Deep Learning Model Improves Radiologists' Performance in Detection and Classification of Breast Lesions

Ying-Shi Sun (sys27@163.com)
Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Department of Radiology, Peking University Cancer Hospital & Institute

Yu-Hong Qu
Peking University Cancer Hospital: Beijing Cancer Hospital

Dong Wang
Peking University

Yi Li
Shunyi maternal and children's hospital

Lin Ye
beijing chaoyang maternal and children's hospital

Jing-Bo Du
Beijing Daxing District People's Hospital: Capital Medical University Daxing Teaching Hospital

Bing Xu
shunyi district hospital

Bao-Qing Li
beijing shijingshan hospital

Xiao-Ting Li
Peking University Cancer Hospital: Beijing Cancer Hospital

Kexin Zhang
Peking University

Yan-Jie Shi
Peking University Cancer Hospital: Beijing Cancer Hospital

Rui-Jia Sun
Peking University Cancer Hospital: Beijing Cancer Hospital

Yichuan Wang
Peking University

Rong Long
Peking University Cancer Hospital: Beijing Cancer Hospital

Dengbo Chen
Peking University Cancer Hospital: Beijing Cancer Hospital

Hai-Jiao Li
Peking University Cancer Hospital: Beijing Cancer Hospital
Liwei Wang
Peking University
Min Cao
Peking University Cancer Hospital: Beijing Cancer Hospital
Yan-Yu Yin
shunyi maternal and children's hospital
Lu-Wen Xing
Beijing Chaoyang maternal and children's hospital
Jie Zhang
Beijing Daxing District People's Hospital: Capital Medical University Daxing Teaching Hospital
Ming-Bin Cheng
shunyi district hospital
Jie Liang
beijing shi jing shan hospital
Gui-Hua Yang
shunyi maternal and children's hospital
Juan Zhu
beijing shi jing shan hospital

Research article

Keywords: Breast cancer, mammography, deep learning, artificial intelligence

Posted Date: August 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-746374/v1

Version of Record: A version of this preprint was published at Chinese Journal of Cancer Research on January 1st, 2021. See the published version at https://doi.org/10.21147/j.issn.1000-9604.2021.06.05.

Abstract

Background: Computer-aided diagnosis using deep learning algorithms has been initially applied in the field of mammography, but there is no large-scale clinical application.

Methods: This study proposed to develop and verify an artificial intelligence model based on mammography. Firstly, retrospectively collected mammograms from six centers were randomized to a training dataset and a validation dataset for establishing the model. Secondly, the model was tested by comparing 12 radiologists' performance with and without it. Finally, prospectively multicenter mammograms were diagnosed by radiologists with the model. The detection and diagnostic capabilities were evaluated using the free-response receiver operating characteristic (FROC) curve and ROC curve.

Results: The sensitivity of model for detecting lesion after matching was 0.908 for false positive rate of 0.25 in unilateral images. The area under ROC curve (AUC) to distinguish the benign from malignant lesions was 0.855 (95% CI: 0.830, 0.880). The performance of 12 radiologists with the model was higher than that of radiologists alone (AUC: 0.852 vs. 0.808, P = 0.005). The mean reading time of with the model was shorter than that of reading alone (80.18 s vs. 62.28 s, P = 0.03). In prospective application, the sensitivity of detection reached 0.887 at false positive rate of 0.25; the AUC of radiologists with the model was 0.983 (95% CI: 0.978, 0.988), with sensitivity, specificity, PPV, and NPV of 94.36%, 98.07%, 87.76%, and 99.09%, respectively.

Conclusions: The artificial intelligence model exhibits high accuracy for detecting and diagnosing breast lesions, improves diagnostic accuracy and saves time.

Trial registration: NCT, NCT03708978. Registered 17 April 2018, https://register.clinicaltrials.gov/prs/app/NCT03708978

Background

Breast cancer is the most common malignant tumor in women [1], and the leading cause of cancer death in women worldwide. Early diagnosis can improve the 5-year survival rate of breast cancer patients from 25% to 99% [2]. Several imaging methods are used to identify suspicious malignant breast lesions, while mammography is the only screening method that has been proved to reduce the mortality of breast cancer [3-4], which can reduce the risk of breast cancer death up to 40% [5-6].

Asian women's mammary glands are denser, reducing the sensitivity of mammography. The large number of breast cancer screening population results in heavy mammography load, and uneven distribution of breast specialists makes difference in the level of mammography diagnosis. A number of studies have pointed out that about 75% of breast biopsies caused by suspicious mammography results are finally confirmed as benign changes [7]. The increase of unnecessary pathological biopsy leads to the waste of medical resources and further aggravates the shortage of medical resources. Therefore, it is highly essential to effectively and accurately detect breast lesions.

Computer-aided detection (CAD) uses computerized algorithms to identify suspicious regions of interest (ROIs) on imaging studies. It can assist radiologists as a second reader in detecting early breast cancer in an efficient way, especially on screening mammograms [8]. Since CAD was proved to improve the detection rate of cancer in 1998, it has been extensively applied thereafter for screening different types of cancer [9]. In spite of improving detection rate, CAD increases false positive rate and true positive rate [10].

In recent years, deep learning (DL), especially convolution neural network (CNN), has remarkably attracted scholars' attention for detection and classification of medical images [11-12]. Numerous machine learning models based on artificial intelligence (AI) have been successfully applied in imaging diagnosis and efficacy evaluation of breast, liver, and rectum [13-16]. Computer-aided models for mammographic breast cancer diagnosis have been proposed [17][18], and studies have shown that DL-based CAD may assist radiologists to improve diagnostic efficiency and reduce their work load [19-21]. However, there is no prospective large-scale clinical study, confirming the clinical practicability of DL-based CAD.

Therefore, the main purpose of the present study was to establish an artificial intelligence assisted diagnosis model based on deep learning method, to evaluate the effectiveness of the model for aiding doctors to obtain better accuracy and less working time, and to finally validated it in real world practice.

Methods

1. Study design and participants

This is a multicenter study, including both retrospective design and prospective design. The study was in accordance with the precepts established by the Helsinki Declaration, and he study protocol was approved by the Ethics Committee of our hospital (2019KT35). The informed consent was waived for the retrospective part, and obtained from all participants for the prospective part (Reg. No. NCT03708978). Our study consisted of three parts, the first part was the retrospective construction of AI system and internal verification. The second part tested whether the diagnosis efficiency of doctors with AI system assistance is higher than that of doctors alone. The third part prospectively verified the effect of the system in multi-center clinical practice.(see figure1)

1.1 Participants of the first part

We retrospectively enrolled patients who were admitted to our hospital for screening clinical symptoms from October 1, 2014 to September 30, 2016. Figure 1 shows the study flowchart. Table S1 showed the study sites and patients enrolled in this part. The inclusion criteria were patients with complete clinical data, mammogram data, and with pathological diagnosis or more than 2 years' follow-up after the first examination. The exclusion criteria included unqualified images required for the segmentation and inconsistency in the location of the lesion between the mammogram and the pathological results.

