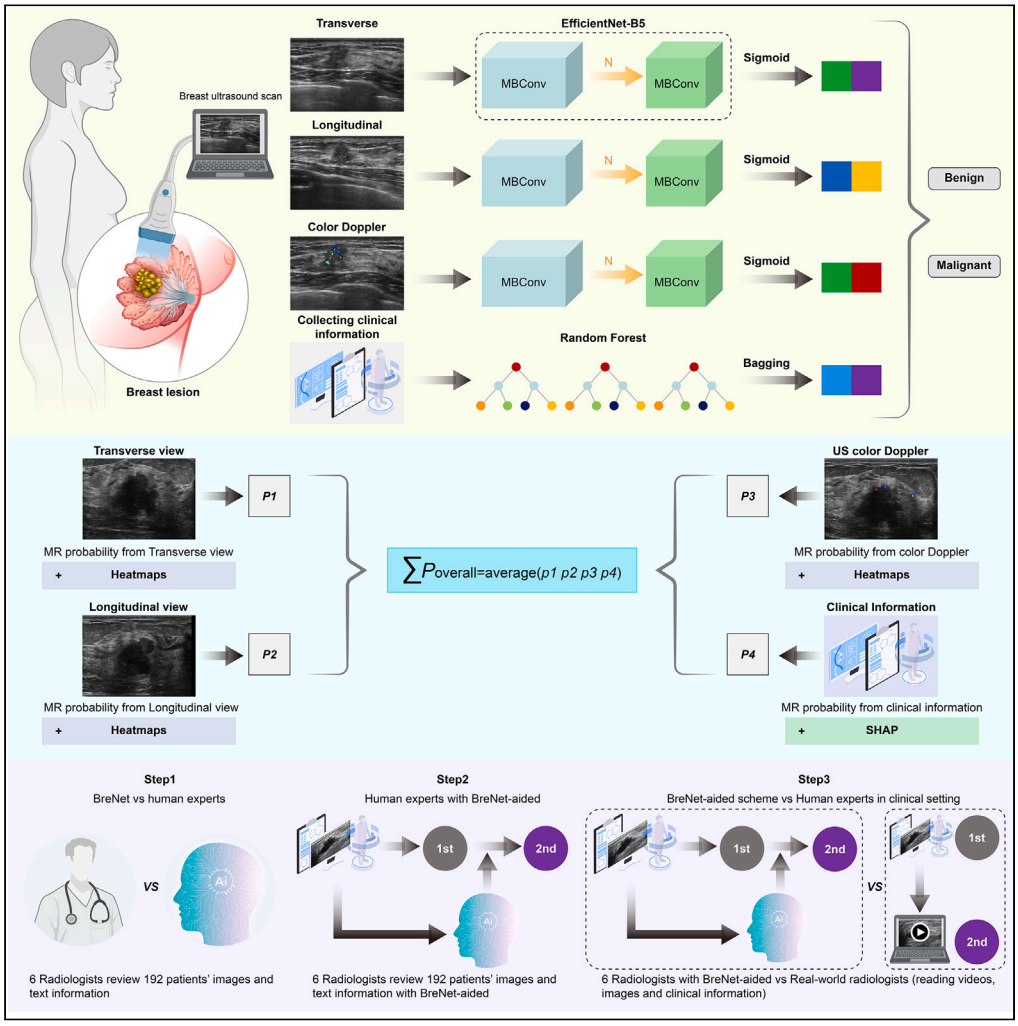
iScience

CellPress
OPEN ACCESS

Article

Assessing breast disease with deep learning model using bimodal bi-view ultrasound images and clinical information



Fengping Liang, Yihua Song, Xiaoping Huang, ..., Yuanqing Li, Yong Ren, Zuofeng Xu

auyqli@scut.edu.cn (Y.L.)
koalary@qq.com (Y.R.)
xuzuofeng77@aliyun.com (Z.X.)

Highlights

Accurate breast ultrasonography diagnosis is associated with the breast cancer prognosis

Deep learning model applied to ultrasound endow great help in breast carcinoma

The BreNet-aided tactic improved radiologists' performance in breast tumor diagnosis

Liang et al., iScience 27, 110279
July 19, 2024
https://doi.org/10.1016/j.isci.2024.110279

iScience

CellPress
OPEN ACCESS

Article

Assessing breast disease with deep learning model using bimodal bi-view ultrasound images and clinical information

Fengping Liang,[1,10] Yihua Song,[1,10] Xiaoping Huang,[2] Tong Ren,[1] Qiao Ji,[1] Yanan Guo,[1] Xiang Li,[1] Yajuan Sui,[1] Xiaohui Xie,[3] Lanqing Han,[4] Yuanqing Li,[5,6,*] Yong Ren,[7,8,9,*] and Zuofeng Xu[1,11,*]

SUMMARY

Breast cancer is the second leading cause of carcinoma-linked death in women. We developed a multimodal deep-learning model (BreNet) to differentiate breast cancer from benign lesions. BreNet was constructed and trained on 10,108 images from one center and tested on 3,762 images from two centers in three steps. The diagnostic ability of BreNet was first compared with that of six radiologists; a BreNet-aided scheme was constructed to improve the diagnostic ability of the radiologists; and the diagnosis of real-world radiologists' scheme was then compared with the BreNet-aided scheme. The diagnostic performance of BreNet was superior to that of the radiologists (area under the curve [AUC]: 0.996 vs. 0.841). BreNet-aided scheme increased the pooled AUC of the radiologists from 0.841 to 0.934 for reviewing images, and from 0.892 to 0.934 in the real-world test. The use of BreNet significantly enhances the diagnostic ability of radiologists in the detection of breast cancer.

INTRODUCTION

Breast cancer has acquired much concentration in recent years due to the dramatic rise in its incidence, which increased the global burden of disease. It is the most commonly diagnosed malignant tumor among women globally, with almost two million women being diagnosed with breast cancer annually.[1] Previous investigation has shown that routine mammography can drastically reduce breast carcinoma-related deaths in Western countries.[2] However, in the case of dense breast, the diagnostic sensitivity of mammography is greatly reduced.[3] In China, the average age at the time of diagnosis of breast carcinoma is 45–55 years, which is considerably younger than that in Western countries. Chinese females have a higher proportion of dense breast tissue.[4] Therefore the diagnostic rate, accuracy, and cost-effectiveness ratio of mammography for diagnosing breast cancer in Chinese women are significantly worse than that of ultrasonography (US).[5] A primary advantage of US is that it does not result in exposure to radiation. US has a supplemental carcinoma detection ratio of approximately 4/1,000 scans, and it can reduce the supplemental carcinoma proportion further.[6] Nevertheless, US is denounced due to its comparatively poor specificity, which results in recalls and biopsies of benign tumors.[7]

Artificial intelligence (AI) has the potential to overturn the existing pattern of cancer diagnosis and management by swiftly screening a large number of images and categorizing them, which is a difficult task for clinicians. Fukushima developed a self-organizing neural network model known as the Neocognitron in 1980.[8] The theory has established the basis of algorithms for convolutional neural networks (CNNs). A CNN was applied to medical images in 2015, which classified fundus photographs into four groups.[9] There is already a large body of high-quality research showing that mammography-based deep learning (DL) model outperforms radiologists at both predicting and diagnosing breast cancer.[10–13] However, high-quality research on ultrasound-based DL models is relatively limited. Qian et al. constructed a polymerization model. Two DL models were used in the AI system: a bi-modal model (inputs with US B-mode and US color Doppler images) and a multimodal model (inputs with US B-mode, US color Doppler, and US elastography images).[14] Clinical parameters can assist to offer complementary information for imaging diagnosis. But all the aforementioned breast-related studies excluded patients' medical histories in the training DL model process.

