with BI-RADS subcategories and pathology results

Table 2 displays the analysis of the base clinical and gray-scale US image features as well as their associations with BI-RADS subcategories, and shows a highly significant difference in patient ages between BI-RADS 2/3 and BI-

Table 2 Base clinical and gray-scale ultrasound imaging features and their associations with BI-RADS 2/3 and BI-RADS 4/5 of breast tumors

Feature name	Total (n=746)	BI-RADS 2/3 (n=291)	BI-RADS 4/5 (n=455)	P value
Age (years)	44.23±12.67	36.64±10.76	49.09±11.35	<0.001**
Dist_LesionToNipple (mm)	11.71±12.10	11.33±12.89	11.95±11.57	0.491
Dist_LesionToSurface (mm)	6.28±5.67	6.05±5.81	6.42±5.57	0.381
Aspect ratio	0.62±0.22	0.56±0.17	0.65±0.24	<0.001**
Tissue composition				<0.001**
Fat	3 (0.40)	2 (0.69)	1 (0.22)	
Fibroglandular	303 (40.62)	198 (68.04)	105 (23.08)	
Heterogeneous	440 (58.98)	91 (31.27)	349 (76.70)	
Shape				<0.001**
Oval	298 (39.95)	251 (86.25)	47 (10.33)	
Round	6 (0.80)	2 (0.69)	4 (0.88)	
Irregular	442 (59.25)	38 (13.06)	404 (88.79)	
Orientation				<0.001**
Parallel	428 (57.37)	261 (89.69)	167 (36.70)	
Not parallel	318 (42.63)	30 (10.31)	288 (63.30)	
Margin				<0.001**
Circumscribed				<0.001**
No	460 (61.66)	26 (8.93)	434 (95.38)	
Yes	286 (38.34)	265 (91.07)	21 (4.62)	
Indistinct				<0.001**
No	565 (75.74)	284 (97.59)	281 (61.76)	
Yes	181 (24.26)	7 (2.41)	174 (38.24)	
Angular				<0.001**
No	520 (69.71)	290 (99.66)	230 (50.55)	
Yes	226 (30.29)	1 (0.34)	225 (49.45)	
Microlobulated				<0.001**
No	529 (70.91)	271 (93.13)	258 (56.70)	
Yes	217 (29.09)	20 (6.87)	197 (43.30)	
Spiculated				<0.001**
No	663 (88.87)	290 (99.66)	373 (81.98)	
Yes	83 (11.13)	1 (0.34)	82 (18.02)	
Posterior features				0.006*
No posterior features	727 (97.45)	291 (100.00)	436 (95.82)	
Enhancement	1 (0.13)	0	1 (0.22)	
Shadowing	15 (2.01)	0	15 (3.30)	
Combined pattern	3 (0.40)	0	3 (0.66)	

Table 2 (*continued*)

Table 2 (*continued*)

Feature name	Total (n=746)	BI-RADS 2/3 (n=291)	BI-RADS 4/5 (n=455)	P value
Echo pattern				0.019
Anechoic	8 (1.07)	7 (2.41)	1 (0.22)	
Complex cystic/solid	17 (2.28)	8 (2.75)	9 (1.98)	
Hypoechoic	711 (95.31)	270 (92.78)	441 (96.92)	
isoechoic	3 (0.40)	1 (0.34)	2 (0.44)	
Heterogeneous	7 (0.94)	5 (1.72)	2 (0.44)	
Micro-calcifications				<0.001**
No	495 (66.35)	287 (98.63)	208 (45.71)	
Yes	251 (33.65)	4 (1.37)	247 (54.29)	
Associate features				<0.001**
Architectural distorted				<0.001**
No	723 (96.92)	291 (100.00)	432 (94.95)	
Yes	23 (3.08)	0	23 (5.05)	
Duct changes				0.969
No	681 (91.29)	266 (91.41)	415 (91.21)	
Yes	65 (8.71)	25 (8.59)	40 (8.79)	
Skin changes				0.276
No	742 (99.46)	291 (100.00)	451 (99.12)	
Yes	4 (0.54)	0	4 (0.88)	
Edema				0.684
No	744 (99.73)	290 (99.66)	454 (99.78)	
Yes	2 (0.27)	1 (0.34)	1 (0.22)	
CDFI features				<0.001**
No blood flow	474 (63.54)	279 (95.88)	195 (42.86)	
Intralesional, Adler I	91 (12.20)	9 (3.09)	82 (18.02)	
Intralesional, Adler II	155 (20.78)	0	155 (34.07)	
Intralesional, Adler III	19 (2.55)	0	19 (4.18)	
Perifocal	7 (0.94)	3 (1.03)	4 (0.88)	

Data are expressed as the mean ± standard deviation or number (percentage). *, P<0.05; **, P<0.01. Adler 0: no obvious blood flow signals; Adler I: one or two small spot-like blood flow signals; Adler II: strip blood flow signals could be seen; Adler III: reticular blood flow signals could be detected. BI-RADS, Breast Imaging Reporting and Data System; CDFI, color Doppler flow imaging; Dist_LesionToNipple, distance from lesion to the nipple; Dist_LesionToSurface, distance from lesion to the breast surface.

