Artificial intelligence-assisted ultrasound image analysis to discriminate early breast cancer in Chinese population: a retrospective, multicentre, cohort study

Jianwei Liao,[a,b,h] Yu Gui,[a,h] Zhilin Li,[b,h] Zijian Deng,[b,h] Xianfeng Han,[b] Huanhuan Tian,[b] Li Cai,[b] Xingyu Liu,[b] Chengyong Tang,[b] Jia Liu,[c] Ya Wei,[d] Lan Hu,[e] Fengling Niu,[f] Jing Liu,[a] Xi Yang,[a] Shichao Li,[a] Xiang Cui,[a] Xin Wu,[a] Qingqiu Chen,[a] Andi Wan,[a] Jun Jiang,[a] Yi Zhang,[a] Xiangdong Luo,[a] Peng Wang,[g,i,***] Zhigang Cai,[b,i,**] and Li Chen[a,i,*]

[a]Department of Breast and Thyroid Surgery, Southwest Hospital of Third Military Medical University, Chongqing, 40038, China
[b]College of Computer and Information Science, Southwest University, Chongqing, 400715, China
[c]Department of Gastroenterology, The First Affiliated Hospital (Southwest Hospital) of Third Military Medical University (Army Medical University), Chongqing, 40038, China
[d]The Third Department of General Surgery, Anyang Cancer Hospital, Henan, 455001, China
[e]Department of General Surgery, The People's Hospital of Dazu, Chongqing, 402360, China
[f]Breast Surgery Department, Tangshan People's Hospital, Tangshan, 063001, China
[g]Centre for Medical Big Data and Artificial Intelligence, Southwest Hospital of Third Military Medical University, Chongqing, 400038, China

Summary

Background Early diagnosis of breast cancer has always been a difficult clinical challenge. We developed a deep-learning model EDL-BC to discriminate early breast cancer with ultrasound (US) benign findings. This study aimed to investigate how the EDL-BC model could help radiologists improve the detection rate of early breast cancer while reducing misdiagnosis.

Methods In this retrospective, multicentre cohort study, we developed an ensemble deep learning model called EDL-BC based on deep convolutional neural networks. The EDL-BC model was trained and internally validated on B-mode and color Doppler US image of 7955 lesions from 6795 patients between January 1, 2015 and December 31, 2021 in the First Affiliated Hospital of Army Medical University (SW), Chongqing, China. The model was assessed by internal and external validations, and outperformed radiologists. The model performance was validated in two independent external validation cohorts included 448 lesions from 391 patients between January 1 to December 31, 2021 in the Tangshan People's Hospital (TS), Chongqing, China, and 245 lesions from 235 patients between January 1 to December 31, 2021 in the Dazu People's Hospital (DZ), Chongqing, China. All lesions in the training and total validation cohort were US benign findings during screening and biopsy-confirmed malignant, benign, and benign with 3-year follow-up records. Six radiologists performed the clinical diagnostic performance of EDL-BC, and six radiologists independently reviewed the retrospective datasets on a web-based rating platform.

Findings The area under the receiver operating characteristic curve (AUC) of the internal validation cohort and two independent external validation cohorts for EDL-BC was 0.950 (95% confidence interval [CI]: 0.909–0.969), 0.956 (95% [CI]: 0.939–0.971), and 0.907 (95% [CI]: 0.877–0.938), respectively. The sensitivity values were 94.4% (95% [CI]: 72.7%–99.9%), 100% (95% [CI]: 69.2%–100%), and 80% (95% [CI]: 28.4%–99.5%), respectively, at 0.76. The AUC for accurate diagnosis of EDL-BC (0.945 [95% [CI]: 0.933–0.965]) and radiologists with artificial intelligence (AI) assistance (0.899 [95% [CI]: 0.883–0.913]) was significantly higher than that of the radiologists without AI assistance (0.716 [95% [CI]: 0.693–0.738]; p < 0.0001). Furthermore, there were no significant differences between the EDL-BC model and radiologists with AI assistance (p = 0.099).

*Corresponding author. Department of Breast and Thyroid Surgery, Southwest Hospital of Third Military Medical University, Chongqing, China.
**Corresponding author. College of Computer and Information Science, Southwest University, Chongqing, China.
***Corresponding author. Centre for Medical Big Data and Artificial Intelligence, Southwest Hospital of Third Military Medical University, Chongqing, China.
    E-mail addresses: chenli@tmmu.edu.cn (L. Chen), czg@swu.edu.cn (Z. Cai), wpxnyy@tmmu.edu.cn (P. Wang).
[h]These authors are considered co–first authors.
[i]These authors are considered corresponding authors.
Translation For the Chinese translation of the Summary, see Supplementary Materials section.

Interpretation EDL-BC can identify subtle but informative elements on US images of breast lesions and can significantly improve radiologists' diagnostic performance for identifying patients with early breast cancer and benefiting the clinical practice.

Funding The National Key R&D Program of China.

Keywords: Artificial intelligence; Ultrasound; Early breast cancer

---

Research in context

Evidence before this study

We searched PubMed with the terms "(deep learning OR machine learning OR artificial intelligence) AND (ultrasound OR ultrasonography) AND early breast cancer" for papers published from database inception up to Dec 31, 2021, with no language restrictions or date restrictions. We found that previous researches were mainly limited to deep learning-based classification of breast lesions, prediction axillary lymph node status in early-stage breast cancer and early prediction of response to neoadjuvant chemotherapy in breast cancer. None of these studies discriminated early breast cancer with ultrasound (US) benign findings.

Added value of this study

We developed an ensemble deep learning model, called EDL-BC to identify high-risk breast lesions in US images. The model

diagnostic performance was verified by multi-centre cohorts. Both in the internal validation cohort and two external validation datasets, our model showed high sensitivity and specificity. To further demonstrate its practicability, six radiologists independently reviewed the retrospective datasets with or without assistance from EDL-BC. The results showed that EDL-BC improved the performance of radiologists when reviewing US images.

Implications of all the available evidence

EDL-BC could identify high-risk lesions in US images with benign findings, which is beneficial to early confirmation with immediate biopsy. Therefore, it can significantly improve the diagnostic performance of radiologists to identify breast cancer at an early stage.

Introduction

Globally, breast cancer has become one of the most common cancers and accounts for one-quarter of all new cancers in female patients. It is also a leading cause of cancer death in women around the world.[1] According to the eighth edition of the American Joint Committee on Cancer staging system, early-stage breast cancer refers to breast cancer stagesIandII,T1-2N0-1M0.[2] Early detection of breast cancer can reduce the corresponding mortality by 40%.[3] Mammography and ultrasound (US) are suitable for addressing this task. However, for women with dense breast tissue, the sensitivity of mammography decreases from 85% to 48–64%. Namely, it is not suitable for all ethnic groups.[4] On the other hand, US can cope with the clinical cases for women with dense breasts. Meanwhile, Chinese women tend to have more dense breasts than Caucasian and Hawaiian groups, which reduces the accuracy of mammography in diagnosis.[4,5] The US is the primary instrument for early breast cancer screening in China due to its convenience, price moderate, non-invasiveness, low radiation, and universality. Accordingly, the US has been applied in differentiation of cysts from solid masses and breast cancer screening.[4,6] Reviews of US images with diagnosed breast cancer have

demonstrated that cancer indicators are visible on merely 10.8% of early examinations interpreted as normal.[7] A variety of factors in the US images related to this phenomenon, including noise, contrast, illumination, and resolution.[8] Until now, a reliable US image processing method still needs to be discovered, especially for an early diagnosis of challenging breast cancer cases with multiple types of modalities. Therefore, this is a meaningful work that the malignant lesions were accurately identified at an earlier stage after analysing diagnosed-benign US lesions.

