Original Article | Breast Imaging

eISSN 2005-8330
https://doi.org/10.3348/kjr.2022.0263
Korean J Radiol 2022;23(12):1241-1250

Korean Journal of Radiology
KJR

Use of Artificial Intelligence for Reducing Unnecessary Recalls at Screening Mammography: A Simulation Study

Yeon Soo Kim[1, 2], Myoung-jin Jang[3], Su Hyun Lee[1, 2], Soo-Yeon Kim[1, 2], Su Min Ha[1, 2], Bo Ra Kwon[4], Woo Kyung Moon[1, 2], Jung Min Chang[1, 2]

[1]Department of Radiology, Seoul National University Hospital, Seoul, Korea; [2]Department of Radiology, Seoul National College of Medicine, Seoul, Korea; [3]Medical Research Collaborating Center, Seoul National University, Seoul, Korea; [4]Department of Radiology, Seoul National University Hospital Healthcare System Gangnam Center, Seoul, Korea

Objective: To conduct a simulation study to determine whether artificial intelligence (AI)-aided mammography reading can reduce unnecessary recalls while maintaining cancer detection ability in women recalled after mammography screening.
Materials and Methods: A retrospective reader study was performed by screening mammographies of 793 women (mean age ± standard deviation, 50 ± 9 years) recalled to obtain supplemental mammographic views regarding screening mammography-detected abnormalities between January 2016 and December 2019 at two screening centers. Initial screening mammography examinations were interpreted by three dedicated breast radiologists sequentially, case by case, with and without AI aid, in a single session. The area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and recall rate for breast cancer diagnosis were obtained and compared between the two reading modes.
Results: Fifty-four mammograms with cancer (35 invasive cancers and 19 ductal carcinomas in situ) and 739 mammograms with benign or negative findings were included. The reader-averaged AUC improved after AI aid, from 0.79 (95% confidence interval [CI], 0.74–0.85) to 0.89 (95% CI, 0.85–0.94) (p < 0.001). The reader-averaged specificities before and after AI aid were 41.9% (95% CI, 39.3%–44.5%) and 53.9% (95% CI, 50.9%–56.9%), respectively (p < 0.001). The reader-averaged sensitivity was not statistically different between AI-unaided and AI-aided readings: 89.5% (95% CI, 83.1%–95.9%) vs. 92.6% (95% CI, 86.2%–99.0%) (p = 0.053), although the sensitivities of the least experienced radiologists before and after AI aid were 79.6% (43 of 54 [95% CI, 66.5%–89.4%]) and 90.7% (49 of 54 [95% CI, 79.7%–96.9%]), respectively (p = 0.031). With AI aid, the reader-averaged recall rate decreased by from 60.4% (95% CI, 57.8%–62.9%) to 49.5% (95% CI, 46.5%–52.4%) (p < 0.001).
Conclusion: AI-aided reading reduced the number of recalls and improved the diagnostic performance in our simulation using women initially recalled for supplemental mammographic views after mammography screening.
Keywords: *Artificial intelligence; Mammography; Screening; Breast cancer*

INTRODUCTION

Mammography remains the principal modality for early breast cancer detection in women with an average risk, and mammography screening has been proven to reduce breast cancer incidence by 40% [1,2]. However, false-positive results are a negative aspect of mammography. On average, 9.6% of screened women returned for additional imaging in a review of 5680743 screening examinations in the National Mammography Database [3]. Biopsy is recommended in 2%–3% of screened women [4,5], and approximately 80% of biopsies subsequently performed are benign [6].

Control of the recall rate is important because of its associated medical costs and patient anxiety [7]. Double reading or the use of digital breast tomosynthesis (DBT) decreases false-positive recalls [8,9] but these efforts are associated with increased cost and reading time [10].

Since the 1990s, computer-aided diagnosis (CAD) systems have been developed and commercially used; however, their

Received: April 14, 2022 **Revised:** September 5, 2022
Accepted: September 29, 2022
Corresponding author: Jung Min Chang, MD, PhD, Department of Radiology, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul 03080, Korea.
• E-mail: imchangjm@gmail.com

evidence is controversial. No benefits of CAD have been reported in screening mammography in large-scale studies [11,12], and radiologists are required to review numerous false-positive CAD marks, which worsens the problem of high recall rate in screening mammographies.

In contrast, recent artificial intelligence (AI) systems have better diagnostic sensitivity and specificity and a higher ability to significantly reduce radiologists' workload [13-16]. As most screenings are negative, AI has the potential to increase the cost and time efficiency in clinical practice by reducing false-positive results, unnecessary examinations, radiation exposure, and patient anxiety. However, the use of AI to assist radiologists in assessing mammographic recall has not yet been thoroughly investigated.

Therefore, this study aimed to conduct a simulation study to determine whether AI-aided mammography reading can reduce unnecessary recalls while maintaining the cancer detection ability in women recalled after mammography screening.

MATERIALS AND METHODS

This study was approved by the Institutional Review Board of our institution (IRB No. 2006-159-1135), which waived the requirement for written informed consent because the data were collected retrospectively and analyzed anonymously.

