Egyptian Journal of Radiology
and Nuclear Medicine

RESEARCH Open Access

Artificial intelligence as an initial reader for double reading in breast cancer screening: a prospective initial study of 32,822 mammograms of the Egyptian population

Sahar Mansour[1,2*], Enas Sweed[2,3], Mohammed Mohammed Mohammed Gomaa[2,4], Samar Ahmed Hussein[2,4], Engy Abdallah[2], Yassmin Mohamed Nada[2], Rasha Kamal[1,2], Ghada Mohamed[2], Sherif Nasser Taha[2,4] and Amr Farouk Ibrahim Moustafa[2,4]

Abstract

Background Although artificial intelligence (AI) has potential in the field of screening of breast cancer, there are still issues. It is vital to make sure AI does not overlook cancer or cause needless recalls. The aim of this work was to investigate the effectiveness of indulging AI in combination with one radiologist in the routine double reading of mammography for breast cancer screening. The study prospectively analyzed 32,822 screening mammograms. Reading was performed in a blind-paired style by (i) two radiologists and (ii) one radiologist paired with AI. A heatmap and abnormality scoring percentage were provided by AI for abnormalities detected on mammograms. Negative mammograms and benign-looking lesions that were not biopsied were confirmed by a 2-year follow-up.

Results Double reading by the radiologist and AI detected 1324 cancers (6.4%); on the other side, reading by two radiologists revealed 1293 cancers (6.2%) and presented a relative proportion of 1·02 ($p < 0·0001$). At the recall stage, suspicion and biopsy recommendation were more presented by the AI plus one radiologist combination than by the two radiologists. The interpretation of the mammogram by AI plus only one radiologist showed a sensitivity of 94.03%, a specificity of 99.75%, a positive predictive value of 96.571%, a negative predictive value of 99.567%, and an accuracy of 99.369% (from 99.252 to 99.472%). The positive likelihood ratio was 387.260, negative likelihood ratio was 0.060, and AUC "area under the curve" was 0.969 (0.967–0.971).

Conclusions AI could be used as an initial reader for the evaluation of screening mammography in routine workflow. Implementation of AI enhanced the opportunity to reduce false negative cases and supported the decision to recall or biopsy.

Keywords Artificial intelligence, Screening mammogram, Breast cancer screening, Workflow, Double reading, Recall rate

*Correspondence:
Sahar Mansour
sahar_mnsr@cu.edu.eg
[1] Women's Imaging Unit, Kasr ElAiny Hospital, Cairo University, Cairo, Egypt
[2] Baheya center for early breast cancer detection and treatment, Cairo, Egypt
[3] Banha University, Banha, Egypt
[4] National Cancer Institute, Cairo Université, Cairo, Egypt

Background

Effective breast cancer screening programs have been developed by utilizing strategies for improving breast cancer prevention, screening, early diagnosis, and treatment in a way that provides a proper outcome, prevents care delays, and minimizes unnecessary waste of medical resources [1].

The majority of screening programs confirm double reading, which is typically followed by actions to help recall decisions for women whose mammograms identify abnormalities [2–4].

A countrywide program in Egypt for screening breast cancer at an early stage examined 28 million women governorate-wide as part of the "100 million Healthy Individuals" presidential campaign and was introduced in July 2019 [5].

Among the barriers to fully implementing the proper strategy and international guidelines for screening breast cancer is the system delay in providing health services in diagnosis and referral for treatment [6]. In addition, there is a global shortage of breast radiologists [7].

Artificial intelligence (AI) has currently emerged as a solution for these challenges [8].

Several retrospective studies based on clinical data sets made use of the consecutive examinations, providing an opportunity to assess AI systems as independent readers for breast cancer on mammograms [8–16]. However, many prospective evaluations need to be created in order to investigate the use of AI in routine daily jobs and to analyze the histopathologic variations of cancers that are correctly identified.

In the current work, a prospective analysis was performed to study the ability to include AI in the routine double reading of screening mammography with one human radiologist to reduce the reading load and support the radiologist's decision of a negative or abnormal mammogram. To date no research had been conducted on this issue in the Middle East or Africa.

Methods

The study is a prospective double-reader analysis that was performed in "Baheya" center of excellence, which is a non-profitable and non-governmental private center for early breast cancer detection and treatment. The current work is an initial experience that was approved by the ethics committee of the research center and included 32,822 mammograms of 16,801 females.

The study was conducted during the following time frames: cases collected in 6 months duration (January 2021 till June 2021); then, the negative assigned mammograms and the unproved benign lesions were followed up from January 2021 till June 2023 (2 years).

Non-eligible cases ($n=5546$) were: (i) known breast cancer patients who were included in the surveillance program ($n=4425$; 3006 performed conservative or reconstructive muscle flap surgeries, 639 had reconstructive autologous implant applications, and 780 had mastectomies); ii) patients with bilateral breast implants for cosmeses ($n=1121$).

Reading was performed in a paired style by two settings: (i) two radiologists and (ii) one radiologist paired with AI. Radiologists were blinded to the abnormality scoring assigned by AI in case the AI was the second reader and the assignments of the other radiologist in the "two-radiologist reader" setting.

The study population included females in the age range of 40–75 years old, with a mean age of $51 \pm SD=9$ years old.

Equipment

A digital mammogram device (Senographe Prestina 3D, GE Healthcare, United Kingdom) was used for the study of the cases. The medio-lateral oblique and the cranio-caudal views were done for each patient.

The used workstation to evaluate mammograms was a two-monochrome 5-megapixel liquid crystal display (2048 × 2560 pixels; 21.3 inches; MFGD5621HD, Barco).

Mammograms were scanned and read by "Lunit INSIGHT MMG," an artificial intelligence solution (Seoul, South Korea, FDA-approved in 2019) for reading mammograms (AI-MMG).

Image analysis and interpretation

Mammograms were reviewed in consensus by three radiologists (with 25 years of experience in breast imaging). Two "double reading" settings were used; in one setting, a human radiologist was paired with a human radiologist, and in the other setting, one radiologist was paired with AI reading. AI functioned as a stand-alone reader.

The included mammograms were sorted as follows: (i) *abnormal mammogram, no cancer*: BI-RADS 2 or 3, (ii) *abnormal mammogram, suspicious*: BI-RADS 4 or 5, (iii) *recall benign, no biopsy*: abnormal mammogram was monitored and subsequently confirmed to be benign after a two-year follow-up period, (iv) *recall suspicious*: abnormal mammogram with suspicious lesions detected on recall by complementary modalities, e.g., digital breast tomosynthesis, breast ultrasound, contrast-enhanced mammography, and/or dynamic post-contrast MR imaging, (v) *recall suspicious, no cancer: the* mammogram was suspicious for cancer, yet complementary modalities presented a benign-looking lesion that was confirmed by follow-up, (vi) *recall suspicious, proven benign:* mammograms with suspicious or malignant looking abnromlities, yet biopsy proved benign pathology, (vii) *recall* *suspicious proven malignant:* biopsy-proven abnormal suspicious mammograms. The initial read was performed by a human radiologist (in the two-radiologist group) and AI (in the reader paired by AI group), followed by a consensus second reading session, which was performed this time by human radiologists in both "double reading" settings and consequently would confirm or doubt the result.

If a discrepancy happened in the paired reading (by one of the two readers or between the reader and AI), a third reader (with 35 years' experience) was considered to decide the outcome (negative, recall, or biopsy).

Recalled female patients were subjected to spot magnification views, digital breast tomosynthesis, breast ultrasonography, contrast-enhanced mammography, and/or dynamic post-contrast MR imaging if malignant-looking (BI-RADS 5) or suspicious (BI-RADS 4) features were discovered [17].

Heatmaps and percentages of suspicion for abnormalities seen on mammograms were produced by AI-MMG.

