Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Impact of AI for Digital Breast Tomosynthesis on Breast Cancer Detection and Interpretation Time

Eun Kyung Park, MD, PhD • SooYoung Kwak, BS, MS • Weonsuk Lee, MSc • Joon Suk Choi, MSc • Thijs Kooi, MSc, PhD • Eun-Kyung Kim, MD, PhD

From Lunit, 374 Gangnam-daero, Gangnam-gu, Seoul 06241, Republic of Korea (E.K.P., S.Y.K., W.L., J.S.C., T.K.); and Department of Radiology, Yongin Severance Hospital, College of Medicine, Yonsei University, Yongin, Republic of Korea (E.K.K.). Received August 14, 2023; revision requested September 24; revision received February 28, 2024; accepted March 20. **Address correspondence to** E.K.P. (email: *epark1001@gmail.com*).

Conflicts of interest are listed at the end of this article.

See also the commentary by Bae in this issue.

Radiology: Artificial Intelligence 2024; 6(3):e230318 • https://doi.org/10.1148/ryai.230318 • Content codes: AI BR OI

Purpose: To develop an artificial intelligence (AI) model for the diagnosis of breast cancer on digital breast tomosynthesis (DBT) images and to investigate whether it could improve diagnostic accuracy and reduce radiologist reading time.

Materials and Methods: A deep learning AI algorithm was developed and validated for DBT with retrospectively collected examinations (January 2010 to December 2021) from 14 institutions in the United States and South Korea. A multicenter reader study was performed to compare the performance of 15 radiologists (seven breast specialists, eight general radiologists) in interpreting DBT examinations in 258 women (mean age, 56 years ± 13.41 [SD]), including 65 cancer cases, with and without the use of AI. Area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and reading time were evaluated.

Results: The AUC for stand-alone AI performance was 0.93 (95% CI: 0.92, 0.94). With AI, radiologists' AUC improved from 0.90 (95% CI: 0.86, 0.93) to 0.92 (95% CI: 0.88, 0.96) ($P = .003$) in the reader study. AI showed higher specificity (89.64% [95% CI: 85.34%, 93.94%]) than radiologists (77.34% [95% CI: 75.82%, 78.87%]) ($P < .001$). When reading with AI, radiologists' sensitivity increased from 85.44% (95% CI: 83.22%, 87.65%) to 87.69% (95% CI: 85.63%, 89.75%) ($P = .04$), with no evidence of a difference in specificity. Reading time decreased from 54.41 seconds (95% CI: 52.56, 56.27) without AI to 48.52 seconds (95% CI: 46.79, 50.25) with AI ($P < .001$). Interreader agreement measured by Fleiss κ increased from 0.59 to 0.62.

Conclusion: The AI model showed better diagnostic accuracy than radiologists in breast cancer detection, as well as reduced reading times. The concurrent use of AI in DBT interpretation could improve both accuracy and efficiency.

Supplemental material is available for this article.

Breast cancer screening programs with mammography reduce mortality by 20%–49% (1–3). In spite of its role in screening, the sensitivity of digital mammography (DM) is limited in female individuals with dense breasts (4). In the past decade or so, digital breast tomosynthesis (DBT) has been shown to improve cancer detection and reduce recall rates (5–10). Use of DBT has steadily increased for routine breast cancer screening, with 83% of Mammography Quality Standards Act (MQSA)–certified facilities reporting DBT units in 2022 (11). However, there is wide variability in recall rates with DBT across radiologists (12). Moreover, DBT interpretation time is almost twice that needed for interpreting DM (13). Additional strategies to increase accuracy and efficiency will allow the potential benefits of DBT to be maximized.

An approach to improving performance in breast cancer detection is to use computer-aided detection (CAD). Traditional CAD with human-designed descriptors has been used to assist mammography interpretation; however, its usefulness and effectiveness have been challenged by subsequent large-scale clinical trials (14–16). High false-positive rates with traditional CAD lead to exhaustion for radiologists and unnecessary additional examinations (14). Within the past 5 years, artificial intelligence (AI)–based CAD was developed using self-learned descriptors, and several studies revealed that the AI algorithm for breast cancer diagnosis at DM showed better diagnostic performance, with a higher cancer detection rate and lower false-positive recall (17–19). Radiologist performance was improved with the aid of AI for DM (17). For DBT, a 2019 study showed that AI CAD could reduce the reading time for DBT with maintained or better performance (20). On the other hand, one study showed a higher recall rate with stand-alone use of AI (21). There remains insufficient evidence on the independent performance of AI for DBT to support implementation into clinical workflow to improve diagnostic accuracy or reduce workload (22).

In this study, we developed and validated a deep learning AI algorithm trained on DBT examinations to detect breast cancer. We investigated the stand-alone diagnostic performance of the AI model and assessed the impact of the model on the diagnostic accuracy of radiologists, reading time, and consistency across readers.

Materials and Methods

This retrospective study was approved by ethics review and the central institutional review board, and the requirement for informed consent was waived. Mammog-

Abbreviations

AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, BI-RADS = Breast Imaging Reporting and Data System, CAD = computer-aided detection, CNN = convolutional neural network, DBT = digital breast tomosynthesis, DM = digital mammography, MQSA = Mammography Quality Standards Act, 2D = two-dimensional

Summary

The developed artificial intelligence system could support radiologists by improving diagnostic accuracy and reducing workload for the diagnosis of breast cancer on digital breast tomosynthesis images.

Key Points

- The artificial intelligence (AI) algorithm developed with robust large-scale digital breast tomosynthesis examinations showed better specificity (89.64%) in breast cancer detection compared with radiologists (77.34%).
- The concurrent use of a qualified AI system led to significant improvement in radiologists' performance (AUC difference: 0.25, P = .003) with reduced reading times (reading time difference: 5.89 seconds, P < .001)
- This study showed the potential of a better screening pathway in breast cancer detection by using AI, with improvement of accuracy and efficiency.

Keywords

Breast, Computer-Aided Diagnosis (CAD), Tomosynthesis, Artificial Intelligence, Digital Breast Tomosynthesis, Breast Cancer, Computer-Aided Detection, Screening

raphy examinations were de-identified and collected according to the Health Insurance Portability and Accountability Act Safe Harbor standard. This work was supported by funds secured by Lunit. The author (E.K.K.) who is not an employee of Lunit had control of the data and information submitted for publication that might present a conflict of interest for authors who are employees of Lunit (E.K.P., S.K., W.L., J.S.C., T.K.).

Study Cohort

A total of 2206 DBT examinations, including 262 cases for the reader study, were retrospectively collected from 14 imaging facilities located in the United States, with a balanced distribution of demographics and cancer characteristics. The study cases were collected separately from the development dataset (described in the Development of AI Model section), and none of the study cases were used for model development. We collected cases with the following criteria: Cancer-positive cases were defined with pathologic confirmation only, benign cases were defined with pathologic confirmation or at least 1 year of follow-up, and normal cases were defined with at least 1 year of follow-up. Studies were eligible for inclusion if they met the following criteria: *(a)* female sex, *(b)* any ethnic origin, *(c)* 22 years old or older, *(d)* acquired with devices from two vendors (Hologic and GE HealthCare), and *(e)* four-view (right craniocaudal, right mediolateral oblique, left craniocaudal, and left mediolateral oblique) images of screening or diagnostic DBT examinations with a full-field DM image or a synthetic two-dimensional (2D) image. Exclusion criteria included the following: *(a)* history of breast cancer, *(b)* previous surgery or vacuum-assisted biopsy, *(c)* presence of a breast implant or pacemaker on the required images, and *(d)* inadequate quality of imaging according to MQSA criteria in terms of positioning, compression, exposure level, dose, contrast, sharpness, noise, and artifacts. Of the 262 cases initially collected for the reader study, four cases were excluded due to inadequate image quality. The final dataset included 2202 DBT examinations, with 1102 (50%) cancer cases, and was used to evaluate stand-alone performance (Fig 1). The ground truth was obtained for each examination by three expert breast imaging radiologists (11, 20, and 15 years of experience) who did not participate in the reader study, where examinations were classified into non-cancer or cancer groups. The three-dimensional location of the lesion at the examination was annotated, and the lesion was classified as a soft-tissue lesion, a calcification, or both. After two radiologists completed the review, the third radiologist, who was the most experienced, made a final decision considering the results of the other two radiologists. To set the reference standards, the radiologists reviewed the collected examinations with relevant clinical supporting data, including radiology and pathology reports. For cancer cases, the relevant cancer characteristic information, including cancer location, size, shape, presence of calcification, and pathologic results, was used for obtaining ground truth.

