Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study

Hyo-Eun Kim, Hak Hee Kim*, Boo-Kyung Han*, Ki Hwan Kim, Kyunghwa Han, Hyeonseob Nam, Eun Hye Lee, Eun-Kyung Kim*

Summary

Background Mammography is the current standard for breast cancer screening. This study aimed to develop an artificial intelligence (AI) algorithm for diagnosis of breast cancer in mammography, and explore whether it could benefit radiologists by improving accuracy of diagnosis.

Methods In this retrospective study, an AI algorithm was developed and validated with 170 230 mammography examinations collected from five institutions in South Korea, the USA, and the UK, including 36 468 cancer positive confirmed by biopsy, 59 544 benign confirmed by biopsy (8827 mammograms) or follow-up imaging (50 717 mammograms), and 74 218 normal. For the multicentre, observer-blinded, reader study, 320 mammograms (160 cancer positive, 64 benign, 96 normal) were independently obtained from two institutions. 14 radiologists participated as readers and assessed each mammogram in terms of likelihood of malignancy (LOM), location of malignancy, and necessity to recall the patient, first without and then with assistance of the AI algorithm. The performance of AI and radiologists was evaluated in terms of LOM-based area under the receiver operating characteristic curve (AUROC) and recall-based sensitivity and specificity.

Findings The AI standalone performance was AUROC 0·959 (95% CI 0·952–0·966) overall, and 0·970 (0·963–0·978) in the South Korea dataset, 0·953 (0·938–0·968) in the USA dataset, and 0·938 (0·918–0·958) in the UK dataset. In the reader study, the performance level of AI was 0·940 (0·915–0·965), significantly higher than that of the radiologists without AI assistance (0·810, 95% CI 0·770–0·850; p<0·0001). With the assistance of AI, radiologists' performance was improved to 0·881 (0·850–0·911; p<0·0001). AI was more sensitive to detect cancers with mass (53 [90%] *vs* 46 [78%] of 59 cancers detected; p=0·044) or distortion or asymmetry (18 [90%] *vs* ten [50%] of 20 cancers detected; p=0·023) than radiologists. AI was better in detection of T1 cancers (73 [91%] *vs* 59 [74%] of 80; p=0·0039) or node-negative cancers (104 [87%] *vs* 88 [74%] of 119; p=0·0025) than radiologists.

Interpretation The AI algorithm developed with large-scale mammography data showed better diagnostic performance in breast cancer detection compared with radiologists. The significant improvement in radiologists' performance when aided by AI supports application of AI to mammograms as a diagnostic support tool.

Funding Lunit.

Lancet Digital Health 2020; 2: e138–48

Published Online February 6, 2020 https://doi.org/10.1016/S2589-7500(20)30003-0

See Comment page e106

*Contributed equally

Lunit, Seoul, South Korea (H-E Kim PhD, K H Kim MD, H Nam MS); **Department of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, South Korea** (H H Kim MD); **Department of Radiology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, South Korea** (B-K Han MD); **Department of Radiology, Research Institute of Radiological Science and Center for Clinical Imaging Data Science, Severance Hospital, Yonsei University College of Medicine, Seoul, South Korea** (K Han PhD, E-K Kim MD); **and Department of Radiology, Soonchunhyang University Hospital Bucheon, Soonchunhyang University College of Medicine, Bucheon, South Korea** (E H Lee MD)

Correspondence to: Prof Eun-Kyung Kim, Department of Radiology, Research Institute of Radiological Science and Center for Clinical Imaging Data Science, Severance Hospital, Yonsei University College of Medicine, Seoul 03722, South Korea **ekkim@yuhs.ac**

Introduction

Multiple randomised controlled studies have shown that mammographic screening significantly reduces breast cancer mortality.[1,2] Despite such efforts, breast cancer is still the most common cancer and the leading cause of cancer-related deaths in women across the world.[3] Other diagnostic methods such as tomosynthesis, ultrasound, or MRI have been proposed, but mammographic screening remains the most commonly used in the world; therefore, accurate reading of mammograms is important to maximise the effectiveness of mammographic screening.

In mammography, 10–30% of breast cancers can be missed, which is commonly attributed to dense parenchyma obscuring lesions, poor positioning, perception error, and interpretation error, among other reasons.[4] It should be noted that efforts to reduce false negatives can sometimes lead to excessive recalls. In the USA, 41% of radiologists showed a higher recall rate than the recommendation, and only 28·6% of the patients who received biopsy were subsequently diagnosed as having cancer.[5] Furthermore, inter-reader variability in breast cancer detection and recall rates is a substantial issue. This implies that interpretation of mammograms is difficult, and extensive experience is required to arrive at an adequate level of interpretive performance in reading mammograms.[6]

Two decades ago, computer-aided detection (CAD) for mammography was developed to assist mammogram interpretation.[7] Early studies have shown that traditional CAD is somewhat beneficial in terms of cancer detection (ie, sensitivity),[8,9] especially in cases of microcalcification.[8] However, its effectiveness has been heavily challenged by

Research in context

Evidence before this study

We searched for studies that used artificial intelligence or deep learning technology, focusing on computer-aided diagnosis of breast cancer in mammography. We searched PubMed for articles published before Jan 2, 2020, with the terms "deep learning" OR "machine learning" OR "artificial intelligence" AND "mammography" AND "breast cancer". We also reviewed a reference list of eligible texts and found several studies on development and validation of artificial intelligence (AI) algorithms. All of the algorithms were developed using mammography data of fewer than 5000 patients with breast cancer. Additionally, most previous studies used data collected from one or two institutions for development of their AI algorithms, and there were no multinational and multicentre studies to cover various imaging devices, scanning conditions, and ethnic diversity. Thus, the previous studies could not verify robustness of the developed AI algorithms, which is the major concern in real-field applications. Regarding performance evaluation metrics, localisation of lesions needs to be assessed to confirm that AI has detected malignant lesions correctly, but most previous studies have only evaluated mammogram-level performance.

Added value of this study

We have developed an AI algorithm that uses the largest breast cancer dataset among known AI algorithms to detect breast cancer. Because the algorithm was trained with data from various institutions, it was able to show comparable performance in validation datasets from different countries. With the aid of the large-scale mammography data, the AI algorithm showed improved diagnostic performance compared with radiologists, especially in early-stage invasive breast cancers. For better understanding of AI behaviour, mammographic features of cancers detected by the AI algorithm were analysed through the comparison study with radiologists.

Implications of all the available evidence

This study shows that AI has the potential to improve early-stage breast cancer detection in mammography. Especially in dense breast areas on a mammogram which pose one of the major difficulties in screening, the performance of radiologists was significantly improved when aided with AI. Such improvements could result in an increase in screen-detected cancers and decrease in interval cancers, which would improve the efficacy of mammography screening. Real-world clinical benefit needs to be evaluated by future prospective studies.

recent large-scale clinical trials, in which CAD has failed to improve radiologists' diagnostic performance.[10–12] Due to its high false-positive rate, radiologists are required to review numerous false-positive marks of CAD, leading to exhaustion and an increase in unnecessary additional examinations.[10]

Radiologists have sought to characterise mammographic differences between cancer and non-cancer by reviewing many images, and cancer-specific mammographic characteristics have been reported and shared with radiologists using their morphological descriptors. Traditional CAD mimics this process. In traditional CAD, however, important information is prone to being lost when designing human-interpretable descriptors. In recent artificial intelligence (AI)-based CAD, the AI algorithm abstracts mammographic features as a descriptor. The difference between human-designed and self-learned descriptors is the main success factor of current deep learning algorithms. It has already been reported that AI can achieve similar performance to experts in medical image analysis.[13,14]

In this study, we developed and validated an AI algorithm to detect breast cancer on mammograms, and explored whether it could improve the performance of radiologists in breast cancer detection.

