Original Article

Deep learning powered breast ultrasound to improve characterization of breast masses: a prospective study

Acta Radiologica
1–11

DOI: 10.1177/02841851251377927
journals.sagepub.com/home/acr

Veenu Singla[1,*], **Dollphy Garg**[1,*], **Sapna Negi**[1], **Nandita Mehta**[1], **T Pallavi**[1], **Sonam Choudhary**[1] and **Abhik Dhiman**[1]

Abstract

Background: The diagnostic performance of ultrasound (US) is heavily reliant on the operator's expertise. Advances in artificial intelligence (AI) have introduced deep learning (DL) tools that detect morphology beyond human perception, providing automated interpretations.

Purpose: To evaluate Smart-Detect (S-Detect), a DL tool, for its potential to enhance diagnostic precision and standardize US assessments among radiologists with varying levels of experience.

Material and Methods: This prospective observational study was conducted between May and November 2024. US and S-Detect analyses were performed by a breast imaging fellow. Images were independently analyzed by five radiologists with varying experience in breast imaging (<1 year–15 years). Each radiologist assessed the images twice: without and with S-Detect. ROC analyses compared the diagnostic performance. True downgrades and upgrades were calculated to determine the biopsy reduction with AI assistance. Kappa statistics assessed radiologist agreement before and after incorporating S-Detect.

Results: This study analyzed 230 breast masses from 216 patients. S-Detect demonstrated high specificity (92.7%), PPV (92.9%), NPV (87.9%), and accuracy (90.4%). It enhanced less experienced radiologists' performance, increasing the sensitivity (85% to 93.33%), specificity (54.5% to 73.64%), and accuracy (70.43% to 83.91%; P <0.001). AUC significantly increased for the less experienced radiologists (0.698 to 0.835 P <0.001), with no significant gains for the expert radiologist. It also reduced variability in assessment between radiologists with an increase in kappa agreement (0.459–0.696) and enabled significant downgrades, reducing unnecessary biopsies.

Conclusion: The DL tool improves diagnostic accuracy, bridges the expertise gap, reduces reliance on invasive procedures, and enhances consistency in clinical decisions among radiologists.

Keywords
Artificial intelligence, deep learning, S-Detect, ultrasound, malignant, breast mass

Date received: 10 March 2025; accepted: 25 August 2025

Introduction

Breast cancer continues to be one of the most common cancers impacting women worldwide, with its incidence and associated mortality showing a concerning upward trend (1). Early detection through screening, accurate diagnosis, and prompt treatment are critical strategies to reduce breast cancer-related fatalities (2,3).

Ultrasound (US) has emerged as a cornerstone in breast cancer diagnosis due to its real-time imaging capability, affordability, lack of radiation, and accessibility (4,5). However, its reproducibility and diagnostic efficacy are heavily reliant on the operator's expertise and acumen (5–7). The introduction of the Breast Imaging Reporting and Data System (BI-RADS) by the American College of Radiology (ACR) sought to systematize the evaluation of breast masses (8). Despite its adoption, variability in interpretation—both among different observers and even within the same observer—remains a significant challenge.

[1]Department of Radiodiagnosis, PGIMER, Chandigarh, India

[*]Equal contributors.

Corresponding author:
Veenu Singla, Professor, Department of Radiodiagnosis, PGIMER, Chandigarh 160012, India.
Email: veenupgi@gmail.com

A wide range of variability from 3% to 94% in estimated malignancy risk for BI-RADS category 4 further contributes to diagnostic inconsistencies, reduced specificity and positive predictive value (PPV). This significantly heightens the reliance on invasive diagnostic procedures that can escalate both patient distress and financial burden (9,10).

Breakthroughs in artificial intelligence (AI) have led to deep learning (DL)-based models capable of detecting morphological features beyond human perception (11). S-Detect (Samsung Medison, Seoul, Republic of Korea) is a commercially obtainable DL- based software that facilitates automatic image segmentation and interpretation of US features of breast masses as per BI-RADS descriptors and gives a dichotomous result of either "possibly benign" or "possibly malignant" (12,13). The US features analyzed by it include orientation, shape, texture, and echopattern; assessed using spatial gray-level dependence matrices, intensity, and gradient magnitude within the mass. Additional features include average gray-level differences between the mass and the surrounding tissue, count of lobulations, and the elliptic-normalized circumference (14,15).

Prior studies have explored S-Detect's role in distinguishing benign from malignant breast masses, with inconsistent results highlighting the need for further research to validate its reliability and effectiveness in clinical use (14–22). The current study had two distinct objectives: first, to assess the diagnostic accuracy and reliability of S-Detect in characterizing breast masses when integrated into the US workflow; and second, to investigate its influence across radiologists with varying levels of experience, particularly analyzing its concordance with expert radiologists. This study also showcases the impact of AI-driven tools in reducing the need for diagnostic interventions and minimizing inter-observer variability, thereby offering an evaluation of its practical value in clinical settings.

Material and Methods

This prospective observational study was conducted in the Department of Radiodiagnosis of a tertiary medical facility between May and November 2024, after approval from the institutional ethics committee.

Inclusion criteria

The study population consisted of patients aged over 18 years having suspicious breast masses detected on US or mammography. In addition, patients with breast masses who were under follow-up for ≥2 years or had previously undergone a biopsy (biopsy done >6 months prior) yielding benign histopathology results were included.

Exclusion criteria

The study excluded pregnant or lactating women, patients undergoing neoadjuvant chemotherapy or radiotherapy, or those who completed these treatments within the past 3 months, large breast masses (>5 cm), and non-mass findings unsuitable for S-Detect assessment. Patients with inconclusive histopathology reports or B3 lesions with uncertain malignant potential were also excluded from the study. B3 lesions were excluded due to their uncertain malignant potential, which could confound classification performance, especially when no definitive surgical outcome is available. However, women at higher risk were not excluded.

A total of 568 patients underwent US examination during the study period. After exclusion, a total of 216 patients with 230 masses were finally included in the study. All the patients underwent US examinations using the Samsung V7 system (Samsung Healthcare, Republic of Korea) with a 2–14 MHz linear probe, including analysis with S-Detect software.

Study protocol and image analysis

To ensure proficiency with the S-Detect software, a training session was conducted for all participating radiologists. The research was carried out in two distinct phases. The first phase involved the on-site prospective acquisition of grayscale US and S-Detect images performed by a breast imaging fellow. The second phase consisted of an off-site evaluation of the images by five independent readers with varying levels of experience in breast radiology.

