Original Article

J Korean Soc Radiol 2026;87(2):314-327
https://doi.org/10.3348/jksr.2024.0144
eISSN 2951-0805

JOURNAL of THE KOREAN SOCIETY of RADIOLOGY

Feasibility of Using an AI System for Breast Ultrasonography Interpretation According to Clinical Expertise: Results of a Pilot Study

임상 경력에 따른 유방 초음파 인공지능 판독 시스템의 활용 효과를 보기 위한 파일럿 연구

Jeeyoun Kim, MD[1], Kyungwha Han, PhD[2], Keum Won Kim, MD[3], Won Hwa Kim, MD[4,5], Jaeil Kim, PhD[5,6], Jung Hyun Yoon, MD[1]*

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

Received November 28, 2024
Revised March 26, 2025
Accepted April 21, 2025
Published Online March 17, 2026

***Corresponding author**
Jung Hyun Yoon, MD
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

Purpose To evaluate the benefits of using a commercially available AI system for breast ultrasonography (US) among readers with varying levels of expertise.

Materials and Methods A total of 285 breast lesions from 141 women who underwent breast US between February 2012 and April 2015 were retrospectively analyzed using a deep-learning-based AI system for lesion detection and diagnosis. Five readers, comprising experienced (two breast radiologists and one breast surgeon) and inexperienced (one gynecologist and one radiology resident) groups, reviewed the grayscale US images in two sessions: without AI assistance (session 1) and with AI assistance after a two-week washout period (session 2). Diagnostic performance was compared between sessions.

Results The mean area under the curve for all readers significantly improved with AI, increasing from 0.885 to 0.927 ($p < 0.001$). The inexperienced group demonstrated significant improvements in mean sensitivity (56.9%–87.5%, $p < 0.001$), negative predictive value (NPV) (77.9%–90.1%, $p < 0.001$), and accuracy (76.1%–84.4%, $p = 0.005$). However, no significant

improvements were observed for the experienced readers (all *p*-values > 0.05).

Conclusion The AI system for breast US significantly enhanced the diagnostic performance of inexperienced readers, augmenting sensitivity, NPV, and accuracy, while experienced readers demonstrated minimal improvement, likely due to their already high baseline performance.

Index terms Breast; Cancer; Ultrasonography; Artificial Intelligence; Computer-Assisted Diagnosis

INTRODUCTION

Breast cancer is the most prevalent cancer and the principal cause of cancer-related deaths among women worldwide (1). Breast ultrasonography (US) serves as an essential imaging tool for detecting breast cancer (2) and is particularly valuable for characterizing symptomatic breast masses or lesions identified through physical examination or other imaging modalities (3). Breast US findings were interpreted according to the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS), which offers a standardized sonographic lexicon and final assessment categories aligned with patient management guidelines (4). Despite efforts for standardization, breast US remains highly operator-dependent, with the interpretation of findings varying considerably among readers (5, 6). Moreover, breast US has been criticized for its high false-positive rate, which can lead to unnecessary biopsies and heightened patient anxiety (7, 8). In this era of advanced technology, AI-based computer-aided diagnosis and detection (CAD) systems have been developed to assist clinicians across various medical imaging fields, including breast US. The role of AI-CAD in breast US has been evaluated in several studies, showing enhanced performance among readers, particularly those with less experience (9-13).

Currently, several AI-based CAD systems are commercially available for breast US and are being integrated into real-world clinical practice (14-21). Most AI systems focus on the differential diagnosis of lesions based on a region of interest (ROI) that is manually delineated by the user. However, this reliance on manual ROI selection can critically affect interpretation outcomes, in addition to introducing variability and potential errors depending on user expertise and lesion conspicuity. To minimize potential variability in the lesion detection process, a weakly supervised algorithm for breast US was developed that eliminates the need for readers to designate the area for analysis (22). As this system automatically identifies and visualizes suspicious lesions, the simultaneous detection and diagnosis process aims to offer proactive decision support through the AI system.

In this pilot study, we assessed the impact of a newly developed AI system that provides simultaneous lesion detection and diagnosis on the diagnostic performance of readers with varying levels of experience in breast imaging.

MATERIALS AND METHODS

This retrospective study was approved by the Institutional Review Boards of two institutions (Severance Hospital, IRB No. 1-2022-0045; Konyang University Hospital, IRB No. KYUH

2022-10-043), and the requirement for informed consent was waived.

STUDY POPULATION

The the database of Konyang University Hospital was searched for women who underwent breast US between February 2012 and April 2015. During this period, all images were obtained using a single US machine (iU22; Philips Medical Systems, Bothell, WA, USA). To collect US images of malignant and benign breast lesions as well as normal breast parenchyma, US examinations were selected based on the following criteria: 1) pathologically confirmed breast masses (size ≥5 mm), 2) masses assessed as benign that remained stable for more than 2 years of follow-up, and 3) negative examinations from women with no diagnosis of breast cancer for more than 2 years. Benign features on breast US were defined according to the US lexicon of the BI-RADS, including oval shape, circumscribed margin, parallel orientation, or simple cysts. In total, breast US examinations from 141 women (mean age, 57.0 years; range, 32–86 years) were reviewed for representative US images of breast lesions. Knowing the pathologic outcome, one radiologist (K.W.K.) who was not included in the reader study, selected a representative image of the diagnosed mass with adequate image quality for AI analysis to be included in this study. A total of 285 breast US images were selected, comprising 54 (18.9%) benign masses, 119 (41.8%) malignancies, and 112 (39.3%) normal breast parenchyma.

