Clinical Application of AI in Mammography: Insights from a Prospective Study

Ebru Yilmaz, MD[1,2] Mustafa Ege Seker, MD[2,3] Nilgun Guldogan, MD[4] Ebru Banu Turk, MD[5] Servet Erdemli, MD[6] Yilmaz Onat Koyluoglu, MD[7] Sehla Nurefsan Sancak, MD[8] Erkin Aribal (Professor)[9]

Rationale and Objectives: This prospective study evaluated the performance of AI in a diagnostic clinic setting, comparing its effectiveness with radiologists of varying experience.

Materials and Methods: The study was conducted at a single center and included 1063 patients undergoing diagnostic or screening mammography. Five radiologists with different experience levels assessed the images using the fifth edition of the BI-RADS lexicon. Standalone AI software assigned risk scores (0–100), with scores above 30.44 considered positive. AI risk assessments were compared with radiologists' BI-RADS scores. Radiologists also re-evaluated AI-positive mammograms as a second look. Ground truth was established through histopathology and two years of follow-up.

Results: Right and left breasts were analyzed separately, and 2126 mammography images were evaluated from 1063 women. A total of 29 cancers were diagnosed in 28 women. Among all examinations, 2.44% (52/2126) were positive, of which 46.15% (24/52) were true positive. Standalone AI detected 82.75% (24/29) of cancers, and the majority voting of radiologists scored positive (BI-RADS 0,4 and 5) in 8% (172/2126) where 89.65% (26/29) of cancers were detected. The AUC score of majority voting was 94.7% (95% CI: 91.1–98.3), and AI was 94.4% (95% CI: 88.5–100). AI was statistically not significantly different than (p=0.79) AUC of the majority voting. The re-evaluation assessment of AI-flagged images achieved an AUC of 94.8% (95% CI: 91.2–98.3), significantly different from the initial evaluation (p=0.015). However, it was not significantly different from AI (p=0.74).

Conclusion: AI algorithms in diagnostic settings can serve as effective CAD systems, aiding in breast cancer detection and reducing inter-reader variability.

Data availability: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Key Words: Breast Cancer; Mammography; AI.

Abbreviations: CAD Computer aided diagnosis, **AI** Artificial Intelligence, **MLO** Mediolateral oblique, **CC** Craniocaudal, **FDA** U.S. Food and Drug Administration, **BI-RADS** Breast Imaging-Reporting and Data System of the American College of Radiology, **ROC** Receiver operating characteristics, **AUC** Area under the curve, **PPV** Positive predictive value, **TP** True positive, **NPV** Negative predictive value, **DCIS** Ductal carcinoma in situ, **IDCa** Invasive ductal carcinoma

Acad Radiol 2025; 32:5016–5027

From the Acibadem Altunizade Hospital, Department of Radiology, Istanbul, Turkey (E.Y., N.G., E.B.T.); University of Wisconsin-Madison, School of Medicine, Department of Radiology, Madison, WI 53792 (M.E.S.); Sultanbeyli Hospital, Department of Radiology, Istanbul, Turkey (S.E.); Acibadem Mehmet Ali Aydinlar University, School of Medicine, Istanbul, Turkey (Y.O.K., S.N.S., E.A.). Received April 6, 2025; revised May 5, 2025; accepted May 12, 2025. **Address correspondence to**: E.A. e-mail: earibal@gmail.com

[1] https://orcid.org/0000–0001-8681–1565
[2] E.Y. and M.E.S. contributed equally
[3] https://orcid.org/0000–0001-7664–5786
[4] https://orcid.org/0000–0001-8322–8374
[5] https://orcid.org/0000–0003-1219–0553
[6] https://orcid.org/0000–0001-5545–768X
[7] https://orcid.org/0000–0002-2776–6650
[8] https://orcid.org/0000–0003-1794–1846
[9] https://orcid.org/0000–0002-5525–8696

https://doi.org/10.1016/j.acra.2025.05.025

INTRODUCTION

Breast cancer is the most common cancer in women. According to 2024 statistical results, the incidence of breast cancer is gradually increasing, but cancer mortality is declining (1). Apart from increased awareness and improvements in treatment, one of the main reasons for the mortality decline is the earlier diagnosis through mammography screening (1). However, the sensitivity of mammography ranges from 80%−98% and reduces to under 70%, as low as 30%−48%, in women with dense breast tissue (2,3). On the other hand, breast cancer miss rate on mammography ranges from 10% to 30% (4). Many reasons can be listed, such as dense parenchyma that obscures a lesion, poor positioning or technique, lesion location outside the field of view, lack of perception of an abnormality present, subtle features of malignancy, or a slowly growing malignancy. Observer errors are another important part of this list (4). A study by Ikeda et al. showed that 10.6% of cases were missing due to observer error (5). Significant causes of radiologist errors are lack of perception and incorrect interpretation (due to lack of experience, fatigue, and inattention) (4). In recent years, double reading (two radiologists or CAD -Computer-aided diagnosis- with radiologists) studies were carried out to reduce observer errors. These studies revealed an increase in breast cancer detection rate by up to 15%. Double reading with two radiologists is not sustainable because of the scarcity of radiologists and economic costs. Also, double reading with a CAD system is not widely used due to increased recall rate and consequent patient anxiety, excessive follow-up, and invasive diagnostic procedures (6–9).

Artificial Intelligence (AI) is a computer system that simulates human intelligence with learning capacity. In recent years the role of AI in medical imaging has rapidly grown and AI has become the latest adjunct tool for cancer detection in breast imaging methods (10,11). Retrospective studies have been conducted on clinical data to validate AI systems in breast cancer screening. One of the most extensive retrospective studies by Larsen et al. demonstrated that AI flagged 80.1% of screen-detected cancers and 30.7% of interval cancers (12). Additionally, Sasaki et al. demonstrated that the area under the curve (AUC) was higher for human readers compared to a standalone AI system (13). This highlights AI's potential to detect more cancers and improve human-based screening when integrated into the screening workflow as a second reader, although some cancers detected by radiologists were missed by AI (12,13). Additional studies further supported the role of AI in screening, showing its contribution to identifying more cancers when used alongside human radiologists (14,15). Moreover, several studies have shown that AI reduces false-negative interpretations without increasing benign biopsy rates (16–19). Recently, prospective studies have started to explore AI's role in screening. One such study showed that replacing one radiologist in double reading with AI led to a 4% higher cancer detection rate, demonstrating the non-inferiority of AI-supported double reading compared to traditional double reading by two radiologists (20,21). Similarly, Lang et al. reported that AI-assisted mammography screening achieved comparable cancer detection rates. These findings provide strong evidence supporting the integration of AI into the mammography screening (20,22).

Most studies have focused on using AI as a second reader or triage tool in screening settings. However, standalone AI usage in diagnostic workflows is not yet approved for clinical practice (23). Prospective diagnostic studies are needed to evaluate the feasibility and effectiveness of standalone AI in mammography interpretation.