The on-site breast imaging fellow acquired cine and static grayscale US images of the breast masses. The static images were recorded in two orthogonal planes. Thereafter, he also performed the analysis of masses using the S-Detect tool. After selecting the appropriate image, the radiologist activated the software to initiate segmentation, where the mass interface was automatically detected by the software; however, it could be manually adjusted. Upon completion of the analysis, the software generated morphological descriptors of the mass in accordance with ACR BI-RADS fifth edition (shape, margin, orientation, pattern, and posterior acoustic features), along with a structured report containing the final results as "possibly benign" or "possibly malignant." In cases where the results differed between the two planes, the final outcome was considered "possibly malignant." These interpretations were systematically recorded. The patients' final histopathology results were also recorded by the on-site fellow.

Subsequently, five readers with varying experience in breast performed off-site independent interpretation of the images. Radiologist 1 (R1) was an "expert" with >15 years of dedicated breast imaging. Radiologist 2 (R2) had "moderate experience" of 3 years in breast imaging. Radiologists 3, 4, and 5 (R3, R4, and R5) were general radiologists having "less experience," with 9 months, 6 months, and 3 months of dedicated breast imaging practice, respectively. In the first session, BI-RADS assessments were based on grayscale images. A follow-up image reading session

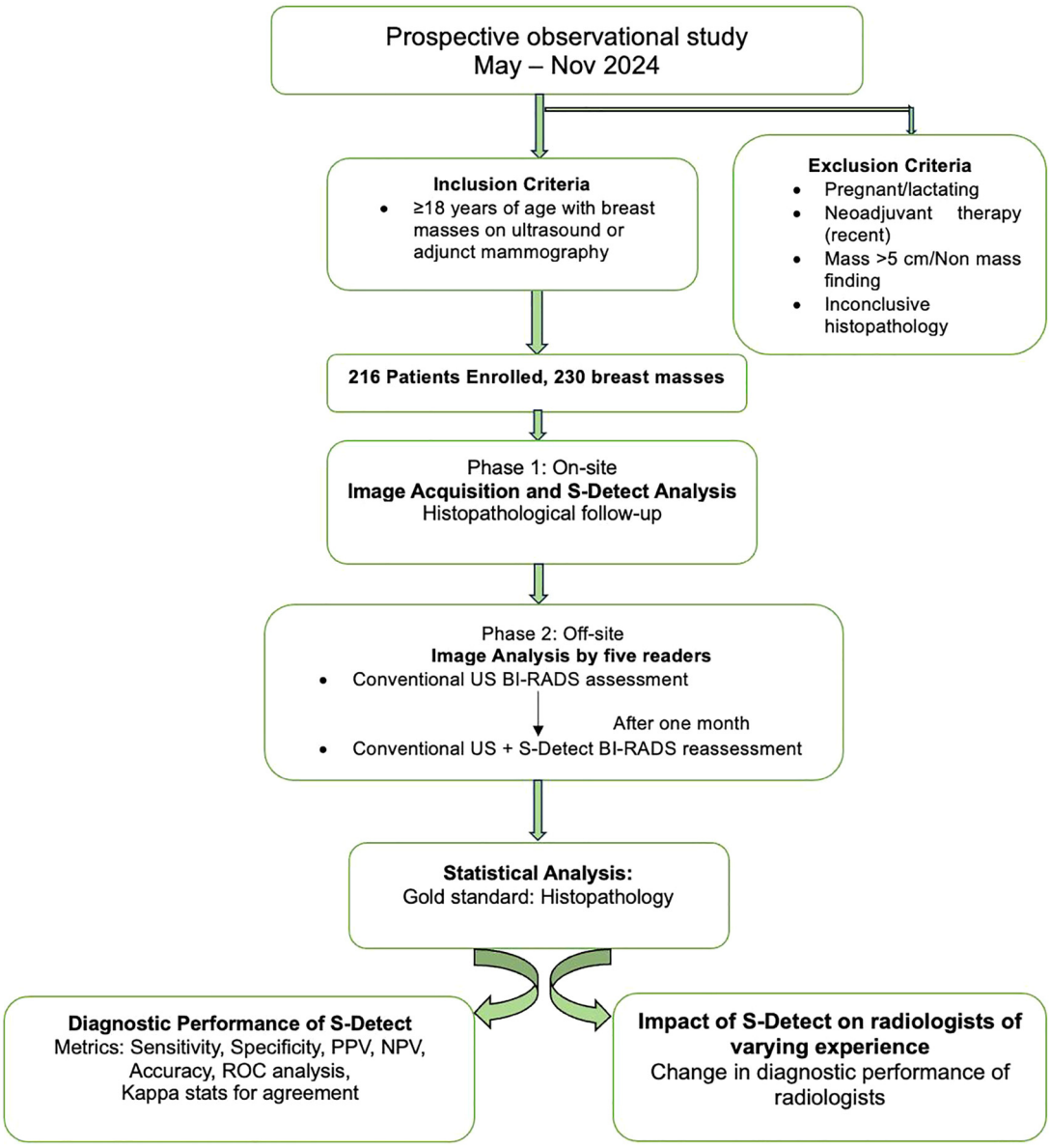

Fig. 1. Flow chart depicting the study protocol.

was conducted at a 1-month interval to reduce the recall bias, during which the masses were reassessed, and final interpretation was given after consulting the S-Detect results. All the readers were blinded to the histopathology results to circumvent bias. Still, they were provided with relevant clinical details (age, personal or family history of breast cancer, and chief complaints) to simulate a real-world clinical setting. Reader responses were recorded in an Excel spreadsheet and were subsequently used for analysis (Fig. 1).

Statistical analysis

The gold standard for diagnosis was final histopathology. However, masses with documented evidence of lesion stability or a reduction in size over more than 2 years were considered benign. For statistical analysis, categories 2 and 3 of BI-RADS were labelled benign, whereas categories 4A through 5 were deemed malignant. The data analysis was carried out by a biostatistician utilizing SPSS version 22.0 (IBM Corp., Armonk, NY, USA). Statistical metrics were calculated to evaluate diagnostic performance, including sensitivity, specificity, PPV, negative predictive value (NPV), receiver operating characteristic (ROC) curve, and the area under the ROC curve (AUC). The significance of changes associated with the addition of the S-Detect was analyzed with paired comparison of ROC analysis using Z statistics. The level of agreement between the radiologists and S-Detect was measured using Cohen's Kappa (κ)

Table 1. Histopathological diagnosis of the 230 breast masses.

