SCIENTIFIC Publishing Limited

Digital Technologies Research and Applications

http://ojs.ukscip.com/index.php/dtra

Article

Application Value of Artificial Intelligence Ultrasound-Assisted Diagnosis System in the Diagnosis of BI-RADS 4 Breast Nodules

Laifu Han [1,2], Delgerdalai Khashbat [3], Bat-Amgalan Byambasuren [4], Jinyan Wang [2], Wei Liu [5], Xiang Zhuang [6], Liwen Feng [7], Xu Han [8], Liguo Liu [2] and Shiirevnyamba Avirmed [9,*]

[1] Advanced School of Administration, Interdisciplinary Training Center, Mongolian National University of Medical Sciences, Ulaanbaatar 14210, Mongolia

[2] Department of Ultrasound, Baotou Cancer Hospital, Baotou 014030, China

[3] Department of Radiology, School of Medicine, Mongolian National University of Medical Sciences, Ulaanbaatar 14210, Mongolia

[4] Department of Supply Logistics and Procurement Services, Mongolian National University of Medical Sciences, Ulaanbaatar 14210, Mongolia

[5] Kezuo Houqi People's Hospital, Tongliao 028100, China

[6] Department of Radiology, Baotou Cancer Hospital, Baotou 014030, China

[7] Department of Pathology, Baotou Cancer Hospital, Baotou 014030, China

[8] Department of Breast Surgery, Baotou Cancer Hospital, Baotou 014030, China

[9] Department of Surgery, School of Medicine, Mongolian National University of Medical Sciences, Ulaanbaatar 14210, Mongolia

[*] Correspondence: shiirevnyamba@mnums.edu.mn

Received: 8 May 2026; **Revised:** 28 May 2026; **Accepted:** 25 June 2026; **Published:** 17 November 2026

Abstract: To investigate the clinical application value of an artificial intelligence (AI) ultrasound-assisted diagnostic system in the diagnosis of Breast Imaging Reporting and Data System (BI-RADS) category 4 breast nodules. A retrospective analysis of 231 breast nodules from 226 patients (aged 26 to 80 years, mean 55.52 ± 12.32 years) was carried out from December 2024 to December 2025. Nodules were divided into a malignant group and a benign group according to the pathological results of biopsy or surgery. Four diagnostic approaches were compared: low-experience ultrasound physicians (<5 years, US), high-experience ultrasound physicians (>15 years, US+), AI-assisted diagnosis alone (AI), and AI-assisted diagnosis combined with low-experience physicians (AI+US), all based on the BI-RADS classification system. A result of BI-RADS 4a or lower was regarded as negative, whereas 4b or above was positive. Sensitivity (SE), specificity (SP), accuracy (ACC), positive predictive value (PPV), negative predictive value (NPV) and area under the receiver operating characteristic curve (AUC) were computed with the pathological results as the gold standard. 128 of 231 nodules were malignant and 103 benign. The AI group achieved SE 82.03%, SP 80.58%, ACC 81.38%, PPV 84.00%, NPV 78.30%, and AUC 0.813 (95% CI: 0.756–0.870), exceeding the US group (all $p < 0.05$ vs. US). The AI ultrasound-assisted diagnostic system shows good efficacy in the diagnosis of BI-RADS 4 breast nodules, especially for small nodules.

Keywords: Breast Nodules; Ultrasonic Examination; Artificial Intelligence

1. Introduction

Breast cancer is the most commonly diagnosed malignancy in women worldwide [1]. In the early stage, the disease usually causes no obvious symptoms, so early breast cancer is often misdiagnosed as a benign breast nodule. Consequently, the detection rate of early-stage breast cancer remains low. Accurate early diagnosis is therefore crucial, as it helps optimize treatment, improve survival, and enhance long-term prognosis [2]. The reliability of early diagnosis depends largely on the imaging modality used to evaluate the breast.

Ultrasound is one of the most widely used modalities for this purpose. Because it involves no ionizing radiation and offers good repeatability, high availability, and real-time imaging at a much lower cost than magnetic resonance imaging (MRI) or positron emission tomography (PET), ultrasound is especially valuable for dense breast tissue, where the sensitivity of mammography is markedly reduced. To standardize image interpretation, the American College of Radiology (ACR) developed the Breast Imaging Reporting and Data System (BI-RADS), a set of criteria intended to minimize interpretive variability and guide clinical management [3]. In practice, however, BI-RADS assessment still depends on image quality and on the experience of the interpreting physician, and both intra- and inter-reader variability remain substantial. This problem is most evident in the BI-RADS 4 category, which spans a very wide malignancy risk (roughly 2% to 95%). Because guidelines recommend biopsy for every BI-RADS 4 lesion, this uncertainty frequently leads to unnecessary invasive procedures such as core-needle biopsy and surgery, causing patient morbidity, psychological distress, and added health-care costs. An accurate and objective ancillary tool is therefore needed to help clinicians stratify BI-RADS 4 nodules more reliably [4].

Artificial intelligence has emerged as a promising tool for this purpose. AI technology, especially deep learning based on convolutional neural networks, is increasingly being applied to medical image detection, segmentation, and classification. Unlike human observers, CNN-based systems can consistently quantify multidimensional sonographic features, including lesion shape, margin characteristics, internal echogenicity, posterior acoustic features, and vascular distribution, while reducing the effects of observer fatigue and subjective variation [5, 6]. A prospective multicenter study involving 901 breast masses demonstrated that ultrasound computer-aided diagnosis achieved favorable diagnostic performance and could help reduce unnecessary biopsies [7]. Several studies have also shown that AI-assisted ultrasound can improve diagnostic performance, with particularly evident benefits for less-experienced physicians. Choi et al. [6] found that computer-aided diagnosis produced the greatest improvement among readers with less experience. Kim et al. [5] reported that S-Detect achieved diagnostic performance comparable to that of dedicated breast radiologists. Lv et al. [8] similarly observed that an AI-based automatic detection system improved the diagnostic performance of physicians with different levels of experience. More recently, Xing et al. [9] demonstrated that combining S-Detect with BI-RADS improved the differential diagnosis of BI-RADS 4 nodules and was particularly beneficial for nodules measuring no more than 20 mm.

Progress in breast imaging and pathological image analysis further demonstrates the potential of deep learning. Hejduk et al. [10] developed a deep convolutional neural network for the automatic detection and BI-RADS classification of lesions on automated breast ultrasound images. Islam et al. [11] reported that a CNN-based framework could effectively classify invasive ductal carcinoma and metastatic breast cancer using histopathological images. Broader investigations of AI in mammography and digital breast tomosynthesis have also demonstrated improvements in lesion detection and risk assessment [12]. In addition, Tsuyuzaki et al. [13] directly compared an AI diagnostic system with radiologists and evaluated the effect of AI assistance, providing further evidence that AI can support breast ultrasound interpretation and improve diagnostic efficiency. Together, these findings underscore the technical maturity and clinical potential of integrating AI into routine breast imaging practice.

