ULTRASONOGRAPHY

Differing benefits of artificial intelligence–based computer-aided diagnosis for breast US according to workflow and experience level

Si Eun Lee[1,2], Kyunghwa Han[3], Ji Hyun Youk[4], Jee Eun Lee[5], Ji-Young Hwang[6], Miribi Rho[1], Jiyoung Yoon[1], Eun-Kyung Kim[1,2], Jung Hyun Yoon[1]

※ Author affiliations appear at the end of this article.

ORIGINAL ARTICLE

https://doi.org/10.14366/usg.22014
pISSN: 2288-5919 • eISSN: 2288-5943
Ultrasonography 2022;41:718-727

Purpose: This study evaluated how artificial intelligence–based computer-assisted diagnosis (AI-CAD) for breast ultrasonography (US) influences diagnostic performance and agreement between radiologists with varying experience levels in different workflows.

Methods: Images of 492 breast lesions (200 malignant and 292 benign masses) in 472 women taken from April 2017 to June 2018 were included. Six radiologists (three inexperienced [<1 year of experience] and three experienced [10–15 years of experience]) individually reviewed US images with and without the aid of AI-CAD, first sequentially and then simultaneously. Diagnostic performance and interobserver agreement were calculated and compared between radiologists and AI-CAD.

Results: After implementing AI-CAD, the specificity, positive predictive value (PPV), and accuracy significantly improved, regardless of experience and workflow (all P<0.001, respectively). The overall area under the receiver operating characteristic curve significantly increased in simultaneous reading, but only for inexperienced radiologists. The agreement for Breast Imaging Reporting and Database System (BI-RADS) descriptors generally increased when AI-CAD was used (κ=0.29–0.63 to 0.35–0.73). Inexperienced radiologists tended to concede to AI-CAD results more easily than experienced radiologists, especially in simultaneous reading (P<0.001). The conversion rates for final assessment changes from BI-RADS 2 or 3 to BI-RADS higher than 4a or vice versa were also significantly higher in simultaneous reading than sequential reading (overall, 15.8% and 6.2%, respectively; P<0.001) for both inexperienced and experienced radiologists.

Conclusion: Using AI-CAD to interpret breast US improved the specificity, PPV, and accuracy of radiologists regardless of experience level. AI-CAD may work better in simultaneous reading to improve diagnostic performance and agreement between radiologists, especially for inexperienced radiologists.

Keywords: Breast neoplasms; Ultrasonography; Diagnosis, Computer-assisted artificial intelligence

Key points: Artificial intelligence-based computer-assisted diagnosis (AI-CAD) can improve the specificity, positive predictive value, and accuracy of radiologists in diagnosing breast cancer on ultrasonography. Inexperienced radiologists may have more benefits in terms of improving the area under the curve. AI-CAD may work better in simultaneous reading to improve diagnostic performance and agreement between radiologists than in sequential reading.

Received: January 28, 2022
Revised: March 11, 2022
Accepted: March 30, 2022

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 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 Oct;41(4):718-727.

Introduction

Ultrasonography (US) is commonly used to evaluate breast abnormalities, especially those that are detected on mammography. Although US has many advantages over mammography, such as being easily available, radiation-free, and cost-effective, it has relatively lower specificity and positive predictive value (PPV) than mammography, which can lead to false-positive recalls and unnecessary biopsies [1]. US examinations and interpretations of US images also rely on the experience level of the examiner and are well-known to be operator-dependent [2]. To overcome observer variability and improve the overall diagnostic performance of breast US, artificial intelligence-based computer-assisted diagnosis (AI-CAD) programs have recently been developed and implemented in clinical practice [3–5].

Several previous studies have demonstrated that the integration of AI-CAD into US improves radiologists' diagnostic performance [6–8], with most US examinations being performed by dedicated breast radiologists from single institutions. However, performers with different training or practice backgrounds and different levels of experience perform and interpret breast US in everyday clinical practice [6,8,9]. To the best of the authors' knowledge, no studies have focused on the analytic results of radiologists from multiple institutions using AI-CAD for breast US. It has also been suggested that diagnostic performance may differ according to the step of US interpretation where AI-CAD is introduced [7], but currently many users refer to AI-CAD arbitrarily, and the stage at which AI-CAD is most effective has not yet been established. Representatively, radiologists may refer to AI-CAD after making a conclusion about a US-detected lesion, which is termed sequential reading, or they may refer to AI-CAD before making a conclusion, which is termed simultaneous reading [7]. Considering that radiologists reach a conclusion by combining various US features, it was hypothesized that the timing of providing AI-CAD results during US image interpretation may have an impact on the final assessment.

Therefore, the purpose of this study was to evaluate and compare diagnostic performance and agreements among radiologists with various levels of training and experience when AI-CAD was used to interpret breast US in different workflows.

Materials and Methods

Compliance with Ethical Standards

This retrospective study was approved by the institutional review board (IRB) of Severance Hospital, Seoul, Korea (1-2019-0027), with a waiver for informed consent.