The scoring percentage included: 100% for definite cancers, 76–99% for probably cancer, 51–75% for possibly cancer, 26–50% for possibly non-cancer, 10–25% for probably non-cancer, and 0–9 for definite non-cancer [18]. The term "low" noted at the bottom of the AI-MMG image suggests a low risk of < 10% of cancer.

A core biopsy was taken using a 14G needle to confirm malignancy, followed by surgical removal of the abnormality. The standard reference was pathology for malignant and 125 benign lesions. Negative cases and benign-looking lesions that were not biopsied were confirmed by stability on interval follow-up for 2 years.

Statistical analysis

Data analysis was performed with commercially available software (IBM SPSS Statistics for Windows version 24.0.2.). Data were summarized using mean, standard deviation, median, minimum, and maximum in quantitative data and using frequency (count) and relative frequency (percentage) for categorical data. Standard diagnostic indices, including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and diagnostic efficacy, were calculated. The correlation of classified groups in relation to the abnormality score by AI was analyzed with the Kruskal–Wallis test. The comparison of abnormality scores among different classification groups was performed with a post hoc Conover correction for multiple comparisons. The significant threshold was set at 0.05. The ROC curve was constructed with an area under curve analysis performed to detect the best cutoff value of AI for the detection of malignant masses. For comparing categorical data, a Chi-square 'χ^2' test was performed. *P*-value less than 0.05 was considered statistically significant. The "Youden" index was used to determine the diagnostic performance of AI in screening breast cancer and estimate the optimal cut point from observations.

Results

The study included 32,822 mammograms of 16,801 females (780 had mastectomies) presented for breast screening.

The eligible mammograms were 20,764 (10,382 females), as shown in Fig. 1.

Pathologically proven lesions were malignant in 1408 (6.8%) and benign in 125 (0.6%).

The included proven carcinoma was diagnosed malignant at first presentation in 1247 (88.6%).

The remaining proved malignant mammograms (11.4%, $n = 161/1408$) were assigned benign/negative at the initial screening examination, then interval cancer was detected by AI on follow-up in 102 lesions (7.2%) in the first year, and extra 59 lesions were detected (4.2%) in the second year, as shown in Fig. 2

Unproved benign-looking lesions were noted in 1454 (7%) mammograms, and negative opinion was assigned for 17,777 (85.6%) mammograms.

Table 1 displays the benign and malignant pathology variants included in the study.

Bilaterality of the same nature of the breast disease (benign vs malignant) detected in 74 females (148 mammograms: 28 proved malignant, 4 proved benign, and 116 benign proved by the stationary course on follow-up).

Multiplicity was detected in 21 malignant proved mammograms (1.5%, $n = 21/1408$).

Double reading by radiologist and AI detected 1324 cancers (6.4%); on the other side, reading by two radiologists revealed 1293 cancers (6.2%) and presented a relative proportion of 1·02 ($p < 0·0001$). The strategy of double reading with one radiologist and AI was superior to double reading by two radiologists (Figs. 3, 4, Table 2)**.**

For mammograms, given the scoring percentage of "probably malignant" (i.e., 76–99%), AI was the positive initial reader ($n = 31$), and the accurate targeting of the biopsy site for confrimed malignancy was AI-based; the intensity of the color hue delineated the most suspicious regions of the abnormality, thereby assisting the radiologist in accurately positioning the biopsy needle within the lesion.

The proportion of abnormal interpretations diagnosed as benign was higher by two the radiologists than by the AI–reader combination. On the other side, the double radiologists diagnosed a lower proportion of suspicious mammograms.

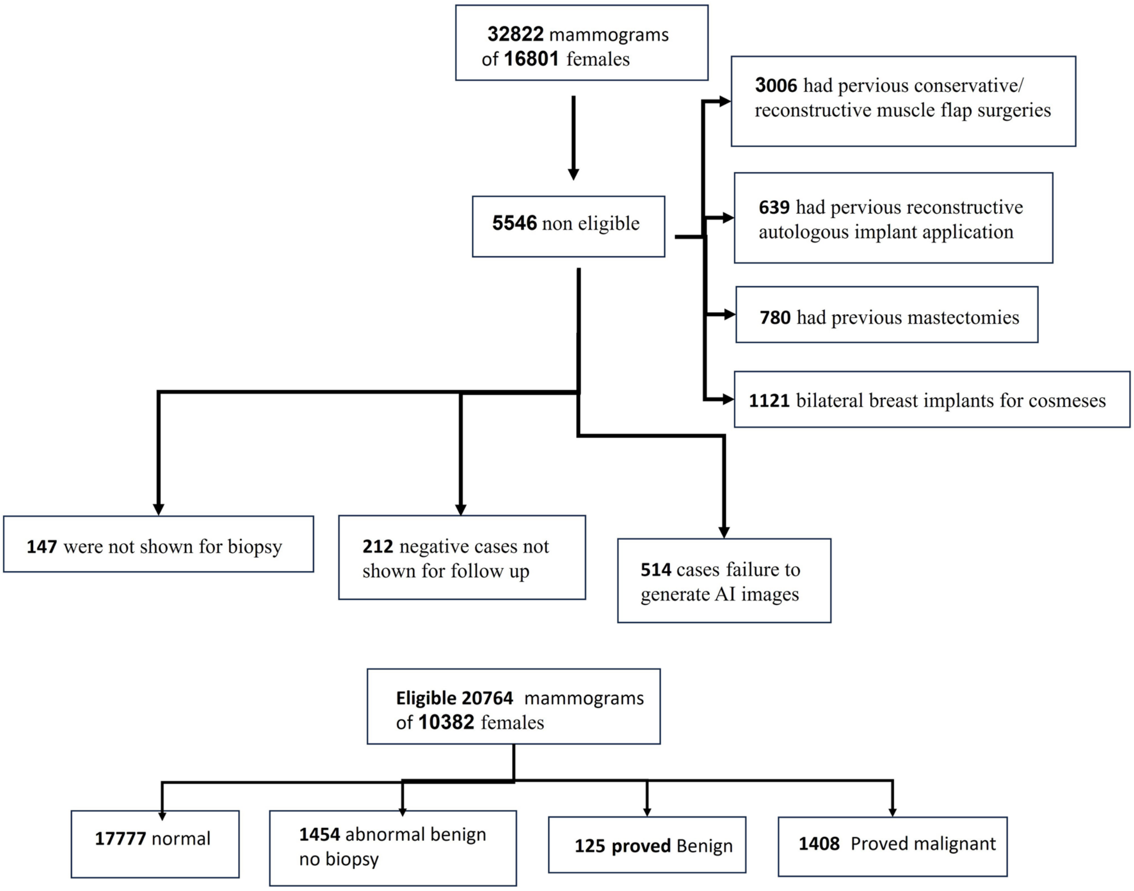
32822 mammograms of **16801** females
5546 non eligible
3006 had pervious conservative/ reconstructive muscle flap surgeries
639 had pervious reconstructive autologous implant application
780 had previous mastectomies
1121 bilateral breast implants for cosmeses
147 were not shown for biopsy
212 negative cases not shown for follow up
514 cases failure to generate AI images
Eligible 20764 mammograms of **10382** females
17777 normal
1454 abnormal benign no biopsy
125 proved Benign
1408 Proved malignant

Fig. 1 Flow chart of the study population selection of eligible and excluded cases

At the recall stage, an ultrasound examination was performed by a radiologist, and the findings were correlated with the abnormality scoring percentage marked by the AI at the screening mammogram. Suspicion and biopsy recommendations were presented more by the AI plus one radiologist combination than by the two radiologists. For the sake of the patient, if lesions looked suspicious, a biopsy was performed even if AI assigned a "low," i.e., < 10%, abnormality score, as shown in Figs. 5, 6.

The study was performed at a breast cancer specialist institute that served the whole nation, and so the included cases were traced over two years. This gave us the opportunity to confirm negative and benign mammograms.

