Differing Benefits of Artificial Intelligence-based Computer-aided Diagnosis (AI-CAD) for Breast US According to Workflow and Experience Level

Si Eun Lee
Yonsei University College of Medicine
Kyunghwa Han
Yonsei University College of Medicine
Ji Hyun Youk
Yonsei University College of Medicine
Jee Eun Lee
Ewha Womans University College of Medicine
Ji-Young Hwang
Hallym University College of Medicine
Miribi Rho
Yonsei University College of Medicine
Jiyoung Yoon
Yonsei University College of Medicine
Eun-Kyung Kim
Yonsei University College of Medicine
Jung Hyun Yoon (lvjenny@yuhs.ac)
Yonsei University College of Medicine

Research Article

Keywords: Breast Neoplasms, Ultrasonography, Diagnosis, Computer-Assisted, Artificial Intelligence

Posted Date: March 31st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-351665/v1

Abstract

Background: To evaluate how artificial intelligence-based computer-assisted diagnosis (AI-CAD) for breast ultrasound (US) influences diagnostic performance and agreements between radiologists with varying experience levels in different workflows.

Methods: From Apr 2017 to Jun 2018, images of 492 breast lesions (200 malignant and 292 benign masses) in 472 women were included. Six radiologists (3 inexperienced with < 1 year of experience, 3 experienced with 10-15 years of experience) individually reviewed US images with and without the aid of AI-CAD, first in sequential and then in independent reading. Diagnostic performances and interobserver agreements were calculated and compared between radiologists and AI-CAD.

Results: After implementing AI-CAD, specificity, PPV and accuracy significantly improved, regardless of experience and workflow (all P<0.001, respectively). Overall area under the receiving operator characteristics curve (AUC) significantly increased in independent reading, but only for inexperienced radiologists. Agreements for 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 independent reading (P<0.001). 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 independent reading than sequential reading (overall: 15.8% and 6.2%, respectively, P<0.001) for both inexperienced and experienced radiologists.

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

Trial registration: retrospectively registered

Introduction

Ultrasound (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 values (PPV) compared with mammography that 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 the diagnostic performance of radiologists [6–8], with most of the US examinations being performed by dedicated breast radiologists from single institutions. However, performers with different training or practicing backgrounds and different levels of experience perform and interpret breast US in everyday clinical practice [6, 8, 9]. To the best of our knowledge, no studies have focused on the analytic results of radiologists from multiple institutions using AI-CAD for breast US. Also, it has been suggested that diagnostic performance will differ according to which step of US interpretation that AI-CAD is introduced in [7], but currently many users refer to AI-CAD arbitrarily and the stage at which AI-CAD proves to be most effective has not yet been established. So, we need to verify how AI-CAD for breast US will influence diagnostic performance according to different workflow processes and experience levels.

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

Materials And Methods

This retrospective study was approved by the institutional review board (IRB) of Severance Hospital, Seoul, South Korea, 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, Korea). Then, the US images were 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 being detected with 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 our study [10]. Mean age of the 472 women was 49.4 ± 10.1 years (range, 25–90 years). Mean size of the 492 breast masses was 14.2 ± 7.5 mm (range, 4–48 mm). Of the 492 breast masses, 409 (83.1%) breast lesions were pathologically diagnosed with US-guided core-needle biopsy (n = 155), vacuum-assisted excision (n = 12), and/or surgery (n = 242). Eighty-three (16.9%) lesions were included based on typically benign US findings that were stable for more than 2 years.

Table 1
Demographics of the 492 breast masses analyzed in this study

	N (%)
Mean size (mm)	14.2 ± 7.5
0-10mm	161 (32.7)
10-20mm	228 (46.3)
≥ 20mm	103 (20.9)
Mean age (years)	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

US images were obtained using a 3-12A linear transducer (RS80A, Samsung Medison, Co., Ltd, Korea). 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 DICOM (Digital Imaging and Communications in Medicine) files and stored in separate hard drives for individual image analysis.

Experience level of the radiologists and workflow with AI-CAD

For the reader study, three inexperienced radiologists (2 radiology residents and 1 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. 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 BI-RADS lexicon and 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 made one of two final assessments: ‘possibly benign’ and ‘possibly malignant’.