Study Design

We developed an AI algorithm to detect breast cancer at DBT, retrospectively investigated its stand-alone performance, and performed a multireader multicase study with a fully crossed design. Stand-alone performance was assessed with 2202 cases, with a cancer prevalence of 50% (1102 cancer, 367 benign, and 733 normal). The reader study was performed with 15 readers and a cancer-enriched dataset (prevalence of 25%) of 258 women (mean age, 56 years ± 13.41 [SD]; 65 with cancer, 65 with benign findings, 128 healthy). Each reader read half of the 258 cases with AI and the other half without AI in the first session and vice versa during the second session, such that each case was read by each reader both with and without AI. Every reader had at least a 4-week washout period between sessions.

Development of AI Model

The AI model used in this study was pretrained with around 30 000 positive and 30 000 negative 2D mammograms and subsequently fine-tuned on 12 810 DBT examinations using accurately determined ground truth labels and expertly annotated tomosynthesis images. The development dataset is described in Figure 2. The mean age of patients included was 50.3 years ± 10.0. The DBT examinations were performed using Hologic (10 436 of 12 810, 81.5%) and GE HealthCare (2374 of 12 810, 18.5%) systems. We collected both screening and diagnostic DBT examinations. To determine the ground truth labels accurately, we collected examinations with the following criteria: cancer-positive examinations defined with pathologic confirmation only, benign examinations defined with pathologic confirmation or at least 1 year of follow-up, and normal examinations defined with at least 1 year of follow-up. The en-

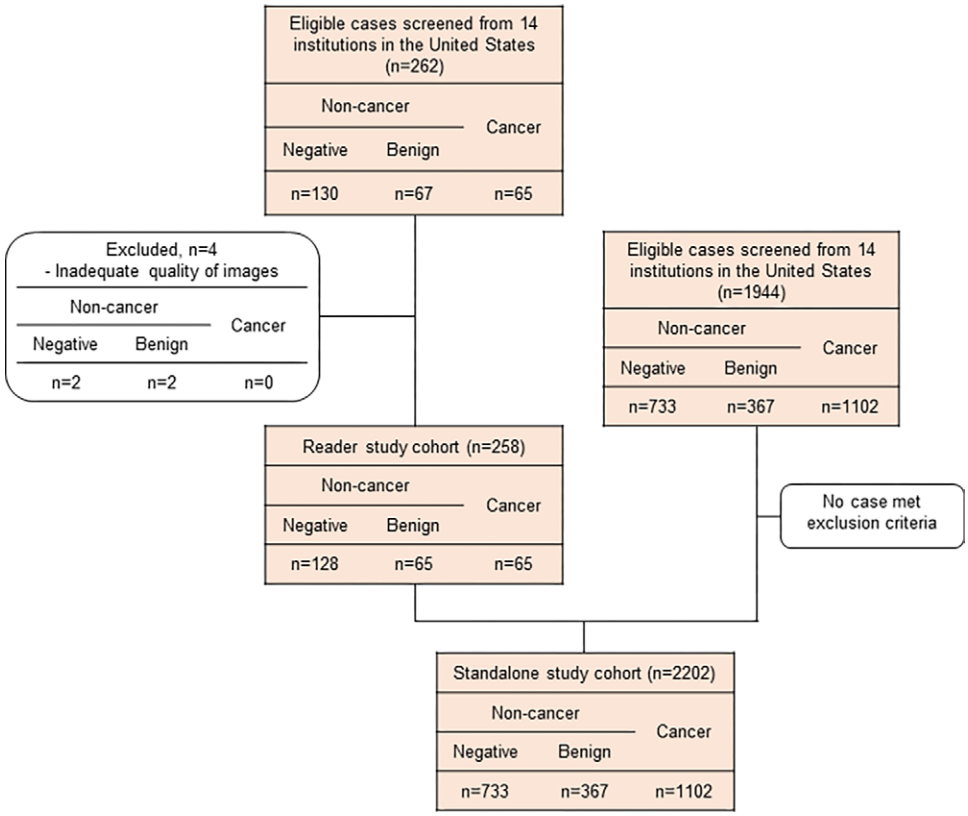

Figure 1: Case selection flowchart in the stand-alone study and the reader study. A total of 262 digital breast tomosynthesis (DBT) examinations were collected according to the inclusion criteria. Four cases were excluded because of inadequate image quality according to Mammography Quality Standards Act criteria in terms of positioning, compression, exposure level, dose, contrast, sharpness, noise, and artifacts. As a result, 258 cases including 65 cancers (25.1%) were enrolled in the reader study. To evaluate stand-alone artificial intelligence performance, 2202 DBT examinations were included, with a balanced distribution of demographics and cancer characteristics.

tire dataset was divided into four sets: a training set for training the AI model, a tuning set for selection of the training scenario, a test set for evaluation of the final model, and an external test set for out-of-domain evaluation of the model with separately collected data from an independent institution. For the purpose of AI algorithm development, 3281 DBT examinations with cancer-positive findings (92.0% of cancers of 3568 total examinations with cancer-positive findings) and 1742 DBT examinations with benign findings (27.7% of 6293 total examinations with benign findings) were annotated by one of 12 radiologists with breast subspecialty by referring to previous radiology and pathology reports. As the model was designed to analyze the location and likelihood of the presence of breast cancer, it was trained as positive only with biopsy-proven cancers. The AI model is based on a deep convolutional neural network (CNN) with a ResNet-34 backbone. A DBT scan is a stack of 2D sections reconstructed from 2D projections taken from multiple angles (23). To train a CNN on large DBT data effectively, we sampled the neighboring sections around the annotated suspicious lesions instead of using the whole DBT stack or a single DBT section. Our experiments showed that using the neighboring sections as input promised higher accuracy than using the single DBT section (24). The model analyzes an input DBT image in three stages: preprocessing, image analysis, and analysis result presentation. The input images are processed through cropping and resizing and analyzed by the deep neural network in the inference server. The analyzed data are converted and generated into a human-readable image for the users to interpret. The model provides heatmaps and an abnormality score per lesion/per breast for each input image (Fig 3). More details on the network architecture and training procedure can be found in the study by Lee et al (24).

An operating cutoff value controlling the trade-off between the sensitivity and specificity of the algorithm was set to achieve 90% sensitivity in the tuning dataset, and this threshold was used for the stand-alone validation and reader studies. We used the per-mammogram abnormality score, which was defined as the maximum of abnormality scores of each of the four views, to evaluate the stand-alone performance of the AI system. A detailed description of the model is given in Appendix S1.