Tumor	n (%)
Benign	
Fibroadenoma	81 (35.2)
Sclerosis adenosis	6 (2.6)
Breast abscess	5 (2.2)
Benign phyllodes	5 (2.2)
Galactocele	3 (1.3)
Intraductal papilloma without atypia	2 (0.9)
Intramammary lymph node	2 (0.9)
Usual ductal hyperplasia	2 (0.9)
Benign biphasic tumor	1 (0.4)
Granuloma	1 (0.4)
Fat necrosis	1 (0.4)
Microcyst cluster	1 (0.4)
Malignant	
Invasive ductal carcinoma no specific type	101 (43.9)
Invasive lobular carcinoma	5 (2.2)
Papillary carcinoma	4 (1.7)
Medullary carcinoma	3 (1.3)
DCIS	2 (0.9)
Metaplastic carcinoma	2 (0.9)
Mucinous carcinoma	2 (0.9)
Adenoid cystic carcinoma	1 (0.4)

DCIS, ductal carcinoma in situ.

statistics as follows: ≤0.20 = slight agreement; 0.21–0.40 = fair agreement; 0.41–0.60 = moderate agreement; 0.61–0.80 = substantial agreement; and those 0.81–1.00 = perfect agreement. Significance was defined as P values <0.05.

Results

Patient demographic and breast mass attributes

A total of 230 breast masses, including 120 malignant masses and 110 benign masses, from 216 patients (3 men, 213 women; mean age = 49.35 ± 16.6 years; age range = 18–78 years) were consecutively enrolled in the study. The mean size of the masses was 18.2 mm ± 8.3 (range = 0.54–4.8 cm). Of the 230 masses evaluated, 209 were confirmed through histopathological examination, while the remaining 21 were classified as benign based on stability in their size over a previous 2-year period. The histopathological distribution is given in Table 1. Benign masses accounted for 47.8% of the cases, while malignant masses comprised 52.2%. Among the benign masses, fibroadenomas were the most common, accounting for 35.2%, while among the malignant masses, the majority were invasive ductal carcinomas of no specific type (IDC NST), comprising 43.9%. Among 230 breast lesions, 217 (94.3%) exhibited uniform S-Detect outcomes across both the orthogonal planes, while 13 (5.6%) had discordant outcomes.

Table 2. Comparison of diagnostic performance of radiologists 1–5 (R1–R5) in decreasing order of expertise on conventional ultrasound and after incorporating S-Detect interpretation showing improvement.

	R1		R2		R3		R4		R5	
Statistics	Conventional	S- Detect (*P*)	Conventional	S- Detect (*P*)	Conventional	S- Detect (*P*)	Conventional	S- Detect (*P*)	Conventional	S- Detect (*P*)
Sensitivity (%)	94.17	96.67 (0.250)	93.33	96.61 (0.125)	91.67	95.83 (0.125)	89.17	92.50 (0.344)	85.0	93.33 (0.006)
Specificity (%)	86.36	88.18 (0.687)	81.82	85.45 (0.289)	70.00	81.82 (<0.001)	57.27	71.82 (<0.001)	54.5	73.64 (<0.001)
PPV (%)	88.28	89.92	84.85	87.88	76.92	85.19	69.48	78.17	66.6	79.43
NPV (%)	93.1%	96.04	91.84	95.92	88.51	94.74	82.89	89.77	76.9	91.01
Accuracy (%)	90.43	92.61 (0.103)	87.83	91.30 (0.022)	81.30	89.13 (<0.001)	73.91	82.61 (<0.001)	70.4	83.91 (<0.001)
AUC (95% CI)	0.903	0.924	0.876	0.911	0.808	0.888	0.732	0.822	0.698	0.835
	(0.857–0.938)	(0.882–0.955)	(0.826–0.915)	(0.866–0.944)	(0.751–0.857)	(0.840–0.926)	(0.670–0.788)	(0.766–0.869)	(0.634–0.756)	(0.780–0.880)

AUC, area under the receiver operating characteristic curve; CI, confidence interval; NPV, negative predictive value; PPV, positive predictive value.

Diagnostic capability of S-Detect and the radiologists with varying expertise

Standalone S-Detect demonstrated high diagnostic performance with sensitivity, specificity, PPV, NPV, and accuracy of 88.3%, 92.7%, 92.9%, 87.9%, and 90.4%, respectively. Table 2 presents the diagnostic metrics achieved by the radiologists, listed in decreasing order of experience in breast imaging from R1 to R5. When comparing the initial grayscale assessments, S-Detect significantly outperformed radiologists R3, R4, and R5 in diagnostic accuracy, though the differences were not significant from R1 and R2. Substantial agreement was observed between S-Detect and radiologists R1 and R2 (kappa values = 0.792 and 0.687, respectively). However, agreement with the less experienced radiologists (R3, R4, and R5) was more inconsistent, with kappa values of 0.644, 0.514, and 0.367, respectively (Table 3).

Impact of S-Detect integration on diagnostic performance across radiologists with varying expertise

Table 2 illustrates the impact of incorporating S-Detect results into the final assessment on diagnostic performance. Upon reassessment, novice radiologists (R3, R4, and R5) made significantly more adjustments than more experienced R1 and R2. While the inclusion of S-Detect resulted in an insignificant improvement in the diagnostic accuracy of R1 ($P > 0.05$), the performance of R2 showed a moderately significant improvement ($P = 0.022$). In contrast, the less experienced radiologists (R3, R4, and R5) showed highly significant improvements ($P < 0.01$) in diagnostic performance after integrated reassessment. ROC analysis demonstrated a clear improvement in the diagnostic performance of R3–R5 after the inclusion of S-Detect, evident from the significant increase in AUC (Fig. 2).

Table 3. Comparative agreement of assessment of breast masses on conventional ultrasound (B-mode) by radiologists 1–5 (R1–R5) with S-detect.

Radiologist	Kappa
R1	0.792
R2	0.687
R3	0.644
R4	0.514
R5	0.376

Table 4. Comparative agreement of radiologists 2–5 (R2–R5) with the expert radiologist R1 on conventional (B-mode) ultrasound and combined (B-mode and S-detect) assessment.