Table 3 represents a demonstration of true positive, true negative, false positive, and false negative mammograms in correlation with follow-up findings for normal ($n = 17{,}777$) and benign-looking lesions ($n = 1454$). Pathology via biopsy or surgery for 125 benign and all malignant lesions was the standard reference.

There was no statistically significant difference in the performance of two readers versus AI plus one reader in interpreting mammograms ($P < 0.0001$), and the effect size value was 3.7.

Cancers that displayed less than 50% abnormality scoring of suspicion ($n = 439$), as shown in Figs. 2, 6, and 7, were: invasive ductal carcinoma in 63.1% ($n = 277$), invasive lobular carcinoma in 20.7% ($n = 91$), ductal carcinoma in situ in 8.2% ($n = 36$), mixed invasive ductal and lobular in 2.1% ($n = 9$), mixed tubular cribriform in 1.8% ($n = 8$), invasive cribriform in 4% ($n = 0.9\%$), invasive mucinous in 0.7 ($n = 3$), invasive tubular in 0.7% ($n = 3$), invasive carcinoma with medullary features in 0.5% ($n = 2$), invasive micropapillary carcinoma in 0.5% ($n = 2$), invasive papillary in 0.2% ($n = 1$), metaplastic carcinoma in 0.2% ($n = 1$), mixed invasive ductal cribriform tubular in 0.2% ($n = 1$), mixed invasive ductal mucinous in 0.2% ($n = 1$).

The double reading strategy by two radiologists displayed a sensitivity of 91.832% (90.277–93.210%), and a specificity of 99.365% (99.242–99.472%), a positive predictive value of 91.312% (89.803–92.617%), a negative

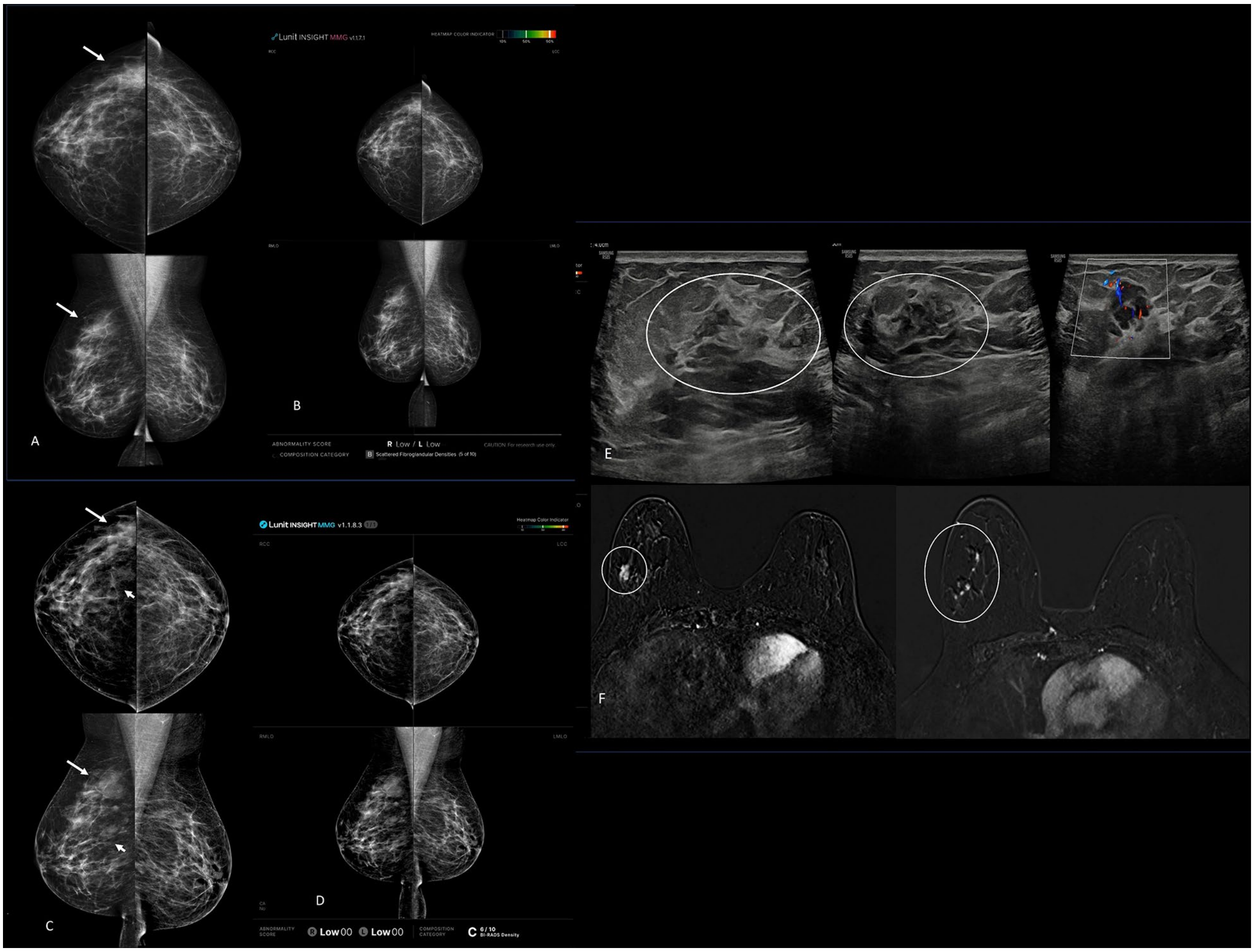

Fig. 2 Right breast multicentric invasive ductal carcinoma grade II was overlooked by the radiologist and AI in the primary screening mammogram and detected by only by the radiologist in the annual follow-up mammogram. **A** Screening digital mammogram revealed right breast upper outer asymmetry (arrow) that represented no markings by in AI-MMG four view image (**B**). On recall, the abnormality was diagnosed as prominent glandular tissue by complementary ultrasound examination and recommended for annual mammographic follow-up. **C** Follow-up digital mammogram showed upper outer suspicious mass in place of the previously detected asymmetry and another deep central tiny area of distortion (arrows). **D** AI-MMG displayed no marking of the right breast newly developed suspicious abnormalities assigned by the radiologist. E. ultrasound images and F. subtraction post-contrast MR images confirming the malignant-looking morphology and multiplicity of the proved carcinoma (circles)

predictive value of 99.406% (99.293–99.501%), and an accuracy of 98.854%. Positive likelihood ratio (LHR) was 144.513, negative LHR was 0.082, and AUC was 0.956 (0.953–0.959).

In data analysis, the most recent AI abnormality scoring value was the one considered for cases on follow-up since different scoring was presented each year, as shown in Figs. 2 and 8.

The interpretation of the mammogram by AI plus one radiologist showed a sensitivity of 94.034% (92.667–95.214%), a specificity of 99.757% (99.677–99.822%), a positive predictive value of 96.571% (95.488% to 97.402%), and a negative predictive value of 99.567% (99.468 to 99.648%) and accuracy of 99.369% (99.252 to 99.472%).

Positive LHR was 387.260, negative LHR was 0.060, and AUC was 0.969 (0.967–0.971).

For screening mammograms, the Youden index (J) was 0.7395, for a cutoff AI scoring value of more than 14%.

The mean AI score for malignant lesions in the current study was 68% (95% confidence interval, 66–70%). Sixty-nine percent of these lesions elicited more than 50% AI scoring ($n = 969/1408$), as shown in Figs. 3, 4.