Data Collection

From April 2017 to June 2018, US images of 639 breast masses in 611 consecutive women were obtained using a dedicated US unit, in which AI-CAD analysis was possible (S-Detect for Breast, Samsung Medison, Co., Ltd., Seoul, Korea). The US images were then reviewed to see if they were of adequate image quality for CAD analysis, and a total of 492 breast lesions (292 benign and 200 malignant masses) in 472 women were finally included for review according to the following indications: (1) masses that were pathologically confirmed with US-guided biopsy or surgery or (2) masses that had been followed for more than 2 years after showing benign features on US (Table 1). The proportion of benign and malignant masses used in preceding research to evaluate the performance of AI-CAD was used to select the 492 breast masses in the present study [10]. The mean age of the 472 women was 49.4±10.1 years (range, 25

Table 1. Clinical characteristics of the 492 breast masses analyzed in this study

	No. (%)
Mean size (mm)	14.2±7.5
0–10	161 (32.7)
10–20	228 (46.3)
≥20	103 (20.9)
Mean age (year)	49.4±10.1
US BI-RADS category	
2	57 (11.6)
3	101 (20.5)
4a	124 (25.2)
4b	22 (4.5)
4c	96 (19.5)
5	92 (18.7)
Pathologic diagnosis	
Benign	292 (59.3)
Stable for more than 2 years	83 (28.4)
Fibroadenoma	99 (33.9)
Fibroadenomatoid hyperplasia	22 (7.5)
Intraductal papilloma	17 (5.8)
Stromal fibrosis	14 (4.8)
Fibrocystic change	13 (4.5)
Others	44 (15.1)
Malignancy	200 (40.7)
Invasive ductal carcinoma	171 (85.5)
Ductal carcinoma *in situ*	14 (7.0)
Invasive lobular carcinoma	11 (5.5)
Tubular carcinoma	4 (2.0)

US, ultrasonography; BI-RADS, Breast Imaging Reporting and Data System.

to 90 years). The mean size of the 492 breast masses was 14.2±7.5 mm (range, 4 to 48 mm). Of the 492 breast masses, 409 breast lesions (83.1%) were pathologically diagnosed with US-guided core-needle biopsy (n=155), vacuum-assisted excision (n=12), and/or surgery (n=242). Eighty-three lesions (16.9%) were included based on typically benign US findings that were stable for more than 2 years.

US images were obtained using a 3-12A linear transducer (RS80A, Samsung Medison, Co., Ltd.). Two staff radiologists (J.H.Y. and E-K. K., 10 and 22 years of experience in breast imaging, respectively) acquired the images. During real-time imaging, representative images of breast masses were recorded and used for the AI-CAD analysis. Images were converted into Digital Imaging and Communications in Medicine files and stored on separate hard drives for individual image analysis. Basic information on the AI-CAD software is provided in Supplementary Data 1.

Experience Level of the Radiologists and Workflow with AI-CAD

For the reader study, three inexperienced radiologists (two radiology residents and one fellow: J.Y., second-year resident; M.R., third-year resident; S.E.L., fellow with less than 1 year of experience in breast imaging) and three experienced breast-dedicated radiologists from different institutions (J.H.Y., J.E.L., and J-Y.H. with 15, 13, and 10 years of experience, respectively) participated in this study. AI-CAD software was set up on each personal computer and each radiologist was initially given 10 separate test images that were not included in the image set for review in order to familiarize themselves with AI-CAD. After the image was displayed with the AI-CAD program, a target point was set at the center of the breast mass by each radiologist, and the program automatically produced a region of interest (ROI) based on the target point. If the ROI was considered inaccurate for analysis by the radiologist, it was adjusted manually. US characteristics according to the Breast Imaging Reporting and Database System (BI-RADS) lexicon, and the final assessments of the masses were automatically analyzed and visualized by the AI-CAD program (Fig. 1). Based on the above data, the AI-CAD program assessed lesions as possibly benign or possibly malignant.

Each radiologist individually evaluated the US images of all 492 breast masses with two separate workflows, sequential reading and simultaneous reading, which took place 4 weeks apart for washout. During sequential reading, each radiologist initially evaluated each of the 492 breast masses according to the BI-RADS lexicon and masses were assigned final assessments from BI-RADS 2 to 5. The radiologists then executed AI-CAD to obtain stand-alone results, which were separately recorded for data analysis. After referring to the analytic results of AI-CAD, each radiologist was asked to reassess the BI-RADS lexicon and final assessment categories, which were also individually recorded for analysis.

During simultaneous reading, radiologists were presented with all 492 images, but in random order, and the AI-CAD results of previous sequential reading were given to the radiologists before image review. As with sequential reading, each radiologist reviewed and recorded data according to the BI-RADS lexicon and final assessments (Fig. 2). Radiologists were blinded to the final pathologic diagnoses of the breast masses and did not have access to clinical patient information or images from mammography or prior US examinations.

Statistical Analysis

The final assessments based on US BI-RADS were divided into two groups for statistical analysis: negative (BI-RADS 2 and 3) and positive (BI-RADS 4a to 5). The diagnostic performance of

A B

Fig. 1. Representative image showing how AI-CAD (S-Detect for Breast) operates.
After the program displays an image for analysis, a target point (green dot on A) is set in the mass center. By clicking the "Calculate" button on the left column of the screen display, a region of interest is automatically drawn along the mass border, with US features (right column) and the final assessment (top blue box) being displayed accordingly (B). AI-CAD, artificial intelligence–based computer-assisted diagnosis.