1.2 Participants of the second part

To determine the effectiveness of the model for improving the accuracy of diagnosis, we collected mammograms from six centers from October 1 to 31, 2015, and conducted a complete cross-sectional evaluation of the developed diagnostic system with participation of 12 radiologists.

Figure S1 illustrates the processes of collection of mammography data and participants' selection. A step-by-step procedure for estimating power and sample size was used that was proposed by Hillis et al. [22] for planned multi-reader receiver operating characteristic (ROC) studies. For 12 evaluators, in which the study efficacy was not less than 0.80, an area under the curve (AUC) difference of 0.05 required 200 mammograms (70 pathologically confirmed malignant cases, 30 pathologically or follow-up confirmed false positive cases, and 100 negative cases).

To ensure adequate mammography to determine the final sample size, we collected at least 14 cancer patients, 6 false-positive patients, and 20 negative patients in each center (data collected from centers E and F were combined due to the small number of cases in those centers). The inclusion and exclusion criteria were shown in figure 1.

To ensure image quality, all cases in this part were reviewed by three radiologists with more than 25 years of experience in mammography. Each case was available for pathology or follow-up. After review, 3 patients with unqualified image quality and 66 patients with very obvious symptoms of breast cancer were excluded.

1.3 Participants of the third part

To further investigate the clinical application of the model, we prospectively applied it in six centers. Patients undergoing mammography in each center were prospectively and consecutively enrolled from April 5, 2018 to May 4, 2018. The inclusion and exclusion criteria were shown in figure 1.There were no specific exclusion criteria in terms of demographic or clinical characteristics for participants without lesions.

2. Quality control of mammogram images

All mammogram images were stored using a picture archiving and communication system (PACS) in digital imaging and communications in medicine (DICOM) format. The two standard views were the craniocaudal (CC) and the mediolateral oblique (MLO). To ensure image quality, all cases were reviewed by 3 radiologists with more than 25 years of experience in mammography.

All pathological results were obtained from the pathology report and reviewed by an experienced pathologist. Pathological tissue was obtained by hollow needle biopsy or surgery and was stained with hematoxylin and eosin (H&E).

3. Radiologist's annotations

Six certified and experienced radiologists, each with an average experience of at least 5 years (range, 5-10 years), read an average of 250,000 mammograms and annotated the images. Six radiologists were trained to read 800 mammograms and began to draw ROI respectively. The delineation principle was as follows: (1) manual delineation along the edge of the lesion; (2) inclusion of all suspicious parts of the tumor in the sketch; (3) the edge included burrs as far as possible; (4) when the label was generated, the characteristics of the lesion were marked according to the Breast Imaging-Reporting and Data System (BI-RADS) (2013 edition), including lesion type (mass, calcification, structural distortion, asymmetry), distribution characteristics, and pathological or follow-up results. In case of doubtfulness, a radiologist will consult with three other experienced radiologists to make a correct decision after discussion.

4. Algorithm development

Following the successful application of DL, we established the model (http://mgshow.yizhun-ai.com/), containing various modules to carry out automatic analysis of mammograms. It contains three deep neural models: the lesion detection module, the lesion matching module, and the malignant degree assessment module, which constitute a complete system for breast lesion analysis (Figure S2). The overview of our system is illustrated in Figure 2.

(a) Lesion detection module

We use Faster R-CNN [23] to detect suspicious lesions in all of the images of one patient. Faster R-CNN is one of the state-of-the-art methods in the area of object detection. Faster R-CNN contains two stages, where the first stage generates box proposals and the second stage refines the box localization and predicts the class of each object. We use ResNet-50 [24] as the backbone network and adopt feature pyramid network to enhance the detector performance of small lesions.

Since the huge size of breast images and the existence of background areas with no information, we first pre-process the mammogram images before sending them to the neural networks. We crop the foreground area of each image by a simple thresholding method and then resize the images to keep spacing = 0.15mm. As shown in 1, the detector takes the four images of different views as inputs, and outputs bounding boxes and lesion classes (i.e., mass and calcification) for detected suspicious lesions. In our problem, mass and calcification can appear at the same location, so we use Sigmoid function to generate the objectivity score for each class instead of SoftMax. This modification allows an object to be identified as both mass and calcification. In practice, if a predicted box has high confidence in both mass and calcification, we will call this lesion a mass with calcification.

(b) Lesion matching module

The matching module is introduced to indicate whether a pair of detected candidates are from different views of the same lesion. In the clinical practice of mammogram examination, it is essential to combine the information of multiple views (MLO and CC). At most of the time, a lesion could be recognized in both MLO and CC views. If a mass can be only found in one view, radiologists may consider that it is caused

by overlapping glands, but not lesion. According to this principle, it is natural to perform false positive reduction by matching the lesions of MLO and CC view in the CAD system.

In our model, we use a neural model to conduct lesion matching. The matching model is after the detector and takes the features of the detected proposals of suspicious lesions as input. We use vertex coordinates, sizes of the proposals, the probabilities of each class, and the depth of proposals in the gland as input features. In the matching process, the model should use the information of all proposals to perform matching, so that we use an attention model [25] to predict the relationship of all lesion pairs. The input of the model is the concatenated features mentions above, and it generates the probability of a real lesion pair for all possible pairs. The lesions with low probabilities will be removed during the output process.

(c) Malignant degree assessment module

We use a CNN based on ResNet [26] to estimate the malignant degrees of lesions. In our model, we treat the malignant degree assessment problem as an ordinal regression problem [24]. Ordinal regression algorithms are to solve multi-class classification problem where the labels have strong ordinal relationships. In our problem, BI-RADS can represent a lesion's degree of malignancy. BI-RADS sometimes provide more information than pathological results, since pathological results only tell us whether a lesion is malignant, but BI-RADS can tell us how malignant a lesion's degree of malignancy. Therefore, we use BI-RADS to train our model. Experimentally, with large amounts of BI-RADS annotations confirmed by experts, we find the performance of our system is better than using the pathological results as labels, even we evaluate the system according to the pathological results.

Following some previous work [24], we use integration of several binary classification problems to solve the ordinal regression problem. We choose ResNet-18 as our backbone, which is one of the state-of-the-art classification models in the area of deep learning [26]. In our data, there are 8 labels ('false positive', 'BI-RADS 2', 'BI-RADS 3', 'BI-RADS 4A', 'BI-RADS 4B', 'BI-RADS 4C', 'BI-RADS 5' and 'BI-RADS 6'). Since there are little lesions which are 'BI-RADS 2' or 'BI-RADS 6' in our training data, we treat 'BI-RADS 2' the same as false positive candidates and merge 'BI-RADS 6' and 'BI-RADS 5'. Therefore, our model outputs 5 logits for each lesion, the first logit predicts whether the BI-RADS of a lesion is larger than 'BI-RADS 3', the second logit predicts whether the BI-RADS of a lesion is larger than 'BI-RADS 4A' and so on. Since we hope the network can output the possibility that a lesion is malignant, we add a fully connected layer to process the result of ordinal regression, which can be seen as a simple linear combination.