Reader Study

All 15 readers are certified by the American Board of Radiology and are qualified to interpret mammograms in accordance with the MQSA. Each reader interpreted more than 500 DBT examinations in the last 2 years. The general reader group included eight readers who dedicated less than 75% of their professional time to breast imaging in the past 3 years. The specialized reader group included the remaining seven readers, who completed a breast imaging fellowship and have devoted

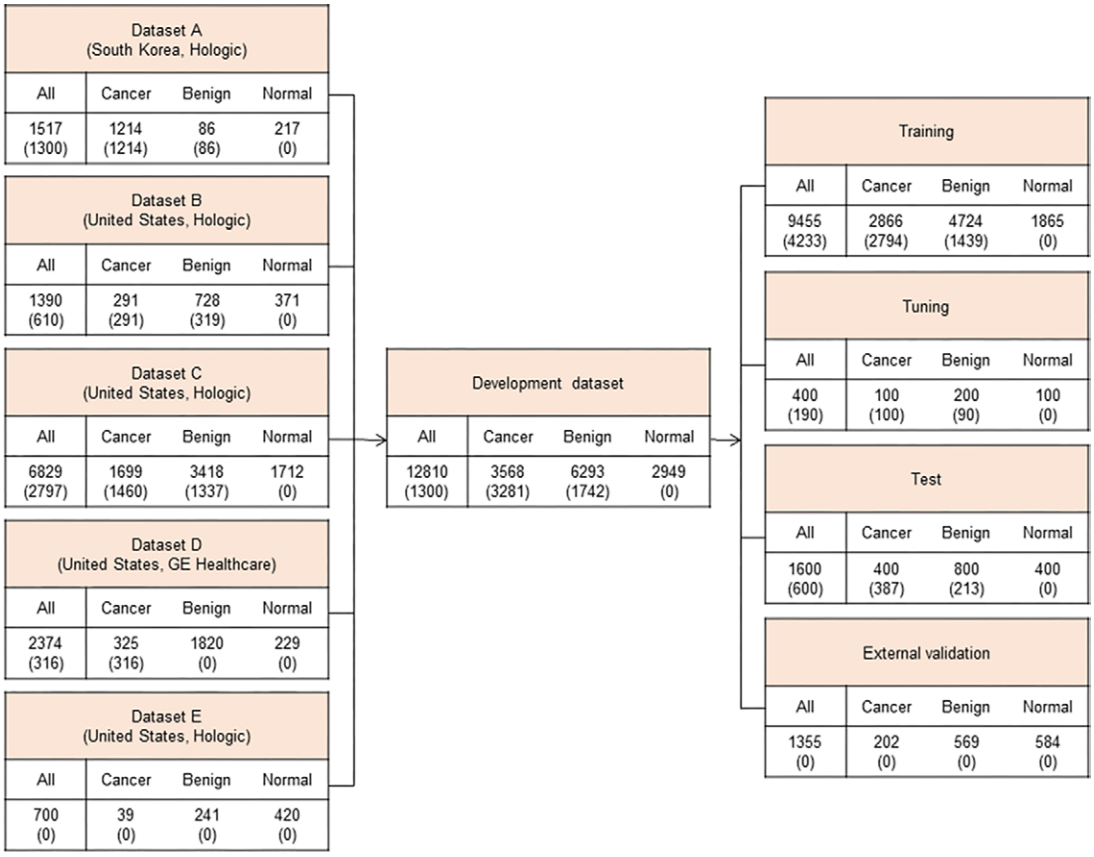

Figure 2: Development dataset. To develop the artificial intelligence algorithm based on deep convolutional neural networks, 12 810 four-view digital breast tomosynthesis (DBT) examinations from five data sources in the United States and South Korea were used. Datasets of DBT examinations from January 2010 to December 2021 in the United States and from January 2012 to January 2018 in South Korea were collected. Values are the numbers of cases, followed by the numbers of annotated cases in parentheses.

75% or more of their professional time to breast imaging in the same period. Each reader reviewed the mammogram and determined whether or not to recall the patient for further workup. Then, each mammogram was graded with two scores: a likelihood of breast cancer on a seven-point assessment scale and a level of suspicion score of 0 to 100 (Appendix S2, Tables S1, S2). Reading time in DBT interpretation was evaluated with the smart navigation function of the in-app viewer of the AI system. Reading time was defined as the time between when the readers clicked "start" to begin viewing the examinations and when they selected the final decision of recall or no recall. Readers were blinded to the reading time measurements for each case to prohibit any potential bias that may occur. The case reading order was randomized separately for each reader.

Statistical Analysis

The number of readers and mammograms required was calculated by the power estimation method (significance level set at 5% and power to 90%) with an effect size of 0.057 based on the area under the receiver operating characteristic curve (AUC) without AI and with AI, from a similar previous reader study (20). For a stand-alone AI performance study, the AUC, sensitivity, and specificity were evaluated. In the reader study, multireader multicase receiver operating characteristic curve analysis was used to account for reader variability and the correlation among readings with and without AI. The difference in AUC was estimated with the nonparametric trapezoidal method and analyzed using a two-sided 95% CI and a *P* value. For the analysis of sensitivity and specificity, a logistic regression with a generalized estimating equation method was applied. Interreader agreement was measured with the Fleiss κ statistic. Reading time improvement was assessed by the difference in reading time in seconds without AI and with AI. Reading time difference and the standard error for each reader, as well as the average reading time difference for total readers and its 95% CI, were analyzed using *P* values. Subgroup analyses stratified by reader type (general radiologist vs breast specialist), manufacturers (GE HealthCare vs Hologic), age (<55 years vs ≥55 years), breast density (fatty [Breast Imaging Reporting and Data System {BI-RADS} class A or B] vs dense [class C or D]), lesion feature (soft tissue vs calcification vs both), and cancer characteristics (invasive vs noninvasive) were performed. The software R (version 4.2.2; R Core Team) was used to perform statistical analyses by a biostatistician (S.K.). A *P* value of less than .05 was deemed statistically significant.

Results

AI Stand-Alone Performance

The overall AUC in the entire stand-alone dataset was 0.93 (95% CI: 0.92, 0.94) with sensitivity and specificity of 88.57%

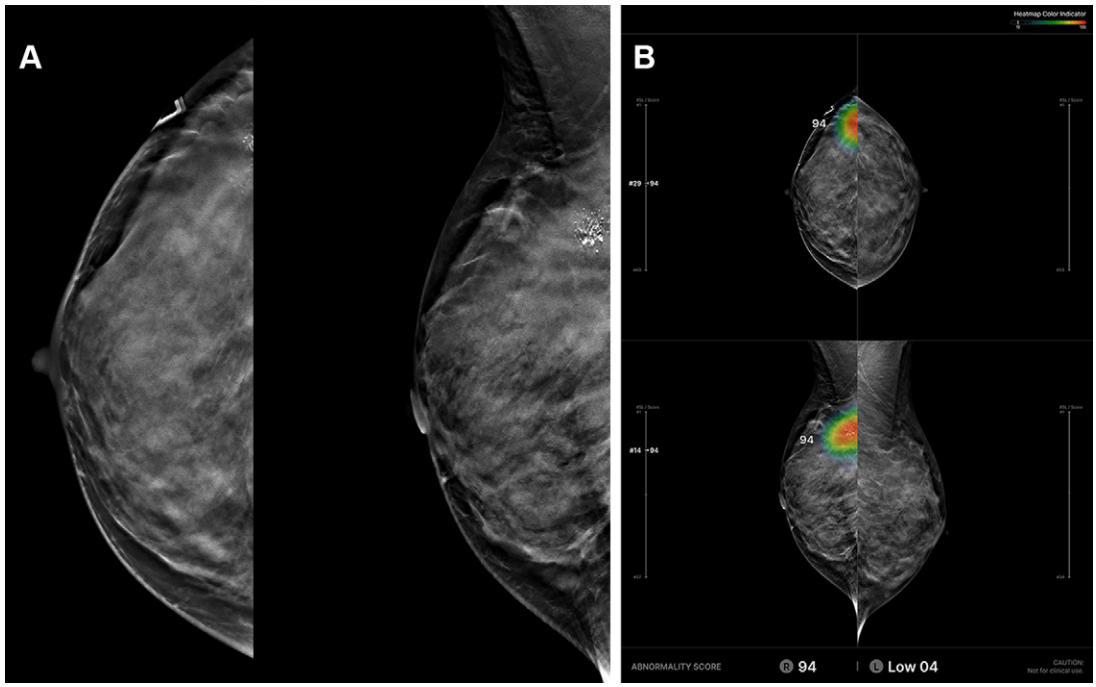

Figure 3: Example of a case with an abnormality analyzed by the artificial intelligence (AI)–based diagnostic support software. **(A)** Based on analysis of the algorithm for an input digital breast tomosynthesis (DBT) examination, **(B)** the resulting diagnostic support software (Lunit INSIGHT DBT) provides heatmap marks (or contour lines) and abnormality scores on suspicious lesions on three-dimensional sections with the Slice Navigation Bar and on synthetic two-dimensional mammography images at corresponding areas in the picture archiving and communication system. Additional dedicated viewer formats (In-App Viewer, Movie) could support the AI system.