Each radiologist individually evaluated the US images of the 492 breast masses with two separate workflows, sequential reading and independent 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 independent reading, radiologists were presented with all 492 images but in random order, and the AI-CAD results 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

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). Diagnostic performances of the radiologists without the assistance of AI-CAD, i.e., 'unaided (U)', with AI-CAD stand-alone (A), and with AI-CAD during sequential reading (R1) and independent reading (R2) were calculated for sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy. Logistic regression with the generalized estimating equations (GEE) method was used to compare diagnostic performances. The area under the receiver operating characteristics (ROC) curve (AUC) was acquired and compared using Multi-Reader Multi-Case Receiver Operating Characteristic (MRMC ROC) by Obuchowski-Rockette et al [11]. Fleiss' kappa (κ) was calculated to analyze the interobserver agreement between radiologists for US descriptors and final assessments, and Cohen's 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 experience level of the radiologists.

Statistical analyses were performed using SAS software (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 performances 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%, $P = 0.002$) compared to the radiologists. AUC was lower with AI-CAD than with the 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 independent reading (all $P < 0.001$, respectively). When independent reading was compared to sequential reading, specificity, PPV, and accuracy were significantly higher in independent reading (all $P < 0.001$, respectively) for both experienced and inexperienced radiologists. AUC did not significantly improve after AI-CAD was implemented in both the sequential and independent reading workflows, 0.908 and 0.913 to 0.895, respectively ($P = 0.093$ and 0.099, respectively).

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

(%)	Unaided (U)	AI-CAD (A)	Sequential (R1)	Independent (R2)	*P*			
					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

PPV: positive predictive value, NPV: negative predictive value, AUC: area under the receiving operator characteristics curve, U: unaided reading, A: AI-CAD result, R1: Sequential reading, R2: Independent reading

95% confidence intervals are given in parentheses.

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 independent reading (all $P < 0.05$, respectively, Table 3). For inexperienced radiologists, AUC increased from 0.868 to 0.891 without statistical significance in sequential reading ($P = 0.108$), but significantly improved from 0.856 to 0.891 in independent reading ($P = 0.027$). For experienced radiologists, AUC did not show significant improvement in both sequential and independent reading ($P = 0.502$ and 0.913).

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

Inexperienced radiologists								
(%)	**Unaided (U)**	**AI-CAD (A)**	**Sequential (R1)**	**Independent (R2)**	P			
					U vs. A	**U vs. R1**	**U vs. R2**	**R1 vs. R2**
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

PPV: positive predictive value, NPV: negative predictive value, AUC: area under the receiving operator characteristics curve, U: unaided reading, A: AI-CAD result, R1: Sequential reading, R2: Independent reading

95% confidence intervals are given in parentheses.

As for changes in final assessments after AI-CAD was integrated into breast US, significantly higher rates of changes were seen in independent reading than sequential reading (overall: 40.8% and 16.8%, respectively, $P < 0.001$). Similar trends were seen for both the experienced and inexperienced groups (all $P < 0.001$, Appendix E1 and E2, online). Moreover, the proportions of change were more significant in the inexperienced group (experienced 35.4% vs. inexperienced 46.2% in independent 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 independent reading than sequential reading (overall: 15.8–6.2%, respectively, $P < 0.001$). Similar trends were seen for both the experienced and inexperienced groups (all $P < 0.001$, Appendix E1 and E2, online).

Interobserver agreement for descriptors and assessments according to BI-RADS

Table 4 summarizes the agreements for 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), the agreements between the six radiologists increased regardless of experience level in both sequential and independent reading. Inexperienced radiologists showed better agreement for all BI-RADS descriptors (echogenicity, shape, margin, orientation and posterior feature) in independent reading than sequential reading (all $P < 0.001$). Agreements for final

assessments were significantly increased for both sequential and independent reading in the inexperienced group ($P = 0.010$ and < 0.001, respectively), while significantly lower agreements were seen for both workflows in the experienced group ($P = 0.042$ and 0.023, respectively).