Table 1 Pathology variants presented in the study

Benign		
UDH with no atypia	59	47.2%
Fat necrosis	13	10.4%
Complex adenoma	12	9.6%
Focal granulomatous mastitis	5	4.0%
Fibroadenosis	21	16.8%
Fibrocystic mastopathy	15	12.0%
Total	125	100%
Malignant		
Adenoid cystic carcinoma	1	0.07%
DCIS	81	5.8%
Encapsulated papillary carcinoma	4	0.3%
IDC	1119	79.5%
ILC	115	8.2%
Invasive carcinoma with medullary features	5	0.4%
Invasive cribriform	10	0.7%
Invasive micropapillary carcinoma	2	0.1%
Invasive mucinous	12	0.9%
Invasive papillary carcinoma	1	0.07%
Invasive tubular	8	0.6%
Metaplastic carcinoma	1	0.07%
Mixed IDC/cribriform /tubular	1	0.07%
Mixed IDC/mucinous	1	0.07%
Mixed invasive ductal and lobular	26	1.8%
Mixed tubular/cribriform	21	1.5%
Total	1408	100.0%

Discussion

Retrospective studies were conducted to evaluate the independent usage of AI systems on consecutive breast screening examinations, but little work has been published for prospective analysis. For true evaluation of AI as a reliable party, radiologists need to practice AI reading of cancer across a large volume of mammogram and learn more about the strengths and challenges of adopting the AI tool in the triage.

Even the most trusted AI tool will fail if it is not properly integrated into existing workflows [19].

The current work is an initial experience that analyzed 16,801 females presented for breast screening from all over the country. The aim was to study the ability to implant AI in the daily screening mammogram workflow to support the radiologist's decision on a negative or abnormal mammogram. AI was evaluated in combination with a human reader (an experienced radiologist) for reading mammograms instead of the routinely used double reading strategy by two radiologists.

Human readers learned about breast cancer through morphology descriptors, e.g., shape, texture, margin, orientation, etc., unlike the AI algorithm, which is provided by several image examples of cancer and teaches itself what it looks like [20].

A large retrospective study used data from real-world deployments included 275,900 mammograms from four mammography equipment vendors collected across seven screening sites in two countries aimed to evaluate AI as an independent reader in the double reading workflow for breast cancer screening. The study underscored the transformative role of AI in breast cancer screening, offering a cost-effective solution to enhance cancer detection rates where sensitivity was comparable to human reader. Specificity and positive predictive value were even superior to the radiologist performance [21].

In this work, double reading by one radiologist plus AI resulted in an increased rate of detected cancer by 2.4% (31/1293) in correlation with double reading by two radiologists. The rate of doing a biopsy was also increased by 3.5% (50/1446), while the number of recalls for benign abnormalities was decreased by 3.2% (50/1541). In view of this, it is suggested that AI and human readers have near-diagnostic performance for cancer, although they may have different ways of recognizing cancer.

Similar observations were made in recent prospective real-world clinical practice, which analyzed the effect of AI when used in the daily screening practice as an assistant reader and found a significant 5–13% increase in the rate of early detection of mostly invasive and small cancerous tumors and a recall rate of 6.7–7.7% [22].

The combination of AI and radiologist double reading will decrease the recall rate for non-cancer abnormalities. Yet biopsy was always the option in case the lesion was suspicious, irrespective of the abnormality scoring percentage of AI, as shown in Figs. 4, 5, and 6. The capability of the human reader to detect breast cancer would increase if artificial intelligence was included in the workflow for screening mammograms [23].

During our experience, 161 interval carcinomas (11.4%, $n=161/1408$) were found. Park et al. [24] retrospective study about detection of missed carcinoma by AI in mammogram was in coordination where the diagnostic rate of AI for the interval cancer was near equal to our work (86.7% vs 87.6%); however, the detection rate was higher than our results (67.2% vs 36.6%). This may be due to the difference in the study design (prospective vs retrospective) and the sample size (1408 vs 204 malignant lesions) between both studies.

A Swedish prospective study [8] calibrated an AI-based retrospective study with the goal of achieving a 2% increase in false positive rate, but in the real routine screening workflow, it produced a 6% increase. This led them to conclude that setting AI statistical indices based on retrospective analysis may not always be sufficient and

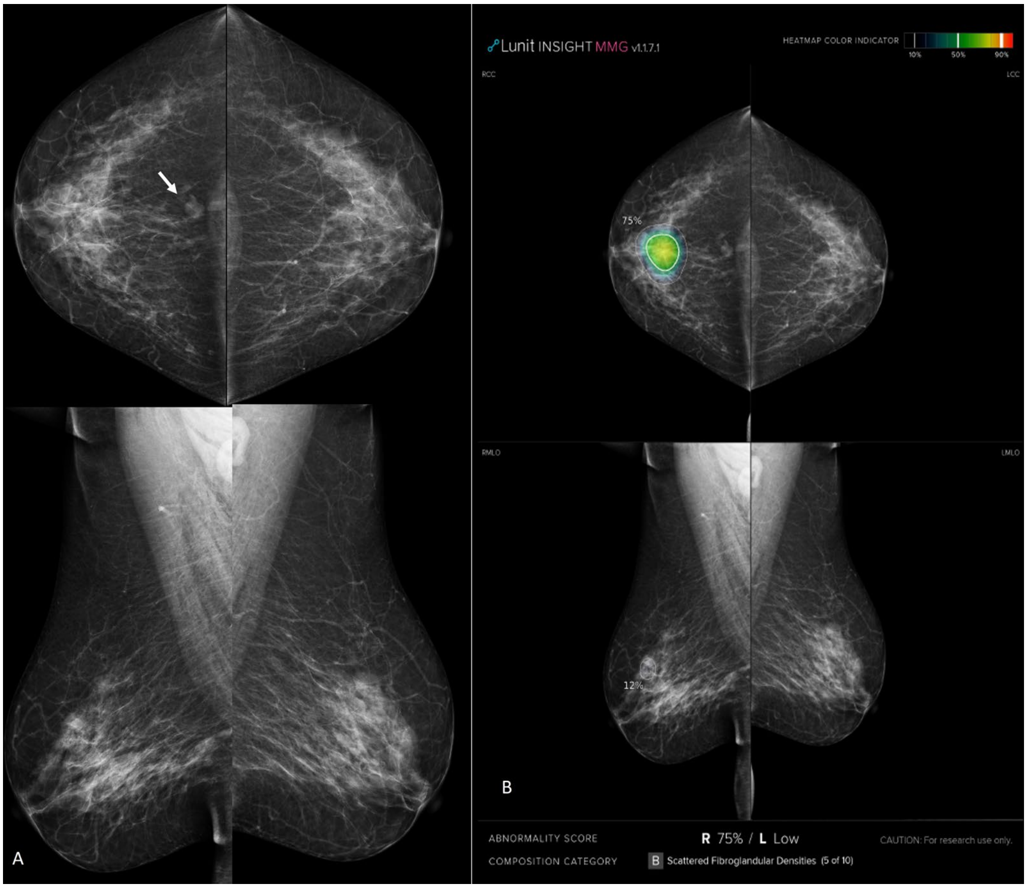

Fig. 3 **A** Right breast carcinoma was overlooked by the radiologist and presented as a deeply seated lobulated mass (arrow) that, on recall, was found to be a cluster of simple cysts. **B** AI-scanned mammograms marked the carcinoma with a precision of 75%. Double reading with AI prevented a delay in diagnosis and, consequently, treatment of the patient, which will potentially be subjected to an altered and unsuitable prognosis

that repeated calibration of the AI solutions in clinical practice is recommended to achieve optimal results.

A comparable prospective multicenter cohort Korean experience was carried to generate a real-world evidence on the benefits and disadvantages of using artificial intelligence-based computer-aided detection/diagnosis (AI-based CADe/x) for breast cancer detection in a population-based screening program for Korean women aged 40 years and older. The population was 32,714 participants enrolled at five different study sites in Korea. Mammography readings were performed with or without the use of AI-based CADe/x, and if recall was required, further diagnostic workup was used for confirming the detected cancers. The study is currently in the patient enrollment phase. The National Cancer Registry Database will be reviewed in 2026 and 2027, and the results of this study are expected to be published in 2027 [25].

