European Journal of Radiology Open 11 (2023) 100509

Contents lists available at ScienceDirect

European Journal of Radiology Open

journal homepage: www.elsevier.com/locate/ejro

Artificial intelligence-based computer-assisted detection/diagnosis (AI-CAD) for screening mammography: Outcomes of AI-CAD in the mammographic interpretation workflow

Jung Hyun Yoon [a], Kyungwha Han [b], Hee Jung Suh [c], Ji Hyun Youk [d], Si Eun Lee [e], Eun-Kyung Kim [e,*]

[a] Department of Radiology, Severance Hospital, Research Institute of Radiological Science, Center for Clinical Imaging Data Science, Yonsei University, College of Medicine, South Korea
[b] Department of Radiology, Center for Clinical Imaging Data Science, Yonsei University, College of Medicine, South Korea
[c] Department of Radiology, Severance Check-up Center, South Korea
[d] Department of Radiology, Gangnam Severance Hospital, Yonsei University, College of Medicine, South Korea
[e] Department of Radiology, Yongin Severance Hospital, Yonsei University, College of Medicine, South Korea

HIGHLIGHTS

- AI-CAD detected 5 (17.9%, 5 of 28) of the 9 cancers overlooked by radiologists.
- 89.0% (577 of 648) of AI-CAD marks were seen on negative examinations, and 267 (41.2%) were considered to be negligible.
- Stand-alone AI-CAD has higher recall rates with comparable sensitivity and CDR compared to the radiologists' interpretation.

ARTICLE INFO

Keywords:
Breast cancer screening
Mammography
Artificial intelligence
Computer-assisted detection
Computer-assisted diagnosis

ABSTRACT

Purpose: To evaluate the stand-alone diagnostic performances of AI-CAD and outcomes of AI-CAD detected abnormalities when applied to the mammographic interpretation workflow.

Methods: From January 2016 to December 2017, 6499 screening mammograms of 5228 women were collected from a single screening facility. Historic reads of three radiologists were used as radiologist interpretation. A commercially-available AI-CAD was used for analysis. One radiologist not involved in interpretation had retrospectively reviewed the abnormality features and assessed the significance (negligible vs. need recall) of the AI-CAD marks. Ground truth in terms of cancer, benign or absence of abnormality was confirmed according to histopathologic diagnosis or negative results on the next-round screen.

Results: Of the 6499 mammograms, 6282 (96.7%) were in the negative, 189 (2.9%) were in the benign, and 28 (0.4%) were in the cancer group. AI-CAD detected 5 (17.9%, 5 of 28) of the 9 cancers that were intially interpreted as negative. Of the 648 AI-CAD recalls, 89.0% (577 of 648) were marks seen on examinations in the negative group, and 267 (41.2%) of the AI-CAD marks were considered to be negligible. Stand-alone AI-CAD has significantly higher recall rates (10.0% vs. 3.4%, $P < 0.001$) with comparable sensitivity and cancer detection rates ($P = 0.086$ and 0.102, respectively) when compared to the radiologists' interpretation.

Conclusion: AI-CAD detected 17.9% additional cancers on screening mammography that were initially overlooked by the radiologists. In spite of the additional cancer detection, AI-CAD had significantly higher recall rates in the clinical workflow, in which 89.0% of AI-CAD marks are on negative mammograms.

* Correspondence to: Department of Radiology, Yongin Severance Hospital, Yonsei University College of Medicine 363, Dongbaekjukjeon-daero, Giheung-gu, Yongin-si, Gyeonggi-do, Republic of Korea. Fax: 82–2-2227–8337
E-mail address: ekkim@yuhs.ac (E.-K. Kim).

https://doi.org/10.1016/j.ejro.2023.100509
Received 7 May 2023; Received in revised form 3 July 2023; Accepted 9 July 2023

1. Introduction

Mammography is currently the most used imaging modality for breast cancer screening with past randomized screening trials showing its contribution to the reduction of breast cancer-related mortality [1,2]. Although commonly used for screening or diagnostic purposes, mammography interpretation is quite difficult and variations in mammography outcomes and accuracy are well acknowledged [3]; the sensitivity for cancer detection in women with mammographically-dense breast has been reported to decrease to 30–48% [4,5]. Effective screening mammography requires imaging from dedicated mammography facilities, and well-trained radiologists with expertise in breast imaging to achieve adequate levels of interpretive accuracy [6–8]. Interpretive accuracy is important because the accurate detection of breast cancer when present (high sensitivity) enables early detection and treatment, while at the same time maintenance of low levels of false-positive recalls (high specificity) is required to prevent unnecessary additional work-up [9].

Considerable effort has been put into finding ways to improve diagnostic accuracy using mammography, with artificial intelligence (AI)-based computer-assisted detection/diagnosis (CAD) currently taking center stage in breast cancer screening. In recent studies, AI-CAD algorithms have shown stand-alone performances comparable to those of radiologists [10–12] or have improved the performances of radiologists by providing interpretative assistance [13–17]. In spite of the promising results, most published studies are results of small, cancer-enriched study samples that is quite different from the cancer prevalence of the real-world screening population. The applicability of AI-CAD to real-world interpretation depends on the abnormality features marked by AI-CAD and their clinical significance, and acknowledging the outcomes of features that are marked as positive by AI-CAD may enhance the utility of the algorithm. However, there are no data that specifies the outcomes of mammographic abnormality features highlighted by AI-CAD. Also, information from prior examinations or supplemental screening modalities are commonly incorporated in mammographic interpretation, and little is known on the outcomes of AI-CAD results when applied to our interpretation workflow.

In this aspect, we simulated a retrospective, cross-sectional study to evaluate the outcomes of AI-CAD detected abnormalities and stand-alone diagnostic performances of AI-CAD when applied to the interpretation workflow of a consecutive, screening population.

2. Materials and methods

This retrospective study was approved by our institutional review board (IRB) with a waiver of informed consent to review medical records and radiologic images.

