Improved breast lesion detection in mammogram images using a deep neural network

Wen Zhou
Xiaodong Zhang
Jia Ding
Lingbo Deng
Guanxun Cheng
Xiaoying Wang

PURPOSE

This study aimed to investigate the effect of using a deep neural network (DNN) in breast cancer (BC) detection.

METHODS

In this retrospective study, a DNN-based model was constructed from a total of 880 mammograms that 220 patients underwent between April and June 2020. The mammograms were reviewed by two senior and two junior radiologists with and without the aid of the DNN model. The performance of the network was assessed by comparing the area under the curve (AUC) and receiver operating characteristic curves for the detection of four features of malignancy (masses, calcifications, asymmetries, and architectural distortions), with and without the aid of the DNN model and by the senior and junior radiologists. Additionally, the effect of utilizing the DNN on diagnosis time for both the senior and junior radiologists was evaluated.

RESULTS

The AUCs of the model for the detection of mass and calcification were 0.877 and 0.937, respectively. In the senior radiologist group, the AUC values for evaluation of mass, calcification, and asymmetric compaction were significantly higher with the DNN model than those obtained without the model. Similar effects were observed in the junior radiologist group, but the increase in the AUC values was even more dramatic. The median mammogram assessment time of the junior and senior radiologists was 572 (357–951) s, and 273.5 (129–469) s, respectively, with the DNN model, and the corresponding assessment time without the model, was 739 (445–1003) s and 321 (195–491) s, respectively.

CONCLUSION

The DNN model exhibited high accuracy in detecting the four named features of BC and effectively shortened the review time by both senior and junior radiologists.

KEYWORDS

Breast cancer, deep neural network, mammography

From the Department of Radiology (W.Z., X.Z., X.W. ✉ wangxiaoying@bjmu.edu.cn), Peking University First Hospital, Beijing, China; Department of Radiology (W.Z., L.D., G.C. ✉ chengguanxun@outlook.com) Peking University Shenzhen Hospital, Shenzhen, China; (J.D.), Beijing Yizhun Medical AI Co., Ltd, Beijing, China.

Breast cancer (BC) is the most common cancer and the second leading cause of cancer deaths in women worldwide,[1] but there is a large difference in the survival rate of BC patients who live in different countries. In particular, the five-year survival rate of BC patients in China is much lower than that in developed countries, such as the United States.[2] One of the main reasons for this discrepancy is the low early diagnosis rate in China.[3] Therefore, the accurate and early diagnosis of BC is critical for early treatment options and for reducing BC mortality in China.

Mammography is the most effective screening method for BC and has been shown to increase the detection rate and reduce the mortality rate of BC.[4-6] Mammography images can clearly show the tissues and glands of the breast as well as the surrounding areas through non-invasive methods. Such images facilitate the identification of lumps, burrs, slight calci-

fications, and cancer spread and metastasis in the breast. Notably, mammography approaches have advantages over similar imaging techniques, such as ultrasound, in detecting microcalcifications.[7] However, it is very difficult to locate and characterize a lesion, and the consistency of doing so across doctors is very poor.[8,9] Lehman et al.[10] reported that the average sensitivity and specificity of reviewing mammography images were 86.9% and 88.9%, respectively. In addition, the false positive and false negative rates of mammography assessment are approximately 7%–12% and 4%–34%, respectively.[11,12] Nevertheless, mammography remains the gold standard for the detection of malignancy, with its high resolution enabling the detection of masses, microcalcifications, asymmetries, and architectural distortions. For the detection of microcalcifications, in particular, mammography has distinct advantages over ultrasound. To detect malignancy, radiologists have to review a large number of images, particularly with digital breast tomosynthesis, which impacts interpretation time. Additionally, because the detection of malignancy depends on factors such as breast density, identifying and accurately localizing a lesion can differ from one physician to another.

Several machine learning algorithms have been applied to the research of mammography data in recent years. In 2014, Wang et al.[13] proposed a breast tumor detection algorithm based on the extreme learning machine, which performs breast tumor edge segmentation for the microscopic detection of a tumor. Similarly, Agrawal et al.[14] used a support vector machine to perform feature extraction on the segmented region in the mammogram X-ray image and then target detection, which effectively segmented the tumor mass region within the normal chest parenchyma. Deep learning approaches used in medical imaging fields leverage the use of more sophisticated algorithms and image processing technology to assess samples with a more refined decomposition of tissue properties. The continuing maturity of deep learning technology can help doctors perform more accurate localization and diagnosis of pathological tissues. These algorithms were found to decrease interpretation time, which facilitates more rapid treatment.

In recent years, many scholars have applied deep learning algorithms to medical image recognition problems.[15,16] Bayramoglu et al.[17] proposed two different architectures based on a convolutional neural network to predict malignant breast tumors. Zhang et al.[18] constructed a two-layer deep learning architecture to automatically extract imaging features for classification, and their model performed well in terms of classification accuracy, sensitivity, and specificity. Mohamed et al.[19] built and trained a convolutional neural network model based on mammography images to accurately and rapidly classify breast density to clarify the risk of BC, and the area under the curve (AUC) of the model classification reached 0.992.