Radiologist	Conventional Ultrasound assessment (Kappa)	Reassessment (conventional ultrasound + S- Detect) (Kappa)
R2	0.717	0.832
R3	0.723	0.858
R4	0.584	0.740
R5	0.459	0.696

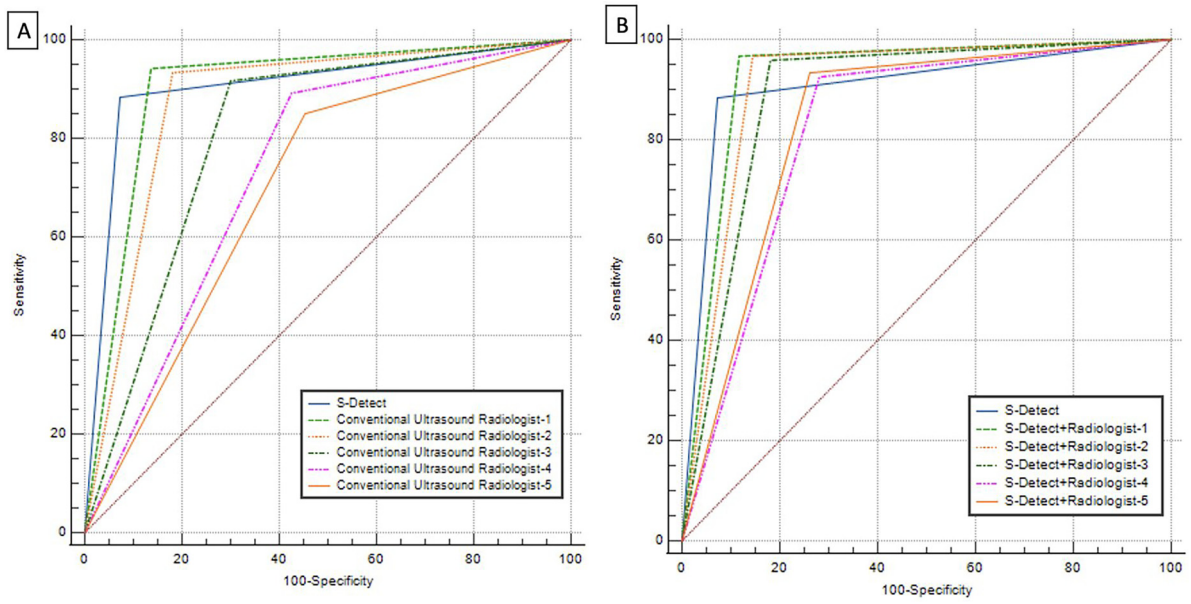

Fig. 2. Comparative ROC analysis. (a) AUC showing the diagnostic performance of radiologists using B-mode ultrasound. (b) AUC shows the diagnostic performance of radiologists using B-mode ultrasound combined with S-Detect. The upward shift in the ROC curve for all radiologists indicates an improvement in diagnostic performance when S-Detect is incorporated. This AUC enhancement was statistically significant for radiologists 2–5 ($P < 0.05$), demonstrating the added value of S-Detect in improving diagnostic accuracy. AUC, area under the curve; ROC, receiver operating characteristic

Table 5. Changes made by radiologists 1–5 (R1–R5) on reassessment after S-detect result analysis.

		HPE final report	
		Malignant	Benign
R1	Upgrade	3 (60.0)	2 (40.0)
	Downgrade	0 (0.0)	4 (100.0)
R2	Upgrade	4 (66.7)	2 (33.3)
	Downgrade	0 (0.0)	6 (100.0)
R3	Upgrade	6 (100.0)	0 (0.0)
	Downgrade	1 (7.14)	13 (92.9)
R4	Upgrade	7 (87.5)	1 (12.5)
	Downgrade	2 (10.5)	17 (89.5)
R5	Upgrade	11 (84.6)	2 (15.4)
	Downgrade	1 (4.2)	23 (95.8)

Values are given as n (%).
Upgrade: radiologist changed the final assessments from BI-RADS 2/3 to BI-RADS 4A–5 after S-Detect combination; downgrade: radiologist changed the final assessments from BI-RADS 4A–5 to BI-RADS 3/2 after S-Detect combination.
HPE, histopathological examination.

Notably, for the least experienced R5, all key metrics demonstrated substantial improvement: sensitivity increased from 85% to 93.33%, specificity rose from 54.5% to 73.6%, PPV improved from 67.11% to 79.4%, NPV grew from 76.9% to 91%, and overall accuracy advanced from 70% to 83.9% ($P<0.001$).

In addition, the use of S-Detect significantly improved the agreement between R1 and the less experienced radiologists (R3, R4, and R5), reducing inter-observer variability (Table 4).

Comparison of S-Detect adjusted final assessments with pathological findings

The role of S-Detect in improving diagnostic accuracy and reducing unnecessary interventions is illustrated in Table 5. With the integration of AI, radiologists accurately downgraded lesions from BIRADS 4A–5 to BIRADS 3/2, which helps minimize unnecessary biopsies. The total correct downgrades (BIRADS 4A–5 to BIRADS 3/2) achieved by radiologists R1,