Artificial intelligence (AI) has made an outstanding contribution to a plethora of clinical challenges in oncology, including tumor diagnosis, treatment, and prognosis.[9] Deep learning (DL) is a subset of Machine learning (ML), which is a subfield of AI. Recently, Since DL is capable of performing feature extraction automatically and dealing with big data, it has achieved unprecedented success in various areas, including image classification, natural language processing, audio recognition, and video analysis.[10,11] DL-assisted image analysis plays a gradually increased role in early cancer detection while enhancing diagnosis accuracy by reducing false positives.[11] DL has also achieved promising outcome in the diagnosis of various tumors,

including liver cancer, breast cancer, colorectal cancer, prostate cancer, non-small cell lung cancer and naso-pharyngeal carcinoma using mammogram, US, computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography-CT (PET-CT).[12–17] Previous studies of DL models have shown their comparable performance to doctors in cancer detection, cancer classification, and cancer grading. By integrating feature extraction and fine-tuning in an automated fashion, DL-based algorithms can guarantee the performance of cancer diagnosis while simplifying the diagnostic workflow and reducing mistakes.[11]

The application of US in breast cancer screening is increasing. Correspondingly, a fast and accurate diagnosis has become a significant problem, which has influence on the entire therapeutic process. DL has made significant progress in analysing medical images, while manual inspection of US images is error-prone, laborious and time-consuming.[18] For instance, Fujioka et al. introduced the GoogLeNet for lesion detection and classification in US images.[19] DL has shown its effectiveness in breast imaging and treatment response of breast cancer. And DL-based diagnostic platforms are becoming more common.[20,21] It has shown the crucial role of DL-based system in classifying benign and malignant breast images.[22]

However, few studies have focused on DL-based early breast cancer detection in the US images. Bearing the above-mentioned analysis in mind, we selected 7421 patients from January 1, 2015 and December 31, 2021 with US benign and 219 patients with breast cancer confirmed. We hypothesized subtle elements within US images that may not be discernible by human visual inspection or revealed by straightforward measurements,[23] and DL can leverage these informative representations to estimate the malignant degree.[6] We developed an ensemble DL model to identify high-risk lesions in US images, which is beneficial to early confirmation with immediate biopsy. The proposed model contains two separate feature extraction modules designed for B-mode and color Doppler US images, respectively. Rather than the manually-crafted embeddings, we employed the DL model to unveil more discriminative representation from US image samples.

This retrospective cohort study was performed on images from patients with early breast cancer with benign US images at Southwest (SW) Hospital (Chongqing, southwestern China; 70% of patients are from Sichuan, Guizhou and Yunnan provinces; 30% are from Chongqing), the Tangshan (TS) People's Hospital (Hebei Province, northern China), and the Dazu (DZ) People's Hospital (Chongqing, southwestern China). We initially trained the proposed model and evaluated its performance on the dataset from SW, split into training and validation sets. We then evaluated the generalizability and robustness of our proposed model using the TS and DZ datasets simultaneously.

Methods

Study design and participants

For SW database, due to repeated visits, male, or declined to allow the use of their medical records for research, we excluded the samples from 338,282 patients. In the remaining 84,016 female patients with US benign lesions, 76,605 (91.2%) patients were suspected breast cancer on mammogram and did not undergo biopsy confirmation or three years of follow-up, 118 (1.6%) patients were <18 years or age unknown, 31 (0.4%) had malignant phyllodes tumors, various malignancies of the non-mammary origin or were pregnant, 231 (3.2%) were diagnosed with evidence of the personal history of breast cancer or ductal carcinoma in situ with clinical signs of a non-mass accompanied by bloody nipple discharge, 172 (2.4%) did not have B-mode or color Doppler images available, and 64 (0.9%) patients had images of poor quality (Fig. 1). The SW database, on which the EDL-BC model was trained and internally validated, consisted of B-mode and color Doppler US images of 7955 lesions from 6795 patients with US benign findings between January 1, 2015 and December 31, 2021 (Fig. 2, Table 1). The doctors involved had an averaged 20 years of clinical experience in breast US imaging. We selected 1–2 lesions from each patient as model input, and each lesion included 2 images. In each patient's US examination, there were 2–4 (mean, 2.3) images collected. In our paper, the training set, internal test set, Tangshan test set and Dazu test set contained 7076 lesions (14152 images), 879 (1758 images), 448 (896 images) and 245 (490 images) lesions in total, respectively. In the SW dataset, the training set consisted of biopsy-confirmed malignant (n = 186), benign lesions (n = 4824; 2015–2021), and benign lesions with 3-year follow-up records (n = 2066; 2015–2017). Meanwhile, the validation set consisted of biopsy-confirmed malignant (n = 18), benign lesions (n = 440; 2021), and benign lesions with 3-year follow-up records (n = 421; 2018). The datasets used to evaluate the generalizability of the proposed approach were selected from 9227 patients in US benign lesions. They contained 448 lesions (10 malignant, 438 benign) from 391 patients within TS between January 1 and December 31, 2021 and 245 lesions (5 malignant, 240 benign) from 235 patients within DZ between January 1 and December 31, 2021.

This study aimed to develop a DL model to recognize the breast cancer risk at early stages in US images. The SW dataset on which the model was trained and internally validated consisted of 7955 US benign lesions from 6795 patients. The generalization ability of the proposed model was evaluated with two datasets from TS (448 lesions from 391 patients) and DZ (245 lesions
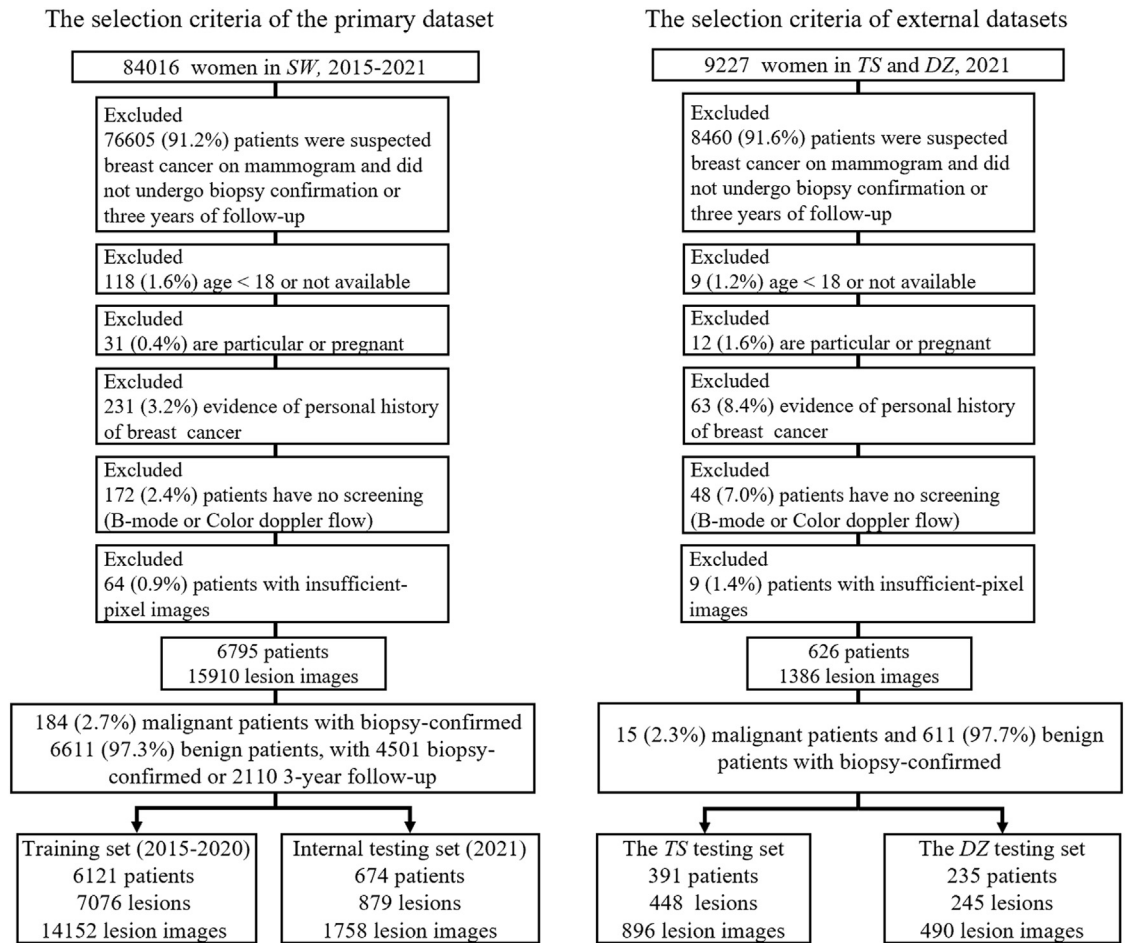
The selection criteria of the primary dataset

84016 women in *SW*, 2015-2021

Excluded
76605 (91.2%) patients were suspected breast cancer on mammogram and did not undergo biopsy confirmation or three years of follow-up