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)	Independent (R2)	U vs. R1	U vs. R2	R1 vs. R2	Unaided (U)	Sequential (R1)	Independent (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.**
	Inexperienced	0.49	0.54	0.64	**0.029**	**< 0.001**	**0.001**	0.41	0.56	0.63	**< 0.001**	**< 0.**
	Experienced	0.48	0.66	0.46	**< 0.001**	0.733	**< 0.001**	0.36	0.81	0.5	**< 0.001**	**< 0.**
Shape	Overall	0.59	0.67	0.70	**< 0.001**	**< 0.001**	**0.015**	0.54	0.81	0.77	**< 0.001**	**< 0.**
	Inexperienced	0.63	0.72	0.83	**< 0.001**	**< 0.001**	**< 0.001**	0.52	0.79	0.83	**< 0.001**	**< 0.**
	Experienced	0.61	0.63	0.66	0.330	0.069	0.263	0.55	0.84	0.7	**< 0.001**	**< 0.**
Margin	Overall	0.29	0.40	0.44	**< 0.001**	**< 0.001**	**0.011**	0.30	0.66	0.54	**< 0.001**	**< 0.**
	Inexperienced	0.33	0.43	0.58	**< 0.001**	**< 0.001**	**< 0.001**	0.33	0.61	0.68	**< 0.001**	**< 0.**
	Experienced	0.32	0.38	0.43	**< 0.009**	**< 0.001**	**0.025**	0.28	0.71	0.41	**< 0.001**	**< 0.**
Orientation	Overall	0.63	0.66	0.73	0.065	**< 0.001**	**< 0.001**	0.57	0.81	0.79	**< 0.001**	0.
	Inexperienced	0.67	0.69	0.85	0.328	**< 0.001**	**< 0.001**	0.61	0.81	0.86	**< 0.001**	**< 0.**
	Experienced	0.62	0.60	0.65	0.523	0.390	0.074	0.52	0.81	0.72	**< 0.001**	**< 0.**
Posterior feature	Overall	0.46	0.64	0.72	**0.001**	**< 0.001**	**< 0.001**	0.46	0.83	0.77	**< 0.001**	**< 0.**
	Inexperienced	0.37	0.51	0.71	**0.001**	**< 0.001**	**< 0.001**	0.42	0.71	0.76	**< 0.001**	**< 0.**
	Experienced	0.54	0.80	0.72	**< 0.001**	**< 0.001**	**< 0.001**	0.51	0.94	0.79	**< 0.001**	**< 0.**
Final assessment	Overall	0.33	0.37	0.35	0.199	**0.007**	**0.027**	0.53	0.64	0.7	**< 0.001**	**< 0.**
	Inexperienced	0.32	0.39	0.36	**0.010**	**< 0.001**	**0.009**	0.55	0.61	0.68	**< 0.001**	**< 0.**
	Experienced	0.41	0.38	0.37	**0.042**	**0.023**	0.344	0.51	0.66	0.73	**< 0.001**	**< 0.**

When we analyzed agreements between radiologists and AI-CAD, the agreements for descriptors and final assessments improved in both workflows.

Discussion

Our results showed that with the aid of AI-CAD, specificity, PPV and accuracy significantly improved regardless of the experience level of the radiologist. Our results were consistent with previous studies that also found significantly improved specificity and PPV with the same AI-CAD program [6, 8, 13-15]. However, AUC did not significantly improve after AI-CAD was implemented, except in independent 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 performances than experienced radiologists without AI-CAD. However, the overall AUCs for both the inexperienced and experienced groups in our study were higher than previous studies (0.868 and 0.922, respectively), which might limit the range for potential improvement. This difference from previous studies may be due to the type and number of images selected for review in our study as the previous studies used video clips for image analysis or pre-selected the CAD interpretation results [13, 14], while we used representative still-images of breast masses with the AI-CAD analysis being performed individually by radiologists.