2.1. Study sample

From January 2016 to December 2017, 9457 routine mammography examinations of 7988 consecutive women who visited a single cancer screening facility were obtained. Among them, mammography images from women who were not followed (n = 2834), mammography images with the incomplete assessment category that were not further investigated (n = 51), mammography images from women who were treated for breast cancer (n = 30) or those who underwent mammoplasty (n = 14), and mammography images not obtained in the routine 4-view positions (n = 29) were excluded. Finally, 6499 four-view, full-field digital mammograms of 5228 women who had 1) undergone biopsy for pathologic diagnosis prior to the next screening round, or 2) had undergone at least one additional round of screening at our facility were included in this study (Fig. 1). In Korea, supplemental screening US is commonly available and included in the screening protocol with mammography, according to the preference of the women or clinicians' recommendation. Mammography examinations were available to the radiologist who performed supplemental screening US. Serial screening mammograms per woman, if performed during the study period were included prior to cancer diagnosis and individually analyzed. Medical records and images of the women were last reviewed on Mar. 22, 2022.

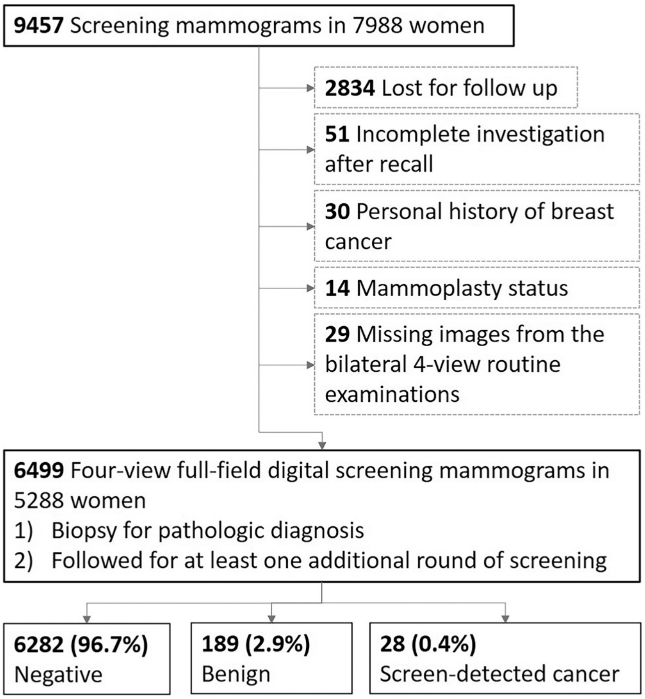
9457 Screening mammograms in 7988 women
2834 Lost for follow up
51 Incomplete investigation after recall
30 Personal history of breast cancer
14 Mammoplasty status
29 Missing images from the bilateral 4-view routine examinations
6499 Four-view full-field digital screening mammograms in 5288 women
1) Biopsy for pathologic diagnosis
2) Followed for at least one additional round of screening
6282 (96.7%) Negative
189 (2.9%) Benign
28 (0.4%) Screen-detected cancer

Fig. 1. Flowchart showing the inclusion of the study sample.

2.2. Mammography examination and interpretation

Digital mammograms were obtained with both craniocaudal and mediolateral oblique views using one of two dedicated mammography devices (Lorad Selenia, Hologic Inc., Danbury, CT, USA). Mammography images were interpreted by three radiologists, one with fellowship training in breast imaging (12 years of experience) and two general radiologists, using the American College of Radiology Breast Imaging Reporting And Data System (ACR BI-RADS) parenchymal density and final assessment categories [18]. The mammograms were interpreted by one radiologist, as double-reading is not routine for mammography interpretation in Korea. Prior examinations, if available, was used for comparison during mammography interpretation. Interpretation results recorded in the radiologic reports were used for data analysis.

2.3. AI-CAD for mammography

For this study, we used an AI-based diagnostic support software dedicated to breast cancer detection on digital mammography (Lunit INSIGHT for Mammography, version 1.1.0.1, Lunit Inc., Seoul, Korea). This AI-CAD was developed with deep convolutional neural networks (CNNs), trained and validated with over 170,000 mammography examinations and tested with an external mammography dataset not used for training or validation [16]. The AI-CAD provided per-breast malignancy scores (range, 0–100%, 100% meaning high likelihood of cancer present) with four-view region-of-interests (ROIs) for suspicious malignant lesions on each input mammogram. BI-RADS final assessment categories are not provided by AI-CAD. Since AI-CAD was not used during initial mammography interpretations, one dedicated breast radiologist (J.H.Y., 13 years of experience) who was not involved in the initial mammographic interpretation retrospectively reviewed the mammograms recalled by AI-CAD to evaluate the abnormal findings detected by the interpreting radiologists and AI-CAD. AI-CAD marks were considered to correlate to the radiologist-detected findings according to the described features on the radiologic reports or abnormalities seen on US examinations when they were located in the same quadrant and showed similar size/imaging features. This radiologist had full access to all medical records during the retrospective review, and prior mammograms.

2.4. *Data and statistical analysis*

Ground truth in terms of cancer or benign diagnosis was confirmed with histopathologic diagnosis via biopsy/surgery. Cancer examinations were divided according to the period of diagnosis to the screening mammograms; 1) screen-detected cancers, defined as cancers detected on screening mammograms, 2) interval cancers, defined as cancers diagnosed in the interval between two consecutive screening rounds after a negative screening examination, 3) next round-detected cancers, defined as cancers detected at or past two screening rounds after negative screens. Since this study aimed to evaluate the detection of cancers at the period of screening examinations, screen-detected and interval cancers were included in the cancer group. A negative examination was defined as mammograms free of screen-detected or interval cancers.

As the ACR BI-RADS lexicon were used for mammography interpretation, BI-RADS 1 and 2 assessments were considered to be 'negative' interpretations while BI-RADS 0, 3, 4, 5 assessments were considered to be 'positive' interpretations. An abnormality score of 10% was used as the cutoff threshold for AI-CAD [16], i.e., AI-CAD marks with abnormality scores ≥ 10% were considered as 'positive' AI-CAD findings, while those with abnormality scores < 10% as 'negative' AI-CAD findings. Areas of the AI-CAD marks were reviewed and categorized into abnormality features defined by the ACR BI-RADS as follows: asymmetry, mass, calcifications, and distortion [18], and 'multiple features' (more than two features present). Cases for which the radiologist could not find a definable abnormality at the AI-CAD mark were categorized as 'not definable' (Supplementary Fig. 1). The significance of the AI-CAD marks was assessed and categorized as 'negligible', defined as AI-CAD marks that the radiologist did not consider significant for recall, and 'need recall', defined as AI-CAD marks that the radiologist considered to warrant further investigation including those found in comparison with prior studies or with additional imaging studies.

