European Radiology
https://doi.org/10.1007/s00330-022-09263-8

BREAST

Ultrasound-based deep learning in the establishment of a breast lesion risk stratification system: a multicenter study

Yang Gu[1] · Wen Xu[1] · Ting Liu[2] · Xing An[2] · Jiawei Tian[3] · Haitao Ran[4] · Weidong Ren[5] · Cai Chang[6] · Jianjun Yuan[7] · Chunsong Kang[8] · Youbin Deng[9] · Hui Wang[10] · Baoming Luo[11] · Shenglan Guo[12] · Qi Zhou[13] · Ensheng Xue[14] · Weiwei Zhan[15] · Qing Zhou[16] · Jie Li[17] · Ping Zhou[18] · Man Chen[19] · Ying Gu[20] · Wu Chen[21] · Yuhong Zhang[22] · Jianchu Li[1] · Longfei Cong[2] · Lei Zhu[23] · Hongyan Wang[1] · Yuxin Jiang[1]

Received: 28 March 2022 / Revised: 3 September 2022 / Accepted: 22 October 2022

Abstract

Objectives To establish a breast lesion risk stratification system using ultrasound images to predict breast malignancy and assess Breast Imaging Reporting and Data System (BI-RADS) categories simultaneously.

Methods This multicenter study prospectively collected a dataset of ultrasound images for 5012 patients at thirty-two hospitals from December 2018 to December 2020. A deep learning (DL) model was developed to conduct binary categorization (benign and malignant) and BI-RADS categories (2, 3, 4a, 4b, 4c, and 5) simultaneously. The training set of 4212 patients and the internal test set of 416 patients were from thirty hospitals. The remaining two hospitals with 384 patients were used as an external test set. Three experienced radiologists performed a reader study on 324 patients randomly selected from the test sets. We compared the performance of the DL model with that of three radiologists and the consensus of the three radiologists.

Results In the external test set, the DL model achieved areas under the receiver operating characteristic curve (AUCs) of 0.980 and 0.945 for the binary categorization and six-way categorizations, respectively. In the reader study set, the DL BI-RADS categories achieved a similar AUC (0.901 vs. 0.933, $p = 0.0632$), sensitivity (90.98% vs. 95.90%, $p = 0.1094$), and accuracy (83.33% vs. 79.01%, $p = 0.0541$), but higher specificity (78.71% vs. 68.81%, $p = 0.0012$) than those of the consensus of the three radiologists.

Conclusions The DL model performed well in distinguishing benign from malignant breast lesions and yielded outcomes similar to experienced radiologists. This indicates the potential applicability of the DL model in clinical diagnosis.

Key Points

- *The DL model can achieve binary categorization for benign and malignant breast lesions and six-way BI-RADS categorizations for categories 2, 3, 4a, 4b, 4c, and 5, simultaneously.*
- *The DL model showed acceptable agreement with radiologists for the classification of breast lesions.*
- *The DL model performed well in distinguishing benign from malignant breast lesions and had promise in helping reduce unnecessary biopsies of BI-RADS 4a lesions.*

Keywords Artificial intelligence · Deep learning · Ultrasonography · Breast neoplasms · Diagnosis

Abbreviations

ACR	American College of Radiology
AI	Artificial intelligence
AUC	Area under the receiver operating characteristic curve
BI-RADS	Breast Imaging Reporting and Data System
CAD	Computer-aided diagnosis
CNN	Convolutional neural network
DL	Deep learning
ETC	External test cohort
ITC	Internal test cohort
ML	Machine learning
NPV	Negative predictive value

Hongyan Wang and Yuxin Jiang contributed equally to this work.

✉ Hongyan Wang
whychina@126.com

✉ Yuxin Jiang
jiangyuxinxh@163.com

Extended author information available on the last page of the article

Published online: 23 November 2022

PPV	Positive predictive value
TC	Training cohort
US	Ultrasound

Introduction

Breast cancer is one of the most common malignant tumors in women [1, 2]. A comprehensive and standardized evaluation should be performed by breast imaging examinations to assess the characteristics, location, size, and number of breast lesions and axillary lymph node status [3]. Ultrasound (US) is a widely used technology that plays an essential role in the whole process of the clinical diagnosis and treatment of breast cancer, such as screening, diagnosis, guided biopsy, identification of lymph node metastasis, and postoperative follow-up [4, 5].

In 2013 the American College of Radiology (ACR) published the Breast Imaging Reporting and Data System (BI-RADS) lexicon for US [6]. The lexicon is a helpful guideline used in clinical practice that provides assessment categories and clinical management recommendations [7, 8]. However, the ACR malignancy classification system for breast lesions on US is affected by relatively low specificity and interobserver variability [9–12]. High false-positive findings result in unnecessary biopsies of lesions that are proven to be benign [13, 14]. However, if no close surveillance or biopsy is performed, it may lead to a delayed breast cancer diagnosis, which would have adverse effects on patient prognosis [15]. The BI-RADS lexicon does not specify which US features should be included in the final classification. The US diagnosis of breast lesions mainly depends on the radiologist's training, level of experience, and expertise. Interobserver differences and discrepancies in clinical management recommendations may result in additional or repeated imaging examinations. Moreover, it is challenging for the human eye to categorize some breast lesions even with the help of the standardized lexicon. Although criticisms remain regarding the use of the US lexicon, the risk of malignancy of US findings has been successfully determined, and an appropriate diagnostic strategy has been established for patients.

Recently, many studies have utilized machine learning (ML) or deep learning (DL) technologies to build binary classification prediction models for benign and malignant breast lesions [16–24]; however, only a few studies have focused on BI-RADS classifications or both [25–29], and some issues remain. First, many studies used training and validation datasets from a single institution to build the models, and there were no independent external test sets to verify the robustness of their model [30, 31]. Second, several studies have reported BI-RADS classifications using artificial intelligence (AI) technologies, but the studies did not show the malignancy of each classification [27–29] or did not demonstrate the diagnostic effectiveness of the system in differentiating benign from malignant breast lesions [27]. Third, most current studies did not focus on making the convolutional neural networks (CNN) concentrate on the content of mass and its surroundings rather than the other areas in images. Fourth, several studies have demonstrated that combining BI-RADS classifications and pathology results could improve the classification performance of DL, but how to optimize the use of relevance between them to improve the classification accuracy is still a challenge [25, 32]. In addition, to our knowledge, few studies have focused on improving the prediction performance of BI-RADS categorizations. Therefore, we established a breast lesion risk stratification system based on DL using US images to conduct binary categorization for benign and malignant breast lesions and six-way BI-RADS categorizations for categories 2, 3, 4a, 4b, 4c, and 5, simultaneously. A large multicenter dataset with pathology information and BI-RADS categories interpreted by board-certified radiologists was used to train our DL model. We proposed a combined input based on lesion delineation to make our model concentrate on the lesion and its surrounding area. To improve the classification accuracy of our model, a consistency supervision mechanism was introduced to ensure that the predictions of the two tasks were consistent. Furthermore, a cross-class loss function was employed to improve the accuracy of BI-RADS categories by penalizing different degrees of misclassification items with different weights.