Study Population and Datasets

In our institution, supplemental diagnostic mammography or DBT was recommended and performed for final assessments of mammographic abnormalities reported as the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) [17] categories C0, C4, and C5, when findings were not evident on routine mammography; these procedures were not performed if decision making was possible using routine views or if mammographic abnormalities were identifiable on concurrent ultrasonography (US). A retrospective review of the screening breast examination database at our two screening centers identified 1626 consecutive women who underwent supplemental diagnostic mammographic views or diagnostic DBT for evaluation of mammography-detected abnormalities between January 2016 and December 2019. Women with screening recalls for supplemental diagnostic mammographic views (e.g., spot compression and/or magnification views) or diagnostic DBT were included in this study. We excluded women with a previous surgical history (n = 157) and those who lacked at least one year of follow-up (n = 676). Finally, 793 women (mean age ± standard deviation [SD], 50 ± 9 years; range, 30–78 years) were included in the study (Fig. 1, Table 1).

Image Acquisition and Interpretation

All imaging data were obtained prospectively as part of routine clinical practice and stored in picture archiving and communication system (PACS) software (M-view, INFINITT Healthcare). All mammographic imaging data were acquired using a full-field digital mammography unit (Selenia Dimensions, Hologic; Senographe 2000D; GE Medical Systems). Standard mammography includes bilateral two-view (mediolateral oblique and craniocaudal) mammography.

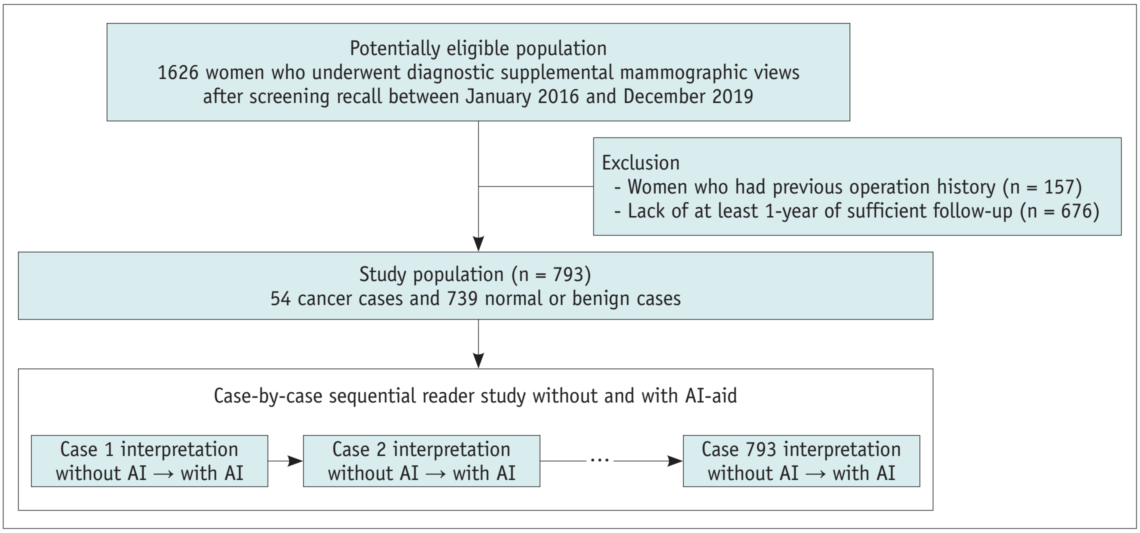
Potentially eligible population
1626 women who underwent diagnostic supplemental mammographic views
after screening recall between January 2016 and December 2019
Exclusion
- Women who had previous operation history (n = 157)
- Lack of at least 1-year of sufficient follow-up (n = 676)
Study population (n = 793)
54 cancer cases and 739 normal or benign cases
Case-by-case sequential reader study without and with AI-aid
Case 1 interpretation
without AI → with AI
Case 2 interpretation
without AI → with AI
...
Case 793 interpretation
without AI → with AI

Fig. 1. Study flow diagram. AI = artificial intelligence

Table 1. Summary of Patient and Lesion Characteristics

Characteristic	Benign (n = 739)	Cancer (n = 54)
Age at screening, years*	49 ± 9 (30–78)	55 ± 10 (32–76)
Breast density[†]		
Almost entirely fatty	8 (1.0)	5 (0.6)
Scattered areas of fibroglandular density	96 (12.1)	9 (1.1)
Heterogeneously dense	364 (45.9)	21 (2.6)
Extremely dense	271 (34.2)	19 (2.4)
Lesion type on screening mammography		
Mass	47 (5.9)	17 (2.1)
Calcification only	357 (45.0)	25 (3.2)
Asymmetry	320 (40.3)	10 (1.3)
Architectural distortion	15 (1.9)	2 (0.3)

Unless otherwise specified, the data are numbers of women with percentages in parentheses. *Data are presented as mean ± standard deviation (range), [†]Breast density was graded according to the American College of Radiology Breast Imaging Reporting and Data System lexicon.

Diagnostic DBT was performed using a full-field digital mammography unit with integrated DBT acquisition (Selenia Dimensions; Hologic). General radiologists interpreted the screening mammography and assessed breast density, image findings, and final assessment categories according to the American College of Radiology BI-RADS.

AI Support System

The AI algorithm used in this study was developed based on deep convolutional neural networks (Lunit INSIGHT MMG, ver. 1.1.4.1; Lunit) using an ResNet-34-based neural network. Details of the development and configuration of the commercially available AI system for breast cancer detection have been described previously [14,18]. The system displays a heat map for areas suspicious for breast cancer in each breast from all four images based on a threshold of 10. It assigns an assessment score for tumor presence of 0–100, where 100 represents the highest level of suspicion. Based on a per-image AI system, four-view heatmaps and a representative abnormality score per breast, which was the maximum abnormality score of the craniocaudal and mediolateral oblique images for each mammography image, were provided. Screening mammography images and AI output results were reviewed using the PACS.