Diagnostic performances of the workflow for mammographic interpretation and stand-alone AI-CAD was calculated according to the following metrics: recall rates, cancer detection rates (CDR), sensitivity and specificity. Logistic regression with generalized estimating equation methods was used for comparison of performance metrics between radiologists and AI-CAD. Statistical analyses was performed using R software (version 4.1.3.; R Foundation for Statistical Computing, Vienna, Austria). *P* values of less than 0.05 were considered to have statistical significance.

3. Results

Table 1
Patient characteristics and imaging features of the screening mammograms.

	Negative	Benign	Cancer	Total
			Screen-detected & interval cancers	
No. of examinations	6282	189	28	6499
Age				
<50 years	3298 (52.5)	136 (72.0)	12 (42.9)	3446 (53.0)
≥ 50 years	2984 (47.5)	53 (28.0)	16 (57.1)	3053 (47.0)
*Mammographic density**				
Fatty breast	934 (14.9)	8 (4.2)	3 (10.7)	945 (14.5)
Dense breast	5348 (85.1)	181 (95.8)	25 (89.3)	5554 (85.5)
Radiologists' interpretation				
BI-RADS 1, 2	6109 (97.2)	161 (85.2)	9 (32.1)	6279 (96.6)
BI-RADS 0, 3–5	173 (2.8)	28 (14.8)	19 (67.9)	220 (3.4)
AI-CAD Abnormality score				
< 10%	5700 (90.7)	146 (77.2)	5 (17.9)	5851 (90.0)
≥ 10%	582 (9.3)	43 (22.8)	23 (82.1)	648 (10.0)
US Examinations				
No US	3048 (48.5)	50 (26.5)	3 (10.7)	3101 (47.7)
Supplemental US	3234 (51.5)	139 (73.5)	25 (89.3)	3398 (52.3)

BI-RADS: Breast Imaging Reporting And Data System, AI-CAD: artificial intelligence-based computer assisted detection/diagnosis, US: ultrasonography
* : fatty breast defined as parenchymal density grades A and B, dense breast as grades C and D

Patient characteristics and imaging features of the 6499 mammograms are summarized according to final diagnosis in Table 1. Of the 6499 mammograms, 6282 (96.7%) were in the negative, 189 (2.9%) were in the benign, and 28 (0.4%) were in the cancer group. Of the 28 cancers (23 invasive cancers, 5 ductal carcinoma *in sit*u (DCIS)), 25 were asymptomatic screen-detected cancers while 3 were interval cancers detected with newly developed symptoms at 7.6, 9.1, and 9.4 months after a negative screen, respectively (Table 2). At the point of reviewing medical records, 14 cancers (12 invasive cancers, 2 DCIS) were detected at or after the next-round screening (mean interval: 21.4 ± 5.0 months, range: 12.3–27.3 months). One patient who was diagnosed as invasive cancer 40.1 months after the initial screen had two consecutive screening mammograms included in the study period. Both screening mammograms were interpreted as negative, and cancer was diagnosed at 24.1 months after the second negative screen. Among the 6471 negative/benign examinations, 1271 (19.6%) were consecutive, next-round screening mammograms. Mean follow-up interval of the 6471 negative/benign mammograms was 40.2 ± 15.3 months (range, 9.3–74.3 months). Among the total study sample, 3398 (52.3%) had supplemental US performed. After US examination, biopsy was recommended in 247 lesions in 243 (4.6%, 243 of 5228) women; 237 underwent US-guided core needle biopsy, 7 US-guided vacuum assisted biopsy and 3 stereotactic biopsy.

For mammographic density, 5554 (85.5%) were assessed as dense breasts and 945 (14.5%) as fatty breasts. Of the study sample, 220 (3.4%) were recalled by the radiologists according to abnormalities detected on the screening mammograms. Six hundred forty-eight (10.0%) mammograms had positive AI-CAD results.

3.1. *Characteristics of the screen-detected/interval cancers and next-round detected cancers*

Among the 28 screen-detected or interval cancers, 19 (67.9%, 8 as BI-RADS 0, 1 as BI-RADS 4b, 1 as BI-RADS 4c, and 9 as BI-RADS 5) had abnormal findings detected by the interpreting radiologists. AI-CAD showed positive results in 18 of the 19 cancer examinations with radiologist-recall (Fig. 2). The remaining 9 cancers were initially assessed as BI-RADS 1 or 2 on screening mammograms, but either had suspicious findings detected on supplemental US (n = 6) or were detected after developing symptoms (n = 3, Table 2). AI-CAD correctly localized 5 (17.9%, 5 of 28) of the 9 cancer examinations that were initially interpretated as negative/benign, that presented as mass with calcifications (n = 2, Fig. 3), asymmetry (n = 2), and distortion (n = 1) at the AI-CAD marks on retrospective review.

Fourteen cancers were diagnosed at or after the next-round screening examinations, i.e., next-round detected cancers (Table 3). Five of the next-round detected cancers were palpable at the time of diagnosis. Retrospective image review of the initial screening mammograms

Table 2
Clinicopathologic features and AI-CAD scores of the 9 screen-detected/interval cancers initially interpreted as negative on screening mammograms.