Recently, the prospective MASAI “Mammography Screening with Artificial Intelligence” trial reported on the outcomes of using AI in screening mammography, which agreed with our results. The study has shown that a strategy of double reading by one radiologist plus AI resulted in an increased cancer detection rate compared with double reading by two radiologists [16].

The cancer detection rate and abnormal interpretation rate were in line with previous retrospective studies [9, 11–15, 26, 27].

Schaffler et al. [28] performed training and validation for an AI algorithm using overall 144,231 screening mammograms (952 cancer positive more than 12 months from screening) used for algorithm training and validation. The group found that combining the algorithm and radiologist assessments resulted in high performance and achieved a high area under the curve of 0.942 with a significantly improved specificity (92.0%) at the same sensitivity of the radiologist.

The double reading strategy by two radiologists displayed a sensitivity of 91.832% (90.277–93.210%), and a specificity of 99.365% (99.242–99.472%) compared to a sensitivity of 94.034% (92.667–95.214%),

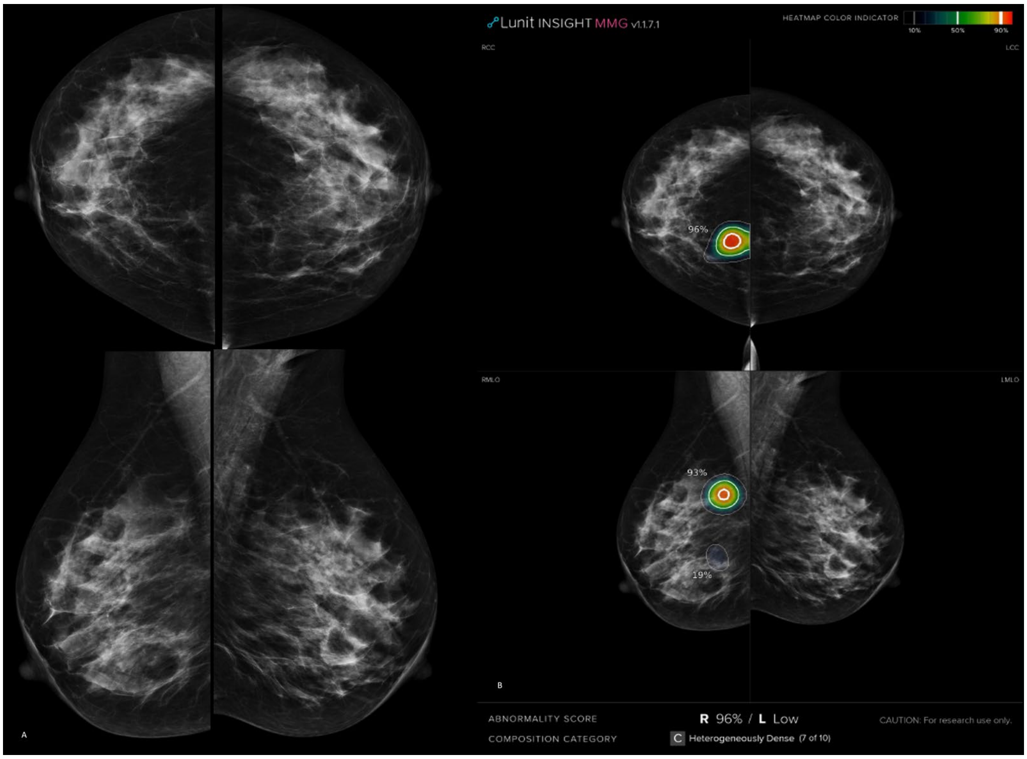

Fig. 4 Right breast deeply seated early invasive ductal carcinoma in a 49-year-old female **A**. Four-view digital mammogram displayed upper inner tiny area of distortion partly obscured by glandular tissue and was missed by the radiologist. **B** The carcinoma was easily targeted by the algorithm in (**B**), assigned a very high "probably cancer" scoring percentage of 96%

Table 2 Flow for screening mammogram included in the study ($n = 20,764$) and the relative proportion for each reader strategy whether two readers or one reader and AI

	Double reading		
	Two radiologist	**AI and one radiologist**	**Relative proportion (95% Confidence Interval)**
Abnormal mammogram; no cancer	1416 (6.82%)	1371 (6.60%)	0.96 (0.94–0.97)
Abnormal mammogram; suspicious	1571 (7.56%)	1616 (7.78%)	1.02 (0.99–1.05)
Recall benign, no biopsy	1541 (7.42%)	1491 (7.18%)	0.96 (0.94–0.97)
Recall suspicious	1446 (6.96%)	1496 (7.20%)	1.03 (0.99–1.12)
Recall suspicious, no cancer	163 (0.78%)	208 (1.00%)	1.27 (1.25–1.29)
Biopsy (total)	1446 (6.96%)	1491 (7.18%)	1.03 (0.98–1.07)
Recall suspicious proven benign	115 (0.55%)	84 (0.40%)	0.73 (0.70–0.77)
Recall suspicious proven cancer	1293 (6.22%)	1324 (6.37%)	1.02 (0.99–1.07)

and a specificity of 99.757% (99.677–99.822%) when the interpretation of the mammogram was performed by AI plus only one radiologist. Positive LHR was 387.260, negative LHR was 0.060, and AUC was 0.969 (0.967–0.971).

Among the strengths of our study: (1) AI was integrated as an independent (junior) reader in the daily workflow of screening mammograms and sorting negative cases from those that required recall. (2) The study was performed by highly expert radiologists in the field of breast imaging, especially mammograms (with more