However, current deep learning approaches in BC research are mostly based on pathological images or algorithm optimization techniques that aim to better segment images. Therefore, it is necessary to establish a reliable model for assessing BC in mammography images that is comparable to a radiologist's assessment. This study investigates the effect of a deep neural network (DNN) on BC detection in clinical practice.

Methods

Study design

The study was approved by the research ethics review board of Peking University (approval number: 2020-011), and informed consent was waived because it was a retrospective study. Mammography images acquired consecutively between April 2020 and June 2020 at a single institution were analyzed, and all of them were anonymized. The exclusion criteria included cases with prior benign and malignant breast surgery, breast reduction, breast augmentation, chemotherapy, radiation therapy, or unknown results from prior biopsies. All mammography analyses were performed by two radiologists experienced in assessing breast mammography images. The lesions were divided into four categories according to the corresponding mammograms, magnetic resonance imaging (MRI), and pathological results. True-positive/negative and false-positive/negative cases were identified by a positive/negative result of the radiologist assessment and confirmation or negation based on MRI and/or pathological evaluation, respectively. Four mammogram images were acquired for each patient and included two images in a mediolateral oblique projection (MLO) and two images in a cranial-caudal projection (CC).

Development of a DNN model

Faster R-CNN was employed as the deep learning framework for model detection, and ResNet50 was used for feature extraction. The feature pyramid network was used to construct new features based on data augmentation techniques. Features were fused in different convolutional layers of the ResNet, ensuring that the model incorporated multi-scale information to improve the ability to detect small lesions. The lesion detection network is shown in Figure 1.

Image resizing for uniform resolution: The size of the input image was converted to a pixel size of 0.15 mm x 0.15 mm. Random cropping was used for data expansion at a rate of 0.8-1.2 times the size of the original image. Images were also randomly flipped horizontally. The model training used four NVIDIA TITAN RTX P8 graphics cards with a configuration of 28 GB video memory and a batch size of four images.

Algorithm optimization: The momentum stochastic gradient descent learning rate was 0.005. The learning rate was adjusted according to the number of iterations using the learning rate scheduler method for learning rate decay. The L2-norm regularization parameter weight decay was 0.0001. The maximum number of iterations was set to 25,000, and the number of warm-start iterations was 500. In the test phase, horizontal and vertical flips were used to expand the data.

Image gray-level normalization: If the original gray value was not compatible with the algorithm for subsequent prediction, the grayscale of the image was normalized to ensure consistency in the gray value range across different images. The grayscale of the segmented region was recorded, and the gray level was linearly mapped according to the statistical results. This procedure was performed so that 90% of the gray value pixels were in the range of 0–1, 5% of the gray value pixels were <0, and 5% of the gray value pixels were >1.

Breast segmentation: The background of the breast mammogram images was re-

Main points

- The use of a deep neural network (DNN) improved breast cancer detection.

- With the help of the DNN, radiologists could more accurately detect tumor mass, calcification, and asymmetric compaction.

- An auxiliary effect of the deep learning model on doctors of different seniority was that it increased the detection accuracy of inexperienced doctors.

- The deep learning model shortened the average mammogram assessment time for both junior and senior radiologists.

moved, and only the breast was retained. The grayscale distribution histogram of the image was recorded, and the threshold value was obtained using the triangle method. The image pixels with a gray level higher than the threshold were segmented as the breast. The minimum rectangular range that contained the breast was then taken as the input of the subsequent module.

Quadrant and depth analysis of the lesion: The relationship between the images in the MLO and CC position was judged according to multiple features. After the detection stage, the location, size, type, and probability of a lesion(s) were obtained. Then, more features of the lesion were analyzed to match lesions more accurately. These features included the quadrant of the lesion and the distance between the lesion and the nipple.

Lesion quadrant division: A mask was used to indicate the location of the lesions, and the classification network was used to classify the MLO lesions. The lesions in the MLO position were divided into five regions: upper, middle, lower, axillary tail, and areola. The lesions in the CC position were divided into four quadrants: outer, middle, inner, and areola.

Lesion depth regression: The whole mammogram image and the mask of the target lesion on the image were spliced together as two channels of the image. The distance from the lesion to the nipple was obtained using the regression network. Distance 0 represented the nipple, and distance 1 represented the pectoralis major muscle.

Focus matching: The lesion features in the CC and MLO positions were combined to predict the probability of two lesions being the same lesion using the GBTD method. To construct a matching probability matrix, each element on the matrix represented the matching probability of the two lesions. The matching relationship between MLO and CC lesions was obtained according to a greedy algorithm. The remaining lesions without matching or with a matching probability that was too small were considered as a single lesion, and no matching relationship was given.

