Diagn Interv Radiol 2021; 27:315–322

© Turkish Society of Radiology 2021

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Application of ultrasonic dual-mode artificially intelligent architecture in assisting radiologists with different diagnostic levels on breast masses classification

Chunxiao Li
Jiajun Li
Tao Tan
Kun Chen
Yi Xu
Rong Wu

PURPOSE
We aimed to compare the diagnostic performance and interobserver variability in breast tumor classification with or without the aid of an innovative dual-mode artificial intelligence (AI) architecture, which can automatically integrate information from ultrasonography (US) and shear-wave elastography (SWE).

METHODS
Diagnostic performance assessment was performed with a test subset, containing 599 images (from September 2018 to February 2019) from 91 patients including 64 benign and 27 malignant breast tumors. Six radiologists (three inexperienced, three experienced) were assigned to read images independently (independent diagnosis) and then make a secondary diagnosis with the knowledge of AI results. Sensitivity, specificity, accuracy, receiver-operator characteristics (ROC) curve analysis and Cohen's κ statistics were calculated.

RESULTS
In the inexperienced radiologists' group, the average area under the ROC curve (AUC) for diagnostic performance increased from 0.722 to 0.765 ($p = 0.050$) with secondary diagnosis using US-mode and from 0.794 to 0.834 ($p = 0.019$) with secondary diagnosis using dual-mode compared with independent diagnosis. In the experienced radiologists' group, the average AUC for diagnostic performance was significantly higher with AI system using the US-mode (0.812 vs. 0.833, $p = 0.039$), but not for dual-mode (0.858 vs. 0.866, $p = 0.458$). Using the US-mode, interobserver agreement among all radiologists improved from fair to moderate ($p = 0.003$). Using the dual-mode, substantial agreement was seen among the experienced radiologists (0.65 to 0.74, $p = 0.017$) and all radiologists (0.62 to 0.73, $p = 0.001$).

CONCLUSION
AI assistance provides a more pronounced improvement in diagnostic performance for the inexperienced radiologists; meanwhile, the experienced radiologists benefit more from AI in reducing interobserver variability.

From the Department of Ultrasound (C.L., R.W. ✉ *wurong7111@163.com*), Shanghai Jiao Tong University School of Medicine, Shanghai General Hospital Shanghai, China; Shanghai Key Lab of Digital Media Processing and Transmission (J.L., K.C., Y.X. ✉ *xuyi@sjtu.edu.cn*), Shanghai Jiao Tong University, Shanghai, China; Department of Mathematics and Computer Science (T.T.), Centre for Analysis, Scientific Computing, and Applications W&I, Eindhoven University of Technology, Eindhoven, Netherlands.

Received 20 January 2020; revision requested 30 March 2020; last revision received 18 May 2020; accepted 26 May 2020.

Published online 19 April 2021.

DOI 10.5152/dir.2021.20018

As a global medical problem, breast cancer is the most commonly diagnosed cancer and the leading cause of cancer death among females (1). The development of imaging technologies have made great contribution to breast cancer diagnosis (2). Ultrasonography (US) is a valuable supplemental scanning modality for women with dense breasts (3–5). Radiologists make final diagnosis based on the morphological characteristics of breast masses according to the interpretation of Breast Imaging Reporting and Data System (BI-RADS) (6). For US BI-RADS classification, interobserver agreement ranging from fair to substantial has been reported among radiologists with different experience levels (7). Therefore, the subjectivity and empirical dependence aroused controversy towards its true diagnostic efficacy. Many studies have shown that the combined use of US-mode and ultrasound elastography has greater diagnostic efficacy than US-mode alone (8–11). However, until now, there was no consistent standard value of shear-wave elastography (SWE) in differentiation between malignant and benign breast masses which might result from the numerous reference parameters including maximum (Emax) and standard deviation (SD) (12). Both US-mode and SWE provide us with abundant uncertain information, which inexperienced radiologists cannot extract from, interpret, and utilize accurately. Therefore,

You may cite this article as: Li C, Li J, Tan T, Chen K, Xu Y, Wu R. Application of ultrasonic dual-mode artificially intelligent architecture in assisting radiologists with different diagnostic levels on breast masses classification. Diagn Interv Radiol 2021; 27:315–322

the standardization and precision diagnosis of US and SWE are the challenges and research hotspots of ultrasound technology currently.