Despite this progress, the value of AI-assisted ultrasound in the diagnostically difficult BI-RADS 4 category has not been adequately evaluated, particularly for small nodules, where conventional ultrasound is most limited. Two questions remain largely unanswered: how the level of AI assistance interacts with the physician's clinical experience, and to what extent AI can compensate for limited expertise. To address these gaps, the present study evaluates the TenD.AI ultrasound-assisted system for distinguishing benign from malignant BI-RADS 4 breast nodules, with attention to nodule-size stratification and the quantitative benefit of AI–physician collaborative diagnosis.

The remainder of the article is organized as follows. Section 2 describes the materials and methods. Section 3 reports the pathological and diagnostic results for the four diagnostic strategies and three nodule-size categories. Section 4 discusses the clinical implications, the underlying mechanisms, the study limitations, and future direc-

tions. Section 5 summarizes the conclusions and contributions.

2. Materials and Methodology

2.1. Study Population

This study consisted of a retrospective analysis of 226 patients (231 nodules) who were diagnosed with BI-RADS 4 category nodules by the breast ultrasound examination at the Baotou Cancer Hospital from December 2024 to December 2025. All patients were first time visitors aged between 26 and 80 years with mean age of 55.52 ± 12.32 years. The maximum nodule diameter varied from 7 to 41 mm having a mean of 22.47 ± 8.96 mm. The mean age did not differ significantly between the benign group (54.38 ± 13.01 years) and the malignant group (56.42 ± 11.78 years) (independent samples t-test; Levene's test for equality of variances: F = 1.23, p = 0.268; t = −1.33, df = 229, p = 0.184). The mean nodule diameter was significantly different between the benign group (20.15 ± 9.24 mm) and the malignant group (24.33 ± 8.37 mm) (Levene's test: F = 2.15, p = 0.144; t = −3.54, df = 229, p = 0.001), reflecting the known correlation between nodule size and the risk of malignancy.

Inclusion criteria: ① The nodules were diagnosed by routine breast ultrasound examination, and the BI-RADS classification was 4; ② The nodules were confirmed by biopsy and surgical pathology; ③ The patients had no other malignant tumor history.

Exclusion criteria: ① Non-primary breast tumors; ② Patients who had undergone surgery or chemotherapy, hormone therapy, or radiotherapy before ultrasound examination; ③ Mental disordered and patients who were unable to cooperate with the examination.

This study was approved by the Ethics Committee of Baotou Cancer Hospital [2025(024)] and the Ethics Committee of the Mongolian National University of Medical Sciences [2025.09.26 25-26/01-02]. The study was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants before the ultrasound examination and data collection.

2.2. Conventional Ultrasound Examination

A conventional ultrasound was independently performed by junior (less than 5 years experience with ultrasound, n = 3 physicians) and senior physicians (more than 15 years experience, n = 2 physicians). The ultrasonic diagnostic equipment used was the Mindray R9 ultrasound diagnostic equipment with broadband linear probe model L14-5WU (frequency 5–14 MHz) (Mindray Co., Ltd., Shenzhen, China) and the Hitachi Preirus ultrasound diagnostic equipment with probe model EUP-L74M (frequency 5–14 MHz) (Hitachi, Ltd., Tokyo, Japan). Both systems ran at similar diagnostic frequencies and scanning protocols were standardised between systems. Patients were examined with both arms lifted, exposing the bilateral breasts and axillary regions, in the supine position.

Initial conventional ultrasound examination was performed by junior doctors according to the standards and diagnostic criteria outlined in the 6th edition of Ultrasound Medicine [6]. Following the identification of the lesion, standard longitudinal and transverse sections were obtained in order to measure the dimensions of the lesion. Acoustic characteristics such as boundary characteristics, morphological regularity, internal echogenicity, presence of calcification foci, Doppler blood flow signals were systematically recorded and stored as dynamic ultrasound image sequences. BI-RADS classification was independently given by junior doctors after real-time assessment, and recorded before consulting AI or senior doctors.

Then, two senior ultrasound physicians independently and blindly re-evaluated the saved dynamic image sequences, without junior physician classifications or the AI outputs. An independent BI-RADS classification was given by each senior physician. If there was discrepancy between the two senior doctors, it was discussed together in a structured manner and consensus was obtained, thereby treating it as the group result of US+ for the case. Junior physician group (US) and senior physician group (US+) were therefore considered as completely different diagnostic groups. Lesions that were BI-RADS 4a or less were put into the negative (benign) group and those that were BI-RADS 4b or more, into the positive (malignant) group.

2.3. Breast Ultrasound AI System

The TenD.AI Medical Technology system (Shenzhen Xinzhixin Medical Technology Co., Ltd., Shenzhen, China) was used for AI-assisted diagnostic platform in this study. The system adopts a deep convolutional neural network

(CNN) architecture comprising multiple stacked convolutional and pooling layers that can capture hierarchical features of images with different spatial scales, allowing to simultaneously capture both the global morphologic features and fine micro-structures in breast nodules. The system has been trained using a large-scale multicenter dataset of over 100,000 annotated breast ultrasound dynamic image sequences that represents a variety of nodule morphologies, sizes, echo patterns, and pathological subtypes. To make the training data more representative of diverse patient groups and scanning devices, data were obtained from various clinical institutions. Standard preprocessing was performed, such as automatic gain normalization, noise artifact suppression, and automated extraction of the lesion region-of-interest (ROI). In order to deal with the class imbalance problem in the training samples, the system used a weighted cross entropy loss function, which weighted the error of the negative class more heavily, and adopted online data augmentation methods like random rotation, horizontal flip, and elastic deformation. Optimisation of the model was carried out using the Adam Optimizer, and also used a learning rate scheduling strategy with warm-up and cosine annealing. Prior to clinical deployment, the system was validated on an independent external hold-out set, reporting an AUC of 0.851 (95% CI: 0.823–0.879) on the external validation set, high sensitivity of 83.6% and specificity of 81.2% for malignancy detection. Average inference time per nodule is about 2–3 s, which can be achieved on a typical clinical workstation (Intel Core i7 processor, 16 GB RAM, onboard NVIDIA GTX 1080 GPU) and allows for real-time integration into the clinical ultrasound workflow without substantial procedure delay.

At the time of the actual examination, the AI software was turned on simultaneously with the standard ultrasound examination, identified and localized the suspicious lesions, and annotated them with bounding boxes based on the CNN algorithm, and then exported BI-RADS classification and malignancy probability scores to the clinical server (**Figure 1**). The AI+US group consisted of junior ultrasound physicians who viewed the AI output after first independently classifying them and then re-assigned final BI-RADS classifications after consulting the AI output. This sequential workflow made sure the AI+US classifications were reflective of true physician-AI integration, not just AI output adoption.