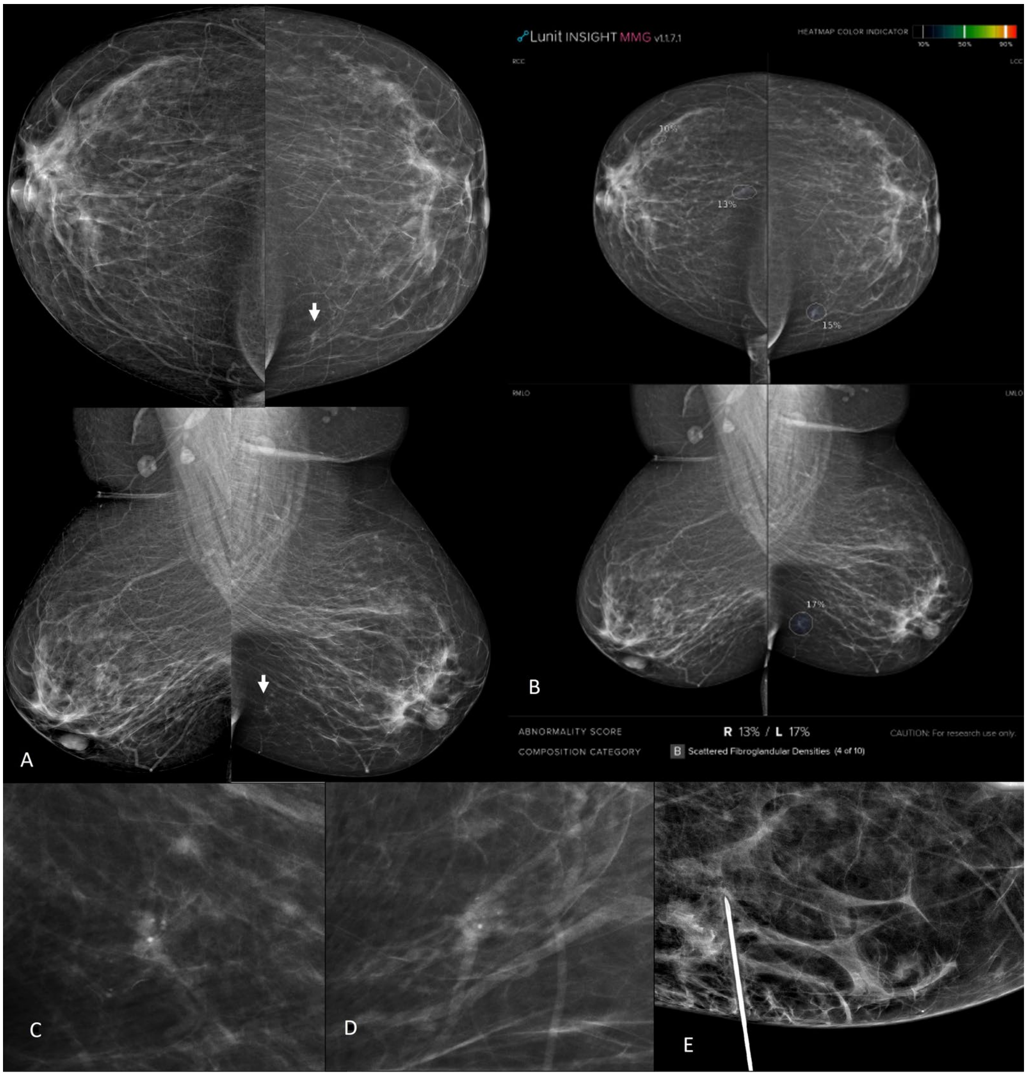

Fig. 5 Screening mammogram of a 52-year-old female. The left breast lower inner suspicious lesion (arrow) proved to be proliferative fibrocystic disease and usual ductal hyperplasia (UDH). **A** A four-image mammogram showed a left lower inner tiny indistinct lesion (arrow) at the pre-mammary fat with calcific foci. **B** AI-scanned four images; a mammogram marked the left suspicious lesion and gave it a score of 17% (probably non-cancer). **C** and **D** Magnification view of the left breast lesion of concern, cranio-caudal (**C**) and medio-lateral oblique (**D**). E: A stereotactic biopsy of the left breast lesion proved to be benign pathology with no evidence of malignancy. Note that the AI marked two areas at the right breast with a low scoring percentage yet looked benign by the radiologist and was subjected to only ultrasound at the time of recall

than 20 years of experience). (3) The first prospective study included biopsy results for suspicious and malignant lesions and 2 years of follow-up for normal and benign (non-proven) diagnosed cases. The availability of tissue sampling and a 2-year follow-up allowed direct calculation of sensitivity and negative predictive value. Thus, there is high certainty that no cancer was present for women who did not have a sample biopsy. (4) The breast imaging center where the study was performed is a center of excellence where the ultrasound and biopsy requirements are available at the same unit of screening mammogram. Such an opportunity supported the completion of the breast cancer screening experience of interpretation, recall, and biopsy (if required) at the same session.

It was previously suggested that the consensus reading of screening mammograms assigned on initial reads, negative by the human reader and positive by AI, be nudged toward a negative decision, provided that a third human reader has already reviewed the images without finding anything suspicious [8]. However, this was found to underestimate the ability of AI in terms of detecting cancer and increasing the rate of missed breast carcinoma.

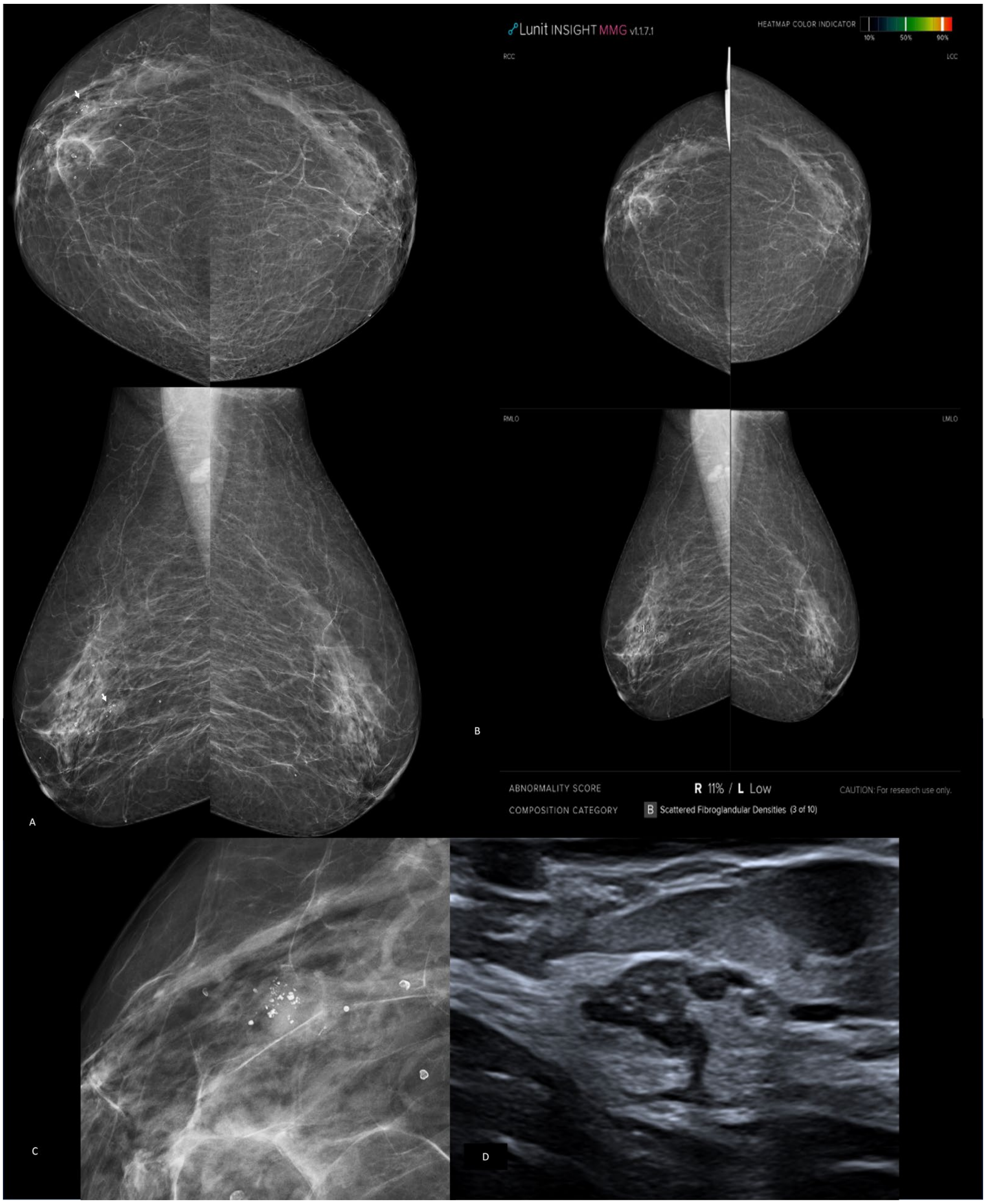
Lunit INSIGHT MMG v1.1.7.1
HEATMAP COLOR INDICATOR
10%
50%
90%
RCC
LCC
RMLO
LMLO
A
B
ABNORMALITY SCORE
R 11% / L Low
CAUTION: For research use only.
COMPOSITION CATEGORY
B Scattered Fibroglandular Densities (3 of 10)
C
D

Fig. 6 Right breast upper outer invasive ductal carcinoma grade II-associated DCIS component 5% solid type in a 55-year-old female (ACR b). **A** A screening mammogram displayed a suspicious cluster of coarse heterogeneous microcalcifications (arrow). **B** The low AI scoring percentage of 11% for the lesion of concern (i.e., probably non-cancer). **C** Magnification view of the microcalcific cluster. **D** Ultrasound of the calcific cluster showed an indistinct soft tissue non-mass with calcifications, which confirmed the suggestion of suspicion for the radiologist and warranted a biopsy

Based on the current work and the number of mammograms flagged by AI for recall and/or biopsy, plus the consequent results, it is suggested that initial consensus readings of negative by the human reader and positive by AI are to be considered for recall even if a third reader interprets the mammogram as negative.