AI holds significant promise in diagnostic workflows as a supportive tool for radiologists. In this study, we aimed to prospectively evaluate the performance of AI in a diagnostic clinic setting and explore its implementation by comparing its effectiveness with radiologists across varying experience levels.

MATERIALS AND METHODS

Study Population

This prospective, single-institution study was performed in a tertiary hospital breast clinic between April and July 2022. The study included 1063 patients who applied for diagnostic or screening mammography. Acıbadem University and Acıbadem Healthcare Institutions Medical Research Ethics Committee approved the study, and written informed consent was obtained for each patient (date, number: 15.10.2020, 2020-22/23). The reporting of this study conforms to STROBE guidelines (24).

The study included women over 40 years old. Two views -mediolateral oblique (MLO) and craniocaudal (CC)- were acquired on full-field digital mammography (Senographe Pristina™, GE Healthcare, United States of America).

The exclusion criteria were as follows: cases with technically inadequate mammography (4 cases); patients with implants (50 cases); patients who had only tomosynthesis images (209 cases) and patients with less than 2 years of follow-up (72 cases). The study flowchart is given in Figure 1.

AI System and Score Assessment

We used Lunit INSIGHT MMG version 1.1.7.1 (Lunit, Seoul, South Korea), a commercially available AI-driven mammography interpretation tool that employs convolutional neural network algorithms. This software has received approval from the U.S. Food and Drug Administration (FDA). It analyzes MLO and CC 2D views of each breast, generating a heatmap to indicate potential cancerous lesions. The AI assigns a score ranging from 1 to 100 to each lesion, reflecting the likelihood of malignancy. Our study's dataset was entirely independent of the AI software's development.

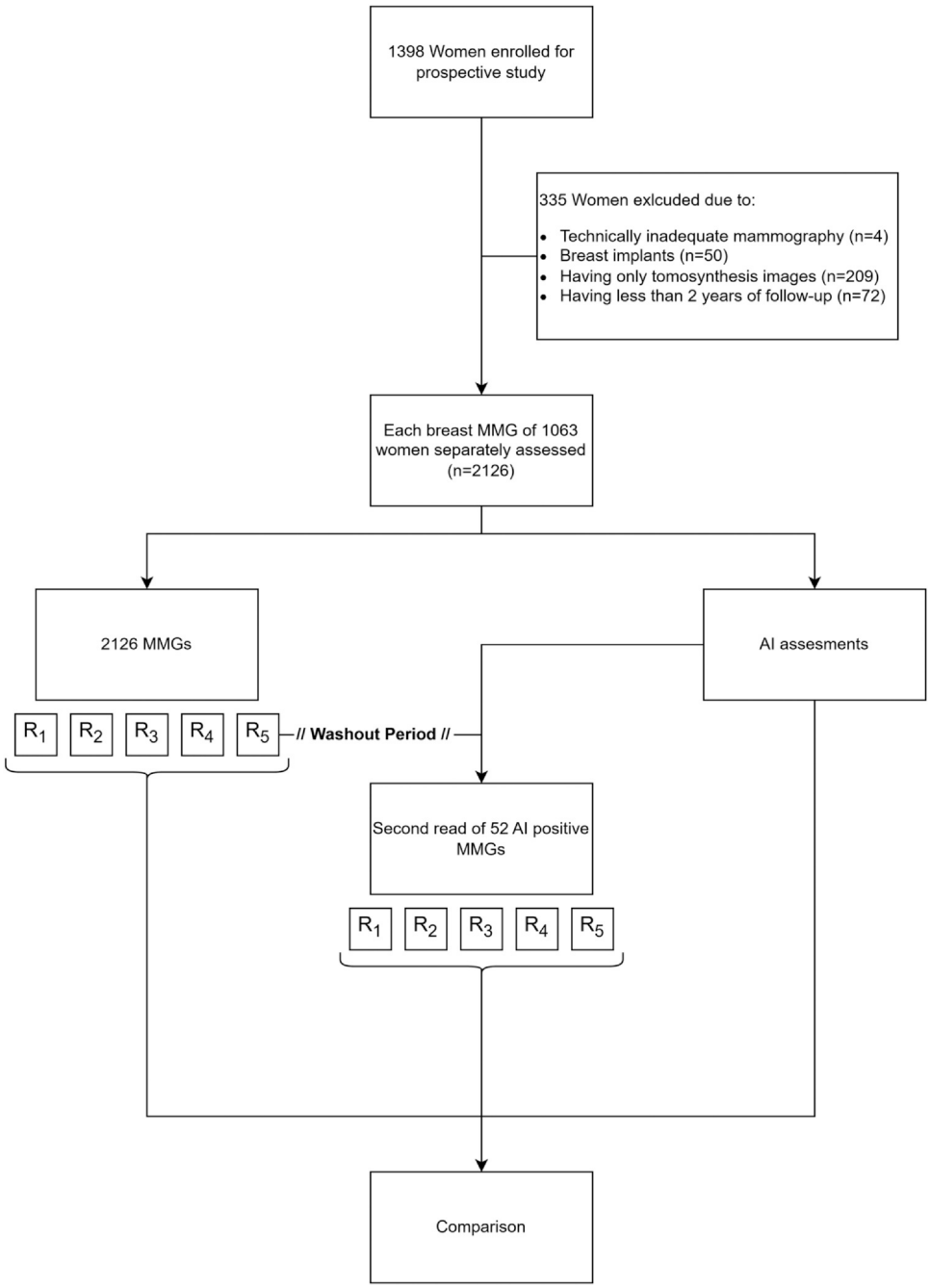

Figure 1. Flowchart of the study. MMG: Mammography, AI: Artificial Intelligence, R_{1-5}: $Radiologist_{1-5}$.

All the mammography images were evaluated by AI software. The software assigned scores between 0 and 100 and a risk score above 30.44 was considered positive (25). The risk assessments of the AI software were compared with the radiologists' Breast Imaging-Reporting and Data System of the American College of Radiology (BI-RADS) scores (26). The BI-RADS scores were dichotomized: BI-RADS 1–2-3 was considered negative, and BI-RADS 0–4-5 was considered positive, as recommended by the BI-RADS lexicon. Ground truth was assessed with histopathology results and two years of follow-up (26).