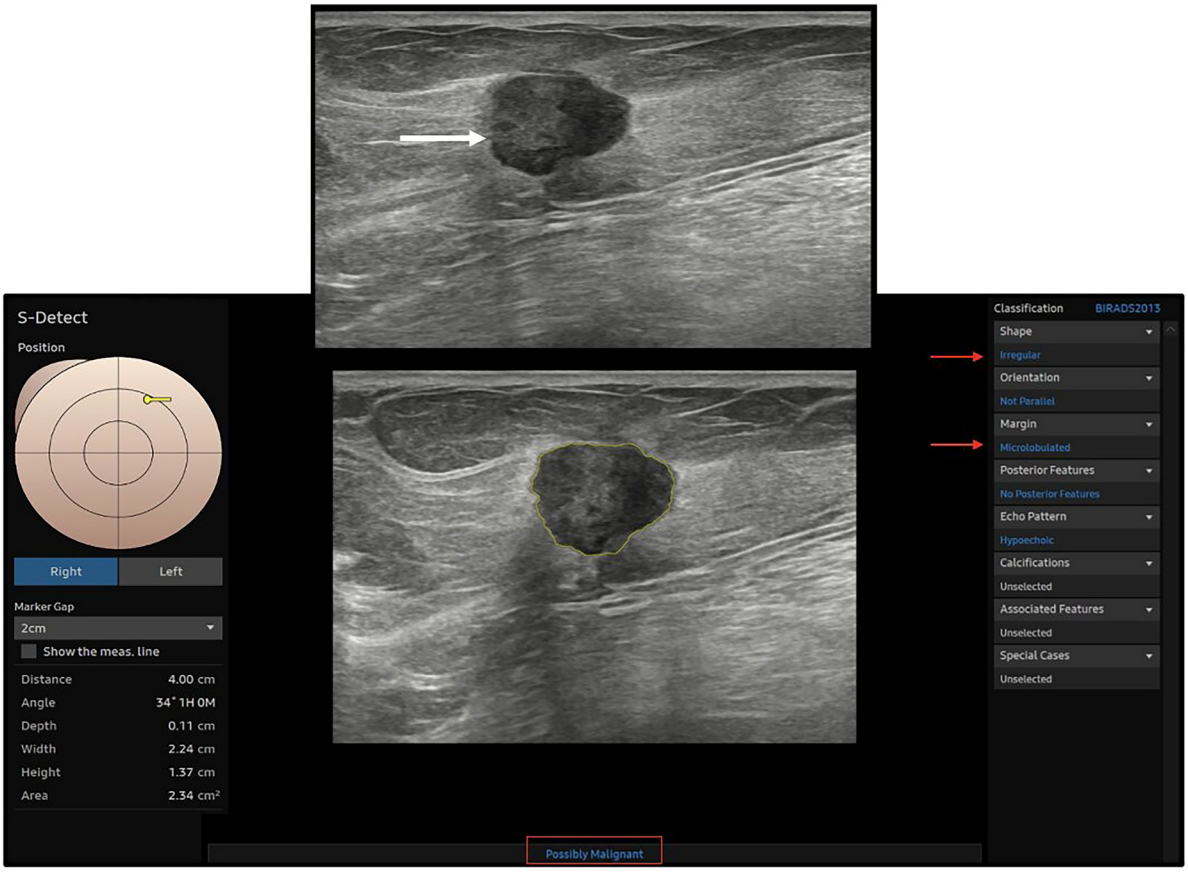

Fig. 3. B-mode ultrasound of a 38-year-old woman revealed a 13 × 22 mm hypoechoic mass (white arrow) in the right breast. The BI-RADS categories assigned on conventional ultrasound by radiologists R1-R5 were 4A, 3, 3, 3, and 3, respectively. The margins were initially assessed as circumscribed by the majority of the radiologists. With S-Detect (yellow), the margins were read as microlobulated, and the mass was interpreted as "possibly malignant." The mass was reclassified by R1–R5 as BI-RADS 4B, 4A, 4B, 4B, and 4A, respectively, on combined reassessment with S-Detect. Core-needle biopsy confirmed invasive ductal carcinoma, Grade 2, ER-, PR-, HER2-, concordant with the S-Detect results, hence a true upgrade.

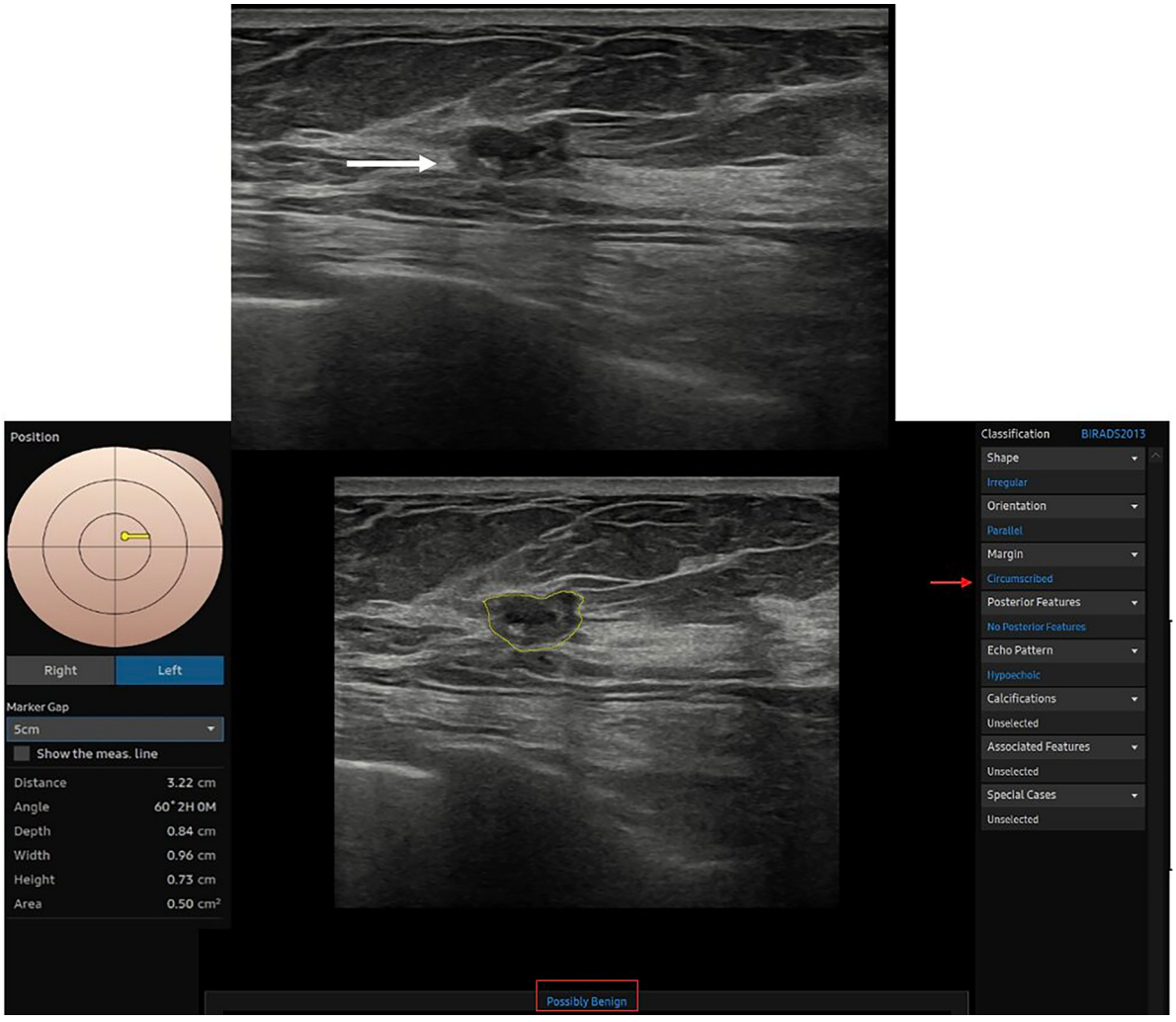

Fig. 4. A 28-year-old woman with mastalgia in the left breast and a positive family history of breast cancer (mother) presented for evaluation. B-mode ultrasound revealed an 8 × 7 mm hypoechoic mass (white arrow) in the left breast. The BI-RADS categories assigned on conventional ultrasound by radiologists R1–R5 were 4A, 4B, 4A, 4A, and 4C, respectively. With S-Detect (yellow), the margins were defined as circumscribed, and the mass was interpreted as "possibly benign." The mass was reclassified by R1–R5 as BI-RADS 3, 3, 3, 3, and 3, respectively, on combined reassessment with S-Detect. Core-needle biopsy confirmed a benign fibroepithelial lesion, concordant with the S-Detect results, leading to a true downgrade.