Materials and methods

Study design and patients

We conducted a multicenter prospective study at 32 tertiary hospitals in China between December 2018 and December 2020. The study was approved by the Institutional Ethics Committee of the principal investigator's hospital (Peking Union Medical College Hospital) and was registered on ClinicalTrials.gov with the number ChiCTR1900023916. Informed consent was obtained from each patient before the US examinations. One lesion per patient was included. The inclusion and exclusion criteria can be found in the Supplementary Materials.

The breast US examinations were carried out using Resona7, Resona7s, Resona7T, Resona8, Resona8T, and DC-80 systems (Shenzhen Mindray BioMedical Electronics, Co., Ltd., Shenzhen, China) equipped with high-frequency linear probes (L14-5, L11-3, L12-3, or L9-3) by board-certified radiologists (the original interpreting radiologists) from multiple hospitals with more than 3 years of experience in breast US. Longitudinal and transverse images of the lesion were acquired with and without caliper measurements. Other sections with suspicious US features of malignancy were acquired selectively. Prospective BI-RADS assessment

categories (the original BI-RADS categories) in actual clinical work were recorded by the original interpreting radiologists according to the fifth edition of the BI-RADS lexicon. BI-RADS 2 and 3 category breast lesions were treated at the request of the patients due to concerns about breast cancer or the surgeons due to family history or other risk factors. Static US images, BI-RADS categories, clinical information, and pathological results were uploaded onto a website (www.nuqcc.cn).

Dataset

B-mode US images without caliper measurements were used for the training and testing of the DL model. The images in the training set were reviewed by one experienced radiologist who specializes in breast US imaging to optimize the BI-RADS categories. A total of 5012 patients with 3220 benign and 1792 malignant breast lesions were included for training and testing, including 52 BI-RADS 2 lesions, 2150 BI-RADS 3 lesions, 842 BI-RADS 4a lesions, 540 BI-RADS 4b lesions, 766 BI-RADS 4c lesions, and 662 BI-RADS 5 lesions. The training cohort (TC) of 4212 patients (3832 patients for training and 380 patients for validation) and the internal test cohort (ITC) of 416 patients were from thirty hospitals. The remaining two hospitals (384 patients) were used as an external test cohort (ETC) for assessing the robustness of our model. The average patient age was 44.05 ± 13.10 years for the TC, 43.81 ± 13.39 years for the ITC, and 40.06 ± 12.40 years for the ETC. The statistics of the TC, the ITC, and the ETC are summarized in Table 1. We also adopted the publicly available Breast Ultrasound Images (BUSI) dataset [33] to test our model.

Development and testing of the DL model

RepVGG [34] showed a favorable accuracy-speed trade-off compared to the state-of-the-art methods and outperformed other architectures in the tasks of binary classification and six-way BI-RADS classifications, as shown in Supplementary Tables 4 and 5. Thus, we chose RepVGG as our backbone network. The overall framework is illustrated in Fig. 1.

We performed three innovations to improve the overall classification accuracies. First, we proposed a combined input that consisted of an original US image, an image containing only the lesion region, and an image containing only the surrounding area of a lesion to make our model focus on lesion and its surrounding areas. Second, a consistency supervision mechanism was introduced to restrict that a benign lesion had a small probability of being predicted as BI-RADS 4c or 5 and a malignant lesion had a slight likelihood of being predicted as

Table 1 Clinical and imaging characteristics of the training set, the internal test set, and the external test set

	TC		ITC		ETC	
	B	M	B	M	B	M
Patients (n)	2671	1541	273	143	276	108
Total patients (n)	4212		416		384	
Total images (n)	11717		1094		1232	
Average age (years)	44.05 ± 13.10 (10–90)		43.81 ± 13.39 (17–83)		40.06 ± 12.40 (13–77)	
Average size (cm)	1.85 ± 1.02 (0.30–7.46)		1.74 ± 0.93 (0.34–5.60)		1.97 ± 0.81 (0.45–4.77)	
Age						
< 40 years	1399	210	144	18	163	15
≥ 40 years	1272	1331	129	125	113	93
Lesion size						
≤ 2 cm	2089	691	228	57	170	64
> 2 cm, ≤ 5 cm	559	818	45	85	106	44
> 5 cm	23	32	0	1	0	0
BI-RADS category						
2	35	0	4	0	13	0
3	1789	5	156	1	198	1
4a	611	72	89	7	59	4
4b	183	286	18	23	5	25
4c	47	624	5	58	1	31
5	6	554	1	54	0	47

TC, training cohort; *ITC*, internal test cohort; *ETC*, external test cohort; *B*, benign; *M*, malignant; *BI-RADS*, Breast Imaging Reporting and Data System