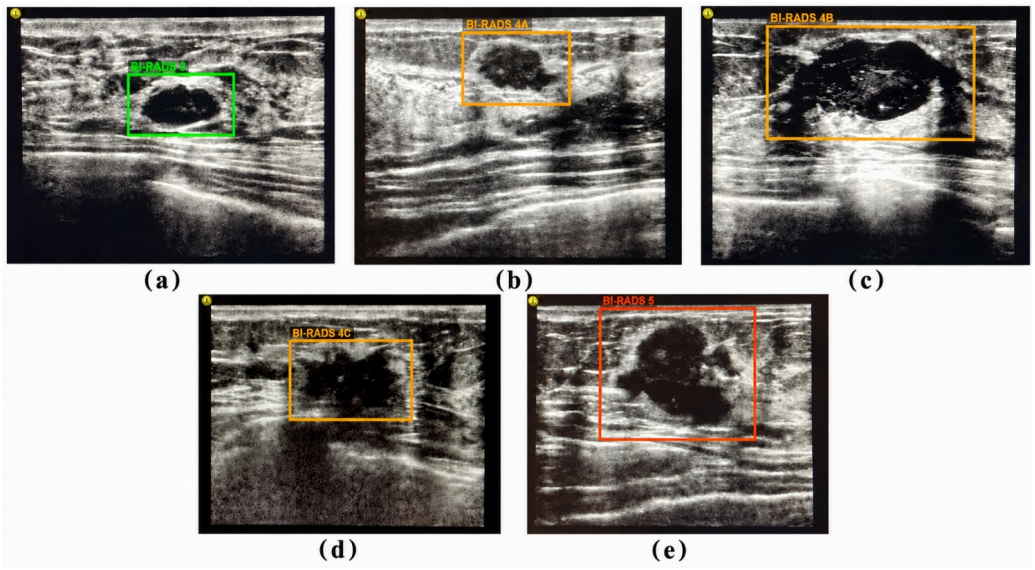


Figure 1. Example of automatic identification of breast nodules and BI-RADS classification by the breast ultrasound AI-assisted diagnostic system. (**a**) indicates BI-RADS category 3, (**b**) indicates BI-RADS category 4a, (**c**) indicates BI-RADS category 4b, (**d**) indicates BI-RADS category 4c, and (**e**) indicates BI-RADS category 5.

2.4. Statistical Analysis

Data were analysed using SPSS 23.0 software (IBM Corp., Armonk, NY, USA). Normality of continuous variables was tested using the Shapiro-Wilk test and homogeneity of variance was tested using Levene's test before parametric testing. For normally distributed continuous variables (age and nodule diameter, both with Shapiro-Wilk: $p > 0.05$), the independent samples *t*-test was used to test for differences between the benign and malignant groups. All

continuous variables used in this study met the normality criteria; otherwise the Mann Whitney U test would have been used to compare variables that were not normally distributed. The chi-square (χ^2) test or Fisher's exact test was used as appropriate for comparing categorical variables, with Fisher's exact test used when cell values were less than five.

SE, SP, ACC, PPV and NPV were computed for each diagnostic method against the gold standard of pathological confirmation. A 95% confidence interval (95% CI) for each binary performance measure were determined by the Wilson score method. ROC curves were generated based on the framework of the pROC package and AUCs with 95% CI were determined nonparametrically using the DeLong method. The DeLong Z-test was used to compare the AUC values of the different diagnostic methods in pairs, and Bonferroni correction was performed for multiple comparisons (corrected $\alpha = 0.0083$), with 6 comparisons made between different methods. Inter-observer agreement was calculated using Cohen's Kappa coefficient (κ) by two different methods: between the AI system and physician groups and between the physician groups, with κ values categorized as follows: <0.20 poor, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and >0.80 almost perfect. Diagnostic accuracy in each nodule size group was compared by chi-square tests and the AUC values for each nodule size group were separately calculated and compared by the DeLong method. Cohen's h was reported as effect size for significant findings of proportional comparisons. Unless otherwise stated, p values < 0.05 (two-tailed) were considered statistically significant.

3. Result

3.1. Pathological Results

A total of 226 patients had 231 breast nodules. Biopsy or surgical pathological examination was used as the gold standard diagnostic procedure for all patients. Pathological examination revealed 103 benign nodules, including 41 cases of breast adenosis, 34 of breast fibroadenoma, 12 of intraductal papilloma, 8 of granulomatous mastitis, 4 of acute or chronic inflammatory cell infiltration of the breast tissues, 2 of breast duct dilation with cyst formation, and 2 of adenosis with local ductal papillary hyperplasia. A total of 128 malignant nodules were found, including 105 cases of invasive ductal carcinoma, 12 cases of ductal carcinoma in situ, 8 cases of invasive lobular carcinoma, and 3 cases of mucinous carcinoma. Considering the BI-RADS 4 nodules in this study, the rate of malignancy was 55.4% (128/231); that is, nearly half of BI-RADS 4 lesions in this study were benign and if used alone, would have been subjected to unnecessary invasive procedure.

3.2. Conventional Ultrasound Examination Diagnostic Results

Table 1 lists diagnostic performance metrics of junior (US) and senior (US+) ultrasound physicians. The US group achieved SE 73.43% (95% CI: 64.9–80.7%), SP 70.88% (95% CI: 61.2–79.1%), ACC 72.29% (95% CI: 65.9–77.9%), PPV 75.80% (95% CI: 67.1–83.0%), NPV 68.22% (95% CI: 58.6–76.5%), and AUC 0.722 (95% CI: 0.657–0.787). The US+ group achieved SE 86.71% (95% CI: 79.6–91.8%), SP 85.44% (95% CI: 77.0–91.5%), ACC 86.15% (95% CI: 81.0–90.3%), PPV 88.09% (95% CI: 80.9–93.1%), NPV 83.81% (95% CI: 75.2–90.2%), and AUC 0.861 (95% CI: 0.813–0.908). In the US+ group all performance measures were significantly superior to the US group (all $p < 0.05$; AUC comparison: $Z = 3.87$, $p < 0.001$). The Cohen's Kappa between US and US+ classifications was $\kappa = 0.612$ (95% CI, 0.521–0.703) indicating substantial agreement between the groups, and the part of the agreement not shown by the kappa was due to the experience-dependent component of diagnostic performance.

Table 1. Comparison of diagnostic efficacy of different ultrasound diagnostic methods in differentiating benign and malignant breast nodules.

