Deep learning applied to two-dimensional color Doppler flow imaging ultrasound images significantly improves diagnostic performance in the classification of breast masses: a multicenter study

Teng-Fei Yu[1], Wen He[1], Cong-Gui Gan[2], Ming-Chang Zhao[2], Qiang Zhu[3], Wei Zhang[4], Hui Wang[5], Yu-Kun Luo[6], Fang Nie[7], Li-Jun Yuan[8], Yong Wang[9], Yan-Li Guo[10], Jian-Jun Yuan[11], Li-Tao Ruan[12], Yi-Cheng Wang[13], Rui-Fang Zhang[14], Hong-Xia Zhang[1], Bin Ning[1], Hai-Man Song[1], Shuai Zheng[1], Yi Li[1], Yang Guang[1]

[1]Department of Ultrasound, Beijing Tian Tan Hospital, Capital Medical University, Beijing 100070, China;
[2]Department of R&D, CHISON Medical Technologies Co., Ltd, Wuxi, Jiangsu 214028, China;
[3]Department of Ultrasound, Beijing Tongren Hospital, Capital Medical University, Beijing 100730, China;
[4]Department of Ultrasound, The Third Affiliated Hospital of Guangxi Medical University, Nanning, Guangxi 9530031, China;
[5]Department of Ultrasound, China-Japan Union Hospital of Jilin University, Changchun, Jilin 130033, China;
[6]Department of Ultrasound, Chinese PLA General Hospital, Beijing 100850, China;
[7]Department of Ultrasound, Lanzhou University Second Hospital, Lanzhou, Gansu 730030, China;
[8]Department of Ultrasound, Xi'an Tangdu Hospital of No. 4 Military Medical University, Xi'an, Shaanxi 710038, China;
[9]Department of Ultrasound, Chinese Academy of Medical Sciences Cancer Institute and Hospital, Beijing 100021, China;
[10]Department of Ultrasound, The Third Military Medical University Southwest Hospital, Chongqing 400038, China;
[11]Department of Ultrasound, Henan Provincial People's Hospital, Zhengzhou city, Henan 450003, China;
[12]Department of Ultrasound, Xi'an Jiaotong University Medical College First Affiliated Hospital, Xi'an, Shaanxi 710061, China;
[13]Department of Ultrasound, Hebei Medical University First Affiliated Hospital, Zhangjiakou, Hebei 075061, China;
[14]Department of Ultrasound, Zhengzhou University First Affiliated Hospital, Zhengzhou, Henan 450052, China.

Abstract

Background: The current deep learning diagnosis of breast masses is mainly reflected by the diagnosis of benign and malignant lesions. In China, breast masses are divided into four categories according to the treatment method: inflammatory masses, adenosis, benign tumors, and malignant tumors. These categorizations are important for guiding clinical treatment. In this study, we aimed to develop a convolutional neural network (CNN) for classification of these four breast mass types using ultrasound (US) images.

Methods: Taking breast biopsy or pathological examinations as the reference standard, CNNs were used to establish models for the four-way classification of 3623 breast cancer patients from 13 centers. The patients were randomly divided into training and test groups ($n = 1810$ *vs.* $n = 1813$). Separate models were created for two-dimensional (2D) images only, 2D and color Doppler flow imaging (2D-CDFI), and 2D-CDFI and pulsed wave Doppler (2D-CDFI-PW) images. The performance of these three models was compared using sensitivity, specificity, area under receiver operating characteristic curve (AUC), positive (PPV) and negative predictive values (NPV), positive (LR+) and negative likelihood ratios (LR−), and the performance of the 2D model was further compared between masses of different sizes with above statistical indicators, between images from different hospitals with AUC, and with the performance of 37 radiologists.

Results: The accuracies of the 2D, 2D-CDFI, and 2D-CDFI-PW models on the test set were 87.9%, 89.2%, and 88.7%, respectively. The AUCs for classification of benign tumors, malignant tumors, inflammatory masses, and adenosis were 0.90, 0.91, 0.90, and 0.89, respectively (95% confidence intervals [CIs], 0.87–0.91, 0.89–0.92, 0.87–0.91, and 0.86–0.90). The 2D-CDFI model showed better accuracy (89.2%) on the test set than the 2D (87.9%) and 2D-CDFI-PW (88.7%) models. The 2D model showed accuracy of 81.7% on breast masses ≤1 cm and 82.3% on breast masses >1 cm; there was a significant difference between the two groups ($P < 0.001$). The accuracy of the CNN classifications for the test set (89.2%) was significantly higher than that of all the radiologists (30%).

Conclusions: The CNN may have high accuracy for classification of US images of breast masses and perform significantly better than human radiologists.

Trial registration: *Chictr.org*, ChiCTR1900021375; http://www.chictr.org.cn/showproj.aspx?proj=33139.

Keywords: Deep learning; Ultrasonography; Breast diseases; Diagnosis

Access this article online

Quick Response Code:

Website: www.cmj.org

DOI: 10.1097/CM9.0000000000001329

Correspondence to: Prof. Wen He, Department of Ultrasound, Beijing Tian Tan Hospital, Capital Medical University, No. 119, West Road of South, 4th Ring Road, Fengtai District, Beijing 100070, China
E-Mail: ttyyus@sina.com

Chinese Medical Journal 2021;134(4)
Received: 03-10-2020 **Edited by:** Ning-Ning Wang

Introduction

Breast cancer is among the most commonly diagnosed cancers and is the main cause of cancer death in women.[1] Early detection of breast cancer with ultrasound (US) reduces the mortality from breast cancer and provides women diagnosed with breast cancer more options for less-aggressive treatment.[2,3] Previous researchers have reported that color Doppler US in combination with B-mode US can improve the diagnosis of breast cancer lesions.[4,5] The main limitation of screening US is its low positive predictive value (PPV), with a large number of false positive results that can lead to unnecessary biopsies or short-term follow-up visits.[6] According to network protocol 6666 of the American College of Radiology and Imaging (ACRIN 6666), the PPV of biopsies after US screening in high-risk women is 7.4% (18 of 242 cases), which is the current benchmark for breast cancer screening in the United States.[7,8] However, the recall for tomosynthesis showed a PPV of only 24% to 37% according to biopsies and 24% to 50% according to magnetic resonance imaging (MRI).[9-11] In addition, the level of medical service in various regions of China varies greatly, and in some primary hospitals the quality of the medical service may not be guaranteed.[12]

Inflammatory masses and adenosis masses are both benign tumors; however, some cases may be misdiagnosed as malignant tumors on US. Sclerosing adenosis (SA) of the breast has sonographic features similar to some malignant tumors,[13] and can present as a solid hypoechoic mass with unclear borders, irregular morphology, and visible calcification. Granulomatous mastitis (GM) is a chronic inflammatory disease that occurs in the lobules of the breast. Clinically, GM lesions can present as a solid mass with unclear borders and are often highly suspected to be breast cancer.[14] In China, breast masses are classified into four categories according to the treatment methods:[15] inflammatory masses, adenosis, benign tumors, and malignant tumors. However, it may be difficult to distinguish the four types on US examination alone.