Excluded
118 (1.6%) age < 18 or not available

Excluded
31 (0.4%) are particular or pregnant

Excluded
231 (3.2%) evidence of personal history of breast cancer

Excluded
172 (2.4%) patients have no screening (B-mode or Color doppler flow)

Excluded
64 (0.9%) patients with insufficient-pixel images

6795 patients
15910 lesion images

184 (2.7%) malignant patients with biopsy-confirmed
6611 (97.3%) benign patients, with 4501 biopsy-confirmed or 2110 3-year follow-up

Training set (2015-2020)
6121 patients
7076 lesions
14152 lesion images

Internal testing set (2021)
674 patients
879 lesions
1758 lesion images

The selection criteria of external datasets

9227 women in *TS* and *DZ*, 2021

Excluded
8460 (91.6%) patients were suspected breast cancer on mammogram and did not undergo biopsy confirmation or three years of follow-up

Excluded
9 (1.2%) age < 18 or not available

Excluded
12 (1.6%) are particular or pregnant

Excluded
63 (8.4%) evidence of personal history of breast cancer

Excluded
48 (7.0%) patients have no screening (B-mode or Color doppler flow)

Excluded
9 (1.4%) patients with insufficient-pixel images

626 patients
1386 lesion images

15 (2.3%) malignant patients and 611 (97.7%) benign patients with biopsy-confirmed

The *TS* testing set
391 patients
448 lesions
896 lesion images

The *DZ* testing set
235 patients
245 lesions
490 lesion images

Fig. 1: **Flow chart of patients' selection.** This study was approved by the institutional review board with a waiver requiring the informed consent and was compliant with the Health Insurance Portability and Accountability Act. Due to the characteristics of retrospective investigation, multimodal ultrasound (US) images are not completely preserved and/or annotations are not integrally labeled in some lesions. To employ the US image datasets from three medical centres, our EDL-BC system was developed and internally validated based on US image.

from 235 patients). All datasets were constructed to reflect the variability across geographically distributed medical centres for different ethnic groups and a variety of US devices, contributing to image variability.[24]

The details on data preprocessing are described in the appendix (p1). All image pairs (color Doppler overlaid on corresponding B-mode images) showed one-to-one correspondence. The proposed model consisted of several fundamental modules, which used structurally identical backbones to unveil underlying features in images, respectively. All breast cancer cases were confirmed by pathological biopsy, and all benign lesions were confirmed by pathological biopsy or 3-year follow-up. The details of the datasets are provided in Fig. 3. The model construction, procedures, and interpretability are described in the appendix (p1–p2).

This study was approved by the ethics committee of the First Affiliated Hospital of Army Medical University [No. (B) KY202264]. The requirement for patient informed consent was waived. The CONSORT-AI guideline was used for reporting in the current study.

Model aided diagnosis

Six radiologists were ignorant of the pathological confirmation of the breast mass status and research aims before the review. Each radiologist independently completed a review process on an online platform. In addition, the radiologists' diagnosis was compared with a reference diagnosis from EDL-BC. If the two results are inconsistent, the radiologists could then choose to adhere to their own diagnosis or adopt the diagnosis from EDL-BC. The final diagnosis was generated from the assistance of EDL-BC.

Statistical analysis

Categorical variables were presented as frequency (percentage). Continuous variables were presented as mean ± standard deviation, and between-group
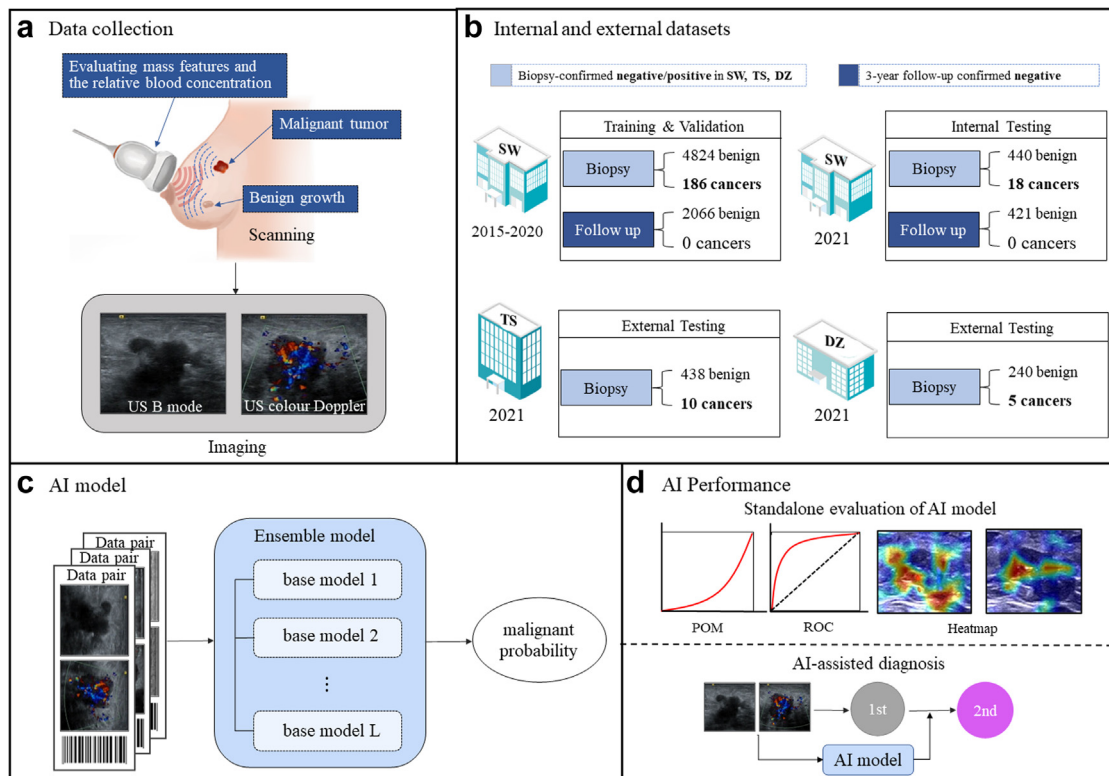
Fig. 2: Data and strategy. a. Conventional ultrasound (US)-based diagnosis, and the patients with US benign findings are generally managed with a short-interval of 6 months follow-up or continued surveillance. **b.** Summary of training and validation datasets. The entire dataset consists of biopsy-confirmed positive/negative samples collected from three medical centres of Southwest (SW), Tangshan (TS), and Dazu (DZ), as well as 3-year follow-up negative samples from SW. **c.** Ensemble model. **d**. AI Performance.

differences were assessed using variance (ANOVA). Models were evaluated based on the following metrics from the five-fold cross validation: the area under the receiver operating characteristic curve (AUC), F-measure to receiver operating characteristic (ROC), and values for kappa accuracy, sensitivity, and specificity. We calculated the following performance evaluation metrics with a 95% confidence interval (CI) using 1000 bootstraps and P < 0.05 against the null hypothesis: (1) Sensitivity (true positive rate), representing the proportion of samples with breast cancer correctly identified as breast cancer: sensitivity = true positive/(true positive + false negative); (2) Specificity (true negative rate), representing the proportion of samples without breast cancer correctly identified as non-breast cancer: specificity = true negative/(true negative + false positive); (3) AUC using to evaluate the efficiency of the proposed model. EDL-BC will generate predictions of the probability of malignancy (POM). That is, for each of the patient's breasts, the system produces a number in a range between 0 and 1. Two-sided P values less than 0.05 were considered indicative of statistical significance. All statistical analyses were performed using the statistical package SPSS (version 19.0, SPSS, Chicago).

Role of the funding source
The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. LH and FN had access to their respective institution's data. The corresponding author had final responsibility for the decision to submit for publication.

Results
Model performance and interpretability
The EDL-BC model was trained and validated using 14,152 B-mode and color Doppler US images of 7076 lesions. The validation set consisted of 3144 images of 1572 lesions from 1300 patients within three hospitals (SW:674TS:391 DZ:235) and was used to evaluate the EDL-BC model in clinical practice. The demographic characteristics of the patients in these three datasets are provided in Table 1. A summary of the US devices used is provided in supplementary Table S1.