Diagnostic method	Benign $N = 103$	Malignant $N = 128$	SE/%	SP/%	ACC/%	PPV/%	NPV/%	AUC
US			73.43	70.88	72.29	75.80	68.22	0.722
Negative group (−)	73	34						
positive group (+)	30	94						
US$^+$			86.71	85.44	86.15	88.09	83.81	0.861
Negative group (−)	88	17						
positive group (+)	15	111						
AI			82.03	80.58	81.38	84.00	78.30	0.813
Negative group (−)	83	23						
positive group (+)	20	105						

Table 1. *Cont.*

Diagnostic method	Benign *N* = 103	Malignant *N* = 128	SE/%	SP/%	ACC/%	PPV/%	NPV/%	AUC
AI+US			88.28	86.41	87.45	88.98	85.58	0.876
Negative group (−)	89	15						
positive group (+)	14	113						

3.3. AI System and Combined AI+US Diagnostic Results

The AI system alone achieved SE 82.03% (95% CI: 74.3–88.2%), SP 80.58% (95% CI: 71.7–87.6%), ACC 81.38% (95% CI: 75.6–86.3%), PPV 84.00% (95% CI: 76.3–89.9%), NPV 78.30% (95% CI: 69.4–85.5%), and AUC 0.813 (95% CI: 0.756–0.870). The AI+US combined approach achieved SE 88.28% (95% CI: 81.5–93.3%), SP 86.41% (95% CI: 78.3–92.3%), ACC 87.45% (95% CI: 82.5–91.4%), PPV 88.98% (95% CI: 82.3–93.7%), NPV 85.58% (95% CI: 77.1–91.8%), and AUC 0.876 (95% CI: 0.829–0.923). The Cohen's Kappa between AI and pathological classification was $\kappa = 0.623$ (95% CI range: 0.533–0.713; substantial agreement), while physician-AI integration (AI+US) had a good agreement with pathological classification ($\kappa = 0.746$; 95% CI range: 0.665–0.827; substantial to almost perfect agreement).

3.4. Comparison among Diagnostic Methods

Full pairwise comparisons can be seen in **Table 1**. The AI group performed significantly better than the US group in SE, SP, ACC and AUC (all $p < 0.05$ before Bonferroni correction; AUC comparison: $Z = 2.14$, $p = 0.032$). There was no significant difference between the performance of the AI group and the US+ group at any of the metrics used after Bonferroni correction (AUC comparison: $Z = 1.23$, $p = 0.218$). The AI+US group was significantly superior to the US group on all measures (all $p < 0.05$; AUC comparison: $Z = 3.12$, $p = 0.002$). All metrics showed numerically higher values in the AI+US group as compared to the US+ group, but did not reach statistical significance (AUC comparison: $Z = 0.89$, $p = 0.372$). The AI+US group showed improvements in all measures compared to AI alone, but after applying Bonferroni correction the difference was not independently statistically significant ($Z = 1.43$, $p = 0.153$, AUC comparison), indicating that the combination of junior physicians and AI offers additive but not independently statistically significant benefits over AI alone for this sample size.

3.5. Diagnostic Results for Different-Sized Breast Nodules

The 231 breast nodules had maximum diameters ranging from 4 to 42 mm, with a mean of 23.5 ± 10.26 mm. Nodules were stratified into three size categories based on maximum diameter: <15 mm (N = 58; 33 benign, 25 malignant), 15–30 mm (N = 103; 46 benign, 57 malignant), and >30 mm (N = 70; 24 benign, 46 malignant). **Table 2** shows the diagnostic accuracy of each method for nodules of different sizes, using the size cutoffs of 15 mm and 30 mm, which are widely cited in the BI-RADS literature and previous AI-assisted ultrasound studies, respectively, and for which nodules <15 mm are the most diagnostically challenging, with conventional sonographic malignancy features being the least reliably expressed [6]. Within the <15 mm subgroup, the US group exhibited the lowest accuracy at 62.07% (95% CI: 49.0–73.7%). The AI and AI+US groups achieved higher accuracy (77.59%, 95% CI: 65.2–87.2%; and 82.76%, 95% CI: 70.6–91.2%, respectively) compared with both US and US+ (75.86%, 95% CI: 63.1–86.0%) groups, although differences did not reach statistical significance on chi-square testing (all $p > 0.05$). AUC values within the <15 mm subgroup were: US 0.658 (95% CI: 0.516–0.800), US+ 0.752 (95% CI: 0.620–0.884), AI 0.789 (95% CI: 0.664–0.914), and AI+US 0.831 (95% CI: 0.715–0.947), as shown in **Figure 2**. The AI+US advantage in small nodule diagnosis was statistically significant with a pairwise AUC comparison between AI+US and US in the <15 mm group achieving a $Z = 1.98$, $p = 0.048$. For the 15–30 mm subgroup, US+ had the highest level of accuracy (88.35%), followed by AI+ (89.32%), US (67.96%) and AI (80.58%), but no significant differences were found between comparisons ($p > 0.05$). With respect to the >30 mm group, there was no statistically significant difference between the different groups ($p > 0.05$), and the US+ group showed the highest accuracy (91.43%), followed by the AI+US group (88.57%), US group (87.14%) and AI group (85.71%). The diagnostic accuracy of the US group in the three size categories was significantly different (62.07% vs. 67.96% vs. 87.14%; $\chi^2 = 14.37$, $p < 0.05$), validating the fact that nodule size has a significant effect on the performance of junior physicians, with the larger nodule size showing much higher accuracy in the diagnosis.

Table 2. Comparison of diagnostic accuracy of four ultrasound diagnostic methods for breast nodules of different sizes.

Diagnostic Method	< 15 mm, *N* = 58			15~30 mm, *N* = 103			> 30 mm, *N* = 70		
	Benign *N* = 33	Malignant *N* = 25	ACC *n* %	Benign *N* = 46	Malignant *N* = 57	ACC *n* %	Benign *N* = 24	Malignant *N* = 46	ACC *n* %
US			62.07			67.96			87.14
Negative group (−)	30	19		27	14		16	1	
positive group (+)	3	6		19	43		8	45	
US^{+}			75.86			88.35			91.43
Negative group (−)	29	10		41	7		18	0	
positive group (+)	4	15		5	50		6	46	
AI			77.59			80.58			85.71
Negative group (−)	30	10		38	12		15	1	
positive group (+)	3	15		8	45		9	45	
AI+US			82.76			89.32			88.57
Negative group (−)	31	8		42	7		16	0	
positive group (+)	2	17		4	50		8	46	