[1]Department of Medical Ultrasound, The Seventh Affiliated Hospital, Sun Yat-sen University, 628 Zhenyuan Road, Shenzhen, China
[2]Department of Ultrasound, Dongguan Songshan Lake Tungwah Hospital, No. 1, Kefa Seventh Road, Songshan Lake Park, Dongguan, China
[3]Section of Epidemiology and Population Science, Department of Medicine, Baylor College of Medicine, Houston, TX, USA
[4]Center for Artificial Intelligence in Medicine, Research Institute of Tsinghua, Pearl River Delta, Guangzhou, China
[5]School of Automation Science and Engineering, South China University of Technology, Guangzhou, China
[6]Research Center for Brain-Computer Interface, Pazhou Lab, Guangzhou, China
[7]Artificial Intelligence and Digital Economy Laboratory (Guangzhou), PAZHOU LAB, No.70 Yuean Road, Haizhu District, Guangzhou, China
[8]Shensi Lab, Shenzhen Institute for Advanced Study, UESTC, Shenzhen, China
[9]The Seventh Affiliated Hospital of Sun Yat-Sen University, Shenzhen, China
[10]These authors contributed equally
[11]Lead contact
*Correspondence: auyqli@scut.edu.cn (Y.L.), koalary@qq.com (Y.R.), xuzuofeng77@aliyun.com (Z.X.)
https://doi.org/10.1016/j.isci.2024.110279

Few investigators have used multi-modal, multi-view ultrasound images and clinical information to construct CNN. However, radiologists usually analyze medical images and correlate the images to the patients' clinical information, such as the age at menopause, family history, and serological examination results, to obtain a precise clinical diagnosis. Therefore, we developed BreNet, a multi-modal DL model, to differentiate breast cancer from benign lesions, and sought to investigate the effect of this model on the radiologists' diagnostic performance, workload, and rate of needless biopsies. We hypothesize that the performance of BreNet would be superior to that of radiologists'.

RESULTS

Baseline characters

The baseline characteristics of the participants and breast tumors are shown in Table S1. The radiologists had 1–12 years of experience in breast disease diagnosis (see also Table S2). The 730 participants were divided into trained (532 participants, 42.8 $\pm$ 11.9 years), internal validation (98 participants, 40.7 $\pm$ 10.9 years), and external validation (100 participants, 38.8 $\pm$ 10.5 years) sets. BreNet, the proposed multimodal DL system, is shown in Figure 1.

Construction of the BreNet system and performances of diagnosis

The related risks of the clinical information variables for breast cancer are shown in Figure 2. Text feature sorting of Shapley additive explanations (SHAP) suggested that breast cancer was highly associated with age at menopause and high estrogen levels. Breast carcinoma showed a poor correlation with the remaining eleven text features.

For the detection of breast cancer, the algorithm had an area under the curve (AUC) of 0.913 (95% confidence interval [CI] 0.891–0.932] for the transverse view, 0.927 (0.891–0.943) for the longitudinal view, 0.955 (0.926–0.973) for color Doppler, 0.945 (0.938–0.965) for multiple images, and 0.985 (0.961–0.992) for clinical information. BreNet showed the best performance, with an AUC of 0.996 (0.983–0.998), which was superior to the performance of the five DL models mentioned previously (Figure 3A). Accuracy was 0.882 (0.857–0.902), 0.891 (0.864–0.910), 0.923 (0.892–0.951), 0.913 (0.891–0.924), 0.981 (0.953–0.994), and 0.983 (0.975–0.992) of transverse view, longitudinal view, color Doppler, multiple images, clinical information, and BreNet, respectively. The BreNet algorithm had a sensitivity and specificity of 0.993 and 0.982, respectively (Table 1). The comprehensive performance of BreNet was superior.

The AUC and accuracy values of BreNet were 0.850 (0.832–0.862) and 0.791 (0.761–0.802) in external test set, respectively (Figure 3B; Table 2). The BreNet model had a sensitivity and specificity of 0.784 and 0.791, respectively (Table 2).

Comparison of the BreNet with human experts in three stages

Two independent test sets of 3,762 images from 198 participants were used to compare the diagnoses of the six radiologists with those made by BreNet. The radiologists were apprised to make a diagnosis per research object, using the patient's breast US images combined with clinical information. The ability to identify breast cancer compared with benign breast tumors was presented as a receiver operating characteristic (ROC) curve, and this ability of BreNet was superior to the ability of the radiologists (Figure 4). The AUC value and accuracy of BreNet were 0.996 (0.983–0.998) and 0.983 (0.975–0.992), respectively, in internal test set, whereas those of the radiologists (reading static images and clinical information) were 0.841 (0.811–0.867) and 0.874 (0.865–0.877), respectively (Table 3). The AUC value and accuracy of BreNet were 0.850 (0.832–0.862) and 0.791 (0.761–0.802), respectively, in the external test set, whereas those of the radiologists were 0.752 (0.735–0.767) and 0.791 (0.761–0.831), respectively (Table 4). In the practical clinical scenario test of BreNet, The AUC of the second diagnosis (reading videos, images, and clinical information) was refined to 0.892 (0.873–0.901) in the internal set and 0.862 (0.823–0.912) in the external set. The pooled AUC of the radiologists with the aid of BreNet (only reviewing static images and clinical information) was superior to that of the radiologists reading static images, videos, and clinical information in the internal set (0.934 [0.928–0.948] vs. 0.892 [0.873–0.901]) (Figure 4A; Table 3). However, the result was opposite (0.804 [0.783–0.829] vs. 0.862 [0.823–0.912]) in the external set (Figure 4B; Table 4).

In the estimation of the BreNet-assisted diagnostic procedure (internal test set), the pooled AUC for the radiologists alone was 0.841 (0.811–0.867), which increased to 0.934 (0.928–0.948) with the aid of BreNet ($p < 0.0001$). The pooled accuracy of the radiologists alone was 0.874 (0.865–0.877), which improved to 0.942 (0.923–0.947) with BreNet aid ($p < 0.0001$). For the senior radiologists, the pooled AUC increased from 0.851 (0.844–0.889) to 0.922 (0.905–0.932) ($p < 0.0001$), whereas it increased from 0.832 (0.801–0.841) to 0.953 (0.933–0.957) for the junior radiologists ($p < 0.0001$). With the aid of BreNet, accuracy of the senior experts increased from 0.893 (0.857–0.896) to 0.931 (0.916–0.936) ($p < 0.0001$), and the κ value rose from 0.714 to 0.832 ($p < 0.0001$). With the aid of BreNet, accuracy of the junior experts increased from 0.853 (0.840–0.869) to 0.953 (0.941–0.959) ($p < 0.0001$), and the κ value rose from 0.661 to 0.893 ($p < 0.0001$) (Figure 4).

The categories created by the BreNet model can be applied to reduce the workloads related to the decision-making course of biopsy while maintaining the administrative criterion. Clinical settings were simulated by excluding biopsies for cases with at high negative predictive value (NPV) and positive predictive value (PPV) when the ACR BI-RADS conformed to the BreNet judgment (Figure 5).

Based on the ACR BI-RADS guidelines, no biopsy was required if the tumors were classified as category $\leq$III. The NPV for predicting benign tumors for these lesions was 97.8% (95% CI 95.2–100). The classification-related NPV was recomputed when the BreNet prediction was also benign. Previous studies have reported NPV ranges from 50% to 98%.[15–17] When the ACR BI-RADS was applied with BreNe-aid, the NPVs of breast lesions category III were as high as 99.0% and 88.0% for category IVa, and 87.0% for category IVb (Figure 5A, see also Table S3). Among the 198 tumors, a median of 115 (IQR, 81–134) biopsies were waived with the assistance of BreNet, owing to the high possibility of predicting the tumor as benign. Previous literature has reported a PPV ranging from 0.8% to 97%.[15–18] With the assistance of BreNet,