Methods

Study design

In this retrospective study, we used data from five institutions to develop and validate an AI algorithm to detect breast cancer on mammograms. We validated the AI algorithm with mammograms from three countries and compared results from the AI algorithm with assessments made by radiologists using separate cancer-enriched mammography data from two institutions.

This study was approved by ethics review and institutional review board from participating institutions, and the requirement for informed consent was waived. Under this approval, mammography examinations were de-identified and collected according to the Health Insurance Portability and Accountability Act Safe Harbor standard.

Development dataset

To develop the AI algorithm for our diagnostic support software, we obtained 170 230 four-view, full-field, digital mammograms (ie, left and right craniocaudal and mediolateral oblique) from five institutions: three in Seoul, South Korea (Yonsei University Severance Hospital, Asan Medical Center, Samsung Medical Center), one in the USA (Wake Radiology Diagnostic Imaging, covering North Carolina), and one in the UK (National Health Service OPTIMAM database; figure 1). The data collection periods were January, 2004–December, 2016, in South Korea; January, 2000–December, 2018, in the USA; and January, 2010–December, 2018, in the UK. The mean age of patients in the datasets was 50·3 years (SD 10·0). The mammograms were done using GE (69·9%), Hologic (28·0%), and Siemens (1·9%) systems, with 0·2% unknown. We included both screening and diagnostic

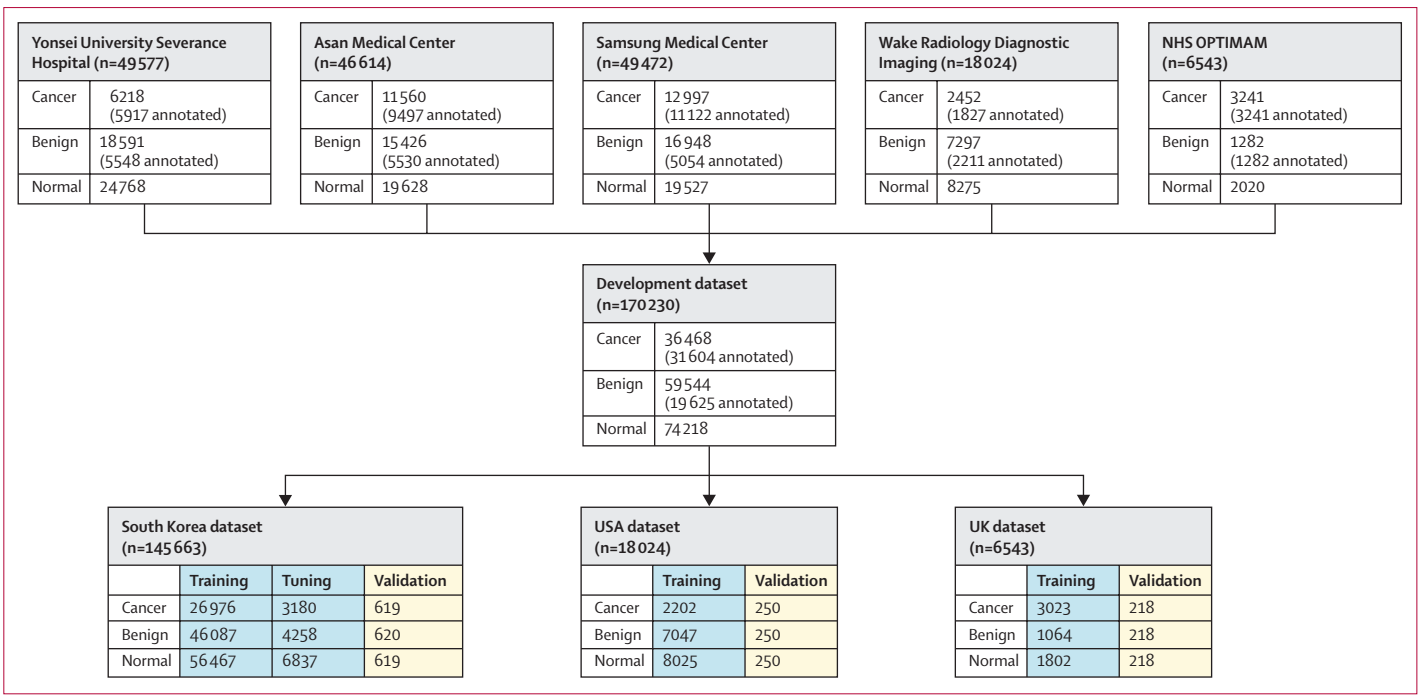

Figure 1: **Development dataset generation and partitioning**
All mammograms are four-view paired (left and right craniocaudal and mediolateral oblique). There was no overlap between categories (cancer, benign, and normal). NHS=National Health Service.

mammograms; ambiguity of the ground-truth label is a major methodological deficiency of previous medical AI studies,[15,16] so we focused on collecting large-scale mammography data with accurate ground-truth label regardless of whether data were obtained for screening or diagnosis. Among 170 230 mammograms, 36 468 (21·4%) were cancer positive confirmed by biopsy, 59 544 (35·0%) were benign confirmed by biopsy (8827 [5·2%]) or at least 1 year of follow-up imaging (50 717 [29·8%]), and 74 218 (43·6%) were normal confirmed by at least 1 year of follow-up imaging (figure 1). For cancer-positive and biopsy-proven benign mammograms, we restricted our data to one mammogram per woman. We allowed multiple mammograms per woman in normal or follow-up-proven benign mammograms in our dataset, but these would have been taken on different dates (ie, independent mammograms). The entire dataset was divided into three sets without patient-level overlap: a training set for training an AI model, a tuning set for selection of the training scenario, and a validation set for evaluation of the final model (figure 1). Once the training scenario was selected using the tuning set, both the training and tuning datasets were used to train a final model.

Reader study dataset

The purpose of the reader study was to assess the applicability of the developed AI model on screening mammography data; as the development dataset contained both screening and diagnostic mammograms, we obtained a separate set of screening data for the reader study. 400 four-view digital mammograms were obtained from two institutions (institution A: Yonsei University Severance Hospital, Seoul, South Korea; institution B: Soonchunhyang University Hospital Bucheon, Bucheon, South Korea; appendix p 2). The reader study dataset was cancer enriched, with cancer prevalence of 50%, similar to a previous study.[17] Readers were not informed of the enrichment levels in the dataset. Data were collected from patient samples between April, 2014, and January, 2018, for institution A and between March, 2009, and September, 2018, for institution B. All the reader study data consist of screening mammograms using GE (50·0%) or Hologic (50·0%) systems. Cancer-positive mammograms in the reader study dataset were either mammography detected or mammography missed but ultrasound detected. Note that examining both mammography and ultrasound at the same time for breast cancer screening is common in South Korea. 80% of the data were randomly selected from each category to meet our sample size of 320 mammograms (320 women; mean age 53·19 years [SD 10·01]) for the reader study (appendix p 2). A summary of the reader study population is shown in the appendix (p 3), including radiological lesion features (soft tissue only: mass, asymmetry, distortion; otherwise, calcification with or without soft tissue), pathological cancer subtypes, lesion size, and Breast Imaging Reporting and Data System (BI-RADS)[18] breast composition categories. The reader study data had a high prevalence of dense breast (categories C and D)