Reader Study with or without AI Aid

Case-by-case sequential reading in the AI-unaided and AI-aided modes was performed by the same radiologists. The radiologist's decision without AI aid was recorded and locked before the AI output was displayed. Subsequently, the AI output was displayed to the radiologists in the same reading session to make a second decision, incorporating the AI results. Radiologists were not allowed to change their data on the AI-unaided reading. Three radiologists with 2, 15, and 11 years of experience in breast imaging independently reviewed all images, with the information that all cases were initially recalled on screening mammography. They were blinded to the prior imaging reports, histological diagnoses, and clinical information.

In the AI-unaided reading, each radiologist reviewed the initial screening mammogram without supplemental mammographic views and determined whether the patient needed to be recalled for supplemental diagnostic mammographic views. In cases of recall decisions, each radiologist assessed the probability of malignancy (POM) of the most suspicious lesion on a 1–100 scale. The location of abnormal findings was marked with an electric indicator, and the images were saved on a PACS. In cases with no recall interpretation, the POM was zero.

During AI-aided reading, each radiologist made a new decision by referring to the AI output results. In cases assigned for recall after the AI-aided reading, each radiologist assessed the POM of the most suspicious lesion on a scale of 1–100. In both sessions, cases assigned for recall (POM ≥ 1) were considered positive screening results and those not assigned for recall were considered negative screening results.

Data Collection and Reference Standard

In the AI reports, maximum pixel-level abnormality scores of 10 points or higher within the corresponding mammographic location of the cancer in at least one view were considered true positives. All markers of the AI program or radiologists at other sites without cancer involvement were considered false positive. Clinical findings and biopsy or surgical results, including tumor size, histological type, nuclear grade, TNM stage, estrogen receptor, progesterone receptor, human epidermal growth factor receptor type 2, and Ki-67 status, were collected from pathology reports (Supplement). Preoperative breast MRI, mammograms obtained after wire localization, or both, were used to determine the reference location of the cancer. More than one year of normal follow-up imaging or histopathologic assessment was used as the reference standard for benign lesions. Two radiologists (with nine years of experience in breast imaging) reviewed the marks reported by the three radiologists and AI software to confirm whether they correctly identified the lesions and classified the radiological manifestations of the recalled cases as mass, calcification only, asymmetry, or architectural distortion in consensus. For the radiologists' assessment, the marks recorded by each radiologist were considered correct if they correctly indicated the corresponding mammographic location in at least one view.

Statistical Analyses

The value was evaluated based on whether the lesion accurately matched recall. In the case of multiple lesions, a positive test result was defined based on the most suspicious lesion. Diagnostic performance was measured by the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and recall rate for all radiologists, with and without AI aid. Performance was evaluated using the AUC and the following measures: sensitivity (the proportion of examinations with a positive assessment among those with a breast cancer diagnosis within the follow-up period), specificity (the proportion of examinations with a negative assessment among those without a breast cancer diagnosis within the follow-up period), and recall rate (the number of examinations with a positive assessment divided by the total number of examinations). To compare the reader-averaged performance of the AI-unaided and AI-aided readings, multireader multicase (MRMC) analysis [19] was used to account for reader variability with the R package MRMCaov [20]. Readers and cases were treated as fixed effects in the MRMC analysis. For reader-specific analysis, the AUC was compared using the DeLong test. Binary measures were compared using the McNemar test, and confidence intervals (CIs) for these binary measures were constructed using the Clopper-Pearson exact binomial CIs. Subgroup analyses were performed according to the breast density and lesion type. Statistical significance was set at $p < 0.05$. All statistical analyses were performed using R (version 3.6.2; R Foundation for Statistical Computing) and SAS (version 9.4; SAS Institute).

RESULTS

Patient and Lesion Characteristics

A total of 793 screening mammograms were assembled, consisting of 54 cancer cases, 182 benign cases with pathological confirmation, and 557 normal cases with normal 1-year follow-up imaging. 54 cancers (median, 1.3

Table 2. Characteristics of the Cancers

Characteristic	Patients with Cancer (n = 54)
Histopathologic type	
IDC	25 (46.3)
DCIS	19 (35.2)
ILC	3 (5.5)
Mixed invasive ductal and lobular carcinoma	2 (3.7)
Microinvasive carcinoma	5 (9.3)
Size, cm*	1.3 (0.1–6.9)
Invasive cancer[†]	1.2 (0.1–5.2)
DCIS	2.0 (0.1–6.9)
TNM stage of cancer[‡]	
T staging[‡]	
pTis	19 (35.2)
pT1	23 (42.6)
pT2	4 (7.4)
pT3	1 (1.9)
N staging[‡]	
pN0	40 (74.1)
pN1	1 (1.9)
pN2	1 (1.9)
M staging	
M0	54 (100.0)
Histologic grade[‡]	
1	4 (7.4)
2	24 (44.4)
3	7 (13.0)
Nuclear grade[‡]	
1	11 (20.4)
2	30 (55.6)
3	11 (20.4)
Hormone status[‡]	
Positive	34 (63.0)
Negative	11 (20.4)
HER2[‡]	
Positive	9 (16.6)
Negative	35 (64.8)
Ki-67[‡]	
Positive	1 (1.9)
Negative	43 (79.6)

Unless otherwise specified, data are presented as the number of women with percentages in parentheses. *Data are presented as medians and ranges, [†]Invasive cancers include IDC, ILC, mixed invasive ductal and lobular carcinoma, and microinvasive carcinoma, [‡]Only patients with available data are presented. DCIS = ductal carcinoma in situ, HER2 = human epidermal growth factor receptor type 2, IDC = invasive ductal carcinoma, ILC = invasive lobular carcinoma cm; range, 0.1–6.9 cm) consisted of 35 invasive cancers (median, 1.2 cm; range, 0.1–5.2 cm) and 19 ductal carcinomas in situ (DCIS) (median, 2.0 cm; range, 0.1–6.9 cm). The clinical and pathological characteristics of 793 women are listed in Tables 1 and 2, respectively.