The classification of benign and malignant lesions: Multi-task learning was used to predict benign and malignant lesions as well as their morphological distribution at the same time. The two tasks promote and complement each other and make the overall performance more accurate than conducting one task alone. The data from 14,811 cases were used for training, comprising 7,519 cases of labeled data and 7,292 cases of unlabeled data. Among the labeled data, the labeled regions of interest included 10,480 masses, 6,358 calcifications, 1,713 asymmetries, and 311 architectural distortions. For each lesion, the category of the lesion, such as mass or calcification, was marked, and the outline of the lesion was drawn. The backbone network of the detection model used transfer learning, and the network was trained with a large amount of ImageNet data that was transferred to the breast detection model. This has been shown to significantly improve the detection performance of the model. The percentages of the training set and the test set were 80% and 20%, respectively.

For robustness, we trained the DNN algorithm in three random 80% partitions of the training set. After the detection of lesions, multi-task learning was used to analyze the shape, edges, and other attributes of the lesions and to predict their benign or malignant nature.

The reading time and breast imaging reporting and data system (BI-RADS) score of senior and junior radiologists with and without the help of a DNN model

A total of 880 images from 220 patients were used to test the impact of the DNN model on radiologist assessment. Two senior radiologists (with 16 and 18 years of mammogram reading experience) and two junior ones (with 1 and 2 years of mammogram reading experience) reviewed all four mammogram images from each study in random order, both with and without the aid of the DNN. The second reading was performed three weeks later. For both readings, the order of images was randomized on an individual assessor basis. All the radiologists were blinded to the patient information. For each patient, the radiologists provided a BI-RADS score according to the following scale: 1 = negative; 2 = benign; 3 = probably benign; 4 = suspicious abnormality (a possibility of malignancy or cancer); and 5 = highly likely to be malignant. Reading times were measured from the opening of a new case to the validation of the lesions and the BI-RADS score. Both the reading time and BI-RADS score were recorded for later analysis.

Statistical analysis

All statistical analyses were performed using IBM SPSS 20.0 (SPSS Inc, Chicago, IL, USA) and MedCalc statistical software (ver-
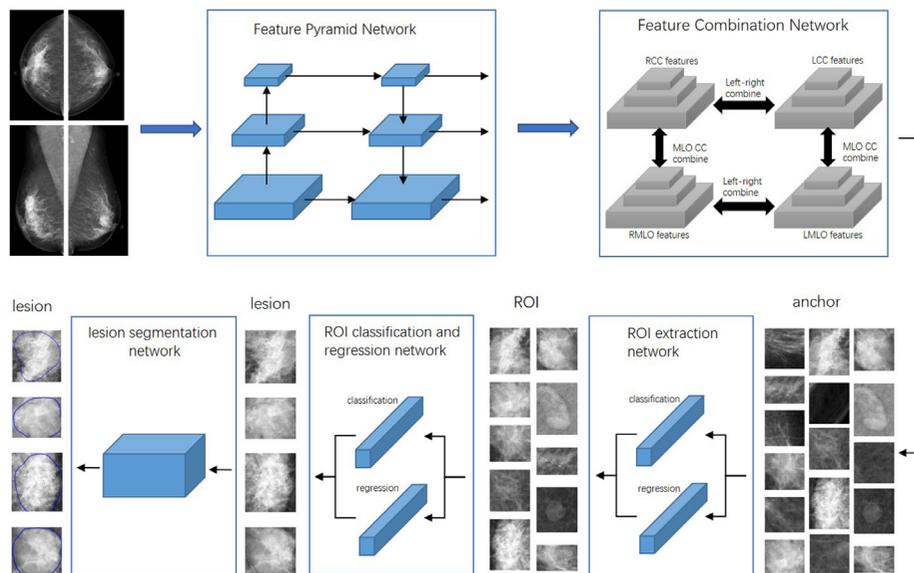

Figure 1. Lesion Detection Network. The input mammograms include two cranial-caudal projection (CC) images and two mediolateral oblique projection (MLO) images, the features of which were extracted using ResNet50. The features were then fused using a Feature Combination Network comprised of right and left MLO and CC images. New features were also constructed with multi-scale information to improve the ability to detect small lesions. Feature pyramid network: Features of the input images were extracted, and several layers of features from fine to abstract were obtained. Feature fusion network: the information from four molybdenum palladium images of the same patient was used to improve the quality of the features and improve lesion detection. This framework included a left and right feature fusion network and MLO/CC feature fusion network. Key area extraction network: a series of anchors were set with different positions, sizes, and aspect ratios using a sliding window. The key area extraction network classified and regressed each anchor to predict whether the anchor was the key area and where the key area was located. ROI, region of interest.

sion 20.026, MedCalc Software). Descriptive statistics of the data are presented with n (%) and are shown as median (min–max) for non-normalized variables. The normality test was determined using the Shapiro–Wilk test. Comparisons between the two groups were performed using the Wilcoxon signed-rank test for variables with a non-normal distribution. Receiver operating characteristic (ROC) curves and the AUC were used to evaluate the performance of the DNN model as well as the senior and junior radiologists, with and without the help of artificial intelligence (AI). The sensitivity, specificity, and Youden index of the ROC curves were calculated, and the highest Youden index was used to determine the cut-off value. Comparisons of the ROC curves were evaluated using a Delong test. The inter-rater agreement of the senior and junior radiologists in terms of the location and BI-RADS assessment was evaluated using a kappa coefficient. The kappa coefficient for the strength of the agreement was categorized as follows: −1, none; 0, poor; 0.0–0.20, slight; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; and 0.81–1: almost perfect.[20] A P value <0.05 was considered statistically significant.