See Online for appendix

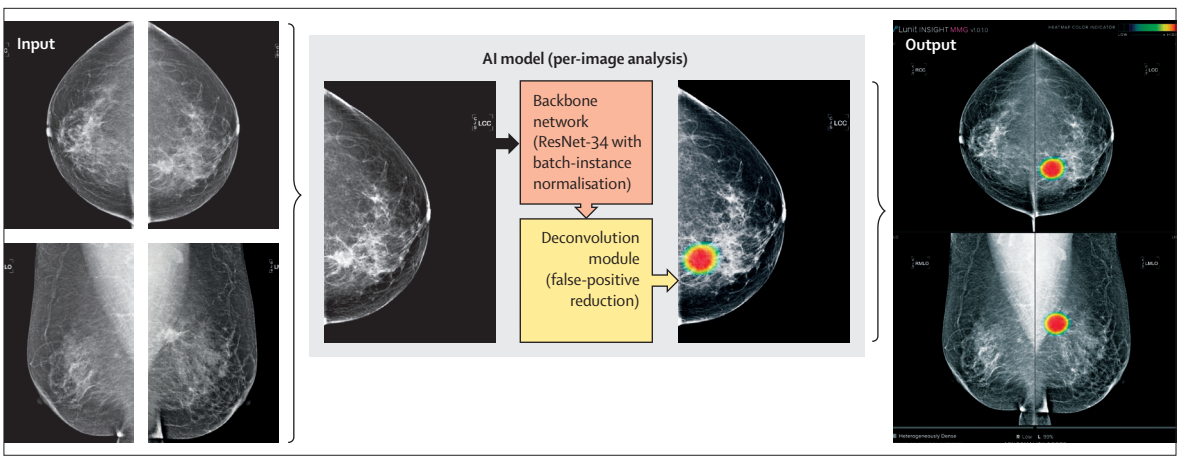

Figure 2: **AI-based diagnostic support software**
AI=artificial intelligence.

mammograms (216 [68%] of 320) and invasive cancers (123 [77%] of 160; appendix p 3).

Development of the AI algorithm

For the purpose of AI algorithm development, 31 604 cancer-positive mammograms (86·7% of cancers) and 19 625 benign mammograms (33·0%) were annotated by one of 12 radiologists with breast subspecialty by referring to previous radiology and pathology reports. For each case included in the mammogram study, per-side information (ie, cancer, benign, or normal) was extracted from the radiology and pathology reports; therefore, all mammograms have image-level labels, and 86·7% of cancer-positive mammograms and 33·0% of benign mammograms have pixel-level labels that indicate the location of lesions.

An AI algorithm was developed on the basis of deep convolutional neural networks (CNNs). ResNet-34, one of the most popular CNN architectures, was used as a backbone network.[19] The algorithm training consists of two stages: patch-level training from scratch for learning low-level features (stage 1), followed by image-level fine-tuning from the stage-1 model for learning high-level context (stage 2). Only lesion-annotated mammograms were used in stage 1 (fully supervised), whereas all mammograms were used in stage 2 (semi-supervised). Batch-instance normalisation[20] and a deconvolution module[21] were additionally adopted to overcome variance of pixel-level characteristics (mainly due to the different imaging acquisition devices) and increase of false positives, respectively. For an input mammogram image (ie, one of the four views), the AI algorithm provides pixel-level abnormality scores as a heatmap (figure 2) and a representative abnormality score, which is the maximum of the pixel-level abnormality scores. The abnormality scores are floating-point values between 0 and 1. Based on a per-image analysis of the algorithm, the resulting diagnostic support software (Lunit INSIGHT MMG) provides four-view heatmaps and an abnormality score per breast (ie, the maximum of the craniocaudal and mediolateral oblique abnormality scores) for each input mammogram (figure 2). Details of the algorithm are specified in the appendix (p 4).

Validation of AI-based diagnostic support software

We used the per-mammogram abnormality score—ie, the maximum of abnormality scores of each of the four-views—to evaluate AI standalone performance, including area under the receiver operating characteristic (ROC) curve (AUROC), sensitivity, and specificity. The cutoff threshold between 0 and 1 for measuring sensitivity and specificity was set to 0·1 to achieve 90% sensitivity in the tuning dataset, and this threshold was also used for validation and the reader study. AI standalone performance was evaluated with three validation datasets from different countries: South Korea (619 cancer-positive, 620 benign, and 619 normal mammograms), the USA (250 cancer-positive, 250 benign, and 250 normal mammograms), and the UK (218 cancer-positive, 218 benign, and 218 normal mammograms; figure 1). To explore how multinational, large-scale datasets affect the performance of AI, we also trained the same AI algorithm using just the South Korea dataset (ie, single nationality) and with subsets of the South Korea dataset (ie, a smaller scale in terms of cancer).

Reader study

A multicentre, observer-blinded study was done with 14 radiologists from four institutions in South Korea (Samsung Medical Center, Seoul; Asan Medical Center, Seoul; Uljin Medical Center, Gyeongsangbuk-do; and Chung-Ang University Hospital, Seoul) using the 320 screening mammograms in the reader study dataset. There was no overlap between the readers' institutions and the data collection institutions, nor between the 14 radiologists in the reader study and the 12 radiologists who annotated the development dataset. The number of

	AUROC	AULROC	Sensitivity*	Specificity*
All (n=3262)	0·959 (0·952–0·966)	0·796 (0·776–0·814)	0·914 (0·897–0·930)	0·860 (0·845–0·875)
South Korea (n=1858)	0·970 (0·963–0·978)	0·775 (0·746–0·804)	0·903 (0·880–0·926)	0·917 (0·901–0·932)
USA (n=750)	0·953 (0·938–0·968)	0·812 (0·774–0·849)	0·936 (0·906–0·966)	0·802 (0·767–0·837)
UK (n=654)	0·938 (0·918–0·958)	0·829 (0·788–0·867)	0·917 (0·881–0·954)	0·768 (0·729–0·808)

95% CIs are given in parentheses. AI=artificial intelligence. AUROC=area under the receiver operating characteristic curve. AULROC=area under the localisation receiver operating characteristic curve. *Sensitivity and specificity were calculated with the cutoff threshold of 0·1 (ie, if the abnormality score is ≥0·1, then positive; otherwise, negative).

Table 1: **Performance of the AI algorithm on three validation datasets**

readers and mammograms required was calculated by the power estimation method (significance level set to 5% and power to 80%)[22] with an effect size of 0·03 from a similar previous study.[17] The 14 radiologists consisted of seven breast specialists and seven general radiologists. Both groups were board-certified radiologists, but general radiologists had not been specifically trained in breast imaging whereas breast specialists had been trained in breast imaging for at least 6 months.