The online demo was shown in appendix. To train the models, the collected mammograms were chronologically divided into training dataset (~80%) and validation dataset (~20%). We trained the models during the first part of our study and further evaluated the established system in the next two parts.

5. Auxiliary efficacy for of the model

We evaluated the effectiveness of the model in detecting and diagnosing mammograms by monitoring the performance of 12 radiologists under different reading conditions (see Figure S3).

The 12 radiologists had an average of 9.5 years (range, 3 to 25 years) of experience with the certificate of Mammography Quality Standards Act, and had read more than 5000 mammograms per year over the past two years.

The 12 radiologists were blinded of any information about the patients, including prior imaging and histopathological reports. The assessment consisted of two stages. Each radiologist received separate training prior to the first evaluation. The purpose of the training was to familiarize radiologists with the evaluation criteria and the functions and operations of the AI-aided diagnosis model. Besides, 12 radiologists were informed that the rate of malignancy in the assessed dataset was higher than clinical practice.

For each case, the radiologists employed the BI-RADS classification (range, 1-5), and labeled the suspicious lesion as benign or malignant, and normal patients without lesion were taken as negative into account. The radiologists scored each case on a difficulty scale of 1-9 (9 represents the highest difficulty).

The evaluation was undertaken on an in-house developed workstation, using a 12-MP Mammography Display System that was calibrated to the medical grayscale standard display function of digital imaging. Radiologists used the AI system to read the film, which can freely adjust the window width and window level, and can scale and shift. Ambient lighting was set to about 45 lux.

6. Prospective clinical applications of the model

Prior to the application of the model in each center, nineteen radiologists in the six centers had participated in the training of the model, in which 200 cases were trained. The median experience in mammography diagnosis was 9.5 years (range, 5–26 years), and the mean number of mammograms read each year during the past 2 years was approximately 6500 (range, 1400–13 000). The purpose of the training was to make all the radiologists proficient in the operating system and application interface, so that they could be used freely in the routine clinical mammography.

The mammography was conducted by radiologists with the DL model at six centers. The model could automatically identify suspicious lesions and percentage of malignancy for reference, and automatically generate structured reports as well. The reading time of each case was automatically recorded by the system. Pathological and follow-up results were taken as the gold standard for the diagnosis of benign and malignant lesions, and three radiologists with more than 20 years of experience were taken as the gold standard for the detection of lesions, so as to observe the clinical effect of the DL model.

7.Statistical analysis

Clopper–Pearson method was applied to calculate the accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the model used to detect and diagnose mammographic lesions (see Appendix). We used the free-response receiver operating characteristic (FROC) curve to indicate the detection ability of the model and further analyze its diagnostic ability in different types of lesions. The ROC curve was plotted, and the AUC was used to evaluate the diagnostic performance of the model. All statistical analyses were bilateral with significance level of 0.05. Statistical analyses were performed using R 3.5.1 programming language.

The end point was to compare the AUC, sensitivity, specificity, and reading time of 12 radiologists who read independently and with the model. $P<0.05$ indicated a statistically significant difference between the two reading conditions. In the present study, if a radiologist did not mark the malignant lesion within the true quadrant of the lesion, the case was modified to be negative by the reader.

The reading time of each case was automatically measured by the workstation software. The paired sample t-test or Wilcoxon rank-sum test were used to compare the average reading time under two different reading conditions (reading alone and reading with the model), and the relationship between reading time and difficulty score was further analyzed. For this analysis, the outlier (defined as more than 1.5 times the standard deviation of the data) was removed.

8.Outcomes and follow-up

Definition of malignant lesions: within 2 years from the time the patient came to the hospital for the first mammogram, the pathological diagnosis of the same lesion as malignant was defined as malignant lesion. Definition of benign lesions :(1) pathological diagnosis of the same lesion within 2 years was benign;(2) the patients were followed up for more than 2 years, and mammography more than 2 years after the first mammography examination indicated benign, without pathological diagnosis. Follow-up plan was in supplementary files.

Results

1. Patients' baseline data

The flowchart of the study design and data collection is shown in Figure 1. Between October 1, 2014 and September 30, 2016, 5350 participants with suspected lesions from PACS of six centers were enrolled, of whom 891 were excluded due to no follow-up data or follow-up for less than two years, and 92 were excluded because their pathological results were not obtained. Of the remaining 4,367 participants assessed for quality control, 97 (2.22%) were excluded due to poor quality of mammography and 151 (3.46%) were excluded due to inconsistency in anatomical location and pathological report. Eventually, 16,476 images of 4119 participants were involved in the analysis, including 2454 patients with malignant lesions and 1665 patients with benign lesions. Among them, pathological results of 3186 patients were achieved through biopsy or surgery. In chronological order, a total of 3389 patients were used for model training from October 1, 2014 to May 31, 2016, and 730 patients were recruited for model verification

from June 1, 2016 to September 30, 2016 (approximately 5:1). The patients' data are summarized in Table S2.

The mean age of 200 patients tested for auxiliary efficacy of the model was 59 years (Table S3), and the detailed pathological types of malignant cases are presented in Table S4.

A total of 5,809 cases of mammography were involved, and 63 cases were excluded according to the exclusion criteria (9 cases had no pathological results, 50 cases had no follow-up results, and 4 cases failed to undergo mammography). The remaining 5,746 cases were included in the analysis. There were 495 patients with malignant lesions, 337 patients with benign lesions, and 4914 negative patients. The prevalence of breast cancer in A-F centers was 15.72%, 5.91%, 7.52%, 3.83%, 11.11%, and 7.10%, respectively. There was no significant difference in the patients' baseline data (Table S5).

2. Results of the first part-validation of the model

When there was a 0.25 false positive rate per image, the overall sensitivity of detection in the validation dataset was 0.828. The sensitivity of detection after matching was 0.908 for false positive rate of 0.25 in unilateral images. Among all the lesions, the AUC of the model to distinguish the benign from malignant lesions was 0.855 (95% confidence interval (CI): 0.830, 0.880). For mass and calcification, the AUC for benign and malignant were 0.865 and 0.841, respectively (Figure 3, Table S6, Figure S4)

3. Results of the second part-comparing clinical data between the model and 12 radiologists

3.1 ROC curve, sensitivity and specificity

The AUC for the model-independent diagnosis was 0.835 (95% CI: 0.819, 0.852). The diagnostic performance of 12 radiologists with the model-assisted was higher than that of 12 radiologists reading alone (AUC: 0.852 vs. 0.808, P = 0.005) (Figure 4, Table S7). The specificity of 12 radiologists with the model-assisted was higher than that of reading alone (88.34% vs. 82.05%, P = 0.005), and there was no significant difference in sensitivity between these two groups of radiologists (68.78% vs. 68.70%, P = 0.937) (Table S8). Figures S7 and S8 showed examples of the correct number of detection and diagnosis changed under different reading conditions. The sensitivity and specificity of the model independent diagnosis were 81.40% and 78.50%, respectively. The sensitivity and specificity of 12 radiologists reading alone are shown in Table S9.