R2, R3, R4, and R5 were 4 (100%), 6 (100%), 13 (92.9%), 17 (89.5%), and 23 (95.8%), respectively. The highest number of correct downgrades were observed with R3 to R5. Furthermore, the correct upgrades, where lesions were reclassified from BIRADS 2/3 to BIRADS 4A–5, were as follows: R1 (3, 60%), R2 (4, 66.7%), R3 (6, 100%), R4 (7, 87.5%), and R5 (11, 86.6%). The inclusion of S-Detect contributed to an increase in the cancer detection rate, particularly for the less experienced radiologists, who demonstrated notable improvements in accurately upgrading masses to higher BI-RADS categories (Figs. 3–6).

Discussion

US remains a vital tool in breast imaging, widely utilized for distinguishing between benign and malignant masses, apart from being an adjunct to mammography (1–4). US serves as the primary imaging modality in resource-limited settings with populations with dense breasts and younger age of peak breast cancer incidence rate. The development of BI-RADS by the American College of Radiology brought much-needed consistency and standardization to the characterization and classification of breast lesions. However, the inherent overlap in imaging features of benign and suspicious masses, coupled with varying levels of expertise among radiologists, continues to challenge the accuracy and reproducibility of US diagnoses. To address these limitations, innovative AI-driven algorithms are being increasingly integrated into modern US systems (5–10). S-Detect is a commercially available software utilizing DL and neural networks trained on vast datasets that independently evaluates the morphological

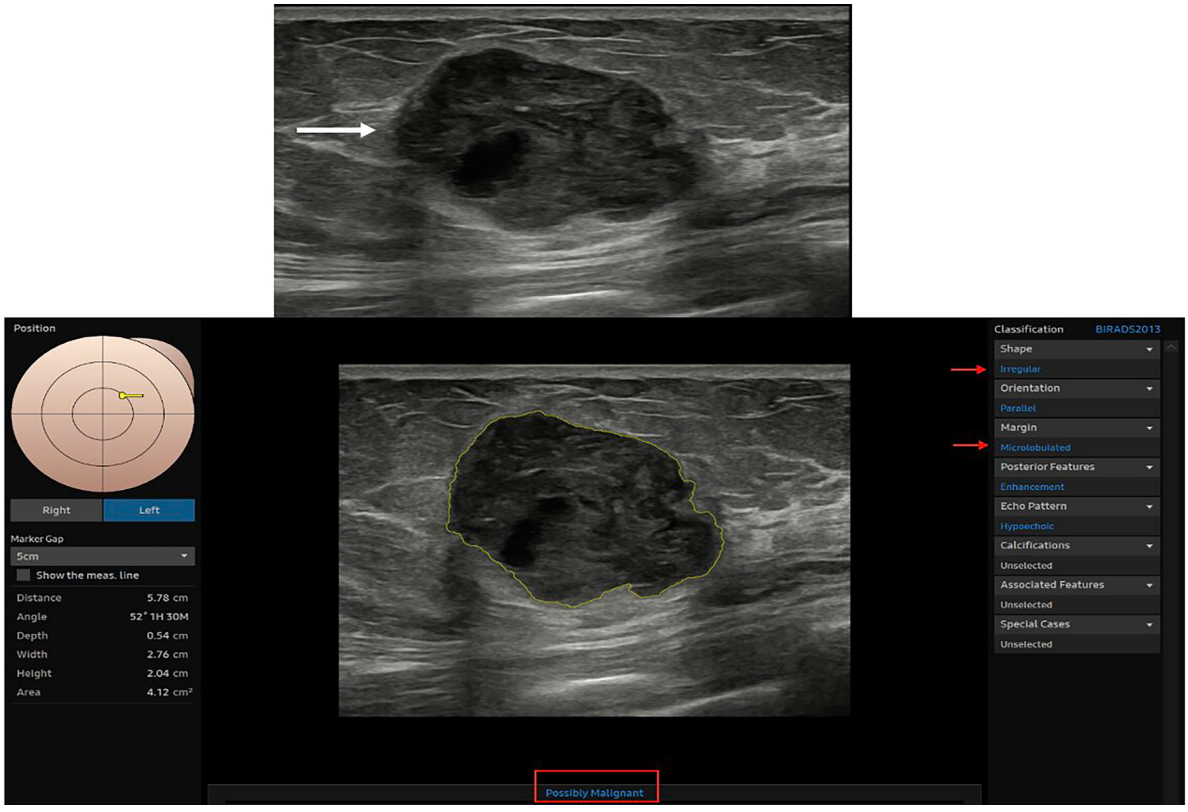

Fig. 5. A 21-year-old woman with a palpable lump in the left breast presented for evaluation. B-mode ultrasound revealed a 27 × 20 mm hypoechoic mass (white arrow) in the left breast. The BI-RADS categories assigned on conventional ultrasound by radiologists R1–R5 were 4B, 3, 4A, 3, and 4A, respectively. With S-Detect (yellow), the margins were read as microlobulated, and the mass was interpreted as "possibly malignant." The mass was reclassified by R1-R5 as BI-RADS 4B, 4A, 4B, 4A, and 4A, respectively, on combined reassessment with S-Detect. Core-needle biopsy confirmed invasive ductal carcinoma, Grade II, ER positive, PR positive, concordant with the S-Detect results, leading to a true upgrade.

features of breast masses using the BI-RADS descriptors and offers an automated, dichotomous classification: "possibly benign" or "possibly malignant" (11–13).

In our study, S-Detect showed high specificity (92.7%) and PPV (92.9%), outperforming radiologists R3, R4, and R5. A study by Cho et al. found that while S-Detect had lower sensitivity, it exhibited higher specificity (90.8%) compared to two radiologists, whose specificities were 49.2% and 55.4%. In their study, the improvement in diagnostic performance was significant, regardless of radiologists' experience (21). However, in our study, the difference in diagnostic performance between S-Detect and the initial assessment by expert radiologist R1 was minimal. R1 demonstrated a high sensitivity (94.17%), NPV, and diagnostic accuracy (91.43%) on the initial grayscale assessment, even without AI assistance.