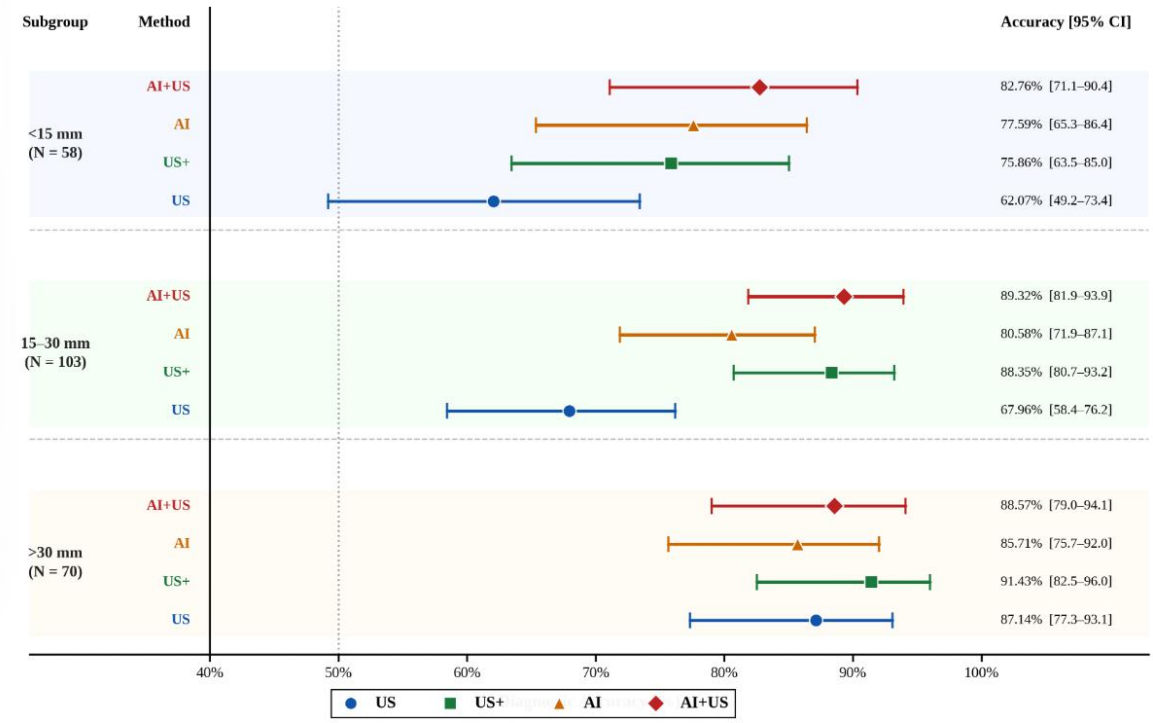


Figure 2. Comparison of diagnostic accuracy curves of four diagnostic methods for breast nodules of different sizes.

4. Discussion

4.1. Clinical Significance of AI in BI-RADS 4 Nodule Assessment

BI-RADS 4 is a diagnostic category that is unique and clinically significant. BI-RADS 4 is an extraordinary broad range of malignancy possibilities (not BI-RADS 3 which is <2%, and not BI-RADS 5 which is >95%), and is broken into three categories of 4a (>2% to ≤10%), 4b (>10% to ≤50%) and 4c (>50% to <95%). The current ACR guidelines suggest that biopsies should be performed for all BI-RADS 4 or higher, but in this study population 44.6% of BI-RADS 4 nodules were found to be benign. This means that a large percentage of patients received invasive procedures which may not have been performed if more accurate pre-biopsy risk stratification had been used. Thus, the use of AI-assisted diagnosis in the BI-RADS 4 evaluation workflow has significant implications for clinical practice and healthcare resources, going beyond diagnostic accuracy metrics.

4.2. AI Diagnostic Performance and Clinical Positioning

The AI system was able to achieve an AUC of 0.813; this score significantly outperformed junior physicians (AUC 0.722, $p = 0.032$) and was not significantly different from senior physicians (AUC 0.861, $p = 0.218$), indicating that the AI system can match the level of diagnostic expertise of intermediate-to-senior level physicians. These results are consistent with the findings of Chen et al. [14] and Xing et al. [15], who also demonstrated that AI-assisted breast ultrasound could improve the diagnostic performance for breast nodules, particularly small lesions. The AI system achieves this by objectively extracting high-dimensional sonographic feature vectors (margin irregularity, internal echo heterogeneity, posterior acoustic shadowing, and microcalcification pattern) with a CNN, free from human fatigue, attentional variation, and cognitive bias [16,17].

The combined AI+US approach had the highest overall AUC of 0.876, significantly higher than the US group ($p = 0.002$) and numerically higher than the US+ group ($p = 0.372$). The findings highlight that junior physicians can significantly reduce the performance gap with their senior counterparts when using AI, achieving outcomes comparable to senior physicians without significant differences, even given the 10+ year experience gap. It is important to highlight, however, that the improvement of AI+US compared with US+ was not statistically significant and, therefore, the findings about the superiority of AI+US over expert physicians should be interpreted with due care. In this study, with this smaller sample size, it could be that a significant difference would come to light with larger groups of patients; on the other hand, it could be that there is a true performance limit that AI-physician collaboration does not systematically surpass.

Of note, in this cohort 55.4% (128/231) of BI-RADS 4 nodules were pathologically confirmed as malignant, while a significant number of patients with BI-RADS 4 on conventional ultrasound underwent unnecessary biopsy or surgery. AI+US outperformed US alone by improving specificity to 86.41% vs. 70.88%, indicating significant potential to reduce false-positive rates, which would directly benefit patient experience, health care costs and resource allocation.

4.3. Superior Performance for Small Nodules

A key and clinically relevant result of this study is the differential performance gain of AI and AI+US in comparison to conventional ultrasound for small nodules (maximum diameter <15 mm). In this sub-population, AI and AI+US obtained 77.59% and 82.76% accuracy, respectively, which were higher than the accuracy of US (62.07%) which was only marginally higher than random accuracy for binary classification. The quantitative statistical differences between AI+US and US in the <15 mm subgroup was statistically significant ($Z = 1.98$, $p = 0.048$), which was not seen in the larger nodule categories.

For small nodules, the percentage of the ultrasound image covered by the nodules is also inherently limited, thus limiting the number of pixels that can be extracted by a human observer for analysis of nodule properties at the lesion margin and within the internal nodule architecture [16,17]. CNN architectures, on the other hand, can use learned convolutional filters, which are optimized to look for patterns associated with malignancy, to systematically evaluate every pixel and can perform consistent and exhaustive image analysis, regardless of lesion size. In this study, a representative case was found where the AI system gave a BI-RADS 4b (borderline malignancy) classification for a 7 mm mucinous carcinoma nodule, whereas both junior and senior physicians classified it as BI-RADS 4a (borderline benign). The AI system classified the micro-invasive margin features identified by the computational image analysis that were not detected by the human observers, and which were more aligned with the pathological diagnosis (**Figure 3**). This finding aligns with the results of Chen et al. [14], who demonstrated that an ultrasound AI-assisted diagnostic system was particularly valuable for diagnosing BI-RADS 4 breast nodules with a maximum diameter of ≤2 cm, and Xing et al. [15], who specifically evaluated S-Detect image analysis for small focal breast lesions.

The gap between the performance of the AI-assisted and human diagnostic methods becomes smaller with the increase in nodule size, especially when the nodule size exceeds 15 mm (15–30 mm and >30 mm) (**Figure 2**). For the >30 mm size group, the US+ group had the highest accuracy (91.43%) and no significant difference was found between the four methods. This convergence is due to the increased availability of macroscopic sonographic malignancy features in larger lesions, which are easily noted and understood by trained human sonographers without computational support: spiculated borders, posterior acoustic shadowing, and prominent vascularity. This implies

that the incremental value of AI is greatest where human visual diagnosis is most difficult—that is, in small nodules.