RADS 4/5 subcategories. Patients with BI-RADS 3 or lower-level tumors were generally younger (P<0.001) and more likely to have a parallel orientation, circumscribed margins, and oval shape (P<0.001). Tumors of BI-RADS 4/5 were found to have larger aspect ratio closer to 1 (P<0.001), mixed glandular types (P<0.001), and were more likely to have instinct micro-lobulated or spiculated margins (P<0.001). Besides, BI-RADS 4/5 tumors are usually accompanied by micro-calcifications (P<0.001), comparatively abundant blood flow signals in CDFI (P<0.001), and architectural

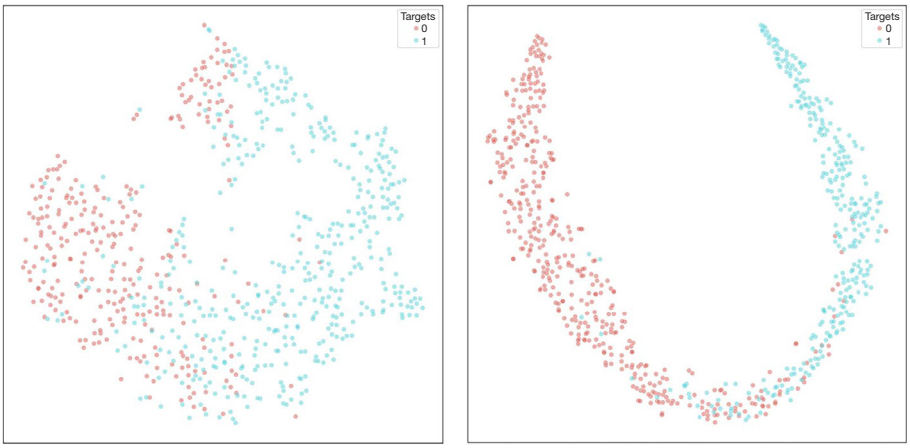

Figure 3 The t-SNE visualization of feature embedding. The left was feature vector-learned from the DLR_LH model based on data_LH dataset; the right was feature vector learned from the DLR_BM model based on data_BM dataset. t-SNE, t-distributed stochastic neighbor embedding approach.

distortion (P<0.001). However, some relevant information was not mentioned in certain US reports. Posterior features (P<0.01) and echo pattern (P<0.05) were characteristic US features. The associations between the US features and the pathological results are analyzed in detail in Table S1, with some US features also showing significant differences. Several studies have attempted to investigate the imaging feature relevance between the BI-RADS risk stratification and pathology results to improve the performance of breast masses classification, but this remains a challenge (21,34).

Testing and performance evaluation of the DLR models

The distribution of US embeddings and DLR models predictions

To further investigate the effectiveness of the learned US representations from the modified ResNet-50 approach, we utilized t-distributed stochastic neighbor embedding (t-SNE) to visualize the distribution of US embeddings. Using the two groups of US image embeddings obtained from the DLR_LH and DLR_BM models as input, *Figure 3* illustrates separately the distributions of the learned 128-dimensional feature vectors, which were dimension-reduced to a two-dimensional (2D) space. The learned features, with different class labels, were distinctly separated into two semantically clusters, indicating that our approach can learn image representations with enhanced discriminant capability.

In addition, we employed confusion matrices to describe the distributions of the DLR model predictions (*Figure 4*). When integrated with the clinical variable (age), the prediction performance of both DLR models was enhanced, bringing it closer to the level of clinicians' discrimination. Moreover, the misclassified cases of the DLR_LH model were mainly concentrated in the BI-RADS 3 and BI-RADS 4A subcategories in this study, which aligns with the clinical diagnostic practice, as these tumors often exhibit similar imaging appearances.

BI-RADS subcategory assessment of the DLR_LH model

The DLR_LH model demonstrated good performance for BI-RADS subcategory prediction. Using only the US feature embeddings, the DLR_LH achieved AUCs of 0.958 and 0.883, accuracies of 89.8% and 78.7%, sensitivities of 92.0% and 81.9%, and specificities of 86.8% and 72.2% in the IV and EV cohorts, respectively. When incorporating the clinical variable, the performance was further improved, resulting in higher AUCs of 0.963 and 0.889, accuracies of 90.6% and 79.6%, sensitivities of 92.0% and 83.3%, and specificities of 88.7% and 72.2%, respectively (*Figure 5* and *Table 3*). The improved performance of our DLR_LH model indicated enhanced alignment with expert diagnostic capability, providing a more reliable reference for initial BI-RADS subcategory predictions.

DLR_BM model in helping reducing unnecessary biopsies of BI-RADS 4A

The DLR_BM model achieved significantly outstanding performance. Experiments showed that the incorporated