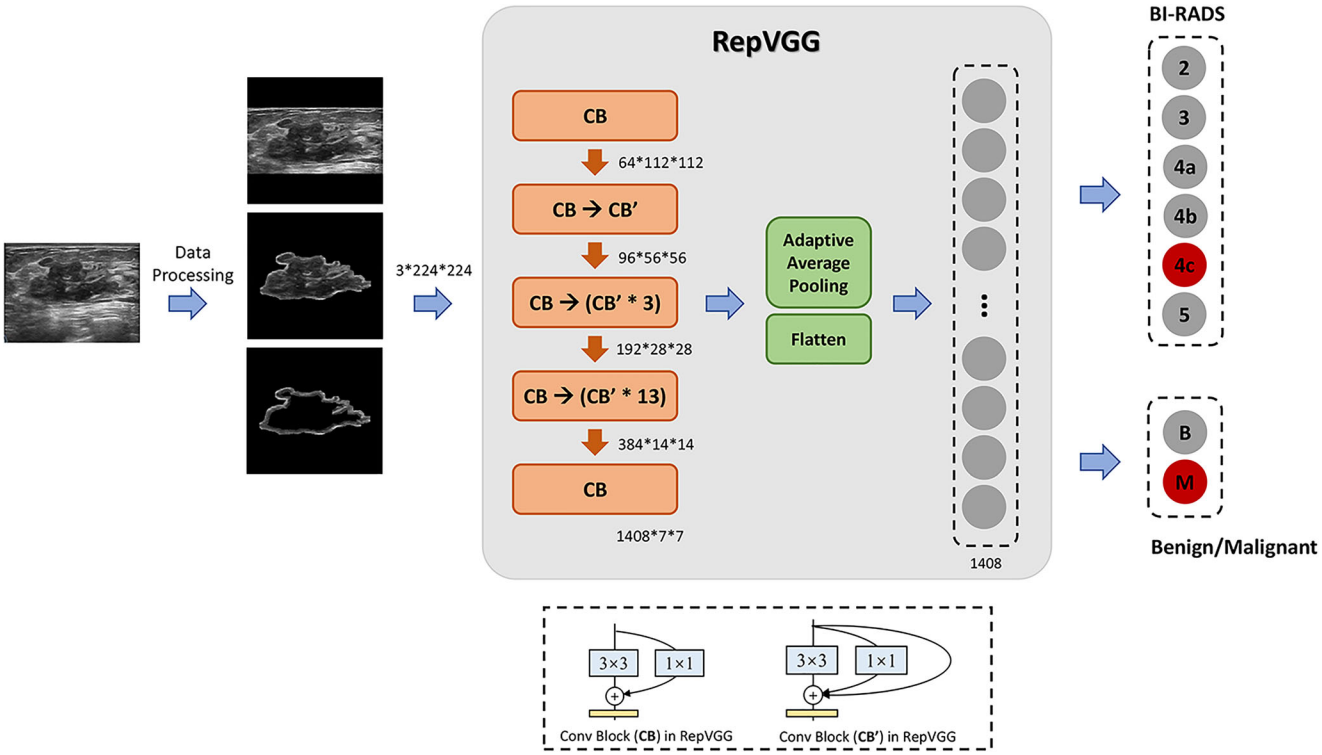

Fig. 1 Overall architecture of the proposed method. The network took the original image, the image containing only the lesion region and the image containing only the margin area of a lesion as input and output the pathologic binary classification and Breast Imaging Reporting and Data System (BI-RADS) categorization probabilities simultaneously

BI-RADS 2 or 3. Third, a cross-class loss function was employed to make the BI-RADS predictions closer to annotations by penalizing different degrees of misclassification items with different weights. Details of the innovations are explained in the Supplementary Materials.

Reader study

We used a subset randomly selected from the test sets for the reader study, including 324 breast lesions (202 benign and 122 malignant lesions), to compare the diagnostic performance between the DL BI-RADS categories and radiologists. All available B-mode US images of the 324 breast lesions were shown to three experienced radiologists (different from the original interpreting radiologists) who specialized in breast imaging. The radiologists were asked to perform a retrospective review of the US images of the breast lesions in the subset and provide a final BI-RADS assessment category (BI-RADS 2, 3, 4a, 4b, 4c, and 5) independently. Radiologists 1, 2, and 3 had more than five years of experience in the US diagnosis of breast lesions. They had no knowledge of the clinical information and pathologic results at the time of review and were blinded to the evaluations made by other radiologists and to the DL model assessment. A consensus of the three radiologists was reached according to the majority results or discussion when involving a discrepancy.

Statistical analysis

Descriptive statistics are presented as counts and frequencies for categorical data and as means (standard deviations) for metric variables. For the DL binary categorization, the outputs of our model were the benign and malignant probabilities. For the DL six-way categorizations, the output of our model was the probability of each BI-RADS category. For each classification task, we chose the highest probability as the predicted result. BI-RADS categories 2 and 3 were considered true negatives, and BI-RADS categories 4 and 5 were considered true positives. The area under the receiver operating characteristic curve (AUC), sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy were used to estimate the classification performance of the DL model. A detailed description of the statistical analyses can be found in the Supplementary Materials.

Results

Testing of the DL model

The distribution of the DL model predictions

We used confusion matrices to describe the distribution of the DL model predictions (Fig. 2). Details are described in the Supplementary Materials.

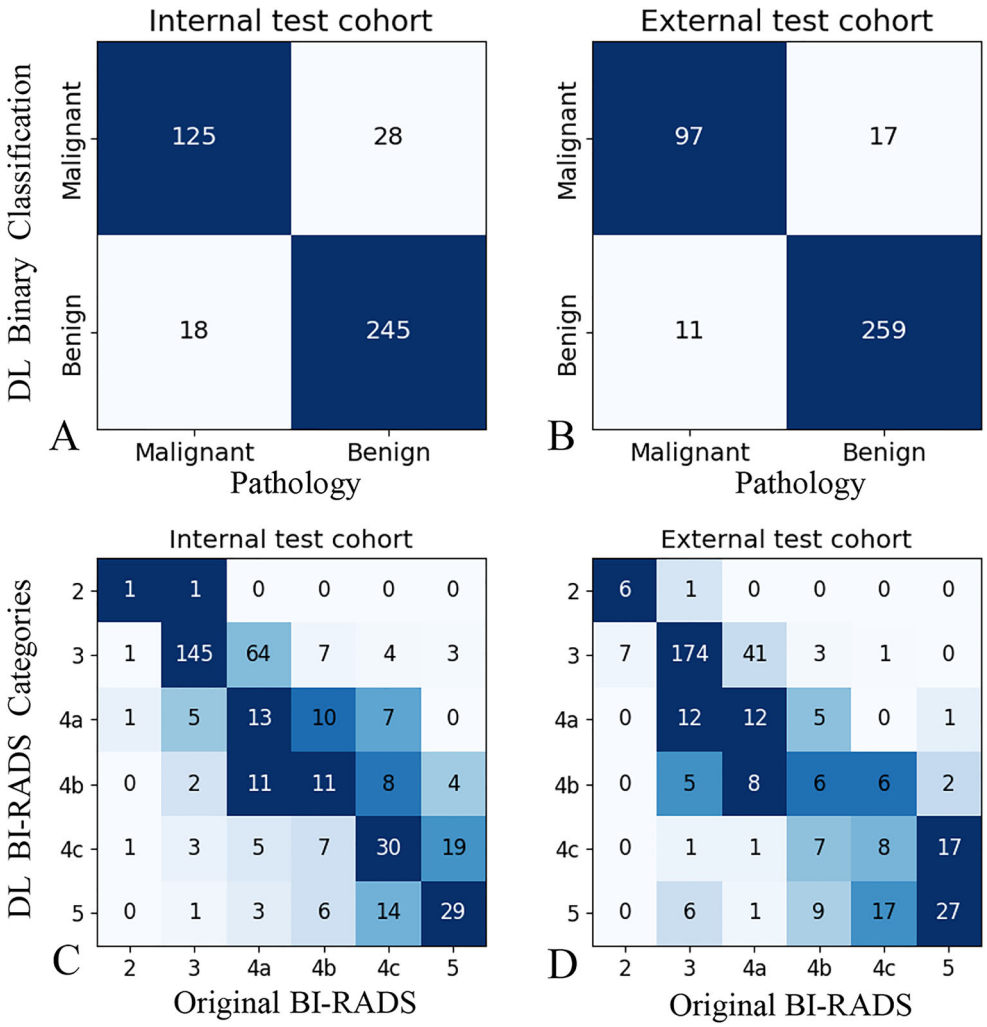

Fig. 2 Distribution of the DL model predictions. Comparison of the DL model binary classification with the pathological findings for the ITC (**A**) and ETC (**B**). Comparison of the DL model BI-RADS assessment with original BI-RADS assessment for the ITC (**C**) and ETC (**D**). DL, deep learning; ITC, internal test cohort; ETC, external test cohort; BI-RADS, Breast Imaging Reporting and Data System

Interobserver variability in the test sets

For DL binary categorization of benign and malignant lesions, the DL model showed substantial agreement ($\kappa = 0.759$, 95% CI: 0.693–0.824) with the pathology results in the ITC and almost perfect agreement ($\kappa = 0.823$, 95% CI: 0.760–0.886) in the ETC. For the DL BI-RADS categories, the DL model showed substantial agreement with the original radiologists in the ITC and ETC (κ = 0.626, 95% CI: 0.575–0.674 and κ = 0.669, 95% CI: 0.619–0.719, respectively).