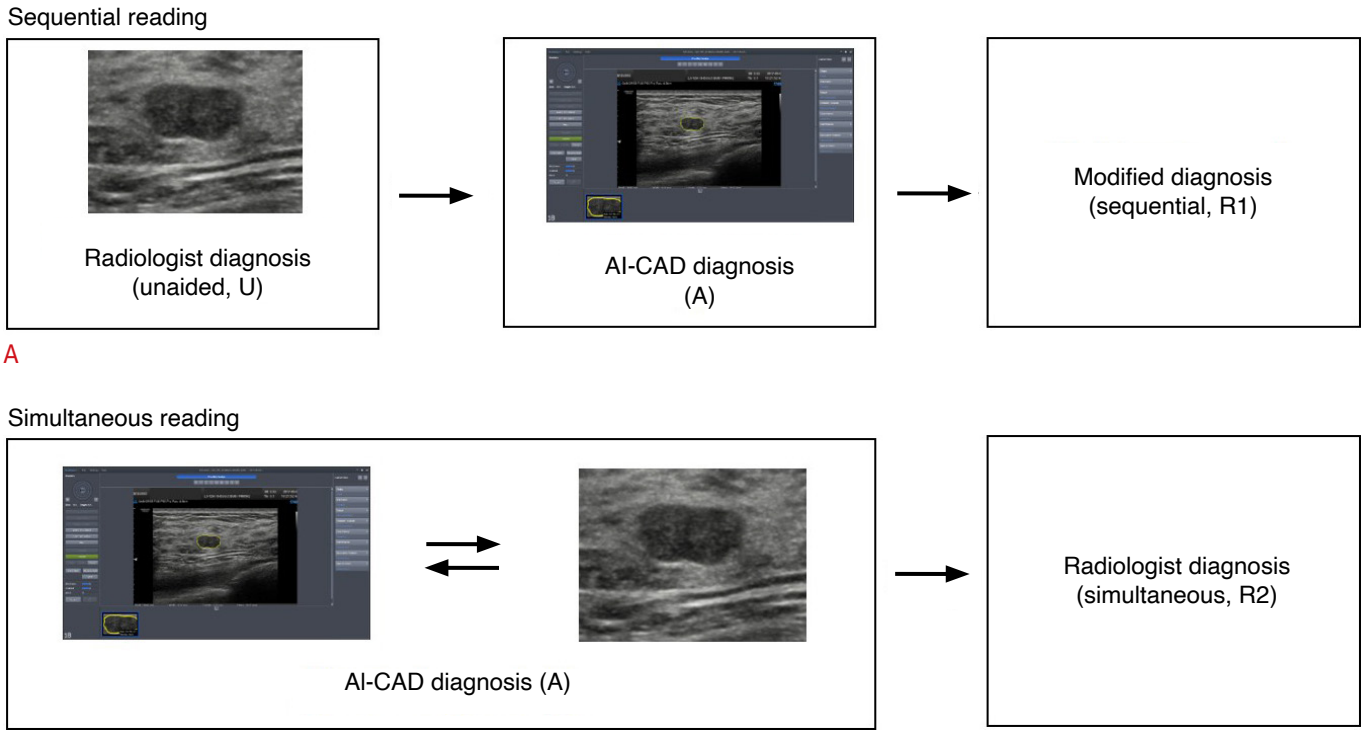


Fig. 2. Schema of the sequential (A) and simultaneous (B) reading workflow.
AI-CAD, artificial intelligence–based computer-assisted diagnosis.

the radiologists without the assistance of AI-CAD, (unaided [U]), with AI-CAD stand-alone (A), and with AI-CAD during sequential reading (R1) and simultaneous reading (R2) was quantified in terms of sensitivity, specificity, PPV, negative predictive value (NPV), and accuracy. Logistic regression with the generalized estimating equation (GEE) method was used to compare diagnostic performance. The area under the receiver operating characteristic curve (AUC) was acquired and compared using the multi-reader multi-case receiver operating characteristic method developed by Obuchowski and Rockette [11]. The conversion rate was defined as the rate of number of changes in final assessments between unaided (U) and aided (R1 and R2, respectively) readings among the total assessments by all six readers, and each subgroup of three inexperienced readers and three experienced readers.

The Fleiss kappa (κ) was calculated to analyze the interobserver agreement between radiologists for US descriptors and final assessments, and the Cohen κ was calculated to analyze the agreement between radiologists and AI-CAD. The κ values were interpreted as follows: 0.00–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, excellent agreement [12]. Logistic regression with the GEE method was used to compare how final assessments changed with the aid of AI-CAD according to each workflow and the radiologists' experience level.

Statistical analyses were performed using SAS version 9.4 (SAS Inc., Cary, NC, USA). All tests were two-sided, and P-values of less than 0.05 were considered to indicate statistical significance.