(95% CI: 86.69%, 90.45%) and 83.00% (95% CI: 80.78%, 85.22%), respectively. Sensitivities for invasive cancers (305 of 1102, 27.7% of cancer cases) and noninvasive cancers (797 of 1102, 72.3%) were 89.18% (95% CI: 85.69%, 92.67%) and 88.33% (95% CI: 86.10%, 90.56%), respectively. Of 1102 cancer cases, 126 (11.4%) cases were missed by AI. A total of 93 (73.8%) of the 126 missed cases were noninvasive cancers. AI demonstrated AUCs of 0.91 (95% CI: 0.88, 0.94) for soft-tissue lesions and 0.79 (95% CI: 0.74, 0.84) for calcification-only lesions. For lesions combined with soft-tissue lesions and calcifications, the AUC was 0.98 (95% CI: 0.96, >0.99), with 98.23% (95% CI: 95.80%, 100.00%) sensitivity. The AUC was comparable across DBT vendors (P = .90) and breast densities (P = .75) (Table 1).

Diagnostic Accuracy of Radiologists with and without AI

Table 2 presents the characteristics of the reader study cohort. Regarding overall diagnostic performance, radiologists without AI had an AUC of 0.90 (95% CI: 0.86, 0.93), and the AUC of radiologists with AI significantly improved to 0.92 (95% CI: 0.88, 0.958) (P = .003) (Table 3). The stand-alone performance of AI in this cohort had an AUC of 0.92 (95% CI: 0.87, 0.97), which was higher than that of radiologists without AI, but showed no statistical significance (P = .24). In the reading panel subgroup analysis, the AUC of general radiologists improved from 0.89 (95% CI: 0.85, 0.93) to 0.92 (95% CI: 0.88, 0.96) (P = .003) when interpreting with AI, which was comparable to the AUC of breast specialists with AI (AUC: 0.92 [95% CI: 0.88, 0.96]) and higher than the AUC for breast specialists without AI (AUC: 0.90 [95% CI: 0.87, 0.94]) (Fig 4). The improvements between AUC without and with AI were found for both dense breasts (P = .005) and fatty breasts (P = .04). According to the lesion features, the AUC for soft-tissue lesions only was 0.83 (95% CI: 0.77, 0.89) for radiologists without AI and 0.88 (95% CI: 0.83, 0.94) for radiologists with AI (P < .001). The AUC of stand-alone AI for the Hologic machine was higher (0.98 [95% CI: 0.95, >0.99]) than that for GE (0.89 [95% CI: 0.82, 0.96]). The AUCs of radiologists with AI for each vendor were similar to stand-alone AI performance. When interpreted with AI, the AUCs were improved regardless of DBT vendor; however, only the AUC of radiologists with AI for Hologic was significantly higher than that of those without AI (P = .008) (Table 4).

Overall, the sensitivity of the radiologists was significantly improved when read with AI (P = .04), with no evidence of a difference in specificity (P = .10) (Table 3). For general radiologists, sensitivity significantly improved with AI from 86.35% (95% CI: 83.39%, 89.30%) to 90.19% (95% CI: 87.64%, 92.75%) (P = .02). For breast specialists, there was no evidence of a difference between sensitivity without AI (84.40% [95% CI: 81.06%, 87.73%]) and with AI (84.84% [95% CI: 81.54%, 88.13%]) (P = .67). Specificities did not change significantly with AI for either general radiologists or breast specialists. In fatty breasts, sensitivity significantly increased with AI from 87.33% (95% CI: 84.26%, 90.41%) to 90.67% (95% CI: 87.98%, 93.35%) (P = .003). In dense breasts, specificity improved with AI from 75.14% (95% CI: 72.59%, 78.68%) to 78.83% (95% CI: 76.43%, 81.23%) (P = .02). Sensitivity for dense breasts when radiologists interpreted with AI (85.14% [95% CI: 82.10%, 88.19%) was not significantly different from sensitivity without

Table 1: Stand-Alone Performance of Artificial Intelligence Algorithm on Validation Dataset

Feature	Case Pool	AUC	Sensitivity (%)	Specificity (%)
All	2022	0.93 (0.92, 0.94)	88.57 (86.69, 90.45)	83.00 (80.78, 85.22)
Vendor				
Hologic	1617/2202 (73.4)	0.93 (0.92, 0.94)	90.67 (88.70, 92.64)	80.15 (77.36, 82.95)
GE HealthCare	585/2202 (26.6)	0.93 (0.91, 0.95)	81.95 (77.33, 86.58)	89.97 (86.67, 93.27)
Breast density				
Fatty	1155/2202 (52.5)	0.93 (0.92, 0.94)	88.06 (85.47, 90.65)	82.97 (79.84, 86.11)
Dense	1047/2202 (47.5)	0.93 (0.91, 0.94)	89.18 (86.45, 91.90)	83.03 (79.89, 86.17)
Lesion type				
Soft-tissue lesion*	901/1399 (64.4)	0.91 (0.88, 0.94)	89.95 (87.83, 92.06)	76.00 (68.51, 83.49)
Calcification	349/1399 (24.9)	0.79 (0.74, 0.84)	76.56 (70.57, 82.55)	68.79 (61.54, 76.04)
Combined	149/1399 (10.7)	0.98 (0.96, >0.99)	98.23 (95.80, 100)	83.33 (71.16, 95.51)
Pathology				
Invasive	305/1102 (27.7)	…	89.18 (85.69, 92.67)	…
Noninvasive	797/1102 (72.3)	…	88.83 (86.10, 90.56)	…

Note.—Data are numbers of cases, with percentages in parentheses for case pool; for AUC, sensitivity, and specificity, data in parentheses are 95% CIs. AUC = area under the receiver operating characteristic curve.
* Normal mammograms were excluded in this subgroup analysis.

AI (P = .40); it was also comparable to sensitivity without AI in fatty breasts (87.33% [95% CI: 84.26%, 90.41%]). For soft-tissue lesions, both the sensitivity and specificity of radiologists significantly improved with AI (P = .007 and .002, respectively). For calcifications, both the sensitivity and specificity of radiologists showed no evidence of a difference with AI (P = .73 and .07, respectively). The stand-alone AI presented comparable sensitivity (86.44% [95% CI: 77.70%, 95.18%]) to readers alone (86.55% [95% CI: 84.31%, 88.80%]) in invasive cancers but showed lower sensitivity (66.67% [95% CI: 28.95%, 100%]) compared with readers (74.44% [95% CI: 65.43%, 83.46%]) in noninvasive cancers. In invasive cancer, when radiologists read with AI, the sensitivity improved to 89.15% (95% CI: 87.10%, 91.20%) (Table S3). Figures 5 and 6 are example cases that were missed without AI but detected with AI.

Interreader agreement (measured by Fleiss κ) of all readers without AI was 0.59, and this increased to 0.62 with AI. Interreader agreement of specialized readers without AI was 0.62, which increased to 0.67 with AI. For general readers, the interreader agreement increased from 0.56 without AI to 0.58 with AI.