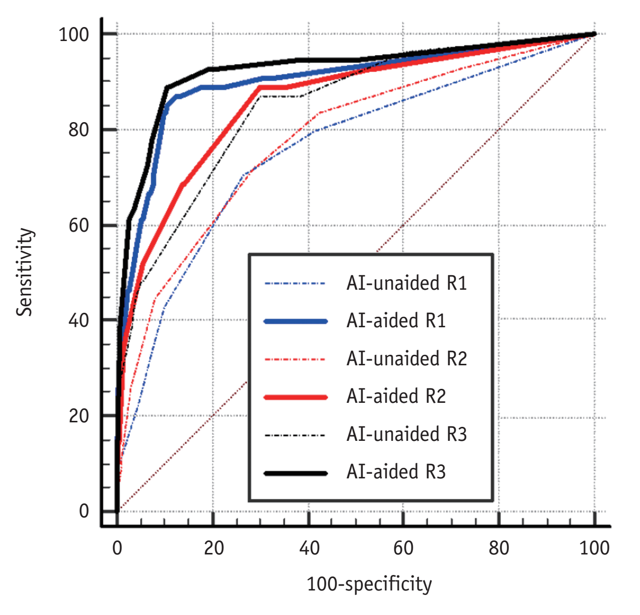

Fig. 2. Receiver operating characteristic curves for AI-unaided and AI-aided reader studies according to each radiologist. AI = artificial intelligence, R = radiologist

Comparison of Diagnostic Performance between AI-Unaided and AI-Aided Readings

Figure 2 shows the receiver operating characteristic (ROC) curve for AI-unaided and AI-aided readings according to each radiologist. The AUC for diagnosing breast cancer ranged from 0.76–0.85 with the AI-unaided reading, and this significantly increased to 0.86–0.92 with the AI-aided reading (p < 0.001 for each reader, Table 3). In the MRMC analysis, which accounted for reader variability, the average AUC also showed a significant increase with an AUC of 0.89 (95% CI, 0.85–0.94) for AI-aided performance compared with AI-unaided performance of 0.79 (95% CI, 0.74–0.85) (p < 0.001).

Sensitivity, Specificity, and Recall Rate

Sensitivity, specificity, and recall rate are presented in Table 3. The average sensitivity was not statistically different between AI-unaided readings and AI-aided readings (89.5% [95% CI, 83.1%–95.9%] vs. 92.6% [95% CI, 86.2%–99.0%]; p = 0.053), although the least

Table 3. Comparison of the AI-Unaided and AI-Aided Reader Studies in Terms of the AUC, Recall Rate, Sensitivity, and Specificity according to Each Radiologist and Average Radiologists

Radiologist	AI-Unaided Reading	AI-Aided Reading	*P*
AUC*			
1	0.76 [0.69–0.83]	0.90 [0.85–0.95]	< 0.001
2	0.78 [0.71–0.85]	0.86 [0.80–0.92]	< 0.001
3	0.85 [0.80–0.90]	0.92 [0.88–0.97]	< 0.001
Average radiologists	0.79 [0.74–0.85]	0.89 [0.85–0.94]	< 0.001
Recall rate†			
1	44.1 (350/793) [40.6–47.7]	36.9 (293/793) [33.6–40.4]	< 0.001
2	73.6 (584/793) [70.4–76.7]	57.1 (453/793) [53.6–60.6]	< 0.001
3	63.3 (502/793) [59.8–66.7]	54.4 (431/793) [50.8–57.9]	< 0.001
Average radiologists	60.4 [57.8–62.9]	49.5 [46.5–52.4]	< 0.001
Sensitivity†			
1	79.6 (43/54) [66.5–89.4]	90.7 (49/54) [79.7–96.9]	0.031
2	92.6 (50/54) [82.1–97.9]	92.6 (50/54) [82.1–97.9]	1.000
3	96.3 (52/54) [87.3–99.5]	94.4 (51/54) [84.6–98.8]	1.000
Average radiologists	89.5 [83.1–95.9]	92.6 [86.2–99.0]	0.053
Specificity†			
1	58.5 (432/739) [54.8–62.0]	67.0 (495/739) [63.5–70.4]	< 0.001
2	27.9 (206/739) [24.7–31.3]	45.7 (338/739) [42.1–49.4]	< 0.001
3	39.4 (291/739) [35.8–43.0]	49.0 (362/739) [45.3–52.7]	< 0.001
Average radiologists	41.9 [39.3–44.5]	53.9 [50.9–56.9]	< 0.001

*Numbers in brackets are the 95% confidence intervals of the AUC values, †Numbers are percentages, raw data are in parentheses, and 95% confidence intervals are in brackets. AI = artificial intelligence, AUC = area under the receiver operating characteristic curve

