ULTRASONOGRAPHY

Improving breast ultrasonography education: the impact of AI-based decision support on the performance of non-specialist medical professionals

Sangwon Lee[1], Hye Sun Lee[2], Eunju Lee[2], Won Hwa Kim[3,4], Jaeil Kim[4,5], Jung Hyun Yoon[1]

[1]Department of Radiology, Severance Hospital, Research Institute of Radiological Science, Yonsei University College of Medicine, Seoul; [2]Biostatistics Collaboration Unit, Yonsei University College of Medicine, Seoul; [3]Department of Radiology, School of Medicine, Kyungpook National University, Kyungpook National University Chilgok Hospital, Daegu; [4]BeamWorks Inc., Daegu; [5]School of Computer Science and Engineering, Kyungpook National University, Daegu, Korea

ORIGINAL ARTICLE

https://doi.org/10.14366/usg.24171
eISSN: 2288-5943
Ultrasonography 2025;44:124-133

Purpose: This study evaluated the educational impact of an artificial intelligence (AI)–based decision support system for breast ultrasonography (US) on medical professionals not specialized in breast imaging.

Methods: In this multi-case, multi-reader study, educational materials, including American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) descriptors, were provided alongside corresponding AI results during training. The AI system presented results in the form of AI-heatmaps, AI scores, and AI-provided BI-RADS assessment categories. Forty-two readers evaluated the test set in three sessions: the first session (S1) occurred before the educational intervention, the second session (S2) followed education without AI assistance, and the third session (S3) took place after education with AI assistance. The area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and overall performance, were compared between the sessions.

Results: The mean sensitivity increased from 66.5% (95% confidence interval [CI], 59.2% to 73.7%) to 88.7% (95% CI, 84.1% to 93.3%), with a statistically significant difference (P<0.001), and the AUC non-significantly increased from 0.664 (95% CI, 0.606 to 0.723) to 0.684 (95% CI, 0.620 to 0.748) (P=0.300). Both measures were higher in S2 than in S1. The AI-achieved AUC was comparable to that of the expert reader (0.747 [95% CI, 0.640 to 0.855] vs. 0.803 [95% CI, 0.706 to 0.900], P=0.217). Additionally, with AI assistance, the mean AUC for inexperienced readers was not significantly different from that of the expert reader (0.745 [95% CI, 0.660 to 0.830] vs. 0.803 [95% CI, 0.706 to 0.900], P=0.120).

Conclusion: The mean AUC and sensitivity improved after incorporating AI into breast US education and interpretation. AI systems with high-level performance for breast US can potentially be used as educational tools in the interpretation of breast US images.

Keywords: Breast; Ultrasound; Breast neoplasms; Artificial intelligence; Education

Key points: The diagnostic performance of non-specialist medical professionals improved following artificial intelligence (AI)–based education on breast ultrasound, with mean sensitivity increasing from 66.5% to 88.7% (P<0.001). With AI assistance, the mean area under the receiver operating characteristic curve for inexperienced readers was not significantly different from that of the expert reader following education; 0.745 (95% confidence interval [CI], 0.660 to 0.830) versus 0.803 (95% CI, 0.706 to 0.900), respectively (P=0.120). An AI system can be an educational tool for less experienced readers.

Received: September 9, 2024
Revised: December 5, 2024
Accepted: December 12, 2024

Correspondence to:
Jung Hyun Yoon, MD, PhD, Department of Radiology, Severance Hospital, Research Institute of Radiological Science, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea

Tel. +82-2-2228-7400
Fax. +82-2-2227-8337
E-mail: lvjenny@yuhs.ac

How to cite this article:
Lee S, Lee HS, Lee E, Kim WH, Kim J, Yoon JH. Improving breast ultrasonography education: the impact of AI-based decision support on the performance of non-specialist medical professionals. Ultrasonography. 2025 Mar; 44(2):124-133.

Introduction

Breast ultrasonography (US) is a safe and readily accessible diagnostic tool for evaluating breast lesions. A recent meta-analysis revealed that the pooled sensitivity and specificity of breast US in diagnostic settings are 74% and 89%, respectively, which are comparable to the performance metrics of mammography [1]. Initially, breast US was used primarily to differentiate between symptomatic or mammography-detected lesions. However, its application has broadened to include supplemental screening for asymptomatic women with mammographically dense breasts [2,3] or those at high risk for breast cancer [4]. Furthermore, the introduction of breast density notification laws in the United States has significantly increased the utilization of US examinations, with increases ranging from 0.5% to 143% (median, 16%) [5].

While breast US is an effective imaging tool, one of its major limitations is that its performance is operator-dependent. The interpretation of US images heavily relies on the training, education, and experience of the operators. To minimize variability among operators, the American College of Radiology (ACR) developed the Breast Imaging Reporting and Data System (BI-RADS) [6], which is actively used in practice. Using the BI-RADS system, interobserver agreements for breast US range from fair to moderate for individual descriptors and are substantial for final assessments [7–9]. This variation among operators underscores the need for dedicated education in breast US, as discrepancies in lesion description and assessment can directly impact patient care.

At present, healthcare providers must either complete an extensive training course [10] or perform a specified number of breast US examinations [11] to qualify for independent practice in breast US. Research has shown that training focused on US BI-RADS increases diagnostic sensitivity, improves the area under the receiver operating characteristic curve (AUC), and increases inter-reader agreement [12–14]. Typically, these training programs are led by human instructors who utilize atlases containing representative US images. These atlases detail specific descriptors, define them, and explain how various combinations of these descriptors lead to particular diagnostic conclusions. Artificial intelligence (AI) has been introduced into breast imaging and is now integrated into practice through commercially available US machines [15]. Given that AI's performance in breast US matches or surpasses that of seasoned radiologists [16–18], the authors hypothesized that AI could also serve as a training tool for medical trainees, especially in regions where access to intensive medical training is scarce [19]. Therefore, the present study explored the effects of using deep-learning-based AI software on the training of medical professionals in interpreting breast US images.