AI SYSTEM FOR BREAST US

A deep learning-based algorithm (CadAI-B for Breast™, BeamWorks Inc., Daegu, Korea) developed to identify lesions on breast US and provide diagnostic recommendations was employed for this study. This AI system employs a deep neural network for detecting regions suspicious of malignancy (computer-aided detection, CADe) and classifies lesions as malignant and non-malignant (computer-aided diagnosis, CADx). The CADe functionality was designed to assist in the identification of suspicious regions within breast US images by highlighting areas of interest through a relevance map. This map provides a pixel-level abnormality score, which is effectively visualized as an "AI-heatmap." The AI heatmap depicts the likelihood of malignancy (color highlighted from BI-RADS 4A and higher), where increased probabilities are denoted in red. The CADx functionality generates a probability score for malignancy, "CadAI-score," ranging from 0%–100%, and provides a BI-RADS assessment using an AI classifier calibrated to align with the ACR BI-RADS categorization system. For instance, a CadAI-score of 50% was aligned with BI-RADS category 4B. A detailed description of the AI system is provided in Supplementary Material. Real-time examples of AI system usage are illustrated in Supplementary Figs. 1 and 2.

IMAGE REVIEW SESSIONS

For this study, five readers with varying levels of experience in breast US were recruited. To assess the impact of the AI system on all potential users, we included physicians who routinely interpreted breast US images in clinical practice. Of the five readers, three (two board-certified radiologists with 5 and 6 years of experience and one breast surgeon with 12 years of experience) were classified in the experienced group with more than 3 years of experience. The

remaining two readers (one gynecologist and one second-year radiology resident) were classified in the inexperienced group with less than 1 year of experience in using breast US in everyday practice.

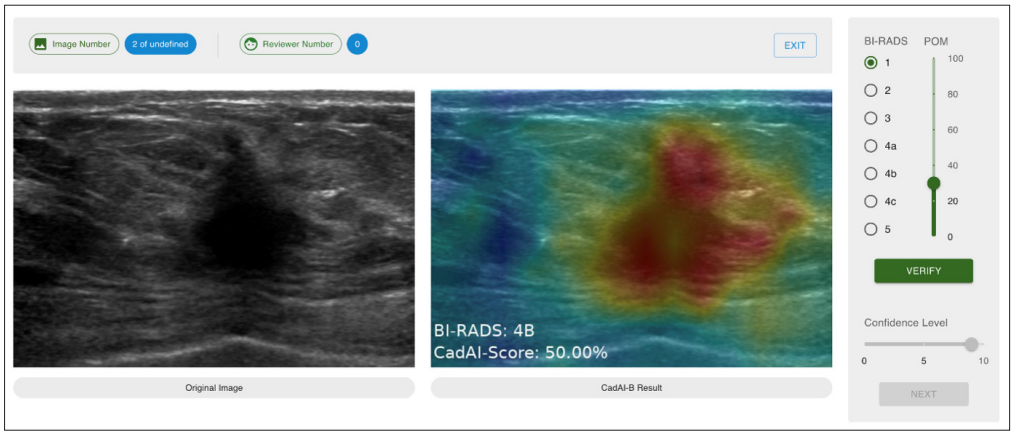
A total of 285 breast US images were uploaded to a web system dedicated to this study. During session 1, the five readers independently reviewed the grayscale US images alone and recorded their interpretations according to the BI-RADS final assessment (categories 1–5, using subcategories 4a, 4b, and 4c) for each US image displayed. After a washout period of 2 weeks, session 2 was initiated, during which the readers re-evaluated the US images, with the AI results displayed concurrently with the grayscale images. After reviewing the original grayscale image and AI results, the readers recorded the final assessment, including the BI-RADS category, probability of malignancy (POM), and confidence level for the AI results (Fig. 1). The confidence level (range: 1–10) reflected the degree of trust that readers placed in the AI results, with 1 indicating the minimal confidence and 10 indicating the maximal confidence.

DATA AND STATISTICAL ANALYSIS

The ground truth for the breast images was established as normal, benign, or malignant based on histopathological findings or clinical outcomes confirmed over a follow-up period exceeding two years. To calculate the positive predictive value (PPV) and negative predictive value (NPV), cases classified as malignant were considered positive, whereas those deemed benign or normal were considered negative.

The area under the receiver operating characteristic (ROC) curve (AUC) was analyzed using a multi-reader multi-case ROC curve model (23). The reader was modeled as a fixed effect, and the jackknife resampling method was employed to estimate the covariance structure within the model. Using BI-RADS 4a, the diagnostic threshold, indices of sensitivity, specificity, PPV, NPV, and accuracy were calculated and compared between the two sessions. Logistic

Fig. 1. Representative image of the web system used for US image review in session 2. For each case, the original grayscale US (left) and the AI analysis result (right) are simultaneously displayed in the web system. AI results are displayed using a semi-transparent colored heatmap overlay, an AI-provided BI-RADS assessment, and an AI-generated probability score for malignancy (CadAI Score) displayed in the bottom left corner. After reviewing the original grayscale US and AI results, readers recorded their assessment using the input system in the upper right corner, according to BI-RADS category, probability of malignancy, and confidence level for the AI results.
BI-RADS = Breast Imaging Reporting and Data System, US = ultrasonography



regression analysis and linear mixed models that included the session as a fixed effect to account for multiple readings per patient were used to compare diagnostic performance and confidence levels, respectively. Further analyses were performed according to experience level (i.e., experienced vs. inexperienced).