The development of artificial intelligence in the medical field has brought new opportunities with the potential to improve the diagnostic accuracy of medical image interpretation while reducing manpower requirements. Artificial intelligence is good at identifying complex patterns in images and quantifying information that humans have difficulty detecting, thereby complementing clinical decision making.[16] Deep learning algorithms have recently become a widely used artificial intelligence method for medical image analysis, being able to use continuous data, nominal data, categorical data, and ordered data simultaneously, taking into account the opinions of the US doctor in the decision-making process. Therefore, the ultrasonologist can work together with an enhanced intelligent algorithm to achieve higher accuracy.[17] The field of deep learning diagnosis of breast cancer is mainly concerned with the differentiation of malignant and benign diagnoses. However, regardless of the size of the mass, it may be difficult to distinguish between malignant tumors, benign tumors, adenosis, and inflammatory masses. In addition, further differentiation of breast masses will be more conducive to guiding clinical treatment, reducing the suffering of patients, and reducing the working load of ultrasonologists.

Therefore, our goal was to train a convolutional neural network (CNN) to distinguish not only between benign and malignant tumors of the breast, but also inflammatory masses and adenosis. We also evaluated the performance of different models trained on two-dimensional (2D) images only, 2D images and color Doppler flow imaging (2D-CDFI), and 2D images, color Doppler flow imaging, and pulsed wave Doppler (2D-CDFI-PW) images.

Methods

Ethical approval

This study was approved by the Ethics Committee and Institutional Review Board of Beijing Tiantan Hospital, Capital Medical University (No. KY2018-099-01). Written informed consent was not required for this study because only fully anonymized US images have been used.

Study overview

This multicenter study used CNNs to classify breast masses. A total of 13 hospitals from nine provinces of China participated in the study and collected US images of breast masses acquired from patients between January 2016 and January 2018. Prior to recruitment, researchers at each site received comprehensive guidance on the research program, including the eligibility criteria and the standardized data collection and interpretation procedures. Figure 1B shows a flowchart of the training and testing protocol. Three models were established to compare the diagnostic accuracy of 2D, 2D-CDFI, and 2D-CDFI-PW images of patients in the training and test sets. The pathological results of breast biopsy or surgery were used as the reference standard against which the CNNs were judged.

Patients and datasets

The patients' treatment process (from 2016.1 to 2019.12) is shown in Figure 1A. If a patient had a grade higher than category 4A (low suspicion for malignancy) according to the Breast Imaging Reporting and Data System (BI-RADS, American College of Radiology [ACR]), a biopsy would typically be performed, with the patient's consent. Breast biopsy was performed using a 16- or 18-G needle (Bard Magnum, GA, USA), and three breast pathologists, each with more than 6 years of experience, examined the biopsy or surgical specimens while blinded to the US and clinical examination results.

Pathological results were classified into benign tumor (I), malignant tumor (II), inflammatory masses (III), or adenosis (IV). Benign tumors include fibroids, intraductal papilloma, benign lobular tumors, and lipomas. Malignant tumors include infiltrative ductal carcinoma, intraductal papillary carcinoma, carcinoma *in situ*, mucinous carcinoma, myeloid carcinoma, and invasive lobular carcinoma. Adenosis include adenoid hyperplasia and SA. If the pathological outcome was a combination of multiple outcomes, the more severe outcome was taken as the classification for this study.

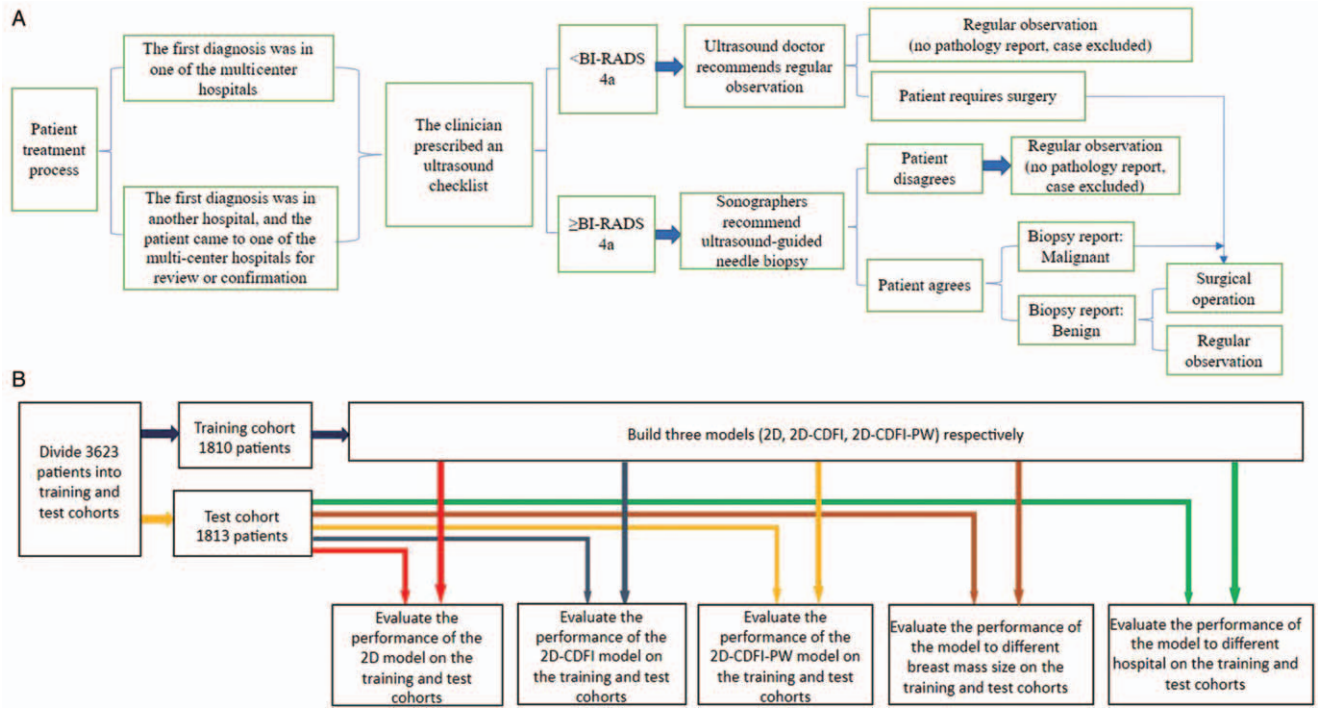

Figure 1: Patient treatment process in flowcharts showing (A) Schematic representation of the patient's clinical pathway from initial clinical diagnosis to ultrasound indications for subsequent biopsy or surgical resection of the lesion; (B) Training and testing protocols. CDFI: Color Doppler flow imaging; PW: Pulse wave; BI-RADS: Breast imaging reporting and data system.

The patient inclusion criteria were: (1) pathological results clearly classified into one of the four categories mentioned above; (2) at least 2D mode US images available, but preferably also CDFI and PW mode images. The exclusion criteria were: (1) a foreign-body in the breast, such as breast augmentation material; (2) other metastatic tumors or co-infection with HIV; (3) measurement markers, arrows, or puncture needles within the image; (4) blurred images or color overflow.