Results

The ROC curves of the models for the four distinct lesion features are shown in Figure 2. The AUCs of the model for mass, calcification, asymmetric compaction, and structural distortion were 0.877 [95% confidence interval (CI), 0.843–0.906], 0.937 (95% CI, 0.910–0.958), 0.697 (95% CI, 0.652–0.740), and 0.624 (95% CI, 0.577–0.669), respectively. The sensitivity values for the detection of the same features were 76.71%, 89.73%, 73.68%, and 99.77%, respectively. Similarly, the specificity values for these four features were 98.66%, 97.68%, 65.76%, and 25.00%, respectively (Table 1).

Figure 3 displays the ROC curves of the senior and junior radiologist assessments with and without the help of the model. The corresponding AUC, specificity, and sensitivity values are listed in Table 1. For senior radiologists, the AUC values of the ROC curves for assessments based on mass with and without the help of the model were 0.926 and 0.909, respectively. Similarly, for junior radiologists, the ROC curves were 0.879 and 0.803 with and without the help of the model, respectively. Regarding calcification, the AUC values of the ROC curves for the senior radiologists were 0.955 and 0.946 with and

without the aid of the model, respectively. In addition, the AUC values of the calcification ROC curves for the junior radiologists were 0.932 and 0.898 with and without the aid of the model, respectively. The AUCs between the radiologists with and without the aid of the model were compared using a Delong test (Table 2). In general, the AUCs of the junior radiologists for mass and calcification were significantly larger with the DNN model than those without the model (both P < 0.001), but there were no significant differ-
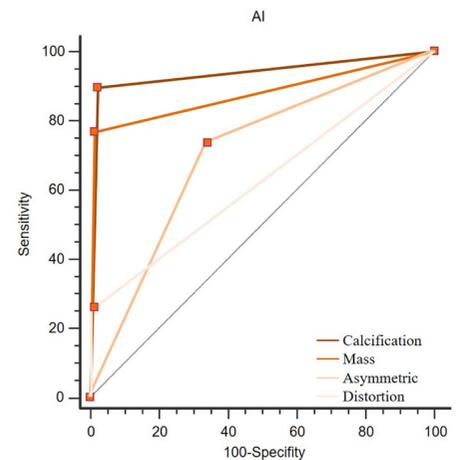

Figure 2. Receiver operating characteristic curve of the model for the detection of mass, calcification, asymmetry, and distortion. AI, artificial intelligence.

Table 1. AUC values and related parameters for the detection of the four lesion features of mass, calcification, asymmetry, and distortion by senior and junior radiologists with and without the assistance of the deep neural network model

Features		AUC	SE	P value	Sensitivity	Specificity	Cut-off
Mass	AI	0.877 (0.843 – 0.906)	0.018	<0.001	76.71	98.66	0.754
	J_noAI	0.803 (0.775 – 0.829)	0.014	<0.001	61.3	99.33	0.606
	J_AI	0.879 (0.856 – 0.900)	0.013	<0.001	77.74	98.15	0.759
	S_noAI	0.909 (0.888 – 0.927)	0.011	<0.001	82.53	99.33	0.819
	S_AI	0.926 (0.906 – 0.942)	0.010	<0.001	85.62	99.5	0.851
Calcification	AI	0.937 (0.910 – 0.958)	0.012	<0.001	89.73	97.68	0.874
	J_noAI	0.898 (0.877– 0.918)	0.011	<0.001	80.27	99.42	0.797
	J_AI	0.932 (0.913 – 0.947)	0.009	<0.001	87.3	99.03	0.863
	S_noAI	0.946 (0.929 – 0.960)	0.008	<0.001	89.73	99.42	0.892
	S_AI	0.955 (0.939 – 0.968)	0.007	<0.001	91.62	99.42	0.910
Asymmetric	AI	0.697 (0.652 – 0.740)	0.038	<0.001	73.68	65.76	0.395
	J_noAI	0.626 (0.593 – 0.658)	0.028	<0.001	34.21	91.01	0.252
	J_AI	0.661 (0.628 – 0.692)	0.030	<0.001	51.32	80.79	0.321
	S_noAI	0.782 (0.753 – 0.808)	0.028	<0.001	61.84	94.46	0.563
	S_AI	0.801 (0.774 – 0.827)	0.026	<0.001	72.37	87.93	0.603
Distortion	AI	0.624 (0.577 – 0.669)	0.065	0.058	25	99.77	0.248
	J_noAI	0.516 (0.482 – 0.549)	0.021	0.516	4.17	98.96	0.031
	J_AI	0.539 (0.505 – 0.572)	0.029	0.179	8.33	99.42	0.078
	S_noAI	0.679 (0.647 – 0.709)	0.051	0.000	37.5	98.26	0.358
	S_AI	0.644 (0.611 – 0.675)	0.047	0.003	29.17	99.54	0.287