Performance evaluation

Evaluation and comparison of the diagnostic performance of the DL model and the original radiologists

A comparison of the diagnostic performance of the DL model and the original radiologists in the test sets is displayed in Fig. 3 and Table 2. The results for the ITC are described in the Supplementary Materials. In the ETC, for the DL binary categorization, the DL model achieved an AUC of 0.980 (95% CI: 0.961–0.992), which was comparable to that obtained by the original radiologists (0.986, 95% CI: 0.969–0.995, $p = 0.4203$). The DL model achieved superior specificity (93.84% vs. 76.45%, $p < 0.0001$), PPV (85.09% vs. 62.21%, $p < 0.0001$), and accuracy (92.71% vs. 82.81%, $p < 0.0001$) for diagnosing breast lesions and inferior sensitivity (89.81% vs. 99.07%, $p = 0.0020$) and NPV (95.93% vs. 99.53%, $p = 0.0118$) compared with the original radiologists. For the DL six-way categorizations, the AUC (0.945, 95% CI: 0.917–0.966) was inferior to that of the original radiologists ($p = 0.0009$). The DL model was superior in specificity (82.61% vs. 76.45%, $p = 0.0464$) and was similar in sensitivity (95.37% vs. 99.07%, $p = 0.1250$), PPV (68.21% vs. 62.21%, $p = 0.2598$), NPV (97.85% vs. 99.53%, $p = 0.1266$), and accuracy (86.20% vs. 82.81%, $p = 0.1480$) to that of the original radiologists. The PPVs of the DL BI-RADS assessment categories for the test tests are described in the Supplementary Materials.

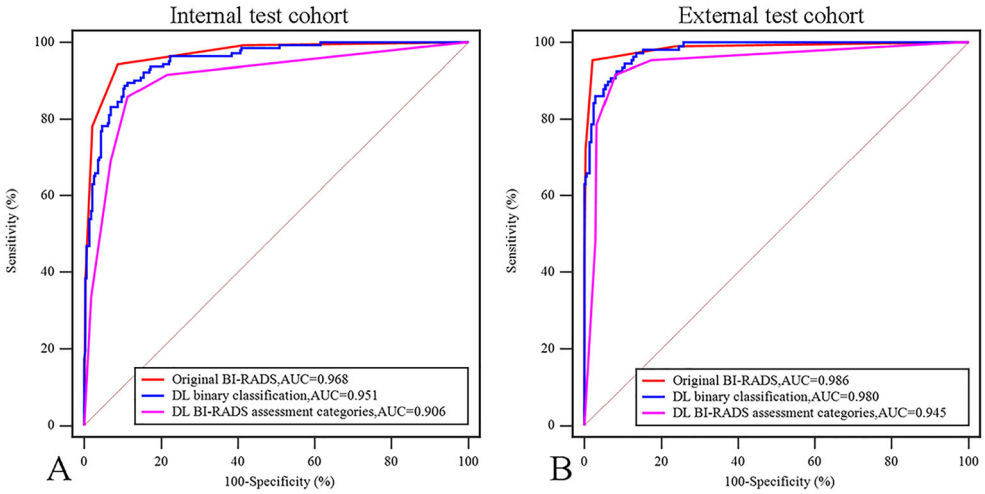

Fig. 3 AUCs for the diagnostic performance of the DL binary classification and BI-RADS assessment categories and the original radiologists in the ITC (**A**) and ETC (**B**). AUCs, areas under the receiver operating characteristic curve; DL, deep learning; BI-RADS, Breast Imaging Reporting and Data System; ITC, internal test cohort; ETC, external test cohort

The utility of DL-assisted diagnosis in reducing unnecessary biopsies

In the ITC, 64.58% (62/96) of BI-RADS 4a breast lesions were downgraded to BI-RADS 3, and 62 biopsies could be avoided with one malignancy missed. In the ETC, 63.49% (40/63) of BI-RADS 4a breast lesions were downgraded to BI-RADS 3, and 40 biopsies could be avoided without any malignancies missed. The biopsy rate for BI-RADS category 4a lesions decreased from 100 to 35.42% in the ITC and from 100 to 36.51% in the ETC compared with the original BI-RADS assessment.

Table 2 Evaluation and comparison of the diagnostic performance of the DL model and the original radiologists

	AUC (95% CI)	SENS (95% CI)	SPEC (95% CI)	PPV (95% CI)	NPV (95% CI)	ACC (95% CI)	F1	MCC
Ro-ITC	0.968 (0.946–0.983)	99.30 (96.17–99.98)	58.61 (52.52–64.51)	55.69 (52.16–59.15)	99.38 (95.77–99.91)	72.60 (68.04–76.83)	0.714	0.565
ITC^{A}	0.951 (0.926–0.970)	87.41 (80.84–92.37)	89.74 (85.52–93.08)	81.70 (75.76–86.44)	93.16 (89.82–95.46)	88.94 (85.53–91.79)	0.845	0.760
p value^{1}	0.0776	< 0.0001*	< 0.0001*	< 0.0001*	0.0027*	< 0.0001*		
ITC^{B}	0.906 (0.874–0.932)	91.61 (85.80–95.59)	78.39 (73.03–83.12)	68.95 (63.79–73.67)	94.69 (91.18–96.85)	82.93 (78.97–86.42)	0.787	0.667
p value^{1}	< 0.0001*	0.0010*	< 0.0001*	0.0045*	0.0117*	< 0.0001*		
p value^{2}	0.0001$^{\#}$	0.2101	< 0.0001$^{\#}$	0.0071$^{\#}$	0.4813	0.0006$^{\#}$		
Ro-ETC	0.986 (0.969–0.995)	99.07 (94.95–99.98)	76.45 (70.99–81.33)	62.21 (57.08–67.08)	99.53 (96.77–99.93)	82.81 (78.66–86.45)	0.764	0.683
ETC^{A}	0.980 (0.961–0.992)	89.81 (82.51–94.81)	93.84 (90.32–96.37)	85.09 (78.19–90.08)	95.93 (93.07–97.63)	92.71 (89.63–95.10)	0.874	0.823
p value^{1}	0.4203	0.0020*	< 0.0001*	< 0.0001*	0.0118*	< 0.0001*		
ETC^{B}	0.945 (0.917–0.966)	95.37 (89.53–98.48)	82.61 (77.61–86.89)	68.21 (62.32–73.58)	97.85 (95.08–99.08)	86.20 (82.34–89.49)	0.795	0.718
p value^{1}	0.0009*	0.1250	0.0464*	0.2598	0.1266	0.1480		
p value^{2}	0.0035$^{\#}$	0.0703	< 0.0001$^{\#}$	0.0016$^{\#}$	0.2196	0.0001$^{\#}$		