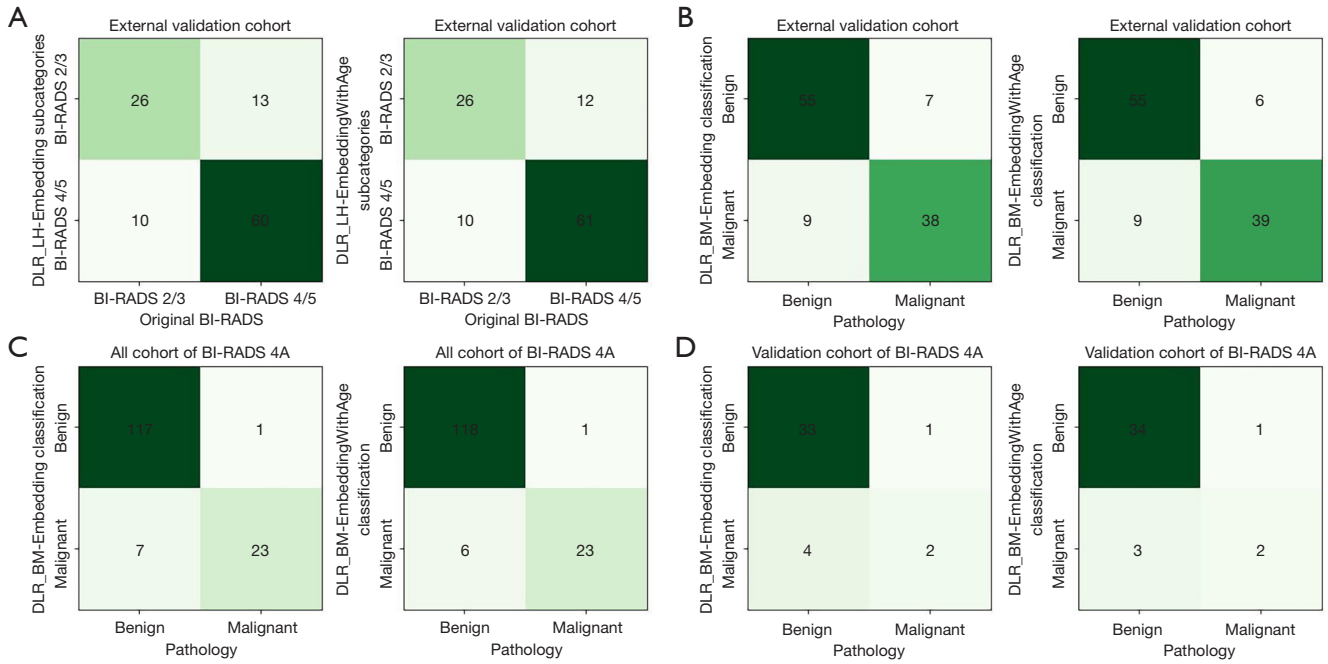

Figure 4 Confusion matrices of DLR model predictions respectively based on the ultrasound embeddings and embedding-age integrated features. Comparisons of DLR-LH model results for BI-RADS subcategories assessment (A) and of DLR-BM model classification results with the pathological findings (B) for the EV cohort. (C) and (D) showed comparisons of DLR-BM model classification results with the pathological findings, respectively for the whole and validation BI-RADS 4A cases. EV, external validation; BI-RADS, Breast Imaging Reporting and Data System; DLR, deep learning radiomics.

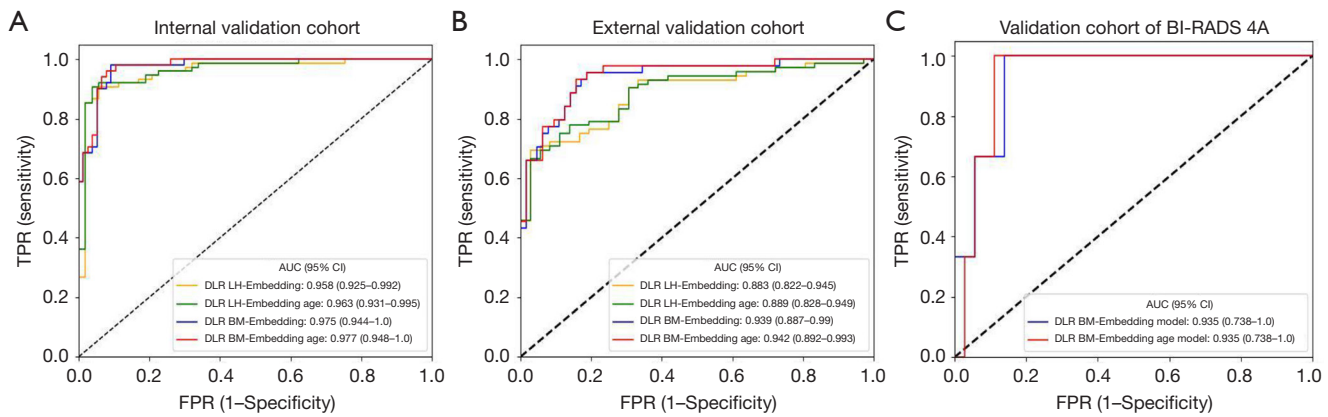

Figure 5 AUCs for BI-RADS subcategories assessment performance (DLR_LH model) and benign and malignant diagnosis performance (DLR-BM model), respectively based on image embedding, embedding-age-integrated features. (A) Results on the internal validation cohort, (B) results on the external validation cohort, (C) results on the validation cohort of BI-RADS 4A lesions. AUCs, areas under the curve; BI-RADS, Breast Imaging Reporting and Data System; TPR, true positive rate; FPR, false positive rate; CI, confidence interval.

Table 3 Performance evaluation and comparison of the DLR-LH and DLR-BM model in IV and EV cohorts, respectively based on image embedding and embedding-age-integrated feature

Model	Feature modality	Cohort	AUC (95% CI)	ACC (95% CI)	SENS (95% CI)	SPEC (95% CI)
DLR-LH	Embedding	IV	0.958 (0.925, 0.992)	89.8 (84.6, 95.1)	92.0 (85.8, 98.2)	86.8 (77.5, 96.1)
		EV	0.883 (0.822, 0.945)	78.7 (70.9, 86.5)	81.9 (72.9, 91.0)	72.2 (67.1, 87.4)
	Embedding_Age	IV	0.963 (0.931, 0.995)	90.6 (85.5, 95.7)	92.0 (85.8, 98.2)	88.7 (79.9, 97.4)
		EV	0.889 (0.828, 0.949)	79.6 (71.9, 87.3)	83.3 (74.6, 92.1)	72.2 (67.1, 87.4)
DLR-BM	Embedding	IV	0.975 (0.944, 1.0)	92.2 (87.5, 96.9)	92.2 (84.6, 99.7)	92.2 (86.1, 98.3)
		EV	0.939 (0.887, 0.99)	85.2 (78.4, 92.0)	84.1 (73.0, 95.2)	85.9 (77.3, 94.6)
	Embedding_Age	IV	0.977 (0.948, 1.0)	93.8 (89.5, 98.0)	94.1 (87.5, 99.0)	93.5 (87.9, 99.1)
		EV	0.942 (0.892, 0.993)	86.1 (79.5, 92.7)	86.4 (75.9, 96.8)	85.9 (77.3, 94.6)
	Embedding	BI-RADS 4A	0.935 (0.738, 1.0)	87.2 (76.2, 98.0)	1.0	86.1 (74.4, 97.8)
	Embedding_Age		0.935 (0.738, 1.0)	89.7 (79.9, 99.6)	1.0	88.9 (78.3, 99.5)