We evaluated the performance of the proposed system by examining AUC) (supplementary Fig. S1 and supplementary Table S2). In the base learner of EDL-BC, the combination of two types of images (B-mode or color Doppler) played an important role in diagnosis

Specifications	Primary dataset (SW)		External datasets	
	Train & validate	Test	TS	DZ
Patients (7421 patients from 3 centers, 199 patients developed cancers)				
Age				
<30	2741 (44.8%)	163 (24.2%)	105 (26.9%)	62 (26.4%)
30–49	3080 (50.3%)	418 (62.0%)	233 (59.6%)	159 (67.6%)
50–69	292 (4.8%)	93 (13.8%)	52 (13.3%)	14 (6%)
≥70	6 (0.1%)	0	1 (0.2%)	0
Diagnostic methods				
Biopsy	4334 (70.8%)	351 (52.1%)	391 (100%)	235 (100%)
Follow-up	1787 (29.2%)	323 (47.9%)	0	0
Malignant type (Patient result)				
Luminal A	78 (46.2%)	6 (40%)	4 (40%)	2 (40%)
Luminal B	30 (17.8%)	2 (13.3%)	3 (30%)	1 (20%)
HER2+	25 (14.8%)	2 (13.3%)	1 (10%)	1 (20%)
TNBC	28 (16.5%)	4 (26.7%)	2 (20%)	1 (20%)
None	8 (4.7%)	1 (6.7%)	0	0
N stage				
N0	148 (87.6%)	14 (93%)	8 (80%)	4 (80%)
N1	21 (12.4%)	1 (7%)	2 (20%)	1 (20%)
TNM stage				
1	88 (52.1%)	11 (73.3%)	4 (40%)	3 (60%)
2	81 (47.9%)	4 (26.7%)	6 (60%)	2 (40%)
Category of Ultrasound				
symptomatic	1028 (16.8)	115 (17.1%)	52 (13.3%)	56 (23.8%)
screen-detected	5093 (83.2%)	559 (82.9%)	339 (86.7%)	179 (76.2%)
Lesions (8648 lesions from 7421 patients, 219 lesions are malignant)				
Lesions size				
<5	285 (4.0%)	116 (13.2%)	17 (3.8%)	12 (4.9%)
5–9.9	1682 (23.8%)	508 (57.8%)	143 (31.9%)	72 (29.4%)
10–19.9	3405 (48.1%)	243 (27.6%)	223 (49.8%)	108 (44.1%)
≥20	1704 (24.1%)	12 (1.4%)	65 (14.5%)	53 (21.6%)
Lesions width				
<5	1984 (28.0%)	534 (60.8%)	124 (27.7%)	73 (29.8%)
5–9.9	3662 (51.8%)	313 (35.6%)	228 (50.9%)	113 (46.1%)
10–19.9	1350 (19.1%)	30 (3.4%)	94 (21.0%)	58 (23.7%)
≥20	80 (1.1%)	2 (0.2%)	2 (0.4%)	1 (0.4%)
Aspect ratio				
≥1	108 (1.5%)	6 (0.7%)	5 (1.1%)	2 (0.8%)
<1	6968 (98.5%)	873 (99.3%)	443 (98.9%)	243 (99.2%)
Boundary				
Clear	6106 (86.3%)	820 (93.3%)	376 (83.9%)	190 (77.6%)
Others	970 (13.7%)	59 (6.7%)	72 (16.1%)	55 (22.4%)
Morphology				
Regular	6346 (89.7%)	848 (96.5%)	396 (88.4%)	200 (81.6%)
Others	730 (10.3%)	31 (3.5%)	52 (11.6%)	45 (18.4%)
Blood Flow Spectrum				
Pulsating	115 (1.6%)	0	5 (1.1%)	3 (1.2%)
Others	6961 (98.4%)	879 (100%)	443 (98.9%)	242 (98.8%)
Mammography				
Lesions can be seen	3399 (48.1%)	202 (22.9%)	181 (40.3%)	103 (42.1%)
Occult	3677 (51.9%)	677 (77.1%)	267 (59.6%)	142 (57.9%)

The primary dataset consisting of the training dataset and the internal dataset, was collected by Southwest Hospital of China. Another two external datasets were collected by the Tangshan People's Hospital (TS, located in Hebei Province, northern of China) and the Dazu People's Hospital (DZ, located in Chongqing), and used for evaluating the proposed EDL-BR3 system. In the table, lesion information was determined using existing screening and diagnostic reports. Note that the training dataset of the primary dataset of SW includes 2015–2020 biopsy-confirmed lesions and 2015–2017 follow-up confirmed lesions, and shows not very consistent features to the internal testing dataset and two external testing datasets. HER2: human epidermal growth factor receptor 2; TNBC: triple negative breast cancer; LN: lymph node; TNM: tumor node metastasis.

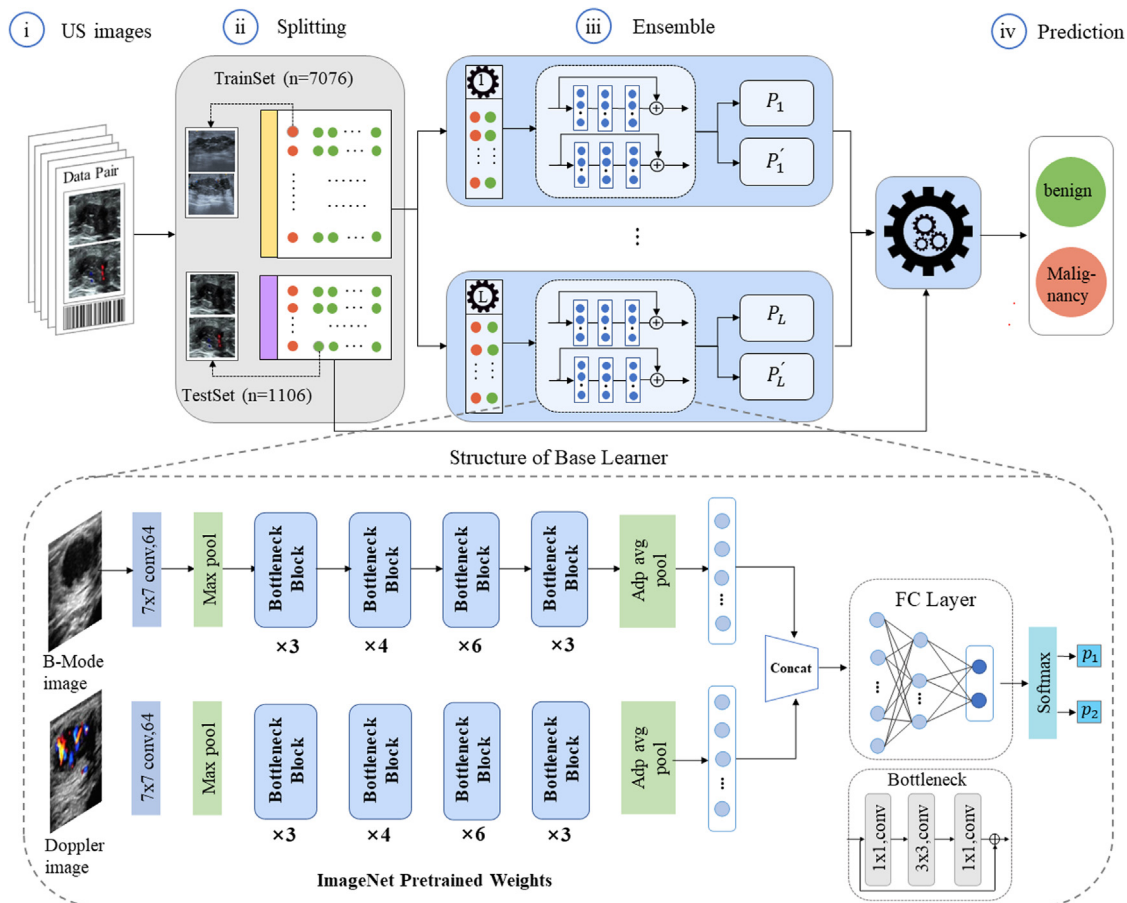*Table 1:* Detailed patients and breast lesion characteristics statistics.