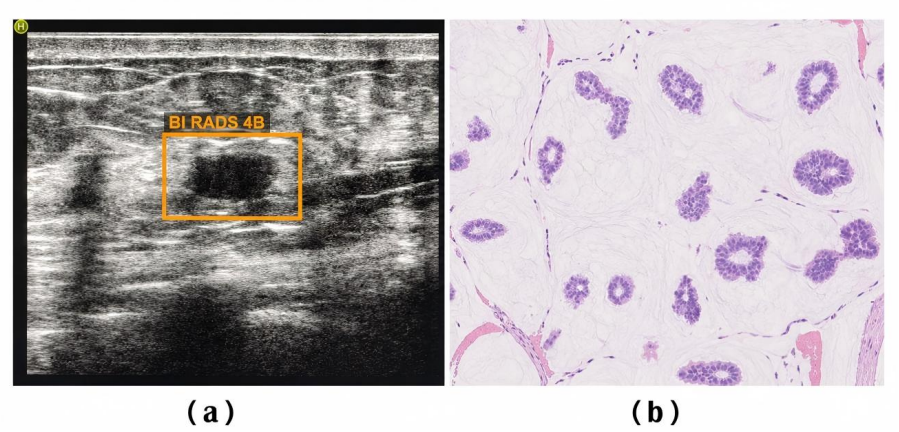


Figure 3. (**a**) The AI-assisted diagnostic system classified the small breast nodule as BI-RADS 4b with a malignant tendency; (**b**) Pathological diagnosis confirmed breast mucinous carcinoma.

4.4. Complementary Value of AI-Physician Collaboration

While the AI system generally performed well, the image-only analysis approach has its limitations, underscoring the critical role of physician-AI collaboration in diagnosis. A representative case of AI misdiagnosis was a 34 mm lesion which was classified as BI-RADS 4c (highly suggestive of malignancy) by the AI system and as BI-RADS 4a (low suspicion) by both physician groups. Pathology report showed granulomatous mastitis (inflammatory cause) which is benign. The purely image-based analysis by the AI system led to misclassification due to the irregular margins and indistinct borders of the lesion, which are characteristics similar to malignancy; importantly, the AI system could not incorporate clinical contextual information that informed the junior and senior physicians to a benign diagnosis, including the patient's young age, acute onset of tenderness on palpation, and systemic inflammatory symptomatology (**Figure 4**). The physician's incorporation of clinical data in the AI+US group successfully overcame the AI misclassification to arrive at the correct benign diagnosis.

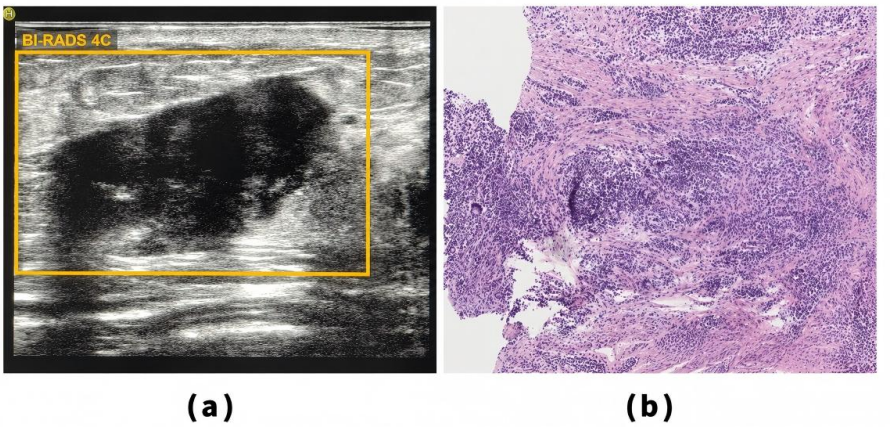


Figure 4. (**a**) The AI-assisted diagnostic system provided a BI-RADS 4c classification suggestive of malignancy; (**b**) Pathological diagnosis confirmed granulomatous mastitis.

This case illustrates a basic and so far unbridgeable constraint of image-only AI systems, that they cannot account for the clinical history, physical examination, inflammatory markers in the lab or epidemiological risk factors in their diagnostic reasoning. In contrast, human doctors, especially those with less experience, might overlook some of the markers of malignancy at the image level that the AI system can identify. This two-way complementarity, with AI picking up on subtle details in images that are not perceptible to the human eye, and physicians filling

in the gaps where AI is wrong, offers the best and most sensible model for deployment of AI in diagnostics in the current setting of clinical practice. The synergy between the AI system and the junior physician is undoubtedly most beneficial for the junior doctors because the AI system can be expected to make up for their lack of pattern recognition experience, and the junior doctors bring the clinical reasoning and patient interaction skills that the AI system cannot. The implications of this finding directly relate to the planning of healthcare resources and the potential use of AI-assisted ultrasound in primary care, community hospitals, and resource-constrained areas where access to senior physician expertise is limited.

4.5. Considerations on Class Imbalance, AI System Characteristics, and Computational Performance

A total of 128 malignant and 103 benign nodules were included in this study, corresponding to proportions of 55.4% and 44.6%, respectively, and indicating a modest class imbalance. According to the technical information provided by the manufacturer, the TenD.AI system was trained using a weighted cross-entropy loss function, which assigned different weights to classification errors from the two classes to mitigate the influence of class imbalance during model training. Online data augmentation, including random rotation, horizontal flipping, and elastic deformation, was also applied to improve data diversity and model robustness. Because the available technical information did not clearly document minority-class oversampling, this strategy was not included in the present interpretation. Weighted loss functions and data-level balancing strategies have been reported to be useful for managing imbalanced datasets in cancer diagnosis [16]. Nevertheless, the operating threshold of an AI model may be influenced by disease prevalence and the intended clinical setting. Therefore, further independent validation is required to determine whether threshold recalibration is necessary when the system is applied to populations with different malignancy prevalence, such as screening populations and patients referred for the assessment of symptomatic breast lesions.

Regarding computational performance, the manufacturer reported an average inference time of approximately 2–3 s per nodule when the system was operated on a clinical workstation equipped with an Intel Core i7 processor, 16 GB of RAM, and an NVIDIA GTX 1080 graphics processing unit. This processing speed suggests that the system may be incorporated into the routine ultrasound workflow without a substantial increase in examination time. However, its real-world processing efficiency and compatibility with different workstation configurations require further independent evaluation. Although a dedicated graphics processing unit is required under the reported configuration, the system does not depend on a GPU cluster or a high-performance computing platform, which may facilitate its deployment in community hospitals and primary healthcare institutions with appropriate hardware resources.

Additional evidence supports the clinical application of AI in breast ultrasound. Shen et al. developed an AI system using a large-scale breast ultrasound dataset and found that AI assistance reduced radiologists' false-positive rates and biopsy recommendations while maintaining sensitivity [17]. O'Connell et al. similarly reported that an AI system achieved sensitivity, specificity, and accuracy greater than 0.80, with performance comparable to that of the better-performing radiologists [18]. These findings support the potential of AI to improve diagnostic consistency and reduce unnecessary invasive procedures.