DL, deep learning; *Ro*, the original radiologists; *ITC*, internal test cohort; *ETC*, external test cohort; *AUC*, area under the receiver operating characteristic curve; *SENS*, sensitivity; *SPEC*, specificity; *PPV*, positive predictive value; *NPV*, negative predictive value; *ACC*, accuracy; *MCC*, Matthews correlation coefficient; *CI*, confidence interval

A: DL binary classification; B: DL BI-RADS assessment categories

p value^{1} : DL model vs original BI-RADS assessment categories

$^{*}p$ value shows statistical difference

p value^{2} : DL binary classification vs. DL BI-RADS assessment categories

$^{\#}p$ value shows statistical difference

Reader study

Evaluation and comparison of the diagnostic performance of the DL BI-RADS categories and the three radiologists in the reader study set

In the reader study set, the DL BI-RADS categories achieved a similar AUC (0.901 vs. 0.933, $p = 0.0632$), sensitivity (90.98% vs. 95.90%, $p = 0.1094$), PPV (72.08% vs. 65.00%, $p = 0.1666$), NPV (93.53% vs. 96.53%, $p = 0.2294$), and accuracy (83.33% vs. 79.01%, $p = 0.0541$) to those of the radiologists in consensus and yielded a higher specificity (78.71% vs. 68.81%, $p = 0.0012$) than that of the radiologists in consensus. The DL BI-RADS categories achieved radiologist-level diagnostic performance and can be made to simulate human decision-making. A comparison of the diagnostic performance of the DL BI-RADS categories and the three radiologists is displayed in Table 3 and Supplementary Figure 5. Examples of discrepant and concordant interpretations between the DL model and the radiologists are shown in Fig. 4.

The PPVs of each BI-RADS category of the DL model and the three radiologists in the reader study set

The DL BI-RADS and the three radiologists failed to achieve the ACR benchmark PPVs for BI-RADS categories 3, 4a, and 4b in the reader study set, which were higher than the expected rate of malignancy. The PPVs for subcategories 4a, 4b, and 4c were 23.08%, 70.37%, and 81.82% for DL BI-RADS and 28.21%, 88.64%, and 95.45% for the radiologists in consensus. The PPVs of each category of the DL model and the three radiologists are shown in Table 4. Interobserver variabilities in the reader study set are described in the Supplementary Materials.

Discussion

We developed a DL model to predict breast malignancy probability and assess BI-RADS categories simultaneously based on US images. For the binary categorization, the DL model showed substantial or almost perfect agreement with the pathology results. For the BI-RADS categories, the DL model achieved moderate or substantial agreement with the radiologists. The DL model performed well in distinguishing benign from malignant breast lesions and held promise in helping reduce unnecessary biopsies of BI-RADS 4a lesions. The BI-RADS classifications made using the DL model yielded similar outcomes compared with radiologists' diagnoses.

Many studies have used AI technology for the differential diagnosis of benign and malignant breast lesions on US and have achieved excellent performance. Qian et al [23] used 10815 multimodal multiview US images of 721 breast lesions to establish a DL model to realize the automatic prediction of breast cancer risk and conducted prospective tests on 152 breast lesions with an AUC of 0.955. Shen et al [24] proposed an AI system to identify malignant lesions on US images and reduced false positives based on 5,442,907 grayscale and color Doppler images from 288,767 breast exams. The AI system achieved an AUC of 0.976 on a test set consisting of 858,636 images from 44,755 exams. However, such classification (binary categorization) did not completely agree with radiologists' diagnoses. It is difficult for radiologists to completely and correctly interpret whether a breast lesion is benign or malignant.

The BI-RADS risk stratification system demonstrates the malignancy of each BI-RADS assessment category and provides appropriate management strategies for patients. Establishing a DL model at a level similar to that of experienced radiologists can not only imitate radiologists to

Table 3 Evaluation and comparison of the diagnostic performance of the DL BI-RADS categories and the three radiologists in the reader study

	AUC (95% CI)	SENS (95% CI)	SPEC (95% CI)	PPV (95% CI)	NPV (95% CI)	ACC (95% CI)
DL BI-RADS	0.901 (0.863–0.931)	90.98 (84.44–95.41)	78.71 (72.42–84.15)	72.08 (66.31–77.20)	93.53 (89.12–96.23)	83.33 (78.82–87.23)
R1	0.885 (0.845–0.918)	95.90 (90.69–98.66)	56.93 (49.79–63.86)	57.35 (53.33–61.28)	95.83 (90.63–98.21)	71.60 (66.36–76.45)
p value	0.4400	0.1460	< 0.0001*	0.0042*	0.3982	< 0.0001*
R2	0.949 (0.919–0.970)	96.72 (91.82–99.10)	62.87 (55.81–69.55)	61.14 (56.73–65.38)	96.95 (92.33–98.82)	75.62 (70.56–80.19)
p value	0.0147*	0.0654	< 0.0001*	0.0329*	0.1775	0.0019*
R3	0.902 (0.864–0.932)	90.16 (83.45–94.81)	78.22 (71.88–83.71)	71.43 (65.67–76.57)	92.94 (88.45–95.77)	82.72 (78.15–86.67)
p value	0.9454	1.0000	1.0000	0.8994	0.8293	0.8830
Rc	0.933 (0.900–0.958)	95.90 (90.69–98.66)	68.81 (61.93–75.13)	65.00 (60.13–69.58)	96.53 (92.14–98.51)	79.01 (74.17–83.32)
p value	0.0632	0.1094	0.0012*	0.1666	0.2294	0.0541

DL, deep learning; *BI-RADS*, Breast Imaging Reporting and Data System; *R*, radiologist; *Rc*, the consensus of the radiologists; *AUC*, area under the receiver operating characteristic curve; *SENS*, sensitivity; *SPEC*, specificity; *PPV*, positive predictive value; *NPV*, negative predictive value; *ACC*, accuracy; *CI*, confidence interval