	Age (years)	Density	Initial radiologist interpretation	Cancer detection modality	Cancer type	AI-CAD abnormality score (%)	Imaging features of cancers on retrospective review	Cancer diagnosis interval (months)*	Pathologic diagnosis	Cancer size (mm)	Axilla LN metastasis
1	74	D	BI-RADS 2	US	Screen-detected	0.71	-	0.47	IDC	9	No
2	49	C	BI-RADS 2	US	Screen-detected	96.02	Mass with calcifications	0.97	IDC	12	No
3	50	D	BI-RADS 2	US	Interval	0.1	-	7.63	IDC	9	No
4	45	C	BI-RADS 2	US	Screen-detected	0.8	-	0.63	IDC	11	No
5	51	D	BI-RADS 2	MG, US	Interval	72.81	Asymmetry	9.12	IDC	20	No
6	48	D	BI-RADS 2	US	Screen-detected	45.69	Asymmetry	0.43	IDC	12	No
7	51	D	BI-RADS 1	MG, US	Interval	92.75	Distortion	9.4	IDC	20	No
8	58	C	BI-RADS 2	US	Screen-detected	0.22	-	0.2	IDC	9	No
9	52	D	BI-RADS 2	US	Screen-detected	91.83	Mass with calcifications	1.53	DCIS	18	No

BI-RADS: Breast Imaging Reporting And Data System, US: ultrasonography, AI-CAD: artificial intelligence-based computer assisted detection/diagnosis, LN: lymph node, IDC: invasive ductal carcinoma, DCIS: ductal carcinoma in situ

* : interval from screening mammography to cancer diagnosis on pathology

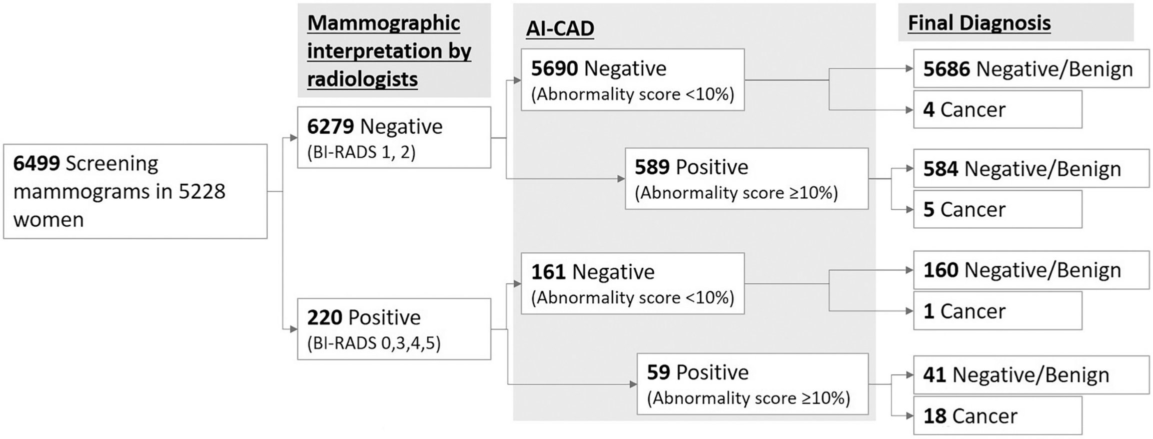

Fig. 2. Diagram of the study population according to radiologists' intial mammographic interpretation, stand-alone AI-CAD analysis and final diagnosis BI-RADS: Breast Imaging Reporting And Data System, AI-CAD: artificial intelligence-based computer-assisted detection/diagnosis.

revealed 4 cases showing subtle abnormalities (1 asymmetry, 1 distortion, 2 calcifications) correlating to the next-round detected cancer site. AI-CAD correctly detected 5 of the next-round detected on the initial screening mammogram that the radiologist had interpreted as negative (Fig. 4).

3.2. *Imaging features and final outcomes of the abnormalities recalled by radiologists vs. AI-CAD*

Imaging features and outcomes of the abnormalities recalled by radiologists and AI-CAD are summarized in Table 4. Radiologists recalled 220 (3.4%) examinations to have abnormal findings needing further investigation in which mass/asymmetry was most common (75.0%). AI-CAD recalled 648 (10.0%) examinations with abnormality scores$\geq$ 10%. When comparing between radiologist-recalled vs. AI-CAD recalled examinations, higher rates of mass/asymmetry (75.0% vs. 51.7%) and calcifications (21.4% vs. 15.1%) were seen in radiologist-recalled examinations, while higher rates of distortion (0.4% vs. 17.7%) and multiple features (3.2% vs. 6.4%) were seen in AI-CAD-recalled examinations ($P < 0.001$). When comparing the cancer rates of the abnormal features, cancer rates of multiple features (85.7% vs. 24.4%, $P = 0.001$) and calcifications (12.8% vs. 3.1%, $P = 0.023$) was significantly higher in the radiologist compared to AI-CAD, respectively. Cancer rates between radiologists vs. AI-CAD for mass/asymmetry (4.2% vs. 2.7%) did not show significant differences ($P = 0.353$).

Of the 648 AI-CAD recalls, 577 (89.0%) were in the negative examination group. Fifty-nine (9.1%) of the AI-CAD marked examinations had no corresponding suspicious abnormality correlating to the AI-CAD marks, 'not definable' marks, in which 1 (1.7%) examination had a biopsy-proven benign (fibrocystic change) mass detected on US that corresponds to the AI-CAD marked region on mammography. The remaining 58 (98.3%) were confirmed as negative on the next screening rounds. Also, 519 (80.1%, 303 asymmetries, 109 distortions, 80 calcifications, 27 mixed features) of the 648 AI-CAD marked lesions were in the negative examination group (Supplementary Fig 2).

When assessing the significance of the AI-CAD marks, 267 (41.2%) abnormalities were considered to be negligable and 381 (58.8%) were at abnormalities that were considered to necessitate recall. Compared with past imaging studies, 240 (63.0%) of the 381 abnormalities marked by AI-CAD were considered benign, showing stability on serial imaging studies.