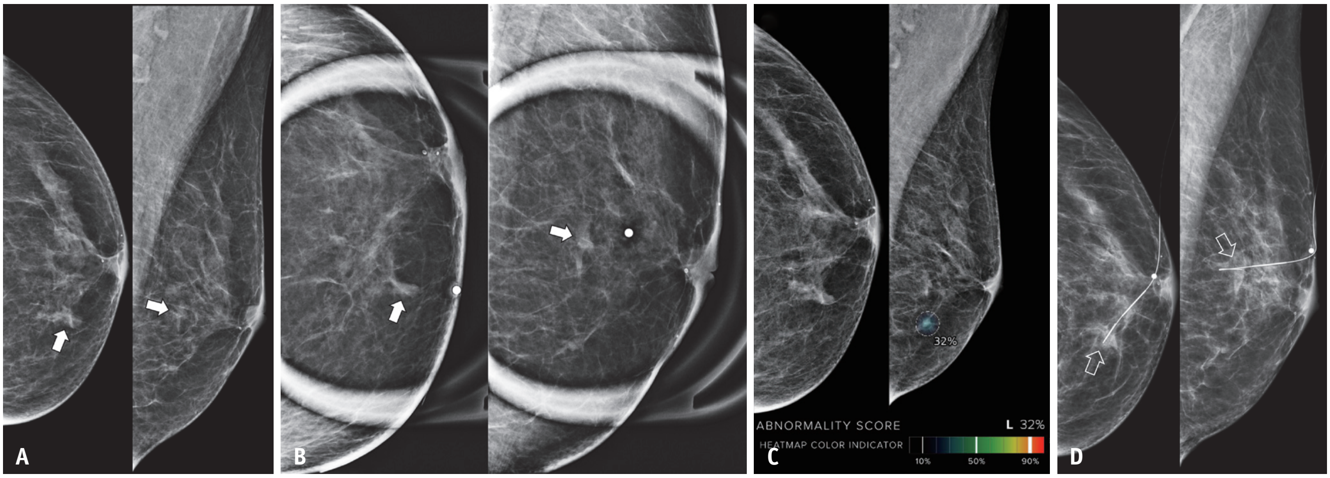

Fig. 3. Images of a 61-year-old woman with ductal carcinoma in situ that was not detected using AI software.
A. Digital mammography screening in a 61-year-old woman with a mass (arrows) in the left upper inner quadrant. **B.** Spot compression magnification images of the mass in the left upper inner quadrant (arrows) at the initial assessment. **C.** AI software assessed the cancer site as normal, but incorrectly identified an asymmetry in the left lower area and assessed a score of 32% on the mediolateral oblique view. **D.** Mammography after ultrasound-guided wire localization revealed true cancer that presented as a mass (empty arrows) in the left upper inner quadrant area recalled in the spot compression magnification view. In the AI-unaided study, one of the three radiologists correctly recalled a mass in the left upper inner quadrant. However, incorrect AI mark led one radiologist to recall the case for an incorrect reason, whereas the other two radiologists missed the cancer on both AI-unaided and AI-aided readings despite the AI result. AI = artificial intelligence

experienced radiologist showed a significant increase in sensitivity (79.6%, 43 of 54 [95% CI, 66.5%–89.4%] vs. 90.7%, 49 of 54 [95% CI, 79.7%–96.9%]; p = 0.031).

Among the 54 cancers, 52 cancers were correctly detected by the AI program, and two cancers (DCISs) were missed by AI. Among the two cancers that were not detected by the AI software, one cancer was also missed by three radiologists, whereas the other cancer was correctly recalled by one radiologist before the AI aid but was not correctly recalled at the lesion level after the AI aid (Fig. 3). In contrast, there was one cancer case that was correctly marked by the AI software but was missed by all three radiologists despite the AI aid. The average specificity was significantly improved in AI-aided reading (41.9% [95% CI, 39.3%–44.5%] vs. 53.9% [95% CI, 50.9%–56.9%]; $p < 0.001$).

All of the radiologists showed significantly decreased recall rate (44.1%–73.6% vs. 36.9%–57.1 %; $p < 0.001$, respectively), with the average recall rate showing significant improvement (60.4% [95% CI, 57.8%–62.9%] vs. 49.5% [95% CI, 46.5%–52.4%]; $p < 0.001$). Of 739 negative or benign cases, all of the radiologists significantly decreased their recall rate after AI aid (41.5%–72.0% vs. 33.0%–54.3%; $p < 0.001$, respectively) (Fig. 4). The average recall rate for negative and benign cases significantly decreased after AI aid (58.1% [95% CI, 55.4%–60.7%] vs. 46.1% [95% CI, 43.1%–49.1%]; $p < 0.001$).

Subgroup Analyses for Performance according to Breast Density and Lesion Types

In both dense and non-dense breasts, the average AUC of dense breasts was significantly increased from 0.79 (95% CI, 0.72–0.85) to 0.89 (95% CI, 0.83–0.95) ($p < 0.001$) and that of non-dense breast was significantly increased from 0.81 (95% CI, 0.71–0.91) to 0.91 (95% CI, 0.81–1.00) (p = 0.007). The average specificities of both dense and non-dense breasts significantly increased after AI aid (42.8% [95% CI, 40.0%–45.7%] vs. 53.9% [95% CI, 50.6%–57.2%]; $p < 0.001$; 36.2% [95% CI, 29.5%–43.0%] vs. 53.8% [95% CI, 45.9%–61.8%]; $p < 0.001$). The detailed sensitivities, specificities, and recall rates are provided in Supplement (Supplementary Table 1).