Statistical analyses were performed using the R package (version 4.3.0, http://www.R-project.org). $p < 0.05$ was considered statistically significant.

Table 1. Demographic Characteristics of the Patients in the Review of 285 Breast US Images

Characteristics	Values
Age, yrs	55.8 ± 12.3
Mean size of masses*, mm	15.4 ± 10.2
Benign	10.1 ± 4.4
Malignant	17.9 ± 11.1
Presence of symptoms	
Absent	195 (68.4)
Palpable	88 (30.9)
Unknown	2 (0.7)
Final diagnosis (per image)	
Negative	112 (39.3)
Benign	54 (18.9)
Malignancy	119 (41.8)
Pathologic diagnosis	
Benign†	54 (18.9)
Adenosis	3 (1.1)
Atypical ductal hyperplasia	1 (0.4)
Fibroadenoma	10 (3.5)
Fibrocystic change	16 (5.6)
Intraductal papilloma	3 (1.1)
LCIS	1 (0.4)
Stromal fibrosis	7 (2.5)
Others‡	3 (1.1)
Typically benign on US	10 (3.5)
Malignant	119 (41.8)
Carcinoma with apocrine differentiation	4 (1.4)
DCIS	19 (6.7)
Invasive carcinoma, no specific type	5 (1.8)
IDC	88 (30.9)
ILC	2 (0.7)
Mucinous carcinoma	1 (0.4)

Data are presented as *n* (%) or mean ± standard deviation.
*Mean size of 173 benign and malignant masses.
†Among 54 benign lesions, 44 lesions were pathologically confirmed and 10 showed typically benign features on US.
‡Benign "Others" include duct ectasia, adenosis with foam cell collection, fat necrosis and histiocytic reaction, and usual ductal hyperplasia and stromal fibrosis.
DCIS = ductal carcinoma in situ, IDC = invasive ductal carcinoma, ILC = invasive lobular carcinoma, LCIS = lobular carcinoma in situ, US = ultrasonography

RESULTS

Table 1 summarizes the demographic features of the 285 US images from the 141 women included in this study. Among 285 US cases, 112 (39.3%) were classified as negative, 54 (18.9%) as benign, and 119 (41.8%) as malignant. The mean diameter of the 173 benign or malignant masses was 15.4 mm (standard deviation: 10.2 mm, range: 6–58 mm). The diameter of malignant masses was significantly larger than that of benign masses (malignant vs. benign: 17.9 ± 11.1 mm vs. 10.1 ± 4.4 mm, $p < 0.001$). Moreover, 88 (30.9%) of the masses were palpable. Of the 54 benign lesions, 44 were surgically confirmed, including five high-risk lesions, one case of atypical ductal hyperplasia, three of intraductal hyperplasia, and one of lobular carcinoma in situ. Of the 119 malignant lesions, 100 were invasive cancers, meanwhile, 19 were ductal carcinomas in situ.

COMPARISON OF AUCS WITH AND WITHOUT AI SUPPORT

Table 2 and Fig. 2 demonstrate the changes in the AUCs before and after using AI for breast US interpretation. For the two inexperienced readers, the AUCs significantly increased with AI support: 0.753 (95% confidence interval [CI]: 0.699, 0.807) to 0.900 (95% CI: 0.862, 0.938) and 0.865 (95% CI: 0.823, 0.908) to 0.922 (95% CI: 0.889, 0.956) ($p < 0.001$ and $p < 0.003$, respectively). For the three experienced readers, the AUCs did not change significantly when AI was used for US interpretation ($p = 0.214$, 0.825, and 0.796, respectively). Additionally, the AUC for standalone AI was 0.916 (95% CI: 0.881, 0.951).

The mean AUC for all readers improved significantly from 0.885 (95% CI: 0.857, 0.913) to 0.927 (95% CI: 0.901, 0.953) ($p < 0.001$) following the use of AI assistance. In the sub-analysis according to experience level, the mean AUC for the inexperienced group revealed a significant increase after using AI, from 0.809 (95% CI: 0.876, 0.850) to 0.911 (95% CI: 0.877, 0.945) ($p < 0.001$). However, the mean AUC of the experienced group did not exhibit significant dif-