Results

Diagnostic Performance of Radiologists after Implementation of AI-CAD

Table 2 summarizes the overall diagnostic performance of the six radiologists and AI-CAD. The AI-CAD program itself showed higher specificity (84.9% vs. 56.6%, P<0.001), PPV (79.7% vs. 60.1%, P<0.001), and accuracy (85.4% vs. 72.4%, P<0.001), with lower sensitivity (86.1% vs. 95.4%, P<0.001) and NPV (89.9% vs. 94.7%, P=0.002) than the radiologists. The AUC was lower with AI-CAD than for the unaided radiologists, but without statistical significance (0.855 vs. 0.895, P=0.050). After applying AI-CAD, the specificity, PPV, and accuracy of the radiologists significantly improved in both sequential reading and simultaneous reading (all P<0.001). When simultaneous reading was compared to sequential reading, specificity, PPV, and accuracy were significantly higher in simultaneous reading (all P<0.001) for both the experienced and inexperienced radiologists. The AUC did not significantly

Table 2. Comparison of diagnostic performance between the six radiologists and AI-CAD according to workflow

	Unaided (U)	AI-CAD (A)	Sequential (R1)	Simultaneous (R2)	P-value			
					U vs. A	U vs. R1	U vs. R2	R1 vs. R2
Sensitivity (%)	95.4 (93.0–97.0)	86.1 (80.7–90.1)	95.2 (92.4–97.0)	93.8 (90.7–96.0)	<0.001	0.725	0.087	0.019
Specificity (%)	56.6 (52.2–60.8)	84.9 (80.6–88.4)	61.8 (57.5–65.8)	68.8 (64.7–72.6)	<0.001	<0.001	0.001	<0.001
PPV (%)	60.1 (55.0–64.9)	79.7 (74.0–84.3)	63.0 (58.0–67.8)	67.3 (62.3–72.0)	<0.001	<0.001	0.001	<0.001
NPV (%)	94.7 (91.8–96.7)	89.9 (85.9–92.9)	94.9 (91.9–96.8)	94.2 (91.2–96.3)	0.002	0.817	0.543	0.178
Accuracy (%)	72.4 (69.1–75.4)	85.4 (82.2–88.1)	75.3 (72.2–78.2)	79.0 (76.0–81.6)	<0.001	<0.001	0.001	<0.001
AUC	0.895 (0.854–0.936)	0.855 (0.825–0.886)	0.908 (0.876–0.941)	0.913 (0.886–0.941)	0.050	0.093	0.099	0.394

95% Confidence intervals are given in parentheses.

AI-CAD, artificial intelligence-based computer-assisted diagnosis; U, unaided reading; A, AI-CAD result; R1, sequential reading; R2, simultaneous reading; PPV, positive predictive value; NPV, negative predictive value; AUC, area under the receiver operating characteristic curve.

Table 3. Comparison of diagnostic performance according to experience level and workflow

	Unaided (U)	AI-CAD (A)	Sequential (R1)	Simultaneous (R2)	P-value			
					U vs. A	U vs. R1	U vs. R2	R1 vs. R2
Inexperienced radiologists								
Sensitivity (%)	93.8 (91.4–96.2)	86.2 (81.5–90.8)	93.8 (91.2–96.5)	92.7 (89.7–95.6)	<0.001	0.999	0.344	0.176
Specificity (%)	58.1 (53.7–62.6)	85.1 (81.1–89.0)	63.4 (59.0–67.8)	70.9 (66.6–75.2)	<0.001	<0.001	<0.001	<0.001
PPV (%)	60.5 (55.5–65.6)	79.8 (74.6–85.0)	63.7 (58.7–68.7)	68.6 (63.5–73.6)	<0.001	<0.001	<0.001	<0.001
NPV (%)	93.2 (90.5–96.0)	90.0 (86.5–93.4)	93.8 (91.0–96.5)	93.4 (90.6–96.1)	0.034	0.572	0.887	0.633
Accuracy (%)	72.6 (69.4–75.8)	85.5 (82.5–88.5)	75.8 (72.6–78.9)	79.7 (76.8–82.7)	<0.001	<0.001	<0.001	<0.001
AUC	0.868 (0.804–0.933)	0.856 (0.825–0.887)	0.891 (0.837–0.945)	0.904 (0.868–0.940)	0.540	0.108	0.027	0.176
Experienced radiologists								
Sensitivity (%)	97.0 (95.0–99.0)	86.0 (81.2–90.8)	96.5 (94.4–98.6)	95.0 (92.6–97.4)	<0.001	0.466	0.037	0.027
Specificity (%)	55.0 (50.2–59.9)	84.8 (80.9–88.8)	60.2 (55.6–64.7)	66.7 (62.4–70.9)	<0.001	<0.001	<0.001	<0.001
PPV (%)	59.6 (54.5–64.7)	79.5 (74.3–84.8)	62.4 (57.4–67.4)	66.1 (61.2–71.1)	<0.001	<0.001	<0.001	<0.001
NPV (%)	96.4 (94.0–98.8)	89.8 (86.3–93.4)	96.2 (93.9–98.5)	95.1 (92.7–97.5)	<0.001	0.761	0.195	0.114
Accuracy (%)	72.1 (68.6–75.6)	85.3 (82.3–88.4)	74.9 (71.7–78.2)	78.2 (75.2–81.2)	<0.001	<0.001	<0.001	<0.001
AUC	0.922 (0.892–0.952)	0.854 (0.823–0.885)	0.925 (0.896–0.955)	0.923 (0.884–0.961)	<0.001	0.502	0.913	0.977

95% Confidence intervals are given in parentheses.