Reading Time Changes with Use of AI

The mean reading time per case was 54.41 seconds for the unaided readings (95% CI: 52.56, 56.27) and 48.52 seconds for the readings with AI (95% CI: 46.79, 50.25) (P < .001); the reading time decreased for nine radiologists (five general, four specialist) and increased for six radiologists (three general, three specialist). The mean reading time decreased with AI for the cancer group (mean reading time difference, 6.33 seconds [95% CI: 2.52, 10.14]; P = .001) and noncancer group (5.75 seconds [95% CI: 3.46, 8.03], P < .001), general readers (7.69 seconds [95% CI: 5.06, 10.32], P < .001) and specialists (3.80 seconds [95% CI: 0.87, 6.73], P = .01), and fatty breasts (4.52 seconds [95% CI: 1.89, 7.16], P = .001) and dense breasts (7.72 seconds [95% CI: 4.80, 10.64], P < .001). When considering the lesion type, the mean reading time significantly decreased only for soft-tissue lesions (9.50 seconds [95% CI: 7.16, 11.85], P < .001). The mean reading time decreased regardless of DBT machine vendor (Table 4).

Discussion

In this study, we investigated the effectiveness of an AI algorithm developed for the detection and diagnosis of breast cancers on DBT images and have shown that it can be used as an effective diagnostic support tool for radiologists when incorporated into the DBT interpretation workflow. The stand-alone AI performance as measured by AUC was higher than that of radiologists alone. When the AI model was used in conjunction with radiologists, their diagnostic performance

Table 2: Characteristics of 258 Digital Breast Tomosynthesis Study Cases in Reader Study

Characteristics	Noncancer			Cancer	Total
	Normal	Benign	Total		
Age (y)					
Mean	56.04	51.97	54.67	60.45	56.12
Range	37–92	30–81	30–92	32–92	30–92
BI-RADS breast density*					
Category A	19/128 (14.8)	4/65 (6.2)	23/193 (11.9)	2/65 (3.1)	25/258 (9.7)
Category B	59/128 (46.1)	37/65 (56.9)	96/193 (49.7)	28/65 (43.1)	124/258 (48.1)
Category C	45/128 (35.2)	20/65 (30.7)	65/193 (33.7)	24/65 (36.9)	89/258 (34.5)
Category D	5/128 (3.9)	4/65 (6.2)	9/193 (4.7)	11/65 (16.9)	20/258 (7.7)
Lesion features					
Soft tissue	…	43/65 (66.2)	43/65 (66.2)	47/65 (72.3)	90/130 (69.2)
Calcifications	…	14/65 (21.5)	14/65 (21.5)	12/65 (18.5)	26/130 (20)
Both	…	5/65 (7.7)	5/65 (7.7)	6/65 (9.2)	11/130 (8.5)
None	…	3/65 (4.6)	3/65 (4.6)	0	3/130 (2.3)
Vendor					
Hologic	73/128 (57)	28/65 (43.1)	101/193 (52.3)	26/65 (40)	127/258 (49.2)
GE HealthCare	55/128 (43)	37/65 (56.9)	92/193 (47.7)	39/65 (60)	131/258 (50.8)
Histologic type					
Invasive cancer	…	…	…	59/65 (90.8)	59/65 (90.8)
Noninvasive cancer	…	…	…	6/65 (9.2)	6/65 (9.2)

Note.—Data are numbers of cases, with percentages in parentheses, unless otherwise indicated. BI-RADS = Breast Imaging Reporting and Data System.
* Breast density was graded according to the American College of Radiology BI-RADS lexicon.

Table 3: Diagnostic Performance of AI and Radiologists without AI and with AI in Reader Study

Parameter	Stand-Alone AI	Without AI (Test 1)	With AI (Test 2)	*P* Value (Test 1 vs Test 2)	*P* Value (AI vs Test 1)
AUC	0.92 (0.87, 0.97)	0.90 (0.86, 0.93)	0.92 (0.88, 0.96)	.003*	.24†
AFROC AUC	0.77 (0.71, 0.83)	0.83 (0.81, 0.85)	0.87 (0.86, 0.89)	<.001*	<.001†
Sensitivity (%)	84.62 (75.84, 93.39)	85.44 (83.22, 87.65)	87.69 (85.63, 89.75)	.04‡	>.99§
Specificity (%)	89.64 (85.34, 93.94)	77.34 (75.82, 78.87)	79.62 (78.15, 81.09)	.10‡	<.001§
Reading time (sec)	…	54.41 (52.56, 56.27)	48.52 (46.79, 50.25)	<.001‖	…
Recall rate	29.07 (23.60, 35.02)	38.48 (36.94, 40.02)	37.34 (35.81, 38.84)	.31‡	.003§

Note.—Data in parentheses are 95% CIs. AFROC = alternative free-response receiver operating characteristic, AI = artificial intelligence, AUC = area under the receiver operating characteristic curve.
* *P* value estimated from analysis of variance F test.
† *P* value estimated from DeLong test.
‡ *P* value estimated from logistic regression with generalized estimating equation method.
§ *P* value estimated from χ^2 test.
‖ *P* value estimated from *t* test.