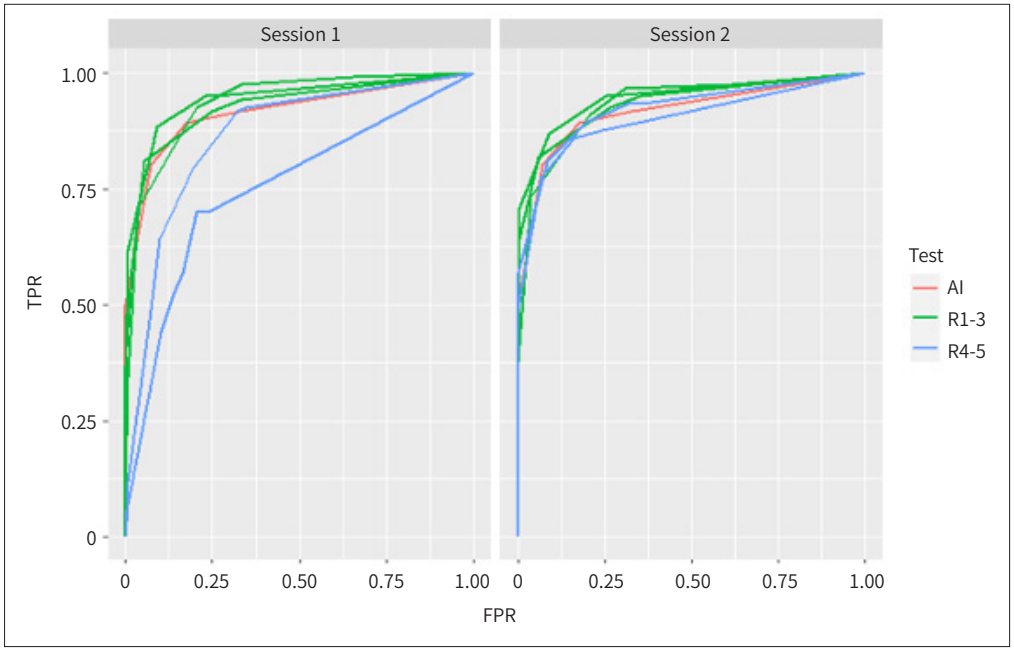
Table 2. Comparison of AUCs for Readers Before and After Using AI in US Interpretation

Reader	AUC		Difference	*p*-Value
	Without AI	With AI		
R1	0.927 (0.895, 0.959)	0.934 (0.904, 0.964)	-0.007 (-0.018, 0.004)	0.214
R2	0.944 (0.917, 0.971)	0.946 (0.919, 0.973)	-0.002 (-0.017, 0.014)	0.825
R3	0.937 (0.910, 0.963)	0.933 (0.904, 0.963)	0.003 (-0.023, 0.029)	0.796
R4	0.753 (0.699, 0.807)	0.900 (0.862, 0.938)	-0.147 (-0.200, -0.095)	<0.001
R5	0.865 (0.823, 0.908)	0.922 (0.889, 0.956)	-0.057 (-0.094, -0.020)	0.003
AI	-	0.916 (0.881, 0.951)	-	-
All readers*	0.885 (0.857, 0.913)	0.927 (0.901, 0.953)	-0.042 (-0.060, -0.024)	<0.001
Inexperienced group†	0.809 (0.768, 0.850)	0.911 (0.877, 0.945)	-0.102 (-0.139, -0.065)	<0.001
Experienced group‡	0.936 (0.912, 0.960)	0.938 (0.914, 0.961)	-0.002 (-0.013, 0.009)	0.749

Data are presented as 95% confidence intervals in parentheses.
*Mean value of all 5 readers.
†Mean value of the 2 inexperienced readers.
‡Mean value of the 3 experienced readers.
AUC = area under the curve, US = ultrasonography

JOURNAL of
THE KOREAN SOCIETY of
RADIOLOGY

Fig. 2. A diagram illustrating the AUC values for the five readers and the AI system. Curves for the experienced readers are displayed in green, inexperienced readers in blue, and AI in red. When compared to the AUC of readers interpreting ultrasonography images without AI (Session 1), the AUC for the inexperienced readers (R4-5, blue line) significantly improved (Session 2), reaching levels comparable to those of the experienced readers (R1-3, green line) using AI.
AUC = area under the curve, FPR = false positive rate, TPR = true positive rate



ferences before and after using AI, 0.936 (95% CI: 0.912, 0.960) vs. 0.938 (95% CI: 0.914, 0.961), respectively (p = 0.749).

DIAGNOSTIC PERFORMANCES WITH AND WITHOUT AI SUPPORT ACCORDING TO EXPERIENCE LEVEL

Table 3 details the diagnostic performances of readers before and after using AI according to experience level. Diagnostic performances of standalone AI are as follows: sensitivity 89.2% (95% CI: 83.6, 94.7), specificity 82.4% (95% CI: 76.6, 88.2), PPV 78.7% (95% CI: 71.8, 85.6), NPV 91.3% (95% CI: 86.7, 95.8), and accuracy 85.3% (95% CI: 81.1, 89.4). The mean sensitivity of all readers significantly increased with AI assistance, rising from 79.5% (95% CI: 74.5, 84.5) to 86.8% (95% CI: 81.9, 91.7) (p = 0.040, Fig. 3), respectively. The inexperienced group demonstrated significant improvements in sensitivity, increasing from 67.9% (95% CI: 61.1, 74.7) to 87.5% (95% CI: 81.8, 93.2), NPV, from 77.9% (95% CI: 72.3, 83.4) to 90.1% (95% CI: 85.4, 94.7), and accuracy, from 76.1% (95% CI: 72.0, 80.3) to 84.4% (95% CI: 80.4, 88.4), after using AI for US interpretation (all p-values <0.05). However, none of the diagnostic indices demonstrated significant differences before and after using AI in the experienced group (all p-values >0.05, respectively).