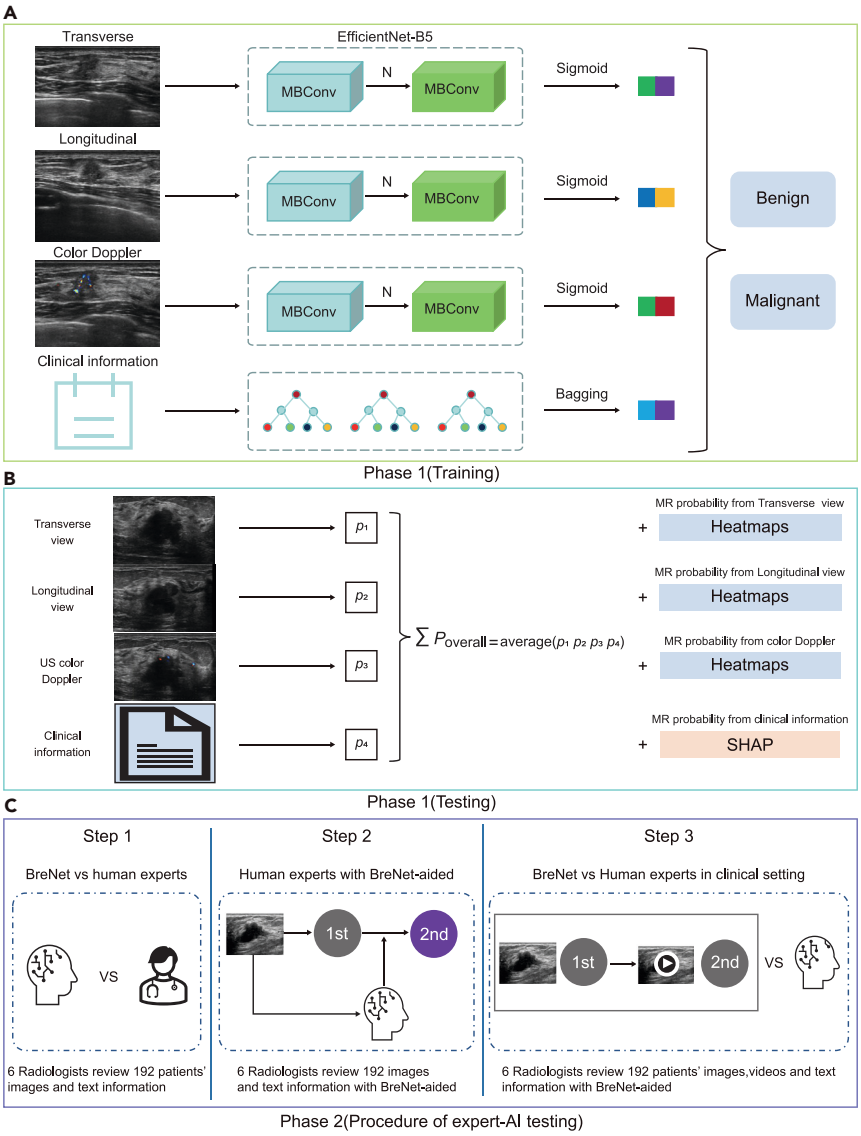


Figure 1. Graphical abstract

(A) A CNN pretrained on the ImageNet dataset of 1,000 categories can be acclimated to drastically rise in the accuracy and curtailed the training period of a network trained on a new dataset of US breast images. The locally connected (convolutional) layers were frozen and transferred into a novel network. Finally, fully connected layers were reconstructed and retrained from random initialization on top of the transferred layers.

(B) BreNet architecture diagram: after preprocessing, ultrasound images are subjected to feature extraction through three different modalities using EfficientNet, with each modality yielding a corresponding prediction probability via a sigmoid activation function. EfficientNet consists of several MBConv blocks, each featuring varying expansion factors, kernel sizes, and strides. Clinical data are processed using a random forest classifier to extract features and generate respective prediction probabilities. Finally, the probability values derived from imaging and clinical data are aggregated to produce the final prediction value for diagnosis.

(C) BreNet was then tested in three steps on the same dataset including an internal and an external set. First, diagnostic ability between human experts with BreNet founded on US static-images was compared. Second, diagnostic ability of human-experts alone and human-experts with BreNet-aided was evaluated founded on US static-images. Third, the first judgment founded on static-images and the second judgment founded on videos were noted down, respectively. Then, we compared the diagnosis of real-world radiologists (simultaneous reading both static images, videos, and clinical information) with radiologists with BreNet-aided (only reading static images and clinical information). Each step was progressed with an interval of more than one month. All the radiologists who participated in the test were blind to the pathological outcomes throughout the progress. CNN, convolutional neural network; MR, malignant rate; MBConv, mobile inverted bottleneck convolution.

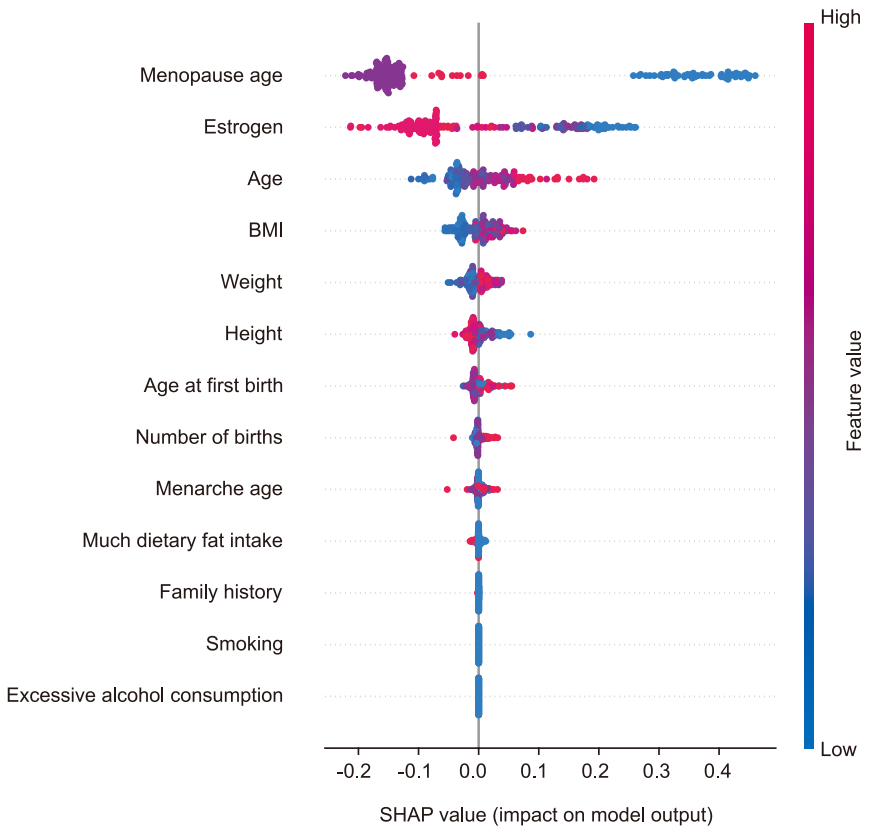


Figure 2. Shapley additive explanations (SHAP) of the clinical information analyzing breast lumps

The x axis is essentially the average amplitude change in the model output when features are "hidden" from the model. The greater the variation in the average amplitude, it presented that the corresponding feature was more important. BMI, body mass index.

the lesions were predicted as malignant, and categories IVb–V had an analogous PPV at 96.7% (Figure 5B, see also Table S4). Among the 198 tumors evaluated, a median of 14 (IQR, 5–22) biopsies were waived with the assistance of BreNet owing to the high possibility of predicting malignancy.

Among the 3,648 diagnoses made by the six experts, the assistance of BreNet helped reduce biopsies from 44.3% to 23.4% ($p < 0.0001$), and the proportion of missed malignancies reduced from 6.5% to 2.2% ($p < 0.0001$; Figure 5C).

In the clinical test setting, the pooled mean duration required by the radiologists to review static images and videos were 4.09 s and 7.45 s, respectively. The total duration required by the radiologists with the assistance of BreNet (only reviewing static images) was less than the duration required by the radiologists to simultaneously review both static images and videos (11.54 s vs. 4.09 s, $p < 0.0001$) (Table 5). The diagnostic performances of both were found to be similar (Figure 4; Table 4). The time required by BreNet for diagnosis was less than milliseconds; therefore, its impact was negligible. Clinical information was collected with Excel in advance, and the time required by senior and junior radiologists to review the images or videos was the same; therefore, it is not listed in the table. However, the time required by the junior radiologists was longer than that required by the senior radiologists to review static images or videos (3.81 s vs. 4.23 s, 6.85 s vs. 7.75 s, respectively; $p < 0.05$) (Table 5).

Interpretability of the BreNet

The heatmap of the US images (including B-mode, and color Doppler images) was incorporated to identify the regions that contributed the most to the predicted diagnosis of the DL model. Heatmaps simplified the evaluation of the region of interest per US imaging mode with clinical significance. The following variables aid in the prediction of the risk of breast carcinoma on ultrasound B-mode images: echogenicity, shape, margin, and echogenic foci. The accent of the heatmap was centered on the areas of the vasculature of the color Doppler images, particularly for breast cancer with rich angiogenesis (see also Figure S2).