Marinovich et al. [29] reported encouraging findings about AI cancer detection with a small but statistically significant reduction in false positive recall although there was a higher abnormal interpretation proportion for AI than radiologist,

Table 3 Chi-squared test and significance levels of reading strategies

	AI plus one reader		
Follow-up/pathology	**Negative**	**Positive**	
Negative	19,309	47	19,356 (93.2%)
Positive	84	1324	1408 (6.8%)
	19,393 (93.4%)	1371 (6.6%)	20,764
Chi-squared	**18,721.937**		
DF	**1**		
Significance level	**P < 0.0001**		
Contingency coefficient	**0.689**		
	Two readers		
Follow-up/pathology	**Negative**	**Positive**	
Negative	19,233	123	19,356 (93.2%)
Positive	115	1293	1408 (6.8%)
	19,348 (93.2%)	1416 (6.8%)	20,764
Chi-squared	**17,177.874**		
DF	**1**		
Significance level	**P < 0.0001**		
Contingency coefficient	**0.673**		

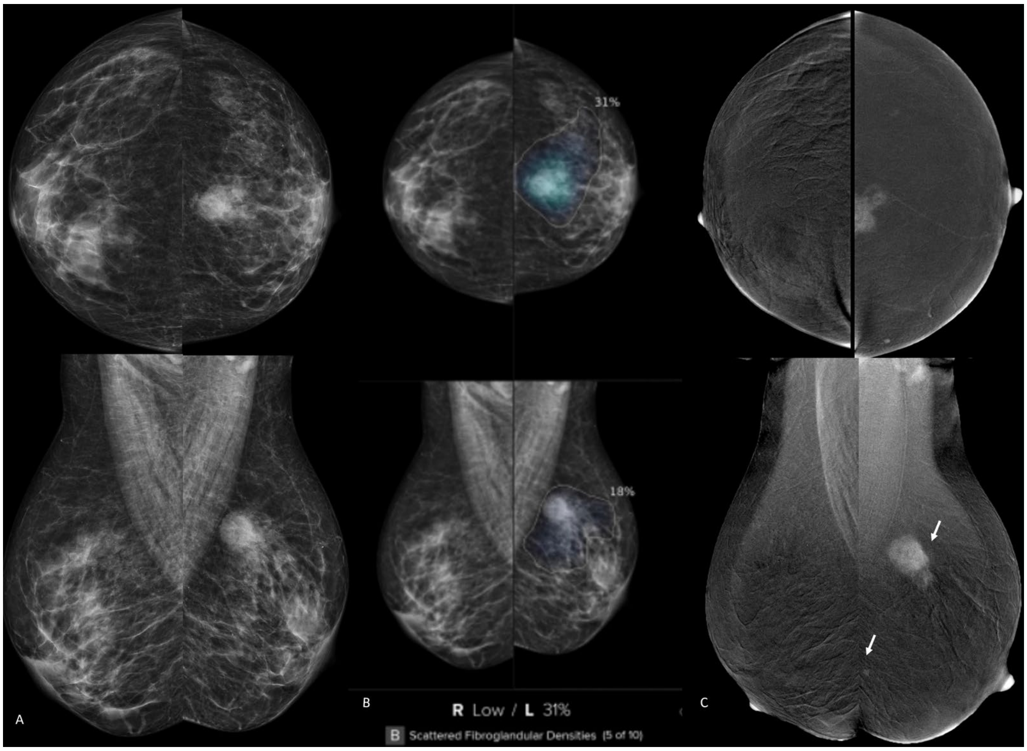

Fig. 7 Left breast invasive ductal carcinoma in a 54-year-old female with dense breast (ACR c). **A** Bilateral breast diseases; right lower inner asymmetry and left upper outer mass were seen on the digital mammogram. **B** AI marked the left breast mass with a score of less than 50%, although it fulfilled the criteria of malignancy on the mammogram for the radiologist. **C** Contrast-enhanced mammogram: the left breast showed malignant-looking upper outer mass, non-mass enhancement, and lower inner focus (multicentric distribution arrows). N.B., the right breast asymmetry was correlated on ultrasound with a simple cyst.

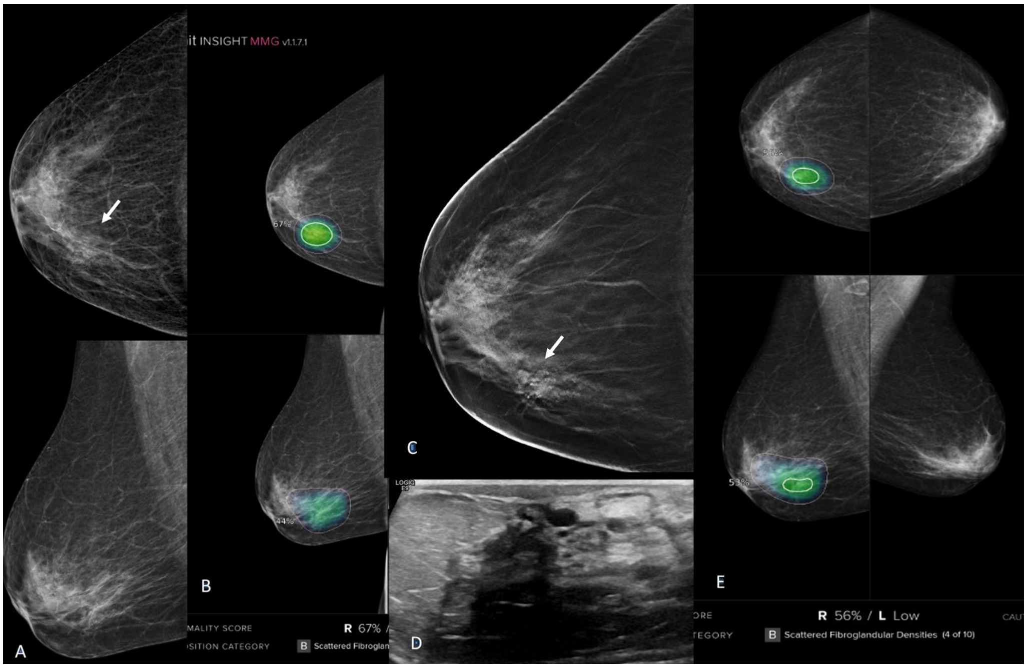

Fig. 8 A postmenopausal 63-year-old female, whose first screening mammogram was in 2017, showed right breast lower inner asymmetry with grouped macrocalcifications, and on recall ultrasound, she displayed clustered microcysts. **A** A digital mammogram in 2021 showed the previously reported right lower inner asymmetry (arrow). **B** A high AI score of 67%, which implied probably cancer and required recalling the patient for an ultrasound and biopsy. On recall, **C** digital breast tomosynthesis examination showed benign features of lobulated contour and macrocalcifications (arrow), and **D** the ultrasound images presented a focal area of adenosis and grouped cysts. There were no suspicious features that warranted a biopsy on (**C**) and (**D**). The pathology report reported usual ductal hyperplasia and fat necrosis which matched with the radiologist's suggestion of benignity. The AI high percentage in 2021 subjected the patient to unnecessary trauma. **E** Follow-up mammogram scanned with AI in 2022 showed a lower scoring percentage (56% versus 67%) of the lesion of interest, which supported the recommendation of ultrasound instead of a re-biopsy

Lesions given an AI abnormality score of more than 14% and less than 68% are candidates for recall and require ultrasound characterization, while lesions given an AI score of more than 68% are to be scheduled for biopsy. AI scoring percentage could be used as a tool for follow-up diagnostic mammograms, as shown in Fig. 8.

It is worth mentioning that study was conducted in Baheya Charity Hospital, which has been actively collaborating with Egypt's national screening program utilizing state-of-the-art technology for breast cancer early detection and comprehensive treatment.