Materials and Methods

Compliance with Ethical Standards

This study employed a retrospective design, and the Institutional Review Board (IRB) of Severance Hospital, Yonsei University (approval No. 4-2024-0766) waived the requirement for informed consent from the participants.

Breast US Education Course for Medical Professionals in Uzbekistan

A 1-day training course on breast US and associated AI software was held in November 2023 at three regional hospitals in Samarkand and Bukhara, Uzbekistan. The educational program was advertised in advance and took place over three consecutive days, with identical content delivered at each hospital. Forty-five medical professionals registered for the program. Basic demographic information, including age, education level, profession, and experience in breast US, was collected from the participants.

Each participating reader was provided with a tablet PC and logged into a web-based education system specifically developed for displaying images and collecting data. Within this system, readers accessed educational materials that introduced definitions for US BI-RADS descriptors, final assessments, and representative US images in the first chapter. The second chapter contained a manual describing the functionality of the AI system for breast US and how it presents its analytical results. In the third chapter, 60 breast US images (education cases) were presented, allowing readers to review the grayscale images alongside their corresponding AI results (AI score and AI-provided BI-RADS assessments) (Fig. 1) and final pathologic diagnoses. The demographic features of the 60 education cases are detailed in Supplementary Table 1. The educational materials were available in both English and Russian.

Test Set Cases for Education and Image Review

The test set comprised 60 grayscale US images of pathologically confirmed masses, typically benign lesions, and normal breast parenchyma (negative findings) (Supplementary Tables 1, 2). It excluded features such as non-mass lesions, calcifications, duct changes, and distortions, as the AI system's performance on these features had not been investigated. The cases used in the test set and the educational materials were not employed for training or tuning the AI system and were collected as an independent external test set. A breast-dedicated radiologist with 15 years of experience (J.H.Y.) reviewed the US images and recorded data using a binary scale (non-cancer vs. cancer) and the US BI-RADS final assessment scale (BI-RADS categories 1–5, including subcategories for BI-RADS category 4).

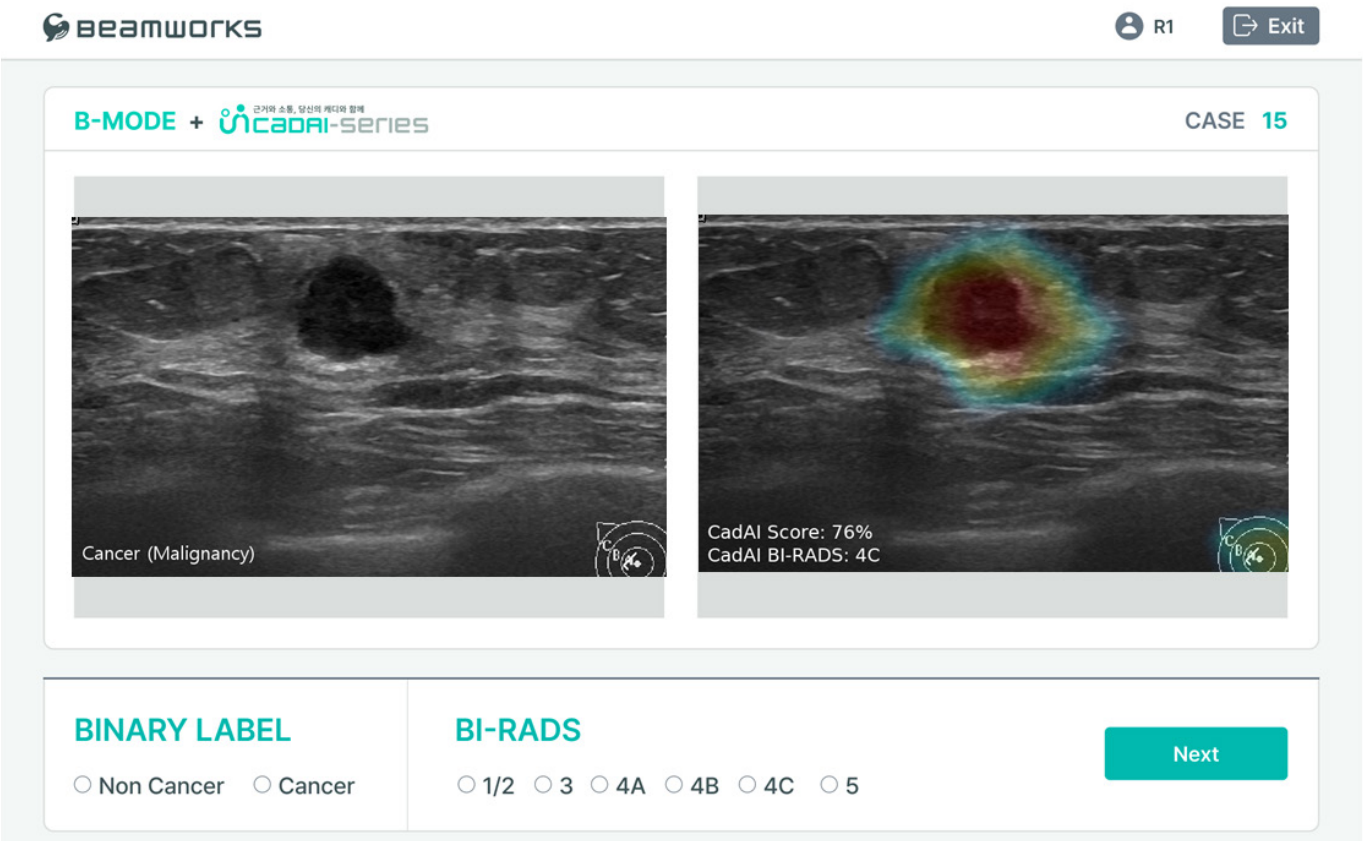

Fig. 1. **Image of a display presented by the web-based system for image review in Session 3.** A grayscale ultrasonography is on the left. Artificial intelligence (AI) analysis results are simultaneously displayed on the right, with a semi-transparent colored heatmap, AI score, and AI-provided Breast Imaging Reporting and Data System (BI-RADS) final assessment shown in the bottom corner. Participants were asked to mark their interpretations according to (1) binary scale (non-cancer vs. cancer), and (2) BI-RADS final assessment with subcategorizations for BI-RADS 4.