Fig. 3: Details of the proposed model. The architecture of our designed base learner, namely Multiple Source Feature Learning Model (MSFLM), where we used two ResNet50 models for dealing with the B-mode US image and the Doppler image, respectively. ResNet50 is a variant of ResNet model which has 48 convolutional layers, one max-pooling, and one average-pooling layer.

performance (supplementary Fig. S2 and supplementary Table S3). We adopted the EDL-EC model with 100 base learners as a baseline to compare with 10, 20, 40, 60, and 80 base learners. The proposed model with 80 base learners achieved an optimal trade-off between performance and overhead. Thus, we leveraged 80 base learners in the proposed DL framework (supplementary Fig. S3). The AUC for the proposed model was significantly better than the average AUC for the 80 base learners in both internal dataset and external datasets (P < 0.05; Fig. 4, supplementary Figs. S3 and S4). The proposed model achieved an AUC of (0.950; 95% CI, 0.909–0.969) on the internal test set, an AUC of (0.956; 95% CI, 0.939–0.971) and an AUC of (0.907; 95% CI, 0.877–0.938) on the external validation set, respectively (Fig. 4a, d, g).

The EDL-BC model showed promising accuracy in identifying malignant lesions in the internal dataset and two external datasets (Fig. 4b, e, h). The sensitivity is of 94.4% from 18 cases for breast cancer in the internal

dataset, and the sensitivity is 80% in the DZ dataset. In these cases, cancer was confirmed by pathological analysis. Specifically, the malignancy detection rates were respectively 94.4% (95% [CI]: 72.7%–99.9%), 100% (95% [CI]: 69.2%–100%), and 80% (95% [CI]: 28.4%–99.5%) in the internal testset and two external testsets with the threshold of 0.76.

Furthermore, we analysed the distribution of predicted probabilities of malignancy (POM), which further confirmed the performance of the proposed model in discrimination (Fig. 4c, f, i). Totally, there were 199 cases of early breast cancer in the three datasets, the number of Luminal A, Luminal B, Human epidermal growth factor receptor (HER) 2 +, and triple-negative breast cancer (TNBC) were 90 (45.23%), 36 (18.09%), 29 (14.57%), and 35 (17.59%), respectively (Table 1). There were 30 malignant cases in the internal validation cohort and two external validation datasets. We classified these cases into different types of moleculars. The number of cases with luminal A, luminal B,
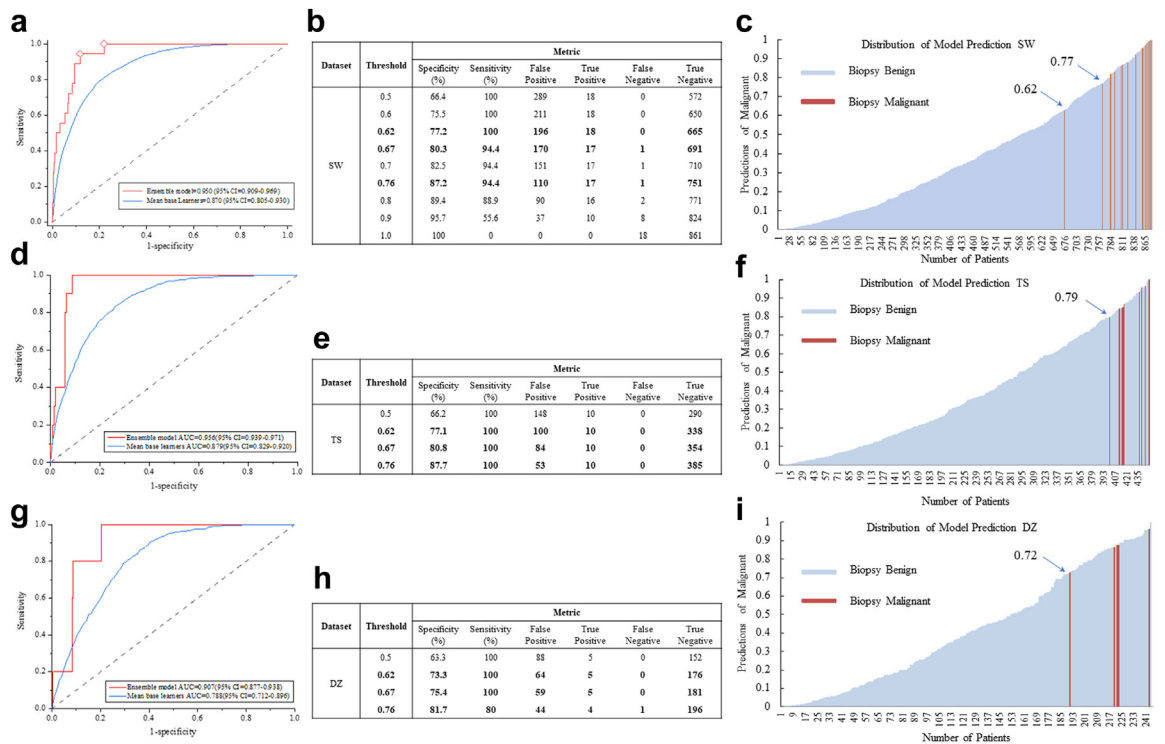
Fig. 4: **Performance of EDL-BC and radiologists in the discrimination of malignant from benign breast lesions. a**. Receiver operating characteristic (ROC) curves for the binary classification of breast lesions. The area under the receiver operating characteristic curve (AUC) scores are reported together with the 95% confidence interval (CI) is on the internal validation set. **b**. Sensitivity, specificity, false positive, true positive, false negative, and true positive evaluated on the internal validation set. **c.** The distribution of predicted probabilities of malignancy (POM) on the internal validation set. **d**. The AUCs are reported together with the 95% CI on the external validation set of Tangshan (TS). **e.** Sensitivity, specificity, false positive, true positive, false negative, and true positive evaluated on the external validation set of TS. **f**. The distribution of predicted POM on the external validation set of TS. **g**. The AUCs are reported together with the 95% CI on the external validation set of Dazu (DZ). **h.** Sensitivity, specificity, false positive, true positive, false negative, and true positive evaluated on the external validation set of DZ. **i**. The distribution of predicted POM on the external validation set of DZ.

HER2-positive and TNBC were 12, 6, 4 and 7, respectively (Table 1). The average of POM in the four molecular subgroups were 0.917, 0.911, 0.922 and 0.827, respectively. The average value of POM in the TNBC group was significantly lower than the remaining groups (p < 0.05) (supplementary Fig. S5).

Fig. 5 provides the examples of true-positive and false-positive outcomes in the SW dataset. These heatmaps can reveal the salient areas in the US images contributing to the prediction of malignant and benign microcalcifications (strong emphasis in red and weak emphasis in blue). They could intuitively explain what the model can learn from the training data by focusing on the lesion areas.

The breasts were highly informative, and neuroactivation corresponding to the lesions contributed significantly to the final assessment of malignancy probability. In previous studies, most heatmap signals did not coincide on B-mode and color Doppler images of the same lesions. The proposed EDL-BC model was sensitive to blood flow signals around and inside of tumors, which were inconspicuous in the color Doppler images. The number of false positives in the internal dataset and the two external datasets were 110, 64 and 48, respectively (Fig. 4b, e, h). In total, 11 false-positive cases of them contained strong signals (9 benign lesions in the nipple and areola area and 2 cases with inflammation; 8 of them were shown in Fig. 5b). When the threshold is set to 0.76, there is 1 false negative in the internal dataset and the DZ dataset, respectively (Fig. 5b and h). More heatmaps of US benign lesions generated by EDL-BC are provided in supplementary Fig. S6 and supplementary Fig. S7.