3.2 Reading time

With the model, the mean reading time of 12 radiologists was significantly shorter than that of 12 radiologists alone (80.18 s vs. 62.28 s, P=0.03). Additionally, with the aid of the model, the reading time of 1 radiologist increased (6.1%), while that of 11 radiologists decreased (range, 9.1-48.3%) (Figure 5a). Of all the reading time, 0.4% (21 of 4800) was defined as an outlier and was excluded from this analysis.

For cases with low difficulty coefficients (1-5 points), with the aid of the model, the average reading time of each case was reduced by 35.2%. On the contrary, for the cases with high difficulty coefficient (6-9 points), the reading time of each case was elevated by 6.5%.

3. Results of the third part-prospective clinical results of the model

With the model, the sensitivity of detection reached 0.887 at false positive rate of 0.25(Figure S5). With the model, the AUC of differentiating benign from malignant lesions was 0.983 (95% CI: 0.978, 0.988) (Figure 6, Table S10). The sensitivity, specificity, PPV, and NPV of diagnosis were 94.36%, 98.07%, 87.76%, and 99.09%, respectively. The AUC of the model diagnosing alone in A-F centers was 0.959, 0.959, 0.986, 0.970, 0.941, and 0.989, respectively (Figure S6). The mean diagnosis time of each mammogram was 94.23 s.

Discussion

In the present study, a mammography-based artificial intelligence model for breast cancer was established, and it was unveiled that the proposed system had superior diagnostic performance, and can assist radiologists to improve the diagnostic accuracy and shorten the diagnosis time. Finally, through prospective multicenter population verification, the system exhibited a satisfactory auxiliary diagnostic performance. To our knowledge, this is the first prospective clinical research in the field of mammography based on AI, and outstanding outcomes could be achieved.

In order to avoid missed diagnosis, an AI-assisted diagnosis model may lead to increase of false-positive rate. In clinical application, a model with high false-positive rate may result in over-testing, interfering with radiologists' attention, consuming radiologist' energy, and increasing patients' psychological anxiety and financial burden. Several AI-based models for mammography were previously reported, some of which were developed for the purposes of detection and classification [27], and some of which were developed based on clinical data [28-29], while none of them tackled the above-mentioned deficiencies. Our model could make a correlation between the two views of lesions. Our model used a matching module to combine the image on the CC position and MLO position to ensure that the detected lesion was a true positive lesion. It can be seen from the data obtained before and after matching that the matching module reduced the false-positive rate, while ensured the sensitivity, and improved the accuracy of differentiation of benign from malignant lesions, indicating the reliable capability of clinical application of the proposed system.

In our study, three different participants were selected for model developing, comparative testing and prospective validation. During the development of the model, we selected the population with suspicious lesions in mammography for better learning. In the comparison test, the 200 mammogram cases were significantly more difficult than those in the usual clinical work, in order to better test the auxiliary ability of the model. In the prospective verification, the cases we collected were as close to the real world as possible, which is more conducive to observing the role of assisted diagnosis system in the real world. The results presented were different due to the differences in the population observed. In the first part of

the study population are suspected cases and therefore the AUC of classification of the model is 0.852, and in the prospective part, participants include clinic diagnosis and screening, as well as breast X-ray negative cases, the results reached 0.983.

In the part of testing whether the model can assist a radiologist to improve the diagnostic performance, we deliberately selected the difficult and differentiated cases. This aimed to monitor radiologists' diagnostic accuracy in diagnosing difficult cases and simple cases, so as to better assess the capability of the model in assisting radiologists for diagnosing and clarifying its clinical application value. With the model, the 12 radiologists' diagnostic performance was higher than that without assistance (0.852 vs. 0.808, P = 0.005). It indicated that radiologists' diagnostic performance can be improved with the DL model. The results showed that the sensitivity of the diagnosis of 12 radiologists reading alone was quite different (38.8-98.6%), which is consistent with result of a previous study [30], and is also one of the important reasons for the implementation of double-reading. This may be related to radiologists' experience. The AUC of the model-independent diagnosis was 0.835 (95% CI: 0.819, 0.852), which was close to some radiologists' diagnostic performance. Therefore, it is feasible to make the model for fast and robust diagnosing patients with breast cancer.

The reading time of 12 radiologists with the model was significantly shorter than that of reading alone. The reading time was shortened in a number of radiologists by up to 50%. We speculated that this might be related to radiologists' experience, and this conclusion was consistent with a previous research's outcome [19]. When there were cases with low-difficulty coefficient (1-5 points), the viewing time was markedly shortened with the aid of the model, indicating that our model can save time and enable radiologists to further concentrate on cases with high-difficulty coefficient, so that radiologists could avoid the possibility of missed diagnosis and misdiagnosis. For cases with high-difficulty, it increased the diagnostic time while the average increase was only 6.5%, which was still within the acceptable range.

The previously reported AI-based models for mammography were partially limited to the detection of lesions [18], and they were partly tested on public datasets [27]. In contrast, the model exhibited high detection and diagnostic efficacy in prospective clinical applications in six different centers. In addition, the model showed high sensitivity and specificity for detection of the two types of lesions (masses and calcifications). In particular, the detection of calcification accompanied with satisfactory results under the background of generally dense glands in Asian women, which greatly shortens the detection time of lesions and saves radiologists' energy in detecting lesions, thereby assisting radiologists to improve the diagnostic efficiency. In addition, the prospective application results of the model achieved in six different centers reflected its universality and practicality.

Despite the above-mentioned outstanding results, this study has several limitations. First, we didn't carry out a prospective multicenter randomized controlled trial to validate the superiority of auxiliary diagnosis with the model compared without it, because it is hard to randomly assign clinical cases through image diagnosis system. Second, there are still a limited number of deficiencies in the matching module (i.e., the network cannot deal with mismatch between lesions and other views). Third, in terms of clinical

applicability, our model was only conducted by training and validation datasets in a large population in mainland China, and its effectiveness in other populations (such as Western countries) remains to be further studied.

Conclusions

We developed an AI-assisted diagnostic model for breast cancer, and demonstrated that it can improve the diagnostic accuracy and shorten the time for breast cancer diagnosis. The clinical application of the model was completed for the first time in prospective multi-centers, which highlighted the effectiveness and applicability of the AI-assisted diagnostic system.