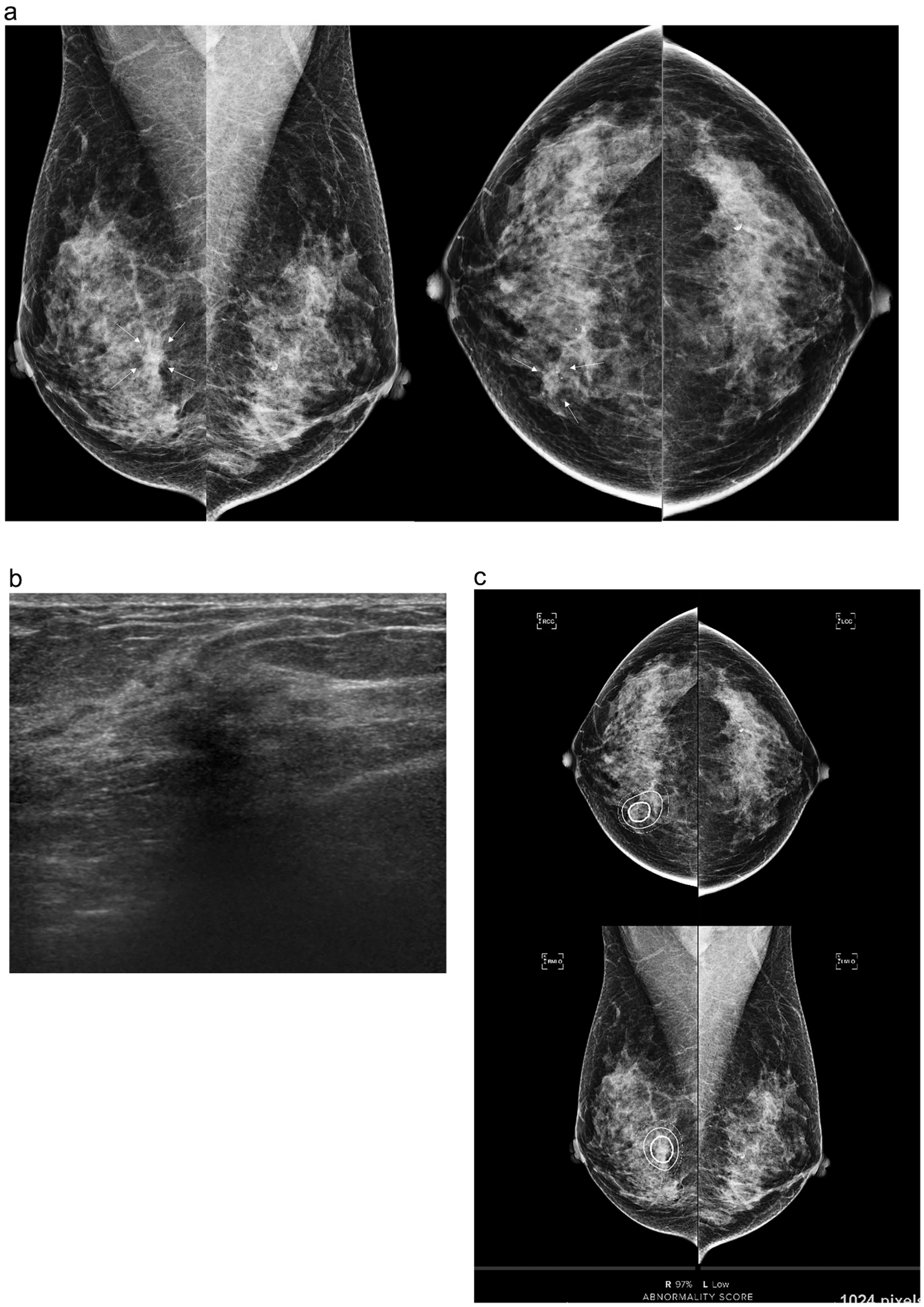

Fig. 3. Representative case of screen-detected cancer initially overlooked by the interpreting radiologist in a 47-year-old woman. (A) The interpreting radiologist had initially overlooked a suspicious mass in right breast (white arrows) and interpreted the examination as BI-RADS 2. (B) Supplemental US performed at the same day revealed a 15-mm sized suspicious mass in the right upper medial quadrant of the breast, pathologically–confirmed as invasive ductal carcinoma. (C) AI-CAD marked the cancer site (white lines) presenting as a suspicious mass with high abnormality score of 97%.

3.3. *Interpretive performances of radiologists vs. stand-alone AI-CAD*

Diagnostic performances of clinical workflow used for mammographic interpretation and AI-CAD are summarized in Table 5. Stand-alone AI-CAD has significantly higher recall rates (10.0% vs. 3.4%, $P < 0.001$) and lower specificity (90.3% vs. 96.9%, $P < 0.001$) compared to the interpreting radiologists. Sensitivity (67.9% vs. 82.1%, $P = 0.086$) and CDRs (2.9/1000 vs. 3.5/1000, $P = 0.102$) did not show significant differences between the radiologists and AI-CAD.

4. Discussion

In this study, we applied a commercially-available AI-CAD software to a retrospectively-collected sample of screening mammograms to simulate the outcomes of AI-CAD analysis results when applied to our interpretation workflow. AI-CAD correctly detected 5 (17.9%) additional cancers of 9 that were initially overlooked by radiologists. Mass/asymmetry or calcifications were more commonly recalled by radiologists, while distortion and multiple features were more commonly recalled by AI-CAD. Although sensitivity and CDRs did not show significant differences between stand-alone AI-CAD and radiologists, AI-CAD had significantly higher recall rates and lower specificity compared to radiologists. Of the 648 AI-CAD recalls, 89.0% (577 of 648) were marks seen in the negative examination group.

Recent studies have consistently reported better performance in breast cancer detection for AI-CAD, either as a stand-alone imaging modality [10,16,19] or as an interpretive assistance tool [13–15]. In our study, approximately 17.9% (5 of 28) of the screen-detected or interval

Table 3
Clinicopathologic features and AI-CAD results of the 14 next-round detected cancers.