AUC, area under the curve; SE, standard error; J_noAI, junior radiologists unaided; J_AI, junior radiologists aided; S_noAI, senior radiologists unaided; S_AI, senior radiologists aided. AI, artificial intelligence.

ences for the senior radiologists ($P = 0.081$ for mass and $P = 0.061$ for calcification). However, the AUCs for asymmetric compaction and structural distortion showed no difference between the radiologists with or without the aid of the model (for asymmetric compaction, $P = 0.244$ for junior radiologists and $P = 0.475$ for senior radiologists; for structural distortion, $P = 0.527$ for junior radiologists and $P = 0.554$ for senior radiologists). On the other hand, the AUCs of the senior radiologist assessments for mass, calcification, asymmetry, and distortion were significantly larger than those of the junior radiologist assessments with ($P < 0.001$, $P = 0.003$, $P < 0.001$, and $P = 0.044$) and without the aid of the model (all $P < 0.001$) (Table 2).

The review times of the radiologists in the aided and unaided scenarios were compared using a Wilcoxon signed-rank test (Table 3). The median reading times of the senior and junior radiologists unaided were 321 (195–491) s and 739 (445–1003) s, respectively. With the help of the model, the median reading times of the senior and junior radiologists fell to 273.5 (129–469) s and 572 (357–951) s, respectively, representing a reduction of 41.9 s (13.6%) for the senior radiologists and 153.5 s (20.5%) for the junior radiologists. The median review times of the senior and junior radiologists were both significantly short-er with the DNN model than those without the model (both $P < 0.001$) (Table 3). Figure 4 shows an example of using AI to help detect linear pleomorphic calcifications in the upper left outer quadrant, which suggests a BI-RADS score of 4C. 4C means high suspicion for malignancy (>50% to <95% likelihood of malignancy).

The inter-rater agreement of the senior and junior radiologists in terms of tumor mass, calcification, asymmetry, and distortion assessment was evaluated using the kappa coefficient. As shown in Table 4, for junior radiologists, the kappa coefficients of mass assessment were 0.836 and 0.676 with and without the help of DNN, respectively, and those of calcification assessment were 0.913 and 0.839 with and without the help of DNN, respectively. These values indicate that the reliability of the junior radiologist assessments regarding mass and calcification can be improved with the help of the DNN model.

Discussion

In the current study, a DNN model was built and found to be helpful in the detection of masses, calcifications, asymmetries, and architectural distortions representing BC. The model was able to significantly shorten the review time of mammogram images by both senior and junior radiologists. Typically, radiologists analyze multiple mammographic images of the same patient, which is time- and energy-consuming. The DNN model proposed in the current work is very promising for clinical application and may be used to help radiologists more efficiently review mammography images, enhancing the accuracy of their diagnosis with the ultimate goal of improving the prognosis of BC. These advantages of the model were further exemplified by the fact that mammogram reading time decreased for both senior and junior radiologists when using AI.

Previous research has used different deep-learning methods to detect BC and has demonstrated a gradual performance improvement.[21-23] The Dialogue for Reverse Engineering Assessments and Methods challenge has tested a large number of mammograms and obtained an AUC of 0.87 with a sensitivity and specificity of 0.81 and 0.8, respectively.[24] Another study focused on categories of breast lesions according to BI-RADS scores using a deep convolutional neural network to analyze mammograms.[25] The sensitivity of this model for the detection of mass, calcification, asymmetry, and compaction was higher than 74% for each feature and is comparable with the Breast Cancer

Table 2. AUC values for the detection of the four lesion features of mass, calcification, asymmetry, and distortion by senior and junior radiologists with and without the assistance of the deep neural network model

Features	J_noAI vs. J_AI	S_AI vs. J_AI	S_noAI vs. J_noAI	S_noAI vs. S_AI
Mass	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P = 0.081$
Calcification	$P < 0.001$	$P = 0.003$	$P < 0.001$	$P = 0.061$
Asymmetry	$P = 0.244$	$P < 0.001$	$P < 0.001$	$P = 0.475$
Distortion	$P = 0.527$	$P = 0.044$	$P < 0.001$	$P = 0.554$

AUC, area under the curve; J_noAI, junior radiologists unaided; J_AI, junior radiologists aided; S_noAI, senior radiologists unaided; S_AI, senior radiologists aided.