Ultrasound examinations

The US instruments employed to acquire the images used in this study included Esaote My Lab (Esaote, Italy), GE LOGIQ E9 (General Electric Co., USA), Hitachi (Hitachi, Ltd., Japan), Mindray Resona 7 (Shenzhen Mindray Bio-Medical Electronics Co., Ltd., China), Samsung (Samsung Medison Co., Ltd. Korea), Siemens (Siemens Healthcare GmbH, USA), Sonoscape (SonoScape Medical Corp., China), Supersonic (Supersonic Imagine, France), and Toshiba (Aplio 500, Aplio i900, CANON Medical Systems Corporation, Japan) systems. The 2D US images were recorded first, followed by the CDFI. PW scanning was performed if there was evidence of significant blood flow around or within the mass. When a mass was too large to display on a single image, it was divided into several parts for image acquisition. One to five images were taken per patient. For this multicenter study, strict quality control was adopted for the entire process, with the operators (who had each performed more than 7000 breast US scans) receiving rigorous training in the use of uniform procedures for 2D-CDFI-PW measurements. About ten doctors with more than 10 years of experience in US operations were hired as quality control personnel to review all 2D-CDFI-PW images and rule out non-qualifying acquisitions. After adoption of the above selection procedures, 3623 patients were selected for this study, with an age range from 11 to 95 years, a mean age of 42.5 years, and a median age of 42 years. The pathology subtypes include fibroids, intraductal papilloma, benign lobular tumors, lipomas, infiltrative ductal carcinoma, intraductal papillary carcinoma, carcinoma *in situ*, mucinous carcinoma, myeloid carcinoma, invasive lobular carcinoma, adenoid hyperplasia, and SA.

CNN for breast mass classification

The CNN was developed using a computer equipped with an Intel Core i7-6850K 3.6-GHz 6-Core processor, 64 Gigabytes of RAM, and NVidia GeForce GTX 1080 TI graphic processing units. The CNN model included two modules: a detection module and a classification module.

The detection module, which is composed of two sub modules, is used to detect the location of breast masses. The first sub module uses ResNet50 to extract feature maps from the input image. The Feature Pyramid Networks (FPN)[18] structure is used to extract multi-scale features because the images of the selected patients were acquired from different hospitals using different US instruments and therefore exhibited different characteristics. The second sub module is a bounding box regression module, which is used to propose and determine the bounding box (rectangle region) in which the breast mass is located. Using feature maps extracted by ResNet50 as the initial candidate maps, nine scale and aspect ratio rectangles are proposed as candidate regions. The candidate regions are then updated by performing boundary regression with the ground truth regions. The final detections are generated by applying a technique known as non-maximum suppression: regions with confidence scores less than 0.5 or intersection over union less than

0.5 are discarded, and only the remaining top K regions are kept. Focal loss[19] is used to solve the extremely unbalanced distribution between positive and negative samples during the regression of bounding boxes. If a breast image of a healthy person is input, there is no bounding box output by the detection module.

The classification module is used to output the type of breast mass detected. The image containing the region defined by the detection module is processed by 12 convolutional layers, five pooling layers, and two fully connected (FC) layers. The output layer is a four-category FC layer, with the four categories corresponding to the pathological classifications of benign tumor (I), malignant tumor (II), inflammatory mass (III), and adenosis (IV) [Figure 2A–2C].

In US imaging, the 2D image contains the intensity information of the lesion area, the CDFI image contains information about the blood flow surrounding the lesion, and the PW image contains information about the frequency spectrum over a specific area within the lesion. To investigate whether additional information can improve the classification performance, three kinds of network models were designed. The 2D model used only 2D images as input and was considered the baseline model. The 2D-CDFI model added CDFI images, and the 2D-CDFI-PW model further added PW images. These three models were designed to match the three steps adopted in the standard operation of Chinese sonographers performing a breast examination, who first use the 2D mode to acquire images, then second use CDFI, then add the third PW mode when accurate blood flow color Doppler signal is detected. The structures of the three models were approximately the same, except for some differences in the classification module. For the 2D-CDFI model, 2D image features and CDFI image features were extracted in turn, then channel concatenation was used to merge these features before final classification using the FC layer. For the 2D-CDFI-PW model, features were still extracted from the 2D and CDFI images and concatenated for classification, but an attention mechanism was then used to introduce PW features into the 2D and CDFI features. The information provided by the PW and CDFI images is considered supplementary to the information from the 2D-mode US images, but the PW image has some characteristics that require special processing. The PW image generally includes two parts, a 2D image and a PW image, which are arranged in left-and-right or up-and-down layouts. Each part only has half the image resolution of the original 2D or 2D-CDFI image. Usually, the 2D image part contains blood flow information overlaid on the 2D image, which results in part of the lesion area being covered by pixels representing blood flow. Therefore, the overall PW image contains complex information, and we considered it reasonable to extract features from the PW image using an attention mechanism. The attention mechanism multiplies PW features with 2D and CDFI features, with the PW features being obtained by global pooling of the features extracted from the PW images with Resnet50. The full model structure is shown in Supplementary Figures 1–3, http://links.lww.com/CM9/A433.

The following parameters were used to train the models: 2000 max iterations; base learning rate of 0.001, decay by 5% per 100 steps, and batch size of 64. The input images were resized to 416 × 416. Data augmentation was performed before the training process, to reduce any potential bias caused by imbalances in the binary classification data.[20] The images in the training set were augmented according to random rotations of −30° to 30° and random scaling of 0.5 to 1.5 times. The image pixel values of all selected patients were normalized to maintain the consistency of pixel distributions across the different hospitals. The stochastic gradient descent (SGD) algorithm