Prospective and model-development studies have also demonstrated the value of deep learning in BI-RADS-based assessment. Qian et al. found that an explainable multimodal deep learning system predicted BI-RADS scores with performance comparable to experienced radiologists [19]. Zhang et al. developed a two-dimensional grayscale ultrasound model that showed potential for differentiating benign from malignant lesions and reducing unnecessary biopsies of BI-RADS 4A lesions [20]. Ciritsis et al. demonstrated that a deep convolutional neural network could classify breast ultrasound lesions according to BI-RADS and potentially standardize observer-dependent assessment [21]. In clinical decision support, Wang et al. reported that adding AI to breast ultrasound could reduce unnecessary biopsies of BI-RADS 4A lesions [22], whereas Amir et al. showed that AI decision support may be particularly valuable for small invasive lobular carcinomas [23]. Nevertheless, multimodal information and clinical context remain important because image-only AI does not invariably outperform expert physicians [24]. Therefore, AI should currently be regarded as an adjunct to, rather than a replacement for, physician interpretation, and further prospective multicenter validation is still required [25].

4.6. Limitations

There are multiple limitations acknowledged in this study. First, there is intrinsic selection bias as this is a retrospective single centre study at a cancer hospital. All 4 nodules identified in this study as BI-RADS 4 were first evaluated by different levels of physicians and the absolute rates of diagnostic performance may not be generalizable to community screening populations, as the prevalence of malignancy may be higher among the patient population at a tertiary cancer center. Second, in this study, the AI+US combined group consisted entirely of junior physicians combined with AI; no data on the performance of senior physicians combined with AI were collected, which limits the characterization of the entire spectrum of effects of the interaction between physicians and AI. Third, given a sample size of 231 nodules, the statistical power for the comparisons within the subgroups was restricted, especially in the <15 mm group (N = 58) for which the confidence intervals were wide, as shown in **Figure 5**. Fourth, two different ultrasound machines were used, and the differences in image characteristics, resolution and artefact profile from these different machines may have resulted in systematic variability in the AI performance that could not be quantified in this study design. Fifth, only the manufacturer's own external validation data were available; independent external validation by other institutions has not been reported. This raises concern about the model's generalizability and the possibility of overfitting. Sixth, the detailed sonographic features (margin characteristics, echo texture parameters, posterior acoustic feature quantification, and vascularity indices) were not analysed quantitatively, which prevented insight into the specific image features that best contributed to the AI performance benefits. Incorporating systematic radiomics analysis in future studies could significantly improve the understanding of the modes of failure of AI and the patterns of decision-making.

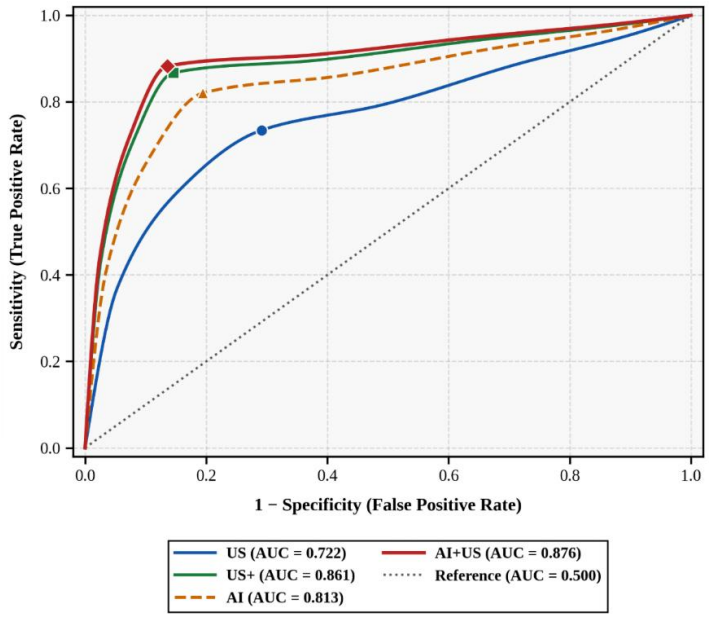


Figure 5. Receiver operating characteristic curves of the four diagnostic methods (US, US+, AI, and AI+US) for breast nodules.

5. Conclusion

The AI ultrasound-assisted diagnostic system is shown to have good diagnostic efficacy for BI-RADS 4 breast nodules, with performance comparable to intermediate-level physicians, and statistical confirmation of its benefits in the detection of small nodules (<15 mm; AUC comparison AI+US vs. US: $p = 0.048$). The performance of the combination of the junior physician conventional ultrasound and AI-assisted-diagnosis is significantly better than that of the junior physicians alone (AUC 0.876 vs. 0.722, $p < 0.05$) and is comparable to that of the senior physicians (AUC 0.861, $p = 0.372$). AI computational image analysis and physician contextualization of these images are complementary technologies that when combined provide an optimal collaborative diagnostic model to overcome the limitations of each method. These results argue for the clinical use of AI-assisted ultrasound imaging systems,

especially in clinical scenarios with a shortage of senior physician expertise and for small breast nodules, where conventional ultrasound imaging is least capable.

This study has made some important contributions:

For BI-RADS 4 nodules, the TenD.AI system was not only more effective than junior physicians (AUC 0.813 vs. 0.722, $p = 0.032$) but also non-inferior to senior physicians (AUC 0.861, $p = 0.218$).

Found that the AUC for AI+US collaboration was significantly superior to that of junior physicians alone (0.876 vs. 0.722, $p = 0.002$) and was not significantly different from the AUC of senior physicians ($p = 0.372$), essentially closing the experience gap.

This is the first statistically significant confirmation of AI+US superiority over junior physicians in the challenging subgroup of small nodules (<15 mm; $Z = 1.98$, $p = 0.048$), supporting AI integration specifically for this group.

Through representative case analysis, they showed how AI image recognition and physician clinical integration can complement each other, with AI showing a benefit in detecting small lesions, and physician showing a benefit in clinical interpretation.

Recognized that the use of AI can be most beneficial for junior physicians and its direct impact on AI deployment in primary care, community hospitals and in resource-limited healthcare settings.

Author Contributions

L.H.: conceptualization, data curation, writing—original draft. J.W.: ultrasound examination, data collection. L.L.: ultrasound examination, data collection. X.Z.: radiology review. L.F.: pathological diagnosis confirmation. X.H.: surgical coordination and pathology liaison. W.L.: data collection and quality control. D.K.: methodology, critical review. B.-A.B.: critical review and validation. S.A.: supervision, project administration, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement

This study was approved by the Ethics Committee of Baotou Cancer Hospital [2025(024)] and the Ethics Committee of the Mongolian National University of Medical Sciences [2025.09.26 25-26/01-02]. The study was conducted in accordance with the Declaration of Helsinki.

Informed Consent Statement

Written informed consent was obtained from all individual participants before the ultrasound examination and data collection.

Data Availability Statement

The datasets generated and analyzed during the current study are not publicly available due to patient privacy considerations but are available from the corresponding author upon reasonable request and subject to institutional data sharing agreements.

Conflicts of Interest

The authors declare no conflict of interest.