IV, internal validation; EV, external validation; AUC, area under the curve; CI, confidence interval; ACC, accuracy; SENS, sensitivity; SPEC, specificity; BI-RADS, Breast Imaging Reporting and Data System.

features of the US embedding and clinical variable yielded improved performance, with AUCs of 0.977 and 0.942, accuracies of 93.8% and 86.1%, sensitivities of 94.1% and 86.4%, and specificities of 93.5% and 85.9% in the IV and EV cohorts (*Figure 5* and *Table 3*), respectively. Here, we particularly focused on whether the DLR models could help reduce unnecessary biopsies of BI-RADS 4A, which have a high FPR in clinical practice. The risk re-stratification process of BI-RADS 4A is outlined in *Figure 6*. Patients with BI-RADS 4A lesions were input into the two DLR models respectively. If the lesions were predicted as low risk by DLR_LH and benign by DLR_BM, a downgrade to BI-RADS 3 was recommended. Conversely, if lesions were assessed as high risk by DLR_LH and malignant by DLR_BM, an upgrade to a higher grade was advised. Ultimately, clinically graded BI-RADS 4A cases (n=148) were selected to verify the effectiveness of the DLR models in helping to reduce unnecessary biopsies (Figure S5). It was revealed that 27.7% (41/148) of BI-RADS 4A breast lesions were downgraded to BI-RADS 3, thus avoiding biopsies. These downgraded lesions were all pathologically confirmed as benign tumors. Furthermore, 54.1% (80/148) of the lesions were recommended to maintain the risk level of BI-RADS 4A, with a malignancy probability of 2.5% (2/80). Meanwhile, 18.2% (27/148) of lesions were advised to upgrade to a higher risk level, with a malignancy probability of 81.5% (22/27). Compared with the clinical diagnosis results, the malignancy rate for BI-RADS 4A decreased from 16.2% to 2.5%, which falls within the lexicon value range of 2–10%. This co-decision strategy for BI-RADS 4A risk re-stratification, combining the DLR_LH and DLR_BM models, draws inspiration from the "expert consultation" process in clinical practice, which comprehensively integrates expert knowledge and AI scores to enhance decision-making accuracy.

Interpretability of the DLR models

In this study, we conducted a three-round reading and two-stage assisted diagnosis process that conformed to the clinical auxiliary diagnosis procedure. To elucidate the decision-making process, we employed Grad-CAM (33), which generated heatmaps to indicate the ROIs to the model (*Figure 7*). The DLR models generally focused on both the border and internal regions of tumors to distinguish the lesion categories. The DLR_LH model focused on the marginal features of BI-RADS 2/3 tumors in most US images, whereas for high-risk tumors, the crucial regions were predominantly inside the tumor, followed by marginal regions. The DLR_BM gave more attention to both marginal and internal regions of benign tumors, while concentrating primarily on the internal regions of malignant tumors. These findings were consistent with clinical diagnostic procedures and demonstrated potential to guide radiologists. However, we need to be cautious about interpreting the heatmaps due to the distribution

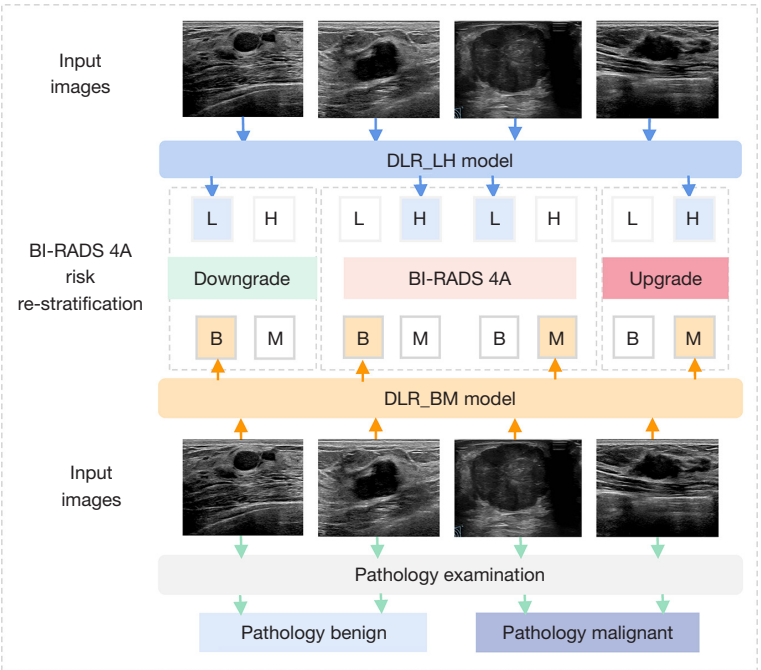

Figure 6 Risk re-stratification procedure of BI-RADS 4A lesions. Input breast ultrasound images with BI-RADS 4A lesions into the DLR_LH and DLR_BM models respectively, when the predicted results were L and B, the downgrade to BI-RADS 3 was recommended, when the predicted results were H and M, the upgrade to a higher grade was recommended. L, BI-RADS 2/3; H, BI-RADS 4/5; B, benign; M, malignant; BI-RADS, Breast Imaging Reporting and Data System.

differences between lesions.