AI System for Breast US

A deep learning-based algorithm (CadAI-B for Breast v.1.0, BeamWorks Inc., Daegu, Korea) was developed to identify lesions in breast US images and provide diagnostic recommendations. This AI system has been approved by the Korean Ministry of Food and Drug Administration for use in assisting with the interpretation of breast US across different vendors and is now commercially available in Korea. The AI system utilizes a deep neural network to detect regions suspicious for malignancy (computer-aided detection, CADe) and to classify these as either malignant or non-malignant (computer-aided diagnosis, CADx) simultaneously. The algorithms of CadAI-B are based on a convolutional neural network architecture and were trained using weakly-supervised learning [20]. The CADe functionality is designed to aid in identifying suspicious regions within breast US images by highlighting these areas on a relevance map. This map provides a pixel-level abnormality score, termed the "AI score," which ranges from 0 to 1 and is effectively visualized as an "AI-heatmap" (Fig. 1). The AI-heatmap indicates the likelihood of malignancy, with color highlighting from BI-RADS category 4A upwards, where higher probabilities are shown in red. The CADx functionality generates a probability of malignancy using an AI classifier that is calibrated to align with the ACR BI-RADS categorization system, referred to as "AI-provided BI-RADS."

The AI model was trained using over 600,000 breast US images, which included both static and cine images. For model tuning, a subset of 1,300 images—1,000 benign and 300 malignant—was selected. The remaining images were utilized for training the model. Subsequently, the trained model was calibrated with the tuning set to ensure that the probability predictions aligned with the ACR BI-RADS framework. The calibrated model can categorize findings into six BI-RADS assessment categories: 1/2, 3, 4A, 4B, 4C, and 5. The model's accuracy was confirmed through comprehensive reviews by expert radiologists. Additional information on the model's development and validation process is available in Supplementary Text 1 and Supplementary Figs. 1 and 2.

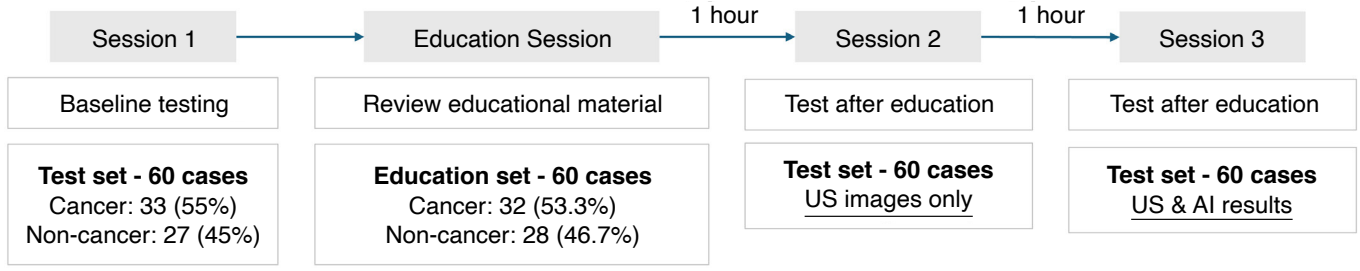

Fig. 2. **Flowchart of the image review sessions.** Participants reviewed the test set before education in session 1 (S1). The education course proceeded with educational material containing definitions and examples of Breast Imaging Reporting and Data System descriptors and final assessments made available with the education image set. After the education session, readers conducted sessions 2 (S2) and 3 (S3) using the same test set. In S3, the artificial intelligence (AI) results were shown simultaneously with the ultrasonography (US) images.

Image Review Sessions and Data Collection

Readers participated in three review sessions (Fig. 2), with a 1-hour interval between each session. Prior to beginning the education course, readers were instructed to individually assess the test set during the first session (S1). Following S1, they independently reviewed the educational material, which included 60 cases from the education set along with AI results. Subsequently, session 2 (S2) commenced, during which readers evaluated the same test set as in S1. In session 3 (S3), participants were required to review the test set again; however, this time grayscale US images were displayed on the left, with AI results simultaneously presented on the right (Fig. 1). The readers were instructed to document their interpretations in the same manner as an experienced reader, first using a binary scale and then providing a US BI-RADS final assessment.

Data and Statistical Analysis

The ground truth for the 60 breast masses in the test set was classified as either benign or malignant based on one of three criteria: (1) typically benign US features, (2) biopsy results, or (3) stability observed over more than 2 years of follow-up.

Diagnostic performance was assessed using AUC, sensitivity, and specificity, which were reported as mean values along with a 95% confidence interval (CI) for each session. These metrics were compared across sessions using the bootstrapping method for the readers, the expert reader, and AI. For BI-RADS assessments, the threshold was established at BI-RADS category 4A, which suggests a potential malignancy and necessitates a biopsy. Readers were categorized based on their experience with breast US in clinical settings. Specifically, experienced readers were those who had over 10 years of experience with breast US. The diagnostic performances of the three sessions were calculated and compared among the reader groups based on their experience levels.