AI-CAD, artificial intelligence-based computer-assisted diagnosis; U, unaided reading; A, AI-CAD result; R1, sequential reading; R2, simultaneous reading; PPV, positive predictive value; NPV, negative predictive value; AUC, area under the receiver operating characteristic curve.

improve after AI-CAD was implemented in both the sequential and simultaneous reading workflows, with changes from 0.908 and 0.913 to 0.895, respectively (P=0.093 and P=0.099, respectively).

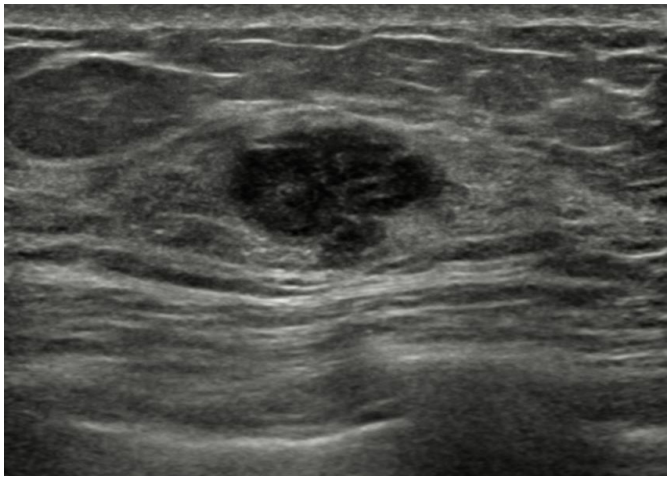
Diagnostic Performance of Radiologists According to Experience Level after Implementation of AI-CAD

When the radiologists were divided according to experience level, specificity, PPV, and accuracy significantly improved in both the experienced and inexperienced groups for both sequential and simultaneous reading (all P<0.05, respectively) (Table 3). For the inexperienced radiologists, the AUC increased from 0.868 to 0.891, but without statistical significance, in sequential reading (P=0.108), while it significantly improved from 0.868 to 0.904 in simultaneous reading (P=0.027) (Fig. 3). For the experienced radiologists, the AUC did not show a significant improvement in either sequential or simultaneous reading (P=0.502 and P=0.913).

As for changes in the final assessments after AI-CAD was integrated into breast US, significantly higher proportions of changes were seen in simultaneous reading than in sequential reading (overall, 40.8% and 16.8%, respectively; P<0.001). Similar trends were seen for both the experienced and inexperienced groups (all P<0.001) (Supplementary Tables 1, 2). Moreover, the proportions of change were more significant in the inexperienced group (experienced 35.4% vs. inexperienced 46.2% in simultaneous reading, P<0.001). The conversion rates for breast masses that were initially BI-RADS 2 or 3 to BI-RADS higher than 4a or vice versa, were also significantly higher in simultaneous reading than in sequential reading (overall, 15.8% to 6.2%, respectively; P<0.001). Similar trends were seen for both the experienced and inexperienced groups (all P<0.001) (Supplementary Tables 1, 2).

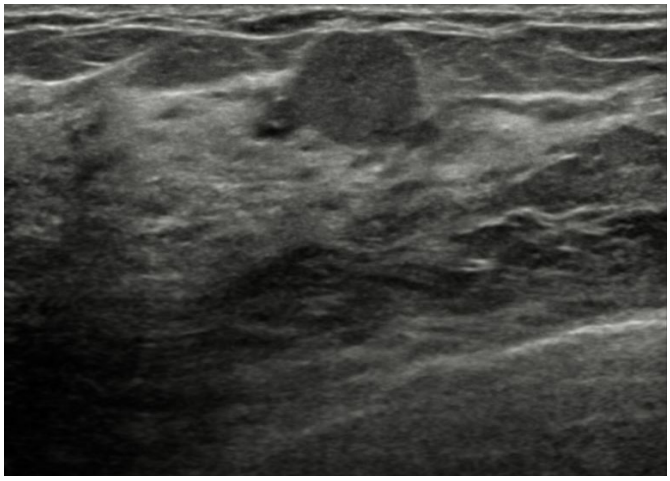
Interobserver Agreement for Descriptors and Assessments According to BI-RADS

Table 4 summarizes the agreement of US descriptors and final assessment categories between radiologists and AI-CAD according to the different workflows. For most descriptors (echogenicity, shape, margin, orientation, and posterior features), agreement between the six radiologists increased regardless of experience level in both



Initial interpretation	BI-RADS 3
AI-CAD	Possibly malignant
Sequential reading	BI-RADS 3
Simultaneous reading	BI-RADS 4a
Final diagnosis	Invasive ductal carcinoma

A



Initial interpretation	BI-RADS 4a
AI-CAD	Possibly benign
Sequential reading	BI-RADS 4a
Simultaneous reading	BI-RADS 3
Final diagnosis	Fibroadenoma