Enhanced diagnosis of radiologists with DLR model assistance

To further confirm whether our models could assist the radiologists' clinical decision-making process, we compared the changes in accuracy, sensitivity, and specificity of the three radiologists before and after the assistance of DLR models, based on the validation samples (n=236), as well as the changes in risk re-stratification of BI-RADS 4A lesions to reduce the FPR (*Table 4*). All readers achieved higher diagnostic accuracy and specificity. Two out of three readers maintained outstanding sensitivity with DLR models assistance, and one junior radiologist (reader 2) achieved diagnostic performance comparable to that of the senior radiologist (reader 3). There was statistical significance in accuracy and specificity before and after implementing the three-round reading and two-stage assisted diagnosis process ($P<0.05$). Namely, the DLR models could significantly enhance radiologists' diagnostic capabilities and narrow the gap between radiologists with different levels of experience. Meanwhile, for the changes in BI-RADS 4A risk re-stratification of lesions diagnosed by each radiologist, we found that 26.8% (15/56), 53.4% (31/58) and 39.2% (20/51) of BI-RADS 4A lesions were downgraded to BI-RADS 3, allowing biopsies to be avoided, with one malignancy missed for reader 1. Additionally, 31, 22, and 20 lesions could be upgraded with malignancy rates of 64.5%, 40.9%, and 45%, respectively. With DLR assistance, the FPR was reduced for all readers, and the malignancy rate of the maintained BI-RADS 4A lesions also decreased, achieving a reduction of about 15% among the junior radiologists. Overall, we observed a positive impact of the DLR models in assisting radiologists to enhance their diagnostic abilities.

Discussion

Although the radiologists considered imaging and non-imaging information comprehensively to find evidence supporting their clinical decisions, the interpretation and risk stratification of breast US lesions by human eyes remained challenging. Due to the considerable variability

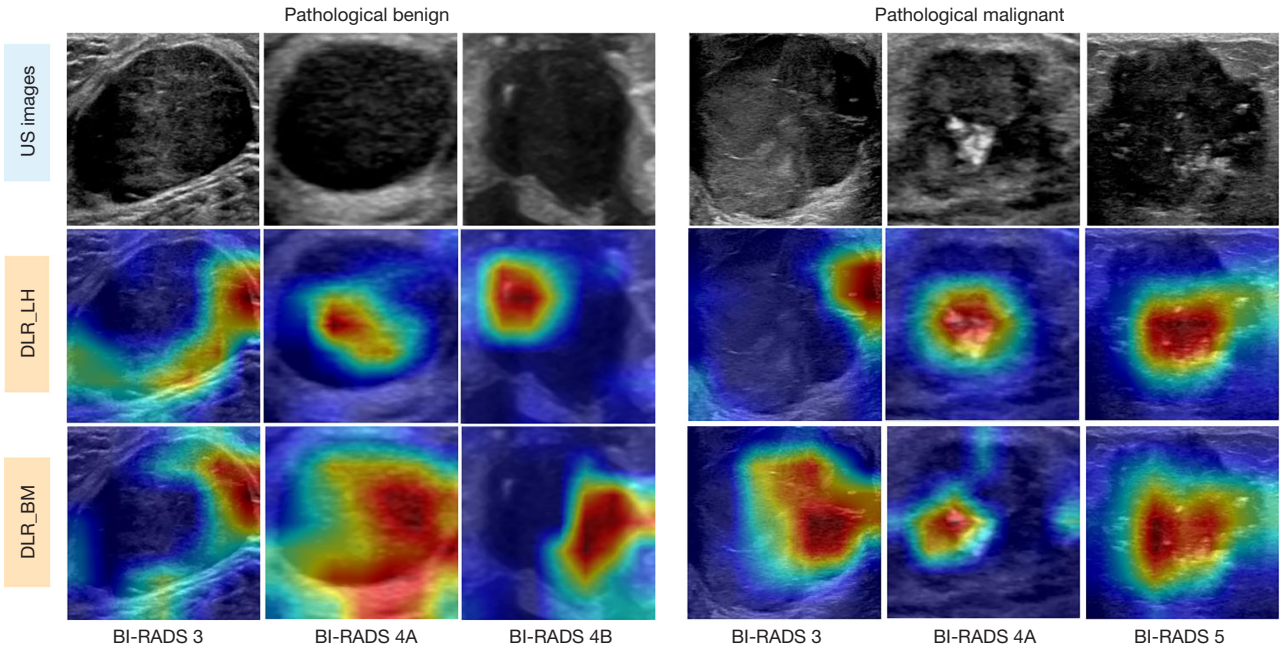

Figure 7 Examples of heatmaps generated by DLR_LH for low-risk level (BI-RADS 2/3) and high-risk level (BI-RADS 4/5) tumors, DLR_BM for benign and malignant tumors, respectively. When US images (first row) are input into DLR_LH, it will give first-stage diagnostic heatmaps to identify the subcategories of BI-RADS 2/3 or BI-RADS 4/5 (second row), and then the DLR_BM gives the second-stage diagnostic heatmaps to distinguish benign from malignant (third row). The DLR models pays more attention to the strong response areas (red areas). US, ultrasound; BI-RADS, Breast Imaging Reporting and Data System; DLR, deep learning radiomics.