The incidence rate of breast cancer expressed by the study cohort was $(1260/16801) \times 100{,}000 \approx 7500$ per 100,000 females, provided that the number of proved cancer cases was 1260 divided by the population at risk ($n = 16{,}801$), which is often expressed per 100,000 people.

Elevated incidence rates of breast cancer may reflect increased prevalence of risk factors, opportunistic or organized mammography screening detections, aging, and growth of population. However, the difference in major risk factors, screening strategies, and population size or structures of different regions led to the disparities in the burden of breast cancer [30].

Future large-scale work is needed to investigate the implementation of AI and cope with the dissemination of new evidence from prospective AI trials for breast screening and population-based screening programs in the context of sites with varying needs, capacities, and screening population characteristics to confirm the extent of achievable improvement in early cancer detection.

Conclusions

Artificial intelligence could be used as an initial reader for the evaluation of screening mammography in routine workflow. Implementation of AI enhanced the opportunity to reduce false negative cases and supported the decision to recall or biopsy.

Abbreviations

AI	Artificial intelligence
AI-based CADe/x	Artificial intelligence-based computer-aided detection/diagnosis
AI-MMG	Artificial intelligence for reading mammograms

ACR	American College of Radiology
AUC	Area under the curve
BI-RADS	Breast Imaging–Reporting and Data System
CI	Confidence interval
LHR	Likelihood ratio
MASAI	Mammography Screening with Artificial Intelligence
NPV	Negative predictive value
PACS	Picture archiving and communication system
PPV	Positive predictive value
ROC	Receiver operating characteristic
SPSS	Statistical Package for the Social Sciences

Acknowledgements
Not applicable.

Author contributions
MS is the guarantor of integrity of the entire study. KR and MS contributed to the study concepts and design. MS, KR, HS, FA, and GM contributed to the literature research. MS, SE, NS, MG, and FA contributed to the clinical studies. MS, GM, SE, and HS contributed to the experimental studies/data analysis. MS and HS contributed to the statistical analysis. MS, KR, AE, NY, and FA contributed to the manuscript preparation. MS, KR, AE, NY, and FA contributed to the manuscript editing. All authors have read and approved the final manuscript.

Funding
No source of funding.

Availability of data and materials
The corresponding author is responsible for sending the used data and materials upon request.

Declarations

Ethics approval and consent to participate
The study was approved by the ethical committee of the Radiology Department of Baheya center of early breast cancer and treatment which is nonprofitable and non-governmental highly specialized multidisciplinary center. Informed consent was obtained.

Consent for publication
All patients included in this research were legible, above 16 years of age. Informed consent from the included patients was obtained.

Competing interests
The authors declare that they have no competing interests.

Received: 5 June 2024 Accepted: 28 August 2024
Published online: 12 September 2024

References

1. Rostom Y, Abdelmoneim S-E, Shaker M et al (2022) Presentation and management of female breast cancer in Egypt. East Mediterr Health J 28(10):725–732
2. European Commission Initiative on Breast Cancer. Screening for women aged 50–69. https://healthcare-quality.jrc.ec.europa.eu/european-breastcancer-guidelines/screening-ages-and-frequencies/women-50-69. Accessed September 2023
3. Hofvind S, Bennett RL, Brisson J et al (2016) Audit feedback on reading performance of screening mammograms: an international comparison. J Med Screen 23(3):150–159
4. Hofvind S, Tsuruda KM, Mangerud G, et al. (2017) The Norwegian Breast Cancer Screening Program, 1996–2016: Celebrating 20 years of organised mammographic screening. Cancer in Norway 2016: cancer incidence, mortality, survival, and prevalence in Norway. Oslo, Norway: Cancer Registry of Norway
5. Wahdan IH (2020) Cost-effectiveness of national breast cancer screening programs in developing countries, with reference to the recent Egyptian Initiative CC BY-SA 4.0. J High Inst Public Health 50(1):1–9
6. Skrundevskiy AN, Omar OS, Kim J et al (2018) Return on investment analysis of breast cancer screening and downstaging in Egypt: implications for developing countries. Value Health Reg Issues 16:22–27
7. Kwee TC, Kwee RM (2021) Workload of diagnostic radiologists in the foreseeable future based on recent scientific advances: growth expectations and role of artificial intelligence. Insights Imaging 12:88
8. Dembrower K, Crippa A, Colón E, ScreenTrustCAD Trial Consortium et al (2023) Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. Lancet Digit Health 5(10):703–711
9. Kim H-E, Kim HH, Han B-K et al (2020) Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Health 2:e138–e148
10. Mansour S, Kamal R, Hashem L et al (2021) Can artificial intelligence replace ultrasound as a complementary tool to mammogram for the diagnosis of the breast cancer? Br J Radiol 94(1128):20210820
11. McKinney SM, Sieniek M, Godbole V et al (2020) International evaluation of an AI system for breast cancer screening. Nature 577:89–94
12. Dembrower K, Wåhlin E, Liu Y et al (2020) Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. Lancet Digit Health 2:e468–e474
13. Mansour S, Soliman S, Kansakar A et al (2022) Strengths and challenges of the artificial intelligence in the assessment of dense breasts. BJR Open 4(1):20220018
14. Leibig C, Brehmer M, Bunk S et al (2022) Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. Lancet Digit Health 4:e507–e519
15. Salim M, Wåhlin E, Dembrower K et al (2020) External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA Oncol 6:1581–1588
16. Lång K, Josefsson V, Larsson A-M et al (2023) 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 24:936–944
17. D'Orsi CJ, Sickles EA, Mendelson EB, et al. (2013) ACR BI-RADS Atlas, fifth edition, breast imaging reporting and data system. American College of Radiology, Reston
18. He K, Zhang X, Ren S et al (2016) Deep residual learning for image recognition. Proc CVPR IEEE 1:770–778
19. Morgan MB, Mates JL (2021) Applications of artificial intelligence in breast imag ing. Radiol Clin North Am 59:139–148
20. Sechopoulosa I, Teuwena J, Manna R (2021) Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: state of the art. Semin Cancer Biol 72:214–225
21. Sharma N, Ng AY, James JJ et al (2023) Multi-vendor evaluation of artificial intelligence as an independent reader for double reading in breast cancer screening on 275,900 mammograms. BMC Cancer 23(1):460
22. Ng AY, Oberije CJG, Ambrózay É et al (2023) Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer. Nat Med 29:3044–3049
23. Lee SE, Han K, Yoon JH et al (2022) Depiction of breast cancers on digital mammograms by artificial intelligence-based computer-assisted diagnosis according to cancer characteristics. Eur Radiol 32:7400–7408
24. Park GE, Kang BJ, Kim SH et al (2022) Retrospective review of missed cancer detection and its mammography findings with artificial-intelligence-based. Comput Aided Diagn Diagn 12(2):387
25. Chang YW, An JK, Choi N et al (2022) Artificial intelligence for breast cancer screening in mammography (AI-STREAM): a prospective multicenter study design in Korea using AI-based CADe/x. J Breast Cancer 25(1):57–68
26. Lotter W, Diab AR, Haslam B et al (2021) Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. Nat Med 27:244–249
27. Rodriguez-Ruiz A, Lång K, Gubern-Merida A et al (2019) Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J Natl Cancer Inst 2019(111):916–922

28. Schaffter T, Buist DSM, Lee CI et al (2020) Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Netw Open 3(3):e200265
29. Marinovich ML, Wylie E, Lotter W et al (2023) Artificial intelligence (AI) for breast cancer screening: breastscreen population-based cohort study of cancer detection. EBioMedicine 90:104498
30. Lei S, Zheng R, Zhang S et al (2021) Global patterns of breast cancer incidence and mortality: a population-based cancer registry data analysis from 2000 to 2020. Cancer Commun 41(11):1183–1194

Publisher's Note