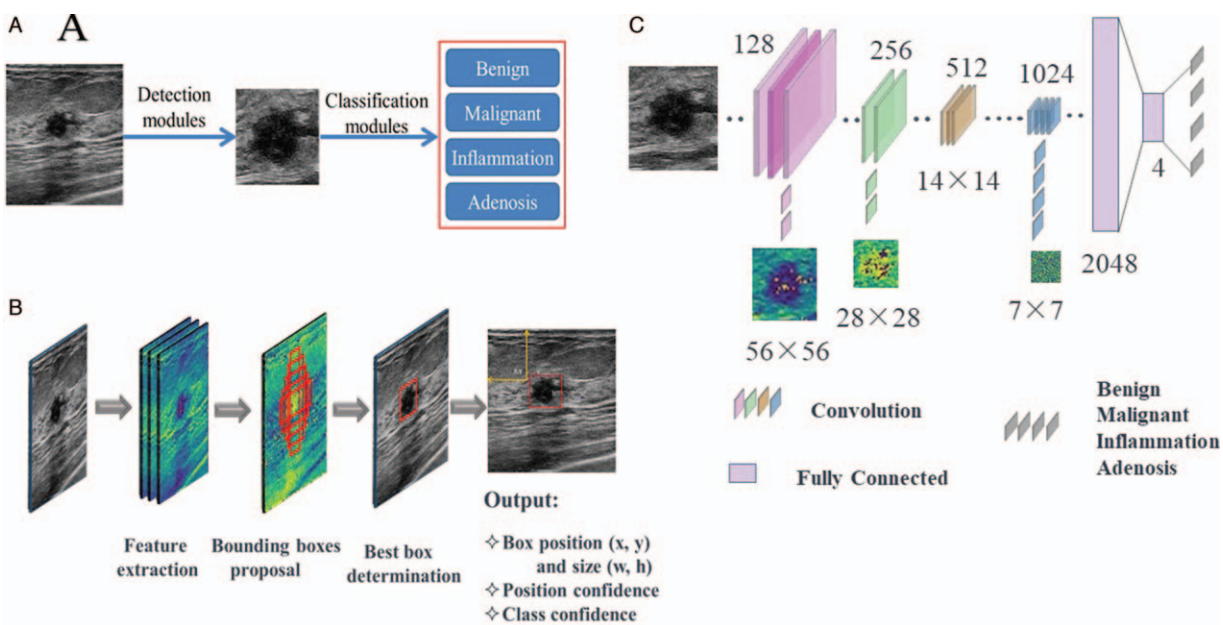

Figure 2: This figure shows the architectures used for building the three models. (A) The whole architecture detection modules and classification modules; (B) The architecture of the detection modules; (C) The architecture of the classification modules.

was used to optimize the models, with an early training stop strategy to avoid over-fitting during the model training.

The integration of different neural networks results in more robustness and better accuracy than that achieved by a single network. The Snapshot Ensemble model provides a simple method for training integrated models.[21] For each of the 2D, 2D-CDFI, and 2D-CDFI-PW models, Snapshot Ensembles were used to train five groups of different model parameters, then the five weak models were combined to achieve a better and more comprehensive strong model.

Statistical analysis

However, to compare the diagnostic performance for different breast masses (benign tumors, malignant tumors, inflammatory masses, and adenosis) between three kinds of images (2D *vs.* 2D-CDFI *vs.* 2D-CDFI-PW), the sensitivity, specificity, positive and negative predictive values, and positive and negative likelihood ratios were calculated for each category for each model, and the areas under the receiver operating characteristics curves (AUCs) were calculated to evaluate the model performance. Further comparisons of these performance measures were made between breast masses of different sizes (size $\leq$ 1 cm *vs.* 1 $<$ size $\leq$ 2 cm *vs.* 2 $<$ size $\leq$ 5 cm *vs.* size $>$ 5 cm). Differences between AUCs were compared using a Delong test. A systematic comparison of the sample composition and performance of each hospital revealed that different results were obtained from images from different hospitals, which may be a result of different distributions. To compare the general ability of the models, we show performance results for data selected from the China-Japan Friendship Hospital of Jilin University (CJ) as an independent validation set. In the calculation of the above values for each category, the corresponding category was processed as a "positive sample," and the other three categories as a "negative sample." All statistical tests were two sided, and *P* value less than 0.05 indicates statistical significance. However, to compare the performance of the model with that of human doctors, 37 experienced ultrasonologists were invited to classify 50 breast mass images randomly selected from the test set into the 4 mass categories. All ultrasonologists had more than 5 years of working experience in breast US. The sample read by each physician was consistent with the AI model, which used all patient cases in the test set. Comparisons were made according to the accuracy of the diagnosis and the total time required to make the diagnosis. All statistical analyses were performed using SPSS for Windows V.20.0 (IBM SPSS Statistics, International Business Machines Corporation, Armonk, NY, USA) and MedCalc (V.11.2; 2011 MedCalc Software bvba, Mariakerke, Belgium).

Results

Baseline characteristics

A total of 5127 patients were initially identified for potential inclusion in this study. From these, 1504 patients were excluded because of the presence of other diseases, antiviral treatment, or unqualified histological, serological, and/or 2D-CDFI-PW results. Thus, 3623 patients with 15,648 images were finally enrolled for analysis. The patients were randomly allocated to training and test sets, with 7835 breast images from 1810 patients used as a training set to train the model and optimize its parameters, and the remaining 7813 images from 1813 patients being used as an independent test set to verify the performance of the generated model. The mean dimension of the tissue samples obtained by needle biopsy was 17.7 mm (all patients). There were 1601 benign masses (I), 1179 malignant masses (II), 572 inflammatory masses (III), and 271 cases of adenosis (IV). The patients' characteristics are summarized in Table 1.

Table 1: A summary of sample composition of each hospital, *n* (%).

Hospitals	All patients	Benign tumor	Malignant tumor	Inflammatory masses	Adenosis
TT	1897 (52.4)	1055 (65.9)	714 (60.6)	110 (19.2)	18 (6.6)
TR	195 (5.4)	76 (4.7)	82 (7.0)	33 (5.8)	4 (1.5)
NN	223 (6.2)	55 (3.4)	50 (4.2)	108 (18.9)	10 (3.7)
CJ	219 (6.0)	81 (5.1)	61 (5.2)	30 (5.2)	47 (17.3)
PLA	158 (4.4)	55 (3.4)	35 (3.0)	18 (3.1)	50 (18.5)
LZ	146 (4.0)	46 (2.9)	35 (3.0)	38 (6.6)	27 (10.0)
TD	135 (3.7)	63 (3.9)	59 (5.0)	9 (1.6)	4 (1.5)
CA	128 (3.5)	37 (2.3)	67 (5.7)	22 (3.8)	2 (0.7)
SW	114 (3.1)	3 (0.2)	1 (0.0)	55 (9.6)	55 (20.3)
HP	122 (3.4)	86 (5.4)	22 (1.9)	14 (2.4)	0 (0.0)
XJ	105 (2.9)	36 (2.2)	53 (4.5)	9 (1.6)	7 (2.6)
HN	103 (2.8)	0 (0.0)	0 (0.0)	102 (17.8)	1 (0.4)
ZZ	78 (2.2)	8 (0.5)	0 (0.0)	24 (4.2)	46 (17.0)
SUM	3623	1601	1179	572	271

TT: Beijing Tiantan Hospital; TR: Beijing Tongren Hospital; NN: The Second Nanning people's Hospital; CJ: China-Japan Friendship Hospital of Jilin University; PLA: Chinese People's Liberation Army General Hospital; LZ: Lanzhou University Second Hospital; TD:Tangdu Hospital; CA: Cancer Hospital Chinese Academy of Medical Sciences; SW: The First Hospital Affiliated to AMU (Southwest Hospital); HP: Henan Provincial People's Hospital; XJ: The First Affiliated Hospital of Xi'an Jiaotong University; HN: The First Affiliated Hospital of Hebei North University; ZZ: The First Affiliated Hospital of Zhengzhou University; SUM: Total sample size of each category; *n*: The percentage of the sample size of this category in the total number of samples of this category in all hospitals.