No.	Age (years)	Density	Initial interpretation	Imaging features of cancer area on initial screening MG	AI-CAD abnormality score (%) on initial screening MG	Cancer detection modality	Imaging features of cancer	Symptoms at cancer diagnosis	Cancer diagnosis interval (months)*	Diagnosis on surgery
1	48	C	BI-RADS 0asymmetry due to parenchymal summation at contralateral breast	Negative	0.06	MG	Newly developed calcifications	No	15.6	DCIS
2	53	C	BI-RADS 2	Benign calcifications	0.1	MG, US	Newly developed mass with calcifications	Palpable	21.9	IDC
3	44	C	BI-RADS 2	Benign calcifications	0.19	MG, US	Newly developed mass	No	18.1	DCIS
4	52	D	BI-RADS 2	Grouped calcifications (FN)	0.39	MG, US	Newly developed mass with calcifications	Palpable	25.6	IDC
5	45	D	BI-RADS 1	Negative	0.58	US	Mass on US	No	17.8	DCIS
6	53	C	BI-RADS 1	Negative	0.68	MG	Newly developed calcifications	No	25.9	DCIS
7	50	D	BI-RADS 1	Negative	0.69	MG, US	Newly developed mass	Palpable	24.1	IDC
8	64	C	BI-RADS 2	Benign calcifications	0.73	US	Mass on US	No	26.1	IDC
9	44	C	BI-RADS 1	Negative	1.41	US	Mass on US	No	27.0	IDC
10	60	B	BI-RADS 2	Asymmetry (FN)	10.1	MG, US	Asymmetry	Palpable	14.0	IDC
11	45	C	BI-RADS 2	Benign calcification	20.8	US	Mass on US	No	27.3	IDC
12	37	C	BI-RADS 2	Benign calcifications	47.38	MG, US	Newly developed mass	Palpable	22.4	IDC
13	53	C	BI-RADS 1	Distortion (FN)	92.88	MG, US	Mass with distortion	No	21.1	IDC
14	51	C	BI-RADS 1	Calcifications (FN)	95.14	MG, US	Calcifications	No	12.3	DCIS

BI-RADS: Breast Imaging Reporting And Data System, FN: false negative interpretation, MG: mammography, US: ultrasonography, AI-CAD: artificial intelligence-based computer assisted detection/diagnosis, IDC: invasive ductal carcinoma, DCIS: ductal carcinoma in situ

* : interval from initial screening mammography to cancer diagnosis on pathology

cancers that were initially interpreted as negative on mammography were marked as positive by AI-CAD. When retrospectively reviewed, these false-negative examinations had correlating subtle, but suspicious abnormalities at the AI-CAD marks (Fig. 3). Similar results were seen in a recent study[20] in that 39.4% (26.3% with minimal signs and 13.1% considered false-negative) of interval cancers had suspicious features in retrospect of which 67.3% had AI scores in the highest range. By highlighting these subtle cancer features, AI-CAD may guide radiologists into giving the marks a second look and considering additional imaging studies for these subtle, but suspicious findings.

On the other hand, 10.0% (648 of 6449) of the examinations were recalled by AI-CAD, significantly higher than the radiologists' recall rates (3.4%). In particular, the recall rate of 3.4% in this study was lower than the recommended acceptable range of 5–12%[21]. Two 'real--world' factors may have contributed to the lower recall rates in radiologists, 1) radiologists having access to prior mammograms for comparison and 2) 52.3% of the study population had supplemental US performed at the same day. Among the 381 AI-CAD marks considered to necessitate recall, 240 (37.0%, of 648 marks) showed stability on prior examinations that would not have been recalled by the radiologists. Even still, AI-CAD recalls have high false-positive rates as approximately 89% (577 of 628, Table 4) of the AI-CAD marks were in negative examination group, and 41.2% of the AI-CAD marks were considered to be negligible during retrospective review. Increased number of AI-CAD marks can lead to overall increase in recalls when using AI-CAD for interpretation as it did when conventional CAD was used for mammographic interpretations [12,22,23], a negative aspect that we should critically consider. The exhaustive number of markings was a major pitfall of conventional CAD in which upto 97.4% of the marks were reported to be dismissed by the radiologists during interpretation [24]. Although AI-CAD has been reported to have 69% reduction in overall false-positive marks compared to conventional CAD [25], how these false-positive AI-CAD marks affect our interpretation performances when integrated into our workflow needs to be investigated by future prospective studies.

Abnormality features detected by the radiologist vs. AI-CAD showed differences; radiologist-recalled abnormality had higher rates of mass/asymmetry (75.0% vs. 51.7%) and calcifications (21.4% vs. 15.1%), while AI-CAD detected abnormalities had higher rates of distortion (17.7% vs. 0.4%) and multiple features (6.3% vs. 3.2%, $P < 0.001$). Cancer rates of multiple features and calcifications recalled by the radiologist were significantly higher compared to AI-CAD, but not for mass/asymmetry. Similar results were seen in a recent report evaluating the AI features of proven cancers on mammography, 79.6% of cancers presenting as mass with calcifications had AI scores$\geq$ 90% [26]. Even so, ground truth-benign or even 'not definable' lesions being recalled by AI-CAD is not uncommon as seen in our results. These findings show that the final outcomes of abnormalities detected by radiologists and AI-CAD somewhat differs. The trends in differences of specific abnormalities detected between humans and AI-CAD may provide insight on how to effectively implement AI-CAD in our practice, and needs further investigation.

There are several limitations to this study. First, this study is of retrospective design, in which a selection bias is inevitable. Of the 9457 consecutive mammography examinations, 2958 (31.2%) were excluded for various reasons. Second, since AI-CAD was not used during the initial