Abbreviations

Artificial intelligence，AI

Area under curve，AUC

Breast imaging reporting and data system, BI-RADS

Computer-aided detection，CAD

Convolution neural network，CNN

Confidence interval，CI

Craniocaudal, CC

Deep learning，DL

Free-response receiver operating characteristic, FROC

Intersection over Union，IOU

Mediolateral oblique, MLO

Negative predictive value, NPV

Positive predictive value, PPV

Receiver operating characteristic, ROC

Declarations

Acknowledgements

The authors thank the patients, the investigators and their research teams, who participated in this study.

Availability of data and materials

Data can be acquired by request to Ying-Shi Sun through email.

Funding

This study was supported by Beijing Municipal Science & Technology Commission (No. Z181100001918001), Beijing Municipal Administration of Hospitals Clinical Medicine Development of Special Funding Support (No. ZYLX201803) and Beijing Hospitals Authority Ascent Plan (Code:20191103).

Authors' contribution

SYS conceived and designed the study. QYH, LY, YL, DJB, XB, LBQ collected and assembled the data. QYH, LY, YL, DJB, XB, LBQ, SYJ, SRJ, LR, LHJ, CM, YYY, XLW, ZJ, CMB, LJ, YGH and ZJ conducted image analysis. WD, LXT, ZKX, WYC, CDB and WLW conducted statistical analysis. All authors contributed to the data interpretation. All authors contributed to the manuscript drafting or manuscript revision for important intellectual content. All authors approved the final version of submitted manuscript.

Ethics approval and consent to participate

The study was in accordance with the precepts established by the Helsinki Declaration, and he study protocol was approved by the Ethics Committee of our hospital (2019KT35). The informed consent was waived for the retrospective part, and obtained from all participants for the prospective part (Reg. No. NCT03708978).

Consent for publication

All authors agreed to this publication.

Competing interests

None declared

References

[1] Chen W, Zheng R, Baade PD, et al. Cancer Statistics in China, 2015.CA CANCER J CLIN 2016;66:115–132.

[2] Hillman BJ, Goldsmith JC. The uncritical use of high-tech medical imaging. N Engl J Med. 2010;363(1):4-6.

[3] Berry DA, Cronin KA, Plevritis SK, et al. Effect of screening and adjuvant therapy on mortality from breast cancer. N Engl J Med. 2005;353(17):1784-92.

[4] Løberg M, Lousdal ML, Bretthauer M, et al.Benefits and harms of mammography screening. Breast Cancer Res. 2015; 17(1): 63.

[5] Myers ER, Moorman P, Gierisch JM, et al.Benefits and Harms of Breast Cancer Screening: A Systematic Review, JAMA. 2015;314(15):1615-34.

[6] Duffy SW, Tabár L, Yen AM, et al. Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women. Cancer. 2020;126(13):2971-2979.

[7] Elmore JG, Nakano CY, Koepsell TD, et al. International variation in screening mammography interpretations in community-based programs. Journal of the National Cancer Institute, 2003;95 (18):1384-1393.

[8] Chan HP, Doi K, Galhotra S, Vyborny CJ, MacMahon H, Jokich PM. Image feature analysis and computer-aided diagnosis in digital radiography. I. Automated detection of microcalcifications in mammography. Med Phys. 1987;14(4):538

[8] Fenton JJ, Xing G, Elmore JG, Bang H, Chen SL, Lindfors KK, Baldwin LM. Short-term outcomes of screening mammography using computer-aided detection: a population-based study of medicare enrollees. Ann Intern Med. 2013;158(8):580.

[10] Taylor P, Potts HW. 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. 2008;44(6):798.

[11] Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 2017; 542(7639):115.

[12] 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(22):2402.

[13] Drukker K, Giger ML, Joe BN, et al. Combined Benefit of Quantitative Three-Compartment Breast Image Analysis and Mammography Radiomics in the Classification of Breast Masses in a Clinical Data Set. Radiology 2019; 290(3):621-628.

[14] Liu Z, Li Z, Qu J, et al.Radiomics of multi-parametric MRI for pretreatment prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer: a multicenter study.Clin Cancer Res 2019;25:3538–3547

[15] Yasaka K, Akai H, Abe O, Kiryu S. Deep Learning with Convolutional Neural Network for Differentiation of Liver Masses at Dynamic Contrast-enhanced CT: A Preliminary Study, Radiology.

2018;286(3):887-896.

[16] Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists, PLoS Med. 2018; 20;15(11):e1002686.

[17] 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(2):331-342.

[18]Kooia T, Litjensa G, van Ginnekena B, et al. Large scale deep learning for computer aided detection of mammographic lesions. Medical Image Analysis 2017; 35:303–312.

[19] Rodríguez-Ruiz A, Krupinski E, Mordang JJ, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, Mann RM.Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. Radiology. 2019;290(2):305-314.

[20] 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(7):434-440.

[21] McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89-94.

[22] Hillis SL, Obuchowski NA, Berbaum KS. Power estimation for mul- tireader ROC methods an updated and unified approach. Acad Radiol 2011;18(2):129–142.

[23] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, 2015; pp. 91-99.

[24] Chen S, Zhang C, Dong M, Le J and Rao M. Using ranking-cnn for age estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5183–5192, 2017.

[25] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L and Polosukhin I. Attention is all you need. In Advances in neural infor- mation processing systems, pages 5998–6008, 2017.

[26] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016; pp. 770-778.

[27] Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with Deep Learning. Sci Rep. 2018;8(1):4165.

[28] 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(2):331-342.

[29] Wu N, Phang J, Park J, et al. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans Med Imaging. 2020;39(4):1184-1194.

[30] Elmore JG, Jackson SL, Abraham L, et al. Variability in Interpretive Performance at Screening Mammography and Associated with Accuracy. Radiology. 2009, 253(3): 641–651.