Currently, no guidelines exist that designate when AI-CAD should be implemented in US interpretation. Thus, we compared two different workflows in this study: sequential and independent reading. For sequential reading, radiologists first assessed lesions without the assistance of AI-CAD and then, were allowed to modify their assessments after the AI-CAD results were made available. In contrast, for independent reading, the AI-CAD results were available to the radiologists from the beginning, and the radiologists assessed the breast masses on US with these results in mind. Our results showed that specificity, PPV and accuracy were higher in independent reading than sequential reading, regardless of the experience level of the radiologist. In addition, the AUC of the inexperienced radiologists significantly increased in independent reading (0.862 to 0.891, P=0.027). A previous study which compared the two different workflows in breast US using a different AI-CAD platform found results similar to ours in that AI-CAD proved to be of better benefit in independent reading for both experienced and inexperienced radiologists, but contrary to our results, this past study showed significantly improved AUC [7]. Based on these findings, we can see that the time point at which the AI-CAD results for breast US are made available can affect the diagnostic performance of radiologists, and this should be considered for the real-world application of AI-CAD.

In addition to diagnostic performance, significantly higher rates of change were seen for BI-RADS categories in independent reading than sequential reading, particularly for radiologists in the inexperienced group. Changes in final assessments 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 or not to perform biopsy. The conversion rates were also significantly higher in independent reading than 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 [16]. In our study, 6 radiologists with variable experience in breast imaging showed fair to substantial agreement for descriptors and final assessments, which were in the value ranges suggested by previous studies [16]. Overall agreements for all BI-RADS lexicons and final assessments improved with AI-CAD. Moreover, independent reading with AI-CAD showed higher agreements between radiologists for shape, margin, orientation, posterior features and final assessments. However, when radiologists were subgrouped according to experience level, agreements for most BI-RADS lexicon did not significantly increase or even slightly decrease for final assessments in experienced radiologists. Agreements in our study were generally lower than previous studies in which AI-CAD improved agreements between radiologists for final assessments [8, 10, 13], possibly due to 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, we selected breast images 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 used for analysis. Analysis of video clips that includes series of images of the entire breast lesion may result in higher interobserver variability arising from selecting the representative image. This may affect the diagnostic performance and interobserver agreement in the multi-reader study. Third, we used the same set of images for sequential and independent 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.

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 independent reading through better diagnostic performance and agreement between radiologists, especially for inexperienced radiologists.

Declarations

1) Ethics approval and consent to participate:

This study was conducted in accordance with the principles expressed in the Declaration of Helsinki. The study protocols were approved by the institutional review board (IRB) of Severance Hospital. IRB also decided to waive the informed consent for this study because it was a retrospective study using anonymized data.

2) Consent for publication:

N/A

3) Availability of data and materials:

The datasets generated and/or analyzed during the current study are not publicly available due to the confidentiality of the data of patient but are available from the corresponding author on reasonable request.

4) Competing interests:

N/A

5) Funding:

This study was supported by a research fund of Samsung Medison, Co., Ltd.

6) Authors' contributions:

JHY and EKK designed and planned the study. JHY, JEL, JYH, MR, JY, and SEL participated in reader study. SEL wrote the main manuscript, and KH performed statistical analysis. All authors read and approved the final manuscript.

7) Acknowledgements:

N/A

Abbreviations

AI-CAD artificial intelligence-based computer-assisted diagnosis

US Ultrasound

AUC area under the receiving operator characteristics curve

PPV positive predictive value

NPV negative predictive value

BI-RADS Breast Imaging Reporting And Data System

ROI region-of-interest

GEE generalized estimating equations

References

1. Berg WA00000, Bandos AI, Mendelson EB, Lehrer D, Jong RA, Pisano ED: **Ultrasound as the Primary Screening Test for Breast Cancer: Analysis From ACRIN 6666**. *Journal of the National Cancer Institute* 2016, **108**(4).
2. Rapelyea JA, Marks CG: **Breast Imaging**. In: *Breast Ultrasound Past, Present, and Future*. Edited by Kuzmiak CM: 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**. *Scientific reports* 2018, **8**(1):13546.
4. Akkus Z, Cai J, Boonrod A, Zeinoddini A, Weston AD, Philbrick KA, Erickson BJ: **A Survey of Deep-Learning Applications in Ultrasound: Artificial Intelligence-Powered Ultrasound for Improving Clinical Workflow**. *Journal of the American College of Radiology : JACR* 2019, **16**(9 Pt B):1318-1328.
5. Tanaka H, Chiu SW, Watanabe T, Kaoku S, Yamaguchi T: **Computer-aided diagnosis system for breast ultrasound images using deep learning**. *Physics in medicine and biology* 2019, **64**(23):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 0110111Radiologists According to Different Levels of Experience**. *Journal of ultrasound in medicine : official journal of the American Institute of Ultrasound in Medicine* 2018, **37**(1):209-216.
7. Barinov L, Jairaj A, Becker M, Seymour S, Lee E, Schram A, Lane E, Goldszal A, Quigley D, Paster L: **Impact of Data Presentation on Physician Performance Utilizing Artificial Intelligence-Based Computer-Aided Diagnosis and Decision Support Systems**. *Journal of Digital Imaging* 2019, **32**(3):408-416.
8. Choi JS, Han BK, Ko ES, Bae JM, Ko EY, Song SH, Kwon MR, Shin JH, Hahn SY: **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 journal of radiology* 2019, **20**(5):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 (Seoul, Korea)* 2017, **36**(1):3-9.
10. Bartolotta TV, Orlando A, Cantisani V, Matranga D, Ienzi R, Cirino A, Amato F, Di Vittorio ML, Midiri M, Lagalla R: **Focal breast lesion characterization according to the BI-RADS US lexicon: role of a computer-aided decision-making support**. *La Radiologia medica* 2018, **123**(7):498-506.
11. Obuchowski NA, Rockette HE: **Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an anova approach with dependent observations**. *Communications in Statistics - Simulation and Computation* 1995, **24**(2):285-308.
12. Landis JR, Koch GG: **The measurement of observer agreement for categorical data**. *biometrics* 1977:159-174.
13. Park HJ, Kim SM, La Yun B, Jang M, Kim B, Jang JY, Lee JY, Lee SH: **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**(3):e14146-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**. *2019* 2019, **21**(3):7.

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 (Seoul, Korea)* 2018, **37**(3):217-225.
16. Schwab F, Redling K, Siebert M, Schötzau A, Schoenenberger C-A, Zanetti-Dällenbach R: **Inter- and Intra-Observer Agreement in Ultrasound BI-RADS Classification and Real-Time Elastography Tsukuba Score Assessment of Breast Lesions**. *Ultrasound in Medicine & Biology* 2016, **42**(11):2622-2629.

Figures

Figure 1

Representative image showing how AI-CAD (S-DetectTM for Breast) operates. After the program displays an image for analysis, a target point (green dot on 1A) 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 (1B).

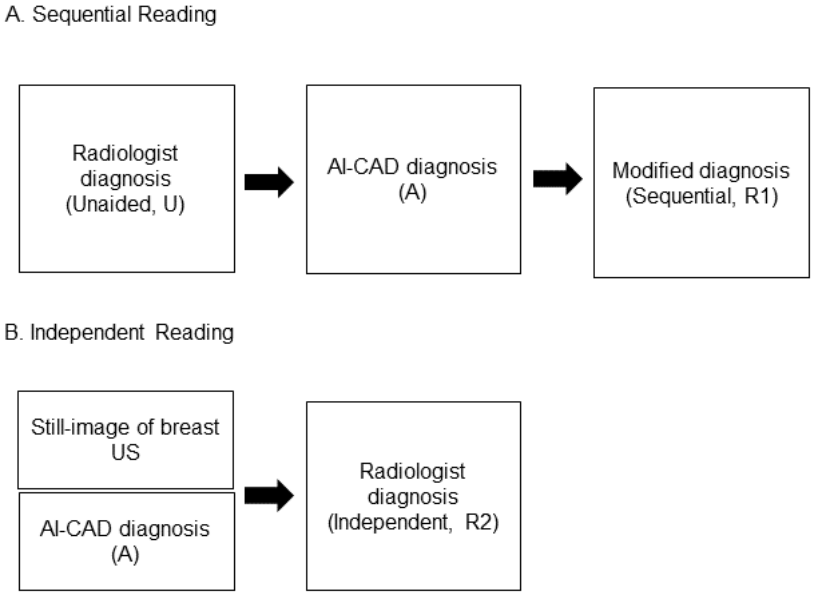

Figure 2

Schema of the sequential and independent reading workflow.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.docx