B

Fig. 3. Representative cases of inexperienced readers with each result by sequential and simultaneous reading.
A. US image of a 56-year-old woman diagnosed with a 16-mm invasive ductal carcinoma is shown. The inexperienced reader initially diagnosed the lesion as BI-RADS 3, and did not change the result after referring to the AI-CAD result of "possibly malignant" in sequential reading. However, after the washout period, the reader diagnosed the lesion as BI-RADS 4a based on simultaneous reading with AI-CAD. B. US image of a 47-year-old woman diagnosed with a 12-mm fibroadenoma is shown. An inexperienced reader initially diagnosed the lesion as BI-RADS 4a, and did not change the result after referring to the AI-CAD result of "possibly benign" in sequential reading. However, after the washout period, the reader diagnosed the lesion as BI-RADS 3 based on simultaneous reading with AI-CAD. US, ultrasonography; BI-RADS, Breast Imaging Reporting and Data System; AI-CAD, artificial intelligence-based computer-assisted diagnosis.

sequential and simultaneous reading. Inexperienced radiologists showed better agreement for all BI-RADS descriptors (echogenicity, shape, margin, orientation, and posterior features) in simultaneous reading than in sequential reading (all P<0.001). The agreement for the final assessments significantly increased for both sequential and simultaneous reading in the inexperienced group (P=0.010 and P<0.001, respectively), while significantly lower agreement was seen for both workflows in the experienced group (P=0.042 and 0.023, respectively).

In an analysis of agreement between radiologists and AI-CAD, the agreement for descriptors and final assessments improved in both workflows.

Discussion

The results of the present study show that with the aid of AI-CAD, specificity, PPV, and accuracy significantly improved regardless of radiologists' experience level. These results are consistent with previous studies that also found significantly improved specificity and PPV with the same AI-CAD program [6,8,13–15]. However, the AUC did not significantly improve after AI-CAD was implemented, except in simultaneous reading with inexperienced radiologists. Some earlier studies found significantly improved AUC when AI-CAD was used for breast US, and this was particularly observed when AI-CAD was used to assist inexperienced radiologists [10,13,14], who initially showed significantly lower diagnostic performance than experienced radiologists without AI-CAD. However, the overall AUCs for both the inexperienced and experienced groups in the present study were higher than reported in previous studies (0.868 and 0.922, respectively), which might limit the range of potential improvement after AI-CAD application. This difference from previous studies may be due to the type and number of images selected for review in this study, as previous studies used video clips for image analysis or pre-selected the CAD interpretation results [13,14], whereas the present study used representative still-images of breast masses with the AI-CAD analysis being performed individually by radiologists.

Currently, there are no guidelines on how AI-CAD should be implemented in breast US interpretation. Therefore, this study compared two different workflows: sequential and simultaneous reading. Sequential reading simulates a workflow where radiologists perform and interpret US examinations, while simultaneous reading

Table 4. Agreements for descriptors between radiologists and AI-CAD

BI-RADS lexicons and category	Radiologists	Among radiologists						Between radiologists and AI-CAD				
		Unaided (U)	Sequential (R1)	Simultaneous (R2)	U vs. R1	U vs. R2	R1 vs. R2	Unaided (U)	Sequential (R1)	Simultaneous (R2)	U vs. R1	R1 vs. R2
Echogenicity	Overall	0.47	0.56	0.54	<0.001	<0.001	0.327	0.39	0.68	0.56	<0.001	<0.001
	Inexperienced	0.49	0.54	0.64	0.029	<0.001	0.001	0.41	0.56	0.63	<0.001	<0.001
	Experienced	0.48	0.66	0.46	<0.001	0.733	<0.001	0.36	0.81	0.50	<0.001	<0.001
Shape	Overall	0.59	0.67	0.70	<0.001	<0.001	0.015	0.54	0.81	0.77	<0.001	<0.001
	Inexperienced	0.63	0.72	0.83	<0.001	<0.001	<0.001	0.52	0.79	0.83	<0.001	<0.001
	Experienced	0.61	0.63	0.66	0.330	0.069	0.263	0.55	0.84	0.70	<0.001	<0.001
Margin	Overall	0.29	0.40	0.44	<0.001	<0.001	0.011	0.30	0.66	0.54	<0.001	<0.001
	Inexperienced	0.33	0.43	0.58	<0.001	<0.001	<0.001	0.33	0.61	0.68	<0.001	<0.001
	Experienced	0.32	0.38	0.43	<0.009	<0.001	0.025	0.28	0.71	0.41	<0.001	<0.001
Orientation	Overall	0.63	0.66	0.73	0.065	<0.001	<0.001	0.57	0.81	0.79	<0.001	0.369
	Inexperienced	0.67	0.69	0.85	0.328	<0.001	<0.001	0.61	0.81	0.86	<0.001	<0.001
	Experienced	0.62	0.60	0.65	0.523	0.390	0.074	0.52	0.81	0.72	<0.001	<0.001
Posterior feature	Overall	0.46	0.64	0.72	0.001	<0.001	<0.001	0.46	0.83	0.77	<0.001	<0.001
	Inexperienced	0.37	0.51	0.71	0.001	<0.001	<0.001	0.42	0.71	0.76	<0.001	<0.001
	Experienced	0.54	0.80	0.72	<0.001	<0.001	<0.001	0.51	0.94	0.79	<0.001	<0.001
Final assessment	Overall	0.33	0.37	0.35	0.199	0.007	0.027	0.53	0.64	0.70	<0.001	<0.001
	Inexperienced	0.32	0.39	0.36	0.010	<0.001	0.009	0.55	0.61	0.68	<0.001	<0.001
	Experienced	0.41	0.38	0.37	0.042	0.023	0.344	0.51	0.66	0.73	<0.001	<0.001