The overall procedure of the reader study is summarised in the appendix (p 5). All mammograms in the study were assessed by every reader, with an inspector who managed and controlled the process to avoid cross-reader consultation. For each mammogram, AI-unaided (test 1) and AI-aided (test 2) readings were done sequentially by the same reader, which is the usual approach for second-reader style observer performance studies.[23] In test 1, each radiologist reviewed a mammogram and made a binary decision of whether it should be recalled—ie, if there existed a suspicious lesion for breast cancer. If recalled, then the radiologist localised the most suspicious lesion for breast cancer by putting a point-mark on the centre of the lesion and graded the mammogram with two scores: probability of malignancy (POM) and likelihood of malignancy (LOM). POM scores are given on a 0–100 scale (0 definite non-cancer, 1–25 probably non-cancer, 26–50 possibly non-cancer, 51–75 possibly cancer, 76–99 probably cancer, 100 definite cancer)[17] and provide a confidence level of a radiologist's reading that a malignant lesion is present in the mammogram. LOM scores are given on a 1–7 scale (1 definite normal, 2 benign, 3 probably benign, 4 low suspicion for malignancy, 5 moderate suspicion for malignancy, 6 high suspicion for malignancy, 7 highly suggestive of malignancy)[24] and provide suspicion scales of malignancy. POM scores were used for evaluation of detection performance and LOM scores were used for evaluation of diagnostic performance. Since BI-RADS assessment categories do not constitute an ordinal scale, it is inappropriate for ROC analysis;[25] hence, LOM, which is modified to be ordinal from the BI-RADS categories, is used in breast imaging.[24] If a radiologist decided not to recall, localisation was not needed; in this case, POM was zero. In test 2, each radiologist modified their original decision in test 1 by referring to the output result of AI. For evaluation of the localisation, two experts (E-KK, EHL) with more than 20 years of experience in breast imaging annotated the location of malignant lesions with a free-form line of contour by referring to the radiology and pathology reports.

To assess the effectiveness of AI, mammogram-level LOM-based AUROC was used as a primary endpoint. Secondary endpoints were mammogram-level POM-based area under the localisation ROC curve (AULROC) and recall-based sensitivity and specificity. Mammographic and pathological characteristics of breast cancers detected by AI and radiologists were also compared. To effectively compare the performance of AI (single) with the readers (multiple), we used a reader representative score: a cancer-positive case was deemed correctly detected by readers if more than half of the readers identified it correctly, whereas it was deemed to be correctly detected by the AI algorithm if the AI prediction score was greater than or equal to 0·1.

Statistical analysis

ROC and localisation ROC curve analyses were done to evaluate the performance of the AI algorithm and radiologists. In the reader study, multireader, multicase ROC curve analysis was used to account for reader variability and the correlation among ratings before and after AI assistance.[22] Readers and cases were treated as random effects, and the non-parametric trapezoidal method was used to estimate AUROC. AULROC was measured with the non-parametric trapezoidal method from the localisation ROC curve—a plot of the x-axis representing a false-positive fraction against the y-axis representing a true-positive localisation fraction.[26] In the AI standalone assessment, localisation was regarded as correct if the location of the maximum of pixel-level abnormality scores was inside the closed free-form line of the reference standard drawn by radiologists. In the reader study, correctness of localisation was determined on the basis of whether reader's point-mark was inside the reference standard. For analysis of sensitivity and specificity, the Clopper-Pearson method was used for estimating 95% CIs, and logistic regression with generalised estimating equation (GEE) method was used for significance testing and for estimating 95% CIs for the difference. Logistic regression with GEE method was also used for comparison between cancers detected by the AI algorithm and radiologists. We did several

	AUROC (diagnosis; LOM based)							AULROC (detection; POM based)						
	AI	Test 1*	Test 2*	Difference in AUROC				AI	Test 1*	Test 2*	Difference in AULROC			
				AI *vs* test 1	p value	Test 2 *vs* test 1	p value				AI *vs* test 1	p value	Test 2 *vs* test 1	p value
Overall														
Cancer (n=160) *vs* non-cancer (n=160)	0·940 (0·915 to 0·965)	0·810 (0·770 to 0·850)	0·881 (0·850 to 0·911)	0·130 (0·091 to 0·169)	<0·0001	0·071 (0·045 to 0·096)	<0·0001	0·812 (0·783 to 0·841)	0·647 (0·576 to 0·718)	0·765 (0·714 to 0·815)	0·165 (0·094 to 0·236)	<0·0001	0·118 (0·075 to 0·161)	<0·0001
Reading panel†														
General: cancer (n=160) *vs* non-cancer (n=160)	0·940 (0·915 to 0·965)	0·772 (0·729 to 0·816)	0·869 (0·834 to 0·903)	0·168 (0·125 to 0·210)	<0·0001	0·096 (0·059 to 0·133)	0·0001	0·812 (0·783 to 0·841)	0·589 (0·508 to 0·670)	0·748 (0·683 to 0·814)	0·223 (0·142 to 0·304)	<0·0001	0·159 (0·101 to 0·218)	<0·0001
Specialist: cancer (n=160) *vs* non-cancer (n=160)	0·940 (0·915 to 0·965)	0·847 (0·809 to 0·886)	0·892 (0·859 to 0·926)	0·093 (0·056 to 0·130)	<0·0001	0·045 (0·027 to 0·063)	<0·0001	0·812 (0·783 to 0·841)	0·705 (0·632 to 0·778)	0·781 (0·723 to 0·839)	0·107 (0·034 to 0·180)	0·0054	0·076 (0·044 to 0·109)	<0·0001
Age														
<50 years: cancer (n=52) *vs* non-cancer (n=63)	0·940 (0·899 to 0·980)	0·764 (0·698 to 0·829)	0·858 (0·804 to 0·912)	0·176 (0·116 to 0·236)	<0·0001	0·095 (0·051 to 0·138)	0·0001	0·804 (0·751 to 0·857)	0·608 (0·501 to 0·716)	0·749 (0·672 to 0·826)	0·196 (0·088 to 0·303)	0·0004	0·141 (0·076 to 0·206)	<0·0001
≥50 years: cancer (n=108) *vs* non-cancer (n=97)	0·945 (0·916 to 0·975)	0·839 (0·797 to 0·880)	0·898 (0·866 to 0·930)	0·107 (0·065 to 0·148)	<0·0001	0·059 (0·036 to 0·082)	<0·0001	0·820 (0·786 to 0·855)	0·670 (0·592 to 0·749)	0·777 (0·719 to 0·835)	0·150 (0·072 to 0·229)	0·0003	0·106 (0·062 to 0·151)	<0·0001
BI-RADS breast composition categories														
Fatty (A or B): cancer (n=44) *vs* non-cancer (n=60)	0·948 (0·901 to 0·996)	0·861 (0·807 to 0·915)	0·905 (0·858 to 0·952)	0·088 (0·028 to 0·147)	0·0040	0·044 (0·015 to 0·073)	0·0041	0·878 (0·833 to 0·923)	0·719 (0·627 to 0·811)	0·789 (0·709 to 0·868)	0·159 (0·068 to 0·251)	0·0008	0·070 (0·019 to 0·120)	0·0086
Dense (C or D); cancer (n=116) *vs* non-cancer (n=100)	0·932 (0·901 to 0·963)	0·782 (0·735 to 0·830)	0·866 (0·828 to 0·903)	0·150 (0·105 to 0·195)	<0·0001	0·083 (0·054 to 0·113)	<0·0001	0·789 (0·753 to 0·825)	0·617 (0·536 to 0·699)	0·752 (0·695 to 0·809)	0·171 (0·090 to 0·252)	0·0001	0·135 (0·088 to 0·182)	<0·0001
Lesion feature‡														
Soft tissue: cancer (n=79) *vs* benign (n=36)	0·919 (0·870 to 0·969)	0·668 (0·593 to 0·744)	0·793 (0·731 to 0·855)	0·251 (0·177 to 0·324)	<0·0001	0·125 (0·078 to 0·172)	<0·0001	0·787 (0·743 to 0·832)	0·539 (0·448 to 0·630)	0·683 (0·619 to 0·746)	0·248 (0·158 to 0·339)	<0·0001	0·143 (0·084 to 0·203)	<0·0001
Calcification: cancer (n=81) *vs* benign (n=28)	0·854 (0·785 to 0·922)	0·767 (0·707 to 0·827)	0·829 (0·772 to 0·885)	0·086 (0·024 to 0·148)	0·0066	0·061 (0·029 to 0·094)	0·0004	0·771 (0·727 to 0·815)	0·676 (0·593 to 0·760)	0·765 (0·707 to 0·824)	0·095 (0·011 to 0·178)	0·027	0·089 (0·046 to 0·132)	<0·0001
Pathological subtype														
Invasive (n=123) *vs* non-cancer (n=160)	0·954 (0·932 to 0·976)	0·806 (0·764 to 0·848)	0·886 (0·854 to 0·918)	0·148 (0·108 to 0·187)	<0·0001	0·080 (0·052 to 0·108)	<0·0001	0·837 (0·805 to 0·868)	0·653 (0·580 to 0·727)	0·790 (0·731 to 0·848)	0·184 (0·106 to 0·261)	<0·0001	0·136 (0·088 to 0·185)	<0·0001
Non-invasive (n=37) *vs* non-cancer (n=160)	0·894 (0·826 to 0·961)	0·823 (0·762 to 0·883)	0·862 (0·805 to 0·919)	0·071 (0·008 to 0·134)	0·027	0·039 (0·011 to 0·068)	0·0094	0·728 (0·660 to 0·797)	0·625 (0·503 to 0·748)	0·682 (0·557 to 0·807)	0·103 (−0·021 to 0·226)	0·10	0·057 (0·004 to 0·109)	0·034