Comparison of diagnostic performance between the different US image combinations

On the test set, the 2D-CDFI model showed better accuracy (89.2%) than the 2D (87.9%) and 2D-CDFI-PW (88.7%) models ($P < 0.005$). This 2D-CDFI model showed AUC, specificity, sensitivity, and PPV of 0.94, 99%, 93%, and 96%, respectively, for benign tumors; 0.96, 100%, 96%, and 93% for malignant tumors; 0.80, 96%, 55%, and 92% for inflammatory masses; and 0.81, 98%, 46%, and 92% for adenosis [Figure 3; Supplementary Table 1, http://links.lww.com/CM9/A434]. The results suggest that adding the color mode image (2D-CDFI) to the 2D image is helpful for increasing the accuracy of classification, but that further addition of the PW mode image (2D-CDFI-PW) reduced the accuracy. This may be due to the low number of PW images available, which resulted in insufficient training of the model, meaning that the model could not fully fit the distribution of the 2D-CDFI-PW images and thus gave lower performance than the 2D-CDFI and 2D models. A confusion matrix of the classifications is shown in Table 2.

However, as a comparison, we selected data from the CJ to form an independent validation set to test the model training on data from other hospitals. The CJ hospital data consist of 219 cases, including 81 benign tumors, 61 malignant tumors, 30 inflammatory masses, and 47 cases of adenosis. The training dataset consisted of 3404 cases, including 1520 benign tumors, 1118 malignant tumors, 542 inflammatory masses, and 224 cases of adenosis. A similar training strategy was applied to the dataset. The 2D model correctly classified 72 of 81 benign tumors, 55 of 61 malignant tumors, 22 of 30 inflammatory masses, and 33 of 47 cases of adenosis, with accuracy of 88.9% for benign tumors, 90.2% for malignant tumors, 73.3% for inflammatory masses, and 70.3% for adenosis. The 2D-CDFI model correctly classified 12 of 14 benign tumors, 50 of 55 malignant tumors, 16 of 21 inflammatory masses, and 18 of 24 cases of adenosis, with accuracy of 0.857, 0.909, 0.762, and 0.750, respectively. These results are lower than those obtained using the models trained with half of the CJ hospital samples. This result suggests that differences in machines and scanning habits across different hospitals lead to different data distributions across different hospitals. There were insufficient PW images to train the 2D-CDFI-PW model, and therefore, only the 2D and 2D-CDFI models were used for this comparison.

Comparison of diagnostic performance between masses of different sizes

The model trained on 2D images was used to compare performance between different mass sizes and hospitals because 2D images were available for the most complete group of patients in terms of age, pathology subtype, and coverage, and could, therefore, best reflect the differences caused by different mass sizes and different hospitals. For benign tumors, a larger size was associated with a higher correct diagnosis rate. For malignant tumors, size did not affect the accuracy of deep learning, even if the maximum diameter of the tumor was less than 1 cm. For inflammatory masses and adenosis, the accuracy rates were lower than for benign and malignant tumors, but this was to be expected. More details are shown in Figure 4 and Supplementary Table 2, http://links.lww.com/CM9/A435.

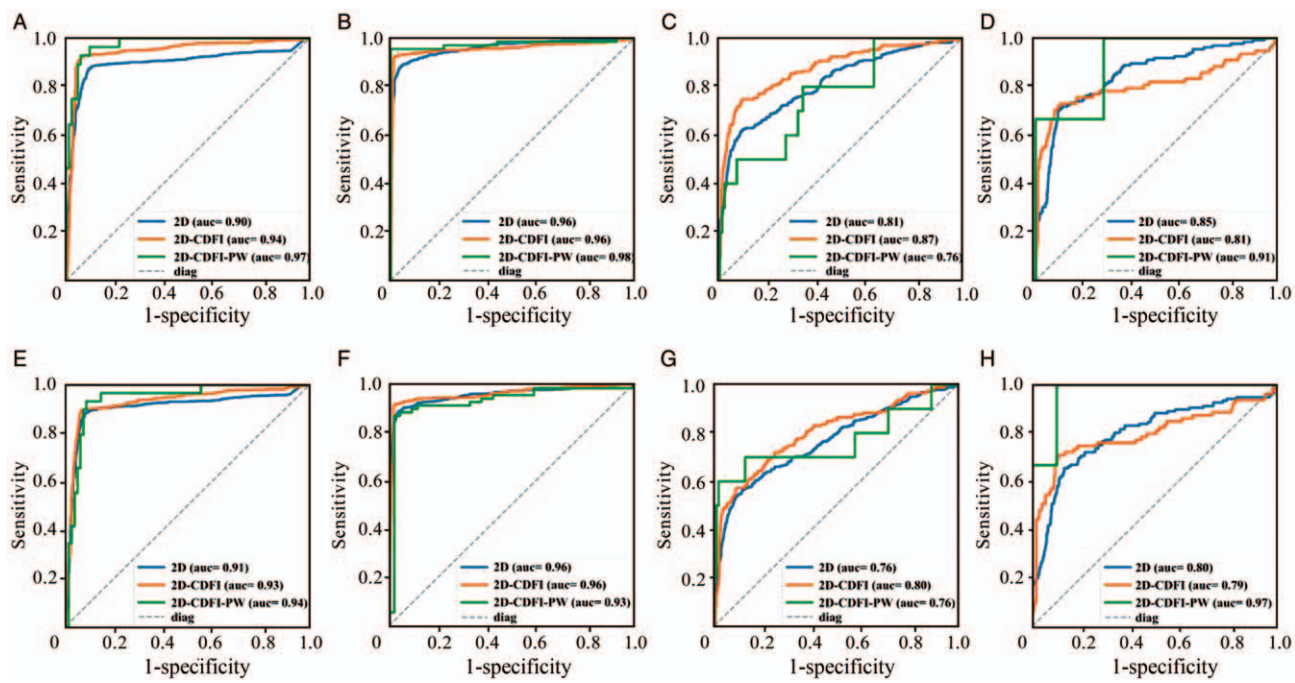

Figure 3: Comparison of ROC curves between 2D ($n = 3623$), 2D-CDFI ($n = 2573$), and 2D-CDFI-PW ($n = 222$) for the assessment of breast masses of (A) Benign tumor (2D: $n = 800$, 2D-CDFI: $n = 589$, 2D-CDFI-PW: $n = 28$); (B) Malignant tumor (2D: $n = 589$, 2D-CDFI: $n = 447$, 2D-CDFI-PW: $n = 69$); (C) Inflammatory masses (2D: $n = 286$, 2D-CDFI: $n = 171$, 2D-CDFI-PW: $n = 10$); (D) Adenosis (2D: $n = 135$, 2D-CDFI: $n = 78$, 2D-CDFI-PW: $n = 3$) in the training cohort; and (E) Benign tumor (2D: $n = 801$, 2D-CDFI: $n = 590$, 2D-CDFI-PW: $n = 29$); (F) Malignant tumor (2D: $n = 590$, 2D-CDFI: $n = 448$, 2D-CDFI-PW: $n = 70$); (G) Inflammatory masses (2D: $n = 286$, 2D-CDFI: $n = 171$, 2D-CDFI-PW: $n = 10$); (H) Adenosis (2D: $n = 136$, 2D-CDFI: $n = 79$, 2D-CDFI-PW: $n = 3$) in the test cohort. CDFI: Color doppler flow imaging; PW: Pulse wave; ROC: Receiver operating characteristic.

Table 2: Confusion matrix of the diagnoses of 2D, 2D-CDFI, and 2D-CDFI-PW models.