The field of biomedical image analysis benefited substantially from rapid developments in artificial intelligence (AI) techniques. Previous studies on traditional computer-aided diagnosis (CAD) for breast cancer have shown that the segmentation of US and SWE images and integration of extracted features achieved better classification performance than their individual uses (13–15). With the CAD system, the average diagnostic performance of radiologists was improved or at least comparable to the independent diagnosis based on breast US images (16, 17). An observational study conducted by Van Zelst et al. (18) concluded that CAD software for automated breast US may speed up the screening time without compromising the screening performance of radiologists. As a revolutionary advance, the deep learning algorithm (convolution neural network for image processing) has a strong feature extraction ability, which can extract higher level features rather than superficial features and provide more possibilities for medical image processing. Recently few studies have shown the great potential of deep learning framework in ultrasound image processing of breast lesions (19–22).

However, the abovementioned observational studies were focused only on a single mode of ultrasonic breast images. Throughout these papers, there was no deep learning model which is capable of integrating US with SWE images of breast lesions rationally. In clinical practice, comprehensive analysis with dual-mode ultrasound is of great importance for breast lesions. To meet the clinical needs, we designed a novel module named "shared latent subspace learning," which captures the inter-mode relationship between SWE-mode and US-mode via a shared-parameter Dense Block that is optimized under adversarial loss and orthogonality constraint. Similar to the normal working mode of radiologists and just like the workflow of CAD software (23), radiologists will combine the diagnostic results given by software and achieve the secondary diagnostic results after comprehensive consideration. Therefore, we conducted a dual-mode reader study where we investigated the benefits of incorporating an innovative AI architecture. We compared the diagnostic performance and interobserver agreement of different radiologists in breast mass classification through independent diagnosis and secondary diagnosis.

Main points

- Dual-mode AI architecture has great potential in clinical dual-mode task of ultrasound techniques in assisting the classification of breast masses.
- The inexperienced radiologists benefit more from AI results in terms of diagnostic performance improvement.
- The experienced radiologists benefit more from AI in reducing interobserver variability.

Methods

Institutional Review Board approval was obtained at our hospital (No. 2019KY055). Written informed consent was waived by the Institutional Review Board.

From September 2018 to February 2019, 2008 images of 395 benign lesions and 959 images of 130 malignant lesions (size, 17.6±10.5 mm; range, 4.2–85 mm) from 525 women (mean age, 47.2±15.4 years; age range, 17–91 years) were collected as the whole dataset used in this study. All case results were confirmed by pathology, with complete information of US and SWE of imaging.

The super linear probe L-15-4 of ultrasound device Aixplorer (Super Imagine) was used for imaging data collection. After full examination of the breast lesions in both of transverse and longitudinal sections, both US and SWE images were saved and then used for further analysis. A target detection labeling tool, LabelImg, was used to crop the region of interest (ROI) from the original images according to the labeling results from radiologists. The labeling results were a set of rectangular boxes to show the potential nodal region, manually delineated and confirmed by two radiologists (with over 5 years of experience in breast ultrasound interpretation). The overall dataset was split into 5 parts, named as Subset 0, 1, 2, 3 and 4. Subsets of 0 to 3 were used for training and validation, while Subset 4 was the holdout test set. All hyperparameters are validated via 4-fold cross validation on subsets of 0 to 3.

Dual-mode artificial intelligence

In order to improve the diagnostic performance of traditional breast cancer CAD system, we designed a dual-mode AI system which integrates complementary information from US-mode and SWE-mode and thus enhances the feature representations of each mode image (Fig. 1). The system uses deep learning convolutional neural networks (CNN) to extract mode-specific feature representations of input images (Fig. 1a). Specifically, we constructed two DenseNets with same network structure but different mode parameters to extract mode-specific features, since DenseNet, as the backbone of our system, has indicated convincing classification accuracy in image recognition tasks.