Table 4 Summary of the changes in the radiologist decision-making process before and after DLR models assistance

Radiologists	Lesions	True	False	Accuracy (%)	Sensitivity (%)	Specificity (%)	Risk re-stratification of BI-RADS 4A				
							Down grade	No change	Up grade	FP_rate (%)	Mal_rate (%)
Reader 1	B (n=138)	98→110↑	40→28↓	82.2→86.4↑*	97.9→95.9	71.0→79.7↑*	14	7	11	55.4→30.3↓	44.6→30.0↓
	M (n=98)	96→94↓	2→4↑				1	3	20		
Reader 2	B (n=138)	66→98↑	72→40↓	69.1→82.6↑*	98.9→98.9	47.8→71.0↑*	31	5	13	84.5→31.0↓	15.5→0↓
	M (n=98)	97→97	1→1				0	0	9		
Reader 3	B (n=138)	78→89↑	66→49↓	72.0→79.2↑*	1.0→1.0	52.2→64.7↑*	20	10	11	80.4→41.2↓	19.6→9.0↓
	M (n=98)	98→98	0→0				0	1	9		

*, a statistically significant difference between radiologist before and after DLR models assistance. Reader 3 is the senior radiologist (>10 years of experience) and reader 1 and reader 2 are junior radiologists (>7 years of experience). The upward arrow (↑) and the downward arrow (↓) represents indicators of the improvement or decrease because of DLR assistance. DLR, deep learning radiomics; BI-RADS, Breast Imaging Reporting and Data System; FP_rate, false positive rate of BI-RADS 4A; Mal_rate, malignant rate in BI-RADS 4A lesions; B, benign; M, malignant.

among patients, overlap in BI-RADS 4 lesions, subjective imaging interpretation, and unsatisfactory consistency among radiologists, breast US has been criticized for contributing to an increased number of false-positive findings and unnecessary biopsies (35), particularly for BI-RADS 4A lesions. We conducted a retrospective study that established co-decision models using breast US images and clinical information to assess the risk probability of breast

lesions. The DLR_LH model aimed to identify the risk level according to BI-RADS-related features consistent with the clinical diagnosis process, whereas the DLR_BM model further extracted features related to pathological ground truth to identify false positive cases in BI-RADS 4A patients. The model co-decision process for BI-RADS 4A lesions ultimately enhanced diagnostic accuracy and reliability. Both DLR models performed well in distinguishing low-risk probability (BI-RADS 2/3) from high-risk probability (BI-RADS 4/5) and demonstrated the ability to reduce unnecessary biopsies of BI-RADS 4A lesions by 27.7%, indicating that patient age has a strong correlation with tumor malignancy. Moreover, we conducted three rounds of radiologist studies. By incorporating model-predicted scores and heatmaps, the DLR models assisted radiologists in making more effective diagnostic decisions.

During breast US examinations, radiologists primarily focus on identifying the presence of lesions and suspicious lesions with a likelihood of malignancy, guided by the BI-RADS lexicon, and provide relevant diagnostic and treatment recommendations. However, inconsistencies have remained in the application of BI-RADS. In this study, we implemented a three-round, two-stage AI-assisted reading process of breast lesions that followed the clinical diagnostic process. The radiologists made an initial decision on each lesion and, if uncertain, refined their decisions using heatmaps and AI scores, which provided clear diagnostic signals. A detailed comparison of the specific changes before and after assistance from DLR models revealed that radiologists downgraded most of the lesions to true benign lesions and upgraded most of the lesions to true malignant lesions. Here, we aimed to explore the practical benefit of model assistance for radiologists, as the ability to assist radiologists is more important than the standalone diagnostic performance of the DLR model. At present, AI systems can be an auxiliary tool to aid efficient diagnosis but cannot fully replace human experts. The supporting information can serve as additional evidence among the multiple sources needed to help radiologists to make a final diagnostic decision in the future. Overall, we found a positive effect and significant benefits of the DLR models in assisting radiologists to enhance their diagnostic capabilities in interpreting breast US examinations, re-stratifying BI-RADS risk, and reducing unnecessary biopsies of BI-RADS 4A lesions. Particularly, with DLR assistance, the gap between radiologists with different levels of experience is narrowed, leading to a significant reduction in the malignancy rate of BI-RADS 4A lesions diagnosed by junior radiologists. Therefore, our model may be a complementary tool in departments lacking senior radiologists.

Previous studies have concentrated on differentiating benign from malignant breast lesions in US images (12,36,37). Shen *et al.* examined the potential of an AI system in US exams using DL models (12) and demonstrated that collaboration between AI and radiologists could reduce requested biopsies by 27.8%. A multi-task learning framework for 3D automated breast US images was proposed to jointly improve the performance of both segmentation and classification tasks (37). These approaches can effectively differentiate benign from malignant tumors; however, the most critical issue is determining appropriate treatment recommendations for patients with abnormal lesions. Consistent with radiologists' diagnoses, some researchers have utilized AI technology to stratify risks within BI-RADS categories. DL models have been developed for lesion detection and risk stratification according to the BI-RADS atlas in automated breast US (23). Additionally, an AI system was constructed to distinguish BI-RADS subcategories to facilitate improve clinical actions (22,38). Moreover, several studies have attempted to investigate the correlation between BI-RADS risk stratification and pathology results to further enhance performance. An integrated method incorporated BI-RADS stratification as auxiliary information within DL models for classifying breast masses based on US images (34). Similarly, another study established a breast lesion risk stratification system to predict breast malignancy and BI-RADS categories simultaneously to reduce unnecessary biopsies of BI-RADS 4A lesions (21). However, BI-RADS risk stratification results are often subject to inter-observer variability in clinical practice, and it is difficult to fully establish the internal correlations between BI-RADS-related features and pathological ground truth features. Moreover, these studies did not take clinical variables into account. In this study, we learned specific feature representations that incorporated clinical variables and achieved superior discriminative ability. Furthermore, through the collaboration of radiologists and DLR models in imaging interpretation within clinical diagnostic scenarios, the diagnostic results can be made more trustworthy and reliable. This collaborative approach is essential for making DL models more acceptable in clinical applications.