Table 3. Junior and senior radiologist review times for unaided and artificial intelligence-aided cases using the Wilcoxon signed-rank test

Group		Review time (seconds)		
		Unaided	Aided	P value
Junior radiologists	Median (min–max)	739 (445–1003)	572 (357–951)	<0.001
Senior radiologists	Median (min–max)	321 (195–491)	273.5 (129–469)	<0.001

Table 4. The consistency between senior and junior radiologists in recognizing mass, calcification, asymmetry, and distortion using the kappa coefficient

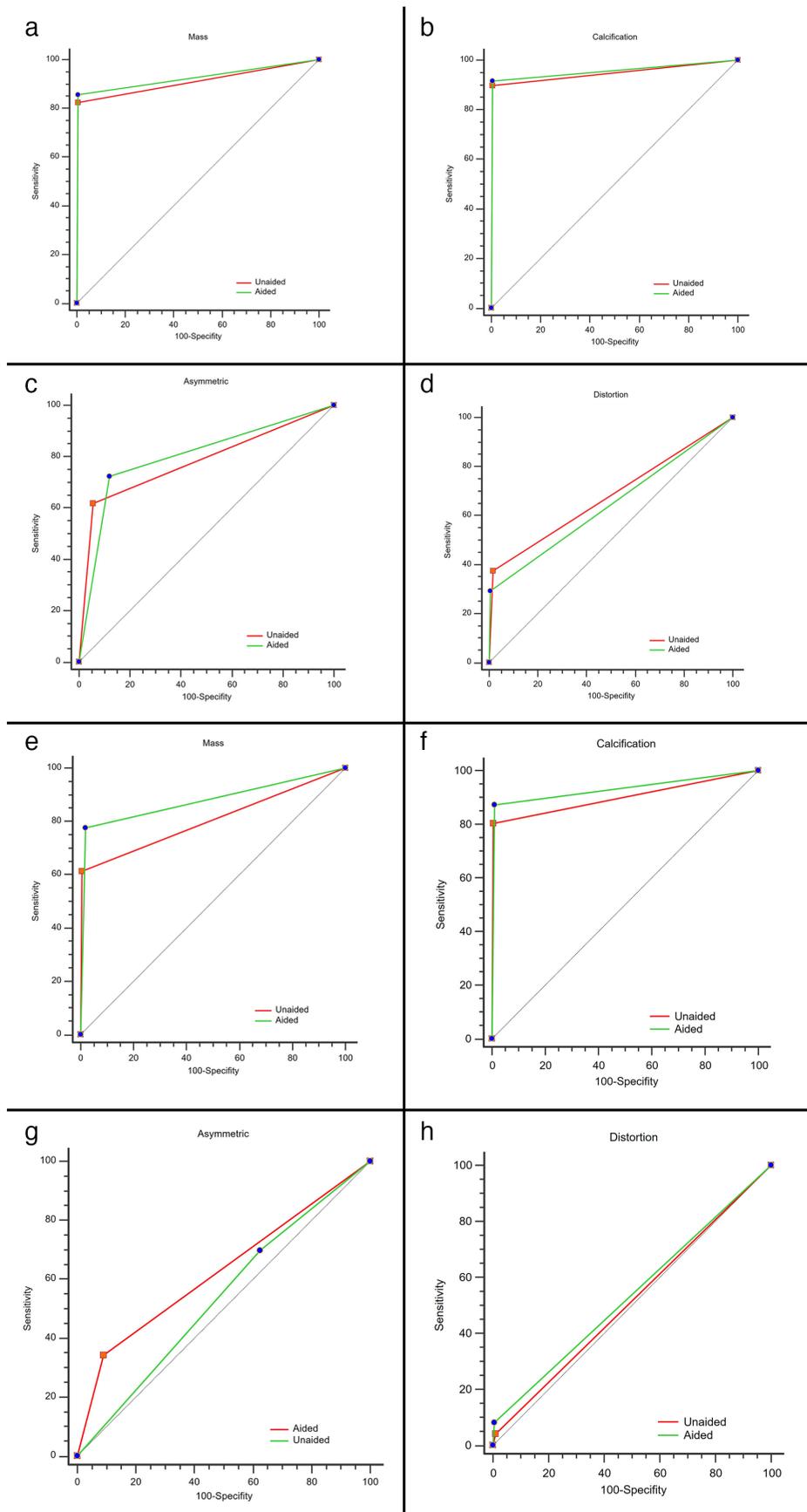|	Senior (without AI)		Senior (with AI)		Junior (without AI)		Junior (with AI)	
	Kappa	P value	Kappa	P value	Kappa	P value	Kappa	P value
Mass	0.861	<0.001	0.900	<0.001	0.676	<0.001	0.836	<0.001
Calcification	0.938	<0.001	0.981	<0.001	0.839	<0.001	0.913	<0.001
Asymmetric	0.495	<0.001	0.343	<0.001	0.298	<0.001	0.303	<0.001
Distortion	0.383	<0.001	0.156	0.001	-0.011	0.814	0.244	<0.001

AI, artificial intelligence.

Figure 3. (a-h) Receiver operating characteristic curves of the diagnosis based on mass **(a, e)**, calcification **(b, f)**, asymmetry **(c, g)**, and structural distortion **(d, h)** by senior and junior radiologists, with or without the help of artificial intelligence.