Statistical analysis was conducted using SAS software (version 9.4, SAS Institute Inc., Cary, NC, USA) and the R package (version 4.0.5, http://www.R-project.org). A P-value of less than 0.05 was considered statistically significant. Authors W.H.K. and J.K. are employees of BeamWorks Inc., which supplied the equipment used in this study. The other authors, who are not employees of the vendor, maintained complete control over the data and the information submitted for publication.

Table 1. **Demographics of the 42 readers in the education program**

Characteristic	No. (%)
Current profession	
Board-certified radiologist	13 (31.0)
Trainee (intern or resident)	17 (40.5)
Technologist	4 (9.5)
Medical student	3 (7.1)
No reply	5 (11.9)
Level of experience in breast US	
<1 year	25 (59.6)
1–10 years	9 (21.4)
>10 years	4 (9.5)
No reply	4 (9.5)

US, ultrasound.

Results

Reader Characteristics

A total of 45 readers participated in an educational program, reviewing breast US images across three sessions using various approaches. Of these, three readers who failed to respond to more than 10% of the 60 cases were excluded from the study. The demographics of the remaining 42 readers are summarized in Table 1. All participants were from Uzbekistan, including 13 board-certified radiologists (31.0%), 17 trainees (interns or residents, 40.5%),

Table 2. Diagnostic performance of the readers and AI according to the reading session

	AI	Expert reader	P-value[a]	S1	S2	S3	P-value[b]	P-value[c]	P-value[d]
Binary assessment (non-cancer vs. cancer)									
AUC	0.747 (0.640–0.855)	0.860 (0.776–0.945)	0.062	0.662 (0.604–0.721)	0.711 (0.640–0.782)	0.768 (0.694–0.842)	0.011	0.005	0.032
Sensitivity (%)	93.9 (85.9–100.0)	75.8 (60.8–90.7)	0.008	69.1 (62.3–76.0)	87.8 (82.4–93.3)	89.0 (83.4–94.7)	<0.001	0.590	0.037
Specificity (%)	55.6 (35.9–75.2)	96.3 (89.0–100.0)	<0.001	63.3 (53.9–72.8)	54.4 (41.4–67.5)	63.4 (49.8–77.0)	0.003	0.001	<0.001
BI-RADS final assessment									
AUC	0.747 (0.640–0.855)	0.803 (0.706–0.900)	0.217	0.664 (0.606–0.723)	0.684 (0.620–0.748)	0.746 (0.659–0.833)	0.297	0.015	0.130
Sensitivity (%)	93.9 (85.9–100.0)	93.9 (85.7–100.0)	>0.999	66.5 (59.2–73.7)	88.7 (84.1–93.3)	90.8 (84.8–96.7)	<0.001	0.401	0.387
Specificity (%)	55.6 (35.9–75.2)	66.7 (49.0–84.4)	0.168	66.4 (57.1–75.7)	48.1 (36.3–60.0)	58.5 (42.3–74.6)	<0.001	0.018	0.225

95% Confidence intervals are in parentheses.

S1: session 1, where readers reviewed the US images before education. S2: session 2, where readers reviewed the US images after education, without AI assistance. S3: session 3, where readers reviewed the US images after education, with AI assistance.

AI, artificial intelligence; AUC, area under the receiving operator characteristic curve; BI-RADS, Breast Imaging Reporting and Data System.

[a]P-value: comparison between the expert reader and AI. [b]P-value: comparison between S1 and S2. [c]P-value: comparison between S2 and S3. [d]P-value: comparison between the expert reader and S3.

and four technologists (9.5%). In terms of clinical experience with breast US, 25 readers (59.6%) had less than one year of experience and were categorized as inexperienced, while four readers (9.5%) had over 10 years of experience and were considered experienced practitioners.

Diagnostic Performance of Readers According to Reading Session

Table 2 summarizes the diagnostic performances of the 42 readers across different reading sessions. When interpreting images using binary assessments, the mean AUC significantly increased in session 2 (S2) compared to session 1 (S1), with values of 0.711 (95% CI, 0.640 to 0.782) vs. 0.662 (95% CI, 0.604 to 0.721), respectively (P=0.011). The mean AUC also showed a significant increase in session 3 (S3) compared to S2, with values of 0.768 (95% CI, 0.694 to 0.842) vs. 0.711 (95% CI, 0.640 to 0.782), respectively (P=0.005). Using the BI-RADS assessment, the mean AUC did not show a significant difference between S2 and S1, with values of 0.684 (95% CI, 0.620 to 0.748) vs. 0.664 (95% CI, 0.606 to 0.723), respectively (P=0.297). However, there was a significant increase in the mean AUC in S3 compared to S2, with values of 0.746 (95% CI, 0.659 to 0.833) vs. 0.684 (95% CI, 0.620 to 0.748), respectively (P=0.015). Changes in AUC according to the training session are visualized in Fig. 3.

Similar trends were observed in the mean sensitivity and specificity for both binary and BI-RADS assessments. Specifically, mean sensitivity showed a significant increase in S2 compared to S1. For the binary assessment, sensitivity rose from 69.1% (95% CI, 62.3 to 76.0%) in S1 to 87.8% (95% CI, 82.4% to 93.3%) in S2. Similarly, for the BI-RADS assessment, sensitivity increased from 66.5% (95% CI, 59.2% to 73.7%) in S1 to 88.7% (95% CI, 84.1% to 93.3%) in S2 (all P<0.001).

Mean specificity significantly decreased in S2 compared to S1; for binary assessment, 54.4% (95% CI, 41.4% to 67.5%) vs. 63.3% (95% CI, 53.9% to 72.8%) and for BI-RADS assessment, 48.1% (95% CI, 36.3% to 60.0%) vs. 66.4% (95% CI, 57.1% to 75.7%), respectively (P=0.003 and P<0.001). In S3, mean specificity significantly increased compared to S2, for both binary (63.4% [95% CI, 49.8% to 77.0%]) and BI-RADS assessments (58.5% [95% CI, 42.3% to 74.6%]) (P=0.001 and P=0.018), but not for mean sensitivity (Fig. 4, Supplementary Fig. 3).