Data are accompanied by 95% CIs (in parentheses) if appropriate. AUROC=area under the receiver operating characteristic curve. AULROC=area under the localisation receiver operating characteristic curve. LOM=likelihood of malignancy. POM=probability of malignancy. BI-RADS=Breast Imaging Reporting and Data System. *Test 1 was AI-unaided radiologist readings and test 2 was subsequent AI-aided radiologist readings. †Reading panel subgroup analysis compared general radiologists with breast specialists. ‡Normal mammograms were excluded in this subgroup analysis.

Table 2: **Diagnostic performance of AI and radiologists and comparisons between AI, AI-aided radiologists, and AI-unaided radiologists**

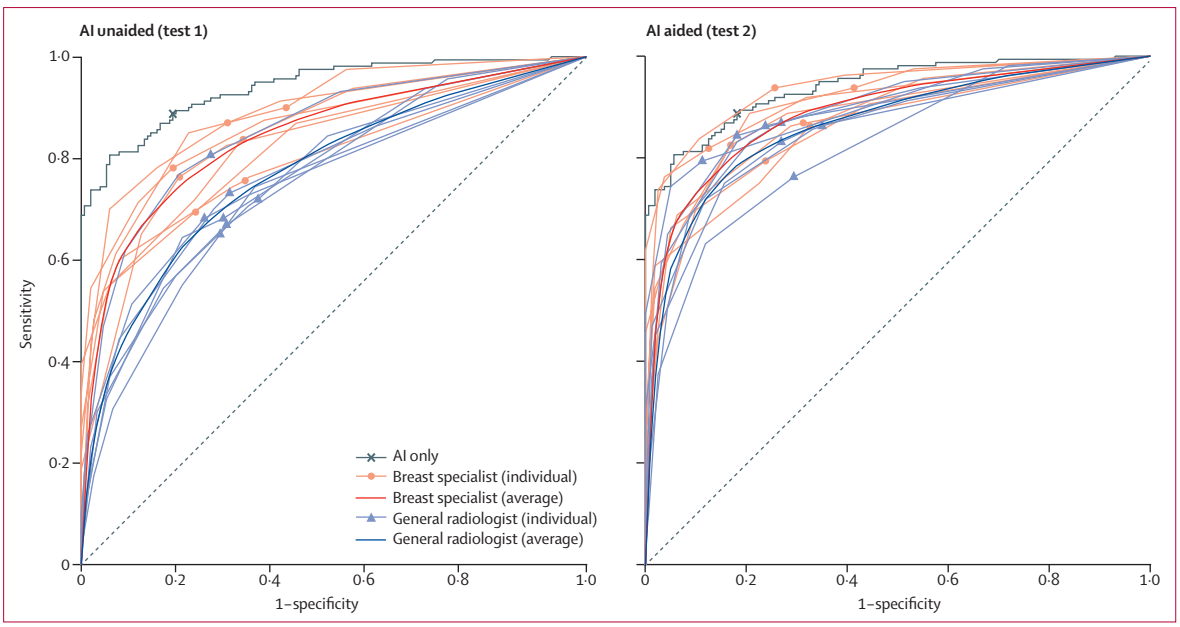

Figure 3: **ROC analysis for AI-unaided and AI-aided diagnosis**
Sensitivity and specificity of each individual (including AI standalone) are marked on each curve. AI=artificial intelligence. ROC=receiver operating characteristic.

subgroup analyses: reading panel (general radiologist *vs* breast specialist), age (<50 years *vs* ≥50 years), BI-RADS composition categories (fatty [class A or B] *vs* dense [class C or D]), lesion feature (soft tissue *vs* calcification), and pathological subtype (invasive *vs* non-invasive). Interaction effects between each subgroup and the assistance of AI were additionally tested on the logistic regression model. SAS (version 9·4) was used for analysis of sensitivity and specificity, PyTorch (version 0.4) was used for deep learning model development and validation, and R (version 3.6.1) was used for the rest of analyses.

Role of the funding source

The funder of the study was involved in collection, management, and analysis of the dataset used in the AI algorithm development; and preparation and review of the manuscript. The corresponding author had full access to most datasets and all summary estimates from each dataset, and had final responsibility for the decision to submit for publication.

Results

When considering AI standalone performance, overall AUROC in the three validation datasets was 0·959 (95% CI 0·952–0·966), whereas the individual performance was 0·970 (0·963–0·978) in the South Korea dataset, 0·953 (0·938–0·968) in the USA dataset, and 0·938 (0·918–0·958) in the UK dataset (table 1).

To explore how patient nationality and data scale affect performance, we trained and validated the algorithm on the South Korean data alone (training: 30 156 cancer, 50 345 benign, 63 304 normal; figure 1). When increasing the number of cancer-positive mammograms included in the training dataset, while maintaining the full set of 50 345 benign and 63 304 normal mammograms, we found that performance continued to improve as the scale of cancer-positive mammograms increased (AUROC of validation set 0·919 [95% CI 0·904–0·935] with a tenth [n=3000] of cancer-positive mammograms; 0·951 [0·940–0·962] with a sixth [n=5000]; 0·962 [0·953–0·972] with a third [n=10 000]; and 0·974 [0·966–0·981] with the full set [n=30 156]). However, when validating the algorithm trained on the full South Korea dataset on the USA and UK datasets, we observed a decrease in performance, with an AUROC of 0·909 (0·887–0·931) for the USA and 0·871 (0·841–0·901) for the UK (table 1).

