BRITISH JOURNAL OF HOSPITAL MEDICINE

ARTICLE

Application Value of Deep Learning-Based AI Model in the Classification of Breast Nodules

Shaogang Zhi[1,*], Xiaoxia Cai[1], Wei Zhou[1], Peipei Qian[1]

[1]Department of Ultrasound, The People's Hospital of Pingyang County, Wenzhou, Zhejiang, China
*Correspondence: zhisg_1983@163.com (Shaogang Zhi)

Abstract

Aims/Background Breast nodules are highly prevalent among women, and ultrasound is a widely used screening tool. However, single ultrasound examinations often result in high false-positive rates, leading to unnecessary biopsies. Artificial intelligence (AI) has demonstrated the potential to improve diagnostic accuracy, reducing misdiagnosis and minimising inter-observer variability. This study developed a deep learning-based AI model to evaluate its clinical utility in assisting sonographers with the Breast Imaging Reporting and Data System (BI-RADS) classification of breast nodules.

Methods A retrospective analysis was conducted on 558 patients with breast nodules classified as BI-RADS categories 3 to 5, confirmed through pathological examination at The People's Hospital of Pingyang County between December 2019 and December 2023. The image dataset was divided into a training set, validation set, and test set, and a convolutional neural network (CNN) was used to construct a deep learning-based AI model. Patients underwent ultrasound examination and AI-assisted diagnosis. The receiver operating characteristic (ROC) curve was used to analyse the performance of the AI model, physician adjudication results, and the diagnostic efficacy of physicians before and after AI model assistance. Cohen's weighted Kappa coefficient was used to assess the consistency of BI-RADS classification among five ultrasound physicians before and after AI model assistance. Additionally, statistical analyses were performed to evaluate changes in BI-RADS classification results before and after AI model assistance for each physician.

Results According to pathological examination, 765 of the 1026 breast nodules were benign, while 261 were malignant. The sensitivity, specificity, and accuracy of routine ultrasonography in diagnosing benign and malignant nodules were 80.85%, 91.59%, and 88.31%, respectively. In comparison, the AI system achieved a sensitivity of 89.36%, specificity of 92.52%, and accuracy of 91.56%. Furthermore, AI model assistance significantly improved the consistency of physicians' BI-RADS classification ($p < 0.001$).

Conclusion A deep learning-based AI model constructed using ultrasound images can enhance the differentiation between benign and malignant breast nodules and improve classification accuracy, thereby reducing the incidence of missed and misdiagnoses.

Key words: breast; nodule; deep learning; artificial intelligence; breast ultrasound; diagnosis

Submitted: 23 January 2025 Revised: 11 March 2025 Accepted: 19 March 2025

How to cite this article:
Zhi S, Cai X, Zhou W, Qian P. Application Value of Deep Learning-Based AI Model in the Classification of Breast Nodules. Br J Hosp Med. 2025. https://doi.org/10.12968/hmed.2025.0078

Introduction

Breast nodules are a common gynaecological condition in women. Their pathogenesis may be associated with endocrine hormone imbalances, heredity, lifestyle factors, and environmental influences. Clinical symptoms include the presence of breast mass, pain or tenderness and nipple discharge. Breast nodules are broadly

classified as benign or malignant depending on their nature. Benign nodules are usually less harmful but can cause symptoms such as pain and discomfort. In contrast, malignant nodules, such as breast cancer, if not treated in time, pose a severe threat to the patient's health and survival (Li et al, 2023). Therefore, early detection, accurate diagnosis, and timely intervention are undoubtedly critical in reducing breast cancer-related morbidity and mortality.

Imaging examinations are the most commonly used non-invasive diagnostic tools for breast nodules (Hong et al, 2022). Given the distinct ultrasonic features of benign and malignant breast nodules, the Breast Imaging Reporting and Data System (BI-RADS) is commonly used to assess the benignity/malignancy risk and classify breast nodules accordingly (Sickles, 2013). However, the positive predictive values for BI-RADS-guided biopsies range from 19.5% to 42.7%, indicating that numerous patients underwent unnecessary invasive biopsies, increasing the risk of complications such as infection and hematoma (Hong et al, 2022). Additionally, in clinical practice, there are no universally accepted classification criteria for BI-RADS categories, especially for breast nodules classified as BI-RADS 3 to 5. This inconsistency contributes to significant variations in diagnostic results among different physicians (Bartolotta et al, 2021; Cheng et al, 2020).

In recent years, numerous studies have shown that artificial intelligence (AI) has significantly improved diagnostic accuracy, reduced missed and misdiagnoses, alleviated physician workload and minimised the differences in diagnostic capabilities among different physicians (Li et al, 2021; Pan et al, 2024; Shen et al, 2021; Sun et al, 2022). Deep learning, currently the most promising machine learning method, can automatically extract image features and perform classification, yielding favourable results in medical image classification (Alzahrani et al, 2024). For example, to distinguish BI-RADS4A lesions, Yi et al (2024) extracted clinical and ultrasound features and integrated them into a predictive model using deep learning algorithms. The performance of the model was evaluated using receiver operator characteristic (ROC) curves, calibration curves, and decision curves. Their findings indicated that with the assistance of deep learning algorithms, radiologists achieved higher area under the curve (AUC) values, improved specificity, and reduced unnecessary biopsy rates by 6.7% and 24% (Yi et al, 2024).

This study aimed to establish a deep learning-based AI model using breast ultrasound images from patients in our hospital to evaluate its clinical application value in assisting ultrasound physicians with BI-RADS breast nodule classification and improving differentiation between benign and malignant breast lesions.

Methods

Clinical Data

A retrospective analysis was conducted on breast ultrasound images from 558 female patients who underwent breast ultrasound examination and subsequent breast surgery at The People's Hospital of Pingyang County between December 2019 and December 2023. All lesions were confirmed through histopathological examination. This study was approved by the Medical Ethics Committee of The People's

Hospital of Pingyang County (Approval No. IRB-2023-04) and complied with the relevant provisions of the Declaration of Helsinki.

Given the retrospective nature of this study and its reliance on available medical records and imaging data, it poses minimal risk to participants. Additionally, due to the large number of cases and the challenges in contacting patients from past records, obtaining informed consent from all participants was impractical. To ensure confidentiality, all patient data were anonymised before analysis. Furthermore, this study aimed to contribute to advancements in clinical practice and patient care and serve the public interest. For the above reasons, informed consent was waived for this retrospective study by The People's Hospital of Pingyang County.

Inclusion Criteria

Patients were included in the study if they met the following criteria:

(1) No biopsy or surgical procedure was performed before the ultrasound examination; (2) Complete clinical and imaging data were available, and the patient demonstrated good compliance; (3) The lesion was clearly visible, and the patient was undergoing breast surgery for the first time; (4) The patient was between 18 and 70 years old.