Comparison of Performance between Standalone AI and the Expert Reader

The diagnostic performances of the standalone AI and expert reader are shown in Table 2. The AUC of AI was comparable to the AUC of the expert reader; for binary assessments, 0.747 (95% CI, 0.640 to 0.855) vs. 0.860 (95% CI, 0.776 to 0.945) (P=0.062), and for BI-RADS assessments, 0.747 (95% CI, 0.640 to 0.855) vs. 0.803 (95% CI, 0.706 to 0.900), respectively (P=0.217). For binary assessments, AI showed a significantly higher sensitivity than the expert reader (93.9% [95% CI, 85.9% to 100%] vs. 75.8% [95% CI, 60.8%

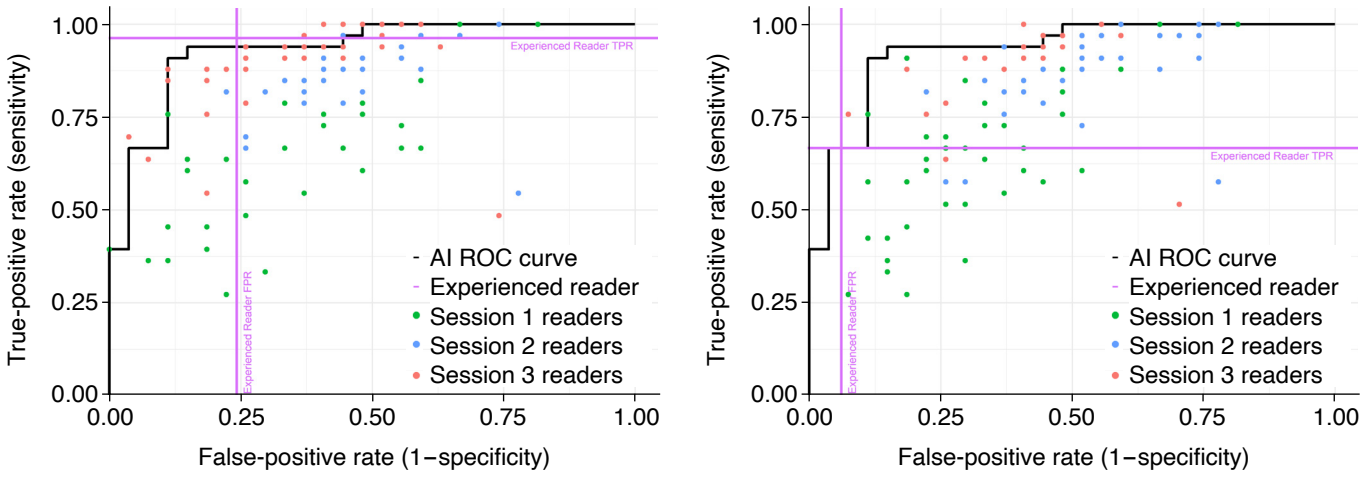

Fig. 3. Receiver operative characteristic (ROC) curve of standalone artificial intelligence (AI), expert reader, and the 42 readers according to reading sessions using binary assessment (A) and Breast Imaging Reporting and Data System assessment (B).
Compared to the performance before education (S1, green dots), the areas under the curves of the 42 readers increased as sessions passed (S2: blue dots, S3: red dots).

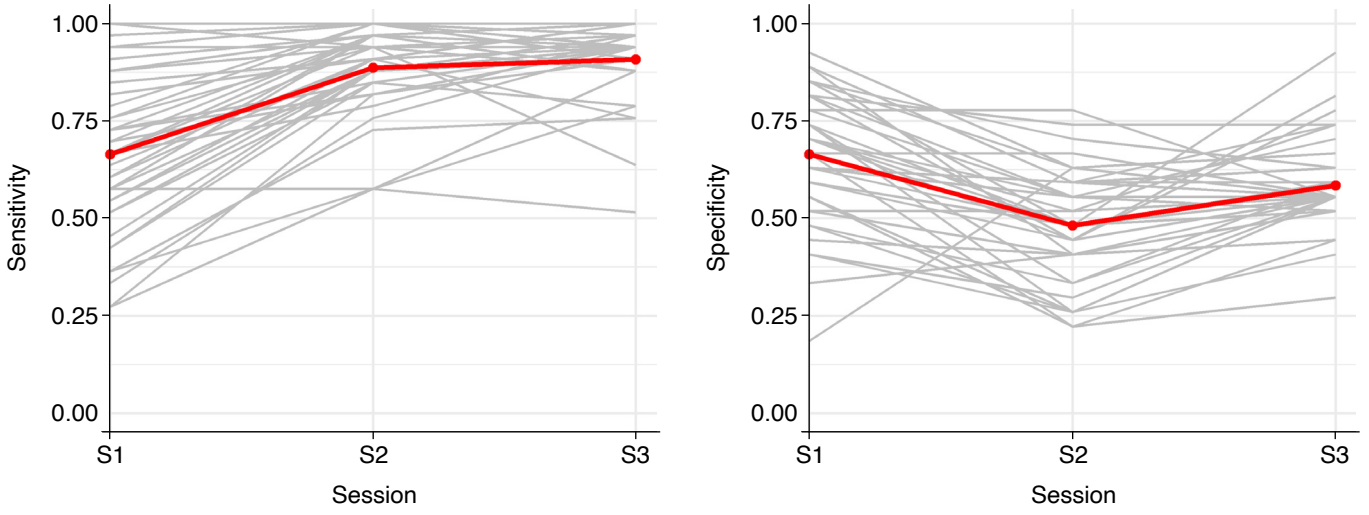

Fig. 4. Spaghetti plots showing the changes in sensitivity (A) and specificity (B) of readers according to the reading session, based on Breast Imaging Reporting and Data System (BI-RADS) assessments.
Gray lines illustrate the change in interpretive skills of each reader according to the reading session, and red lines show the mean sensitivity and specificity of the 42 readers. Binary assessments showed a similar trend to the BI-RADS assessments (given as Supplementary Fig. 3).