CONFIDENCE LEVELS FOR THE AI ASSESSMENTS OF US IMAGES

The mean confidence level of all readers in the AI results was 7.96 ± 2.30. For experienced readers, the mean confidence level in the AI results was 7.25 ± 2.33, which was significantly

Table 3. Diagnostic Performances of Readers Before and After Using AI for Breast US Interpretation According to Experience Level

	Without AI	With AI	Difference	*p*-Value
Sensitivity, %				
All readers*	79.5 (74.5, 84.5)	86.8 (81.9, 91.7)	-7.3 (-14.3, -0.3)	0.040
Inexperienced group†	67.9 (61.1, 74.7)	87.5 (81.8, 93.2)	-19.6 (-28.5, -10.7)	<0.001
Experienced group‡	87.2 (82.5, 91.9)	86.4 (81.4, 91.4)	0.8 (-6.1, 7.7)	0.813
Specificity, %				
All readers*	85.8 (81.8, 89.8)	85.7 (81.4, 90.0)	0.1 (-5.7, 6.0)	0.968
Inexperienced group†	82.1 (77.1, 87.2)	82.1 (76.6, 87.7)	0 (-7.5, 7.5)	>0.999
Experienced group‡	88.3 (84.6, 91.9)	88.1 (84.2, 91.9)	0.2 (-5.1, 5.5)	0.941
PPV, %				
All readers*	80.3 (74.4, 86.2)	81.5 (75.7, 87.3)	-1.2 (-9.5, 7.1)	0.770
Inexperienced group†	73.4 (66.0, 80.8)	78.1 (71.3, 84.8)	-4.6 (-14.7, 5.4)	0.365
Experienced group‡	84.4 (79.2, 89.6)	84.1 (78.7, 89.5)	0.3 (-7.2, 7.9)	0.926
NPV, %				
All readers*	85.2 (80.9, 89.5)	89.9 (86.0, 93.9)	-4.8 (-10.6, 1.1)	0.113
Inexperienced group†	77.9 (72.3, 83.4)	90.1 (85.4, 94.7)	-12.2 (-19.4, -4.9)	<0.001
Experienced group‡	90.5 (86.7, 94.3)	89.9 (85.9, 93.9)	0.6 (-4.9, 6.1)	0.837
Accuracy, %				
All readers*	83.2 (80.1, 86.3)	86.2 (83.0, 89.4)	-3.0 (-7.5, 1.5)	0.189
Inexperienced group†	76.1 (72.0, 80.3)	84.4 (80.4, 88.4)	-8.2 (-14.0, -2.4)	0.005
Experienced group‡	87.8 (84.9, 90.8)	87.4 (84.3, 90.4)	0.5 (-3.8, 4.7)	0.829

Data are presented as 95% confidence intervals in parentheses.
*Mean value of all readers.
†Mean value of the inexperienced group.
‡Mean value of the experienced group.
NPV = negative predictive value, PPV = positive predictive value, US = ultrasonography

lower than that of inexperienced readers, 9.03 ± 1.78 ($p < 0.001$). Table 4 summarizes the mean confidence scores according to the final pathological diagnosis and compares them between reader subgroups. The difference in confidence levels between the experienced and inexperienced groups varied significantly according to the final pathology (p for interaction = 0.020). According to the sub-analysis of each final diagnosis, both the experienced and inexperienced groups demonstrated significantly higher confidence for negative images compared to those containing either benign or malignant lesions (all p-values < 0.05). Experienced readers exhibited higher mean confidence levels for malignant lesions than for benign lesions 7.042 (95% CI: 6.794, 7.289) vs. 6.164 (95% CI: 5.791, 6.536), respectively ($p = 0.0001$). Inexperienced readers did not demonstrate a significant difference in mean confidence levels for benign and malignant lesions, 8.425 (95% CI: 7.999, 8.850) vs. 8.891 (95% CI: 8.608, 9.174), respectively ($p = 0.074$).

DISCUSSION

In this pilot study, we evaluated a recently developed CADe/x system for breast US that auto-

matically detects and characterizes lesions according to the ACR BI-RADS final assessments used in our practice, and assessed its impact on readers' diagnostic performance across various experience levels. The mean AUC of the readers significantly increased after AI use

Fig. 3. Diagnostic performances of standalone AI and readers in breast ultrasonography interpretation stratified by AI support and reader experience level. Mean values and 95% confidence intervals (error bars) are displayed. *p < 0.05, †p < 0.001.
NPV = negative predictive value, PPV = positive predictive value