Model	Ground truth	Prediction, *n*			
		Inflammatory masses	Adenosis	Benign tumor	Malignant tumor
2D model	Inflammatory masses ($n = 572$)	406	85	72	9
	Adenosis ($n = 271$)	71	181	17	2
	Benign tumor ($n = 160$)	36	89	1451	25
	Malignant tumor ($n = 1179$)	37	54	35	1053
2D-CDFI model	Inflammatory masses ($n = 342$)	252	47	40	3
	Adenosis ($n = 157$)	39	110	6	2
	Benign tumor ($n = 1179$)	17	57	1097	8
	Malignant tumor ($n = 895$)	26	40	17	812
2D-CDFI-PW model	Inflammatory masses ($n = 20$)	16	1	2	1
	Adenosis ($n = 6$)	1	5	0	0
	Benign tumor ($n = 57$)	2	4	49	2
	Malignant tumor ($n = 139$)	2	3	1	133

Ground truth, number of samples in each category identified by the doctor; Prediction, number of samples for each category predicted by the models. 2D: Two-dimensional images only; CDFI: Color Doppler flow imaging; PW: Pulsed wave.

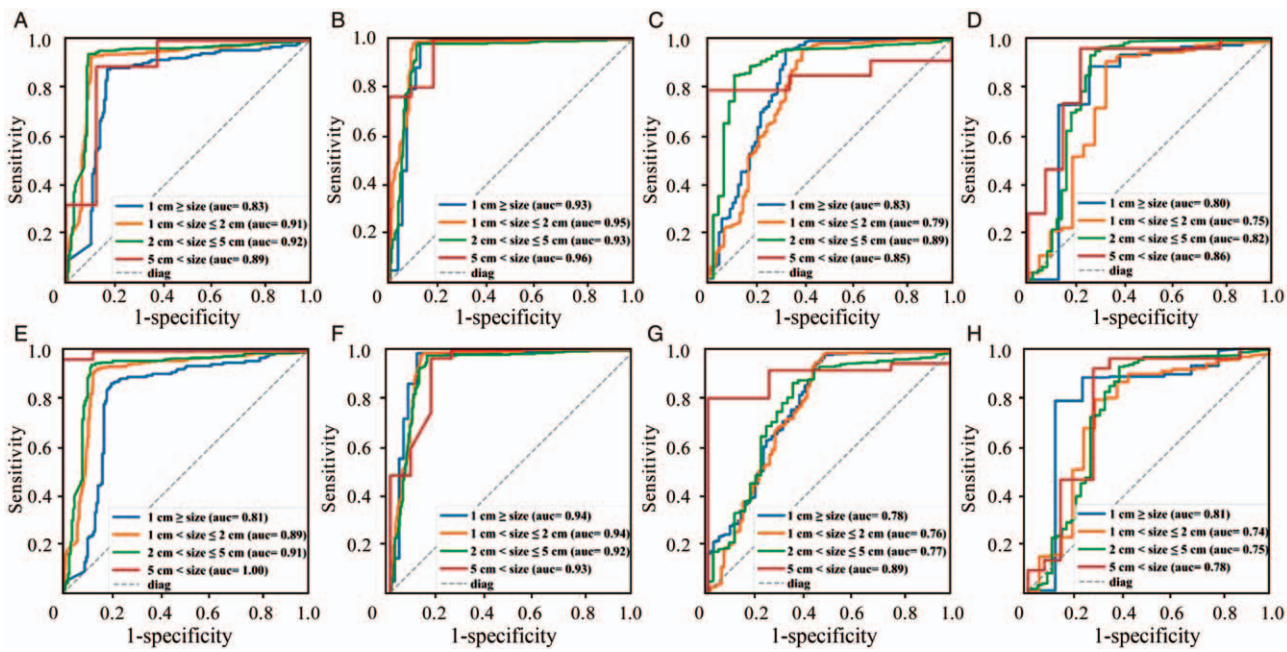

Figure 4: Comparison of receiver operating characteristic (ROC) curves for the assessment of breast masses of: (A) size $\leq$ 1 cm (Benign: $n = 165$, Malignant: $n = 55$, Inflammation: $n = 88$, Adenosis: $n = 8$); (B) 1–2 cm (Benign $n = 370$, Malignant: $n = 223$, Inflammation: $n = 91$, Adenosis: $n = 22$); (C) 2–5 cm (Benign: $n = 186$, Malignant: $n = 213$, Inflammation: $n = 46$, Adenosis: $n = 51$); (D) $\geq$5 cm (Benign: $n = 8$, Malignant: $n = 11$, Inflammation: $n = 3$, Adenosis: $n = 14$) in the training cohort, and (E) size $\leq$ 1 cm (Benign: $n = 166$, Malignant: $n = 55$, Inflammation: $n = 88$, Adenosis: $n = 8$); (F) 1–2 cm (Benign: $n = 371$, Malignnat: $n = 223$, Inflammation: $n = 91$, Adenosis: $n = 22$); (G) 2–5 cm (Benign: $n = 186$, Malignant: $n = 213$, Inflammation: $n = 46$, Adenosis: $n = 51$); (H) $\geq$5 cm (Benign: $n = 8$, Malignant: $n = 12$, Inflammation: $n = 4$, Adenosis: $n = 15$) in the test cohort.

Comparison of diagnostic performance between hospitals

The TT hospital was the most numerically optimized in all aspects, which may be related to the TT hospital having provided more cases than any of the other hospitals. More details are shown in Figure 5 and Supplementary Table 3, http://links.lww.com/CM9/A436.

Comparison of the classification of breast masses between the CNN and radiologists

The CNN showed superior performance to all 37 ultrasonologists for the classification of 50 breast mass images randomly selected from the test set, showing a shorter time and higher accuracy. The accuracy of the CNN model was 89.2%, the CPU processing time 1 s, and the GPU processing time 400 ms. For the 37 ultrasonologists the average time was 314 s, the average accuracy was 30%, the shortest time was 83 s, and the highest accuracy was 45%.

Discussion

In this study, we have used a CNN algorithm to classify 2D, 2D-CDFI, and 2D-CDFI-PW images of breast masses into four categories. The 2D-CDFI model showed better accuracy than the 2D and 2D-CDFI-PW models. There was no