Image Analysis

Assessments of radiologists and AI were recorded for each breast of the women. Breast cancer ground truths were decided with histopathology assessments and negative cases were decided with 2-year negative follow-up. Five radiologists (three radiologists with 5–10 years of experience, one resident, and one with more than 20 years of breast imaging experience in breast radiology) evaluated the images blinded. Radiological findings were evaluated under the guidance of the fifth edition of BI-RADS (26). The

TABLE 1. Demographic Characteristics Distribution by Breast Densities

Breast Density	Type A	Type B	Type C	Type D	P Value	All Breast Compositions
Number of Women, (%)	70 (6.6%)	278 (26.2%)	616 (57.9%)	99 (9.3%)	-	1063
Age, *year*[*] Mean	59.5 (±10.8)	56.4 (±11.5)	52.8 (±12.3)	50.5 (±13.8)	<0.001[a]	54 (±12.4)
Median	58.5 (15.5)	55 (17)	49 (15)	45 (12)		50 (18)
40–49	15 (2.9%)	96 (18.5%)	339 (65.3%)	69 (13.3%)	-	519
50–59	23 (9.4%)	85 (34.7%)	128 (52.2%)	9 (3.7%)	-	245
60–69	17 (12.6%)	53 (39.2%)	61 (45.2%)	4 (3%)	-	135
≥70	15 (9.1%)	44 (26.8%)	88 (53.7%)	17 (10.4%)	-	164

[*] First line is mean (± standard deviation), second line is median (interquartile range)
[a] Differences between Type B-C-D were assessed with Kruskal-Wallis test

evaluation was made only with current mammography images; no additional information, such as prior mammograms, clinical findings, or breast cancer risk assessment, was included. Also, each radiologist underwent assessments solely without additional imaging such as ultrasonography, tomosynthesis, magnification, or AI. After sole mammography evaluation responsible radiologist of the case gave final clinical decision with all of the extra information. The responsible radiologist alternated among all radiologists by case basis. As a second step, radiologists re-evaluated all AI-positive mammography images as a second look with AI findings following a one-year washout period. In discordance, the final assessment was done by majority voting of all five radiologists. AI-negative images were considered BI-RADS 2. Mammography density was classified according to the BI-RADS lexicon 5th edition system by radiologists with visual assessment. The final assessment is decided by majority voting of all five radiologists. AI system analyzes breast density and generates quantitative density assessment. AI density score ranges from 1–10.

Statistical Analysis

All analyses were done with the R statistical language (Austria, R Core Team, version 4.1.0). A confidence level of 0.95 was considered significant. Distribution of variables assessed by Kolmogorov-Smirnov test, skewness, kurtosis, and Q-Q plot. Age of women with different breast densities compared with the Kruskal-Wallis test. The receiver operating characteristics (ROC) curve and area under the curve (AUC) were analyzed. Statistical differences in the ROC curves were compared with the DeLong test.

Cancer detection rate, accuracy, sensitivity, specificity, positive predictive value (PPV, Positive predictive value of recall), negative predictive value (NPV), and recall rate of the radiologists, AI, and re-evaluation results were calculated. Differences in accuracy, sensitivity, specificity, and recall rate between AI and radiologists were evaluated with the Two-Proportions Z test. PPV and NPV were compared with the statistical package of Stock et al (27). The inter-reader agreements among radiologists in BI-RADS were evaluated using weighted Cohen's kappa (28). Fleiss' Kappa was calculated and compared for prior and reevaluation readings (29). The kappa scores were interpreted as follows: a kappa score of < 20, a poor agreement; 21–40, a fair agreement; 41–60, a moderate agreement; 61–80, a good agreement; and 81–100, an excellent agreement (28).

RESULTS

Right and left mammograms were evaluated separately, and 2126 mammography images of 1063 women (mean age: 53.98, SD: ± 12.35, median age: 50, IQR: 18 years) were evaluated. The demographics are given in Table 1. A total of 29 cancers were diagnosed in 28 women. Interval cancer was not detected in the 2-year follow-up period.

Among all examinations, 2.44% (52/2126) had an AI score above 30.44, of which 46.15% (24/52) were true positive. Standalone AI detected 82.75% (24/29) of cancers; 70.83% (17/24) were invasive, and 29.16% (7/24) were ductal carcinoma in situ (DCIS). The majority voting of radiologists scored positive (BI-RADS 0,4 and 5) 8% (172/2126) of patients and were able to detect 89.65% (26/29) cancers; 73.07% (19/26) were invasive cancers and 26.92% (7/26) were DCIS. The median diameter of cancers was 21 mm. Details of the histopathology of cases are given in Table 2.

Radiologists were able to detect 26 and standalone AI was able to detect 24 of 29 cancers. Only one was marked with AI and a risk score of 13; the other four cases were not flagged with AI. All radiologists and AI missed three cases (two DCIS and one invasive ductal carcinoma -IDCa-) Of these three cases one was detected only on MRI, and the other two were detected on ultrasound. In the two ultrasound-detected cases, the mammographic density was type C, and the masses were obscured by fibroglandular tissue (Table 3). The case, which was detected by MRI only, was a non-mass lesion subsequently diagnosed as DCIS. AI failed

TABLE 2. Characteristics of Detected Cancers

	All Detected Cancer (n=29)	Majority Voting of Radiologists (n=26)	AI Detected Cancer (n=24)
Pathological Subtype			
DCIS	9	7	7
IDCa	19	18	16
ILCa	1	1	1
Molecular Subtype			
Luminal A&B	15	14	12
HER2 Positive	3	3	3
Triple Negative	2	2	2
Tumor Diameter (mm)*,[a]	20 (14)	21 (13.75)	21 (13.75)

* Values were given with median and interquartile range in parentheses

[a] Post-surgery largest diameters were recorded, AI: Artificial intelligence, DCIS: Ductal carcinoma in situ, IDCa: Invasive ductal carcinoma, ILCa: Invasive lobular carcinoma, Her2: human epidermal growth factor receptor 2

TABLE 3. Histopathological Characteristics of Cancers Missed by AI or Radiologists, and AI

Age	Tumor Size	Missed by Only AI or Radiologists and AI	Histopathologic Diagnosis	Subtype of the Tumors	Breast Density	Imaging Features
85	17 mm	Radiologists and AI	IDC	Luminal b (ER/PR +, Her2 -,Ki 67 25%)	Type C	Detected by ultrasound as an irregular mass
57	35 mm	Radiologists and AI	DCIS	High-grade DCIS with comedo necrosis	Type C	Detected by MRI as non-mass enhancement
38	27 mm	AI-only	IDC	Luminal b (ER/PR +, Her2 -,Ki 67 18%)	Type D	Focal asymmetry
51	13 mm	AI-only	IDC	Luminal a (ER/PR +, Her2 -,Ki 67 12%)	Type C	Focal asymmetry
51	5 mm	Radiologists and AI	DCIS	Low-grade DCIS	Type C	Detected by ultrasound as a mass

to detect the other two cases, diagnosed as IDCa. In these cases, mammography density was type C, and the lesions appeared as focal asymmetries (Fig 2).

Distribution of breast density assessments of radiologists across breast density assessments of AI are given in Sankey plot (Fig 3). Radiologists and AI density assessment showed moderate agreement (0.602). The dataset used in this study was rich in dense breasts (67.2% of mammograms were type C and D breast density).