Exclusion Criteria

Patients were excluded from the study if they met the following conditions:

(1) Pregnant or lactating women; (2) Individuals with mental illness or those unable to communicate normally; (3) Patients with poor-quality or missing imaging data; (4) Patients who had previously undergone anti-cancer treatments such as surgery for breast disease, chemotherapy, or radiation therapy, before the ultrasound examination; (5) Patients with breast nodules exceeding a maximum diameter of 5 cm.

Data Set Source

The imaging data were acquired using a colour Doppler ultrasound diagnostic system (EPIQ 7 Ultrasound System, Philips Ultrasound, Inc., Bothell, WA, USA), with an L12-5 linear array probe (50 mm, Philips Ultrasound, Inc., Bothell, WA, USA), operating at a frequency range of 5–12 MHz. During examination, patients were positioned in the supine position, with both breasts fully exposed within the scanning field of view. Continuous multi-section scans were performed with the nipple as the centre, with each scan overlapping the previous one by 1/3. In two-dimensional mode, the location, size, and structural characteristics of the breast nodule were assessed. The imaging was then switched to colour Doppler mode to explore the internal structures of the nodule and blood flow conditions. The obtained longitudinal and transverse images were stored for further analysis. In total, 1026 breast nodules and 4123 breast ultrasound images were collected.

AI Model

Construction of AI Model Based on Deep Learning

A total of 1026 breast nodule ultrasound images from 558 female patients diagnosed with breast diseases at The People's Hospital of Pingyang County between

December 2019 and December 2023 were collected. The image dataset was randomly divided into a training set, a validation set, and a test set in a 7:1.5:1.5 ratio for AI model development.

In this study, a convolutional neural network (CNN) was used to construct a deep learning-based AI model. The AI model comprised two parts: a lesion identification module and a lesion classification module. The AI-assisted diagnostic software used in this study was a portable, intelligent ultrasound diagnostic system jointly developed by Fudan University Cancer Hospital, the School of Information Science and Technology at Fudan University, Shanghai University, and Vision Hawk Intelligence Technology (Shanghai) Co., Ltd. Firstly, ultrasound images from the training set were input into the AI diagnostic system to generate lesion benignity and malignancy parameters. Subsequently, the validation set was analysed using the AI-assisted diagnostic system, and the optimal malignancy parameters were adjusted based on the pathological examination results.

The lesion recognition module processes the entire input ultrasound image to determine the location of the lesion. In this study, a model adapted from the RetinaNet network was primarily used (Lin et al, 2020). To accommodate variations in ultrasound diagnostic devices and image resolutions, input ultrasound images were rescaled to three sizes with different resolutions of 224 × 224, 112 × 112, and 56 × 56 to extract multi-scale features. The extracted features were then combined and subjected to multi-scale feature extraction. This extraction process uses modules composed of 3 × 3 convolutional layers, max-pooling, and rectified linear unit (ReLU) activation functions. Finally, rectangular regions were applied to the 7 × 7 extracted feature map, with scaling ratios of 1/2, 1, and 2, aspect ratios of $2^{1/2}$, 1, and 2^2 at each pixel to generate rectangular boxes of varying sizes and aspect ratios. These boxes were then analysed through region regression analysis to align with the true lesion area.

The lesion recognition module, which adopts a residual network structure for feature extraction and classification, processed the entire input ultrasound image to identify the lesion location. Images were input at 224 × 224 resolution and processed through modules comprising 3 × 3 convolutional layers, max pooling layers, and ReLU activation functions for feature extraction. Residual modules were incorporated within each processing unit to enhance the expressive capability of the model. Finally, a fully connected layer was used, followed by a softmax logistic regression model for benign and malignant classification.

During the training of the above network model, to mitigate overfitting due to insufficient data, "Dropout" and "Weight decay" regularisation techniques were incorporated to optimise model performance. The "Dropout" module assigns a certain probability for neuron outputs to be zero, while the "Weight decay" module primarily prevents excessively large weight values by introducing a regularisation term.

Training and Validation

The dataset was randomly split into a training set (70%, 718 cases), a validation set (15%, 154 cases), and a test set (15%, 154 cases) following a 7:1.5:1.5 ratio. The

training set was used to optimise the learnable parameters of the model, while the validation set was used to fine-tune the malignancy parameters after selecting parameters from the training set, with the objective of constructing an optimal model based on final evaluation results. To prevent overfitting and enhance the generalisation of the model, data augmentation techniques were applied to the ultrasound images, including elastic transformations, scaling, translation, and horizontal flipping.

AI Model-Assisted Physician Classification

A random subset of breast nodule ultrasound images from the test set was selected. Five independent physicians interpreted and classified the images according to the BI-RADS system in a blinded manner (Spak et al, 2017). Two weeks later, the same five physicians re-evaluated the same images, this time incorporating the AI model's diagnostic results, and reclassified them using the BI-RADS system. Notably, the five physicians in this classification task are the same individuals referenced in the physician adjudication section described later in the study.

Physician Adjudication

Five breast ultrasound diagnostic physicians with 8, 11, 12, 15, and 19 years of experience in breast ultrasound diagnosis independently performed blinded image interpretation with reference to the BI-RADS system, classifying and labelling the breast ultrasound images. This study included BI-RADS categories 3 to 5, including the subcategories of category 4 (4a, 4b, and 4c). The classification results from the five physicians were collected, and a senior physician with 21 years of experience in breast ultrasound diagnosis served as the arbitrator. The adjudication principles were as follows: When three or more of the five physicians reached a consensus on classification, the majority result was considered final. When more than three of the five physicians disagreed on the BI-RADS classification of the lesion, the arbitrating physician made the final decision.

Observation Indices

Clinical data were collected, including age, number of nodules, lesion location, marital and childbearing status, education level, menopausal history, family history, and imaging features related to nodules. The imaging parameters included maximum lesion diameter, BI-RADS classification, form, margins, echogenicity, posterior acoustic pattern, blood flow characteristics, structural distortion, and microcalcification.

Using ultrasound images from the training set (70%, 718 cases) as a reference, an AI model was established following the outlined workflow (Fig. 1). The BI-RADS classifications assigned by the AI model and physician adjudication (ultrasound) were recorded, and the diagnostic performance of the AI model and physician adjudication (ultrasound) for benign and malignant breast nodules in the training set ultrasound images was analysed. The validation set (15%, 154 cases) was subsequently evaluated using the AI model, and the BI-RADS classifications from

the AI model and physician adjudication (ultrasound) were recorded to further assess their diagnostic accuracy for benign and malignant breast nodules.