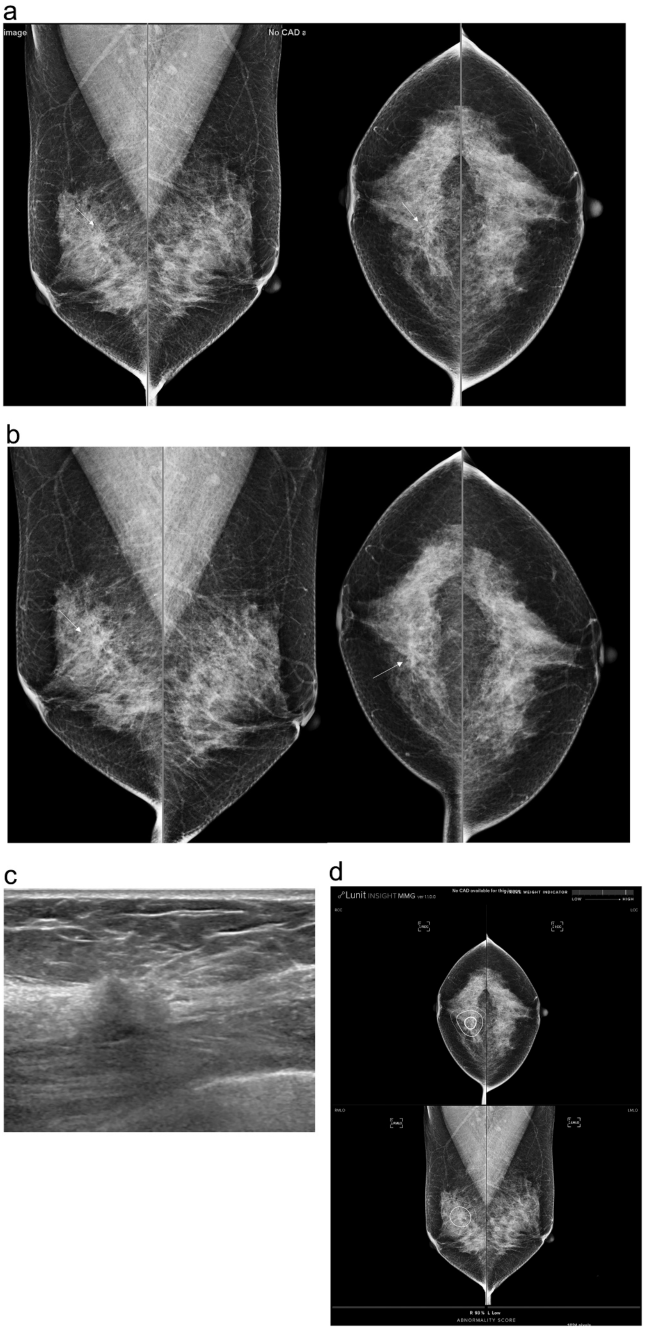

Fig. 4. Representative case of next-screen detected cancer in a 53-year-old woman. (A) The interpreting radiologist had overlooked the subtle distortion in the right breast (white arrows) and initially assessed this examination as BI-RADS 1. The patient had screening mammogram (B) and supplemental US (C) performed 19.9 months later in which the distortion in the right upper medial quadrant was detected (white arrows). Supplemental US performed at the next-round screen revealed a 13-mm sized suspicious mass in the right upper medial quadrant correlating to the distortion on mammography, pathologically–confirmed as invasive ductal carcinoma. (D) Retrospective analysis of the AI-CAD on the initial screening mammogram marked the cancer site with high abnormality score of 93%.

Table 4
Imaging features and outcomes of abnormalities recalled by the radiologist and AI-CAD.

Radiologist	Negative	Benign	Screen-detected cancers	Next round-detected cancer	Total
Abnormal feature					
Not definable	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Mass/ Asymmetry	138 (83.6)	19 (11.5)	7 (4.2)	1 (0.6)	165 (75.0)
Distortion	1 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (0.4)
Calcifications	33 (70.2)	8 (17.0)	6 (12.8)	0 (0.0)	47 (21.4)
Multiple features	0 (0.0)	1 (14.3)	6 (85.7)	0 (0.0)	7 (3.2)
Total	172	28	19	1	220
AI-CAD*	**Negative**	**Benign**	**Screen-detected cancers**	**Next round-detected cancer**	**Total**
Abnormal feature					
Not definable	58 (98.3)	1 (1.7)	0 (0.0)	0 (0.0)	59 (9.1)
Mass/ Asymmetry	303 (90.4)	22 (6.6)	9 (2.7)	1 (0.3)	335 (51.7)
Distortion	109 (94.8)	3 (2.6)	1 (0.9)	2 (1.7)	115 (17.7)
Calcifications	80 (81.6)	13 (13.3)	3 (3.1)	2 (2.0)	98 (15.1)
Multiple features	27 (65.9)	4 (9.8)	10 (24.4)	0 (0.0)	41 (6.4)
Total	577	43	23	5	648

* : recall defined according to abnormality score≥ 10%

Table 5
Diagnostic performances of radiologist vs. stand-alone AI-CAD.

	Radiologists' interpretation	AI-CAD	*P*
Recall rate (%)	3.4 (2.9, 3.8) [220/6499]	10.0 (9.2, 10.7) [648/6499]	< 0.001
CDR (per 1000)	2.9 (1.6, 4.2) [19/6499]	3.5 (2.1, 5.0) [23/6499]	0.102
Sensitivity (%)	67.9 (50.6, 85.2) [19/28]	82.1 (68.0, 96.3) [23/28]	0.086
Specificity (%)	96.9 (96.5, 97.3) [6270/6471]	90.3 (89.6, 91.1) [5846/6471]	< 0.001

95% confidence intervals are in parentheses, raw data are in brackets

AI-CAD: artificial intelligence-based computer assisted detection/diagnosis, CDR: cancer detection rate

Recall rates: number of positive examinations divided by the total screening examinations.

Cancer detection rates (CDR): number of cancers detected per 1000 examinations.

Sensitivity: number of positive examinations with cancer diagnosis within the screening round divided by all cancers within the same period.

Specificity: number of negative examinations with no pathologic diagnosis of cancers within the screening round divided by all examinations with no cancer diagnosis within the same period.

interpretation, the interpretive performances of radiologists before and after using AI-CAD could not be obtained. Also, the interaction between the radiologists and AI-CAD during the interpretation process could not be evaluated in the current study design. Third, patients included in this study consisted of Korean women from a single screening clinic, and results from multinational, multicenter cohorts may differ from this study. In addition, approximately 85.5% of the study sample had mammographically-dense breast, which may have affected the detection of cancers for both the radiologist and AI-CAD. Third, the decision of negative/benign examinations were made based on the results of next round screening examinations (1.7%, 104 of 6282) or benign pathologic diagnosis at the time of screening examinations (13.7%, 26 of 189) with follow up of less than 12 months. Last, 52.3% of the women included had supplemental screening US. Results may have differed if evaluated on a study sample using mammography only for screening.