Multi-factorial exploration of artificial intelligence assistance in clinical practice

To evaluate the influence of EDL-BC on clinical practice, we invited six radiologists with more than 20 years of clinical experience to analyse both B-mode and color-Doppler images with or without assistance from EDL-
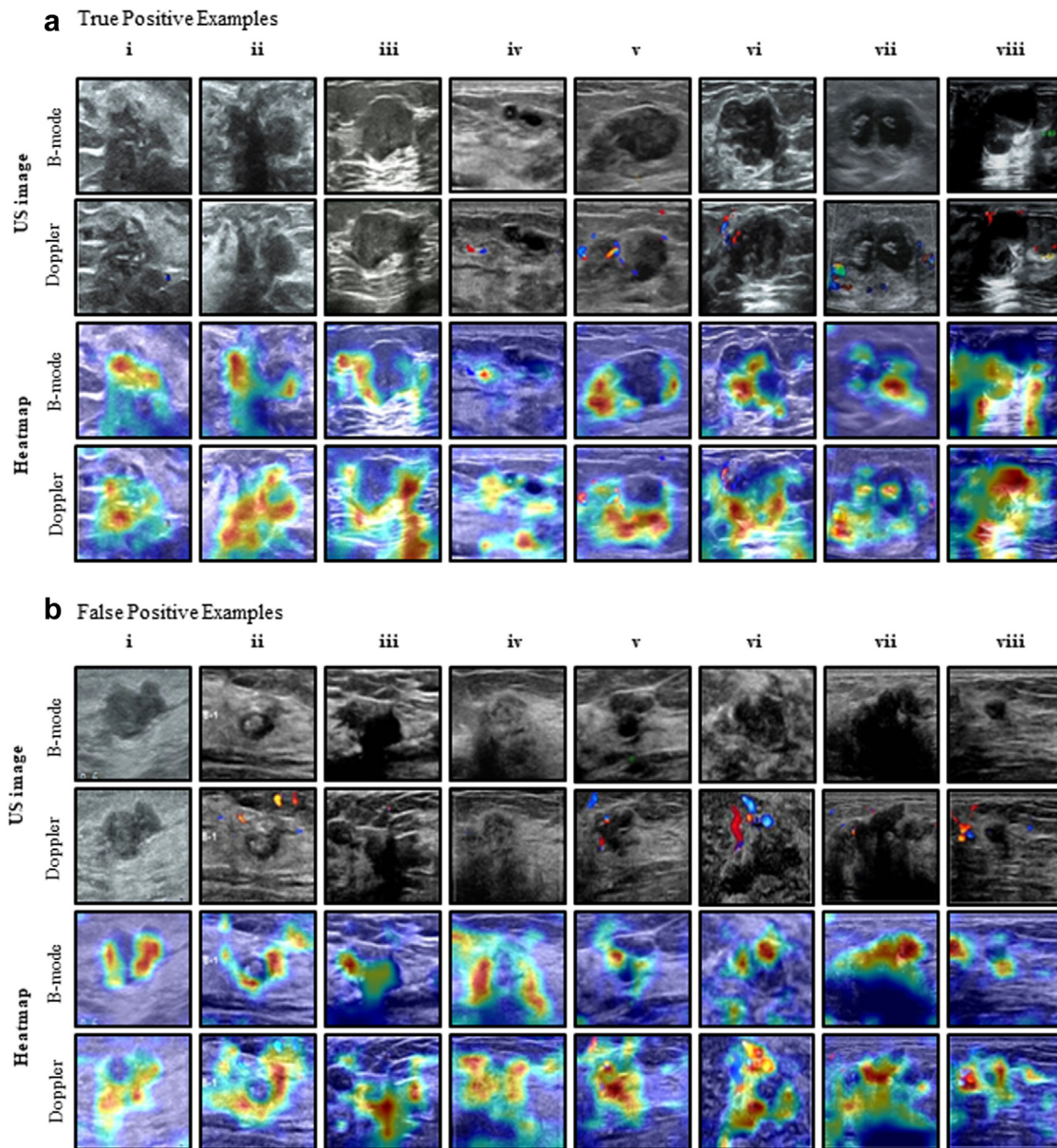
Fig. 5: **Visualization of the attention heatmaps of breast lesions in the B-mode ultrasound (US) images and Doppler US images from Southwest (SW) cohort.** These heatmaps are generated using by EDL-BC with Grad-CAM. **a**. Heatmaps of true positive breast lesions, which both EDL-BC model and pathology confirmed as malignancy. **b**. Heatmaps of false positive breast lesions, which EDL-BC predicted as malignancy while pathological results were benign. Note that the first row represents the B-mode images and the second row shows Color Doppler images of patient lesions, the last two rows are the corresponding EDL-BC ensembled heatmaps of two types of US images.

BC. As mentioned above, the review process was realized on an online rating platform and 1572 pairs of images from 3 centres were incorporated.

As shown in Table 2, supplementary Table S4 and supplementary Fig. S8, the proposed EDL-BC achieved an AUC of 0.945 (95% CI: 0.933–0.965). With the addition of the threshold, the specificity increased from 0.753 (CI: 0.741–0.795) to 0.945 (CI: 0.933–0.957), the

value of Kappa was increased from 10.1% to 21.5%, the F1-score was increased from 12.3% to 23.2%, while the sensitivity decreased from 0.969 (CI: 0.842–0.999) to 0.454 (CI: 0.281–0.636). The radiologist can find 15.2%–30.3% of breast cancers in the test sets (in a separate manner). Diagnostic performance for each individual radiologist, the sensitivity was 15.2%–30.3% and the specificity was 95.9–98.2%. On average, the radiologists

achieved an AUC of 0.716 (CI: 0.693–0.738). The sensitivity, the value of Kappa and F1-score of these radiologists was merely 0.121 (CI: 0.034–0.282), 0.139 and 0.146, respectively, but the corresponding specificity was better than the proposed approach using any threshold (0.753–0.945 compared to 0.991). It can be ascribed to the characteristics of the datasets. In addition, the radiologists favored predicting these lesions to be benign.

These experimental results did demonstrate the complementarity between radiologists and the proposed DL-based platform. With the assistance of EDL-BC (threshold set to 0.76), the radiologists achieved an AUC of 0.899 (CI: 0.883–0.913), a Kappa of 0.359 and a F1-score of 0.370, and the sensitivity was 0.545 (CI: 0.364–0.719), which was immensely superior to the outcome without AI assistance (p < 0.0001) (as shown in Table 2). There were no significant differences between the EDL-BC model and radiologists with AI assistance (p = 0.099) (Table 2). For each individual radiologist, the Kappa value and F1-score of radiologists with AI assistance were higher than those of radiologists without AI assistance (supplementary Table S4). Namely, the performance of malignant lesion detection had been significantly enhanced.

In general, the experimental results suggested that our model could be potentially valuable to assist radiologists in interpreting the manually indistinguishable breast US images. With the assistance of our model, under the premise of insignificant decrease in specificity, the sensitivity of radiologists was significantly improved, indicating the application prospect of our model in clinical practice.

Discussion

This is an early work of DL in screening the patients with early breast cancer with US images for assisting the radiologists in clinical practice. Breast cancer is the most commonly diagnosed cancer in women, severely threatening women's health globally. In addition, breast cancer could be successfully cured if detected in its early stage. AI plays a vital role in breast cancer screening and detection, which can reduce the workload of the radiologists while making up for the inexperience and skill deficiency of beginners. AI models can find details in medical images that human visual inspection cannot and automatically make a quantitative judgment. DL has been extensively employed in image detection and classification due to its advantages, including being accurate, fast, and reproducible. Many studies have been reported regarding the detection and classification of lesions in breast US images using AI models.[6,25–29] Our study has several strengths. First, prior research has primarily focused on differentiating between benign and malignant breast lesions, hence evaluating AI systems only on the images which contain either benign or malignant lesions.[6,27–29] In this work, we aim to develop a DL model to identify high-risk lesions in US images with benign findings, which is beneficial to early confirmation with immediate biopsy. Thus, it can further improve the detection rate of early breast cancer. Second, most studies used only training and validation sets from one institution without an independent external test set.[25,28,29] In order to eliminate the differences in the disease spectrum in different centres, two external test sets were included for our model validation. Third, unlike other studies that only included B-mode US images,[27,29] our EDL-BC contains two separate feature extraction modules designed for B-mode and color Doppler US images. Finally, different from previous work that was built on the top of a conventional image classification model of SEResNet18,[6] our work proposed an ensemble model consisting of 80 base learners (each base learner was built on the top of ResNet50) by considering the nature of unbalanced

	AUROC (95% CI)	P value	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV	NPV	Kappa	F1-score
Composition of test dataset (1562 data pairs from 3 centres)									
AI only									
0.62	0.945 (0.933–0.956)	–	0.769 (0.741–0.795)	0.969 (0.842–0.999)	0.753 (0.731–0.775)	0.082	0.999	0.122	0.156
0.67	0.945 (0.933–0.956)	–	0.803 (0.778–0.824)	0.969 (0.842–0.999)	0.754 (0.732–0.776)	0.095	0.999	0.141	0.173
0.76	0.945 (0.933–0.956)	–	0.867 (0.847–0.885)	0.909 (0.757–0.981)	0.842 (0.823–0.860)	0.127	0.999	0.203	0.232
0.80	0.945 (0.933–0.956)	–	0.898 (0.878–0.916)	0.878 (0.718–0.966)	0.899 (0.883–0.914)	0.158	0.997	0.230	0.258
0.90	0.945 (0.933–0.956)	–	0.944 (0.929–0.957)	0.454 (0.281–0.636)	0.945 (0.933–0.957)	0.178	0.989	0.231	0.254
Radiologists without AI assistance									
	0.716 (0.693–0.738)	<0.001[φ]	0.973 (0.965–0.978)	0.121 (0.034–0.282)	0.991 (0.985–0.995)	0.148	0.982	0.139	0.146
Radiologists with AI assistance									
	0.899 (0.883–0.913)	<0.001[Θ] 0.099[§]	0.974 (0.963–0.982)	0.545 (0.364–0.719)	0.982 (0.974–0.988)	0.244	0.990	0.359	0.370

The primary dataset consisting of the internal dataset, was collected by Southwest Hospital of China. Another two external datasets were collected by the Tangshan People's Hospital (TS, located in Hebei Province, northern China) and the Dazu People's Hospital (DZ, located in Chongqing), and used for evaluating the proposed AI system. P value φ were radiologists without AI compared to AI only, Θ were radiologists with AI compared to radiologists without AI, § were radiologists with AI compared to AI only.