The architecture of DenseNet utilized in the proposed AI system is illustrated in Fig. 1b. The whole DenseNet consists of four stacked blocks (i.e., DenseBlock), each of which extracts different levels of features from the input image, from shallow texture features to deep semantic features. The features are calculated through a series of convolution operations which are represented as four internal nodes contained in each DenseBlock. The most remarkable feature of DenseNet is that any two nodes in each DenseBlock are directly connected, which indicates that the shallow features can be reused in the subsequent feature calculation process, and thereby enhances the robustness of feature representation. Therefore, we choose to leveraging the power of dense connectivity and utilize DenseNet as the backbone network of the image feature extraction module in the proposed AI system.

Moreover, the last DenseBlock (i.e., DenseBlock 4) in either DenseNet shares the same set of parameters, being optimized to learn a shared latent subspace and extract the shared features from both modes. The main purpose of learning a shared latent subspace is to align features of different ultrasound modalities in a higher dimensional spatial embedding, and thereby learn the joint distribution between these two modal features. This shared latent subspace is constructed to learn the cross-modal relationship, where the complementary information from two modalities can be expressed by modeling the joint distribution density function of the data. Specifically, the main technical route to learn such a subspace is to decouple features into mode-specific features and shared features between different ultrasound modalities, shown as two independent branches of each mode in Fig. 1a.