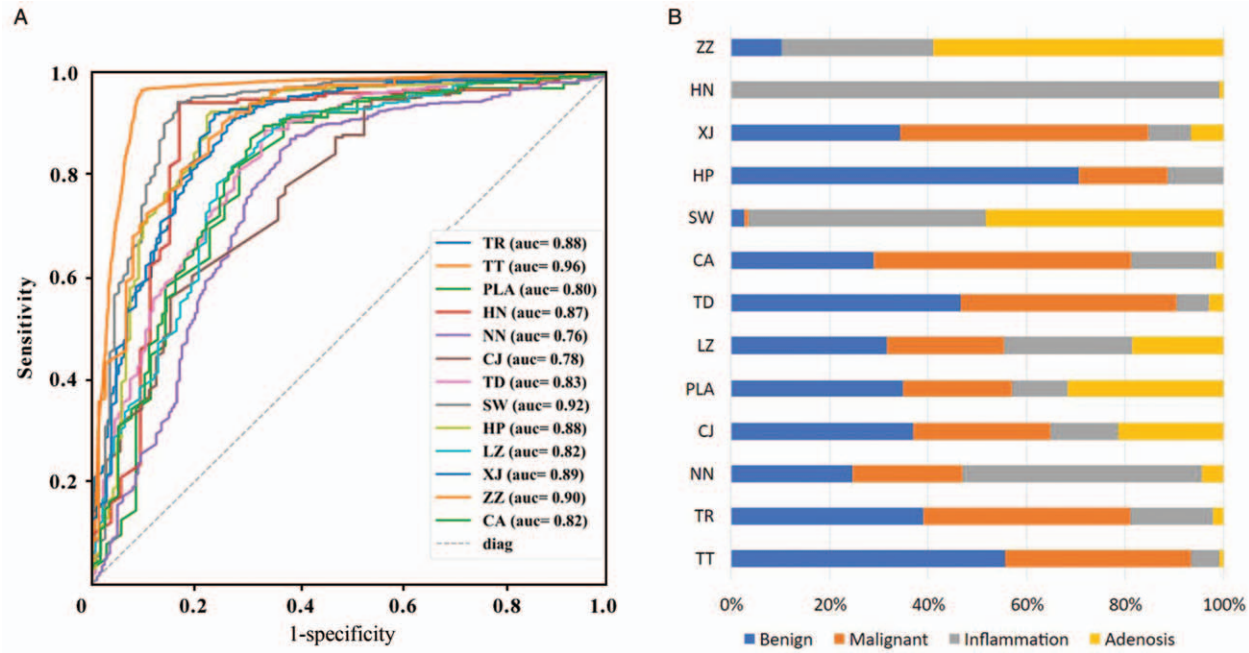

Figure 5: Comparison of assessments of breast masses between different hospitals using (A) ROC curves; (B) Benign, malignant, inflammation, and adenosis category sample size ratios. ROC: Receiver operating characteristic; TT: Beijing Tiantan Hospital; TR: Beijing Tongren Hospital; NN: The Second Nanning people's Hospital; CJ: China-Japan Friendship Hospital of Jilin University; PLA: Chinese People's Liberation Army General Hospital; LZ: Lanzhou University Second Hospital; TD:Tangdu Hospital; CA: Cancer Hospital Chinese Academy of Medical Sciences; SW: The First Hospital Affiliated to AMU (Southwest Hospital); HP: Henan Provincial People's Hospital; XJ: The First affiliated Hospital of Xi'an Jiaotong University; HN: The First Affiliated Hospital of Hebei North University; ZZ: The First Affiliated Hospital of Zhengzhou University.

significant difference in the performance of our model according to different sizes of breast masses. After processing by the CNN model, the background and lesion areas were clearly distinguishable. The model achieved better accuracy than 37 ultrasonologists. As described in the sample selection process, our samples were selected from historical patients who obtained a definite diagnosis after biopsy, with patients without a definite diagnosis being excluded. Therefore, our model may be at risk of low diagnostic accuracy for patients who are difficult to diagnose. Additionally, the data may be affected by the pathological distribution of historical case data, such as in invasive ductal carcinoma, where fibroadenoma pathological types account for the majority of cases, and some cases may also have been affected by concomitant diseases. This bias may lead to a risk of higher diagnostic accuracy for samples belonging to the majority pathological type and lower diagnostic accuracy for samples belonging to less frequent pathological types. When we divided the data sets, we tried our best to consider the data distributions of the training and validation sets, to ensure that they were similar. The classification accuracy was better for benign and malignant tumors than for inflammatory masses and adenosis, which may reflect our data distribution.

Ultrasound is among the most commonly used methods for the screening of breast lumps in China. The ultrasonologist typically examines about 150 patients a day, which is a heavy workload. Neural networks can learn from every new case they deal with and can apply the acquired knowledge to future US diagnoses.[22-24] Therefore, the application of artificial intelligence to US-assisted diagnosis is of great significance for reducing the workload of doctors and providing a rapid diagnosis of breast masses. Until now, research on US classification of breast masses has focused on differentiating benign from malignant tumors. We here propose four classifications for breast masses: benign tumors, malignant tumors, inflammatory masses, and adenosis, which is more in line with the clinical diagnostic process [Supplementary Figure E4, http://links.lww.com/CM9/A433]. Human doctors show excellent performance in distinguishing benign tumors from malignant tumors on US images, with an accuracy rate of about 84%,[25] whereas the accuracy rate of the doctors who performed best on the four-category classification problem was only 45%. It is also difficult to distinguish inflammatory masses from adenosis on more expensive radiological methods, such as MRI or molybdenum target imaging. This indicates that most clinical cases require biopsy or pathological examination to confirm the diagnosis. Therefore, we tried to identify the four types of masses simultaneously during diagnosis, which is of great importance for improving the efficiency of clinical diagnosis and treatment. Our results show that the classification accuracy for inflammatory masses and adenosis was less than that for benign and malignant lesions, which may be a result of the lower number of samples available for training in the inflammatory masses and adenosis categories.

As a multicenter study, we collected US data from 3623 patients from 13 hospitals. This is the largest breast US dataset ever reported, and it contains high quality annotations using pathological results as the reference standard. The dataset also includes PW mode images, which have not been mentioned in previous studies. Our research supplements the small amount of data used in

previous studies, and validates the applicability of deep learning models created using small amounts of breast US data to the analysis of large datasets.

In this study, we provided not only the commonly used 2D-CDFI diagnosis and processing methods, but also introduced PW mode images as model inputs rarely. This is more in line with the clinical diagnostic process. However, because of the limitations of the current PW mode data, we were not able to conclude that PW can significantly improve diagnostic results. Still, the PW images acquired as an auxiliary method in the clinical diagnostic process contain additional spectral information. The same structural methods were used to build the different models to allow the contribution of images acquired using different modes to be compared. There is currently no other research model focusing on PW images, and we expect to further improve this model in future research work.

This study used datasets acquired using US equipment from different manufacturers. As different US devices may have different signal processing and image optimization parameters, many traditional auxiliary diagnostic models are only suitable for use with a single model of US device. Therefore, we included images acquired on US equipment from different manufacturers, which will also help with the implementation of future models.

Because this study reports preliminary research results, it is subject to the following limitations. We have not fully explored the features of various patterns of images and established targeted models. We did not investigate whether the classification results change according to the distribution of the characteristics of the device, the characteristics of the doctor, or the characteristics of the patients. We expect to provide a more detailed discussion and analysis of various details of the dataset in future research.

Conclusion

The author would like to conclude that using 2D-CDFI US images, the deep CNN model achieved better results in the diagnosis of breast masses than observations by experienced ultrasonologists. In addition to benign and malignant breast masses, the CNN model could accurately diagnose inflammatory and adenopathy breast masses. Without the need for special equipment, the CNN model's diagnostic recommendations based on US images could reduce predetermined biopsies, simplify the workload of US doctors, and enable targeted and refined treatment.