Surveillance Consortium benchmark used in another study that exhibited a sensitivity of 75%.[10]

In this study, with the assistance of the DNN model, the radiologists were able to recognize features such as mass, calcification, asymmetry, and distortion with high sensitivity and specificity. Akselrod-Ballin et al.[26] reported that a deep learning model was able to detect 48% of the false-negative findings missed by radiologists and confirmed by surgical pathology with a sensitivity of 87%. According to the results from the present study, the time that senior and junior radiologists spent on diagnosis was significantly reduced with the use of the model, especially for junior radiologists. Therefore, it was established that a deep learning algorithm trained by experts in the field was able to better assist less experienced radiologists who are at particular risk of making diagnostic errors.

The analysis described in this paper showed that the proposed model can be used to assist junior radiologists and help improve their performance in identifying lesions when reviewing mammograms. Additionally, this study showed that when junior radiologists were provided with the assistance of the trained model, their ability to detect breast lesions significantly improved, thus diminishing diagnostic errors and improving efficiency.

There are several limitations to this study that should be noted. First, the DNN model used in the current work was verified in a dataset acquired from the same mammography vendor and manufacturer. Furthermore, the patients from whom the data were collected were all from the Peking University Shenzhen Hospital. Therefore, the results presented here must be validated using images from different vendors and populations. Second, only mammograms were analyzed in this study to improve lesion detection, and clinical information was not utilized. In clinical practice, radiologists usually review a patient's clinical history and symptoms before making a diagnosis based on mammographic data. Therefore, it would be useful to analyze whether the use of clinical information and imaging data together could help radiologists make even more accurate diagnoses. In addition, the number of patients in this study with asymmetry and structural distortion was small, which may have affected the results to a certain extent.

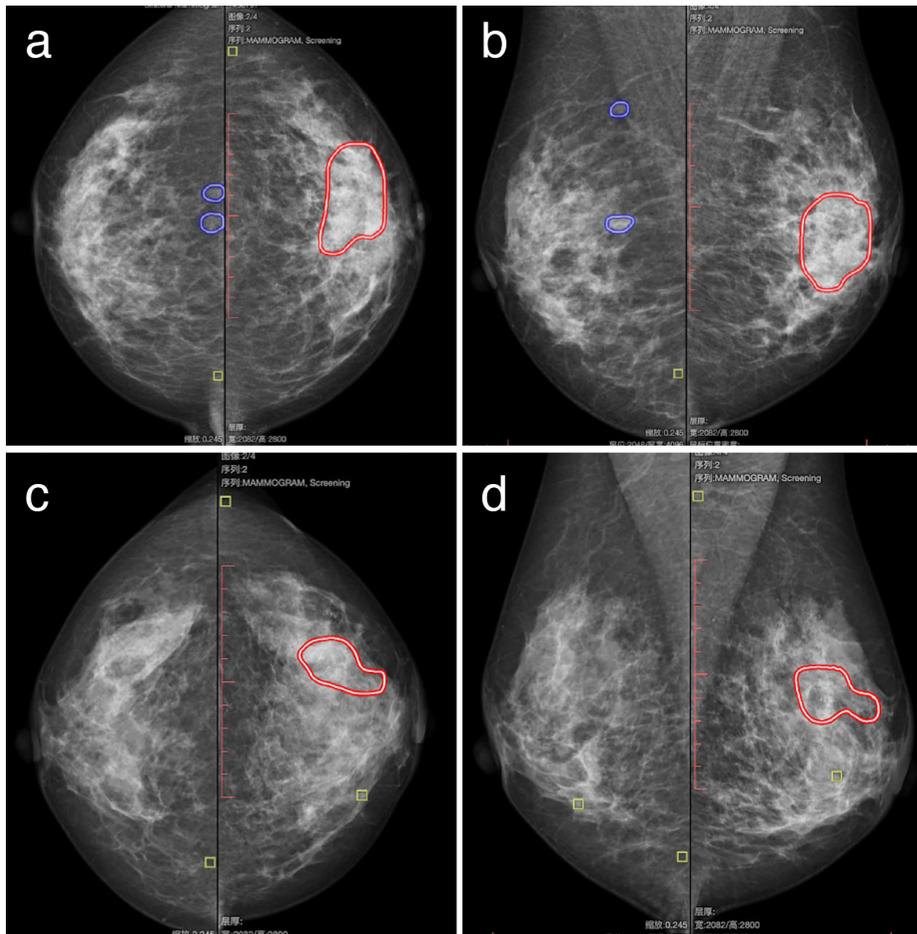In conclusion, a DNN model was developed and validated using a dataset of mam-

Figure 4. (a-d) Panels **(a, b)** show mammograms of a 48-year-old patient. The top-left panel **(a)** shows the craniocaudal view, and the top-right panel **(b)** shows the mediolateral oblique view of the breast lesions. An irregular high-density mass with calcification is seen in the upper outer quadrant of the left breast with a cluster of pleomorphic calcifications, and the assessment using artificial intelligence (AI) analysis is a breast imaging reporting and data system (BI-RADS) 5 (red circle). Panels **(c, d)** show mammograms of a 51-year-old patient. The bottom-left panel **(c)** shows the craniocaudal view, and the bottom-right panel **(d)** shows the mediolateral oblique view of the breast lesions. Fine linear pleomorphic calcification can be seen in the upper outer quadrant of the left breast, and the assessment using AI analysis is a BI-RADS 4C (red circle).