to 90.7%]) (P=0.008), but significantly lower specificity (55.6% [95% CI, 35.9% to 75.2%)] vs. 96.3% [95% CI, 89.0% to 100%]), respectively (P<0.001). For the BI-RADS assessments, sensitivity and specificity were not significantly different between AI and the expert reader (P>0.999 and P=0.168, respectively).

Diagnostic Performance According to Experience Level

Table 3 demonstrates the diagnostic performances of readers subgrouped according to experience level. In inexperienced readers, AUCs for both binary and BI-RADS assessments were significantly higher in S3 than in S2 (P=0.006 and P=0.010, respectively). The AUCs of experienced readers did not differ between S2 and S3 (all P>0.05, respectively). Regardless of the level of experience, mean sensitivity for both binary and BI-RADS assessments significantly increased in S2 compared to S1 (all P<0.001, respectively), while mean specificity significantly decreased in S2 (all P<0.05,

Table 3. Diagnostic performance of the readers according to experience level

	Expert reader	Inexperienced (n=38)				Experienced (n=4)			
		S1	S2	S3	P-value[a]	S1	S2	S3	P-value[b]
Binary assessment									
AUC	0.860 (0.776–0.945)	0.661 (0.602–0.719)	0.709 (0.638–0.780) P=0.018[c]	0.759 (0.684–0.833) P=0.006[d]	0.029	0.675 (0.598–0.752)	0.727 (0.640–0.814) P=0.115[c]	0.795 (0.719–0.872) P=0.050[d]	0.108
Sensitivity (%)	75.8 (60.8–90.7)	71.1 (64.4–77.7)	88.8 (83.4–94.1) P<0.001[c]	89.6 (84.1–95.1) P=0.716[d]	0.033	50.8 (39.6–61.9)	78.8 (68.8–88.8) P<0.001[c]	84.1 (75.8–92.4) P=0.309[d]	0.147
Specificity (%)	96.3 (89.0–100.0)	61.1 (51.5–70.7)	53.1 (40.1–66.2) P=0.012[c]	62.2 (48.4–75.9) P=0.001[d]	<0.001	84.3 (73.4–95.1)	66.7 (52.1–81.2) P=0.001[c]	75.0 (62.3–87.7) P=0.094[d]	<0.001
BI-RADS final assessment									
AUC	0.803 (0.706–0.900)	0.664 (0.607–0.722)	0.681 (0.618–0.745) P=0.401[c]	0.745 (0.660–0.830) P=0.010[d]	0.120	0.663 (0.583–0.744)	0.711 (0.626–0.796) P=0.074[c]	0.756 (0.652–0.859) P=0.288[d]	0.264
Sensitivity (%)	93.9 (85.7–100.0)	67.9 (60.8–74.9)	89.5 (85.2–93.7) P<0.001[c]	90.4 (84.5–96.2) P=0.714[d]	0.329	53.0 (41.9–64.1)	81.1 (71.1–91.0) P<0.001[c]	94.7 (87.6–100.0) P=0.006[d]	0.840
Specificity (%)	66.7 (49.0–84.4)	65.0 (55.8–74.2)	46.8 (34.9–58.7) P<0.001[c]	58.7 (42.8–74.5) P=0.005[d]	0.232	79.6 (67.9–91.3)	61.1 (47.0–75.3) P<0.001[c]	56.5 (37.2–75.7) P=0.508[d]	0.185

95% Confidence intervals are in parentheses.

S1: session 1, where readers reviewed the US images before education. S2: session 2, where readers reviewed the US images after education, without AI assistance. S3: session 3, where readers reviewed the US images after education, with AI assistance.

AUC, area under the receiving operator characteristic curve; BI-RADS, Breast Imaging Reporting and Data System; AI, artificial intelligence.

[a]P-value: comparison between expert and S3 of inexperienced readers. [b]P-value: comparison between expert and S3 of experienced readers. [c]P-value: comparison between S1 and S2. [d]P-value: comparison between S2 and S3.

respectively). For session 3, experienced readers showed significantly increased sensitivity (P=0.006), while inexperienced readers showed significantly increased specificity (P=0.005).

Compared to the expert reader, inexperienced readers showed significantly lower mean AUC in S3 for binary assessments (P=0.029), whereas experienced readers did not (P=0.108). The mean AUC of S3 for BI-RADS assessments did not significantly differ between readers regardless of their level of experience compared to the expert reader.

Discussion

This study investigated the educational impact of an AI system on breast US for medical professionals who lacked specific training in breast imaging. The findings revealed that the mean AUC for readers significantly improved after integrating AI into their education on breast US images (S2) and providing AI results to aid in interpretation (S3). Post-education, readers demonstrated increased mean sensitivity with the aid of AI, although specificity initially declined after the educational intervention but improved with AI assistance. In scenario S3, concerning BI-RADS assessments, readers achieved mean diagnostic performances comparable to that of an expert reader, irrespective of their initial experience level.