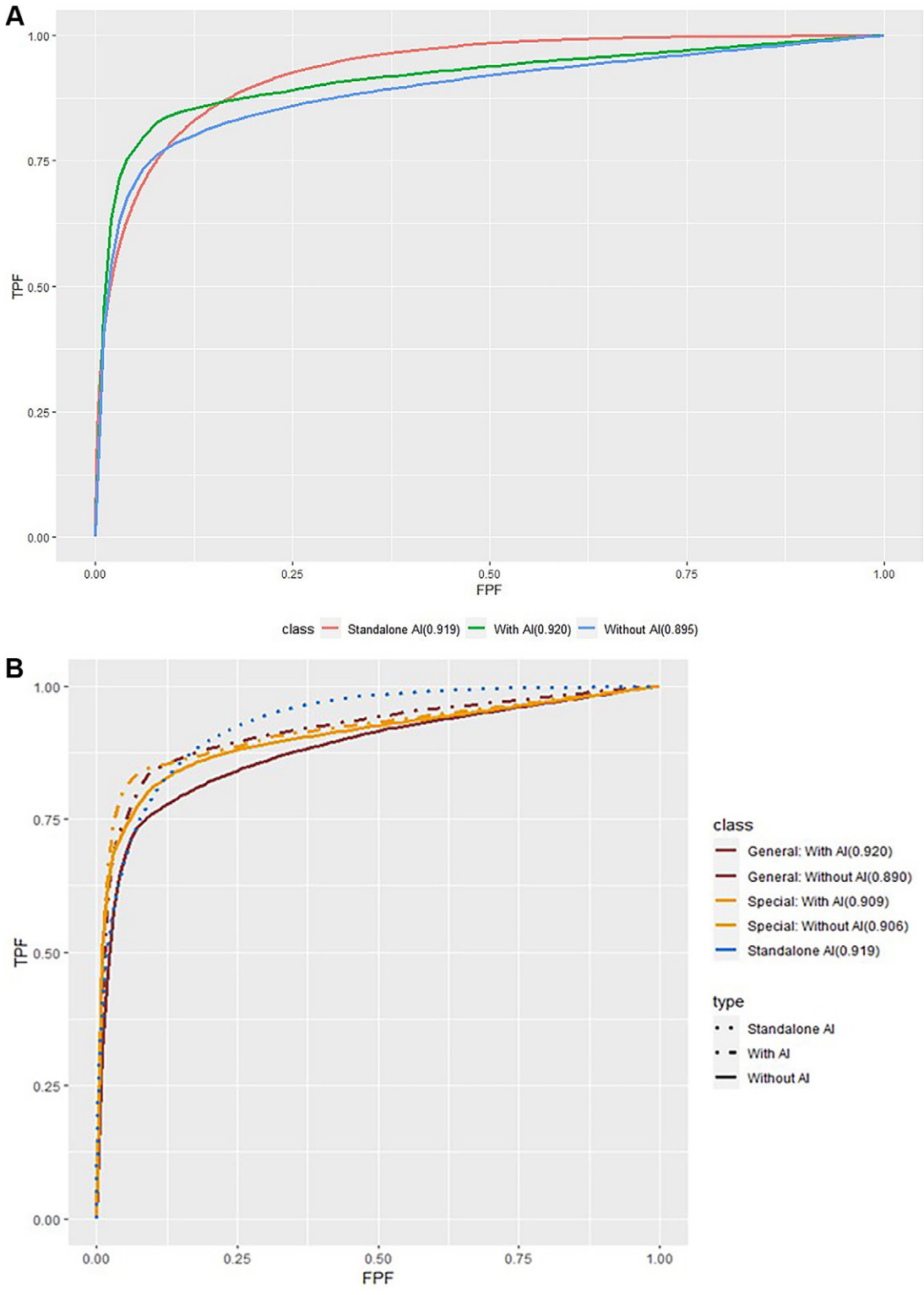

Figure 4: Receiver operating characteristic analysis with and without artificial intelligence (AI). As characteristic of a medical working career in breast imaging, breast specialists had been trained in breast imaging and had 75% or more of their professional time devoted to breast imaging in the last 3 years, whereas general radiologists had not been specifically trained in breast imaging and had less than 75% of their professional time devoted to breast imaging in the last 3 years. FPF = false-positive fraction (1 − specificity), TPF = true-positive fraction (case-level sensitivity).

Table 4: Subgroup Analysis and Comparisons between AI, Radiologists without AI, and Radiologists with AI

		Radiologists without AI		Radiologists with AI		
Feature	Stand-Alone AI	AUC	Reading Time (sec)	AUC	Reading Time (sec)	*P* Value*
Reading panel						
General	0.92 (0.87, 0.97)	0.89 (0.85, 0.93)	59.21 (56.37, 62.06)	0.92 (0.88, 0.96)	51.52 (49.16, 53.89)	.009
Specialist	0.92 (0.87, 0.97)	0.906 (0.87, 0.94)	48.84 (46.59, 51.08)	0.91 (0.86, 0.96)	45.04 (42.50, 47.57)	.79
Age						
<55 years	0.92 (0.83, >0.99)	0.92 (0.87, 0.98)	56.62 (53.97, 59.26)	0.94 (0.88, >0.99)	50.22 (47.98, 52.45)	.06
≥55 years	0.93 (0.87, 0.98)	0.89 (0.84, 0.94)	52.36 (49.76, 54.97)	0.92 (0.87, 0.97)	46.95 (44.33, 49.56)	.01
Breast density						
Fatty	0.95 (0.90, >0.99)	0.91 (0.86, 0.96)	53.85 (51.35, 56.36)	0.93 (0.89, 0.98)	49.33 (46.89, 51.77)	.04
Dense	0.89 (0.80, 0.97)	0.88 (0.82, 0.94)	55.16 (52.40, 57.92)	0.91 (0.85, 0.97)	47.44 (45.04, 49.83)	.005
Lesion features						
Soft tissue	0.89 (0.82, 0.97)	0.83 (0.77, 0.89)	51.75 (49.61, 53.90)	0.88 (0.83, 0.94)	42.25 (40.51, 43.99)	<.001
Calcifications	0.85 (0.69, >0.99)	0.82 (0.68, 0.96)	48.03 (45.42, 50.63)	0.84 (0.70, 0.97)	47.80 (45.09, 50.51)	.42
Both	1.00 (>0.99, >0.99)	0.99 (0.98, 1.00)	44.09 (40.36, 47.83)	0.99 (0.98, >0.99)	42.13 (38.20, 46.06)	.80
Vendor						
Hologic	0.98 (0.95, >0.99)	0.91 (0.85, 0.97)	58.25 (55.03, 61.46)	0.94 (0.89, 0.99)	50.97 (48.12, 53.82)	.008
GE HealthCare	0.89 (0.82, 0.96)	0.89 (0.83, 0.94)	50.76 (48.84, 52.69)	0.90 (0.85, 0.96)	46.20 (44.18, 48.21)	.08

Note.—Data in parentheses are 95% CIs. AI = artificial intelligence, AUC = area under the receiver operating characteristic curve.
* AUCs of radiologists without AI and with AI were compared using an analysis of variance F test.

improved significantly and reading times were significantly reduced. Interreader agreement in both general and specialist readers increased with the use of AI.

There is no standard for how to use AI in clinical workflow; however, robust independent performance of AI is essential to diverse implementations. According to the systematic review by Yoon et al (22), the pooled AUC of stand-alone performances of AI systems from previous DBT studies was 0.90 (standard error, 0.011). Although it is not ideal to compare in different settings, our study showed sufficient performance of the AI model, with an AUC of 0.93 (95% CI: 0.92, 0.94). Few studies have reported the stand-alone performance of diverse AI systems with sensitivity and specificity (20,25). Conant et al (20) reported 91% sensitivity and 41% specificity for 260 cases, and Shoshan et al (25) showed 96% sensitivity and 36% specificity, whereas our AI system had results of 88.57% sensitivity and 83.00% specificity. In our reader study results, AI for DBT had higher specificity (89.64% [95% CI: 85.34%, 93.94%]) than both general radiologists (74.09% [95% CI: 71.91%, 76.28%], $P < .001$) and breast specialists (81.05% [95% CI: 81.90%, 85.83%], $P = .005$), with comparable sensitivities.

There have been several studies on the use of AI systems for the interpretation of DBT examinations. Several studies investigated how to reduce workload for DBT studies with the concurrent use of AI or an AI-based simulation strategy (20,25–27). With consideration of the performance of radiologists in other reader studies with an AUC of 0.79 (standard error, 0.020) (22), although readers in our study had higher performance measured by an AUC of 0.90 (95% CI: 0.86, 0.93), the improvements we found with concurrent use of AI showed its feasibility in terms of diagnostic accuracy and efficiency. For general radiologists, the AUC significantly improved when interpreting with AI (0.92 [95% CI: 0.88, 0.96]) and was comparable to that of breast specialists without AI (0.91 [95% CI: 0.87, 0.94]) or with AI (0.91 [95% CI: 0.86, 0.96]). Our analysis also showed performance according to cancer subtypes, breast density, and lesion features at mammography to understand the impact of our AI system. Significantly improved AUCs and reduced reading times were observed in both fatty and dense breasts, with