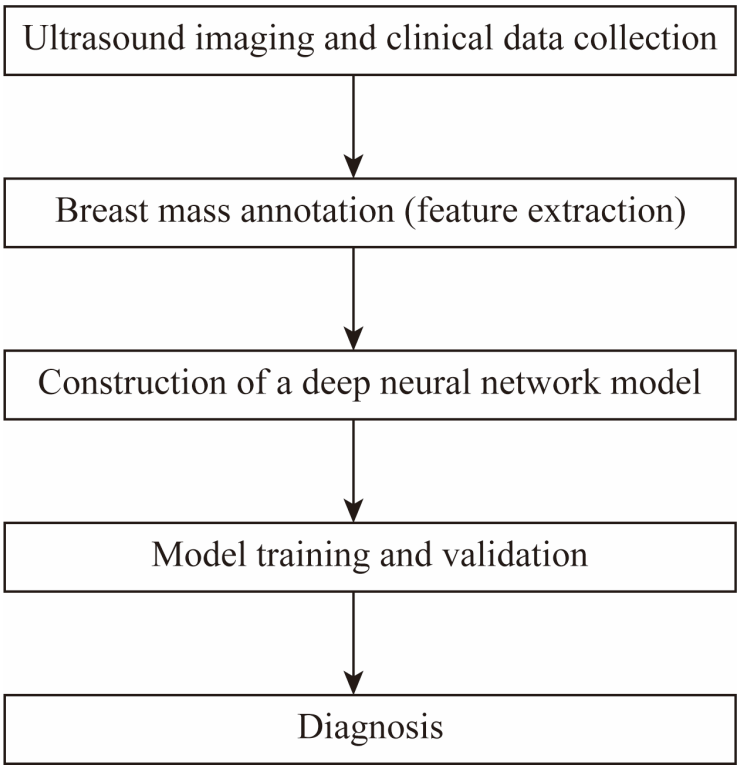

Fig. 1. Flowchart of the study.

To further evaluate the diagnostic value of the AI model for breast cancer, pathological results were used as the gold standard. Physicians performed BI-RADS classification on the 154 test set images before and after incorporating AI model assistance. ROC curve analysis was conducted to compare the sensitivity, specificity, accuracy, and AUC of the diagnostic methods.

Statistical Analysis

Data were processed using SPSS 22.0 software (IBM, Armonk, NY, USA). Categorical variables were expressed as frequency (%), and group differences were evaluated using the chi-square (χ^2) test or Fisher's exact test. Continuous variables that conform to a normal distribution were expressed as mean $\pm$ standard deviation ($\bar{x} \pm s$). The Kappa consistency test was used to assess the consistency between physician adjudication (ultrasound) and the AI model. To determine the diagnostic efficacy of the detection methods, ROC curves were constructed, and the AUC was compared. A p-value < 0.05 was considered statistically significant.

Results

Basic Characteristics

The patients' ages ranged from 21 to 69 years, with a mean age of 45.36 ± 7.85 years. The disease duration ranged from 6 to 14 months, with an average duration

of 10.20 ± 1.39 months. Table 1 shows the baseline characteristics of the study population. No significant differences were detected between the benign and malignant groups regarding patient age ($p = 0.128$), lesion site ($p = 0.055$), marital and childbearing status ($p = 0.085$), and educational level ($p = 0.153$). However, significant differences were observed in the number of nodules ($p < 0.001$), menopausal history ($p = 0.019$), and family history of breast disease ($p = 0.006$). These results suggest that the occurrence of breast nodules is significantly correlated with menstrual history, marital and childbearing status, and family history of breast disease.

Based on postoperative pathological results, among the 1026 breast nodules from 558 patients, 765 were benign, including 273 fibroadenomas, 226 fibroadenoses, 115 fibrocystic breast diseases, 62 intraductal papillomas, 34 stromal collagenizations, 29 inflammatory lesions, and 26 other benign lesions. The remaining 261 nodules were malignant, including 116 non-specific invasive breast cancers, 73 ductal carcinomas in situ, 10 intraductal papillary carcinomas, 9 medullary carcinomas, 31 malignant phyllodes tumours, and 22 mucinous carcinomas. In the comparison of clinicopathological characteristics, significant differences were observed between the malignant and benign groups for maximum tumour diameter ($p < 0.001$), BI-RADS classification ($p < 0.001$), forms ($p < 0.001$), nodule margins ($p < 0.001$), echogenicity ($p < 0.001$), posterior acoustic pattern ($p = 0.022$), blood flow characteristics ($p < 0.001$), structural distortion ($p < 0.001$), and microcalcifications ($p = 0.038$) ($p < 0.01$, Table 2).

Diagnostic Efficacy of the Model

An AI model based on deep learning was developed using a CNN. Based on the physician adjudication classification results, the coincidence rates between the AI model's classifications and the physician assessments in the test set were 94.07% for category 3 (222/236), 85.02% for category 4a (210/247), 75.36% for category 4b (52/69), 90.22% for category 4c (83/92), and 91.89% for category 5 (68/74) (Table 3).

The lower accuracy for category 4b nodules may be due to their moderate probability of malignancy (10%–50%) and their morphological similarity to benign or malignant nodules (such as category 4c), which makes the model prone to misjudgment when distinguishing category 4b nodules. Additionally, the display of category 4b nodules by ultrasound images may have limitations, such as blurred boundaries and uneven echoes, making it difficult for the model to extract enough effective information from the images, ultimately contributing to a lower coincidence rate for category 4b nodules.

Using pathological results as the gold standard (171 malignant and 547 benign cases), the AI model and physician adjudication (ultrasound) showed high consistency with pathological findings for the training set of breast ultrasound images ($p < 0.001$). Moreover, the diagnostic consistency between the AI model's results and the pathological results was slightly higher than that of physician adjudication (ultrasound) (Table 4). The AUC for the model's diagnosis of malignant breast nodules was 0.899 (95% CI: 0.867–0.932, $p < 0.001$), with a sensitivity of 85.38%, specificity of 95.06%, and an overall accuracy of 92.76%, indicating that the model

Table 1. Baseline characteristics of patients with breast nodules [n (%)].