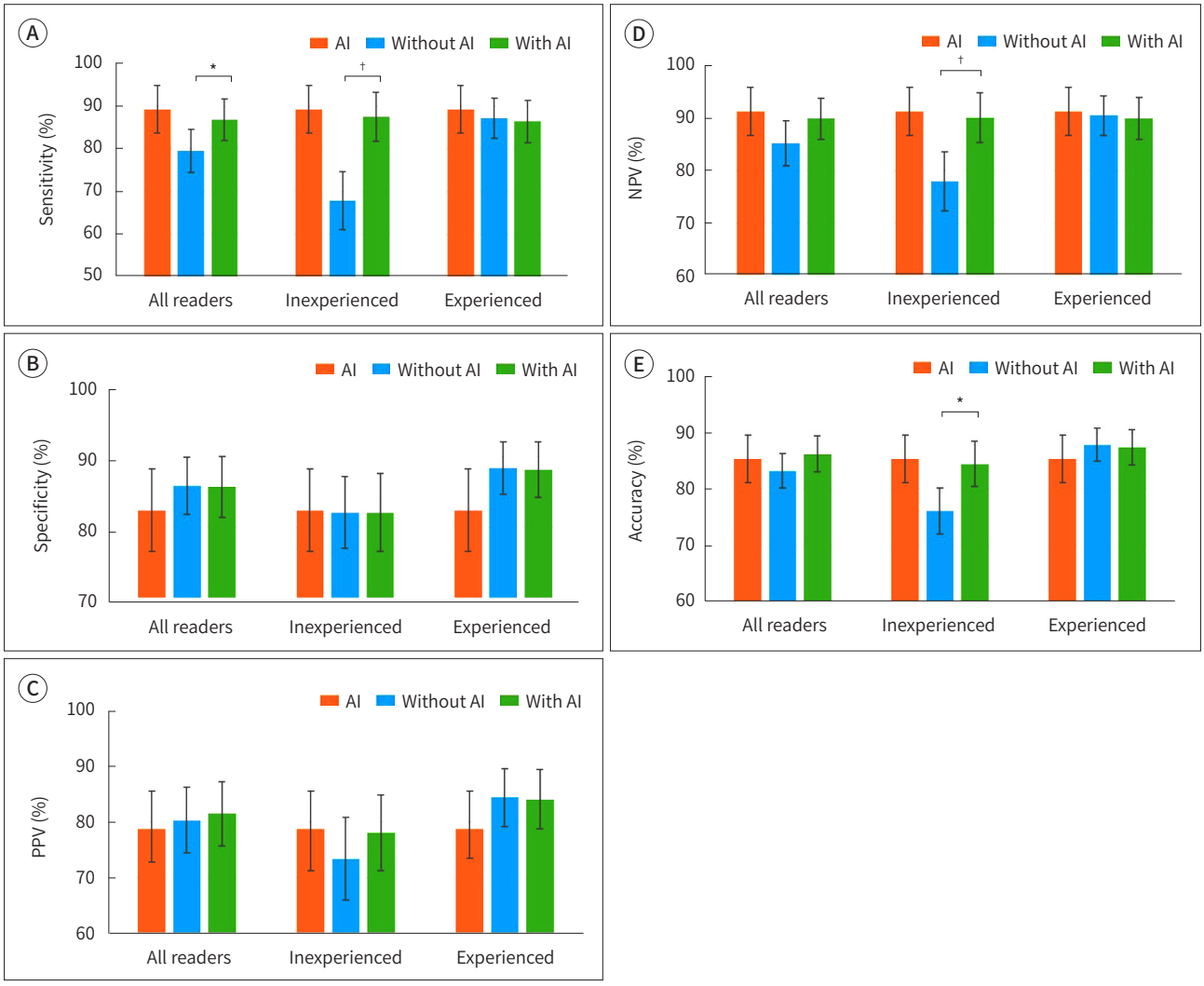


Table 4. Mean Confidence Levels of Readers in AI-Assisted Assessments According to Final Pathology

	Final Diagnosis			p^*	$p^\dagger$	$p^\ddagger$
	Negative	Benign	Malignant			
Inexperienced group	9.469 (9.176, 9.761)	8.425 (7.999, 8.850)	8.891 (8.608, 9.174)	<0.001	0.005	0.074
Experienced group	7.988 (7.732, 8.245)	6.164 (5.791, 6.536)	7.042 (6.794, 7.289)	<0.001	<0.001	<0.001

Data are presented as 95% confidence intervals in parentheses.
*Comparison between negative and benign results.
†Comparison between negative and malignant results.
‡Comparison between benign and malignant results.

(0.885–0.927, $p < 0.001$). When analyzed by the experience level, the AUC of inexperienced readers significantly increased after using AI, whereas the AUC of experienced readers remained unchanged. Inexperienced readers demonstrated significant improvements in mean sensitivity, NPV, and accuracy after using AI for US interpretation; however, in experienced readers, no significant differences were observed in the diagnostic indices with AI support.

The results of our study align with those of previous studies where CAD systems enhanced AUCs specifically for inexperienced readers (12, 14, 15, 24). In our results, the AUCs of the two inexperienced readers significantly improved from 0.753 to 0.900 and 0.865 to 0.922, respectively ($p < 0.001$ and 0.003, respectively), approaching the standalone AI's AUC of 0.916 (Fig. 2). Consistent with the findings of the previous studies, the AUCs of the experienced readers did not change significantly with the use of AI. This aligns with our expectations that AI would have less impact on experienced readers, who had already achieved high diagnostic performance (AUC without AI: 0.927, 0.944, and 0.937, respectively), exceeding the performance of the standalone AI system.