DISCUSSION

The BreNet-aided tactic increased the diagnostic performance of human experts when reviewing the images and clinical information only as well as when reviewing images, videos, and clinical information in the clinical scenario. The amalgamation of the ACR BI-RADS with BreNet

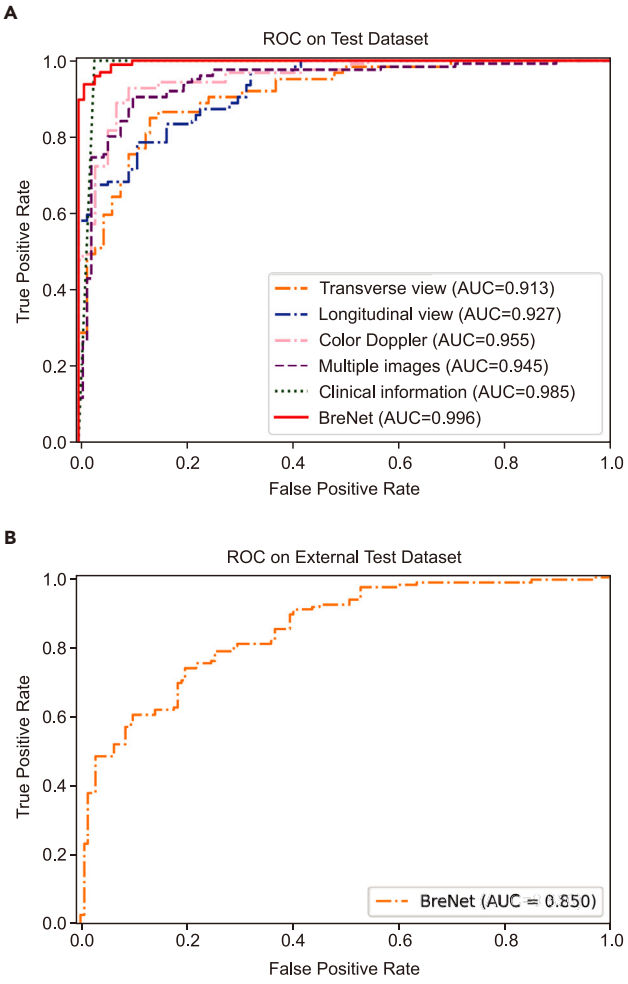


Figure 3. Performance of the CNNs in predicting the risk of breast carcinoma on the prospective dataset

(A) In internal test dataset, a total of six deep CNNs were further compared, including transverse view, longitudinal view, color Doppler, multiple images, clinical information, and our polymerization mode (BreNet).

(B) In external test dataset, the performance of the BreNet in predicting the risk of breast carcinoma.

enhanced PPV and NPV, which has the potential to reduce the number of unnecessary biopsies. The performance of BreNet in the external testing set was inferior to that in the internal testing set. This outcome can be attributed to that the external dataset comprised more subtypes of breast cancer. This may have made the US features of the tumors atypical. The findings of the present study, that the DL system outperformed human experts, are compatible with those of previous studies.[14,19] In real-world settings, many confounding factors exceed the

Table 1. Performance of the CNNs in predicting the risk of breast carcinoma on the prospective internal test dataset

	AUC	F1	ACC	PPV	NPV	SENS	SPEC
Transverse	0.913 (0.891–0.932)	0.842	0.882 (0.857–0.902)	0.954	0.793	0.693 (0.667–0.723)	0.981 (0.958–0.992)
Longitudinal	0.927 (0.891–0.943)	0.853	0.891 (0.864–0.910)	0.946	0.794	0.763 (0.741–0.783)	0.961 (0.935–0.972)
Color Doppler	0.955 (0.926–0.973)	0.911	0.923 (0.892–0.951)	0.942	0.891	0.864 (0.837–0.882)	0.962 (0.947–0.973)
Multiple images	0.945 (0.938–0.965)	0.914	0.913 (0.891–0.924)	0.883	0.915	0.902 (0.885–0.922)	0.903 (0.881–0.923)
Clinical information	0.985 (0.961–0.992)	0.983	0.981 (0.953–0.994)	0.945	0.984	0.992 (0.975–0.995)	0.971 (0.949–0.991)
BreNet	0.996 (0.983–0.998)	0.981	0.983 (0.975–0.992)	0.954	0.994	0.993 (0.984–0.996)	0.982 (0.961–0.992)

AUC, area under the curve; F1, a weighted average of the PPV and SENS; ACC, accuracy; PPV, positive predictive value; NPV, negative predictive value; SENS, sensitivity; SPEC, specificity.

Table 2. Performance of the BreNet in predicting the risk of breast carcinoma on the prospective external test dataset

	AUC	F1	ACC	PPV	NPV	SENS	SPEC
BreNe	0.850 (0.832–0.862)	0.782	0.791 (0.761–0.802)	0.782	0.803	0.784 (0.756–0.889)	0.791 (0.766–0.806)

AUC, area under the curve; F1, a weighted average of the PPV and SENS; ACC, accuracy; PPV, positive predictive value; NPV, negative predictive value; SENS, sensitivity; SPEC, specificity.

domains of these ideal algorithms. Clinical decisions must be overseen by healthcare professionals, even if the DL model is verified to have a better diagnostic ability. Consequently, the most vital characteristic of this DL system is to enhance performance by aiding human experts in clinical decisions. In the internal dataset, the diagnostic ability of the real-world radiologists was inferior to that of the BreNet-assisted radiologists, whereas the opposite was the case in the external dataset. These findings suggest that the radiologists' diagnostic level is relatively stable between external and internal datasets, with of the AUC ranging from 0.86 to 0.89. The AUC of BreNet decreased from 0.93 to 0.80 in the internal and external datasets, indicating that there may be overfitting of the existing model.

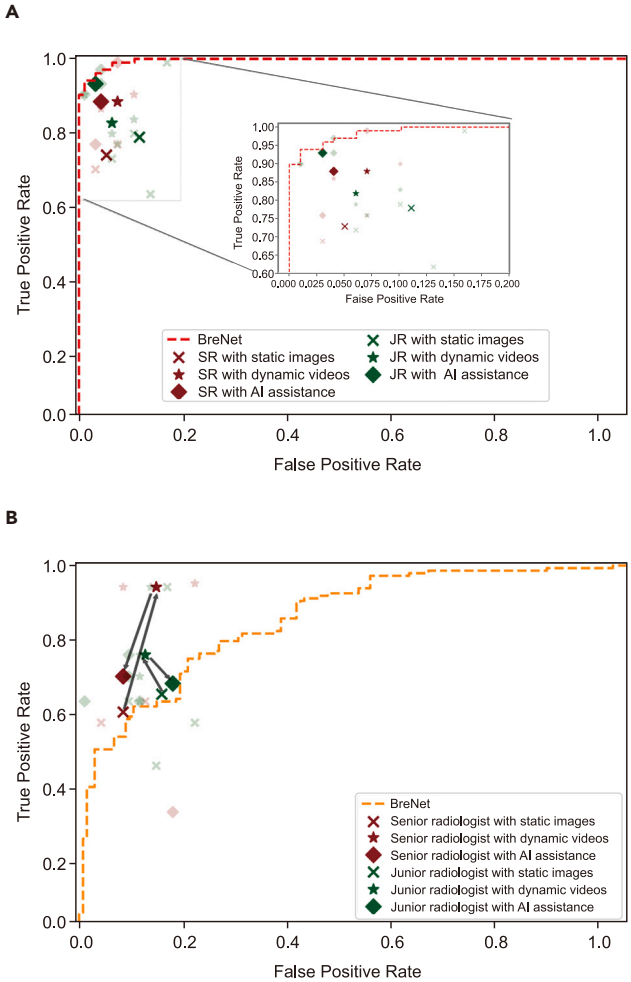


Figure 4. Diagnostic performance of BreNet and radiologists in testing sets for distinction of malignant from benign breast lumps

(A) Diagnostic performance of BreNet compared to every specialist in the internal test set. Light colored forks, stars, and squares denote diagnostic SENSs and SPECs of each specialists using static images plus clinical information, clinical scenario (static images, videos, and clinical information), and BreNet-assisted, respectively. The dark green fork, star, and square indicate the pooled SENSs and SPECs of all junior radiologists using static images plus clinical information, clinical scenario, and BreNet-assisted, respectively. The crimson fork, star, and square indicate the pooled SENSs and SPECs of all senior radiologists using static images plus clinical information, clinical scenario, and BreNet-assisted, respectively.