Compared to the baseline reading session, sensitivity significantly increased for both binary and BI-RADS assessments after education, regardless of the participants' experience levels. The findings align with those of a previous study that evaluated the impact of educating radiology residents on US BI-RADS. In that study, sensitivity not only improved significantly post-training but also remained high for an extended period [12]. Similarly, the ACRIN 6666 investigator study reported enhanced performance in breast US interpretation when participants received immediate feedback on BI-RADS features and histopathologic results. Notably, those in the lowest quartile of initial performance exhibited the most significant improvement in sensitivity, with no changes to AUC [13]. The definitions of US BI-RADS descriptors and final assessments primarily focus on features indicative of malignancy. Therefore, the observed trend of increased detection sensitivity following education on US BI-RADS is consistent and expected.

Past studies have typically used specialized educational

materials, including breast US images or video clips, along with expert consensus on image descriptors or final assessments to train participants [12,13,21]. In contrast, this education program provided AI analysis results, which included AI scores and BI-RADS final assessments generated by AI for specific US images. This approach differs significantly from previous methods that either involved didactic sessions focused on US BI-RADS [13] or provided a large number of example cases, with some sessions reviewing approximately 100 cases [12,21]. This method, which solely utilized a set of US cases accompanied by corresponding AI results, led to increased sensitivity and AUC among breast US readers. Given that there were no significant differences in performance between standalone AI and expert readers in BI-RADS assessments, it is evident that AI can offer educational examples that are as effective as the traditional expert consensus data historically used to train healthcare providers in US interpretation. Furthermore, the education of US practitioners is crucial for ensuring standardized interpretations and minimizing unnecessary false-positive results in breast US, particularly in regions with scarce medical resources. With its robust standalone performance, AI holds promise for enhancing educational opportunities in areas where access to both medical resources and educational programs is limited.

After the educational session, the use of AI to assist in US interpretation (S3) resulted in increased AUC and specificity for both binary and BI-RADS assessments. Sensitivity remained high at 93.9% with AI, which may have initially enhanced the readers' sensitivity, but this came at the cost of reduced specificity post-education. When AI was employed to aid interpretation, specificity significantly improved, allowing readers to disregard the false-positive interpretations they had previously made in S2 without AI assistance. The enhancement in specificity with AI support aligns with the findings of previous studies [22–24]. Furthermore, after AI assistance (S3), AUC, sensitivity, and specificity were not significantly different from the performance levels of an expert reader in BI-RADS assessments, even for inexperienced readers (all $P>0.05$) (Table 3). Given that BI-RADS assessment is a standardized method for US interpretation in routine practice, the findings indicate that employing high-performance AI software can elevate the proficiency of inexperienced readers to that of an expert radiologist.

This study has several limitations. First, the same test set was used for each session (S1 to S3), which may have introduced bias due to memory effects. Additionally, due to time constraints, there was no sufficient washout period between sessions, potentially further contributing to memory bias. Second, the AI analysis results were presented in three formats—AI-heatmap, AI score, and AI-provided BI-RADS—to the readers simultaneously. It is difficult to determine which aspect of the AI system was most influential in educating the readers or influencing their decision-making. Third, the educational materials included examples of BI-RADS descriptors and an image set with concurrent AI results. This introduces uncertainty regarding which element primarily contributed to the observed changes in performance. It is possible that results might have varied if only AI results were used for educational purposes; however, this approach would lack realism as readers require a basic understanding of US BI-RADS descriptors and assessment to interpret breast US images effectively. Lastly, the study tested a single expert reader and a single AI system. The results might vary with the inclusion of different expert readers or AI systems with varying performance levels. Large, multicenter studies are needed to validate these findings.

In conclusion, the mean AUC and sensitivity were enhanced when AI was utilized to train and support medical professionals in interpreting breast US images. AI systems demonstrating high-level performance in breast US could serve as effective educational tools for interpreting these images. With the aid of AI in image interpretation, the performances of inexperienced readers closely matched that of the expert reader.

ORCID: Sangwon Lee: https://orcid.org/0000-0003-3089-8491; Hye Sun Lee: https://orcid.org/0000-0001-6328-6948; Eunju Lee: https://orcid.org/0009-0003-7271-8310; Won Hwa Kim: https://orcid.org/0000-0001-7137-9968; Jaeil Kim: https://orcid.org/0000-0002-9799-1773; Jung Hyun Yoon: https://orcid.org/0000-0002-2100-3513

Author Contributions

Conceptualization: Lee S, Lee HS, Lee E, Kim WH, Yoon JH. Data acquisition: Lee S, Kim WH, Kim J. Data analysis or interpretation: Lee HS, Lee E, Kim J, Yoon JH. Drafting of the manuscript: Lee S, Kim WH, Kim J, Yoon JH. Critical revision of the manuscript: Lee S, Lee HS, Lee E, Yoon JH. Approval of the final version of the manuscript: all authors.

Conflict of Interest

Authors W.H.K. and J.K. are the CEOs of BeamWorks Inc., which provided the equipment for this study. The remaining authors had full control of the data and information submitted for publication.

Acknowledgments

This study was supported by the Korea Medical Device Development Fund grant funded by the Korean government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 1711197554, RS-2023-00227526). The authors would like to thank Tae-Du Jung and Hyeryeong Son from the International Healthcare Business Team of Kyungpook National

University Chilgok Hospital for their support in this study.

Supplementary Material

Supplementary Text 1. Description of the artificial intelligence (AI) system for breast ultrasound (CadAI-B for Breast) (https://doi.org/10.14366/usg.24171).

Supplementary Table 1. Demographics of the education and test set used in this study (https://doi.org/10.14366/usg.24171).

Supplementary Table 2. Distribution of cases according to AI findings and final pathology (https://doi.org/10.14366/usg.24171).

Supplementary Fig. 1. Real-time mode of CadAI-B for breast (https://doi.org/10.14366/usg.24171).

Supplementary Fig. 2. Frame-captured ultrasound images with CadAI-B for breast (https://doi.org/10.14366/usg.24171).