AI Use Statement

During the preparation of this work, the authors used ChatGPT (OpenAI) for English-language refinement, grammatical checking, and improvement of non-native English expressions. The authors subsequently reviewed and edited the content as necessary and take full responsibility for the final content of the published article. ChatGPT

was not used for study design, data collection, statistical analysis, interpretation of the results, or generation of scientific conclusions.

References

1. Siegel, R.L.; Miller, K.D.; Wagle, N.S.; et al. Cancer statistics, 2023. *CA Cancer J. Clin.* **2023**, *73*, 17–48.
2. Burstein, H.J.; Curigliano, G.; Thürlimann, B.; et al. Customizing local and systemic therapies for women with early breast cancer: The St. Gallen International Consensus Guidelines for treatment of early breast cancer 2021. *Ann. Oncol.* **2021**, *32*, 1216–1235.
3. Lee, J. Practical and illustrated summary of updated BI-RADS for ultrasonography. *Ultrasonography* **2016**, *36*, 71–81.
4. Zhao, J.Q.; Diao, Z.; Xu, Q.; et al. New development of ultrasound medicine in the era of artificial intelligence. *Acad. J. Second Mil. Med. Univ.* **2019**, *40*, 478–482. (in Chinese)
5. Kim, K.; Song, M.K.; Kim, E.K.; et al. Clinical application of S-Detect to breast masses on ultrasonography: A study evaluating the diagnostic performance and agreement with a dedicated breast radiologist. *Ultrasonography* **2017**, *36*, 3–9.
6. Choi, J.H.; Kang, B.J.; Baek, J.E.; et al. Application of computer-aided diagnosis in breast ultrasound interpretation: Improvements in diagnostic performance according to reader experience. *Ultrasonography* **2018**, *37*, 217–225.
7. Wei, Q.; Yan, Y.J.; Wu, G.G.; et al. The diagnostic performance of ultrasound computer-aided diagnosis system for distinguishing breast masses: A prospective multicenter study. *Eur. Radiol.* **2022**, *32*, 4046–4055. [CrossRef]
8. Lv, S.; Zhang, Y.; Zhang, M.; et al. The auxiliary role of an artificial intelligence-based automatic detection system in the diagnosis of small breast lumps by physicians with different experience. *Chin. J. Med. Ultrasound (Electronic Edition)* **2022**, *19*, 983. (in Chinese)
9. Xing, B.; Gu, C.; Fu, C.; et al. Diagnostic performance of ultrasound S-Detect technology in evaluating BI-RADS-4 breast nodules ≤20 mm and >20 mm. *BMC Cancer* **2025**, *25*, 1306. [CrossRef]
10. Hejduk, P.; Marcon, M.; Unkelbach, J.; et al. Fully automatic classification of automated breast ultrasound imaging according to BI-RADS using a deep convolutional neural network. *Eur. Radiol.* **2022**, *32*, 4868–4878. [CrossRef]
11. Islam, T.; Hoque, M.E.; Ullah, M.; et al. CNN-based deep learning approach for classification of invasive ductal and metastasis types of breast carcinoma. *Cancer Med.* **2024**, *13*, e70069. [CrossRef]
12. Geras, K.J.; Mann, R.M.; Moy, L. Artificial intelligence for mammography and digital breast tomosynthesis: Current concepts and future perspectives. *Radiology* **2019**, *293*, 246–259. [CrossRef]
13. Tsuyuzaki, S.; Fujioka, T.; Yamaga, E.; et al. Evaluation of AI diagnostic systems for breast ultrasound: Comparative analysis with radiologists and the effect of AI assistance. *Jpn. J. Radiol.* **2025**, *43*, 1643–1651. [CrossRef]
14. Chen, R.; Wu, S.; Guo, J.; et al. Application value of an ultrasound artificial intelligence-assisted diagnostic system in the diagnosis of BI-RADS 4 breast nodules with a maximum diameter ≤2 cm. *J. Nav. Med. Univ.* **2024**, *45*, 592–598. (in Chinese)
15. Xing, B.; Chen, X.; Wang, Y.; et al. Evaluating breast ultrasound S-Detect image analysis for small focal breast lesions. *Front. Oncol.* **2022**, *12*, 1030624. [CrossRef]
16. Fotouhi, S.; Asadi, S.; Kattan, M.W. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *J. Biomed. Inform.* **2019**, *90*, 103089. [CrossRef]
17. Shen, Y.; Shamout, F.E.; Oliver, J.R.; et al. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. *Nat. Commun.* **2021**, *12*, 5645. [CrossRef]
18. O'Connell, A.M.; Bartolotta, T.V.; Orlando, A.; et al. Diagnostic performance of an artificial intelligence system in breast ultrasound. *J. Ultrasound Med.* **2022**, *41*, 97–105. [CrossRef]
19. Qian, X.; Pei, J.; Zheng, H.; et al. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. *Nat. Biomed. Eng.* **2021**, *5*, 522–532. [CrossRef]
20. Zhang, N.; Li, X.T.; Ma, L.; et al. Application of deep learning to establish a diagnostic model of breast lesions using two-dimensional grayscale ultrasound imaging. *Clin. Imaging* **2021**, *79*, 56–63. [CrossRef]
21. Ciritsis, A.; Rossi, C.; Eberhard, M.; et al. Automatic classification of ultrasound breast lesions using a deep convolutional neural network mimicking human decision-making. *Eur. Radiol.* **2019**, *29*, 5458–5468. [CrossRef]
22. Wang, X.Y.; Cui, L.G.; Feng, J.; et al. Artificial intelligence for breast ultrasound: An adjunct tool to reduce

excessive lesion biopsy. *Eur. J. Radiol.* **2021**, *138*, 109624. [CrossRef]
23. Amir, T.; Coffey, K.; Sevilimedu, V.; et al. A role for breast ultrasound artificial intelligence decision support in the evaluation of small invasive lobular carcinomas. *Clin. Imaging* **2023**, *101*, 77–85. [CrossRef]
24. Pfob, A.; Sidey-Gibbons, C.; Barr, R.G.; et al. The importance of multi-modal imaging and clinical information for humans and AI-based algorithms to classify breast masses (INSPiRED 003): An international, multicenter analysis. *Eur. Radiol.* **2022**, *32*, 4101–4115. [CrossRef]
25. Wu, G.G.; Zhou, L.Q.; Xu, J.W.; et al. Artificial intelligence in breast ultrasound. *World J. Radiol.* **2019**, *11*, 19–26. [CrossRef]

Copyright © 2026 by the author(s). Published by UK Scientific Publishing Limited. This is an open access article under the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Publisher's Note: The views, opinions, and information presented in all publications are the sole responsibility of the respective authors and contributors, and do not necessarily reflect the views of UK Scientific Publishing Limited and/or its editors. UK Scientific Publishing Limited and/or its editors hereby disclaim any liability for any harm or damage to individuals or property arising from the implementation of ideas, methods, instructions, or products mentioned in the content.