Radiologists showed AUC scores between 91.6–95.4%. The AUC score of majority voting showed 94.7% (95% CI: 91.1–98.3). AI software performance (94.4%, 95% CI: 88.5–100) was statistically not significantly different than (p=0.79) AUC of the majority voting. Re-evaluated MMGs achieved an AUC of 94.8% (95% CI: 91.2–98.3), statistically significantly different from the initial evaluation (p=0.015). However, it was not statistically different from AI (p=0.74). ROC curves of the initial evaluation, AI, and re-evaluation of MMGs are given in Figure 4. ROC curves of all radiologists in both initial evaluation and re-evaluation are given in Supplementary 1. Initial evaluation and AI software achieved 92% (95% CI: 91–93%) and 98% (98% CI: 98–99%) accuracy, 90% (95% CI: 73–98%), and 83% (95% CI: 64–94%) sensitivity, 92% (95% CI: 91–93%) and 99% (95% CI: 98–99%) specificity, 14% (95% CI: 9–19%) and 45% (95% CI: 32–60%) PPV, 100% (95% CI: 100–100%) and 100% (95% CI: 99–100%) NPV respectively where radiologists and AI found 24 (82.8%) and 26 (89.7%) cancers (2 radiologists only) respectively. Re-evaluation with AI showed 92% (95% CI: 91–94%) accuracy, 90% (95% CI: 73–98%) sensitivity, 93% (95% CI: 91–94%) specificity, 14% (95% CI: 9–20%) PPV, 100% (95% CI: 100–100%) NPV with 26 (89.7%) cancers. Details of diagnostic performances are given in Table 4. Further detailed explanation was given for false positive cases of AI in Table 5 (Fig 5) Prior and reevaluation assessment showed both poor to good inter-reader agreement. Further details of pairwise agreements are given in Figure 6. Priors and reevaluations showed no statistically significant difference (45.8 and 46 Fleiss Kappa score, p=0.62).

DISCUSSION

This prospective study in a diagnostic setting showed similar diagnostic performance between major voting of radiologists and AI (AUC score 94.7% for radiologists and 94.4% for AI), which was not elevated after re-evaluation of images with AI

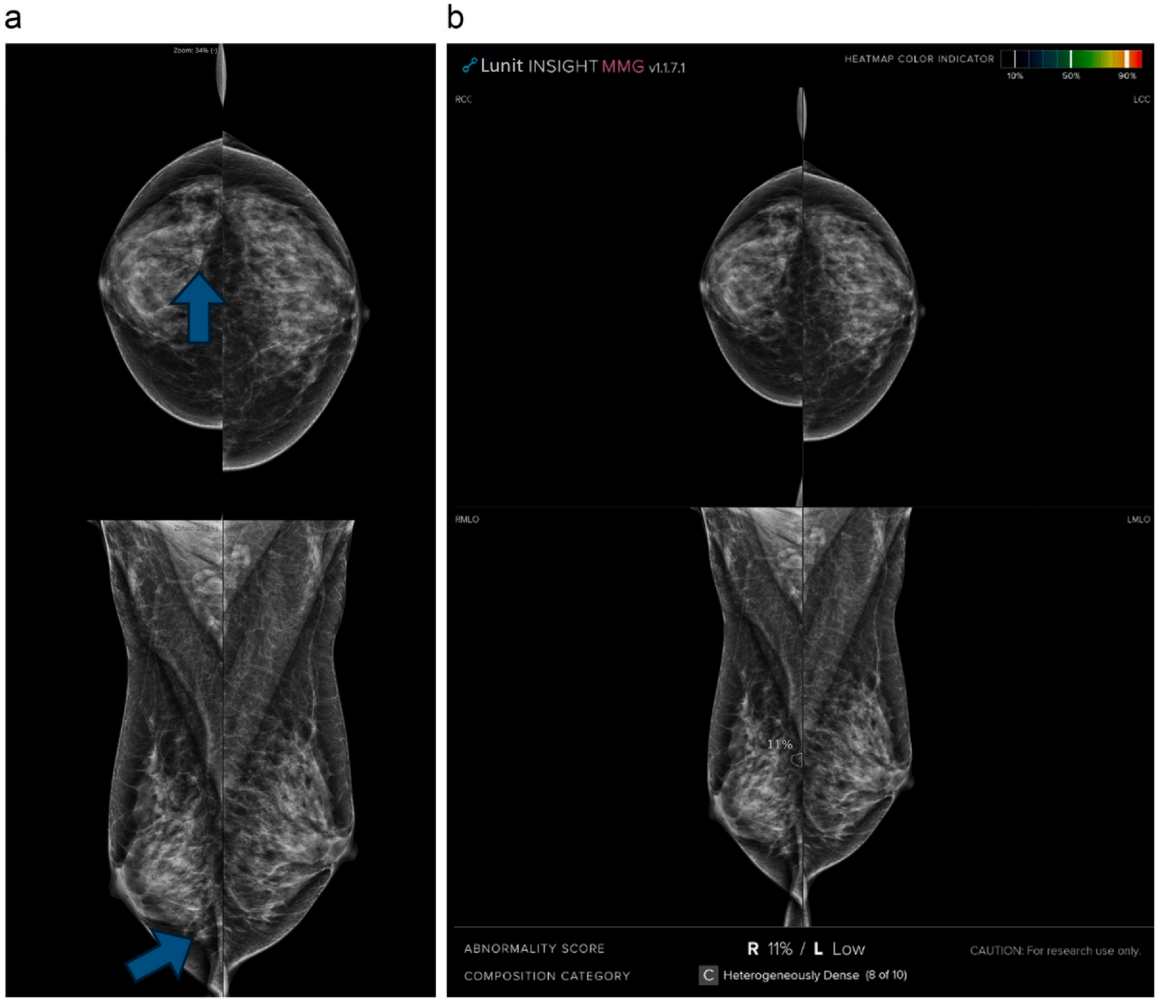

Figure 2. 50-year-old woman with diagnosed invasive ductal carcinoma in the lower outer part of the right breast. (**a**) Full-field digital mammogram (bilateral craniocaudal and mediolateral oblique view). Asymmetric density and minimal distortion were reported in the outer lower part of the right breast by all radiologists and BI-RADS assessment category 4 was assigned (arrows). (**b**) Outline areas and scores are shown in the AI-generated evaluation. The pathological area was not flagged by artificial intelligence.

markings (AUC: 94.8%). Recent studies show the standalone AI sensitivity ranging from 76.9% to 89.7% in screening environments (12,16,30). In this study, standalone AI showed a sensitivity of 83% lower than that of radiologists at 90% similar to Sasaki's et al. study that reported a sensitivity of AI (85%) lower than radiologists (89%) (13). This difference can be attributed to the AI's independent evaluation of mammograms without the ability to compare the right and left sides or prior studies. In two cases, the AI system failed to detect focal asymmetries that can only be detected through a comparative analysis of both breasts. This outcome is consistent with findings in the literature, which suggest that such discrepancies arise from differences in the methods used by AI and radiologists to evaluate mammograms based on varying image features (20). Integrating AI with human evaluation outperforms both standalone AI and the double reading by two radiologists, suggesting that incorporating AI into clinical workflows can significantly enhance overall diagnostic accuracy (14,15,31). On the other hand, AI displayed higher specificity and lower recall rates compared to radiologists ($p < 0.001$). Notably, the recall rate of AI in our study was remarkably low at 2.5%, compared to the significantly higher recall rate of radiologists at 8.9%. However, our dataset included a high proportion of dense breasts, which may have contributed to the increased radiologists' recall rate. Nevertheless, after re-evaluation, the recall rate remained unchanged (8.6%) despite eliminating some AI-generated false positives (Table 4). This finding is consistent with previous reports and underscores the need for careful consideration of AI's role in reducing recall rates and improving diagnostic accuracy (16,20,22). In our study, we analyzed false positive AI markings with risk scores above the threshold and outlined the reasons for these false positives in Table 5. Almost all false positives were attributed to limitations of the AI algorithm, including its inability to use information from prior mammograms, patient reports, or anamnesis, as well as its lack of capability to compare the right and left sides of the breast. After re-evaluation of AI-