Despite the contributions of our study, it has some limitations. First, although our study contained a relatively large dataset acquired from two device suppliers, the patients

were gathered by one single institution from medically underdeveloped regions of China; subsequent research requires a multi-center approach and larger sample size to validate the models. Second, we focused on evaluating BI-RADS risk re-stratification using static US imaging; however, breast US dynamic video can record lesions in real-time and provide more comprehensive information from multiple dimensions. Finally, an increasing number of clinicians have recognized the diagnostic value of multi-modal US and clinical information in identifying breast lesions (39,40). Utilizing multi-modal learning to combine features from various imaging modalities and related risk factors, such as menstrual history and family history, on a larger dataset would enhance diagnostic accuracy.

In summary, we collected data from 746 patients with breast lesions to explore the diagnostic capabilities of AI. To our knowledge, this is the first attempt to incorporate both US imaging and clinical variables in the development of DLR models, aiming to classify breast tumors as BI-RADS 3 or lower and BI-RADS 4A or higher, while simultaneously distinguishing benign from malignant BI-RADS 4A lesions, to reduce the high FPR and provide clinical treatment recommendations. Through a three-round reader study, the DLR models performed well in breast US imaging interpretation and BI-RADS risk re-stratification, demonstrating promising potential to help reduce unnecessary biopsies of BI-RADS 4A lesions. This indicates the potential applicability of the DLR models in clinical diagnostic practice.

Acknowledgments

We would like to acknowledge all participating investigators who contributed to this study.

Footnote

Reporting Checklist: The authors have completed the TRIPOD+AI reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-24-580/rc

Funding: This study was jointly supported by the National Natural Science Foundation of China (Nos. 62061023 and 61961037) and the Gansu Provincial Science and Technology Plan Project (No. 23JRRA1799).

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-580/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Ethics Committee of the Gansu Provincial Cancer Hospital (No. A202303090008) and the requirement for informed consent for this retrospective analysis was waived.

References

1. Giaquinto AN, Sung H, Miller KD, Kramer JL, Newman LA, Minihan A, Jemal A, Siegel RL. Breast Cancer Statistics, 2022. CA Cancer J Clin 2022;72:524-41.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021;71:209-49.
3. Smolarz B, Nowak AZ, Romanowicz H. Breast Cancer-Epidemiology, Classification, Pathogenesis and Treatment (Review of Literature). Cancers (Basel) 2022;14:2569.
4. Rebolj M, Assi V, Brentnall A, Parmar D, Duffy SW. Addition of ultrasound to mammography in the case of dense breast tissue: systematic review and meta-analysis. Br J Cancer 2018;118:1559-70.
5. Brem RF, Lenihan MJ, Lieberman J, Torrente J. Screening breast ultrasound: past, present, and future. AJR Am J Roentgenol 2015;204:234-40.
6. Feig S. Cost-effectiveness of mammography, MRI, and ultrasonography for breast cancer screening. Radiol Clin North Am 2010;48:879-91.
7. Fallis AG. ACR BI-RADS® Atlas Breast. American College of Radiology 2013.