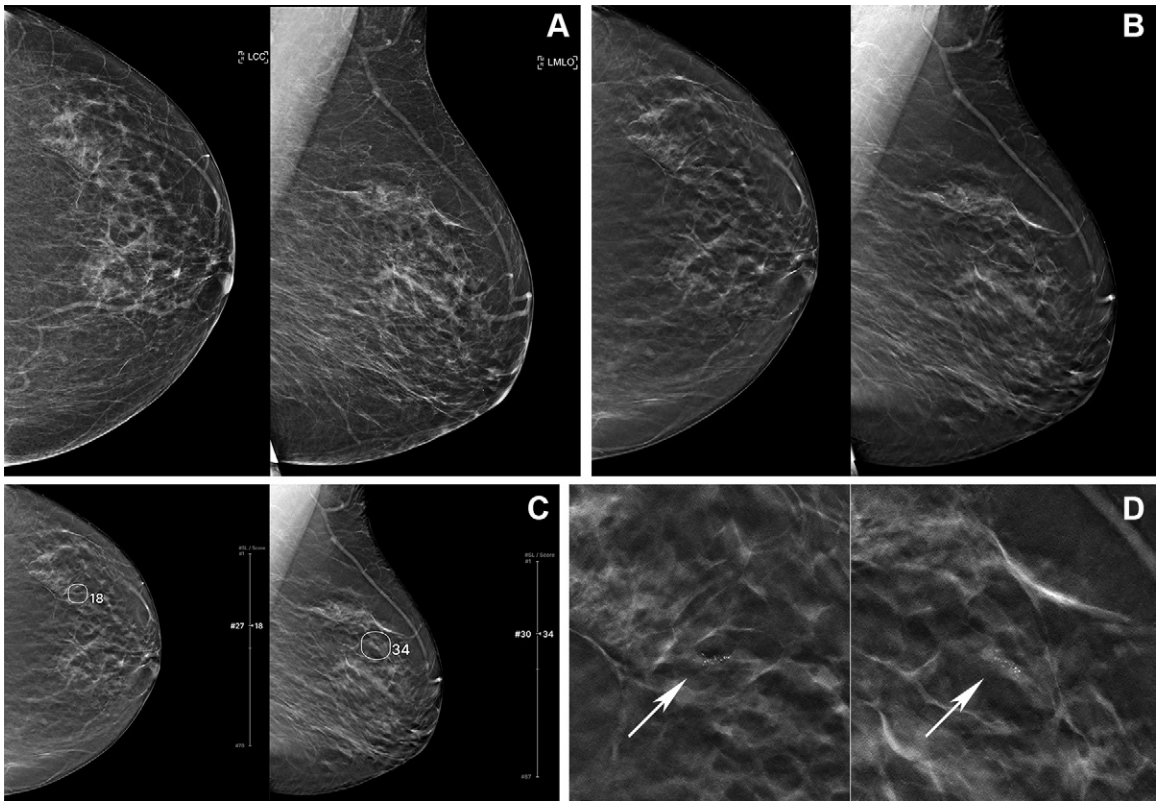

Figure 5: Digital mammography (DM) and digital breast tomosynthesis (DBT) images in a 61-year-old woman. **(A)** Left craniocaudal (LCC) and left mediolateral oblique (LMLO) views at DM and **(B)** LCC and LMLO views at DBT. In the reader study, six of 15 radiologists interpreted this case as negative for cancer. **(C)** The artificial intelligence (AI) algorithm for DBT images outlined the lesion with abnormality scores, which represent the likelihood of the presence of breast cancer, of 18 and 34, and five of the six readers changed their decision when aided by AI. **(D)** Zoomed-in LCC and LMLO DBT images show grouped fine and linear microcalcifications (arrows) in the left upper outer quadrant, which was histopathologically proven to be ductal carcinoma in situ.

increased sensitivity in fatty breasts and increased specificity in dense breasts. Improvements in sensitivity and specificity were greater for cases with soft-tissue lesions than for calcification-only lesions, suggesting that AI can be useful for detecting and diagnosing soft-tissue lesions at DBT. A previous reader study using an enriched cohort (20) demonstrated increased radiologist performance and a reduction in reading time with the concurrent use of a DBT AI system. They reported that improvements in sensitivity were greater for lesions with calcifications only than for soft-tissue lesions. This difference in performance across lesion types might come from different characteristics of independent AI systems. Conant et al (20) used an AI system with higher overall sensitivity (91% sensitivity, 41% specificity) and higher sensitivity in cases with only calcifications, whereas this study used a model with higher specificity (85% sensitivity, 90% specificity in the reader study cohort) and higher sensitivity in cases with soft-tissue lesions. To our knowledge, there has been no study to investigate and compare the impact of diverse AI systems on supporting DBT interpretation for radiologists; however, there is the potential to have greater gains in the concurrent use of AI by understanding its independent performance and the characteristics of AI systems.

The acquisition parameters for DBT vary substantially among different mammography vendors (28), and it is important for AI systems to be robust to such variations. To our knowledge, there has been no published study to show the use of AI in DBT with different DBT machines. Our AI showed comparable AUCs across the two different vendors; however, there was relatively low sensitivity and high specificity for the cases from the GE machine. In the reader study cohort, 26 (40.0% of cancer) cancer cases were interpreted by the radiologists when read in the clinic as BI-RADS category 0 (n = 25) or 3 (n = 1). Of these 26 studies, the majority (23 of 26, 88.5%) were from GE machines. For this subgroup, the sensitivity (21 of 26, 80.8%) of AI was not low when considering that it is challenging for radiologists as well to make an accurate diagnosis for cases with indeterminate suspicious findings. The readers' sensitivity and specificity for the cases from the GE machine had similar tendencies to those of AI. Although the development dataset of this AI system had fewer GE (18.5%) than Hologic machine cases (81.5%), the study cohort characteristics may be related to the differences between DBT vendors.

This study had several limitations. First, the reader study was performed with a cancer-enriched dataset. In the past few years, the use of AI for DBT in a retrospective screening cohort was investigated in a few studies (25,26,29), but it is still unclear whether the AI system would have led to a better screening strategy in real clinical practice. Second, the laboratory effect in reader studies (30) is well known; therefore, it may cause differences between actual and experimental settings. Third, while

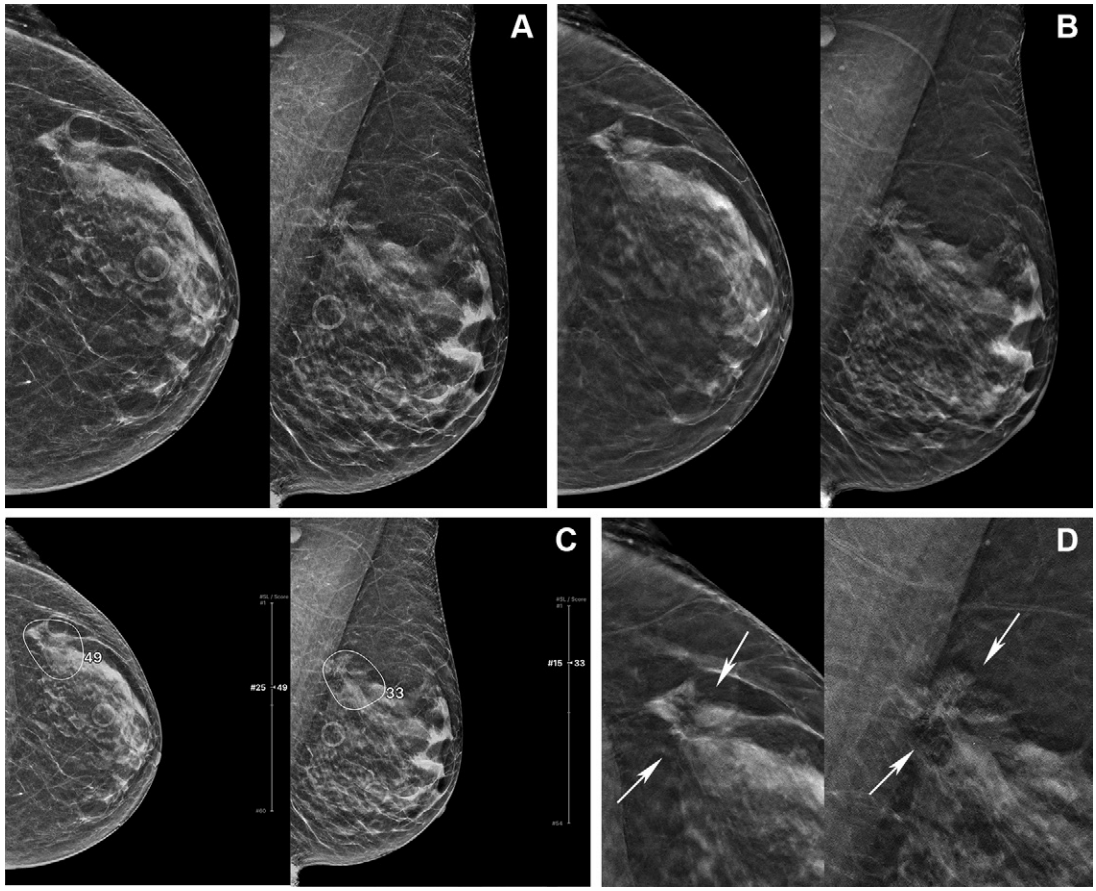

Figure 6: Digital mammography (DM) and digital breast tomosynthesis (DBT) images in a 72-year-old woman. **(A)** Left craniocaudal (LCC) and left mediolateral oblique (LMLO) views at DM and **(B)** LCC and LMLO views at DBT. In the reader study, four of 15 radiologists interpreted this case as negative for cancer. **(C)** The artificial intelligence (AI) algorithm for DBT images outlined the lesion with abnormality scores, which represent the likelihood of the presence of breast cancer, of 49 and 33, and all four readers changed their decision when aided by AI. **(D)** Zoomed-in LCC and LMLO DBT images show architectural distortion (arrows) in the left upper outer quadrant, which was histopathologically proven to be invasive ducal carcinoma.

we tried to curate a dataset with a balanced distribution, it still had limitations for evaluating the AI performance from two different vendors and making comparisons according to the characteristics of the study cohort through the subgroup analysis. Last, the AI system reads only a single examination, whereas, in clinical practice, radiologists typically compare the current study to previously recorded studies (31), which has been shown to improve the specificity of radiologists at 2D digital mammography. Despite not using this information, the AI system showed excellent performance; however, future work could include the addition of prior studies to further boost performance, especially increased specificity.