Table 2: The diagnostic performance of EDL-BC alone, radiologists alone, and EDL-BC-assisted radiologists.

negative/positive samples. This study focused on developing a DL model to identify subtle changes in an early stage of breast cancer which is accessible to mis-diagnosed by radiologists. Moreover, US has shown better sensitivity than mammography for breast cancer detection regardless of age group.[30]

The experimental results indicated that our EDL-based AI system could enhance the accuracy and sensitivity of radiologists. According to these findings, we make two recommendations for implementing the DL model in clinical practice. Firstly, our EDL-BC model could improve the accuracy of diagnosis. The diagnosis of US images is often subject to inter-observer vari-abilities, especially in non-academic centres. It has been reported that the variability in sensitivity ranges from 54.9% to 100%, and specificity ranges from 23.3% to 94.2% radiologists.[31–34] In our study, the sensitivity of six radiologists was 0.121, which is lower than those of each radiologist. In our case, each doctor provides a clear judgment of benign or malignant with the given lesion, so the malignant probability is either 0 or 1. Then, with a given malignant lesion, the probability value of this lesion will increase by 1/6, if a doctor judges correctly. The malignant probability of the given lesion is the accumulation of probability value of all radiologists. Finally, considering our AI system employs the threshold of 0.76 to classify the benign lesions and the malignant lesions, roughly corresponding to the cases of (at least) 5 radiologists judge the given positive lesion as malignant (with the malignant probability of 5/6). As a result, we yield a low sensitivity of 0.121 for all radiol-ogists. With the assistance of EDL-BC, the accuracy of radiologists increased from 0.716 to 0.899 (Table 2). EDL-BC also improved the performance of radiologists in clinical practice. The combined accuracy of EDL-BC and radiologists using the DL model in the diagnosis of breast cancer in US images was significantly better than that of US doctors (P < 0.0001) (Table 2). Mean-while, the diagnostic accuracy of US doctors for breast cancer assisted by the DL model was significantly higher than that of the non-assisted AI group (P < 0.0001) (Table 2). The accuracy of breast cancer diagnosis by AI-assisted doctors was slightly lower than that of the DL model, and there was no significant difference between the two groups (P = 0.099) (Table 2). Therefore, the employment of DL in breast cancer screening and detection is of great significance. Our proposed AI sys-tem can be used to assess breast lesions comparable to that of experienced human experts. Note that clinical decisions should be supervised by clinicians even though EDL-BC is reported to have superior perfor-mance. Secondly, EDL-BC could contribute to discrim-inating against the high risk of early breast cancer, reducing the misdiagnosis rate, and avoiding unnec-essary invasive biopsies. Notably, the most crucial role of the proposed model is improving diagnostic accuracy by assisting clinicians.

The salient regions in the heatmaps from various imaging modalities were beneficial in discriminating malignant lesions and aiding clinicians in understand-ing the decisions made by AI. For instance, the tumor vessel density is proportional to the tumor size and pathological severity.[35] Early breast cancer can be char-acterized by a high density of blood vessels with disor-dered distribution.[24] Vessel distribution is equivalent in the cores and peripheries of benign lesions but concentrated toward the centres of malignant lesions had greater vascularization towards their centre.[36] Almost 99% of malignant lesions, and only 4% of benign lesions, had detectable vascularization on color Doppler images superimposed on B-mode images.[37] AI decision making is derived from the morphological and texture features extracted from the images. In this study, we speculate that the EDL-BC model could capture subtle changes in blood flow signals in color Doppler images, which plays an important role in identifying malignant lesions and represents a significant advantage over diagnosis by US doctors.

In addition, we found two cases of early breast cancer whose US images did not change in size and shape within three years. However, EDL-BC suggested that the malignant value was high before pathology finally confirmed breast cancer. It is suggested that breast cancer cannot be excluded even if the size or morphology does not change in the patients with US benign during follow-up. Furthermore, there are four subtypes of breast cancer according to molecular cate-gories: luminal A, luminal B, HER 2 over-express, and TNBC. In 199 cases of early breast cancer, Luminal A 45.23%, Lumina B 18.09%, HER2+ 14.57%, and TNBC 17.59%. About 40% of the 199 cases belong to Luminal A, which is consistent with the distribution of Luminal A in the breast cancer population. As reported in pre-vious studies, luminal A is the most common molecular subtype, representing 40%–50% of breast cancers.[38] Our results showed that EDL-BC has different prediction for different breast cancer subtypes. We can observe that the malignant predictive value of our model for TNBC is lower than that of another molecular subtype, and the differences in these comparisons were significant (p < 0.05) (supplementary Fig. S5). In addition, the performance of our model for TNBC is unsatisfactory since TNBC was more likely to be misinterpreted as benign in the US images. Meanwhile, the mass lesions of TNBC were characterized by circumscribed margins, were markedly hypoechoic, and were less likely to show posterior shadowing.[39]

The proposed DL model produced 11 false-positive cases with strong signals detected in the internal and external datasets. In 9 cases of them, the lesions were located in the nipple and areola, and the distance be-tween the nipple and the lesion was less than 30 mm. The other 2 cases contained inflammatory masses. The enhancement of the blood flow signal can be discovered

in the thermogram generated by the proposed model in all 11 cases and its mechanism needs further study. Furthermore, the false-negative predictions made by the EDL-BC model at malignancy thresholds of 0.628 and 0.769, were diagnosed by US doctors as fibroadenoma and adenosis in the internal dataset (supplementary Fig. S9). With the threshold set to 0.5, the lesions were classified as breast cancer, indicating that the EDL-BC model is biased toward malignancy prediction. Lack of the trust of human experts is a barrier to the deployment of DL in clinical practice. To cope with the black-box nature of DL,[40] we highlightened the visually interpretable features (i.e., heatmaps) in the EDL-BC model to make the model's outcome understandable and increase the likelihood of its application.

This study also has several limitations. First, it was retrospective, which resulted in variation in the class distribution among datasets, with a mixture of consecutive series of patients and convenience samples. Second, we did not have data about risk factors, including a family history of breast cancer and *BRCA* gene test results. Third, all the patients in the training and validation datasets in this study have undergone pathological examination. Therefore, the accuracy and stability of the DL model depend on the quality of cytopathological diagnosis. The change in the incidence of breast cancer among people in different regions may significantly affect the malignant predictive value among people or may reduce the potential universality of the results. Moreover, potential biases exist concerning the selection of radiologists and data, including the exclusion of low-quality images and normal scans. Therefore, future validation in a clinical prospective large-scale screening cohort is needed.

In conclusion, the EDL-BC significantly improved the diagnostic accuracy on breast nodule differentiation and could decrease the number of invasive biopsy. The proposed model can extract the morphological features from early breast lesions, conduct effective and objective image analysis, and provide precise outcome for early breast cancer. The performance of this model in clinical cases shows its effectiveness in assisting the doctors to improve the diagnostic accuracy.

Contributors
JWL, ZGC, XDL, and LC designed the research goals and aims. JWL, YG, ZJD, HHT, LC, ZGC, and XFH designed the model. JWL, ZGC, and LC designed the evaluation methodology. YG, JWL, ZJD, ZLL, PW, XFH, and CYT developed the software for the deep learning model. PW, XW, JL, JL, XY, SCL, XC, QQC, ADW, JJ, YZ, YW, LH, and FLN curated the datasets. YG, JL, ZLL, ZJD, and HHT performed the analysis. HHT created the illustrations. JWL, YY, XFH, XDL, ZGC, and LC wrote the manuscript with the assistance and feedback of all other co-authors. JWL and LC conceived and directed the project. JWL, CL, LH and FLN have accessed and verified data. All authors have read and agreed to publish the paper.