Unlike previous studies that reported improvements in the specificity and accuracy with AI support (9-11), our study demonstrated enhancements in the mean sensitivity, NPV, and accuracy among inexperienced readers. At the same time, no significant changes were observed in any diagnostic indices for experienced readers (Fig. 3, Supplementary Fig. 3). These distinct findings may be attributed to the AI-generated results. The CAD systems described in previous studies presented results in dichotomized categories (possibly benign or possibly malignant) or four-scale classifications (benign, probably benign, suspicious, and probably malignant), with the regions of interest designated either manually by the reader or marked by the system using squares (Supplementary Table 1). Unlike prior CAD systems, the AI system used in this study automatically displays analysis results using colored heatmaps that visually represent the level of suspicion along with the specific "AI-provided BI-RADS" categories, including detailed subcategories within BI-RADS 4 (Fig. 1). This intuitive and visually detailed presentation may have contributed to increased sensitivity by enhancing the readers' ability to recognize subtle findings, especially for those with less experience. Our results suggest that detailed visual integration of BI-RADS categories through an AI-generated heatmap can meaningfully enhance reader sensitivity and improve interpretation quality. Further studies are essential to explore how various display modes may affect the diagnostic performance and interpretation consistency, potentially leading to an optimized CAD design for clinical use.

To assess the value readers attributed to the AI results, we analyzed their confidence levels for each image's AI-generated output. The mean confidence level of experienced readers was significantly lower than that of inexperienced readers (7.25 vs. 9.03, $p < 0.001$), demonstrating that experienced readers relied less on AI results compared to inexperienced readers. When analyzed according to the final pathology, both the experienced and inexperienced groups demonstrated significantly higher confidence levels for negative images than for those with either benign or malignant lesions (all p-values < 0.001) (Table 4). The high NPV of the standalone AI system (91.3%) may explain the elevated confidence levels observed, which likely contributed to the increase in the mean NPV across all readers (Fig. 3D). In addition, experienced readers had significantly higher mean confidence levels for images containing malignant lesions compared to those with benign lesions ($p < 0.001$). This indicates that AI-generat-

ed outputs may serve as complementary tools for experienced readers, particularly to reinforce their assessment of malignant lesions.

Several recent publications have reported three Food and Drug Administration-approved, commercially available AI-CAD systems for handheld breast US (Supplementary Tables 1, 2). Three studies reported that the AUCs of inexperienced readers increased significantly after using AI to interpret breast US images (11, 12, 24). In contrast to these studies, one evaluating performance changes among nine breast radiologists reported no significant differences in the mean AUC in the original AI setting (16). Specifically, AI-CAD systems increase the specificity and accuracy of users, although the degree of improvement varies depending on the experience level (9-12, 14). In addition to enhancing diagnostic performance, studies have demonstrated that using AI reduces the frequency of unnecessary biopsies (12, 13), improves intra- and inter-observer agreement (10, 11, 15, 24), and reduces interpretation time, thereby enhancing work efficiency (14). Based on the results of previous research and our pilot study, AI, irrespective of the specific system used, has the potential to enhance the diagnostic accuracy for inexperienced readers while also improving workflow efficiency in daily practice.

This study has several limitations. First, the study employed a retrospective design using static US images rather than dynamic videos, which did not replicate a true clinical setting. Additionally, the readers were not provided with information such as patient symptoms, risk factors, and results from other imaging modalities. Second, representative US images of pathologically diagnosed masses with adequate image quality were selected from a consecutive study population, which may have caused a selection bias. As this was a pilot study evaluating the effect on readers with varying levels of experience, this limitation was unavoidable, though it may not fully reflect real-world outcomes. Third, the AI analysis results are presented in three formats: AI heatmap, AI score, and AI-provided BI-RADS, all displayed simultaneously to the readers. In the present setting, assessing which information influenced the readers' interpretations the most was difficult. Third, the data were collected from a small number of patients with a high malignancy rate, which lacked generalizability. Future prospective studies with large sample sizes and AI systems integrated into actual clinical environments are required.

This pilot study demonstrated that the AI system for breast US significantly enhanced the diagnostic performance, particularly for inexperienced readers, by improving the sensitivity, NPV, and accuracy. Although experienced readers exhibited minimal improvement due to their high diagnostic accuracy, the AI system's visual outputs may have contributed to improved detection among less experienced readers. Future studies with large sample sizes and integration into real-world clinical workflows are necessary to validate these findings and optimize the implementation of AI in breast US interpretation.

Supplementary Materials

The Supplement is available with this article at http://doi.org/10.3348/jksr.2024.0144.

Author Contributions

Conceptualization, Y.J.H.; data curation, K.K.W., K.W.H., K.J., Y.J.H.; formal analysis, K.J., H.K., Y.J.H.; funding acquisition, Y.J.H.; investigation, Y.J.H.; methodology, H.K., Y.J.H.; project administration, Y.J.H.; resources, Y.J.H.; software, K.W.H., K.J.; supervision, Y.J.H.; validation, Y.J.H.; visualiza-

tion, Y.J.H.; writing—original draft, K.J., K.K.W., K.W.H., K.J., Y.J.H.; and writing—review & editing, K.J., H.K., Y.J.H.

Conflicts of Interest

Authors Won Hwa Kim and Jaeil Kim are the CEOs of BeamWorks Inc., which provided the equipment for this study. The authors did not have the control of the data and information submitted for publication. The remaining authors have declared no conflicts of interest.