Regarding lesion types, all of the average AUCs were significantly increased from 0.78 (95% CI, 0.69–0.88) to 0.90 (95% CI, 0.80–0.99) for mass (p = 0.001), from 0.83 (95% CI, 0.76–0.91) to 0.92 (95% CI, 0.88–0.97) for calcification only ($p < 0.001$), and from 0.71 (95% CI, 0.55–0.87) to 0.82 (95% CI, 0.67–0.97) for asymmetry or architectural distortion ($p < 0.001$). In addition, the average recall rates for calcification only and asymmetry or architectural distortion significantly decreased from 64.0% (95% CI, 60.3%–67.6%) to 58.4% (54.1%–62.6%) and from 54.9% (95% CI, 51.1%–58.6%) to 37.2% (95% CI, 32.9%–41.4%) ($p < 0.001$, respectively). The average specificities for calcification only and asymmetry or

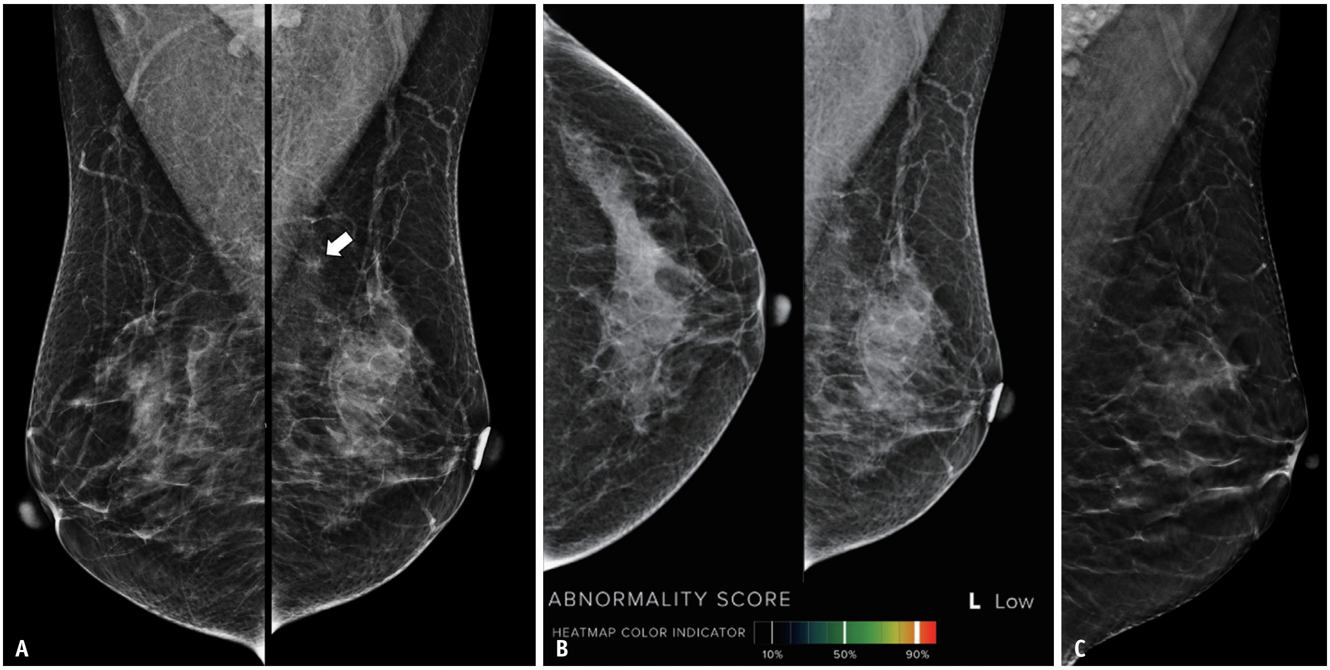

Fig. 4. Images of a 54-year-old woman with asymmetry assessed as negative using AI software.
A. Screening digital mammography showing asymmetry (arrow) in the upper left breast. **B.** The AI software assessed this as negative. **C.** Digital breast tomosynthesis images were obtained for the left breast and no definite lesions were identified at the site of asymmetry on digital mammography. This finding was stable for more than three years. In the AI-unaided reading, all three radiologists recalled this case because of asymmetry; none of them recalled after AI aid. AI = artificial intelligence

architectural distortion significantly increased from 38.1% (95% CI, 34.3%–41.9%) to 44.4% (95% CI, 39.9%–48.8%) and from 46.2% (95% CI, 42.4%–50.0%) to 64.7% (95% CI, 60.4%–68.9%) ($p < 0.001$, respectively). Detailed AUCs, sensitivities, specificities, and recall rates for each lesion type are described in Supplement (Supplementary Table 2).

DISCUSSION

Many efforts have been made to develop AI software to help radiologists interpret screening mammography findings [21,22]. However, the incremental value of AI software in reducing the number of recalls and supplemental diagnostic mammographic examinations while maintaining cancer detection ability remains unclear and requires further investigation. According to our study findings, among 793 women recalled by general radiologists for supplemental views, on average, 60.4% were recalled by dedicated breast radiologists on retrospective re-reading, and this was further lowered to 49.5% by the AI aid without loss of sensitivity in our simulation. The potential of AI to reduce false-positive recalls may provide an efficient way to diagnose negative cases, leading to workload reduction.