8. Raza S, Chikarmane SA, Neilsen SS, Zorn LM, Birdwell RL. BI-RADS 3, 4, and 5 lesions: value of US in management--follow-up and outcome. Radiology 2008;248:773-81.
9. Menezes GLG, Pijnappel RM, Meeuwis C, Bisschops R, Veltman J, Lavin PT, van de Vijver MJ, Mann RM. Downgrading of Breast Masses Suspicious for Cancer by Using Optoacoustic Breast Imaging. Radiology 2018;288:355-65.
10. Spak DA, Plaxco JS, Santiago L, Dryden MJ, Dogan BE. BI-RADS® fifth edition: A summary of changes. Diagn Interv Imaging 2017;98:179-90.
11. Lee JM, Arao RF, Sprague BL, Kerlikowske K, Lehman CD, Smith RA, Henderson LM, Rauscher GH, Miglioretti DL. Performance of Screening Ultrasonography as an Adjunct to Screening Mammography in Women Across the Spectrum of Breast Cancer Risk. JAMA Intern Med 2019;179:658-67.
12. Shen Y, Shamout FE, Oliver JR, Witowski J, Kannan K, Park J, et al. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. Nat Commun 2021;12:5645.
13. Kim WH, Moon WK, Kim SM, Yi A, Chang JM, Koo HR, Lee SH, Cho N. Variability of breast density assessment in short-term reimaging with digital mammography. Eur J Radiol 2013;82:1724-30.
14. Bachar N, Benbassat D, Brailovsky D, Eshel Y, Glück D, Levner D, Levy S, Pecker S, Yurkovsky E, Zait A, Sever C, Kratz A, Brugnara C. An artificial intelligence-assisted diagnostic platform for rapid near-patient hematology. Am J Hematol 2021;96:1264-74.
15. Han SS, Kim YJ, Moon IJ, Jung JM, Lee MY, Lee WJ, Won CH, Lee MW, Kim SH, Navarrete-Dechent C, Chang SE. Evaluation of Artificial Intelligence-Assisted Diagnosis of Skin Neoplasms: A Single-Center, Paralleled, Unmasked, Randomized Controlled Trial. J Invest Dermatol 2022;142:2353-2362.e2.
16. Zhang Z, Wang Y, Zhang H, Samusak A, Rao H, Xiao C, Abula M, Cao Q, Dai Q. Artificial intelligence-assisted diagnosis of ocular surface diseases. Front Cell Dev Biol 2023;11:1133680.
17. Jabeen K, Khan MA, Alhaisoni M, Tariq U, Zhang YD, Hamza A, Mickus A, Damaševičius R. Breast Cancer Classification from Ultrasound Images Using Probability-Based Optimal Deep Learning Feature Fusion. Sensors (Basel) 2022.
18. Inan MSK, Alam FI, Hasan R. Deep integrated pipeline of segmentation guided classification of breast cancer from ultrasound images. Biomed Signal Process Control 2022;75:103553.
19. Civilibal S, Cevik KK, Bozkurt A. A deep learning approach for automatic detection, segmentation and classification of breast lesions from thermal images. Expert Syst Appl 2023;212:118774.
20. Ragab M, Albukhari A, Alyami J, Mansour RF. Ensemble Deep-Learning-Enabled Clinical Decision Support System for Breast Cancer Diagnosis and Classification on Ultrasound Images. Biology (Basel) 2022.
21. Gu Y, Xu W, Liu T, An X, Tian J, Ran H, et al. Ultrasound-based deep learning in the establishment of a breast lesion risk stratification system: a multicenter study. Eur Radiol 2023;33:2954-64.
22. Hayashida T, Odani E, Kikuchi M, Nagayama A, Seki T, Takahashi M, et al. Establishment of a deep-learning system to diagnose BI-RADS4a or higher using breast ultrasound for clinical application. Cancer Sci 2022;113:3528-34.
23. Hejduk P, Marcon M, Unkelbach J, Ciritsis A, Rossi C, Borkowski K, Boss A. Fully automatic classification of automated breast ultrasound (ABUS) imaging according to BI-RADS using a deep convolutional neural network. Eur Radiol 2022;32:4868-78.
24. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A, Aerts HJ. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48:441-6.
25. Zheng X, Yao Z, Huang Y, Yu Y, Wang Y, Liu Y, Mao R, Li F, Xiao Y, Wang Y, Hu Y, Yu J, Zhou J. Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. Nat Commun 2020;11:1236.
26. Gu J, Pan J, Hu J, Dai L, Zhang K, Wang B, He M, Zhao Q, Jiang T. Prospective assessment of pancreatic ductal adenocarcinoma diagnosis from endoscopic ultrasonography images with the assistance of deep learning. Cancer 2023;129:2214-23.
27. Pesapane F, De Marco P, Rapino A, Lombardo E, Nicosia L, Tantrige P, Rotili A, Bozzini AC, Penco S, Dominelli V, Trentin C, Ferrari F, Farina M, Meneghetti L, Latronico A, Abbate F, Origgi D, Carrafiello G, Cassano E. How Radiomics Can Improve Breast Cancer Diagnosis and Treatment. J Clin Med 2023;12:1372.
28. Byra M, Galperin M, Ojeda-Fournier H, Olson L, O'Boyle M, Comstock C, Andre M. Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. Med

Phys 2019;46:746-55.
29. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. Data Brief 2020;28:104863.
30. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA; 2016:770-8.
31. Mercado CL. BI-RADS update. Radiol Clin North Am 2014;52:481-7.
32. Elezaby M, Li G, Bhargavan-Chatfield M, Burnside ES, DeMartini WB. ACR BI-RADS Assessment Category 4 Subdivisions in Diagnostic Mammography: Utilization and Outcomes in the National Mammography Database. Radiology 2018;287:416-22.
33. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy; 2017:618-26.
34. Xing J, Chen C, Lu Q, Cai X, Yu A, Xu Y, Xia X, Sun Y, Xiao J, Huang L. Using BI-RADS Stratifications as Auxiliary Information for Breast Masses Classification in Ultrasound Images. IEEE J Biomed Health Inform 2021;25:2058-70.
35. Lazarus E, Mainiero MB, Schepps B, Koelliker SL, Livingston LS. BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value. Radiology 2006;239:385-91.
36. Zhong L, Shi L, Zhou L, Liu X, Gu L, Bai W. Development of a nomogram-based model combining intra- and peritumoral ultrasound radiomics with clinical features for differentiating benign from malignant in Breast Imaging Reporting and Data System category 3-5 nodules. Quant Imaging Med Surg 2023;13:6899-910.
37. Zhou Y, Chen H, Li Y, Liu Q, Xu X, Wang S, Yap PT, Shen D. Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images. Med Image Anal 2021;70:101918.
38. Ji H, Zhu Q, Ma T, Cheng Y, Zhou S, Ren W, et al. Development and validation of a transformer-based CAD model for improving the consistency of BI-RADS category 3-5 nodule classification among radiologists: a multiple center study. Quant Imaging Med Surg 2023;13:3671-87.
39. Huang R, Lin Z, Dou H, Wang J, Miao J, Zhou G, Jia X, Xu W, Mei Z, Dong Y, Yang X, Zhou J, Ni D. AW3M: An auto-weighting and recovery framework for breast cancer diagnosis using multi-modal ultrasound. Med Image Anal 2021;72:102137.
40. Arya N, Saha S. Multi-modal advanced deep learning architectures for breast cancer survival prediction. Knowl Based Syst 2021;221:106965.

Cite this article as: Lu X, Lu Y, Zhao W, Qi Y, Zhang H, Sun W, Zhang H, Ma P, Guan L, Ma Y. Ultrasound-based deep learning radiomics for multi-stage assisted diagnosis in reducing unnecessary biopsies of BI-RADS 4A lesions. Quant Imaging Med Surg 2025;15(3):2512-2528. doi: 10.21037/qims-24-580