Interestingly, on reassessment with S-Detect there was improvement in diagnostic sensitivity, specificity, PPV, NPV, and overall accuracy across all radiologists. The most pronounced improvements were seen in less experienced radiologists (R3–R5), highlighting S-Detect's ability to standardize interpretations and bridge the performance gap between novice and expert radiologists. The substantial agreement between S-Detect and R1 (kappa value = 0.792) emphasizes the consistency of this tool with expert judgment. However, lower agreement with R4 and R5 (kappa values = 0.514 and 0.376) highlights its potential role in guiding clinical decisions for practitioners still developing their expertise. The agreement between less experienced radiologists and the expert radiologist improved from fair to substantial after incorporating AI assistance. In another research study by Wei et al., the integration of S-Detect led to more changes in assessments by less experienced radiologists compared to their experienced counterparts. Amateur radiologists showed significant improvements in accuracy, specificity, PPV, NPV, and AUC ($P<0.05$), while no significant changes were observed for experienced radiologists ($P>0.05$) (23). In our study, both the novice radiologists (R3, R4, R5) and the one with moderate experience (R2) benefited. Studies by Kim et al. and Bartolotta et al. also reiterated similar results (15,19,20).

Inter-observer variability remains a significant challenge in US imaging, especially for BI-RADS category 4 lesions,

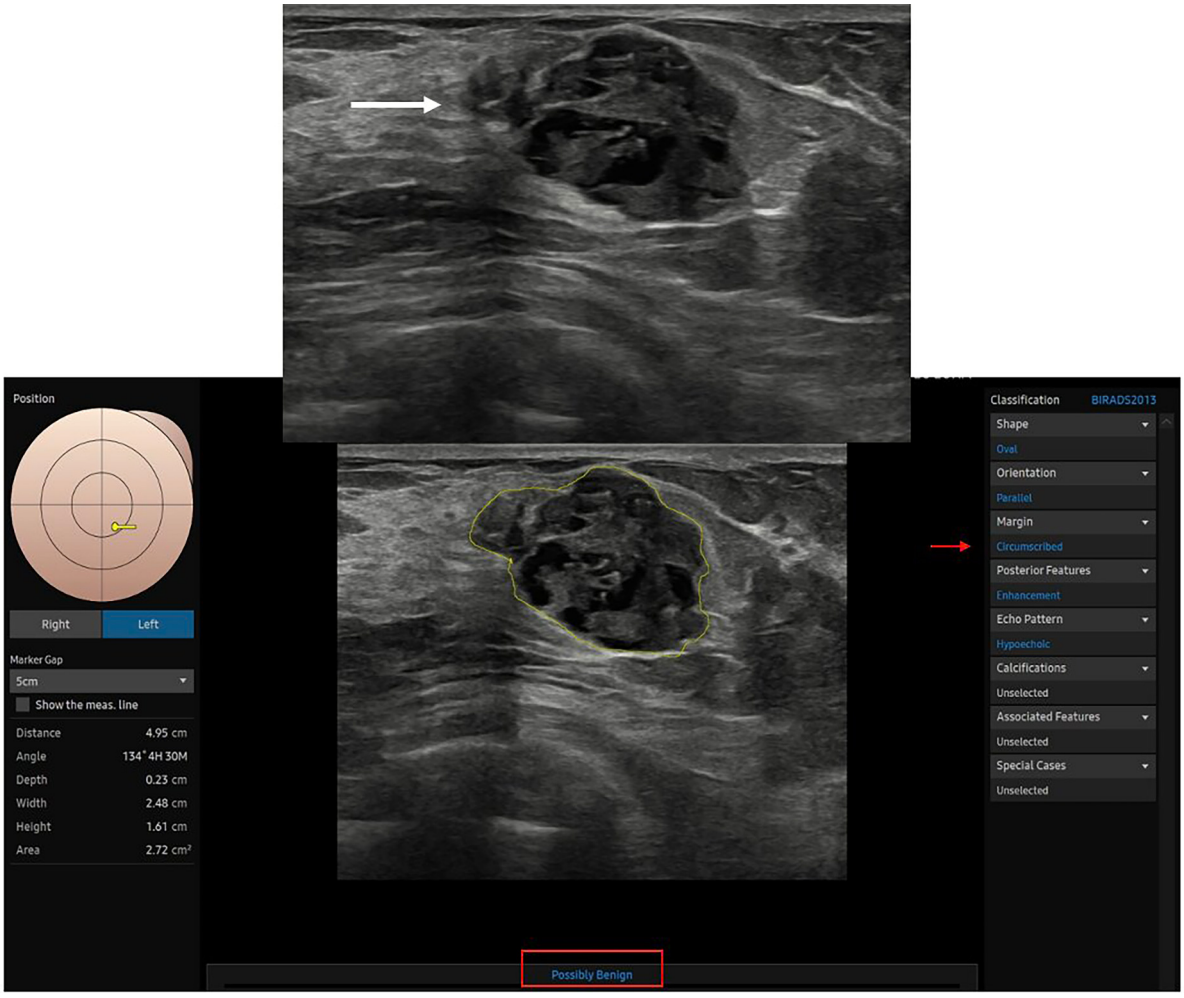

Fig. 6. A 38-year-old woman with a palpable lump in the left breast presented for evaluation. B-mode ultrasound revealed a 24 × 16 mm heterogeneously hypoechoic mass (white arrow) in the left breast. The BI-RADS categories assigned on conventional ultrasound by radiologists R1–R5 were 3, 4A, 4A, 4B, and 4B, respectively. With S-Detect (yellow), the margins were read as circumscribed, and the mass was interpreted as "possibly benign." The mass was reclassified by R1–R5 as BI-RADS 3, 4A, 3, 3, and 3 on combined reassessment with S-Detect. Core-needle biopsy confirmed a benign fibroadenoma, concordant with the S-Detect results, leading to a true downgrade.

which have a wide likelihood for malignancy (range = 2%–95%). DL-powered automated segmentation and morphological analysis mitigate this variability, providing a robust framework for consistent lesion characterization. By leveraging its structured BI-RADS descriptors and dichotomous classification system, radiologists can reduce dependence on subjective interpretations, leading to more standardized reporting. In addition, the improvement in specificity (92.7% with S-Detect versus 53%–86% across radiologists R1–R5) suggests its potential to reduce reliance on diagnostic interventions, thus alleviating patient distress and reducing healthcare costs.

For example, benign lesions in BI-RADS 4A category, often subjected to unnecessary biopsies, could be safely downgraded with S-Detect, avoiding invasive procedures while maintaining diagnostic accuracy. In our study, the correct downgrades by various radiologists were R1 (n = 4), R2 (n = 6), R3 (n = 13), R4 (n = 17), and R5 (n = 23). This highlights the tool's ability to enhance decision-making, particularly for less experienced radiologists and avoid overtreatment.