In conclusion, the AI model we developed showed better diagnostic accuracy than radiologists in breast cancer detection. Our reader study demonstrated that the concurrent use of AI in DBT interpretation has the potential to improve both accuracy and efficiency. To have promising evidence for the impact of AI in breast cancer screening, prospective studies are needed to validate our findings with diverse cohorts and screening programs.

Acknowledgments: The authors acknowledge medical support, including educating annotators in the reader study from Ambika Seth, MD, who is a medical director at Lunit; managing clinical research from Seungjin Rhee, BS, who is a medical writer at Lunit; Kyungjee Min, MS, who is a clinical research manager at Lunit; technical support from Seokhwan Hwang, BS, who is a backend engineer at Lunit; and Heesung Yang, BS, who is a solution engineer at Lunit.

Author contributions: Guarantor of integrity of entire study, **E.K.P.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **E.K.P., W.L., T.K.**; clinical studies, **E.K.P., J.S.C.**; experimental studies, **W.L., T.K.**; statistical analysis, **E.K.P., S.Y.K., T.K.**; and manuscript editing, **E.K.P., S.Y.K., W.L., E.K.K.**

Disclosures of conflicts of interest: **E.K.P.** Funding from Lunit; employee and stockowner of Lunit. **S.Y.K.** Employee of Lunit. **W.L.** Employee of Lunit; stock/stock options in Lunit. **J.S.C.** Employee of Lunit; stock and stock options in Lunit. **T.K.** Funding from Lunit; support for attending meetings/travel from Lunit; patents planned, issued, or pending with Lunit; stock/stock options in Lunit. **E.K.K.** No relevant relationships.

References

1. Hakama M, Coleman MP, Alexe DM, Auvinen A. Cancer screening: evidence and practice in Europe 2008. Eur J Cancer 2008;44(10):1404–1413.
2. Paci E; EUROSCREEN Working Group. Summary of the evidence of breast cancer service screening outcomes in Europe and first estimate of the benefit and harm balance sheet. J Med Screen 2012;19(Suppl 1):5–13.
3. Duffy SW, Vulkan D, Cuckle H, 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(9):1165–1172.

4. Carney PA, Miglioretti DL, Yankaskas BC, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. Ann Intern Med 2003;138(3):168–175.
5. Bernardi D, Macaskill P, Pellegrini M, et al. Breast cancer screening with tomosynthesis (3D mammography) with acquired or synthetic 2D mammography compared with 2D mammography alone (STORM-2): a population-based prospective study. Lancet Oncol 2016;17(8):1105–1113.
6. Ciatto S, Houssami N, Bernardi D, et al. Integration of 3D digital mammography with tomosynthesis for population breast-cancer screening (STORM): a prospective comparison study. Lancet Oncol 2013;14(7):583–589.
7. Aase HS, Holen ÅS, Pedersen K, et al. A randomized controlled trial of digital breast tomosynthesis versus digital mammography in population-based screening in Bergen: interim analysis of performance indicators from the To-Be trial. Eur Radiol 2019;29(3):1175–1186.
8. Caumo F, Zorzi M, Brunelli S, et al. Digital breast tomosynthesis with synthesized two-dimensional images versus full-field digital mammography for population screening: Outcomes from the verona screening program. Radiology 2018;287(1):37–46.
9. Conant EF, Barlow WE, Herschorn SD, 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(5):635–642.
10. Conant EF, Beaber EF, Sprague BL, et al. Breast cancer screening using tomosynthesis in combination with digital mammography compared to digital mammography alone: a cohort study within the PROSPR consortium. Breast Cancer Res Treat 2016;156(1):109–116.
11. U.S. Food and Drug Administration. MQSA national statistics. https://www.fda.gov/radiation-emitting-products/mqsa-insights/mqsa-national-statistics. Accessed June 1, 2022.
12. Sprague BL, Coley RY, Kerlikowske K, et al. Assessment of radiologist performance in breast cancer screening using digital breast tomosynthesis vs digital mammography. JAMA Netw Open 2020;3(3):e201759.
13. 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(1):49–56.
14. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. N Engl J Med 2007;356(14):1399–1409.
15. Lehman CD, Wellman RD, Buist DS, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med 2015;175(11):1828–1837.
16. Fenton JJ, Abraham L, Taplin SH, et al. Effectiveness of computer-aided detection in community mammography practice. J Natl Cancer Inst 2011;103(15):1152–1161.
17. Kim HE, Kim HH, Han BK, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Health 2020;2(3):e138–e148.
18. 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(10):1581–1588.
19. Dembrower K, Wåhlin E, Liu Y, 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(9):e468–e474.
20. Conant EF, Toledano AY, Periaswamy S, et al. Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis. Radiol Artif Intell 2019;1(4):e180096.
21. Romero-Martín S, Elías-Cabot E, Raya-Povedano JL, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. Stand-alone use of artificial intelligence for digital mammography and digital breast tomosynthesis screening: a retrospective evaluation. Radiology 2022;302(3):535–542.
22. Yoon JH, Strand F, Baltzer PAT, et al. Standalone AI for breast cancer detection at screening digital mammography and digital breast tomosynthesis: a systematic review and meta-analysis. Radiology 2023;307(5):e222639.
23. Sechopoulos I. A review of breast tomosynthesis. Part I. The image acquisition process. Med Phys 2013;40(1):014301.
24. Lee W, Lee H, Lee H, Park EK, Nam H, Kooi T. Transformer-based deep neural network for breast cancer classification on digital breast tomosynthesis images. Radiol Artif Intell 2023;5(3):e220159.
25. Shoshan Y, Bakalo R, Gilboa-Solomon F, et al. Artificial intelligence for reducing workload in breast cancer screening with digital breast tomosynthesis. Radiology 2022;303(1):69–77.
26. 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(1):57–65.
27. van Winkel SL, Rodríguez-Ruiz A, Appelman L, et al. Impact of artificial intelligence support on accuracy and reading time in breast tomosynthesis image interpretation: a multi-reader multi-case study. Eur Radiol 2021;31(11):8682–8691.
28. Yoon JH, Han K, Suh HJ, Youk JH, Lee SE, Kim EK. Artificial intelligence-based computer-assisted detection/diagnosis (AI-CAD) for screening mammography: Outcomes of AI-CAD in the mammographic interpretation workflow. Eur J Radiol Open 2023;11:100509.
29. Pinto MC, Rodriguez-Ruiz A, Pedersen K, et al. Impact of artificial intelligence decision support using deep learning on breast cancer screening interpretation with single-view wide-angle digital breast tomosynthesis. Radiology 2021;300(3):529–536.
30. Gur D, Bandos AI, Cohen CS, et al. The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. Radiology 2008;249(1):47–53.
31. Hayward JH, Ray KM, Wisner DJ, et al. Improving screening mammography outcomes through comparison with multiple prior mammograms. AJR Am J Roentgenol 2016;207(4):918–924.