Supplementary Fig. 3. Spaghetti plots showing the changes of sensitivity and specificity of readers according to reading session, based on binary assessments (https://doi.org/10.14366/usg.24171).

References

1. Tadesse GF, Tegaw EM, Abdisa EK. Diagnostic performance of mammography and ultrasound in breast cancer: a systematic review and meta-analysis. J Ultrasound 2023;26:355-367.
2. Yang L, Wang S, Zhang L, Sheng C, Song F, Wang P, et al. Performance of ultrasonography screening for breast cancer: a systematic review and meta-analysis. BMC Cancer 2020;20:499.
3. Berg WA, Bandos AI, Mendelson EB, Lehrer D, Jong RA, Pisano ED. Ultrasound as the primary screening test for breast cancer: analysis from ACRIN 6666. J Natl Cancer Inst 2016;108:djv367.
4. Berg WA, Blume JD, Cormack JB, Mendelson EB, Lehrer D, Bohm-Velez M, et al. Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. JAMA 2008;299:2151-2163.
5. Huang S, Houssami N, Brennan M, Nickel B. The impact of mandatory mammographic breast density notification on supplemental screening practice in the United States: a systematic review. Breast Cancer Res Treat 2021;187:11-30.
6. Mendelson E, Bohm-Velez M, Berg WA. ACR BI-RADS ultrasound. In: D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA, eds. ACR BI-RADS atlas, Breast Imaging Reporting and Data System. 5th ed. Reston, VA: American College of Radiology, 2013;128-130.
7. Lee HJ, Kim EK, Kim MJ, Youk JH, Lee JY, Kang DR, et al. Observer variability of Breast Imaging Reporting and Data System (BI-RADS) for breast ultrasound. Eur J Radiol 2008;65:293-298.
8. Lee YJ, Choi SY, Kim KS, Yang PS. Variability in observer performance between faculty members and residents using Breast Imaging Reporting and Data System (BI-RADS)-ultrasound, fifth edition (2013). Iran J Radiol 2016;13:e28281.
9. Park CS, Lee JH, Yim HW, Kang BJ, Kim HS, Jung JI, et al. Observer agreement using the ACR Breast Imaging Reporting and Data System (BI-RADS)-ultrasound, First Edition (2003). Korean J Radiol 2007;8:397-402.
10. Monticciolo DL, Rebner M, Appleton CM, Newell MS, Farria DM, Sickles EA, et al. The ACR/society of breast imaging resident and fellowship training curriculum for breast imaging, updated. J Am Coll Radiol 2013;10:207-210.
11. Education and Practical Standards Committee; European Federation of Societies for Ultrasound in Medicine and Biology. Minimum training requirements for the practice of medical ultrasound. Ultraschall Med 2006;27:79-105.
12. Yoon JH, Lee HS, Kim YM, Youk JH, Kim SH, Jeong SH, et al. Effect of training on ultrasonography (US) BI-RADS features for radiology residents: a multicenter study comparing performances after training. Eur Radiol 2019;29:4468-4476.
13. Berg WA, Blume JD, Cormack JB, Mendelson EB. Training the ACRIN 6666 Investigators and effects of feedback on breast ultrasound interpretive performance and agreement in BI-RADS ultrasound feature analysis. AJR Am J Roentgenol 2012;199:224-235.
14. Ortiz-Perez T, Trevino EJ, Sepulveda KA, Hilsenbeck SG, Wang T, Sedgwick EL. Does formal instruction about the BI-RADS ultrasound lexicon result in improved appropriate use of the lexicon? AJR Am J Roentgenol 2013;201:456-461.
15. Brunetti N, Calabrese M, Martinoli C, Tagliafico AS. Artificial intelligence in breast ultrasound: from diagnosis to prognosis: a rapid review. Diagnostics (Basel) 2022;13:58.
16. Mango VL, Sun M, Wynn RT, Ha R. Should we ignore, follow, or biopsy? Impact of artificial intelligence decision support on breast ultrasound lesion assessment. AJR Am J Roentgenol 2020;214:1445-1452.
17. Lai YC, Chen HH, Hsu JF, Hong YJ, Chiu TT, Chiou HJ. Evaluation of physician performance using a concurrent-read artificial intelligence system to support breast ultrasound interpretation. Breast 2022;65:124-135.
18. Berg WA, Gur D, Bandos AI, Nair B, Gizienski TA, Tyma CS, et al. Impact of original and artificially improved artificial intelligence-based computer-aided diagnosis on breast US interpretation. J Breast Imaging 2021;3:301-311.
19. Berg WA, Lopez Aldrete AL, Jairaj A, Ledesma Parea JC, Garcia CY, McClennan RC, et al. Toward AI-supported US triage of women with palpable breast lumps in a low-resource setting. Radiology 2023;307:e223351.

20. Kim J, Kim HJ, Kim C, Lee JH, Kim KW, Park YM, et al. Weakly-supervised deep learning for ultrasound diagnosis of breast cancer. Sci Rep 2021;11:24382.
21. Youk JH, Jung I, Yoon JH, Kim SH, Kim YM, Lee EH, et al. Comparison of inter-observer variability and diagnostic performance of the fifth edition of BI-RADS for breast ultrasound of static versus video images. Ultrasound Med Biol 2016;42:2083-2088.
22. Lee SE, Han K, Youk JH, Lee JE, Hwang JY, Rho M, et al. Differing benefits of artificial intelligence-based computer-aided diagnosis for breast US according to workflow and experience level. Ultrasonography 2022;41:718-727.
23. Cho E, Kim EK, Song MK, Yoon JH. 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.
24. Choe YH. A glimpse on trends and characteristics of recent articles published in the Korean Journal of Radiology. Korean J Radiol 2019;20:1555-1561.