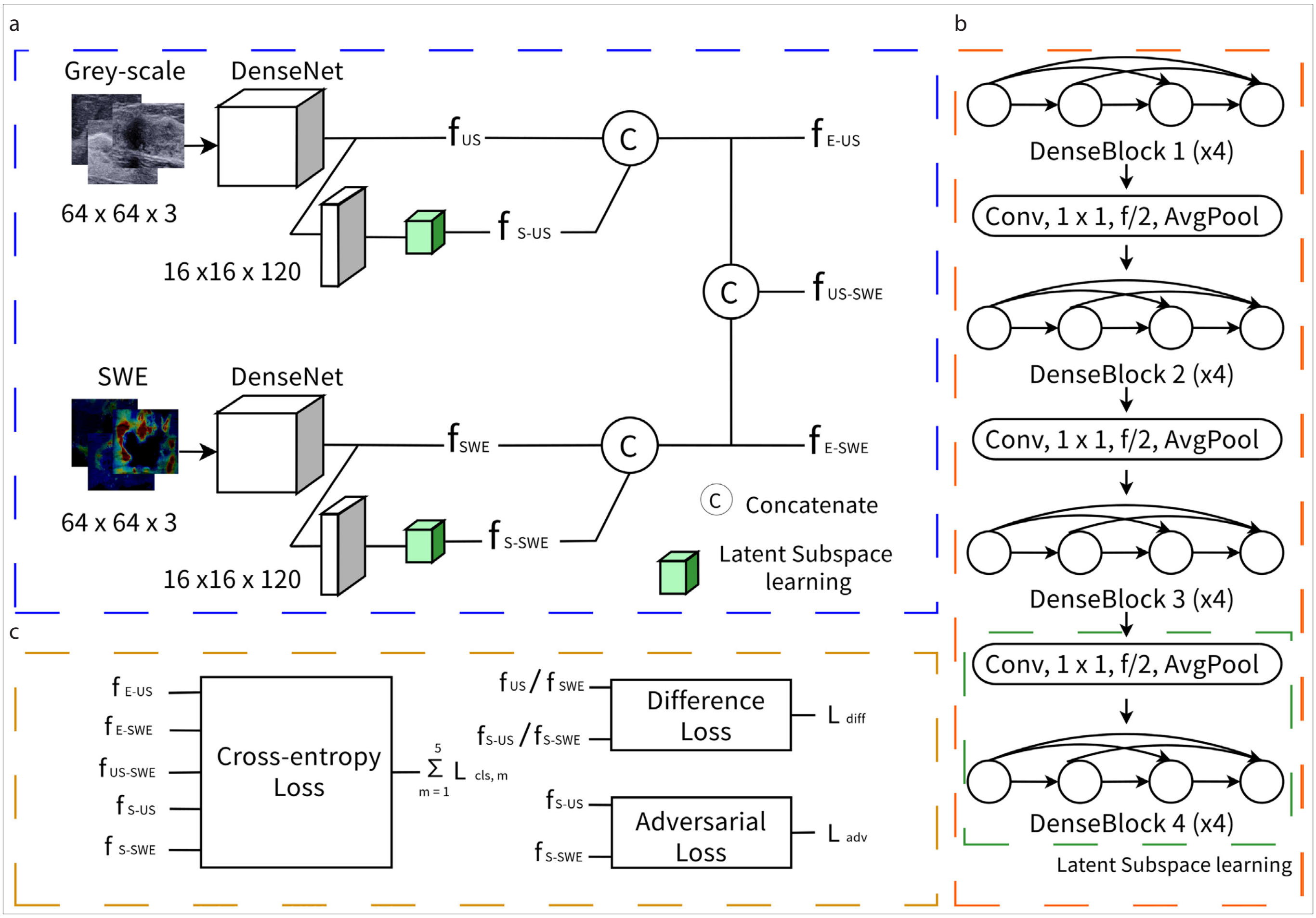


Figure 1, a–c. The architecture of the proposed system **(a–c)**. The three parts show the three kinds of extracted feature sets that were used for classification outputs **(a)**, the illustration of the framework in DenseNet **(b)**, and the illustration of each loss function used during training **(c)**.

Through optimizing the adversarial loss L_{adv}, difference loss L_{diff} and cross-entropy loss $L_{cls,m}$, where adversarial loss forces "DenseBlock 4" or the so called "mode-shared encoder" to extract only the mode-invariant features, difference loss prevents redundant mode-invariant features from being present in both latent mode-shared subspace and latent mode-specific subspace, and cross-entropy loss measures the similarity between the predicted results and the actual results, we successfully learn a shared latent subspace of two ultrasound modalities, capturing their potential inter-mode relationship. Another important motivation to decouple mode-specific features and shared features is to depress feature redundancy problems that may reduce model's classification accuracy. Finally, it is natural and effective to fuse these two decoupled feature vectors by concatenation before being fed into feature classifiers.

Furthermore, a triple-task learning framework makes our system capable of generating three benign-malignant classification results based on different extracted features, including enhanced US image features (f_{E-US}) which is composed of US-specific features and US-shared features, enhanced SWE image features (f_{E-SWE}), which is composed of SWE-specific features and SWE-shared features and the fused US-SWE dual-mode features using concatenation of f_{E-US} *and* f_{E-SWE}, together with two categorical probabilities to indicate the effectiveness of shared features (f_{S-US}, f_{S-SWE}) (Fig 1c).

The overall architecture of our proposed system is illustrated in Fig 1. The system is flexible to deal with both dual-mode data and single-mode data, while the existing breast cancer CAD systems are limited to either dual-mode inputs or single-mode inputs. Moreover, the design of feature concatenation enables our system to provide reliable classification results in cases when one single mode information is missing or obscure.

Study design and data analysis

Six radiologists who had performed more than 1000 breast US and 150 breast SWE examinations in the last 1-2 years were involved in this reader study. Because 5 years of experience is usually required before promotion to attending doctor, we divided them into inexperienced group (reader 1, reader 2, and reader 3) and experienced group (reader 4, reader 5, and reader 6). General US experience of six radiologists was 2, 4, 5, 9, 11 and 15 years. For inexperienced and experienced readers, numbers of breast US imaging studies (performing and interpreting) they evaluated per year was more than 1000 and 2000, respectively.