In the reader study, overall diagnostic performance of radiologists was AUROC 0·810 (95% CI 0·770–0·850), compared with the AI standalone performance of 0·940 (0·915–0·965; p<0·0001; table 2). All of the readers' ROC curves were inside the AI standalone ROC curve (figure 3). When aided by AI (ie, test 2), radiologists' performance was significantly improved to 0·881 (0·850–0·911; p<0·0001; table 2; figure 3; appendix pp 6–7).

In the reading panel subgroup analysis, the AUROC of general radiologists improved from 0·772 (95% CI 0·729–0·816) to 0·869 (0·834–0·903; p=0·0001) when aided by AI, achieving comparable performance to the breast specialist group (table 2). The improvement in AUROC between AI-unaided and AI-aided radiologists was more noticeable in dense breasts, with a difference of 0·083 (0·054–0·113; p<0·0001), whereas it was 0·044 (0·015–0·073; p=0·0041) in fatty breasts (table 2). When considering lesion features, the AUROC difference between standalone AI and AI-unaided radiologists was

	AI*	Test 1†	Test 2†	Difference			
				AI *vs* test 1	p value	Test 2 *vs* test 1	p value
Sensitivity							
Overall (n=160)	88·75% (82·80 to 93·19)	75·27% (73·43 to 77·04)	84·78% (83·22 to 86·24)	13·48 (8·65 to 18·32)	<0·0001	9·51 (6·86 to 12·16)	<0·0001
Reading panel‡							
General (n=160)	88·75% (82·80 to 93·19)	70·54% (67·77 to 73·19)	83·21% (80·89 to 85·36)	18·21 (12·97 to 23·46)	<0·0001	12·68 (9·24 to 16·12)	<0·0001
Specialist (n=160)	88·75% (82·80 to 93·19)	80·00% (77·54 to 82·31)	86·34% (84·19 to 88·30)	8·75 (3·80 to 13·70)	0·0005	6·34 (3·99 to 8·68)	<0·0001
Age							
<50 years (n=52)	90·38% (78·97 to 96·80)	74·31% (70·98 to 77·45)	85·71% (82·96 to 88·18)	16·07 (7·41 to 24·74)	0·0003	11·40 (6·35 to 16·46)	<0·0001
≥50 years (n=108)	87·96% (80·30 to 93·43)	75·73% (73·48 to 77·87)	84·33% (82·39 to 86·12)	12·24 (6·43 to 18·04)	<0·0001	8·60 (5·54 to 11·66)	<0·0001
BI-RADS composition categories							
Fatty (A or B; n=44)	86·36% (72·65 to 94·83)	79·22% (75·80 to 82·36)	84·09% (80·96 to 86·89)	7·14 (0·09 to 14·20)	0·047	4·87 (1·89 to 7·85)	0·0013
Dense (C or D; n=116)	89·66% (82·63 to 94·54)	73·77% (71·56 to 75·89)	85·04% (83·21 to 86·74)	15·89 (9·84 to 21·94)	<0·0001	11·27 (7·85 to 14·69)	<0·0001
Lesion feature§							
Soft tissue (n=79)	89·87% (81·02 to 95·53)	71·43% (68·67 to 74·08)	83·09% (80·75 to 85·26)	18·44 (11·08 to 25·81)	<0·0001	11·68 (7·45 to 15·87)	<0·0001
Calcification (n=81)	87·65% (78·47 to 93·92)	79·01% (76·52 to 81·35)	86·42% (84·29 to 88·36)	8·64 (2·53 to 14·76)	0·0056	7·41 (4·24 to 10·58)	<0·0001
Pathology							
Invasive (n=123)	90·24% (83·58 to 94·86)	75·55% (73·45 to 77·57)	86·59% (84·88 to 88·16)	14·69 (9·11 to 20·27)	<0·0001	11·03 (7·80 to 14·27)	<0·0001
Non-invasive (n=37)	83·78% (67·99 to 93·81)	74·32% (70·33 to 78·03)	78·76% (74·99 to 82·21)	9·46 (-0·05 to 18·97)	0·051	4·44 (0·96 to 7·92)	0·013
Specificity							
Overall (n=160)	81·87% (75·02 to 87·51)	71·96% (70·05 to 73·82)	74·64% (72·79 to 76·43)	9·91 (3·69 to 16·13)	0·0018	2·68 (1·33 to 4·03)	<0·0001
Reading panel‡							
General (n=160)	81·87% (75·02 to 87·51)	71·61% (68·87 to 74·23)	75·54% (72·91 to 78·03)	10·27 (3·73 to 16·81)	0·0021	3·93 (1·92 to 5·94)	0·0001
Specialist (n=160)	81·87% (75·02 to 87·51)	72·32% (69·60 to 74·92)	73·75% (71·07 to 76·31)	9·55 (3·32 to 15·79)	0·0027	1·43 (0·05 to 2·81)	0·043
Age							
<50 years (n=63)	77·78% (65·54 to 87·28)	61·68% (58·38 to 64·90)	64·17% (60·91 to 67·34)	16·10 (5·69 to 26·51)	0·0024	2·49 (-0·21 to 5·20)	0·071
≥50 years (n=97)	84·54% (75·78 to 91·08)	78·65% (76·37 to 80·80)	81·44% (79·27 to 83·48)	5·89 (-1·72 to 13·50)	0·13	2·80 (1·44 to 4·16)	<0·0001
BI-RADS composition categories							
Fatty (A or B; n=60)	93·33% (83·80 to 98·15)	79·52% (76·63 to 82·20)	83·21% (80·51 to 85·68)	13·81 (4·74 to 22·88)	0·0028	3·69 (2·20 to 5·18)	<0·0001
Dense (C or D; n=100)	75·00% (65·34 to 83·12)	67·43% (64·90 to 69·88)	69·50% (67·01 to 71·90)	7·57 (-0·73 to 15·87)	0·074	2·07 (-0·12 to 4·02)	0·038
Lesion feature§							
Soft tissue; benign (n=36)	83·33% (67·19 to 93·63)	44·25% (39·86 to 48·70)	49·21% (44·76 to 53·66)	39·09 (24·11 to 54·06)	<0·0001	4·96 (1·96 to 7·96)	0·0012
Calcification; benign (n=28)	42·86% (24·46 to 62·82)	52·30% (47·22 to 57·33)	49·74% (44·69 to 54·81)	−9·44 (−26·96 to 8·08)	0·29	−2·55 (−7·41 to 2·31)	0·30

Data are n or n (95% CI). AI=artificial intelligence. BI-RADS=Breast Imaging Reporting and Data System. *Sensitivity and specificity of AI were calculated using the cutoff threshold of 0·1 (ie, if the abnormality score of AI is greater than or equal to 0·1, then positive; otherwise, negative). †Test 1 was AI-unaided radiologist readings and test 2 was subsequent AI-aided radiologist readings. ‡Reading panel subgroup analysis compared general radiologists with breast specialists. §Normal mammograms were excluded in this subgroup analysis.

Table 3: **Sensitivity and specificity of AI and radiologists and comparisons between AI, AI-aided radiologists, and AI-unaided radiologists**

0·251 (0·177–0·324; p<0·0001) in soft tissue and 0·086 (0·024–0·148; p=0·0066) in calcification, suggesting that AI can be effective for discriminating soft tissue lesions from breast parenchyma. The overall trend of AULROC was similar to AUROC (table 2).