Characteristic	Benign group (n = 400)	Malignant group (n = 158)	χ^2	*p*-value
Age (years)			7.159	0.128
21–30	29 (7.25)	4 (2.53)		
31–40	106 (26.50)	34 (21.52)		
41–50	174 (43.50)	76 (48.10)		
51–60	82 (20.50)	40 (25.32)		
61–69	9 (2.25)	4 (2.53)		
Number of nodules			58.861	<0.001
1	35 (8.75)	55 (34.81)		
≥2	365 (91.25)	103 (65.19)		
Lesion site			5.800	0.055
Left	112 (28.00)	33 (20.89)		
Right	124 (31.00)	43 (27.22)		
Double	164 (41.00)	82 (51.90)		
Marital and childbearing status			4.936	0.085
Unmarried, no children	89 (22.25)	22 (13.92)		
Married, no children	81 (20.25)	36 (22.78)		
Married with children	230 (57.50)	100 (63.29)		
Education level			2.039	0.153
Undergraduate and above	129 (32.25)	61 (38.61)		
Specialist and below	271 (67.75)	97 (61.39)		
Menopause status			5.494	0.019
Postmenopausal	130 (32.50)	68 (43.04)		
Premenopausal	270 (67.50)	90 (56.96)		
Family history of breast disease			7.417	0.006
Present	68 (17.00)	43 (27.22)		
Absent	332 (83.00)	115 (72.78)		

has good diagnostic efficacy (Table 5 and Fig. 2). No significant difference was observed between the AI model and physician adjudication classification results (p = 0.396).

Validation of the Diagnostic Model

To validate the ability of the AI model to classify the trained BI-RADS categories, 154 breast ultrasound images from the validation set were analysed using the model. Based on physician adjudication classification results, the compliance rates of AI model classification for various types of nodules in the verification set were 94.23% (49/52) for category 3, 97.06% (33/34) for category 4a, 87.50% (21/24) for category 4b, 95.45% (21/22) for category 4c, and 95.45% (21/22) for category 5 (Table 6). These findings indicate that the diagnostic accuracy of the trained AI model to classify breast nodules is highly consistent with physician adjudication classification. Using pathological results as the gold standard (43 malignant and 111 benign cases), the AI model and physician adjudication (ultrasound) demonstrated high consistency with pathological results for the test set of breast ultrasound im-

Table 2. Sonographic characteristics of benign and malignant breast nodules [n (%)].

Characteristic	Benign group (n = 765)	Malignant group (n = 261)	χ^2	*p*-value
Maximum tumour diameter (cm)			50.447	<0.001
≥3	327 (42.75)	178 (68.20)		
<3	438 (57.25)	83 (31.80)		
BI-RADS classification				<0.001
3	408 (53.33)	0 (0.00)		
4a	224 (29.28)	45 (17.24)		
4b	133 (17.39)	37 (14.18)		
4c	0 (0.00)	97 (37.16)		
5	0 (0.00)	82 (31.42)		
Form			73.876	<0.001
Regular	505 (66.01)	93 (35.63)		
Irregular	260 (33.99)	168 (64.37)		
Nodule margin			119.388	<0.001
Smooth	306 (40.00)	36 (13.79)		
Blurred	197 (25.75)	38 (14.56)		
Lobulated	74 (9.67)	46 (17.62)		
Angular	100 (13.07)	91 (34.87)		
Burr	88 (11.50)	50 (19.16)		
Echogenicity				<0.001
Hyperechoic	0 (0.00)	26 (9.96)		
Cystic-solid	188 (24.58)	35 (13.41)		
Hypoechoic	382 (49.93)	69 (26.44)		
Isoechoic	77 (10.07)	20 (7.66)		
Heterogeneous	118 (15.42)	111 (42.53)		
Posterior acoustic features			5.224	0.022
Attenuation	296 (38.69)	122 (46.74)		
No change	469 (61.31)	139 (53.26)		
Doppler blood flow pattern			118.036	<0.001
Absent	376 (49.15)	89 (34.10)		
Internal vascularity	120 (15.69)	127 (48.66)		
Peripheral vascularity	269 (35.16)	45 (17.24)		
Structural distortion			54.976	<0.001
Present	357 (46.67)	191 (73.18)		
Absent	408 (53.33)	70 (26.82)		
Microcalcifications			4.311	0.038
Present	371 (48.50)	146 (55.94)		
Absent	394 (51.50)	115 (44.06)		

ages ($p < 0.001$). Moreover, the consistency between the AI model's results was significantly higher than that of physician adjudication (ultrasound) (Table 7).

Taking pathological results as the gold standard, the ROC curves for BI-RADS classification by the AI model and physician adjudication in the validation set images are shown in Fig. 3. The sensitivity, specificity, and accuracy of the AI model were 88.37%, 93.69%, and 92.21%, respectively. The AUC area for the AI model

Table 3. Analysis of BI-RADS classification by the AI model on training set images (cases).

Physician adjudication classification (Ultrasound)	AI model classification					Total
	3	4a	4b	4c	5	
3	222	7	4	1	2	236
4a	18	210	13	3	3	247
4b	7	7	52	2	1	69
4c	0	0	2	83	7	92
5	0	0	3	3	68	74
Total	247	224	74	92	81	718

Note: Both conventional ultrasound and AI-assisted diagnosis consider BI-RADS categories 4c to 5 as malignant, and the same as below. BI-RADS, Breast Imaging Reporting and Data System; AI, artificial intelligence.

Table 4. Comparison of AI model and physician adjudication (ultrasound) in diagnosing breast nodule benignity and malignancy based on training set ultrasound images.

Pathological examination	Physician adjudication (Ultrasound)		AI model	
	Benignity	Malignancy	Benignity	Malignancy
Benignity (n = 547)	515	32	520	27
Malignancy (n = 171)	37	134	25	146
Total	552	166	545	173
Kappa	0.732		0.801	
p-value	<0.001		<0.001	

was 0.914 (95% CI: 0.856–0.972, $p < 0.001$), with no statistically significant difference compared to physician adjudication classification results ($p = 0.06$, Table 8 and Fig. 3).

The Auxiliary Role of the AI Model in Physician Diagnosis

Comparison of Physicians' Diagnostic Efficacy Before and After AI Model Assistance

Using pathological results as the gold standard, the ROC curves for BI-RADS classification performed by five physicians before and after the AI model assistance in analysing 154 randomly selected test set images are shown in Table 9 and Fig. 4. The diagnostic accuracy of doctors improved slightly after AI model assistance, but the differences were not statistically significant ($p = 0.256$). This may be attributed to the extensive experience of the five doctors in breast ultrasound diagnosis and prior training in BI-RADS classification standards before image interpretation.

Consistency of BI-RADS Classification Among Five Physicians Before and After AI Model Assistance

A weighted Kappa coefficient analysis was performed to assess the consistency of BI-RADS classification among the five physicians before and after AI model assistance (Table 10). When physicians independently classified nodules using

Table 5. Diagnostic value of the AI model and physician adjudication (ultrasound) for assessing breast nodule benignity and malignancy based on training set ultrasound images.

Detection Method	AUC	SE	95% CI	Sensitivity (%)	Specificity (%)	Accuracy (%)
Physician adjudication (Ultrasound)	0.889	0.016	0.857–0.920	78.36	94.15	90.39
AI model	0.899	0.017	0.867–0.932	85.38	95.06	92.76

Comparison: No statistically significant difference between AI and physician adjudication ($p = 0.396$). AUC, area under the curve.