Figures

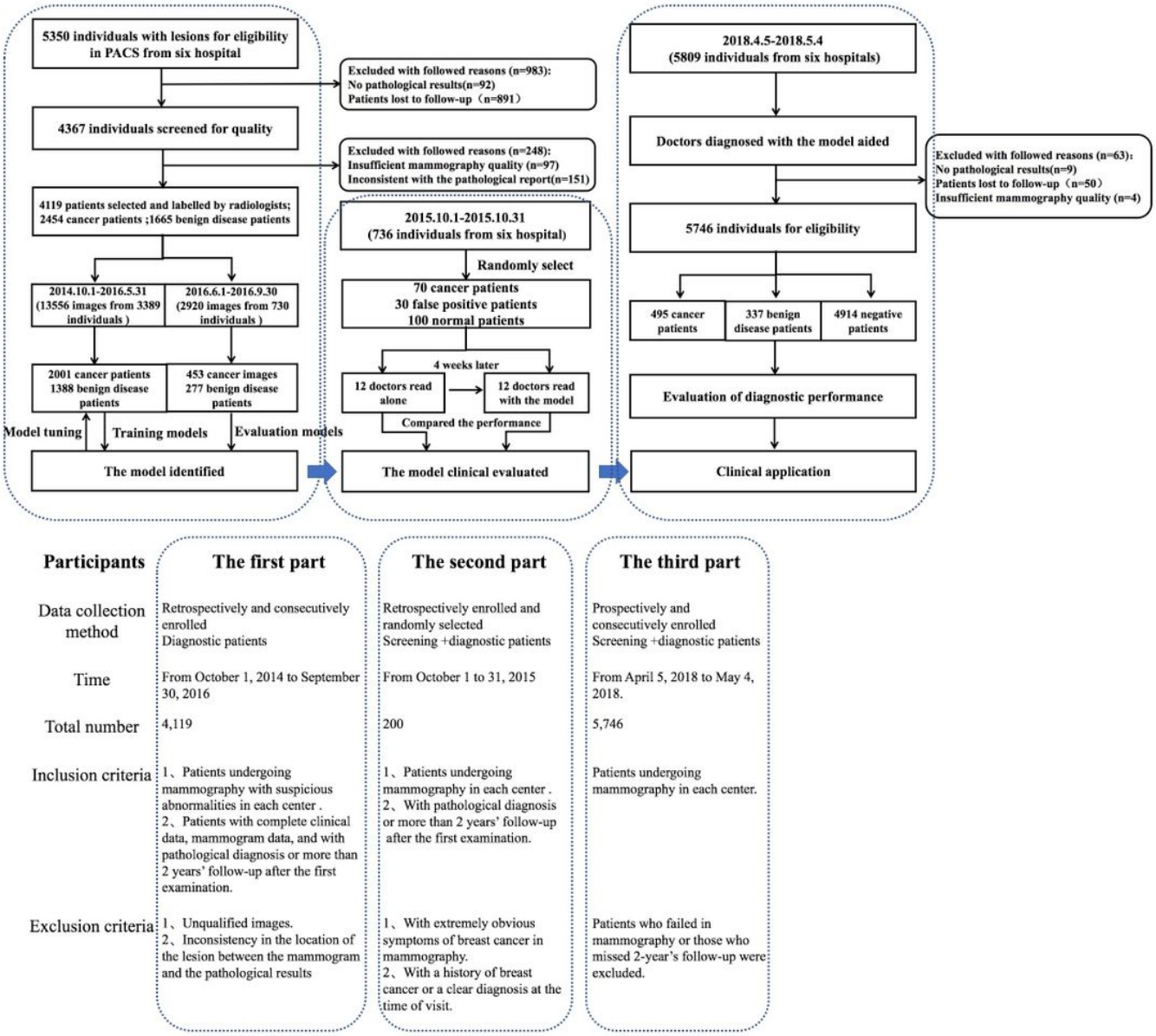

Figure 1

Workflow diagram for the development and application of the model.

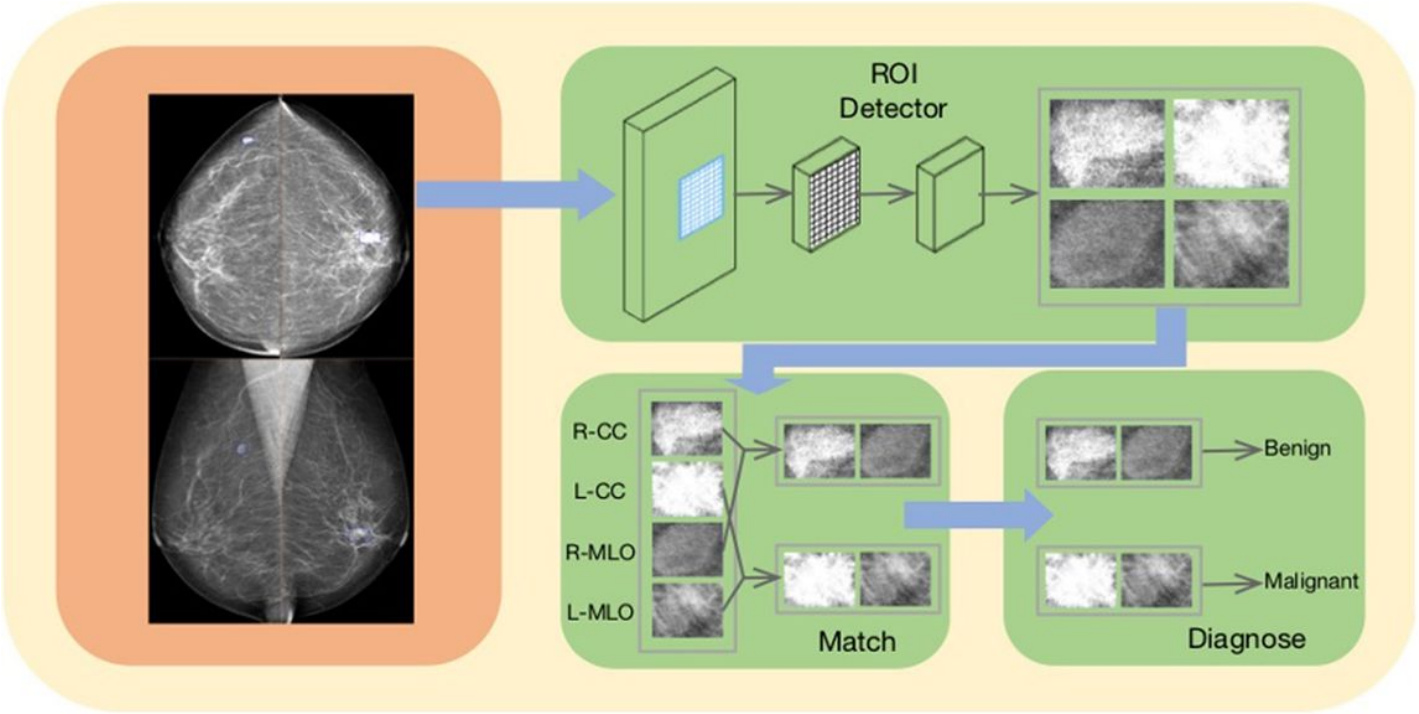

Figure 2

The pipeline of the model.