Funding

This study was supported by the grants from the National Key Research and Development Program of China (No.2016YFC0104801), National Natural Science Foundation of China (No. 81730050).

Conflicts of interest

None.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394–424. doi: 10.3322/caac.21492.
2. Bahl M, Barzilay R, Yedidia AB, Locascio NJ, Yu L, Lehman CD. High-risk breast lesions: A machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision. Radiology 2018;286:810–818. doi: 10.1148/radiol.2017170549.
3. Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, *et al*. Breast cancer screening with mammography: Overview of Swedish randomised trials. Lancet 1993;341:973–978. doi: 10.1016/0140-6736(93)91067-v.
4. Cho N, Jang M, Lyou CY, Park JS, Choi HY, Moon WK. Distinguishing benign from malignant masses at breast US: Combined US elastography and color Doppler US-influence on radiologist accuracy. Radiology 2012;262:80–90. doi: 10.1148/radiol.11110886.
5. Ozdemir A, Ozdemir H, Maral I, Konus O, Yucel S, Isik S. Differential diagnosis of solid breast lesions – Contribution of Doppler studies to mammography and gray scale imaging. J Ultras Med 2001;20:1091–1101. doi: 10.7863/jum.2001.20.10.1091.
6. Sprague BL, Stout NK, Schechter C, van Ravesteyn NT, Cevik M, Alagoz O, *et al*. Benefits, harms, and cost-effectiveness of supplemental ultrasonography screening for women with dense breasts. Ann Intern Med 2015;162:157–166. doi: 10.7326/m14-0692.
7. Berg WA, Zhang Z, Lehrer D, Jong RA, Pisano ED, Barr RG, *et al*. Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. JAMA 2012;307:1394–1404. doi: 10.1001/jama.2012.388.
8. Engmann NJ, Golmakani MK, Miglioretti DL, Sprague BL, Kerlikowske K. Breast Cancer Surveillance Consortium. Population-Attributable Risk Proportion of Clinical Risk Factors for Breast Cancer. JAMA Oncol 2017;3:1228–1236. doi: 10.1001/jamaoncol.2016.6326.
9. Kuhl CK, Schrading S, Strobel K, Schild HH, Hilgers RD, Bieling HB. Abbreviated breast magnetic resonance imaging (MRI): First postcontrast subtracted images and maximum-intensity projection – A novel approach to breast cancer screening with MRI. J Clin Oncol 2014;32:2304–2310. doi: 10.1200/jco.2013.52.5386.
10. Lång K, Andersson I, Rosso A, Tingberg A, Timberg P, Zackrisson S. Performance of one-view breast tomosynthesis as a stand-alone breast cancer screening modality: results from the Malmö Breast Tomosynthesis Screening Trial, a population-based study. Eur Radiol 2016;26:184–190. doi: 10.1007/s00330-015-3803-3.
11. Tagliafico AS, Calabrese M, Mariscotti G, Durando M, Tosto S, Monetti F, *et al*. Adjunct screening with tomosynthesis or ultrasound in women with mammography-negative dense breasts: Interim report of a prospective comparative trial. J Clin Oncol 2016;34:1882–1888. doi: 10.1200/jco.2015.63.4147.
12. Fullman N, Yearwood J, Abay SM, Abbafati C, Abd-Allah F, Abdela J, *et al*. Measuring performance on the Healthcare Access and Quality Index for 195 countries and territories and selected subnational locations: A systematic analysis from the Global Burden of Disease Study 2016. Lancet 2018;391:2236–2271. doi: 10.1016/s0140-6736(18)30994-2.
13. Chen YL, Chen JJ, Chang C, Gao Y, Wu J, Yang WT, *et al*. Sclerosing adenosis: Ultrasonographic and mammographic findings and correlation with histopathology. Mol Clin Oncol 2017;6:157–162. doi: 10.3892/mco.2016.1108.
14. Heer R, Shrimankar J, Griffith CDM. Granulomatous mastitis can mimic breast cancer on clinical, radiological or cytological examination: a cautionary tale. Breast 2003;12:283–286. doi: 10.1016/s0960-9776(03)00032-8.
15. Zuo W, Yu J. Ruxian Jibingxue [in Chinese]. Beijing: People's Medical Publishing House, 2019: 3-991.
16. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, *et al*. Artificial intelligence in cancer imaging: clinical challenges and applications. Cancer J Clin 2019;69:127–157. doi: 10.3322/caac.21552.
17. Ghosh A. Artificial intelligence using open source BI-RADS data exemplifying potential future use. J Am Coll Radiol 2019;16:64–72. doi: 10.1016/j.jacr.2018.09.040.

18. Lin TY, Dollar P, Girshick R, He KM, Hariharan B, Belongie S. Feature pyramid networks for object detection. USA: IEEE, 2017: 936-944 [2020-04-20]. https://ieeexplore.ieee.org/document/8099589
19. Lin TY, Goyal P, Girshick R, He KM, Dollar P. Focal loss for dense object detection. IEEE I Conf Comp Vis 2017;2999–3007. doi: 10.1109/cvpr.2017.106.
20. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–357. doi: 10.1613/jair.953.
21. Huang G, Li Y, Pleiss G, Liu Z, Hopcroft J, Weinberger K. Snapshot ensembles: Train 1, get M for free. Int Conf Learning Represent 2017. France: OpenReview.net, 2017 [2020-04-30]. https://openreview.net/forum?id=BJYwwY9ll
22. Gao Y, Zhang ZD, Li S, Guo YT, Wu QY, Liu SH, *et al*. Deep neural network-assisted computed tomography diagnosis of metastatic lymph nodes from gastric cancer. Med J 2019;132:2804–2811. doi: 10.1097/CM9.0000000000000532.
23. Liu SL, Li S, Guo YT, Zhou YP, Zhang ZD, Li S, *et al*. Establishment and application of an artificial intelligence diagnosis system for pancreatic cancer with a faster region-based convolutional neural network. Med J 2019;132:2795–2803. doi: 10.1097/CM9.0000000000000544.
24. Ding L, Liu GW, Zhao BC, Zhou YP, Li S, Zhang ZD, *et al*. Artificial intelligence system of faster region-based convolutional neural network surpassing senior radiologists in evaluation of metastatic lymph nodes of rectal cancer. Med J 2019;132:379–387. doi: 10.1097/CM9.0000000000000095.
25. Tan KP, Mohamad Azlan Z, Rumaisa MP, Siti Aisyah Murni MR, Radhika S, Nurismah MI, *et al*. The comparative accuracy of ultrasound and mammography in the detection of breast cancer. Med J Malaysia 2014;69:79–85.

How to cite this article: Yu TF, He W, Gan CG, Zhao MC, Zhu Q, Zhang W, Wang H, Luo YK, Nie F, Yuan LJ, Wang Y, Guo YL, Yuan JJ, Ruan LT, Wang YC, Zhang RF, Zhang HX, Ning B, Song HM, Zheng S, Li Y, Guang Y. Deep learning applied to two-dimensional color Doppler flow imaging ultrasound images significantly improves diagnostic performance in the classification of breast masses: a multicenter study. Chin Med J 2021;134:415–424. doi: 10.1097/CM9.0000000000001329