In conclusion, AI-CAD detected 17.9% of additional cancers that were initially overlooked by the radiologists in a consecutive, screening population. In spite of the additional cancer detection, stand-alone AI-CAD had significantly higher recall rates to the radiologists' interpretation, in which 89% of AI-CAD marks are proven as negative. Abnormal features detected on mammography differ between human readers and AI-CAD and how these differences affect our interpretation should be investigated in depth for consideration of implementing AI-CAD in our interpretation workflow.

Funding statement

No relevant financial support was involved with this study.

Ethical statement

This study was approved by our institutional review board (IRB) with a waiver of informed consent to review medical records and radiologic images.

CRediT authorship contribution statement

Jung Hyun Yoon: Conceptualization, Data curation, Investigation, Project administration. **Kyunghwa Han**: Formal analysis, Writing – review & editing. **Hee Jung Suh**: Data curation, Writing – review & editing. **Ji Hyun Youk**: Writing – review & editing. **Si Eun Lee**: Writing – review & editing. **Eun-Kyung Kim**: Conceptualization, Data curation, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

Authors have no conflicts of interest regarding this manuscript.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ejro.2023.100509.

References

[1] S.W. Duffy, L. Tabar, H.H. Chen, et al., The impact of organized mammography service screening on breast carcinoma mortality in seven Swedish counties, Cancer 95 (2002) 458–469.

[2] L. Nystrom, I. Andersson, N. Bjurstam, J. Frisell, B. Nordenskjold, L.E. Rutqvist, Long-term effects of mammography screening: updated overview of the Swedish randomised trials, Lancet (Lond., Engl.) 359 (2002) 909–919.

[3] R.D. Rosenberg, B.C. Yankaskas, L.A. Abraham, et al., Performance benchmarks for screening mammography, Radiology 241 (2006) 55–66.

[4] M.T. Mandelson, N. Oestreicher, P.L. Porter, et al., Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers, J. Natl. Cancer Inst. 92 (2000) 1081–1087.

[5] R.J. Hooley, K.L. Greenberg, R.M. Stackhouse, J.L. Geisel, R.S. Butler, L. E. Philpotts, Screening US in patients with mammographically dense breasts: initial experience with Connecticut Public Act 09-41, Radiology 265 (2012) 59–69.

[6] J.G. Elmore, S.L. Jackson, L. Abraham, et al., Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy, Radiology 253 (2009) 641–651.

[7] D.L. Miglioretti, C.C. Gard, P.A. Carney, et al., When radiologists perform best: the learning curve in screening mammogram interpretation, Radiology 253 (2009) 632–640.

[8] S. Taplin, L. Abraham, W.E. Barlow, et al., Mammography facility characteristics associated with interpretive accuracy of screening mammography, J. Natl. Cancer Inst. 100 (2008) 876–887.

[9] I. Theberge, S.L. Chang, N. Vandal, et al., Radiologist interpretive volume and breast cancer screening accuracy in a Canadian organized screening program, J. Natl. Cancer Inst. (106) (2014) djt461.

[10] A. Rodriguez-Ruiz, K. Lang, A. Gubern-Merida, et al., Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists, J. Natl. Cancer Inst. (2019), https://doi.org/10.1093/jnci/djy222.

[11] T. Kooi, G. Litjens, B. van Ginneken, et al., Large scale deep learning for computer aided detection of mammographic lesions, Med. Image Anal. 35 (2017) 303–312.

[12] J.H. Yoon, E.K. Kim, Deep Learning-Based Artificial Intelligence for Mammography, Korean J. Radio. 22 (2021) 1225–1239.

[13] A. Rodriguez-Ruiz, E. Krupinski, J.J. Mordang, et al., Detection of breast cancer with mammography: effect of an artificial intelligence support system, Radiology 290 (2019) 305–314.

[14] M. Salim, E. Wåhlin, K. Dembrower, et al., External Evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms, JAMA Oncol. (2020), https://doi.org/10.1001/jamaoncol.2020.3321.

[15] T. Schaffter, D.S.M. Buist, C.I. Lee, et al., Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms, JAMA Netw. Open 3 (2020), e200265.

[16] H.-E. Kim, H.H. Kim, B.K. Han, et al., Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study, Lancet Digit. Health 2 (2020) e138–e148.

[17] L.R. Lamb, C.D. Lehman, A. Gastounioti, E.F. Conant, M. Bahl, Artificial Intelligence (AI) for screening mammography, from the AI special series on AI applications, AJR Am. J. Roentgenol. (2022), https://doi.org/10.2214/ajr.21.27071.

[18] American College of Radiology. Breast imaging reporting and data system, 5th ed., Reston, VA: American College of Radiology, 2013.

[19] S.M. McKinney, M. Sieniek, V. Godbole, et al., International evaluation of an AI system for breast cancer screening, Nature 577 (2020) 89–94.

[20] K. Lång, S. Hofvind, A. Rodríguez-Ruiz, I. Andersson, Can artificial intelligence reduce the interval cancer rate in mammography screening? Eur. Radiol. 31 (2021) 5940–5947.

[21] C.D. Lehman, R.F. Arao, B.L. Sprague, et al., National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium, Radiology 283 (2017) 49–58.

[22] P. Taylor, H.W.W. Potts, Computer aids and human second reading as interventions in screening mammography: Two systematic reviews to compare effects on cancer detection and recall rate, Eur. J. Cancer 44 (2008) 798–807.

[23] M.J. Morton, D.H. Whaley, K.R. Brandt, K.K. Amrami, Screening mammograms: interpretation with computer-aided detection–prospective evaluation, Radiology 239 (2006) 375–383.

[24] T.W. Freer, M.J. Ulissey, Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center, Radiology 220 (2001) 781–786.

[25] R.C. Mayo, D. Kent, L.C. Sen, M. Kapoor, J.W.T. Leung, A.T. Watanabe, Reduction of false-positive markings on mammograms: a retrospective comparison study using an artificial intelligence-based CAD, J. Digit Imaging 32 (2019) 618–624.

[26] S.E. Lee, K. Han, J.H. Yoon, J.H. Youk, E.K. Kim, Depiction of breast cancers on digital mammograms by artificial intelligence-based computer-assisted diagnosis according to cancer characteristics, Eur. Radiol. (2022), https://doi.org/10.1007/s00330-022-08718-2.