(B) Diagnostic performance of BreNet compared to every specialist in the external test set. The content referred to by forks, stars, and squares are the same as Figure 4A. SENSs, sensitivities; SPECs, specificities.

Table 3. The diagnostic performance of BreNet alone, radiologists alone, and BreNet-aided radiologists in internal dataset

	AUC	F1	K	ACC	PPV	NPV	SENS	SPEC
BreNet	0.996 (0.983–0.998)	0.981	0.952	0.983 (0.975–0.992)	0.954	0.994	0.993 (0.984–0.996)	0.982 (0.961–0.992)
Radiologists alone (reviewing static images and clinical information)								
All	0.841 (0.811–0.867)	0.863	0.692	0.874 (0.865–0.877)	0.814	0.903	0.754 (0.721–0.771)	0.921 (0.903–0.957)
Senior	0.851 (0.844–0.889)	0.882	0.714	0.893 (0.857–0.896)	0.864	0.892	0.736 (0.719–0.744)	0.95 (0.933–0.968)
Junior	0.832 (0.801–0.841)	0.843	0.661	0.853 (0.840–0.869)	0.753	0.913	0.782 (0.761–0.799)	0.89 (0.879–0.917)
Radiologists alone (reviewing static images, dynamic videos, and clinical information)								
All	0.892 (0.873–0.901)	0.902	0.774	0.913 (0.881–0.923)	0.854	0.941	0.853 (0.840–0.871)	0.934 (0.920–0.942)
Senior	0.901 (0.881–0.923)	0.903	0.763	0.913 (0.879–0.918)	0.841	0.948	0.881 (0.859–0.892)	0.930 (0.911–0.938)
Junior	0.872 (0.861–0.879)	0.904	0.771	0.914 (0.889–0.935)	0.853	0.934	0.821 (0.807–0.861)	0.943 (0.931–0.950)
Radiologists with BreNet-aided								
All	0.934 (0.928–0.948)	0.940	0.861	0.942 (0.923–0.947)	0.912	0.963	0.902 (0.879–0.934)	0.965 (0.952–0.973)
Senior	0.922 (0.905–0.932)	0.933	0.832	0.931 (0.916–0.936)	0.891	0.952	0.884 (0.869–0.887)	0.960 (0.942–0.968)
Junior	0.953 (0.933–0.957)	0.950	0.893	0.953 (0.941–0.959)	0.924	0.974	0.931 (0.883–0.942)	0.970 (0.957–0.978)

AUC, area under the curve; F1, a weighted average of the PPV and SENS; ACC, accuracy, PPV, positive predictive value, NPV, negative predictive value; SENS, sensitivity; SPEC, specificity.

Based on these findings, we recommend the use of AI models in clinical practice. First, US imaging is frequently subject to inter-observer inconsistencies, especially when performed by young radiologists at remote areas, with the difference in the Fleiss kappa value ranging from 0.37–0.78.[20–22] BreNet provides a consistent solution for this problem. With the assistance of BreNet, the pooled accuracy of the junior radiologists increased from 83% to 95% in the internal test dataset. Second, the ACR BI-RADS with the aid of BreNet could be effective for the exemption of needless biopsy. The PPV and NPV in our survey were both >95%, suggesting that supervision or direct surgical resection without biopsy was feasible. Third, the BreNet system has the potential to reduce the workload of radiologists, especially in mass-scanning scenarios. It can also be made available in remote and backward areas with finite or even non-existent radiologists. Breast-scanning procedures performed by nurses can only capture static high-resolution images of the mass. Under these conditions, the diagnostic performance of nurses with the aid of BreNet was comparable with that of professional radiologists. If the AI system predicts malignancy, the patient can be transferred to a hospital for further assessment.

In this study, we combined clinical information with bi-view US B-mode and US color Doppler images to construct an AI system for interpreting breast tumors that is closer to clinical practice. Previous studies were used images only (even single-view ultrasonic B-mode image) to construct a DL model, which is unlike the workflow of human experts for diagnosing diseases.[9,19,23] Moreover, the DL system was used to diagnose US breast tumors with AUC values ranging from 0.833–0.955.[14,24,25] Becker et al. used a DL model to categorize breast carcinoma

Table 4. The diagnostic performance of BreNet alone, radiologists alone, and BreNet-aided radiologists in external dataset

	AUC	F1	K	ACC	PPV	NPV	SENS	SPEC
BreNet	0.850 (0.832–0.862)	0.782	0.553	0.791 (0.761–0.802)	0.782	0.803	0.784 (0.756–0.889)	0.791 (0.766–0.806)
Radiologists alone (reviewing static images and clinical information)								
All	0.752 (0.735–0.767)	0.781	0.484	0.791 (0.761–0.831)	0.563	0.923	0.623 (0.582–0.646)	0.874 (0.845–0.938)
Senior	0.760 (753–0.783)	0.812	0.532	0.822 (0.801–0.842)	0.643	0.921	0.594 (0.578–0.612)	0.921 (0.886–0.941)
Junior	0.742 (0.731–0.750)	0.763	0.433	0.773 (0.755–0.790)	0.482	0.922	0.641 (0.617–0.652)	0.854 (0.834–0.881)
Radiologists alone (reviewing static images, dynamic videos, and clinical information)								
All	0.862 (0.823–0.912)	0.873	0.622	0.872 (0.857–0.902)	0.591	0.963	0.842 (0.802–0.901)	0.873 (0.852–0.889)
Senior	0.911 (0.882–0.954)	0.890	0.673	0.881 (0.879–0.911)	0.610	0.992	0.941 (0.922–0.956)	0.861 (0.845–0.873)
Junior	0.823 (0.801–0.872)	0.864	0.574	0.874 (0.850–0.877)	0.583	0.944	0.753 (0.733–0.789)	0.884 (0.876–0.902)
Radiologists with BreNet-aided								
All	0.804 (0.783–0.829)	0.813	0.592	0.850 (0.839–0.860)	0.672	0.932	0.682 (0.665–0.709)	0.930 (0.927–0.939)
Senior	0.811 (0.805–0.833)	0.844	0.594	0.844 (0.823–0.846)	0.654	0.930	0.691 (0.681–0.732)	0.924 (0.915–0.925)
Junior	0.792 (0.778–0.801)	0.822	0.591	0.854 (0.848–0.866)	0.681	0.934	0.673 (0.661–0.679)	0.932 (0.920–0.941)

AUC, area under the curve; F1, a weighted average of the PPV and SENS; ACC, accuracy; PPV, positive predictive value; NPV, negative predictive value; SENS, sensitivity; SPEC, specificity.

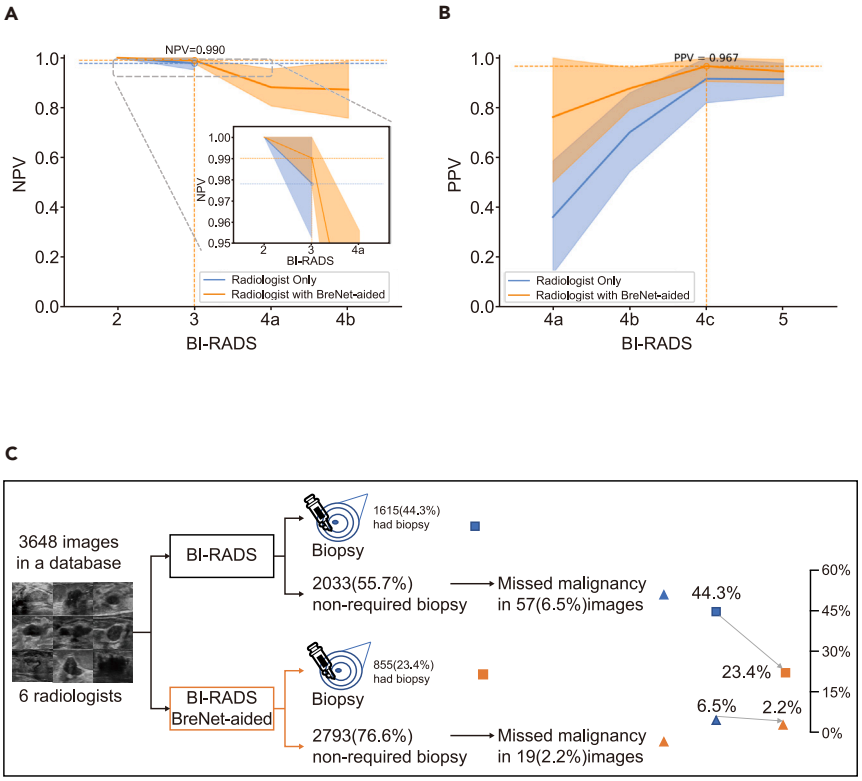


Figure 5. Recommendation for biopsy with BreNet-aided

(A) Tendency of NPV as per ACR BI-RADS diagnosis is benign. The shadow zones are 95% CI.