* *p* value shows statistical difference

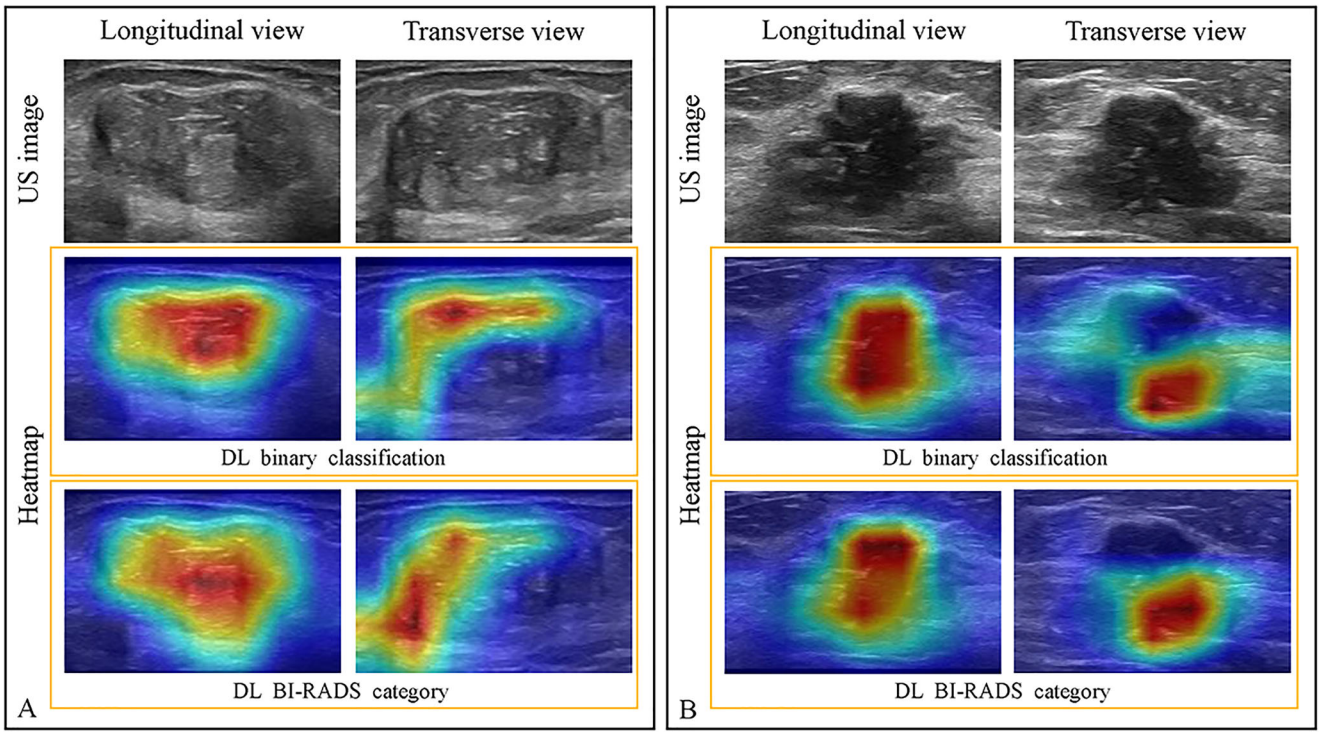

Fig. 4 Examples of discrepant and concordant interpretations between the DL model and the radiologists. The images of the lesion on B-mode US and heatmaps. **A** Example of discrepant interpretation. A 30-year-old woman with palpable breast mass, which was confirmed as fibroadenoma. This lesion was classified as BI-RADS category 4a, 4b, and 3 by the three radiologists, respectively. After discussion, the consensus of the three radiologists was BI-RADS 4a. The predicted results of the DL model were benign for binary classification and 3 for the BI-RADS category. **B** Example of concordant interpretation. A 73-year-old woman with palpable breast mass, which was confirmed as invasive ductal carcinoma. This lesion was classified as BI-RADS category 4b, 4c, and 4c by the three radiologists, respectively. The consensus of the three radiologists was BI-RADS 4c. The predicted results of the DL model were malignant for binary classification and 4c for the BI-RADS category. DL, deep learning; US, ultrasound; BI-RADS, Breast Imaging Reporting and Data System

categorize breast lesions but also has the potential to be applied in actual clinical breast US examinations. Xing et al [25] incorporated the six BI-RADS categories within a DL framework in breast mass US classification. The authors also integrated the BI-RADS estimations and benign and malignant discrimination for classification to improve sensitivity while maintaining a high specificity. Shen et al [26] proposed a weighted computer-aided classification (CAC) system, which was trained using the classification result of radiologists and the pathologic results, to classify the breast masses into BI-RADS categories 3, 4, and 5. Ciritsis et al [28] evaluated a deep CNN with 1019 US images from 582 patients for the classification of US breast lesions into BI-RADS 2, 3, and 4-5 and achieved good classification performance (AUC = 0.838 on ITC, AUC = 0.870 on ETC). Qian et al [29] developed a neural network model based on US B-mode and color Doppler images to classify breast masses into four BI-RADS categories. The authors used radiologists' BI-RADS category assessment as the reference standard to train and validate the model and suggested that a combination of radiologist clinical experience and pathologic results is needed to optimize the model. Our DL model incorporated pathologic findings and radiologist's experience based on BI-RADS. The DL model can not only classify benign and malignant breast tumors but also classify breast lesions into six categories based on US images. The diagnostic accuracy of the DL BI-RADS categories was inferior to the DL binary categorization, but it was higher than that of the prospective BI-RADS assessments. In brief, the DL model showed acceptable diagnostic performance. It is expected that our model could be used as a

Table 4 The PPV of each BI-RADS category of the DL model and the three radiologists in the reader study set

BI-RADS category	PPV (%)				
	DL BI-RADS	R1	R2	R3	Rc
2	0	0	NA	NA	NA
3	6.55	4.20	3.05	7.06	3.47
4a	23.08	30.00	12.00	45.59	28.21
4b	70.37	77.97	84.62	84.62	88.64
4c	81.82	88.89	90.54	97.30	95.45
5	89.13	100	100	100	100

NA, not applicable; *DL*, deep learning; *BI-RADS*, breast imaging reporting and data system; *PPV*, positive predictive value; *R*, radiologist; *Rc*, the consensus of the radiologists

'second opinion' to support the radiologist's decision and mimic the clinical diagnosis of radiologists.

Although several studies have reported BI-RADS classifications using AI technologies, the studies did not show the PPV for each classification [27–29]. The PPVs of categories 3 and 5 ranged from 2.2 to 6.0% [35–37] and from 79.3 to 99.6% [35–38], respectively. Different studies have reported varied PPVs of BI-RADS 4a, 4b, and 4c, ranging from 6.1 to 30.5%, from 25.4 to 88.5%, and from 41.0 to 90.7%, respectively [35–41], which might be due to the different prevalence of breast cancer, the distribution of patient characteristics and the interpretation by radiologists with various experience levels [36, 37, 42]. Shen et al [26] reported that the PPVs in categories 3, 4, and 5 based on the results of a weighted CAC system were 1.63%, 40.23%, and 94.74%, respectively. Xing et al [25] used two publicly available datasets to demonstrate the effectiveness of their proposed model. The PPVs for categories 2, 3, 4a, 4b, 4c, and 5 (calculated according to the tables provided in the article) were 0%, 6.67–16.05%, 23.33–37.04%, 0–40%, 61.11–80.95%, and 93.33–97.10%, respectively. The PPVs of our DL model BI-RADS classification fall within the range of the results of these studies.