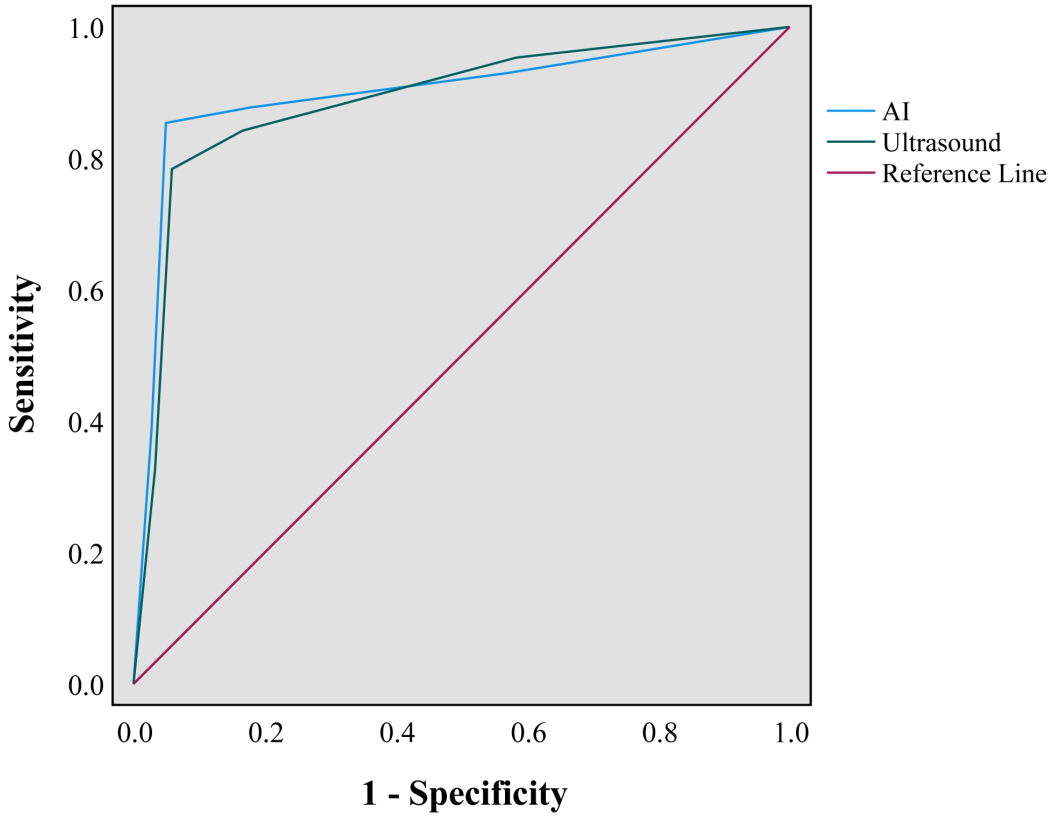

Fig. 2. ROC curve for differentiating benign and malignant lesions in the training set images using the AI model and physician adjudication classification (ultrasound). ROC, receiver operating characteristic.

BI-RADS, the overall Kappa coefficient was 0.415, indicating moderate consistency. After AI model assistance, the Kappa coefficient increased to 0.702, reflecting strong consistency with a significant difference ($p < 0.001$). After grouping according to BI-RADS classification, Kappa values increased across all groups, indicating improved classification agreement.

Changes in Physician BI-RADS Classification After AI Model Assistance

After AI model assistance, five physicians re-classified the BI-RADS categories, with varying degrees of classification upgrades and downgrades (Tables 11,12). In the images of benign nodules, physicians made a total of 71 reclassifications,

Table 6. Analysis of BI-RADS classification by the AI model on the validation set ultrasound images (cases).

Physician adjudication (Ultrasound)	AI model classification					Total
	3	4a	4b	4c	5	
3	49	2	1	0	0	52
4a	0	33	1	0	0	34
4b	0	0	21	3	0	24
4c	0	0	1	21	0	22
5	0	0	1	0	21	22
Total	49	35	25	24	21	154

Table 7. Comparison of AI model and physician adjudication (ultrasound) in diagnosing breast nodule benignity and malignancy in the validation set of ultrasound images.

Pathological examination	Physician adjudication (Ultrasound)		AI model	
	Benignity	Malignancy	Benignity	Malignancy
Benignity (n = 111)	102	9	104	7
Malignancy (n = 43)	8	35	5	38
Total	110	44	109	45
Kappa	0.728		0.809	
p-value	<0.001		<0.001	

with the highest proportion being downgraded from category 4 to category 3 (57.75%, 41/71). In the images of malignant nodules, physicians made a total of 50 reclassifications, with the highest proportion being upgraded from category 4 to category 5 (70.00%, 35/50).

Discussion

To date, there is no standardised classification for imaging features in breast ultrasound BI-RADS diagnosis. Due to variations in physicians' work experience and knowledge levels, the subjective differences in classification results remain substantial in clinical practice (Wen et al, 2021). Given that the probability of malignancy for category 4 nodules ranges from 2% to 95%, physicians generally recommend histological and pathological examination to minimise the risk of missed diagnoses and misdiagnoses, which may lead to unnecessary biopsies of certain benign nodules (Zhao et al, 2020). In this study, we developed an AI model based on deep learning to assist physicians in BI-RADS classification, aiming to enhance diagnostic consistency in ultrasound imaging and reduce unnecessary biopsies for some benign category 4 nodules.

AI has demonstrated promising results in the diagnosis of benign and malignant breast nodules. However, its application in BI-RADS classification and subclassification remains limited. Huang et al (2019) proposed a two-stage CNN model to classify breast ultrasound images according to BI-RADS, achieving coincidence rates of 99.8% (category 3), 94.0% (category 4a), 73.4% (category 4b), 92.2% (cat-

Table 8. Diagnostic value of the AI model and physician adjudication (ultrasound) for assessing the benignity and malignancy of breast nodules in the validation set of ultrasound images.

Detection Method	AUC	SE	95% CI	Sensitivity (%)	Specificity (%)	Accuracy (%)
Physician adjudication (Ultrasound)	0.898	0.030	0.839–0.958	81.40	91.89	88.96
AI model	0.914	0.030	0.856–0.972	88.37	93.69	92.21

Comparison: No statistically significant difference between AI and physician adjudication ($p = 0.060$).