Overall, when aided by AI, sensitivity and specificity of the radiologists was increased (table 3). AI-unaided radiologists showed higher sensitivity in fatty breasts, but performance improvement when aided by AI was greater in dense breasts (p=0·0082 in interaction test), with sensitivities becoming comparable (table 3). When aided by AI, specificity was increased in fatty breasts, but the increase was non-significant in dense breasts (table 3). When analysing by lesion features, AI assistance improved sensitivity in calcification and in soft tissue (table 3). However, specificity in calcification was decreased, although the decrease was not significant (table 3). Specificity of standalone AI in soft tissue was 39·09 percentage points (95% CI 24·11–54·06) higher than AI-unaided radiologists; additionally, the specificity of radiologists in soft tissue was increased with the assistance of AI (table 3).

Of the 160 cancers, 142 (89%) were detected by AI with an abnormality score of at least 0·1 and 122 (76%) were detected by more than half of the reader group (table 4). With these thresholds, AI was significantly better than readers at detecting cancers with mass or distortion mammographic features (table 4). AI detected 73 (91%) of 80 T1 cancers and 104 (87%) of 119 node-negative cancers with an abnormality score of at least 0·1 whereas 59 (74%) T1 cancers and 88 (74%) node-negative cancers were detected by more than half of readers (table 4).

	Detected by AI (abnormality score ≥0·1)	Detected by more than half of readers	Detected by both	Missed by both	p value (AI *vs* readers)
All (n=160)	142 (89%)	122 (76%)	117 (73%)	13 (8%)	0·0002
Dominant imaging feature					
Mass (n=59)	53 (90%)	46 (78%)	45 (76%)	5 (8%)	0·044
Calcifications (n=81)	71 (88%)	66 (81%)	63 (78%)	7 (9%)	0·14
Distortion or asymmetry (n=20)	18 (90%)	10 (50%)	9 (45%)	1 (5%)	0·023
T stage (size)					
T0 (in-situ; n=37)	31 (84%)	28 (76%)	26 (70%)	4 (11%)	0·27
T1 (≤20 mm; n=80)	73 (91%)	59 (74%)	57 (71%)	5 (6%)	0·0039
T2 (tumour >20 mm but ≤50 mm; n=27)	25 (93%)	23 (85%)	22 (81%)	1 (4%)	0·34
Unknown (n=16)	13 (81%)	12 (75%)	12 (75%)	3 (19%)	0·30
N stage (lymph node)					
Negative (n=119)	104 (87%)	88 (74%)	84 (71%)	11 (9%)	0·0025
Positive (n=24)	23 (96%)	20 (83%)	20 (83%)	1 (4%)	0·064
Unknown (n=17)	15 (88%)	14 (82%)	13 (76%)	1 (6%)	0·57
Cancer subtype					
Luminal A (n=35)	31 (89%)	30 (86%)	29 (83%)	3 (9%)	0·57
Luminal B (n=76)	67 (88%)	53 (70%)	51 (67%)	7 (9%)	0·0038
HER2 positive (n=10)	8 (80%)	7 (70%)	7 (70%)	2 (20%)	0·29
Triple negative (n=24)	23 (96%)	18 (75%)	18 (75%)	1 (4%)	0·012
Unknown (n=15)	13 (87%)	13 (87%)	12 (80%)	0	0·56

Data are n (%). Table shows cancers detected or missed by more than half of readers (>7) and in which the abnormality score of AI is greater than or equal to the predefined threshold of 0·1. AI=artificial intelligence.

Table 4: **Mammographic and pathologic features of breast cancer detected or missed by AI and radiologists**

Discussion

In this study, we have shown that an AI algorithm for detecting breast cancer can be used as an effective diagnostic support tool for radiologists in mammography interpretation. It showed 0·938–0·970 of AUROC on multiple validation datasets collected from five institutions in South Korea, the USA, and the UK. It also showed significantly better performance than 14 radiologists in 320 independent mammograms, resulting in a significant improvement in radiologists' AI-aided diagnostic performance.

Breast cancer has heterogeneous appearances, ranging from obvious masses with spiculated margins to subtle asymmetry or faint microcalcification, leading to difficulties in accurate diagnosis and consistent interpretation of mammography. Deep learning is known to be superior to traditional machine learning algorithms for various recognition tasks. Rich feature representations directly learned from large-scale data are not limited by human-designed features, which could allow recognition of various cancer-specific radiological appearances accurately. Inter-reader performance variation is another problem in screening mammography.[5,6,27] For example, sensitivity in breast cancer detection has been shown to vary from 74·5% to 92·3%.[27] Software is robust to human variation, so deep learning might contribute to reducing the variability in radiologists' diagnostic performance. Although feasibility has to be shown in prospective clinical trials, AI is expected to help breast cancer screening in mammography by increasing cancer detection and decreasing false-positive recalls.

In this study, more than 30 000 pathologically proven cancer-positive mammograms—the largest scale of cancer data among mammography-related AI studies[28–31]—were collected from various institutions in different countries. Our experiments showed that multinational large-scale data—and especially the scale of the cancer data—are important for robustness of AI. In terms of data quality, 87% of cancer and 33% of benign mammograms were annotated at pixel level by radiologists. All of the cancer data and a portion of benign data were pathologically proven cases, so the algorithm could be trained to discriminate the subtle difference between benign tumours and malignancy. We restricted our cancer data to one mammogram per each patient with cancer, meaning 36 468 cancer-positive mammograms were obtained from 36 468 patients. Thanks to the high-quality multinational large-scale data, our AI algorithm consistently showed excellent performance in various validation datasets.

Our AI algorithm was observed to have the following characteristics. First, AI showed superior performance in

breast cancer detection. The reader study showed that AI detects more cancers with mammographic features of mass, architectural distortion, and asymmetry than radiologists. Traditional CAD is known to have poor performance in detection of cancers with distortion or asymmetry,[32] which suggests that AI can lead to significant improvement in diagnostic performance of radiologists by overcoming the problems of traditional CAD. Second, AI showed better performance than radiologists in detection of early-stage invasive cancers. Of 21 T1 cancers and 31 node-negative cancers missed by the reader group, 16 and 20, respectively, were detected by AI. Although the real clinical value needs to be confirmed by prospective studies, these results suggest that early detection of breast cancers by AI might contribute to a reduction of interval cancer and improvement of outcomes for patients with breast cancer. Lastly, the diagnostic performance of AI was less affected by breast density than was the performance of radiologists. Radiologists' performance can decrease with dense breasts, since dense parenchymal tissue is more likely to mask cancer lesions in mammograms.[33] The sensitivity difference of AI between fatty and dense breasts was much smaller than that of radiologists, leading to a significant improvement of radiologists' AI-aided performance in dense breasts.

A similar study on AI for breast cancer screening has recently been published,[34] in which an AI algorithm was shown to be superior to radiologists in terms of interpretive performance. However, the reader study employed in that study was limited in terms of clinical implication, as the effect of the AI algorithm on radiologists' interpretive performance was not directly evaluated. The relative strength of the AI algorithm investigated in our study includes higher calibre data used for training, both in terms of quantity (ie, 30 000 cancer cases with 87% annotated by breast specialists) and quality (ie, data from five institutions across both white and Asian populations compared with data from two institutions representing only white participants).