(B) Trend of PPV based on ACR BI-RADS when diagnosis is malignant. The shadow zones are 95% CI.

(C) Breast lesion management aided by BreNet. The BreNet-aided tactic lessened biopsies from 44.3% to 23.4% ($p < 0.0001$), meanwhile missed malignancy diminished from 6.5% to 2.2%. CI, confidence interval; ACR, American College of Radiology; NPV, negative predictive value; PPV, positive predictive value; BI-RADS, Breast Imaging Reporting and Data System.

and compared its performance with those of two radiologists and one intern radiologist.[26] In the present study, BreNet was developed using prospective data and its ability was compared with that of six specialists (two seniors, three juniors, and one intern radiologist). Thus, the composition of the human expert group was more reasonable. Although many methods[14,19,27] have been further studied to improve the diagnostic ability of the DL algorithm, the lack of proof of interpretability for AI and the absence of guidance for related clinical decision-making have been a challenge for clinicians. Therefore, a clinical test was conducted to address these issues. In other words, radiologists simultaneously reviewed static images, movies, and clinical information with the aid of BreNet. A heatmap strategy was used to visualize the regions

Table 5. The time-consumption of reviewing static images and (or) dynamic videos

	Reviewing static images (s)	Reviewing dynamic videos (s)	p
Radiologists only			
All	4.09 [3.95, 4.23]	7.45 [6.55, 7.85]	NA
Senior	3.81 [3.58, 4.03]	6.85 [6.63, 7.34]	<0.05[a]
Junior	4.23 [4.05, 4.41]	7.75 [7.37, 8.12]	<0.05[a]
Radiologists with BreNet-aided			
All	4.09 [3.95, 4.23]	NA	<0.0001[b]
Senior	3.81 [3.58, 4.03]	NA	<0.0001[b]
Junior	4.23 [4.05, 4.41]	NA	<0.0001[b]

The time-consumption of BreNet for diagnosing is less than milliseconds, so we did not compute that. Collected text information with Excel in advance, the time-consumption of senior and junior radiologists was the same to skim this, so it was not listed in the table.

[a]Comparison between senior and junior radiologists in simulated clinical setting.

[b]Time of BreNet-aided process with Radiologists alone in simulated clinical setting.

of interest in the DL model. Clinical information was analyzed using the SHAP strategy. Figure 2 shows that the age at menopause age and estrogen level were the main factors related to breast cancer, which is consistent with the findings of epidemiological studies.[28] Family history is widely accepted as a high-risk factor for the incidence of breast carcinoma.[4,28] However, there was no significant correlation was observed between family history and the incidence of breast cancer in the present study.

The use of BreNet significantly enhanced the diagnostic accuracy of human experts in distinguishing breast cancer. BreNet has the potential to reduce the rate of needless biopsies and radiologists' workload. These findings support the integration of AI models like BreNet in clinical practice to enhance diagnostic efficiency and accuracy.

Limitations of the study

This study has some limitations. First, the performance of BreNet in recognizing breast cancer subtypes (invasive ductal carcinoma, invasive lobular carcinoma, and carcinoma *in situ* et al.) were not analyzed owing to the relatively small sample size as breast carcinoma subtypes are impacted by clinical decision-making. Decreasing the detection of fewer clinically meaningful carcinomas might lead to the development of a DL model that can achieve an equilibrium between both advantages and detriments.[29] Second, all patients in the external test set were selected from one hospital in South China, and the reliability and generalizability of the AI model should be further investigated in the future. Third, BreNet was developed on prospectively static images. Future studies should use dynamic videos to train the AI model and validate it in multi-center, multi-race, and multi-region settings.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Ethic statement
  - Design and dataset
  - Outcomes
- METHOD DETAILS
  - Deep learning and machine learning algorithms
  - Construction of BreNet
  - Interpretability
  - Review by six human experts for comparison
  - Model evaluation
  - Model environment
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Statistical analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110279.

ACKNOWLEDGMENTS

We would like to thank Wei Wang (the Department of Medical Ultrasonics, Ultrasonics Artificial Intelligence X-Laboratory, Institute of Diagnostic and Interventional Ultrasound, First Affiliated Hospital of Sun Yat-sen University) for consulting on the design of the study. We also thank Qiuhan Lu, MM, and Guozhi Jiang, PhD (School of Public Health, Sun Yat-sen University), for their help with the statistical analysis consulting. This study was funded by grants from the Shenzhen Science and Technology Program (no. JCYJ20220530145001002) and the Seventh Affiliated Hospital, Sun Yat-sen University Clinical Research 735 Program (no. ZSQY735202210). The study was registered at www.chictr.org.cn (no. ChiCTR2300072061).

AUTHOR CONTRIBUTIONS

Conceptualization: F.L., Z.X., Y.R., Y.L., and Q.J. Data curation: X.H., F.L., Y. Song, and T.R. Formal analysis: Y.R., F.L., Y.L., and L.H. Funding acquisition: Z.X. and Y.R. Investigation: Y. Song, T.R., F.L., Y.G., X.L., X.X., and Y. Sui. Methology: F.L. and Y.R. Project administration: F.L. and Z.X. Resources: X.H., F.L., and Z.X. Software: Y.R. and F.L. Supervision: Z.X., Y.R., and Y.L. Validation: F.L. and Y.R. Visualization: F.L. and Y.R. Writing – original draft: F.L., Y.R., and Y. Song. Writing – review & editing: Z.X. and Y.L.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 21, 2024
Revised: May 19, 2024
Accepted: June 13, 2024
Published: June 18, 2024