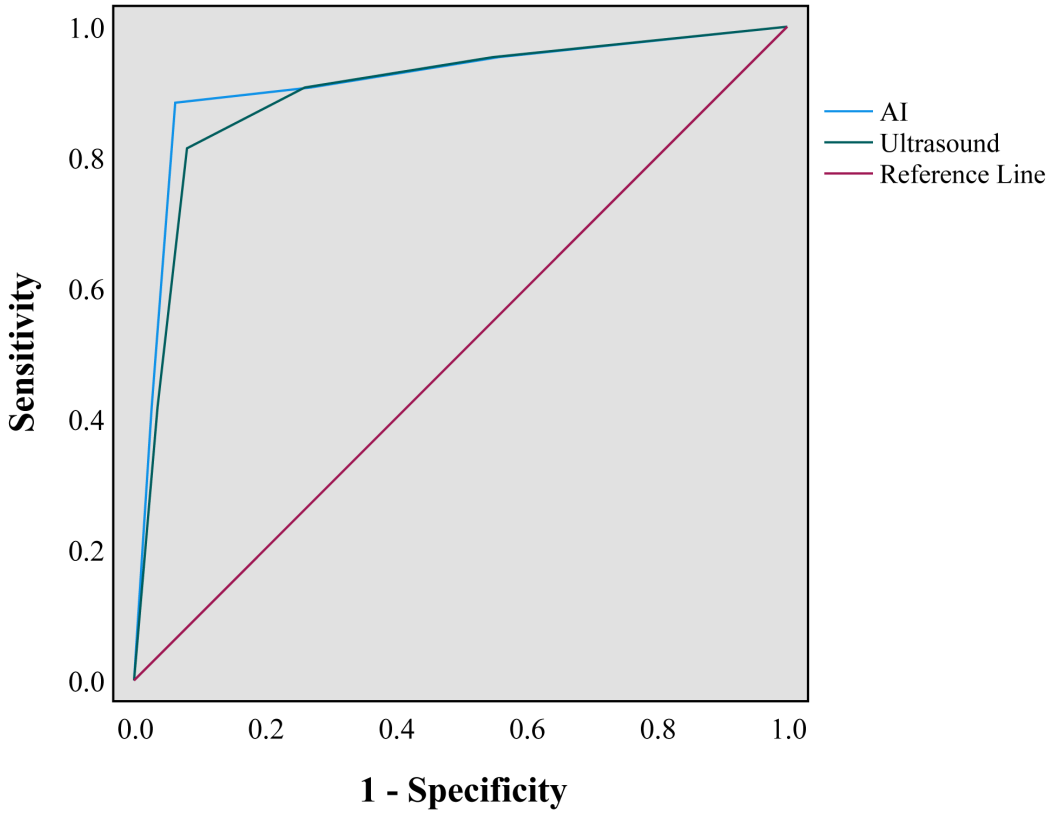

Fig. 3. ROC curve for differentiating benign and malignant lesions in the validation set images using the AI model and physician adjudication classification (ultrasound).

egory 4c), and 87.6% (category 5). These results are comparable to the diagnostic coincidence rate of the AI model established in this study. Additionally, no statistically significant difference was observed between the AUC of the AI model and that of the physician adjudication, indicating that the AI model exhibits high diagnostic efficiency in BI-RADS classification. Therefore, AI-based software demonstrates high accuracy in analysing breast ultrasound image features, consistent with findings from Wang et al (2024).

Previous studies have shown that AI models can assist young doctors or less experienced physicians in improving the accuracy of classifying nodules and reducing unnecessary biopsies (Wang et al, 2021; Xing et al, 2024; Yang et al, 2023). In this study, the diagnostic efficiency of doctors improved slightly with AI assistance

Table 9. Diagnostic performance of five physicians before and after AI model assistance in BI-RADS 3–5 breast nodule classification.

Detection method	AUC	SE	95% CI	Sensitivity (%)	Specificity (%)	Accuracy (%)
Before AI assistance	0.904	0.026	0.853–0.954	80.85	91.59	88.31
After AI assistance	0.926	0.023	0.880–0.972	89.36	92.52	91.56

Comparison: No statistically significant difference between AI-assisted and non-assisted physicians ($p = 0.256$).

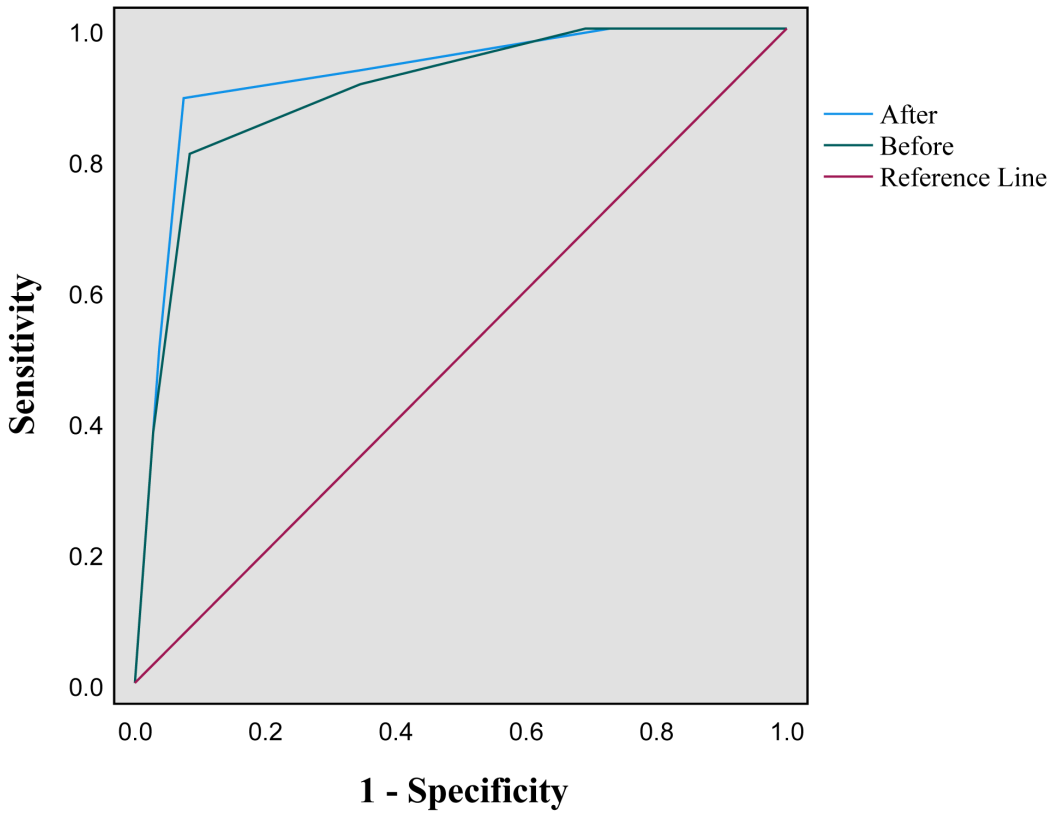

Fig. 4. ROC curve for physician performance before and after AI assistance in diagnosing BI-RADS category 3–5 breast nodules.

compared to independent diagnosis, though the difference was not statistically significant. However, since doctors assessed only 154 images with AI assistance, the sample size was relatively small. Therefore, the impact of AI on the diagnostic efficiency of more experienced sonographers may not be fully reflected.