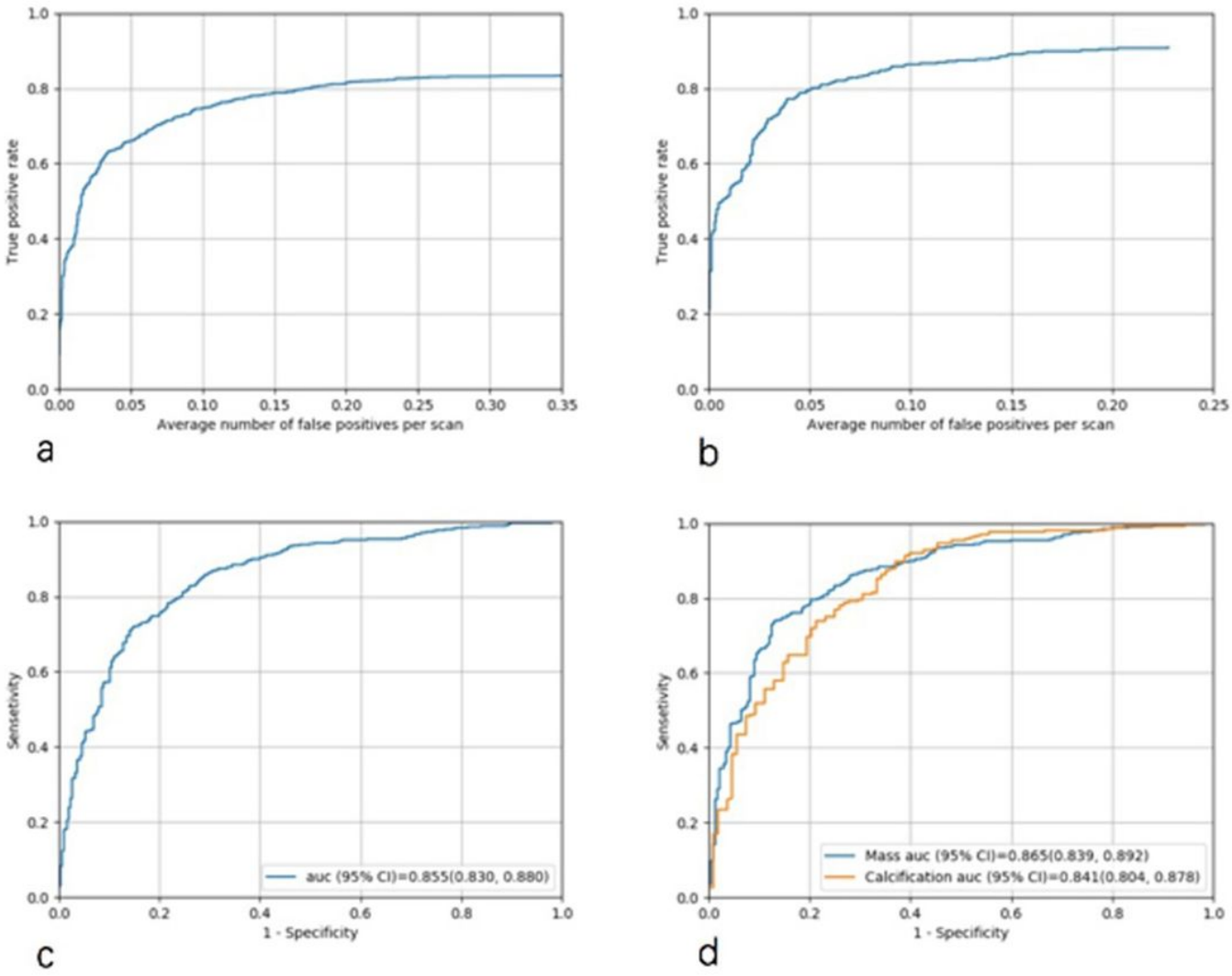

Figure 3

FROC and ROC curves in the validation dataset. (a) FROC curve for detection; (b) FROC curve for detection after matching; (c) ROC curve for distinguishing benign from malignant lesions; (d) ROC curve of the model to differentiate benign from malignant lesions for calcification and masses, respectively (Note: Only the detected lesions were considered, and the test result and the IOU marked by the radiologists were > 0.25). FROC: free-response receiver operating characteristic; ROC: receiver operating characteristic; IOU: intersection over Union.

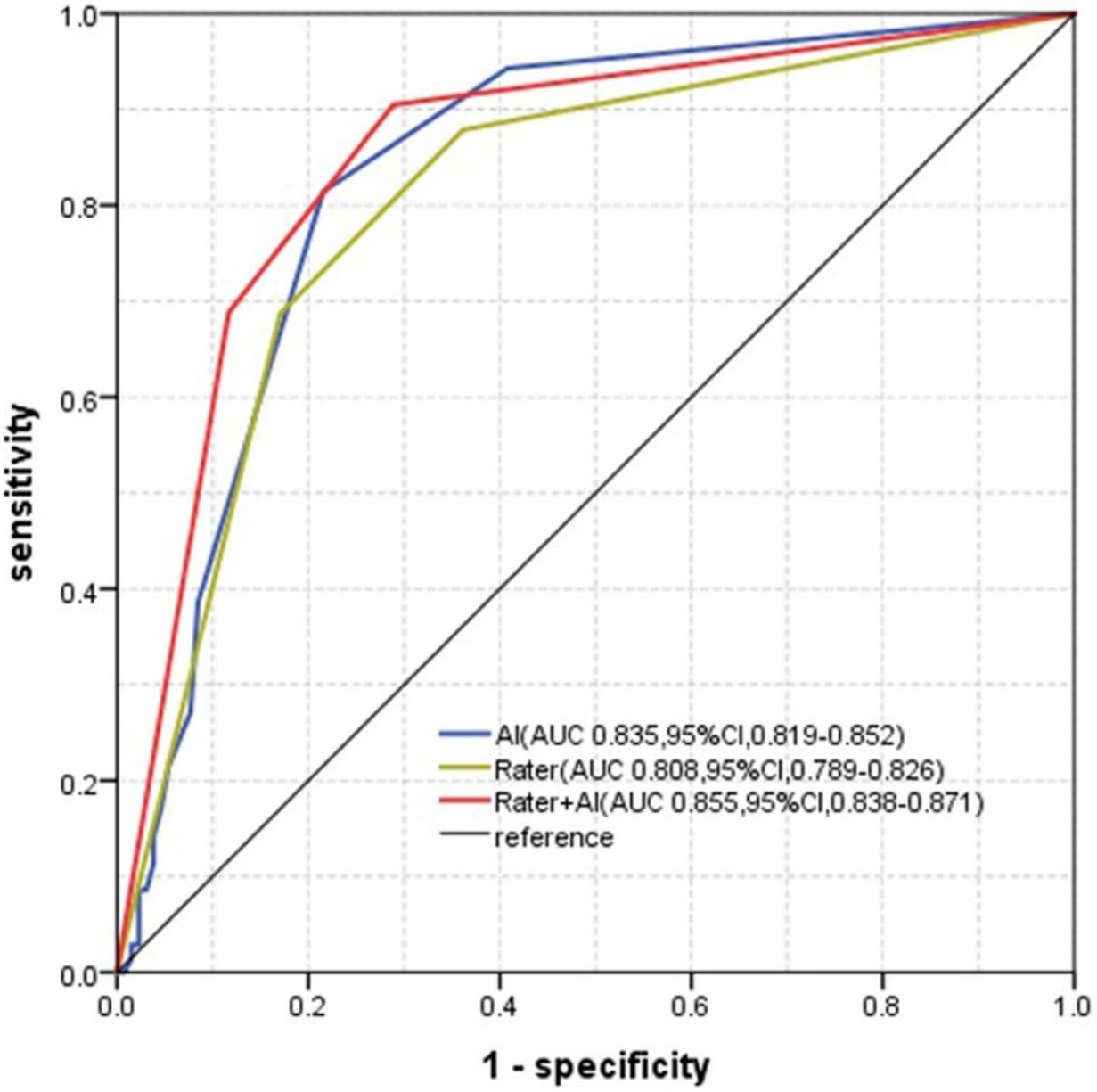

Figure 4

ROC curves of diagnosis assisted by the model and diagnosis alone for 12 radiologists. ROC: receiver operating characteristic.