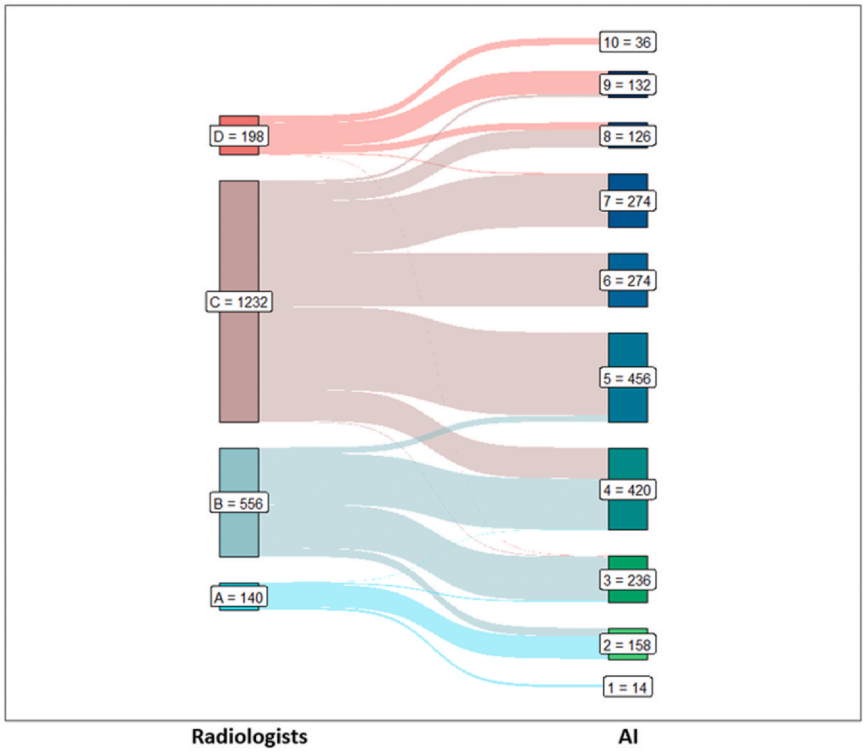

Figure 3. Sankey plot of the radiologists' and artificial intelligence's density assessment.

flagged cases, AI false positive cases were solved with radiologist assessment. The issue of AI-generated false positives can be addressed in the future with advancements in AI-based algorithms highlighted in the literature, incorporating multi-source data, such as prior imaging and clinical information, is expected to significantly reduce these false positives (13,15,23,32). On the other hand, our study showed a higher specificity at 99% with AI evaluation.

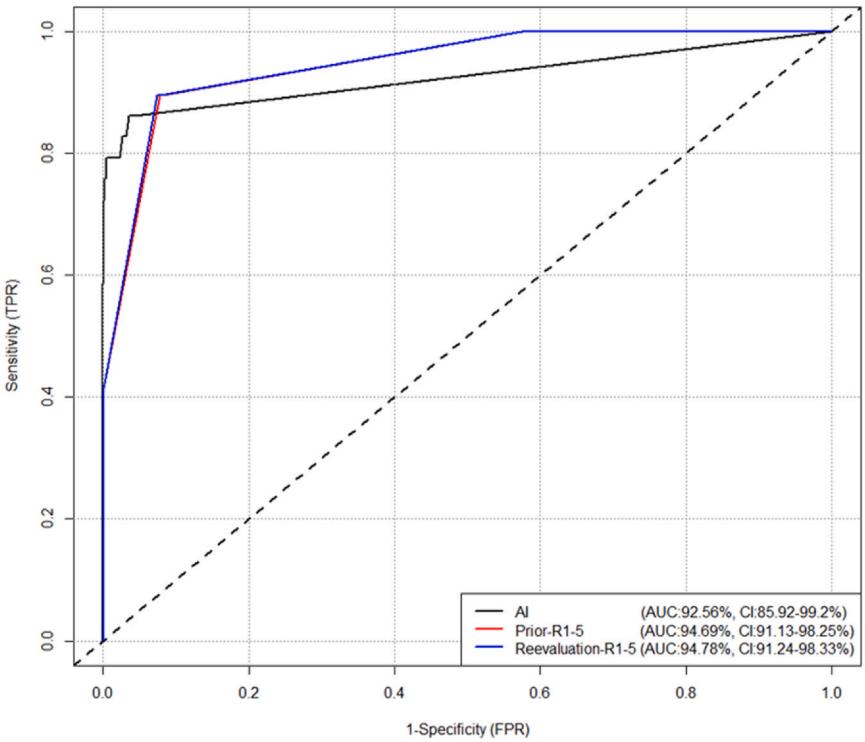

Figure 4. Receiver operating characteristic curves of the initial evaluation, artificial intelligence, and re-evaluation of MMGs.

TABLE 4. Diagnostic Performances of Various Modalities and Combined Approaches: A Comparative Analysis

Metrics	Modality			Comparison (P Value)		
	Prior	AI	Reevaluation	Prior vs AI	Prior vs Reevaluation	AI vs Reevaluation
Cancer Detection Rate*	12.2	11.3	12.2	0.89	1	0.89
Accuracy	92% (CI:91–93%)	98% (CI:98–99%)	92% (CI:91–94%)	< 0.001	0.72	< 0.001
Sensitivity	90% (CI:73–98%)	83% (CI:64–94%)	90% (CI:73–98%)	0.70	1	0.70
Specificity	92% (CI:91–93%)	99% (CI:98–99%)	93% (CI:91–94%)	< 0.001	0.72	< 0.001
PPV	14% (CI:9–19%)	45% (CI:32–60%)	14% (CI:9–20%)	< 0.001	0.008	< 0.001
NPV	100% (CI:100–100%)	100% (CI:99–100%)	100% (CI:100–100%)	**0.21**	**< 0.001**	**0.21**
Recall rate	8.9%	2.5%	8.6%	< 0.001	0.75	< 0.001

* Per 1000 screening, AI: Artificial intelligence, PPV: Positive predictive value of recall, TP/(number of positive screening examinations), NPV: Negative predictive value.