AI-CAD, artificial intelligence-based computer-assisted diagnosis; U, unaided reading; R1, sequential reading; R2, simultaneous reading.

simulates a workflow where sonographers perform US examinations first, and interpreting radiologists review the scanned images using the results of AI-CAD analysis. The results of this study suggest that AI-CAD may work better where radiologists interpret scans performed by sonographers, especially for inexperienced radiologists, a finding that clinicians should consider when implementing AI-CAD for breast US in practice. In addition to the clinical workflow, the two workflows can be considered in terms of technical availability: sequential reading simulates using AI-CAD embedded on a picture Archiving and Communication Systems), while simultaneous reading simulates using AI-CAD embedded on US equipment in real-time examinations. How to effectively implement AI-CAD in this workflow is as complex as the heterogeneity of the workflow itself, and along with these results, the authors anticipate that future studies will provide guidelines on how to effectively integrate AI-CAD for breast US according to different workflows.

The results of the present study showed that specificity, PPV, and accuracy were higher in simultaneous reading than in sequential reading, regardless of the radiologists' experience level. In addition, the AUC of the inexperienced radiologists significantly increased in simultaneous reading (0.868 to 0.904, P=0.027). A previous study that compared the two different workflows in breast US using a different AI-CAD platform found results similar to these, in that AI-CAD proved to be more beneficial in simultaneous reading for both experienced and inexperienced radiologists [7]. The differences in performance according to sequential and simultaneous reading may be due to (1) radiologists' less flexible acceptance of contrary results by AI-CAD after they have a certain diagnosis in mind during sequential reading, and (2) the "bandwagon effect," which refers to the tendency to align one's opinion with AI-CAD [16]. These factors may explain the significant improvement of the AUC in simultaneous reading. These findings indicate that the time point at which the AI-CAD results for breast US are made available can affect radiologists' diagnostic performance, and this should be considered for the real-world application of AI-CAD.

In addition to diagnostic performance, significantly higher proportions of change were seen for BI-RADS categories in simultaneous reading than in sequential reading, particularly for radiologists in the inexperienced group. Changes in the final assessment from BI-RADS 2 or 3 to BI-RADS higher than 4a or vice versa are important, as they can lead to critical decisions on whether to perform biopsy. The conversion rates were also significantly higher in simultaneous reading than in sequential reading for both experienced and inexperienced radiologists, suggesting that the type of workflow in which AI-CAD is implemented can also influence the clinical management of patients, as was seen in a previous study [10].

Prior studies have reported considerable variability among radiologists in the evaluation of the US BI-RADS lexicon and final assessments [17]. In this study, six radiologists with various levels of experience in breast imaging showed fair to substantial agreement for descriptors and final assessments, which were in the value ranges suggested by previous studies [17]. The overall agreement for all BI-RADS lexicons and final assessments improved with AI-CAD. Moreover, simultaneous reading with AI-CAD showed higher agreement between radiologists for shape, margin, orientation, posterior features, and the final assessments. However, when radiologists were subgrouped according to experience level, the agreement for most BI-RADS lexicon items did not significantly increase, or even slightly decreased, for the final assessments made by experienced radiologists. The agreement in this study was generally lower than in previous studies, in which AI-CAD improved the agreement between radiologists for final assessments [8,10,13], due to the categorization/subcategorization of BI-RADS 4 and the inclusion of many radiologists from different training backgrounds and institutions.

This study has several limitations. The most notable one is its retrospective data collection from a single institution. However, in order to reflect real-world practice, breast images were selected from a consecutive population according to the benign-malignant ratio and the proportion of BI-RADS final assessments found for real-time US in preceding research using AI-CAD [10]. Second, pre-selected static images of breast masses were analyzed. An analysis of video clips that includes a series of images of the entire breast lesion may result in higher interobserver variability arising from the selection of the representative image. This may affect the diagnostic performance and interobserver agreement in a multi-reader study. Third, the same set of images was used for sequential and simultaneous reading. Although there was a 4-week washout period between the two reading processes, some images may have stuck in the radiologists' memory, and this might have affected their assessments. Last, using the cutoff of BI-RADS 3/4a for a binary classification may have influenced the calculated diagnostic performance, and using different cutoffs may have led to different results.

In conclusion, using AI-CAD to interpret breast US improves the specificity, PPV, and accuracy of radiologists regardless of experience level. More improvements may be seen when AI-CAD is implemented in simultaneous reading through better diagnostic performance and agreement between radiologists, especially for inexperienced radiologists.