Data sharing statement
All codes used for model development and training is available at: http://112.74.95.118:8888/down/1GTcJYCR8HSr.

The website of the retrospective datasets on a web-based rating platform is available at: http://birads.ssdlab.cn/admin.

The trained model is available for research use upon request. All datasets were used under the license of the respective hospital systems for the current study and are not publicly available.

References
1 WHO. *WHO report on cancer: setting priorities, investing wisely and providing care for all*. 2020.
2 Cancer AJCo. *AJCC cancer staging manual*. New York, NY: Springer; 2016.
3 Duggan C, Trapani D, Ilbawi AM, et al. National health system characteristics, breast cancer stage at diagnosis, and breast cancer mortality: a population-based analysis. *Lancet Oncol*. 2021;22(11):1632–1642.
4 Shen S, Zhou Y, Xu Y, et al. A multi-centre randomised trial comparing ultrasound vs mammography for screening breast cancer in high-risk Chinese women. *Br J Cancer*. 2015;112(6):998–1004.
5 Mandelson MT, Oestreicher N, Porter PL, et al. Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. *J Natl Cancer Inst*. 2000;92(13):1081–1087.
6 Qian X, Pei J, Zheng H, et al. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. *Nat Biomed Eng*. 2021;5(6):522–532.
7 Sato K, Tamaki K, Tsuda H, et al. Utility of axillary ultrasound examination to select breast cancer patients suited for optimal sentinel node biopsy. *Am J Surg*. 2004;187(6):679–683.
8 Sadoughi F, Kazemy Z, Hamedan F, Owji L, Rahmanikatigari M, Azadboni TT. Artificial intelligence methods for the diagnosis of breast cancer by image processing: a review. *Breast Cancer*. 2018;10:219–230.
9 Szolovits P, Patil RS, Schwartz WB. Artificial intelligence in medical diagnosis. *Ann Intern Med*. 1988;108(1):80–87.
10 Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018;2(10):719–731.
11 Chen ZH, Lin L, Wu CF, Li CF, Xu RH, Sun Y. Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine. *Cancer Commun*. 2021;41(11):1100–1115.
12 Lotter W, Diab AR, Haslam B, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat Med*. 2021;27(2):244–249.
13 Das A, Acharya UR, Panda SS, Sabut S. Deep learning based liver cancer detection using watershed transform and Gaussian mixture model techniques. *Cognit Syst Res*. 2019;54:165–175.
14 Mohsen HE-DE-S, El-Horbaty E-SM, Salem A-BM. Classification using deep learning neural networks for brain tumors. *Future Comput Inform J*. 2018;3(1):68–71.
15 Yuan Z, Xu T, Cai J, et al. Development and validation of an image-based deep learning algorithm for detection of synchronous peritoneal carcinomatosis in colorectal cancer. *Ann Surg*. 2022;275(4):e645–e651.
16 Prostate cancer diagnosis using deep learning with 3D multiparametric MRI. In: Liu SZH, Feng Y, Li W, eds. *Medical imaging 2017: computer-aided diagnosis; 2017: International Society for Optics and Photonics;* 2017.
17 Deng K, Wang L, Liu Y, et al. A deep learning-based system for survival benefit prediction of tyrosine kinase inhibitors and immune checkpoint inhibitors in stage IV non-small cell lung cancer patients: a multicenter, prognostic study. *eClinicalMedicine*. 2022;51:101541.
18 Fujioka T, Mori M, Kubota K, et al. The utility of deep learning in breast ultrasonic imaging: a review. *Diagnostics*. 2020;10(12).

19 Fujioka T, Kubota K, Mori M, et al. Distinction between benign and malignant breast masses at breast ultrasound using deep learning method with convolutional neural network. *Jpn J Radiol*. 2019;37(6):466–472.

20 Le EPV, Wang Y, Huang Y, Hickman S, Gilbert FJ. Artificial intelligence in breast imaging. *Clin Radiol*. 2019;74(5):357–366.

21 Liu Y, Wang Y, Wang Y, et al. Early prediction of treatment response to neoadjuvant chemotherapy based on longitudinal ultrasound images of HER2-positive breast cancer patients by Siamese multi-task network: a multicentre, retrospective cohort study. *eClinicalMedicine*. 2022;52:101562.

22 Liu Q, Qu M, Sun L, Wang H. Accuracy of ultrasonic artificial intelligence in diagnosing benign and malignant breast diseases: a protocol for systematic review and meta-analysis. *Medicine (Baltim)*. 2021;100(50):e28289.

23 Sloun VRJG, Regev C, Yonina CE. Deep learning in ultrasound imaging. *Proc IEEE*. 2019;108:11–29.

24 Guo R, Lu G, Qin B, Fei B. Ultrasound imaging technologies for breast cancer detection and management: a review. *Ultrasound Med Biol*. 2018;44(1):37–70.

25 Shen Y, Shamout FE, Oliver JR, et al. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. *Nat Commun*. 2021;12(1):5645.

26 Gu Y, Xu W, Liu T, et al. Ultrasound-based deep learning in the establishment of a breast lesion risk stratification system: a multicenter study. *Eur Radiol*. 2022;33(4):2954–2964.

27 Gu Y, Xu W, Lin B, et al. Deep learning based on ultrasound images assists breast lesion diagnosis in China: a multicenter diagnostic study. *Insight Imag*. 2022;13(1):124.

28 Zhang N, Li XT, Ma L, Fan ZQ, Sun YS. Application of deep learning to establish a diagnostic model of breast lesions using two-dimensional grayscale ultrasound imaging. *Clin Imag*. 2021;79:56–63.

29 Zhao Z, Hou S, Li S, et al. Application of deep learning to reduce the rate of malignancy among BI-RADS 4A breast lesions based on ultrasonography. *Ultrasound Med Biol*. 2022;48(11):2267–2275.

30 Bitencourt A, Daimiel Naranjo I, Lo Gullo R, Rossi Saccarelli C, Pinker K. AI-enhanced breast imaging: where are we and where are we heading? *Eur J Radiol*. 2021;142:109882.

31 Fleury EFC, Marcomini K. Breast elastography: diagnostic performance of computer-aided diagnosis software and interobserver agreement. *Radiol Bras*. 2020;53(1):27–33.

32 Turnaoglu H, Haberal KM, Arslan S, Yavuz Colak M, Ulu Ozturk F, Uslu N. Interobserver and intermethod variability in data interpretation of breast strain elastography in suspicious breast lesions. *Turk J Med Sci*. 2021;51(2):547–554.

33 Yoon JH, Kim MH, Kim EK, Moon HJ, Kwak JY, Kim MJ. Interobserver variability of ultrasound elastography: how it affects the diagnosis of breast lesions. *AJR Am J Roentgenol*. 2011;196(3):730–736.

34 Dong Y, Zhou C, Zhou J, Yang Z, Zhang J, Zhan W. Breast strain elastography: observer variability in data acquisition and interpretation. *Eur J Radiol*. 2018;101:157–161.

35 Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*. 2017; 2017. p. 618–626.

36 Song SE, Cho N, Chu A, et al. Undiagnosed breast cancer: features at supplemental screening US. *Radiology*. 2015;277(2):372–380.

37 Hooley RJ, Greenberg KL, Stackhouse RM, Geisel JL, Butler RS, Philpotts LE. Screening US in patients with mammographically dense breasts: initial experience with Connecticut Public Act 09-41. *Radiology*. 2012;265(1):59–69.

38 Voduc KD, Cheang MC, Tyldesley S, Gelmon K, Nielsen TO, Kennecke H. Breast cancer subtypes and the risk of local and regional relapse. *J Clin Oncol*. 2010;28(10):1684–1691.

39 Schrading S, Kuhl CK. Mammographic, US, and MR imaging phenotypes of familial breast cancer. *Radiology*. 2008;246(1):58–70.

40 Moon HJ, Kim EK, Kwak JY, Yoon JH, Kim MJ. Interval growth of probably benign breast lesions on follow-up ultrasound: how can these be managed? *Eur Radiol*. 2011;21(5):908–918.