While previous studies have shown an overall increased sensitivity with the additional use of AI support in the interpretation of screening mammography [21-23], only the least experienced radiologist in this study showed significantly increased sensitivity with AI aid. Since our study population consisted of women who required additional diagnostic mammographic views after mammography screening, the added value of sensitivity was not observed, except for the least experienced reader. Even lower sensitivity after AI aid was observed in a recent study that evaluated the diagnostic performance of screening recalls in women, which misled radiologists to underdiagnose cancer while reducing unnecessary follow-ups [16]. Similarly, a statistically significant specificity improvement was noted in our study, and the false negative AI result led the reader not to recall the case that was initially recalled without AI aid. Although there was no statistically significant decrease in cancer detection noted by experienced readers, reducing the number of recalls can result in a higher threshold for the perception of small cancers; thus, caution is needed to maximize the added value of AI software. As noted in this study, there was a wide variability in recall rates between readers. In the reader study, 44.1%–73.6% of original recalled cases by general radiologists were recalled by dedicated breast radiologists. Inter-observer variability in the recall rates of screening mammography has been observed in previous studies [24-26]. We expect that AI-CAD will reduce the inter-observer variability in the recall rate by reducing the overall number of recalls.

In this study, among 793 women with screening recalls, 675 (85.1%) had dense breasts. Mammography is generally less sensitive in women with dense breasts, and patients with high fibroglandular tissue volumes have a higher mean number of false-positive mass marks than those with low fibroglandular tissue volumes [27]. An improvement in the average sensitivity was noted in dense breasts, and a reduction in screening recall and improvement in the average AUC and specificity were observed, regardless of breast density. Our results show that AI-CAD can be helpful in the mammography interpretation of dense breast tissue to overcome its limited sensitivity. Further investigation of AI-CAD use in women with dense breasts in larger populations is warranted.

Our study had some limitations. First, this was a retrospective reader study with an enriched cancer population; it used a single AI vendor and data were collected from two sites. Although our reader simulated clinical practice, the results cannot be directly applied to a real screening scenario. Second, this study was conducted by radiologists in only one country. As screening practices, recall rates, diagnostic approaches, and selection of supplemental imaging vary substantially worldwide [28,29], the effect of the AI system on radiologists may vary depending on the geographic region and local policies. Third, we only included patients in whom additional mammographic images were obtained from those who had mammographic abnormalities. Patients who had suspicious mammographic findings but did not undergo additional mammographic views, or who underwent diagnostic US or MRI without supplemental mammographic images were excluded. Consequently, the majority of our cases had subtle mammographic abnormalities, which may have resulted in underestimation of AI. Fourth, the reading environment used in this study was different from that used in the daily practice. There was no restriction on reading time, which might have caused performance differences between clinical and experimental settings [30]. In addition, since the readers were already aware of the study design in which screening mammography recalls were included, the sensitivity of this study would be overestimated. Although the patients in this study were recalled by general radiologists, their performance of the general radiologists after AI aid was not evaluated in this study. Fifth, since the 1–100 scale of cancer probability was assigned in recalled cases instead of the BI-RADS final assessment category, diagnostic performance based on BI-RADS final assessment and changes in management decisions cannot be assessed.

In conclusion, AI-aided reading reduced the number of recalls and improved the diagnostic performance in our simulation using women initially recalled for supplemental mammographic views after mammography screening. Larger prospective population-based screening studies should be performed to validate these findings and evaluate the role of AI aids in reducing additional diagnostic imaging.

Supplement

The Supplement is available with this article at https://doi.org/10.3348/kjr.2022.0263.

Availability of Data and Material

The datasets generated or analyzed during the study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

Author Contributions

Conceptualization: Yeon Soo Kim, Jung Min Chang. Data curation: Yeon Soo Kim, Myoung-jin Jang, Jung Min Chang. Formal analysis: Yeon Soo Kim, Myoung-jin Jang, Jung Min Chang. Funding acquisition: Jung Min Chang. Investigation: Yeon Soo Kim, Myoung-jin Jang, Su Hyun Lee, Soo-Yeon Kim, Su Min Ha, Jung Min Chang. Methodology: Yeon Soo Kim, Myoung-jin Jang, Jung Min Chang. Project administration: Jung Min Chang. Resources: Yeon Soo Kim, Jung Min Chang. Software: Yeon Soo Kim, Jung Min Chang. Supervision: Jung Min Chang. Validation: Yeon Soo Kim, Jung Min Chang. Visualization: Yeon Soo Kim, Jung Min Chang. Writing—original draft: Yeon Soo Kim, Myoung-jin Jang, Jung Min Chang. Writing—review & editing: all authors.

ORCID iDs

Yeon Soo Kim
https://orcid.org/0000-0003-1838-202X
Myoung-jin Jang
https://orcid.org/0000-0001-8123-5001

Su Hyun Lee
https://orcid.org/0000-0002-0171-8060
Soo-Yeon Kim
https://orcid.org/0000-0001-8915-3924
Su Min Ha
https://orcid.org/0000-0002-1833-0919
Bo Ra Kwon
https://orcid.org/0000-0002-7687-8800
Woo Kyung Moon
https://orcid.org/0000-0001-8931-3772
Jung Min Chang
https://orcid.org/0000-0001-5726-9797

Funding Statement

This study was supported by a research grant from DongKook Lifescience Co., Ltd., Republic of Korea, and Seoul National University Hospital (grant no. 06-2020-2840).