TABLE 5. Analysis of false positive cases with artificial intelligence

AI Score	Radiologist Assessment (BI-RADS Lexicon Score)	Diagnosis
49%	2	Vascular calcification
33%	2	Opacity with obscured margin; no changes found compared with previous mammography
48%	2	Opacity with obscured margin; no changes found compared with previous mammography
46%	2	Microcalcification; no changes found compared with previous mammography
43%	2	Scattered monomorphic microcalcification; no changes found compared with previous mammography
39%	2	Linear microcalcifications; plasma cell mastitis
66%	2	Scattered monomorphic microcalcification; same with contralateral breast and no changes found compared with previous mammography
32%	3	Focal asymmetry; no pathology on US evaluation. No changes in 1 year follow-up
54%	3	Group punctate microcalcifications; Stabil on 1-year follow-up
55%	2	Architectural distortion; secondary to postoperative changes.
35%	2	Scattered monomorphic microcalcifications; no changes found compared with previous mammography
58%	0	Asymmetric density with monomorphic microcalcifications; MRI normal
38%	2	Periareolar heterogeneous microcalcifications; secondary to operation
41%	2	Postoperative heterogeneous macro and microcalcifications
40%	2	Skin calcifications
36%	2	Asymmetric density; symmetric with contralateral breast
40%	2	Focal asymmetry; no changes found compared with previous mammography
35%	2	Asymmetric density; symmetric with contralateral breast
54%	3	Group round monomorphic microcalcifications; no changes found compared with previous mammography
32%	2	Focal asymmetry with scattered heterogeneous calcifications; no pathology on US evaluation. No changes found compared with previous mammography
57%	2	Linear vascular microcalcifications
44%	0	Grouped punctate microcalcifications; progression with previous mammography. MRI recommended.
41%	2	Focal asymmetry, macrocalcifications and marker according to previous benign biopsy; no pathology on US evaluation. No changes found compared with previous mammography

A recent study by Lauritzen et al. reported that AI-based screening sensitivity was non-inferior to that of radiologists (P =.02) while achieving higher specificity (33). The higher specificity of AI is a key strength and should be carefully considered during clinical evaluation to help reduce recall rates and improve specificity in human readings.

Our AUC for AI is compatible with prior studies reporting AI AUC values between 0.840 and 0.959 (14,34,35). In our study, AUC for radiologists ranged from 91.6% to 95.2%. Notably, the radiologist with the highest AUC score had the highest experience in breast radiology. Inter-reader performance variability is a major problem in mammography

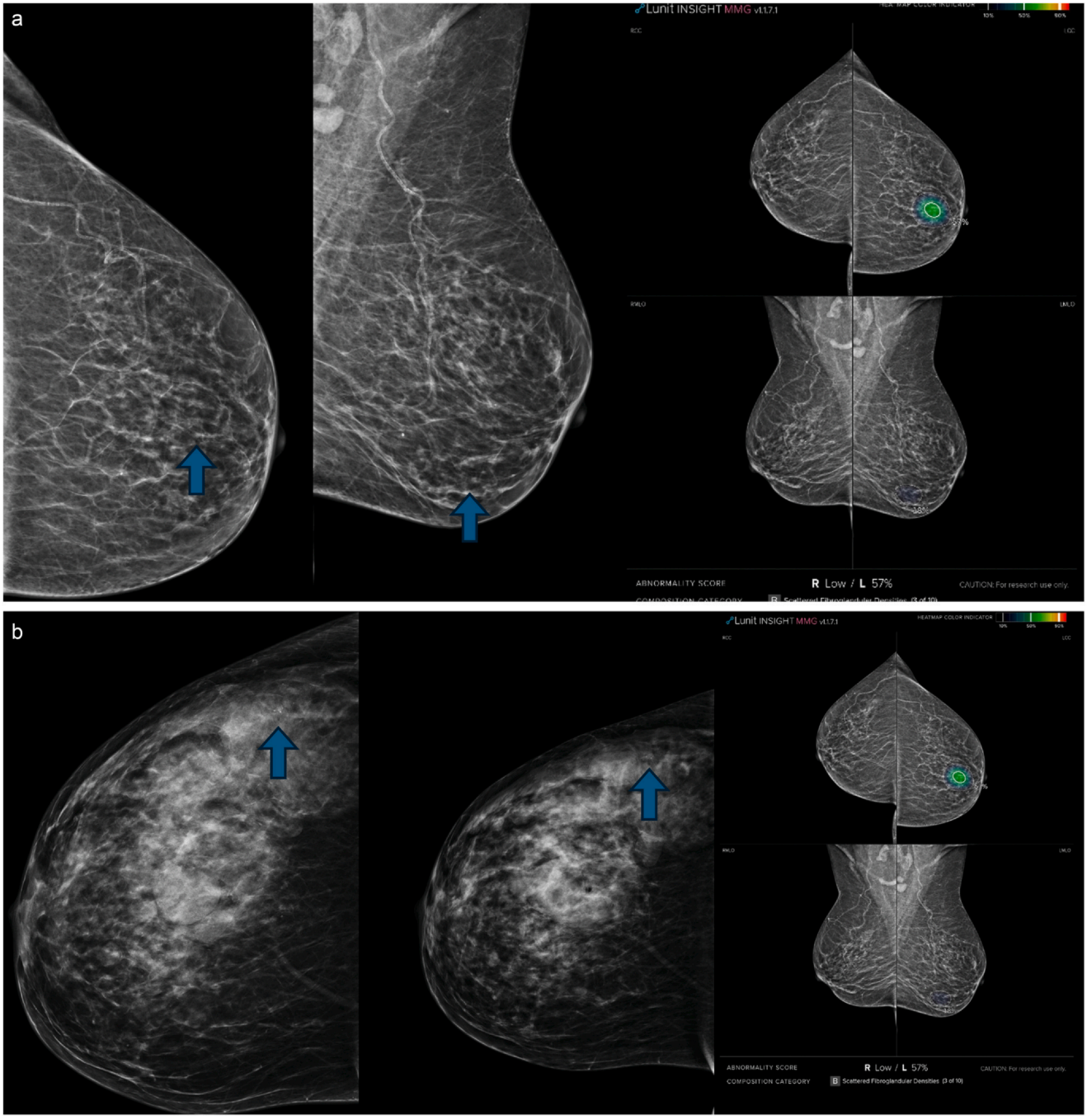

Figure 5. Two different examples of AI-flagged cases with calcifications. (**a**) Screening mammogram of a 72-year-old woman. Left craniocaudal (CC) and mediolateral oblique (MLO) views show linear vascular microcalcifications in the inner lower quadrant and BI-RADS assessment category 2 was assigned by all radiologists (arrow). Outline areas and scores are shown in the AI-generated evaluation. The vascular calcifications were flagged by AI with a score of 57% (above the threshold of 30.44). (**b**) Screening mammogram of a 52-year-old woman. Right CC view and the prior CC are seen with the AI-generated image. A cluster of monomorphic microcalcifications is seen in the outer upper quadrant of the right breast showing no change compared to the prior image. This area was flagged by AI with a score of 54% (above the threshold of 30.44).

assessment and sensitivity in breast cancer detection has been shown to vary from 74.5% to 92.3% (36). Consequently, this study showed that AI performance was comparable to that of experienced radiologists and holds the potential to reduce inter-reader variability.