mograms to improve the detection of BC by radiologists. The model was especially successful at detecting the mass, calcification, asymmetric compaction, and structural distortion of BC lesions. With the assistance of the model, both senior and junior radiologists were able to recognize a lesion within a shorter review time.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. 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(6):394-424. [CrossRef]

2. Feuer EJ, Rabin BA, Zou Z, et al. The surveillance, epidemiology, and end results cancer survival calculator SEER*CSC: validation in a managed care setting. *J Natl Cancer Inst Monogr*. 2014;2014(49):265-274. [CrossRef]

3. Feng RM, Zong YN, Cao SM, Xu RH. Current cancer situation in China: good or bad news from the 2018 Global Cancer Statistics? *Cancer Commun (Lond)*. 2019;39(1):22. [CrossRef]

4. Hersch J, Jansen J, McCaffery K. Decision-making about mammographic screening: pursuing informed choice. *Climacteric*. 2018;21(3):209-213. [CrossRef]

5. Duffy SW, Vulkan D, Cuckle H, et al. Effect of mammographic screening from age 40 years on breast cancer mortality (UK Age trial): final results of a randomised, controlled trial. *Lancet Oncol*. 2020;21(9):1165-1172. [CrossRef]

6. Carles M, Martínez-Alonso M, Pons A, et al. The effect of information about the benefits and harms of mammography on women's decision-making: study protocol for a randomized controlled trial. *Trials*. 2017;18(1):426. [CrossRef]

7. Zhang K, Lu Q, Hua J, Xu J, Wu G. Positive predictive value of mammographic lymphography in the evaluation of patients with breast cancer: a preliminary study. *Acad Radiol*. 2016;23(10):1278-1282. [CrossRef]

8. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *N Engl J Med*. 1994;331(229):1493-1499. [CrossRef]

9. Svahn TM, Macaskill P, Houssami N. Radiologists' interpretive efficiency and variability in true- and false-positive detection when screen-reading with tomosynthesis (3D-mammography) relative to standard mammography in population screening. *Breast*. 2015;24(6):687-693. [CrossRef]

10. 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(1):49-58. [CrossRef]

11. Nelson HD, Fu R, Cantor A, Pappas M, Daeges M, Humphrey L. Effectiveness of breast cancer screening: systematic review and meta-analysis to update the 2009 U.S. preventive services task force recommendation. *Ann Intern Med*. 2016;164:244-255. [CrossRef]

12. Huynh PT, Jarolimek AM, Daye S. The false-negative mammogram. *Radiographics*. 1998;18:1137-1154. [CrossRef]

13. Wang Z, Yu G, Kang Y, Zhao Y, Qu Q. Breast tumor detection in digital mammography based on extreme learning machine. *Neurocomputing*. 2014;128:175-184. [CrossRef]

14. Agrawal P, Vatsa M, Singh R. Saliency based mass detection from screening mammograms. *Signal Processing*. 2014;99:29-27. [CrossRef]

15. McBee MP, Awan OA, Colucci AT, et al. Deep Learning in radiology. *Acad Radiol*. 2018; 25:1472-1480. [CrossRef]

16. Ueda D, Shimazaki A, Miki Y. Technical and clinical overview of deep learning in radiology. *Jpn J Radiol*. 2019;37(1):15-33. [CrossRef]

17. Bayramoglu N, Kannala J, Heikkila J. Deep learning for magnification independent breast cancer histopathology image classification. 2016 23rd ed. International Conference on *Pattern Recognition (ICPR);* 2016. [CrossRef]

18. Zhang Q, Xiao Y, Dai W, et al. Deep learning based classification of breast tumors with shear-wave elastography. *Ultrasonics*. 2016;72:150-157. [CrossRef]

19. Mohamed AA, Berg WA, Peng H, Luo Y, Jankowitz RC, Wu S. A deep learning method for classifying mammographic breast density categories. *Med Phys*. 2018;45:314-321. [CrossRef]

20. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med.* 2005;37:360-363. [CrossRef]

21. Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal.* 2017;35:303-312. [CrossRef]

22. Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Guevara Lopez MA. Representation learning for mammography mass lesion classification with convolutional neural networks. *Comput Methods Programs Biomed.* 2016;127:248-257. [CrossRef]

23. Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Invest Radiol.* 2017;52:434-440. [CrossRef]

24. D.R.E.A.M. The Digital Mammography DREAM Challenge. https://www.synapse.org/#!Synapse:syn4224222/wiki/401743

25. Geras KJ, Wolfson S, Shen Y, et al. High-resolution breast cancer screening with multi-view deep convolutional neural networks. 2017.

26. Akselrod-Ballin A, Chorev M, Shoshan Y, et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology.* 2019;292:331-342. [CrossRef]