REFERENCES

1. Duffy SW, Vulkan D, Cuckle H, Parmar D, Sheikh S, Smith RA, et al. Effect of mammographic screening from age 40 years on breast cancer mortality (UK Age trial): final results of a randomised, controlled trial. *Lancet Oncol* 2020;21:1165-1172
2. Monticciolo DL, Malak SF, Friedewald SM, Eby PR, Newell MS, Moy L, et al. Breast cancer screening recommendations inclusive of all women at average risk: update from the ACR and Society of Breast Imaging. *J Am Coll Radiol* 2021;18:1280-1288
3. Lee CS, Sengupta D, Bhargavan-Chatfield M, Sickles EA, Burnside ES, Zuley ML. Association of patient age with outcomes of current-era, large-scale screening mammography: analysis of data from the National Mammography Database. *JAMA Oncol* 2017;3:1134-1136
4. Løberg M, Lousdal ML, Bretthauer M, Kalager M. Benefits and harms of mammography screening. *Breast Cancer Res* 2015;17:63
5. Blanchard K, Colbert JA, Kopans DB, Moore R, Halpern EF, Hughes KS, et al. Long-term risk of false-positive screening results and subsequent biopsy as a function of mammography use. *Radiology* 2006;240:335-342
6. Lehman CD, Arao RF, Sprague BL, Lee JM, Buist DS, Kerlikowske K, et al. National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. *Radiology* 2017;283:49-58
7. Lee CS, Parise C, Burleson J, Seidenwurm D. Assessing the recall rate for screening mammography: comparing the medicare hospital compare dataset with the national mammography database. *AJR Am J Roentgenol* 2018;211:127-132
8. Conant EF, Barlow WE, Herschorn SD, Weaver DL, Beaber EF, Tosteson ANA, et al. Association of digital breast tomosynthesis vs digital mammography with cancer detection and recall rates by age and breast density. *JAMA Oncol* 2019;5:635-642
9. Marinovich ML, Hunter KE, Macaskill P, Houssami N. Breast cancer screening using tomosynthesis or mammography: a meta-analysis of cancer detection and recall. *J Natl Cancer Inst* 2018;110:942-949
10. Dang PA, Freer PE, Humphrey KL, Halpern EF, Rafferty EA. Addition of tomosynthesis to conventional digital mammography: effect on image interpretation time of screening examinations. *Radiology* 2014;270:49-56
11. Fenton JJ, Taplin SH, Carney PA, Abraham L, Sickles EA, D'Orsi C, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med* 2007;356:1399-1409
12. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 2015;175:1828-1837
13. Dembrower K, Wåhlin E, Liu Y, Salim M, Smith K, Lindholm P, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. *Lancet Digit Health* 2020;2:e468-e474
14. Kim HE, Kim HH, Han BK, Kim KH, Han K, Nam H, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Health* 2020;2:e138-e148
15. Raya-Povedano JL, Romero-Martín S, Elías-Cabot E, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. AI-based strategies to reduce workload in breast cancer screening with mammography and tomosynthesis: a retrospective evaluation. *Radiology* 2021;300:57-65
16. Yi C, Tang Y, Ouyang R, Zhang Y, Cao Z, Yang Z, et al. The added value of an artificial intelligence system in assisting radiologists on indeterminate BI-RADS 0 mammograms. *Eur Radiol* 2022;32:1528-1537
17. D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA. *ACR BI-RADS atlas: breast imaging reporting and data system*. Reston, VA: American College of Radiology, 2013
18. Kim EK, Kim HE, Han K, Kang BJ, Sohn YM, Woo OH, et al. Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study. *Sci Rep* 2018;8:2762
19. Obuchowski Jr NA, Rockette Jr HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an anova approach with dependent observations. *Commun Stat B: Simul Comput* 1995;24:285-308
20. Smith BJ, Hillis SL. Multi-reader multi-case analysis of variance software for diagnostic performance comparison of imaging modalities. *Proc SPIE Int Soc Opt Eng* 2020;11316:113160K

https://doi.org/10.3348/kjr.2022.0263

21. Rodríguez-Ruiz A, Krupinski E, Mordang JJ, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. *Radiology* 2019;290:305-314
22. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577:89-94
23. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst* 2019;111:916-922
24. Klompenhouwer EG, Duijm LE, Voogd AC, den Heeten GJ, Nederend J, Jansen FH, et al. Variations in screening outcome among pairs of screening radiologists at non-blinded double reading of screening mammograms: a population-based study. *Eur Radiol* 2014;24:1097-1104
25. Sharpe RE Jr, Venkataraman S, Phillips J, Dialani V, Fein-Zachary VJ, Prakash S, et al. Increased cancer detection rate and variations in the recall rate resulting from implementation of 3D digital breast tomosynthesis into a population-based screening program. *Radiology* 2016;278:698-706
26. Kim SH, Lee EH, Jun JK, Kim YM, Chang YW, Lee JH, et al. Interpretive performance and inter-observer agreement on digital mammography test sets. *Korean J Radiol* 2019;20:218-224
27. Engelken F, Bremme R, Bick U, Hammann-Kloss S, Fallenberg EM. Factors affecting the rate of false positive marks in CAD in full-field digital mammography. *Eur J Radiol* 2012;81:e844-e848
28. Taylor-Phillips S, Wallis MG, Jenkinson D, Adekanmbi V, Parsons H, Dunn J, et al. Effect of using the same vs different order for second readings of screening mammograms on rates of breast cancer detection: a randomized clinical trial. *JAMA* 2016;315:1956-1965
29. Timmers JM, den Heeten GJ, Adang EM, Otten JD, Verbeek AL, Broeders MJ. Dutch digital breast cancer screening: implications for breast cancer care. *Eur J Public Health* 2012;22:925-929
30. Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty LA, Ganott MA, et al. The “laboratory” effect: comparing radiologists’ performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* 2008;249:47-53