One of the most important limitations of evaluation with mammography is dense fibroglandular tissue. Mammographic sensitivity is significantly reduced in women with dense fibroglandular breast tissue due to the 'masking effect' of overlapping parenchymal structures, which can

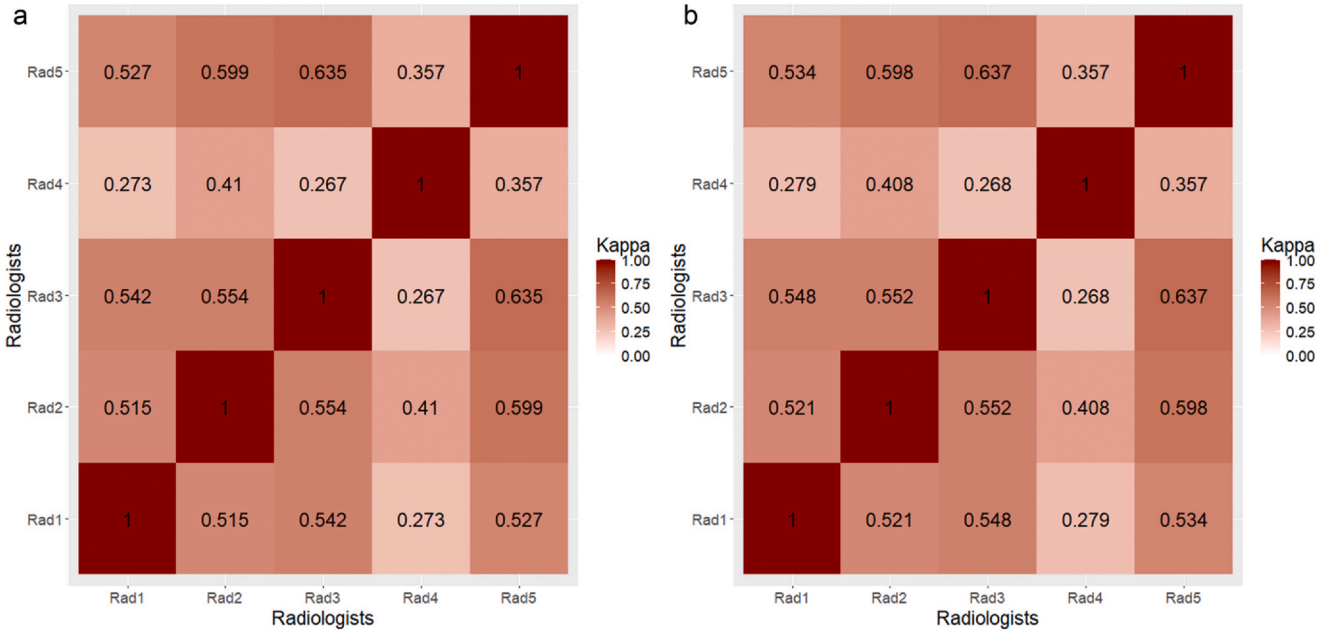

Figure 6. Prior (**a**) and reevaluation (**b**) pairwise Kappa score inter-reader agreement matrices.

obscure underlying malignancies (37). As a result, supplemental imaging modalities such as US and MRI have been increasingly recommended to improve cancer detection rates (37–39). In our study, three cases were not detected by either AI or radiologists because of the 'masking effect'. Among these cases, two were detected by ultrasonography while the remaining case was not visualized on ultrasound and was subsequently detected only on MRI (Table 3).

Our study had several limitations. First, the data were derived from a single institution, which may limit the generalizability of the findings. Second, all mammograms were obtained using a single vendor's mammography device, potentially introducing device specific biases. Third, the majority of the mammograms were normal, and the number of cancer cases was relatively low. Specifically, there were 29 cancer cases among 1063 women, equating to 2.7 cancers per 100 cases. Fourth, determining the risk score threshold of the AI algorithm presented another limitation. Although the threshold was set based on retrospective data repeated calibrations might be necessary (25). However, it is noteworthy that only one case was missed by AI with a risk score of 13, which was significantly below the threshold, and the remaining missed cases were not flagged by AI at all. Finally, we did not evaluate the performance of double reading by a radiologist combined with AI. Instead, in the second step of our analysis, we focused solely on AI-positive cases and identified the reasons for the false-positive findings of the AI system.

In conclusion, this prospective study conducted in a diagnostic setting provides valuable insights. It highlights the implementation of AI CAD systems in daily clinical practice. Our findings demonstrate that standalone AI exhibits diagnostic accuracy comparable to that of an experienced radiologist, suggesting its potential integration of AI as a second reader into daily workflows to reduce inter-reader variability in diagnostic settings. However, further research involving larger patient cohorts is needed to more accurately evaluate the contribution of AI to routine diagnostic clinical practice.

FUNDING

The AI software and the server running the software were provided by Lunit Inc (Lunit, Seoul, South Korea), for research purposes.

ETHICAL STATEMENT

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

The local ethics committee approved this study informed consent was obtained from all subjects involved in the study (Acıbadem University and Acıbadem Healthcare Institutions Medical Research Ethics Committee) (date, number: 15.10.2020, 2020–22/23).

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Each author has made substantial contributions to the conception or design of the work; or the acquisition, analysis, or

interpretation of data; or the creation of new software used in the work. Each author has approved the submitted version (and any substantially modified version that involves the author's contribution to the study). Each author has agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature. Please find the detailed list of each author's contribution to the present work below:

Conceptualization, E.A.; methodology, E.Y., M.E.S., and E.A.; Formal analysis, M.E.S.; Data curation, E.Y., M.E.S, N.G, E.B.T., S.E., Y.O.K., S.N.S., and E.A.; Writing—original draft preparation, E.Y., M.E.S. and E.A.; Writing—review and editing, E.Y. and E.A.; Visualization, E.Y. and M.E.S.; Supervision, E.A.

DECLARATION OF COMPETING INTEREST

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Erkin Aribal reports equipment, drugs, or supplies was provided by Lunit, Seoul, South Korea. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENTS

Not applicable.

APPENDIX A. SUPPORTING INFORMATION

Supplemental data associated with this article can be found in the online version at doi:10.1016/j.acra.2025.05.025.