In addition, the correct upgrades recorded were R1 (n = 3), R2 (n = 4), R3 (n = 6), R4 (n = 7), and R5 (n = 11), further highlighting its role in improving the identification of malignant lesions and boosting the cancer detection rate. Intriguingly, the agreement of all the less experienced radiologists with the expert radiologists improved from fair to substantial after integrating the software. Zhao et al. reported that S-Detect correctly downgraded multiple BI-RADS 4A lesions to "possibly benign." Combining residents' evaluations with S-Detect improved specificity (46.02%–76.11%) and AUC (0.71–0.85; $P<0.001$), aligning with our findings (17). The results were in accordance with our study. Similar results were also found in a study by Park et al. (24).

Our findings are particularly impactful in low-resource settings, where trained breast radiologists are scarce. AI assistance significantly improves diagnostic accuracy, reduces observer variability, and bridges the expertise gap among radiologists. The ability of AI can be harnessed not only to provide support, but also to train young radiologists.

The present study has some limitations. First, S-Detect cannot autodetect microcalcifications—a critical marker for malignancy—which may limit its sensitivity in some cases. Enhancing the algorithm for microcalcification analysis could improve its accuracy. Second, the study excluded elastography or Doppler analysis, which might complement this AI tool's capabilities and reduce the observer variability. Third, performance in non-mass findings and large masses was not studied. In addition, for inclusion as benign, we accepted core needle biopsy-proven benign lesions performed at least 6 months prior, which is a relatively shorter follow-up**.** This approach was intended to balance diagnostic certainty with clinical practice, where benign pathology with imaging concordance is routinely followed. Because most lesions in our study underwent histopathological verification to establish a definitive diagnosis, our cohort contained a higher proportion of malignant cases than would typically be seen in a general breast imaging or screening population. This design choice ensured diagnostic certainty but also introduced selection bias. Although this does not undermine sensitivity and specificity, the predictive values and overall accuracy reported may be higher than what would be observed in a general clinical population with a lower cancer prevalence. Lastly, single-center data limit generalizability. Advancing AI algorithms, integrating multi-modal approaches, and broader validation could unlock its potential as a robust tool for precise breast cancer diagnostics in diverse clinical settings.

In conclusion, integration of DL tools into US workflow enhances diagnostic performance, particularly for novitiate radiologists who tend to under-/overestimate the mass morphology. It bridges the experience gap, reduces interobserver variability, minimizes unnecessary biopsies, and standardizes evaluations. For seasoned breast radiologists, AI-assisted tools can serve as a valuable second opinion, providing an additional layer of validation, and boosting diagnostic confidence.

Declaration of conflicting interest

The authors disclosed no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Veenu Singla https://orcid.org/0000-0002-7351-4964
Dollphy Garg https://orcid.org/0000-0002-9550-9736
Sapna Negi https://orcid.org/0009-0009-1617-4594
Nandita Mehta https://orcid.org/0009-0003-0623-4879

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74:229–263.
2. Ren W, Chen M, Qiao Y, et al. Global guidelines for breast cancer screening: a systematic review. Breast 2022;64:85–99.
3. Gegios AR, Peterson MS, Fowler AM. Breast cancer screening and diagnosis: recent advances in imaging and current limitations. PET Clin 2023;18:459–471.
4. Tagliafico AS, Calabrese M, Mariscotti G, et al. Adjunct screening with tomosynthesis or ultrasound in women with mammography-negative dense breasts: interim report of a prospective comparative trial. J Clin Oncol 2016;34:1882–1888.
5. Chang JM, Koo HR, Moon WK. Radiologist-performed hand-held ultrasound screening at average risk of breast cancer: results from a single health screening center. Acta Radiol 2015;56:652–658.
6. Bahl M, Chang JM, Mullen LA, et al. Artificial intelligence for breast ultrasound: AJR expert panel narrative review. AJR Am J Roentgenol 2024;4:e2330645.
7. Villa-Camacho JC, Baikpour M, Chou SH. Artificial intelligence for breast US. J Breast Imaging 2023;5:11–20.
8. Eghtedari M, Chong A, Rakow-Penner R, et al. Current status and future of BI-RADS in multimodality imaging, from the AJR special series on radiology reporting and data systems. AJR Am J Roentgenol 2021;216:860–873.
9. Abdullah N, Mesurolle B, El-Khoury M, et al. Breast imaging reporting and data system lexicon for ultrasound: interobserver agreement for assessment of breast masses. Radiology 2009;252:665–672.
10. Lazarus E, Mainiero MB, Schepps B, et al. BI-RADS lexicon for ultrasound and mammography: interobserver variability and positive predictive value. Radiology 2006;239:385–391.
11. Zhao C, Xiao M, Ma L, et al. Enhancing performance of breast ultrasound in opportunistic screening women by a deep learning-based system: a multicenter prospective study. Front Oncol 2022;12:804632.
12. 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.
13. Song P, Zhang L, Bai L, et al. Diagnostic performance of ultrasound with computer-aided diagnostic system in detecting breast cancer. Heliyon 2023;9:e20712.
14. Chan HP, Hadjiiski LM, Samala RK. Computer-aided diagnosis in the era of deep learning. Med Phys 2020;47:e218–e227.
15. Bartolotta TV, Orlando AA, Di Vittorio ML, et al. S-Detect characterization of focal solid breast lesions: a prospective analysis of inter-reader agreement for US BI-RADS descriptors. J Ultrasound 2021;24:143–150.

16. Wang XY, Cui LG, Feng J, et al. Artificial intelligence for breast ultrasound: an adjunct tool to reduce excessive lesion biopsy. Eur J Radiol 2021;138:109624.
17. Zhao C, Xiao M, Liu H, 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.
18. Yongping L, Zhou P, Juan Z, et al. Performance of computer-aided diagnosis in ultrasonography for detection of breast lesions less and more than 2 cm: prospective comparative study. JMIR Med Inform 2020;8:e16334.
19. Kim K, Song MK, Kim EK, 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 2016;36:3.
20. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.
21. Cho E, Kim EK, Song MK, et al. Application of computer-aided diagnosis on breast ultrasonography: evaluation of diagnostic performances and agreement of radiologists according to different levels of experience. J Ultrasound Med 2018;37:209–216.
22. Li N, Liu W, Zhan Y, et al. Value of S-detect combined with multimodal ultrasound in differentiating malignant from benign breast masses. Egypt J Radiol Nucl Med 2024;55:22.
23. Wei Q, Zeng SE, Wang LP, et al. The added value of a computer-aided diagnosis system in differential diagnosis of breast lesions by radiologists with different experience. J Ultrasound Med 2022;41:1355–1363.
24. Park HJ, Kim SM, La Yun B, et al. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of breast masses on ultrasound: added value for the inexperienced breast radiologist. Medicine (Baltimore) 2019;98:e14146.