ORCID iDs

Jeeyoun Kim https://orcid.org/0009-0009-2104-919X
Kyungwha Han https://orcid.org/0000-0002-5687-7237
Keum Won Kim https://orcid.org/0000-0002-7312-5483
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

Funding

This study was supported by a Korea Medical Device Development Fund grant from 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).

REFERENCES

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, 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. Berg WA, Blume JD, Cormack JB, Mendelson EB, Lehrer D, Böhm-Vélez 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
3. Dempsey PJ. The history of breast ultrasound. *J Ultrasound Med* 2004;23:887-894
4. American College of Radiology. *ACR BI-RADS atlas: breast imaging reporting and data system.* 5th ed. Reston: American College of Radiology 2013
5. 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
6. Lazarus E, Mainiero MB, Schepps B, Koelliker SL, Livingston LS. BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value. *Radiology* 2006;239:385-391
7. Berg WA, Zhang Z, Lehrer D, Jong RA, Pisano ED, Barr RG, et al. Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. *JAMA* 2012;307:1394-1404
8. 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
9. 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
10. Choi JH, Kang BJ, Baek JE, Lee HS, Kim SH. Application of computer-aided diagnosis in breast ultrasound interpretation: improvements in diagnostic performance according to reader experience. *Ultrasonography* 2018;37:217-225
11. 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
12. Wei Q, Zeng SE, Wang LP, Yan YJ, Wang T, Xu JW, 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

13. He P, Chen W, Bai MY, Li J, Wang QQ, Fan LH, et al. Deep learning-based computer-aided diagnosis for breast lesion classification on ultrasound: a prospective multicenter study of radiologists without breast ultrasound expertise. *AJR Am J Roentgenol* 2023;221:450-459
14. 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
15. 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
16. 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
17. Browne JL, Pascual MÁ, Perez J, Salazar S, Valero B, Rodriguez I, et al. AI: can it make a difference to the predictive value of ultrasound breast biopsy? *Diagnostics (Basel)* 2023;13:811
18. Zhao C, Xiao M, Ma L, Ye X, Deng J, Cui 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
19. Wang X, Meng S. Diagnostic accuracy of S-Detect to breast cancer on ultrasonography: a meta-analysis (PRISMA). *Medicine (Baltimore)* 2022;101:e30359
20. Bahl M, Chang JM, Mullen LA, Berg WA. Artificial intelligence for breast ultrasound: AJR expert panel narrative review. *AJR Am J Roentgenol* 2024;223:e2330645
21. Fruchtman Brot H, Mango VL. Artificial intelligence in breast ultrasound: application in clinical practice. *Ultrasonography* 2024;43:3-14
22. 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
23. Hillis SL, Berbaum KS, Metz CE. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. *Acad Radiol* 2008;15:647-661
24. Park HJ, Kim SM, La Yun B, Jang M, Kim B, Jang JY, 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

J Korean Soc Radiol 2026;87(2):314-327

JOURNAL of THE KOREAN SOCIETY of RADIOLOGY

임상 경력에 따른 유방 초음파 인공지능 판독 시스템의 활용 효과를 보기 위한 파일럿 연구

김지연[1] · 한경화[2] · 김금원[3] · 김원화[4,5] · 김재일[5,6] · 윤정현[1]*

목적 임상 경험 수준에 따라 유방 초음파 판독에서 상용화된 인공지능(이하 AI) 시스템 사용의 이점을 비교하였다.

대상과 방법 2012년 2월부터 2015년 4월까지 141명의 여성에서 얻은 285개의 유방 병변을 대상으로 연구를 진행하였다. 딥러닝 기반 AI 시스템으로 병변을 탐지 및 진단하였으며, 숙련자(유방 영상의 2명, 외과의 1명)와 비숙련자(산부인과 의사 1명, 영상의학과 전공의 1명)로 구성된 5명의 판독의가 초음파 이미지를 판독하였다. 첫 번째 세션에서는 AI 없이, 두 번째 세션에서는 AI 판독 결과를 참고하여 동일 이미지를 판독하였다. 각 세션의 판독 진단 성능을 계산 및 비교하였다.

결과 AI 사용 후 판독의 전체의 평균 ROC 곡선하 면적(area under the receiver operating characteristic curve; 이하 AUC)은 0.885에서 0.927로 유의미하게 향상되었다($p < 0.001$). 비숙련자는 민감도(56.9%에서 87.5%), 음성 예측도(77.9%에서 90.1%), 정확도(76.1%에서 84.4%)에서 유의미한 향상을 보였으나, 숙련자는 진단 지표에서 유의미한 차이를 보이지 않았다.

결론 AI 시스템은 비숙련자 판독의들의 AUC, 민감도, 음성 예측도, 정확도를 향상시켰으나, 숙련자 그룹에서는 유의미한 변화가 없었다.

[1]연세대학교 의과대학 세브란스병원 영상의학과, 방사선의과학연구소,
[2]연세대학교 의과대학 영상의학교실, 방사선의과학연구소,
[3]건양대학교 의과대학 건양대학교병원 영상의학과,
[4]경북대학교 의과대학 칠곡경북대학교병원 영상의학과,
[5]빔웍스,
[6]경북대학교 컴퓨터학부