REFERENCES

1. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA: Cancer J Clin 2024; 74:12–49. https://doi.org/10.3322/caac.21820
2. Kolb TM, Lichy J, Newhouse JH. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. Radiology 2002; 225:165–175. https://doi.org/10.1148/radiol.2251011667
3. Buchberger W, Geiger-Gritsch S, Knapp R, et al. Combined screening with mammography and ultrasound in a population-based screening program. Eur J Radiol 2018; 101:24–29. https://doi.org/10.1016/j.ejrad.2018.01.022
4. Majid AS, de Paredes ES, Doherty RD, et al. Missed breast carcinoma: pitfalls and pearls. Radiographics 2003; 23:881–895. https://doi.org/10.1148/rg.234025083
5. Ikeda DM, Andersson I, Wattsgård C, et al. Interval carcinomas in the malmö mammographic screening trial: radiographic appearance and prognostic considerations. AJR Am J Roentgenol 1992; 159:287–294. https://doi.org/10.2214/ajr.159.2.1632342
6. Gilbert FJ, Astley SM, Gillan MGC, et al. Single reading with computer-aided detection for screening mammography. N Engl J Med 2008; 359:1675–1684. https://doi.org/10.1056/NEJMoa0803545
7. Anderson ED, Muir BB, Walsh JS, Kirkpatrick AE. The efficacy of double reading mammograms in breast screening. Clin Radiol 1994; 49:248–251. https://doi.org/10.1016/s0009-9260(05)81850-1
8. Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994; 191:241–244. https://doi.org/10.1148/radiology.191.1.8134580
9. Arıbal E. Future of breast radiology. Eur J Breast Health 2023; 19:262–266. https://doi.org/10.4274/ejbh.galenos.2023.2023-8-3
10. Tang X. The role of artificial intelligence in medical imaging research. BJR Open 2019; 2:20190031. https://doi.org/10.1259/bjro.20190031
11. Sorin V, Sklair-Levy M, Glicksberg BS, et al. Deep learning for contrast enhanced mammography - a systematic review. Acad Radiol 2024. https://doi.org/10.1016/j.acra.2024.11.035
12. Larsen M, Aglen CF, Lee CI, et al. Artificial intelligence evaluation of 122 969 mammography examinations from a population-based screening program. Radiology 2022; 303:502–511. https://doi.org/10.1148/radiol.212381
13. Sasaki M, Tozaki M, Rodríguez-Ruiz A, et al. Artificial intelligence for breast cancer detection in mammography: experience of use of the ScreenPoint Medical Transpara system in 310 Japanese women. Breast Cancer 2020; 27:642–651. https://doi.org/10.1007/s12282-020-01061-8
14. Kim H-E, Kim HH, Han B-K, 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. https://doi.org/10.1016/S2589-7500(20)30003-0
15. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J Natl Cancer Inst 2019; 111:916–922. https://doi.org/10.1093/jnci/djy222
16. Sharma N, Ng AY, James JJ, et al. Multi-vendor evaluation of artificial intelligence as an independent reader for double reading in breast cancer screening on 275,900 mammograms. BMC Cancer 2023; 23:460. https://doi.org/10.1186/s12885-023-10890-7
17. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020; 577:89–94. https://doi.org/10.1038/s41586-019-1799-6
18. Kizildag Yirgin I, Koyluoglu YO, Seker ME, et al. Diagnostic performance of AI for cancers registered in a mammography screening program: A retrospective analysis. Technol Cancer Res Treat 2022; 21:15330338221075172. https://doi.org/10.1177/15330338221075172
19. Çelik L, Aribal E. The efficacy of artificial intelligence (AI) in detecting interval cancers in the national screening program of a middle-income country. Clin Radiol 2024; 79:e885–e891. https://doi.org/10.1016/j.crad.2024.03.012
20. Dembrower K, Crippa A, Colón E, et al. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. Lancet Digit Health 2023; 5:e703–e711. https://doi.org/10.1016/S2589-7500(23)00153-X
21. Yoen H, Jang M-J, Yi A, et al. Artificial intelligence for breast cancer detection on mammography: factors related to cancer detection. Acad Radiol 2024; 31:2239–2247. https://doi.org/10.1016/j.acra.2023.12.006
22. Lång K, Josefsson V, Larsson A-M, et al. Artificial intelligence-supported screen reading versus standard double reading in the mammography screening with artificial intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. Lancet Oncol 2023; 24:936–944. https://doi.org/10.1016/S1470-2045(23)00298-X
23. Díaz O, Rodríguez-Ruíz A, Sechopoulos I. Artificial Intelligence for breast cancer detection: technology, challenges, and prospects. Eur J Radiol 2024; 175:111457. https://doi.org/10.1016/j.ejrad.2024.111457
24. von Elm E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008; 61:344–349. https://doi.org/10.1016/j.jclinepi.2007.11.008
25. Seker ME, Koyluoglu YO, Ozaydin AN, et al. Diagnostic capabilities of artificial intelligence as an additional reader in a breast cancer screening program. Eur Radiol 2024; 34:6145–6157. https://doi.org/10.1007/s00330-024-10661-3

26. Sickles E, D'Orsi CJ, Bassett L. ACR BI-RADS mammography. ACR BI-RADS Atlas, Breast Imaging Reporting and Data System. 5th edition., Reston, VA: American College of Radiology (ACR),; 2013.
27. Stock C., Hielscher T., Discacciati A. (2023) DTComPair: Comparison of Binary Diagnostic Tests in a Paired Study Design.
28. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 20:37–46. https://doi.org/10.1177/001316446002000104
29. Vanbelle S. Asymptotic variability of (multilevel) multirater kappa coefficients. Stat Methods Med Res 2019; 28:3012–3026. https://doi.org/10.1177/0962280218794733
30. Bergan MB, Larsen M, Moshina N, et al. AI performance by mammographic density in a retrospective cohort study of 99,489 participants in BreastScreen Norway. Eur Radiol 2024; 34:6298–6308. https://doi.org/10.1007/s00330-024-10681-z
31. Larsen M, Aglen CF, Hoff SR, et al. Possible strategies for use of artificial intelligence in screen-reading of mammograms, based on retrospective data from 122,969 screening examinations. Eur Radiol 2022; 32:8238–8246. https://doi.org/10.1007/s00330-022-08909-x
32. Çelik L, Güner DC, Özçağlayan O, et al. Diagnostic performance of two versions of an artificial intelligence system in interval breast cancer detection. Acta Radiol 2023; 64:2891–2897. https://doi.org/10.1177/02841851231200785
33. Lauritzen AD, Rodríguez-Ruiz A, von Euler-Chelpin MC, et al. An artificial intelligence-based mammography screening protocol for breast cancer: outcome and radiologist workload. Radiology 2022; 304:41–49. https://doi.org/10.1148/radiol.210948
34. Salim M, Wåhlin E, Dembrower K, et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA Oncol 2020; 6:1581–1588. https://doi.org/10.1001/jamaoncol.2020.3321
35. Rodríguez-Ruiz A, Krupinski E, Mordang J-J, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology 2019; 290:305–314. https://doi.org/10.1148/radiol.2018181371
36. Elmore JG, Jackson SL, Abraham L, et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology 2009; 253:641–651. https://doi.org/10.1148/radiol.2533082308
37. Boyd NF, Guo H, Martin LJ, et al. Mammographic density and the risk and detection of breast cancer. N Engl J Med 2007; 356:227–236. https://doi.org/10.1056/NEJMoa062790
38. Melnikow J, Fenton JJ, Whitlock EP, et al. Supplemental screening for breast cancer in women with dense breasts: a systematic review for the U.S. Preventive Services Task Force. Ann Intern Med 2016; 164:268–278. https://doi.org/10.7326/M15-1789
39. Berg WA. Rationale for a trial of screening breast ultrasound: American College of Radiology Imaging Network (ACRIN) 6666. AJR Am J Roentgenol 2003; 180:1225–1228. https://doi.org/10.2214/ajr.180.5.1801225