Before using AI model assistance, the overall Kappa coefficient among physicians was low, indicating poor consistency, which aligns with previous studies. Jales et al (2013) studied the interobserver variability in the BI-RADS classification of category 4 nodules, with Kappa coefficients of 0.48 and 0.58, showing moderate agreement. In comparison, the consistency of BI-RADS classification for categories 3, 4a, and 4b in this study was lower than that of category 4, with Kappa coefficients of 0.424, 0.365, and 0.293, respectively. This may be attributed to the lack of clear classification standards for categories 3, 4a, and 4b, leading

Table 10. Consistency analysis of BI-RADS classification of breast nodules among five physicians before and after AI assistance.

BI-RADS classification	Kappa coefficient		*p*-value
	Before AI assistance	After AI assistance	
Category 3	0.424	0.857	<0.001
Category 4a	0.365	0.714	<0.001
Category 4b	0.293	0.646	<0.001
Category 4c	0.540	0.623	<0.001
Category 5	0.519	0.629	<0.001
Overall consistency	0.415	0.702	<0.001

Table 11. Changes in BI-RADS classification of benign breast nodules by five physicians before and after assistance.

Physician	Total cases (n)	Upgrade		Downgrade	
		Category 3 → Category 4	Category 4 → Category 5	Category 4 → Category 3	Category 5 → Category 4
1	107	7 (6.54)	1 (0.93)	4 (3.74)	1 (0.93)
2	107	7 (6.54)	0 (0.00)	12 (11.21)	0 (0.00)
3	107	5 (4.67)	0 (0.00)	9 (8.41)	0 (0.00)
4	107	5 (4.67)	0 (0.00)	10 (9.35)	0 (0.00)
5	107	2 (1.87)	1 (0.93)	6 (5.61)	1 (0.93)

to variations in physicians' interpretations of the ultrasound characteristics of the same nodule. Additionally, since all cases included in this study involved patients undergoing surgical treatment, selection bias may have influenced the results.

Following AI model assistance, the overall Kappa coefficient increased to 0.702, indicating a strong level of consistency. The agreement among physicians across different BI-RADS categories also improved, suggesting that AI-assisted classification can reduce subjective differences arising from differences in individual experience and interpretation of ultrasound features. By minimising differences in classification results, AI assistance alleviates unnecessary mental stress for patients and enhances the reliability of clinical decision-making.

This study primarily analysed the upward and downward reclassification of BI-RADS categories 3 to 5 but did not statistically evaluate reclassification within the four category 4 subgroups. The rationale is that, according to the American Society of Radiology guidelines, nodules in categories 3, 4a, and 4b can be followed up and observed, while those in category 4c and above require histological biopsy for definitive diagnosis. Therefore, even if reclassification occurs within category 4 subgroups, the clinical management recommendation remains an invasive biopsy.

Overall, the AI model demonstrated a valuable auxiliary role in BI-RADS classification. It effectively downgrades some benign nodules from category 4 to category 3, reducing unnecessary biopsies and alleviating the psychological burden on patients. At the same time, it upgrades some malignant nodules from category 4 to

Table 12. Changes in BI-RADS classification of malignant breast nodules by five physicians before and after AI assistance.

Physician	Total cases (n)	Upgrade		Downgrade	
		Category 3 → Category 4	Category 4 → Category 5	Category 4 → Category 3	Category 5 → Category 4
1	47	1 (2.13)	9 (19.15)	0 (0.00)	3 (6.38)
2	47	0 (0.00)	6 (12.77)	0 (0.00)	2 (4.26)
3	47	0 (0.00)	8 (17.02)	1 (2.13)	3 (6.38)
4	47	0 (0.00)	5 (10.64)	0 (0.00)	2 (4.26)
5	47	0 (0.00)	7 (14.89)	2 (4.26)	1 (2.13)

category 5, enhancing diagnostic accuracy, accelerating the diagnosis and treatment process, and preventing delays in patient management. These findings highlight the potential clinical utility of the AI model in improving diagnosis and treatment processes.

However, this study is limited by its sample size, which may impact the representativeness of different breast nodule types and the generalizability of the results. Additionally, the lack of longitudinal data precludes the ability to track changes in nodules over time. Furthermore, this study only included ultrasound images and classified them according to individual static images. In clinical practice, breast nodule assessment typically integrates patient-specific factors such as age, symptoms, and family alongside real-time dynamic scanning, elastography, and contrast-enhanced ultrasound for a comprehensive evaluation. The absence of these additional diagnostic modalities may contribute to lower diagnostic specificity and accuracy in the performance of the AI model. Moreover, unlike traditional machine learning methods, deep learning-based AI models are usually complex and involve numerous parameters. Their prediction results often lack interpretability, posing a major bottleneck to their clinical adoption.

Future research should focus on establishing more accurate and efficient AI models that can be integrated into routine diagnostic and screening workflows. Incorporating multimodal imaging, including dynamic ultrasound images, could provide a more comprehensive BI-RADS classification of nodules. Additionally, conducting multi-centre studies with an increased number of cases will improve model performance and explore its application across different healthcare settings and specialities.

Conclusion

In summary, an AI model based on deep learning, constructed using ultrasound images, can enhance the differentiation between benign and malignant breast nodules and improve nodule classification, reducing the incidence of missed and incorrect diagnoses. Additionally, the diagnostic performance of the model can be further enhanced in the future by optimising AI algorithm parameters, implementing alternative computational approaches, or designing novel CNN architectures.

Key Points

- The classification of imaging signs for breast ultrasound BI-RADS diagnosis remains non-standardised, leading to unnecessary misdiagnoses and missed diagnoses that affect patient treatment.
- AI technology improves the sensitivity and accuracy of breast ultrasound image interpretation, reducing misdiagnoses with high sensitivity (89.36%) and specificity (92.52%) based on different imaging characteristics of breast nodules.
- AI-assisted classification helps minimise subjective differences among physicians in classification due to variations in personal experience or interpretation of ultrasound signs, avoiding unnecessary mental stress on patients caused by inconsistencies in classification results.
- AI-assisted diagnosis improves inter-physician consistency without reducing diagnostic efficiency, effectively reducing unnecessary biopsies for benign nodules while improving the accuracy of malignant nodules' classification levels.