REFERENCES

1. Cancer Tomorrow. https://gco.iarc.fr/tomorrow/en/dataviz/trends?types=0&sexes=2&mode=cancer&group_populations=0&multiple_populations=0&multiple_cancers=1&cancers=20&populations=900&apc=cat_ca20v1.5_ca23v-1.5.
2. Tabár, L., Dean, P.B., Chen, T.H.-H., Yen, A.M.-F., Chen, S.L.-S., Fann, J.C.-Y., Chiu, S.Y.-H., Ku, M.M.-S., Wu, W.Y.-Y., Hsu, C.-Y., et al. (2019). The incidence of fatal breast cancer measures the increased effectiveness of therapy in women participating in mammography screening. Cancer *125*, 515–523. https://doi.org/10.1002/cncr.31840.
3. Stomper, P.C., D'Souza, D.J., DiNitto, P.A., and Arredondo, M.A. (1996). Analysis of parenchymal density on mammograms in 1353 women 25-79 years old. Am. J. Roentgenol. *167*, 1261–1265. https://doi.org/10.2214/ajr.167.5.8911192.
4. Fan, L., Strasser-Weippl, K., Li, J.-J., St Louis, J., Finkelstein, D.M., Yu, K.-D., Chen, W.-Q., Shao, Z.-M., and Goss, P.E. (2014). Breast cancer in China. Lancet Oncol. *15*, E279–E289. https://doi.org/10.1016/s1470-2045(13)70567-9.
5. Shen, S., and Sun, Q. (2018). Current status and suitable mode evaluation of breast carcinoma screening in chinese women. Med. J. Peking Union Med. Coll. Hosp. *9*, 298–302. https://doi.org/10.3969/j.issn.1674-9081.2018.04.003.
6. Ohuchi, N., Suzuki, A., Sobue, T., Kawai, M., Yamamoto, S., Zheng, Y.-F., Shiono, Y.N., Saito, H., Kuriyama, S., Tohno, E., et al. (2016). Sensitivity and specificity of mammography and adjunctive ultrasonography to screen for breast cancer in the Japan Strategic Anti-cancer Randomized Trial (J-START): a randomised controlled trial. Lancet *387*, 341–348. https://doi.org/10.1016/s0140-6736(15)00774-6.
7. Mann, R.M., Hooley, R., Barr, R.G., and Moy, L. (2020). Novel Approaches to Screening for Breast Cancer. Radiology *297*, 266–285. https://doi.org/10.1148/radiol.2020200172.
8. Fukushima, K. (1980). Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybernetics *36*, 193–202. https://doi.org/10.1007/bf00344251.
9. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA, J. Am. Med. Assoc. *316*, 2402–2410. https://doi.org/10.1001/jama.2016.17216.
10. McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al. (2020). International evaluation of an AI system for breast cancer screening. Nature *577*, 89–94. https://doi.org/10.1038/s41586-019-1799-6.
11. Larsen, M., Olstad, C.F., Koch, H.W., Martiniussen, M.A., Hoff, S.R., Lund-Hanssen, H., Solli, H.S., Mikalsen, K.Ø., Auensen, S., Nygård, J., et al. (2023). AI Risk Score on Screening Mammograms Preceding Breast Cancer Diagnosis. Radiology *309*, e230989. https://doi.org/10.1148/radiol.230989.
12. Hickman, S.E., Payne, N.R., Black, R.T., Huang, Y., Priest, A.N., Hudson, S., Kasmai, B., Juette, A., Nanaa, M., Aniq, M.I., et al. (2023). Mammography Breast Cancer Screening Triage Using Deep Learning: A UK Retrospective Study. Radiology *309*, e231173. https://doi.org/10.1148/radiol.231173.
13. Donnelly, J., Moffett, L., Barnett, A.J., Trivedi, H., Schwartz, F., Lo, J., and Rudin, C. (2024). AsymMirai: Interpretable Mammography-based Deep Learning Model for 1-5-year Breast Cancer Risk Prediction. Radiology *310*, e232780. https://doi.org/10.1148/radiol.232780.
14. Qian, X., Pei, J., Zheng, H., Xie, X., Yan, L., Zhang, H., Han, C., Gao, X., Zhang, H., Zheng, W., et al. (2021). Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. Nat. Biomed. Eng. *5*, 522–532. https://doi.org/10.1038/s41551-021-00711-2.
15. Raza, S., Chikarmane, S.A., Neilsen, S.S., Zorn, L.M., and Birdwell, R.L. (2008). BI-RADS 3, 4, and 5 lesions: Value of US in management - Followup and outcome. Radiology *248*, 773–781. https://doi.org/10.1148/radiol.2483071786.
16. Orel, S.G., Kay, N., Reynolds, C., and Sullivan, D.C. (1999). BI-RADS categorization as a predictor of malignancy. Radiology *211*, 845–850. https://doi.org/10.1148/radiology.211.3.r99jn31845.
17. Mendelson, E.B., Böhm-Vélez, M., Berg, W.A., Whitman, G.J., Feldman, M.I., Madjar, H., Rizzatto, G., Baker, J.A., Zuley, M., Thomas Stavros, A., et al. (2013). ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System.
18. Lazarus, E., Mainiero, M.B., Schepps, B., Koelliker, S.L., and Livingston, L.S. (2006). BI-RADS lexicon for US and mammography: Interobserver variability and positive predictive value. Radiology *239*, 385–391. https://doi.org/10.1148/radiol.2392042127.
19. Shen, Y., Shamout, F.E., Oliver, J.R., Witowski, J., Kannan, K., Park, J., Wu, N., Huddleston, C., Wolfson, S., Millet, A., et al. (2021). Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. Nat. Commun. *12*, 5645. https://doi.org/10.1038/s41467-021-26023-2.
20. Abdullah, N., Mesurolle, B., El-Khoury, M., and Kao, E. (2009). Breast Imaging Reporting and Data System Lexicon for US: Interobserver Agreement for Assessment of Breast Masses. Radiology *252*, 665–672. https://doi.org/10.1148/radiol.2523080670.
21. Webb, J.M., Adusei, S.A., Wang, Y., Samreen, N., Adler, K., Meixner, D.D., Fazzio, R.T., Fatemi, M., and Alizad, A. (2021). Comparing deep learning-based automatic segmentation of breast masses to expert interobserver variability in ultrasound imaging. Comput. Biol. Med. *139*, 104966. https://doi.org/10.1016/j.compbiomed.2021.104966.
22. Calas, M.J.G., Almeida, R.M.V.R., Gutfilen, B., and Pereira, W.C.A. (2010). Intraobserver interpretation of breast ultrasonography following the BI-RADS classification. Eur. J. Radiol. *74*, 525–528. https://doi.org/10.1016/j.ejrad.2009.04.015.
23. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell *172*, 1122–1131.e9. https://doi.org/10.1016/j.cell.2018.02.010.
24. Cheng, J.-Z., Ni, D., Chou, Y.-H., Qin, J., Tiu, C.-M., Chang, Y.-C., Huang, C.-S., Shen, D., and Chen, C.-M. (2016). Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. Sci. Rep. *6*, 24454. https://doi.org/10.1038/srep24454.
25. van Zelst, J.C., Tan, T., Mann, R.M., and Karssemeijer, N. (2020). Validation of radiologists' findings by computer-aided detection (CAD) software in breast cancer detection with automated 3D breast ultrasound: a concept study in implementation of artificial intelligence software. Acta Radiol. *61*, 312–320. https://doi.org/10.1177/0284185119858051.
26. Becker, A.S., Mueller, M., Stofel, E., Marcon, M., Ghafoor, S., and Boss, A. (2018). Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. Br. J. Radiol. *91*, 20170576. https://doi.org/10.1259/bjr.20170576.
27. Qi, X., Zhang, L., Chen, Y., Pi, Y., Chen, Y., Lv, Q., and Yi, Z. (2019). Automated diagnosis of breast ultrasonography images using deep neural networks. Med. Image Anal. *52*, 185–198. https://doi.org/10.1016/j.media.2018.12.006.
28. Sun, Y.-S., Zhao, Z., Yang, Z.-N., Xu, F., Lu, H.-J., Zhu, Z.-Y., Shi, W., Jiang, J., Yao, P.-P., and Zhu, H.-P. (2017). Risk Factors and Preventions of Breast Cancer. Int. J. Biol. Sci. *13*, 1387–1397. https://doi.org/10.7150/ijbs.21635.
29. Taylor-Phillips, S., Seedat, F., Kijauskaite, G., Marshall, J., Halligan, S., Hyde, C., Given-Wilson, R., Wilkinson, L., Denniston, A.K., Glocker, B., et al. (2022). UK National Screening Committee's approach to

reviewing evidence on artificial intelligence in breast cancer screening. Lancet Digit. Health *4*, E558–E565. https://doi.org/10.1016/S2589-7500(22)00088-7.

30. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vis. *115*, 211–252. https://doi.org/10.1007/s11263-015-0816-y.
31. Lundberg, S.M., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Adv. Neural Info. Process.Syst *30*, 1.
32. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int. J. Comput. Vis. *128*, 336–359. https://doi.org/10.1007/s11263-019-01228-7.
33. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). TensorFlow: A System for Large-Scale Machine Learning. In 12th USENIX symposium on operating systems design and implementation (OSDI 16), pp. 265–283.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Python	Version 3.7.8	https://www.python.org/downloads/release/python-378/
ImageNet	Russakovsky et al.[30]	https://www.image-net.org/
SHapley Additive exPlanations	Lundberg et al.[31]	https://github.com/shap
Grad-CAM	Selvaraju et al.[32]	https://keras.io/examples/vision/grad_cam/
Tensorflow	Martín Abadi et al.[33]	https://www.tensorflow.org/
Breast cancer diagnosis network (BreNet)	Code for this study	https://github.com/koalary/breast_ultrasound_images/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Zuofeng Xu (xuzuofeng77@aliyun.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The ultrasonic data reported in this study cannot be deposited in a public repository because of the hospital regulation restrictions and patient privacy concerns. Anonymized data reported in this paper will be shared by the lead contact upon reasonable request.
- All original code has been deposited at https://github.com/koalary/breast_ultrasound_images and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Ethic statement

The survey was conducted in accordance with the Declaration of Helsinki (revised in 2013). This bi-centre prospective study was approved by the Institutional Review Board of the Seventh Affiliated Hospital, Sun Yat-sen University (No. KY-2020–075-01). Informed consent was obtained from all participants.