This study has several limitations. First, the reader study was done with a cancer-enriched dataset, which has different cancer prevalence to real-field data (50% in the reader study dataset *vs* <1% in real-field data). Medical AI has been increasingly studied as deep learning technology becomes mainstream, but most studies have methodological deficiencies[15,16]—eg, in a systematic review of deep learning performance in medical imaging, only four of 82 studies considered diagnostic performance in an algorithm-plus-clinician scenario.[15] Although our study compared the diagnostic performance of humans, AI-aided humans, and standalone AI using an independent set of external data in a strict reader study format, the real clinical value needs to be investigated further via prospective clinical studies with the same prevalence of the real-world clinical setting. Second, our AI algorithm does not take into account clinical factors such as family history or symptoms, which might limit comprehensive analysis. Third, the reading environment of this study was different from that of daily practice, especially in terms of the proportion of cancer cases. There was no restriction on reading time, but it was noted that reading volume affects diagnostic performance.[35] These factors might cause performance difference between clinical and experimental setting.[36] Although a radiologist's cancer detection rate is expected to be lower in daily practice than in the experimental setting, these factors should be controlled in future studies.

In conclusion, the AI algorithm we developed with large-scale high-quality data showed better diagnostic performance than radiologists in breast cancer detection from mammograms. More importantly, the diagnostic performance of radiologists was significantly improved with the assistance of AI. This result shows that AI can be used as an effective diagnostic support tool for breast cancer detection, which is worth evaluating in prospective clinical trials.

For access to the **OPTIMAM database** see https://medphys.royalsurrey.nhs.uk/omidb/getting-access

For **PyTorch** see https://pytorch.org

For the **online demo** see https://insight.lunit.io/mmg

Contributors

E-KK and H-EK conceived and designed the study. H-EK and HN developed and validated the algorithms, with the help of clinical advice from E-KK, HHK, B-KH, and KHK. E-KK, HHK, and B-KH collected and curated data for AI development. E-KK and EHL collected and curated data for the reader study. E-KK, H-EK, and KH designed the reader study protocol. KH did the statistical analysis. E-KK, H-EK, and KHK interpreted the results of the validation study. H-EK and KHK wrote the initial draft. All authors subsequently edited the report. E-KK, HHK, and B-KH supervised the project.

Declaration of interests

H-EK, KHK, and HN are employees of Lunit, the funder of the study. All other authors declare no competing interests.

Data sharing

Additional documents related to this study are available on request to the corresponding author. The datasets from Yonsei University Severance Hospital, Asan Medical Center, Samsung Medical Center, Wake Radiology, and Soonchunhyang University Hospital Bucheon were used under license for the current study and are not publicly available. Applications for access to the OPTIMAM database can be made online. The code used to train the AI model are dependent on annotation, infrastructure, and hardware, so cannot be released. However, all experimental and implementation details that can be shared are described in detail in appendix. Several major components of our work are available in the PyTorch open source repository. The AI algorithm developed from this study is available through the commercial product, Lunit INSIGHT MMG, and can be freely experienced through an online demo).

Acknowledgments

This study was funded by Lunit.

References

1 Tabar L, Vitak B, Chen HH, Yen MF, Duffy SW, Smith RA. Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. *Cancer* 2001; **91:** 1724–31.

2 Tabar L, Yen MF, Vitak B, Chen HH, Smith RA, Duffy SW. Mammography service screening and mortality in breast cancer patients: 20-year follow-up before and after introduction of screening. *Lancet* 2003; **361:** 1405–10.

3 Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018; **68:** 394–424.

4 Majid AS, de Paredes ES, Doherty RD, Sharma NR, Salvador X. Missed breast carcinoma: pitfalls and pearls. *Radiographics* 2003; **23:** 881–95.

5 Lehman CD, Arao RF, Sprague BL, et al. National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. *Radiology* 2017; **283:** 49–58.

6 Miglioretti DL, Gard CC, Carney PA, et al. When radiologists perform best: the learning curve in screening mammogram interpretation. *Radiology* 2009; **253:** 632–40.

7 Giger ML, Chan HP, Boone J. Anniversary paper: history and status of CAD and quantitative image analysis: the role of medical physics and AAPM. *Med Phys* 2008; **35:** 5799–820.

8 Warren Burhenne LJ, Wood SA, D'Orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology* 2000; **215:** 554–62.

9 Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology* 2001; **220:** 781–86.

10 Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med* 2007; **356:** 1399–409.

11 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:** 1828–37.

12 Fenton JJ, Abraham L, Taplin SH, et al. Effectiveness of computer-aided detection in community mammography practice. *J Natl Cancer Inst* 2011; **103:** 1152–61.

13 Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017; **318:** 2199–210.

14 Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; **316:** 2402–10.

15 Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digital Health* 2019; **1:** e271–97.

16 Houssami N, Kirkpatrick-Jones G, Noguchi N, Lee CI. Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. *Expert Rev Med Devices* 2019; **16:** 351–62.

17 Nikitin V, Filatov A, Bagotskaya N, Kil I, Lossev I, Losseva N. Improvement in ROC curves of readers with next generation of mammography CAD. C-2315. ECR 2014. DOI:10.1594/ecr2014/C-2315.

18 Sickles E, D'Orsi CJ, Bassett LW, et al. ACR BI-RADS mammography. In: ACR BI-RADS Atlas, 5th edn. Reston, VA: American College of Radiology, 2013.

19 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proc CVPR IEEE* 2016; **1:** 770–78.

20 Nam H, Kim H-E. Batch-instance normalization for adaptively style-invariant neural networks. *Adv Neur In* 2018; **1:** 2563–72.

21 Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. *Proc IEEE I Conf Comp Vis* 2015; **1:** 1520–28.

22 Hillis SL, Berbaum KS, Metz CE. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. *Acad Radiol* 2008; **15:** 647–61.

23 Obuchowski NA, Bullen JA. Statistical considerations for testing an AI algorithm used for prescreening lung CT images. *Contemp Clin Trials Commun* 2019; **16:** 100434.

24 Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med* 2005; **353:** 1773–83.

25 Pepe MS. Chapter 4.5: the ROC for ordinal tests. In: The statistical evaluation of medical tests for classification and prediction. New York, NY: Oxford University Press, 2003: 85–92.

26 Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. *Med Phys* 1996; **23:** 1709–25.

27 Elmore JG, Jackson SL, Abraham L, et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology* 2009; **253:** 641–51.

28 Rodriguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. *Radiology* 2019; **290:** 305–14.

29 Akselrod-Ballin A, Chorev M, Shoshan Y, et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology* 2019; **292:** 331–42.

30 Wu N, Phang J, Park J, et al. Deep neural networks improve radiologists' performance in breast cancer screening. *arXiv* 2019; published online March 20. https://arxiv.org/abs/1903.08297 (preprint).

31 Rodriguez-Ruiz A, Lang K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst* 2019; **111:** 916–22.

32 Baker JA, Rosen EL, Lo JY, Gimenez EI, Walsh R, Soo MS. Computer-aided detection (CAD) in screening mammography: sensitivity of commercial CAD systems for detecting architectural distortion. *AJR Am J Roentgenol* 2003; **181:** 1083–88.

33 Freer PE. Mammographic breast density: impact on breast cancer risk and implications for screening. *Radiographics* 2015; **35:** 302–15.

34 McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020; **577:** 89–94.

35 Haneuse S, Buist DS, Miglioretti DL, et al. Mammographic interpretive volume and diagnostic mammogram interpretation performance in community practice. *Radiology* 2012; **262:** 69–79.

36 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:** 47–53.