Studies have reported that only fair interobserver agreement (κ = 0.14-0.32) was obtained for the BI-RADS categories [11, 36, 43], while other studies achieved moderate (κ = 0.48–0.53) or substantial (κ = 0.67) agreement in assessing the final category between different radiologists [35, 39, 44]. Shen et al [26] reported substantial (κ = 0.644) agreement between the weighted CAC system and the radiologists. Ciritsis et al [28] used small datasets to evaluate the interrater reliability and showed strong (ITC, κ = 0.79) and almost perfect (ETC, κ = 0.90) interrater agreement for three-way classifications (BI-RADS 2, 3, and 4–5) between the deep CNN and the reference standard. In our study, it was acceptable that the agreement was moderate or substantial between DL BI-RADS categories and radiologists.

There are several limitations in our study. First, the labelled BI-RADS classifications of US images used to train the model relied on a subjective evaluation by radiologists based on their experience and differences between observers are inevitable. Second, the distribution of lesions in each classification was uneven. Future studies with additional breast lesions assigned different categories are conducive to enriching the database and developing the DL model. Third, only B-mode US images were used to develop the DL model. Integration with multi-model information could improve the performance of the model.

In conclusion, the DL model performed well in distinguishing benign from malignant breast lesions and held promise in helping reduce unnecessary biopsies of BI-RADS 4a lesions. Our DL BI-RADS model showed radiologist-level performance in diagnosing breast tumors, which indicates the potential applicability of the DL model in clinical diagnosis.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-022-09263-8.

Funding This work is supported by the Beijing Natural Science Foundation (7202156), and the Foundation of International Health Exchange and Cooperation Center NHC PRC (ihecc2018C0032-2).

Declarations

Guarantor The scientific guarantors of this publication are Hongyan Wang and Yuxin Jiang.

Conflict of interest Four of the authors are engineers in Shenzhen Mindray Bio-Medical Electronics Co., Ltd, which provides the ultrasound system and technical support to our research.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was obtained from all patients before they underwent US.

Ethical approval Institutional Review Board approval was obtained.

Methodology
- prospective
- diagnostic study
- multicenter study

References

1. Harbeck N, Gnant M (2017) Breast cancer. Lancet 389:1134–1150
2. Lei S, Zheng R, Zhang S et al (2021) Breast cancer incidence and mortality in women in China: temporal trends and projections to 2030. Cancer Biol Med 18:900–909
3. Zheng C, Yu ZG, Chinese Society of Breast S (2021) Clinical practice guidelines for pre-operative evaluation of breast cancer: Chinese Society of Breast Surgery (CSBrS) practice guidelines 2021. Chin Med J 134:2147–2149
4. Hooley RJ, Scoutt LM, Philpotts LE (2013) Breast ultrasonography: state of the art. Radiology 268:642–659
5. Chang JM, Leung JWT, Moy L, Ha SM, Moon WK (2020) Axillary nodal evaluation in breast cancer: state of the art. Radiology 295:500–515
6. D'Orsi C, Sickles E, Mendelson E, Morris E et al (2013) ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. 5th ed. American College of Radiology, Reston, VA
7. Raza S, Chikarmane SA, Neilsen SS, Zorn LM, Birdwell RL (2008) BI-RADS 3, 4, and 5 lesions: value of US in management–follow-up and outcome. Radiology 248:773–781
8. Raza S, Goldkamp AL, Chikarmane SA, Birdwell RL (2010) US of breast masses categorized as BI-RADS 3, 4, and 5: pictorial review of factors influencing clinical management. Radiographics 30:1199–1213