ORCID: Si Eun Lee: https://orcid.org/0000-0002-3225-5484; Kyunghwa Han: https://orcid.org/0000-0002-5687-7237; Ji Hyun Youk: https://orcid.org/0000-0002-7787-780X; Jee Eun Lee: https://orcid.org/0000-0003-4451-7661; Ji-Young Hwang:

https://orcid.org/0000-0002-1414-6233; Miribi Rho: https://orcid.org/0000-0002-1703-7657; Jiyoung Yoon: https://orcid.org/0000-0003-2266-0803; Eun-Kyung Kim: https://orcid.org/0000-0002-3368-5013; Jung Hyun Yoon: https://orcid.org/0000-0002-2100-3513

※ Author affiliations
[1]Department of Radiology, Research Institute of Radiological Science, Severance Hospital, Yonsei University College of Medicine, Seoul; [2]Department of Radiology, Research Institute of Radiological Science, Yongin Severance Hospital, Yonsei University College of Medicine, Yongin; [3]Department of Radiology, Research Institute of Radiological Science, Center for Clinical Imaging Data Science, Yonsei University College of Medicine, Seoul; [4]Department of Radiology, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul; [5]Department of Radiology, Ewha Womans University College of Medicine, Seoul; [6]Department of Radiology, Kangnam Sacred Heart Hospital, Hallym University College of Medicine, Seoul, Korea

Author Contributions
Conceptualization: Yoon JH. Data acquisition: Youk JH, Lee JE, Hwang JY, Rho M, Yoon J, Yoon JH. Data analysis or interpretation: Lee SE, Han K, Yoon JH. Drafting of the manuscript: Lee SE. Critical revision of the manuscript: Lee SE, Han K, Youk JH, Lee JE, Hwang JY, Rho M, Yoon J, Yoon JH. Approval of the final version of the manuscript: all authors.

Conflict of Interest
No potential conflict of interest relevant to this article was reported.

Supplementary Material
Supplementary Data 1. Basic information of S-Detect for breast (https://doi.org/10.14366/usg.22014).

Supplementary Table 1. Distribution of changes in final assessments (https://doi.org/10.14366/usg.22014).

Supplementary Table 2. The range of changes in final assessments, calculated as the percentage of number of cases with changed final assessments per total number of cases reviewed by the radiologists (https://doi.org/10.14366/usg.22014).

References

1. 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.
2. Rapelyea JA, Marks CG. Breast imaging. In: Kuzmiak CM, ed. Breast ultrasound past, present, and future. London: IntechOpen, 2017.
3. Lee SE, Han K, Kwak JY, Lee E, Kim EK. Radiomics of US texture features in differential diagnosis between triple-negative breast cancer and fibroadenoma. Sci Rep 2018;8:13546.
4. Akkus Z, Cai J, Boonrod A, Zeinoddini A, Weston AD, Philbrick KA, et al. A survey of deep-learning applications in ultrasound: artificial intelligence-powered ultrasound for improving clinical workflow. J Am Coll Radiol 2019;16:1318-1328.
5. Tanaka H, Chiu SW, Watanabe T, Kaoku S, Yamaguchi T. Computer-aided diagnosis system for breast ultrasound images using deep learning. Phys Med Biol 2019;64:235013.
6. 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.
7. Barinov L, Jairaj A, Becker M, Seymour S, Lee E, Schram A, et al. Impact of data presentation on physician performance utilizing artificial intelligence-based computer-aided diagnosis and decision support systems. J Digit Imaging 2019;32:408-416.
8. Choi JS, Han BK, Ko ES, Bae JM, Ko EY, Song SH, et al. Effect of a deep learning framework-based computer-aided diagnosis system on the diagnostic performance of radiologists in differentiating between malignant and benign masses on breast ultrasonography. Korean J Radiol 2019;20:749-758.
9. Kim K, Song MK, Kim EK, Yoon JH. Clinical application of S-Detect to breast masses on ultrasonography: a study evaluating the diagnostic performance and agreement with a dedicated breast radiologist. Ultrasonography 2017;36:3-9.
10. Bartolotta TV, Orlando A, Cantisani V, Matranga D, Ienzi R, Cirino A, et al. Focal breast lesion characterization according to the BI-RADS US lexicon: role of a computer-aided decision-making support. Radiol Med 2018;123:498-506.
11. Obuchowski NA, Rockette HE Jr. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an anova approach with dependent observations. Commun Stat Simul Comput 1995;24:285-308.
12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.
13. 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.
14. Lee J, Kim S, Kang BJ, Kim SH, Park GE. Evaluation of the effect of computer aided diagnosis system on breast ultrasound for inexperienced radiologists in describing and determining breast lesions. Med Ultrason 2019;21:239-245.
15. 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.

16. Le EP, Wang Y, Huang Y, Hickman S, Gilbert FJ. Artificial intelligence in breast imaging. Clin Radiol 2019;74:357-366.
17. Schwab F, Redling K, Siebert M, Schotzau A, Schoenenberger CA, Zanetti-Dallenbach R. Inter- and intra-observer agreement in ultrasound BI-RADS classification and real-time elastography Tsukuba score assessment of breast lesions. Ultrasound Med Biol 2016;42:2622-2629.