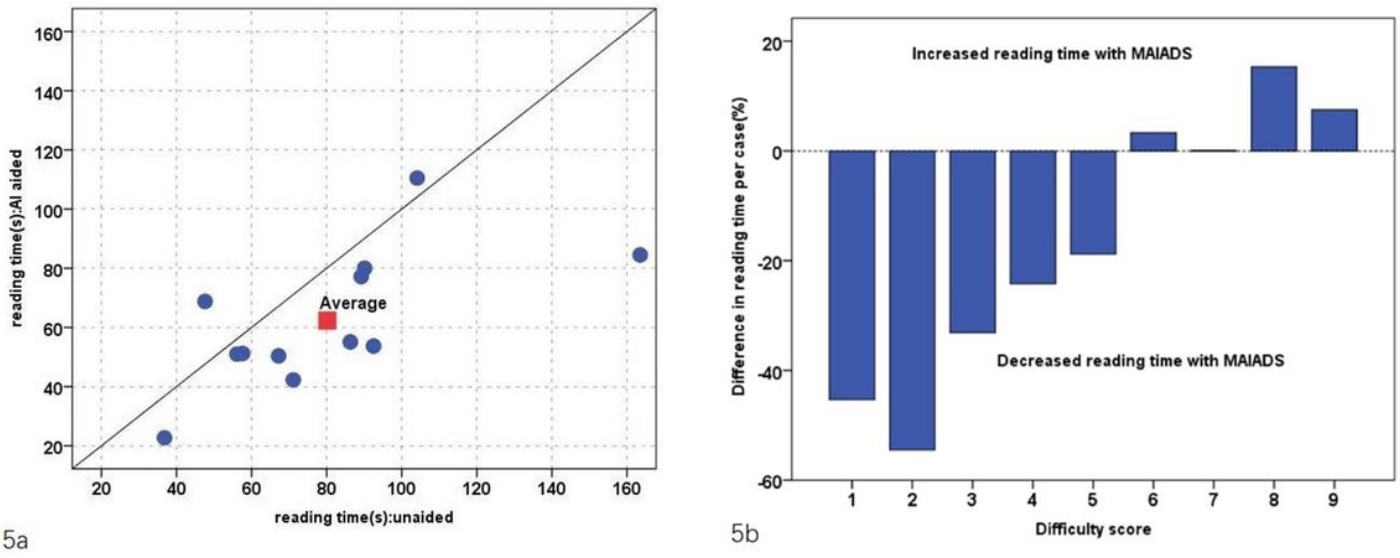

Figure 5

Reading times. (a) The graph shows the reading time (circle) and mean reading time (square) of each case for each radiologist. (b) The bar graph depicts the difference in reading time caused by different difficulty scores.

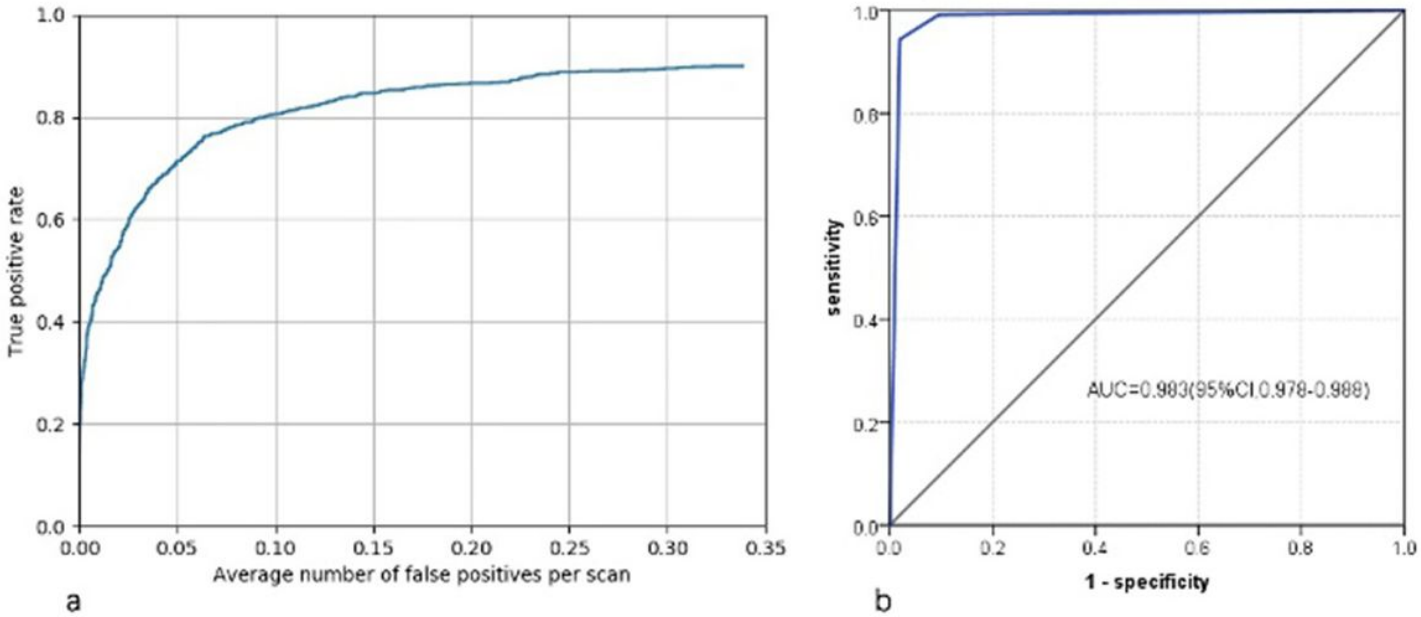

Figure 6

Diagnostic performance of radiologists with the aid of the model in the prospective multicenter clinical application. (a)FROC curve; (b) ROC curve for differentiating malignant lesions. FROC: free-response receiver operating characteristic; ROC: receiver operating characteristic.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- figureS1.jpg
- figureS3.jpg
- figureS4.jpg
- figureS5.jpg
- figureS6.jpg
- figureS7.jpg
- figureS2.jpg
- figureS8.jpg
- supplementaryfiles.docx