Availability of Data and Materials

All the data of this study are included in this article.

Author Contributions

SZ had the original conception of the work. XC collected the clinical data. WZ and PQ performed the research. SZ drafted the manuscript. All authors contributed to the important editorial changes in the manuscript. All authors read and approved the final manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

This study was approved by the Medical Ethics Committee of The People's Hospital of Pingyang County (No. IRB-2023-04) and complied with the relevant provisions of the Declaration of Helsinki. The waiver of informed consent in this section was granted by The People's Hospital of Pingyang County.

Acknowledgement

The author would like to acknowledge the support and assistance of The Pingyang County People's Hospital.

Funding

This study is supported by the Wenzhou Municipal Science and Technology Plan Project (Y2023556).

Conflict of Interest

The authors declare no conflict of interest.

References

Alzahrani NM, Henry AM, Clark AK, Al-Qaisieh BM, Murray LJ, Nix MG. Dosimetric impact of contour editing on CT and MRI deep-learning autosegmentation for brain OARs. Journal of Applied Clinical Medical Physics. 2024; 25: e14345. https://doi.org/10.1002/acm2.14345

Bartolotta TV, Orlando AAM, Di Vittorio ML, Amato F, Dimarco M, Matranga D, et al. S-Detect characterization of focal solid breast lesions: a prospective analysis of inter-reader agreement for US BI-RADS descriptors. Journal of Ultrasound. 2021; 24: 143–150. https://doi.org/10.1007/s40477-020-00476-5

Cheng M, Tong W, Luo J, Li M, Liang J, Pan F, et al. Value of contrast-enhanced ultrasound in the diagnosis of breast US-BI-RADS 3 and 4 lesions with calcifications. Clinical Radiology. 2020; 75: 934–941. https://doi.org/10.1016/j.crad.2020.07.017

Hong ZL, Chen S, Peng XR, Li JW, Yang JC, Wu SS. Nomograms for prediction of breast cancer in breast imaging reporting and data system (BI-RADS) ultrasound category 4 or 5 lesions: A single-center retrospective study based on radiomics features. Frontiers in Oncology. 2022; 12: 894476. https://doi.org/10.3389/fonc.2022.894476

Huang Y, Han L, Dou H, Luo H, Yuan Z, Liu Q, et al. Two-stage CNNs for computerized BI-RADS categorization in breast ultrasound images. Biomedical Engineering Online. 2019; 18: 8. https://doi.org/10.1186/s12938-019-0626-5

Jales RM, Sarian LO, Torresan R, Marussi EF, Alvares BR, Derchain S. Simple rules for ultrasonographic subcategorization of BI-RADS®-US 4 breast masses. European Journal of Radiology. 2013; 82: 1231–1235. https://doi.org/10.1016/j.ejrad.2013.02.032

Li J, Bu Y, Lu S, Pang H, Luo C, Liu Y, et al. Development of a Deep Learning-Based Model for Diagnosing Breast Nodules With Ultrasound. Journal of Ultrasound in Medicine. 2021; 40: 513–520. https://doi.org/10.1002/jum.15427

Li L, Deng H, Ye X, Li Y, Wang J. Comparison of the diagnostic efficacy of mathematical models in distinguishing ultrasound imaging of breast nodules. Scientific Reports. 2023; 13: 16047. https://doi.org/10.1038/s41598-023-42937-x

Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020; 42: 318–327. https://doi.org/10.1109/TPAMI.2018.2858826

Pan H, Shi C, Zhang Y, Zhong Z. Artificial intelligence-based classification of breast nodules: a quantitative morphological analysis of ultrasound images. Quantitative Imaging in Medicine and Surgery. 2024; 14: 3381–3392. https://doi.org/10.21037/qims-23-1652

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. Nature Communications. 2021; 12: 5645. https://doi.org/10.1038/s41467-021-26023-2

Sickles EA. ACR BI-RADS® Atlas, Breast imaging reporting and data system. American College of Radiology. 2013; 39.

Spak DA, Plaxco JS, Santiago L, Dryden MJ, Dogan BE. BI-RADS fifth edition: A summary of changes. Diagnostic and Interventional Imaging. 2017; 98: 179–190. https://doi.org/10.1016/j.diii.2017.01.001

Sun P, Feng Y, Chen C, Dekker A, Qian L, Wang Z, et al. An AI model of sonographer's evaluation+ S-Detect + elastography + clinical information improves the preoperative identification of benign and malignant breast masses. Frontiers in Oncology. 2022; 12: 1022441. https://doi.org/10.3389/fonc.2022.1022441

Wang XY, Cui LG, Feng J, Chen W. Artificial intelligence for breast ultrasound: An adjunct tool to reduce excessive lesion biopsy. European Journal of Radiology. 2021; 138: 109624. https://doi.org/10.1016/j.ejrad.2021.109624

Wang Z, Xu C, Zhou J, Wang Y, Xu Z, Hu F, et al. Accuracy of breast ultrasound image analysis software in feature analysis: a comparative study with sonographers. Scientific Reports. 2024; 14: 30724.

https://doi.org/10.1038/s41598-024-79773-6

Wen W, Liu J, Wang J, Jiang H, Peng Y. A National Chinese Survey on Ultrasound Feature Interpretation and Risk Assessment of Breast Masses Under ACR BI-RADS. Cancer Management and Research. 2021; 13: 9107–9115. https://doi.org/10.2147/CMAR.S341314

Xing B, Fu C, Yang Z. Diagnosis of Benign and Malignant Breast Nodules by Conventional Ultrasound in Combination with S-Detect Technology and Elastic Imaging. Journal of the College of Physicians and Surgeons–Pakistan. 2024; 34: 1154–1157. https://doi.org/10.29271/jcpsp.2024.10.1154

Yang L, Zhang B, Ren F, Gu J, Gao J, Wu J, et al. Rapid Segmentation and Diagnosis of Breast Tumor Ultrasound Images at the Sonographer Level Using Deep Learning. Bioengineering. 2023; 10: 1220. https://doi.org/10.3390/bioengineering10101220

Yi M, Lin Y, Lin Z, Xu Z, Li L, Huang R, et al. Biopsy or Follow-up: AI Improves the Clinical Strategy of US BI-RADS 4A Breast Nodules Using a Convolutional Neural Network. Clinical Breast Cancer. 2024; 24: e319–e332.e2. https://doi.org/10.1016/j.clbc.2024.02.003

Zhao C, Xiao M, Liu H, Wang M, Wang H, Zhang J, et al. Reducing the number of unnecessary biopsies of US-BI-RADS 4a lesions through a deep learning method for residents-in-training: a cross-sectional study. BMJ Open. 2020; 10: e035757. https://doi.org/10.1136/bmjopen-2019-035757