Design and dataset

This bi-centre, prospective, diagnostic investigation used US images, clinical information, and video datasets of consecutive patients (Han Chinese and of East Asian Decent) from two hospitals in South China, from August 2020 to June 2023. All masses were pathologically diagnosed. The inclusion criteria were as follows: women aged $\geq$18 years; tumours of the American College of Radiology Breast Imaging Reporting and Data System (ACR BI-RADS) categories II–V, among which categories III–V had pathologically-confirmed diagnosis. ACR BI-RADS category II includes breast cysts confirmed using US. The exclusion criteria were as follows: pregnant or lactating women; history of previous breast surgery; diffuse lesions; large artefacts or poor image resolution; and $\geq$3 tumours in the ipsilateral breast, with pathological results that could not be distinguished among the lumps. The time between US examinations and biopsy (or resection) was managed within 2 weeks, without any clinical intervention throughout the period. The sample size of the study consulted prior investigation.[14]

Outcomes

The main indicator used in this study was the area under the curve (AUC) for the diagnosis of breast tumours. The minor indicators were accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for the diagnosis of breast neoplasms. The post-hoc analysis included the diagnostic accuracy of the BreNet-aided biopsy strategy and the time required by the radiologist for diagnosing the tumour.

METHOD DETAILS

Deep learning and machine learning algorithms

A CNN was used to model and extract the features from static US images. The transverse, longitudinal, and colour Doppler modes were trained separately using models based on EfficientNet-B5. Before training, each model was initialised using the fully trained parameters from the ImageNet dataset,[30] and the final output of the model was modified to two neurons to match the benign and malignant dichotomous tasks. The optimiser used in the model was Adam with a learning rate of 0·0008. The last-layer activation and loss functions were set to sigmoid and binary cross-entropy, respectively. After beginning the training, the corresponding models with the highest accuracy were saved in the 100 epochs training cycle, and three ultrasonic image models corresponding to the three modes were obtained.

A machine-learning algorithm was used to model and extract clinical features. The data was pre-processed by filling-in some of the vacant values, and standardising the data to a range of 0–1 before training. A random forest model was subsequently used for modelling and analysis, and the importance of the model features was sorted using the SHapley Additive exPlanations (shap, MIT) visualisation method.[31]

Construction of BreNet

Different modes of US images are used to describe different visual angles of the lesions. Three CNN models were created for the transverse, longitudinal, and colour Doppler views, and each model was used to predict the probability of benign and malignant tumours in this mode. The imaging prediction probabilities of the transverse, longitudinal, and colour Doppler features were obtained using the average output probabilities of the three CNN models. A final prediction probability that combines the three types of images and clinical information was obtained by averaging the prediction probability of the imaging model and the clinical information model. This polymer model was named BreNet.

Interpretability

Visual gradient-weighted class activation mapping (Grad-CAM) technology[32] was used to generate heat maps that highlighted the areas of each US image for classification. This step may increase interpretability to some extent and may be useful as an approximate visual diagnosis for presentation to radiologists. The most important clinical features for benign and malignant prediction were obtained by ranking the importance of the model features using the SHapley Additive exPlanations (shap, MIT) visualisation method.

Review by six human experts for comparison

Six radiologists (two senior and four junior radiologists) participated in this study to evaluate the level of BreNet-assisted diagnosis. The tests were divided into three stages (Figure 1). First, the diagnostic ability of the human experts with the aid of BreNet for the assessment of static US images were compared. Second, the diagnostic abilities of the human experts alone and with the aid of BreNet were evaluated. Third, the first judgment based on static images and the second judgment based on videos were recorded. The diagnoses of real-world radiologists (simultaneous reading of static images, videos, and clinical information) were subsequently compared with those of the BreNet-aided radiologists (only reading static images and clinical information). The radiologists who participated in the test were blinded to the other participants' diagnoses, and made the diagnosis independently. Each stage progressed at an interval of more than one month.

Model evaluation

The performances of models were evaluated using confusion matrix, positive predictive value (precision), sensitivity (recall), F1 score, Fleiss kappa coefficient, the receiver operating characteristic (ROC) curve and AUC. The F1 score is the harmonic mean of precision and recall. Only when both precision and recall are high will F1 be high. Kappa coefficient is used to check consistency and can also be used to measure the accuracy of classification, but the calculation of kappa coefficient is based on confusion matrix. The AUC typically varies from 0.5 to 1, with 1 indicating perfect classifier performance and 0.5 indicating random performance.

The US images and clinical information of all 630 patients included in the study were collected from the seventh affiliated hospital, Sun Yat-sen University. Ultrasonic images include three views: transverse, longitudinal, and color Doppler view. The clinical information includes 13 clinical features, such as body mass index, family history, menarche age, menopause age, estrogen and so on. The final diagnosis of all patients is determined by pathology. Before modeling, the data at the patient level are randomly divided into training dataset and internal test dataset according to the proportion of 7:3. In addition, according to the same standards and requirements, we collected the ultrasonic images, videos, and clinical information of 100 patients in Dongguan Tungwah Hospital as external dataset (see also Figure S1).

In clinic practice of breast mass managing, a vital decision after ACR BI-RADS classifying is whether succeeding biopsy is indicated. Considering ACR BI-RADS, masses that categories 2 or 3 do not require biopsy, in which case the possibility of malignancy (NPV) is low ($\leq$2%). While masses with a classification of 4c or more had a great possibility of being malignant and a PPV of pathological biopsy is testified to be between 50–95%.[17] If the BreNet-aided scheme can obtain a comparable PPV as biopsy, the biopsy may be avoided. To mitigate the clinical encumbrance caused by increasing need of biopsy, we virtual the clinical setting by omitting biopsy for masses at a great possibility of being benign (NPV) or of high malignant risk (PPV), according to radiologist ACR BI-RADS and BreNet-aided diagnosis.

We analyzed the NPV of lumps with ACR BI-RADS categories 2 and 3. We also assessed the NPV of lumps with ACR BI-RADS classification 4, diagnosed as benign by the BreNet-aided. We anticipated the classification can be enlarged without a loss of NPV compared to ACR BI-RADS classification 2 and 3; consequently, biopsy can be avoided for lumps adjudicated as benign within this enlarged classification. For

category 5 masses, we thought that surgical resection should be performed without biopsy. For the rest neoplasms, medical suggestion was followed to the ACR BI-RADS guide.

In addition, we also computed another secondary indicator which is the time consumption of radiologist for diagnosis. We hypothesized that human-experts spend less time for reading static images with BreNet-aided than the total time for reviewing both static images, videos, and text information without BreNet-aided, and the diagnostic performance of the former is not inferior to the latter. It will illustrate that AI-assisted human diagnosis can effectively reduce the workload of radiologist without sacrificing diagnostic performance.

Model environment

To train and evaluate our models, we used an Ubuntu 18.04 computer (Canonical, Ltd, London, United Kingdom) using TensorFlow2.0 (Google Inc)[33] backed within Python 3.7 language. The hardware components are as follows: a Tesla V100 GPU with 32-GB memory on one DGX1 server (Intel Xeon E5-2698 v4 at 2.20 GHz, 512 GB memory, and 7-TB hard drive; NVIDIA).

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis

The data were analysed using IBM SPSS Statistics package (version 26·0). Descriptive statistics were described as frequencies and percentages for classified variables and means and standard deviations for continuous arguments. Differences among means were evaluated using the t-test and amid binary categorical variables using the chi-square test. Values are given as n(%) or median (25^{th}, 75^{th} percentile) or mean $\pm$ SD. A receiver operating characteristic curve was created for each testing set by drafting the true positive rate against the true negative rate and varying the predicted probability threshold. The AUC values were computed, using roc function from pROC package (version 1.18.0). The Delong's test was used to compare the AUCs between the BreNet and human experts. A two-by-two confusion matrix with the numbers of true positives, false positives, false negatives, and true negatives was created for each diagnosis. The accuracy, sensitivity, specificity, PPV, and NPV were computed according to the confusion matrix for breast carcinoma detection, using epiR package (version 2.0.63). Additional analyses confirmed BreNet's superior performance in both internal and external datasets. The inter-radiologists agreement and Fleiss' κ value were estimated for each testing set using the kappam.fleiss function, R software (version 4·2·3) to assess the agreement between the diagnosis of malignant and benign tumours. The intraclass correlation coefficient was estimated using the icc function of the irr package (version 4·2·3) to estimate the agreement between the six radiologists concerning the six ACR BI-RADS classifications. $P<0·05$ was considered to indicate a statistically significant difference.