9. Berg WA, Cosgrove DO, Dore CJ et al (2012) Shear-wave elastography improves the specificity of breast US: the BE1 multinational study of 939 masses. Radiology 262:435–449
10. Lazarus E, Mainiero MB, Schepps B, Koelliker SL, Livingston LS (2006) BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value. Radiology 239:385–391
11. Abdullah N, Mesurolle B, El-Khoury M, Kao E (2009) Breast imaging reporting and data system lexicon for US: interobserver agreement for assessment of breast masses. Radiology 252:665–672
12. Menezes GLG, Pijnappel RM, Meeuwis C et al (2018) Downgrading of breast masses suspicious for cancer by using optoacoustic breast imaging. Radiology 288:355–365
13. Berg WA, Blume JD, Cormack JB et al (2008) Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. JAMA 299:2151–2163
14. Nothacker M, Duda V, Hahn M et al (2009) Early detection of breast cancer: benefits and risks of supplemental breast ultrasound in asymptomatic women with mammographically dense breast tissue. A systematic review. BMC Cancer 9:335
15. Berg WA (2020) Reducing unnecessary biopsy and follow-up of benign cystic breast lesions. Radiology 295:52–53
16. Han S, Kang HK, Jeong JY et al (2017) A deep learning framework for supporting the classification of breast lesions in ultrasound images. Phys Med Biol 62:7714–7728
17. Zhuang Z, Yang Z, Zhuang S, Joseph Raj AN, Yuan Y, Nersisson R (2021) Multi-features-based automated breast tumor diagnosis using ultrasound image and support vector machine. Comput Intell Neurosci 2021:9980326
18. Shia WC, Lin LS, Chen DR (2021) Classification of malignant tumours in breast ultrasound using unsupervised machine learning approaches. Sci Rep 11:1418
19. Shia WC, Chen DR (2021) Classification of malignant tumors in breast ultrasound using a pretrained deep residual network model and support vector machine. Comput Med Imaging Graph 87:101829
20. Romeo V, Cuocolo R, Apolito R et al (2021) Clinical value of radiomics and machine learning in breast ultrasound: a multicenter study for differential diagnosis of benign and malignant lesions. Eur Radiol 31:9511–9519
21. Huo L, Tan Y, Wang S et al (2021) Machine learning models to improve the differentiation between benign and malignant breast lesions on ultrasound: a multicenter external validation study. Cancer Manag Res 13:3367–3379
22. Kalafi EY, Jodeiri A, Setarehdan SK et al (2021) Classification of breast cancer lesions in ultrasound images by using attention layer and loss ensemble in deep convolutional neural networks. Diagn (Basel) 11:1859
23. Qian X, Pei J, Zheng H et al (2021) Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. Nat Biomed Eng 5:522–532
24. Shen Y, Shamout FE, Oliver JR et al (2021) Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. Nat Commun 12:5645
25. Xing J, Chen C, Lu Q et al (2021) Using BI-RADS stratifications as auxiliary information for breast masses classification in ultrasound images. IEEE J Biomed Health Inform 25:2058–2070
26. Shen WC, Chang RF, Moon WK (2007) Computer aided classification system for breast ultrasound based on Breast Imaging Reporting and Data System (BI-RADS). Ultrasound Med Biol 33:1688–1698
27. Huang Y, Han L, Dou H et al (2019) Two-stage CNNs for computerized BI-RADS categorization in breast ultrasound images. Biomed Eng Online 18:8
28. Ciritsis A, Rossi C, Eberhard M, Marcon M, Becker AS, Boss A (2019) Automatic classification of ultrasound breast lesions using a deep convolutional neural network mimicking human decision-making. Eur Radiol 29:5458–5468
29. Qian X, Zhang B, Liu S et al (2020) A combined ultrasonic B-mode and color Doppler system for the classification of breast masses using neural network. Eur Radiol 30:3023–3033
30. Zhang H, Han L, Chen K, Peng Y, Lin J (2020) Diagnostic efficiency of the breast ultrasound computer-aided prediction model based on convolutional neural network in breast cancer. J Digit Imaging 33:1218–1223
31. Qi X, Zhang L, Chen Y et al (2019) Automated diagnosis of breast ultrasonography images using deep neural networks. Med Image Anal 52:185–198
32. Liu J, Li W, Zhao N et al (2018) Integrate domain knowledge in training CNN for ultrasonography breast cancer diagnosis. Springer International Publishing, Cham, pp 868–875
33. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A (2020) Dataset of breast ultrasound images. Data Brief 28:104863
34. Ding X, Zhang X, Ma N, Han J, Ding G, Sun J (2021) RepVGG: Making VGG-style ConvNets Great Again. https://arxiv.org/abs/2101.03697. Accessed 19 Apr 2021
35. Spinelli VMA, Teixeira DCJ, Rauber A, Varella IS, Fleck JF, Moreira LF (2018) Role of BI-RADS ultrasound subcategories 4A to 4C in predicting breast cancer. Clin Breast Cancer 18:e507–e511
36. Stavros AT, Freitas AG, deMello GGN et al (2017) Ultrasound positive predictive values by BI-RADS categories 3-5 for solid masses: an independent reader study. Eur Radiol 27:4307–4315
37. Fu CY, Hsu HH, Yu JC et al (2011) Influence of age on PPV of sonographic BI-RADS categories 3, 4, and 5. Ultraschall Med 32(Suppl 1):S8–S13
38. Yoon JH, Kim MJ, Moon HJ, Kwak JY, Kim EK (2011) Subcategorization of ultrasonographic BI-RADS category 4: positive predictive value and clinical factors affecting it. Ultrasound Med Biol 37:693–699
39. Lee HJ, Kim EK, Kim MJ et al (2008) Observer variability of Breast Imaging Reporting and Data System (BI-RADS) for breast ultrasound. Eur J Radiol 65:293–298
40. Jales RM, Sarian LO, Torresan R, Marussi EF, Alvares BR, Derchain S (2013) Simple rules for ultrasonographic subcategorization of BI-RADS(R)-US 4 breast masses. Eur J Radiol 82:1231–1235
41. He P, Cui LG, Chen W, Yang RL (2019) Subcategorization of ultrasonographic BI-RADS category 4: assessment of diagnostic accuracy in diagnosing breast lesions and influence of clinical factors on positive predictive value. Ultrasound Med Biol 45:1253–1258
42. Hu Y, Yang Y, Gu R et al (2018) Does patient age affect the PPV3 of ACR BI-RADS Ultrasound categories 4 and 5 in the diagnostic setting? Eur Radiol 28:2492–2498
43. Lee YJ, Choi SY, Kim KS, Yang PS (2016) Variability in observer performance between faculty members and residents using Breast Imaging Reporting and Data System (BI-RADS)-Ultrasound, Fifth Edition (2013). Iran J Radiol 13:e28281
44. Park CS, Kim SH, Jung NY, Choi JJ, Kang BJ, Jung HS (2015) Interobserver variability of ultrasound elastography and the ultrasound BI-RADS lexicon of breast lesions. Breast Cancer 22:153–160

Affiliations

Yang Gu[1] · Wen Xu[1] · Ting Liu[2] · Xing An[2] · Jiawei Tian[3] · Haitao Ran[4] · Weidong Ren[5] · Cai Chang[6] · Jianjun Yuan[7] · Chunsong Kang[8] · Youbin Deng[9] · Hui Wang[10] · Baoming Luo[11] · Shenglan Guo[12] · Qi Zhou[13] · Ensheng Xue[14] · Weiwei Zhan[15] · Qing Zhou[16] · Jie Li[17] · Ping Zhou[18] · Man Chen[19] · Ying Gu[20] · Wu Chen[21] · Yuhong Zhang[22] · Jianchu Li[1] · Longfei Cong[2] · Lei Zhu[23] · Hongyan Wang[1] · Yuxin Jiang[1]

1 Department of Ultrasound, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, No.1 Shuai Fu Yuan, Dong Cheng District, Beijing 100730, China

2 Department of Medical Imaging Advanced Research, Beijing Research Institute, Shenzhen Mindray Bio-Medical Electronics Co., Ltd., Beijing, China

3 Department of Ultrasound, The Second Affiliated Hospital of Harbin Medical University, Harbin, China

4 Department of Ultrasound, The Second Affiliated Hospital of Chongqing Medical University & Chongqing Key Laboratory of Ultrasound Molecular Imaging, Chongqing, China

5 Department of Ultrasound, Shengjing Hospital of China Medical University, Shenyang, China

6 Department of Medical Ultrasound, Fudan University Shanghai Cancer Center, Fudan University, Shanghai, China

7 Department of Ultrasonography, Henan Provincial People's Hospital, Zhengzhou, China

8 Department of Ultrasound, Shanxi Bethune Hospital, Shanxi Academy of Medical Sciences, Taiyuan, China

9 Department of Medical Ultrasound, Tongji Hospital, Tongji Medical College of Huazhong University of Science and Technology, Wuhan, China

10 Department of Ultrasound, China-Japan Union Hospital of Jilin University, Changchun, China

11 Department of Ultrasound, The Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China

12 Department of Ultrasonography, First Affiliated Hospital of Guangxi Medical University, Nanning, China

13 Department of Medical Ultrasound, The Second Affiliated Hospital, School of Medicine, Xi'an Jiaotong University, Xi'an, China

14 Department of Ultrasound, Union Hospital of Fujian Medical University, Fujian Institute of Ultrasound Medicine, Fuzhou, China

15 Department of Ultrasound, Ruijin Hospital, Shanghai Jiaotong University, School of Medicine, Shanghai, China

16 Department of Ultrasonography, Renmin Hospital of Wuhan University, Wuhan, China

17 Department of Ultrasound, Qilu Hospital, Shandong University, Jinan, China

18 Department of Ultrasound, The Third Xiangya Hospital of Central South University, Changsha, China

19 Department of Ultrasound Medicine, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

20 Department of Ultrasonography, The Affiliated Hospital of Guizhou Medical University, Guiyang, China

21 Department of Ultrasound, The First Hospital of Shanxi Medical University, Taiyuan, China

22 Department of Ultrasound, The Second Hospital of Dalian Medical University, Dalian, China

23 Department of Medical Imaging Advanced Research, Shenzhen Mindray Bio-Medical Electronics Co., Ltd., Shenzhen, China