The readers first read ultrasound images independently; then, after an interval of 7 days, they did a secondary diagnosis with the aid of AI (secondary-diagnosis mode). At first, all readers were trained by reading 20 series of breast images (including US-mode and SWE-mode) in which 10 were read just by independent-diagnosis mode, and 10 through secondary-diagnosis mode. In secondary-diagnosis mode, radiologists read breast images without AI first, then combined the AI results to make the final diagnosis. In the following observer study, first, all readers were asked to read data from our internal test dataset (Subset 4) independently, and then, all readers made the other diagnosis after reading AI interpretations. The image numbers and sequences of the same patient were presented in random ways to different readers. Based on the BI-RADS category and reference threshold of SWE value set by instrument manufacture, readers classified breast lesions into categories ranging from 3 to 5: first, Emax ≥60 kPa was set as cutoff value; second, BI-RADS 4a mass downgraded to BI-RADS 3 combined with Emax ≤40 kPa, BI-RADS 3 mass upgraded to BI-RADS 4a combined with Emax ≥50 kPa. Breast lesions of category 3 were regarded as benign, and those of categories 4a, 4b, 4c and 5 were regarded as malignant. In total, we got four diagnosis results for each breast lesion manually (2 ultrasound modes × 2 reading modes). For AI system, the diagnostic result of multiple pictures of one mass was given by the definition of majority rule. Finally, all lesions were dichotomized into 0 for benign and 1 for malignant for diagnostic performance assessment. The pathological diagnosis after surgery or needle biopsy was used as reference standard.

Statistical power estimation

In the test data set, there were 186 positive and 413 negative samples. For real-world radiologists, on US-mode, assuming the AUC of independent-diagnosis mode to be 0.70, the AUC of secondary-diagnosis mode to be 0.75, the statistical power was 0.61. On dual-mode, assuming the AUC of independent-diagnosis mode to be 0.80, the AUC of secondary-diagnosis mode to be 0.85, the statistical power was 0.71. Alpha was set to be 0.05 in both cases, power estimation were calculated with PASS 11.

Statistical analysis

The descriptive data of mass size and patient age were shown as mean ± standard deviation. The diagnosis performance of independent-diagnosis mode and secondary-diagnosis mode of all readers was assessed according to four measures: sensitivity, specificity, accuracy and the area under the ROC curve. With the paired diagnostic results (without AI and with AI), the two diagnostic modes were evaluated on the same set of patients, therefore, we used McNemar's test for dependent proportions. The McNemar test was applied on sensitivity, specificity, accuracy and the z-test was applied on AUC to measure the performance differences. ROC curve for the inexperienced and experienced radiologists were produced by group averaging (mean with standard error), respectively. The AUC values of all the six radiologists with or without AI were given as median (min–max). Inter-observer variability was assessed with the Weighted Cohen's kappa (κ) statistics: slight agreement (≤0.2), fair agreement (0.21–0.40), moderate agreement (0.41–0.60), substantial agreement (0.61–0.80), or almost perfect agreement (0.81–1.00). We assessed the weighted κ values between readers at two different diagnostic mode. For the six readers, a total of 15 pairs of observers yielded 15 κ values. Observer agreement analysis was performed for three observer groups: between the inexperienced and experienced radiologists (9 pairs of observers yielded 9 κ values), within inexperienced radiologists (3 pairs of observers yielded 3 k values) and within experienced radiologists (3 pairs of observers yielded 3 κ values). Average κ values of each group without and with AI were calculated. Statistical analyses other than Az comparison were computed using the SPSS software (version 20.0), while Medcalc software (Version 16.2) was employed to make Az comparison. $p < 0.05$ was regarded as statistically significant.

Table 1. Comparison of AUCs between different reading groups and AI

	US-mode		Dual-mode	
	Independent-D	Secondary-D	Independent-D	Secondary-D
AI vs. Inexperienced	0.781 vs. 0.722	0.781 vs. 0.765	0.892 vs. 0.794	0.892 vs. 0.834
p	0.116	0.641	<0.001	0.007
AI vs. Experienced	0.781 vs. 0.812	0.781 vs. 0.833	0.892 vs. 0.858	0.892 vs. 0.866
p	0.264	0.047	0.062	0.081
AI vs. All readers	0.781 vs. 0.793	0.781 vs. 0.812	0.892 vs. 0.835	0.892 vs. 0.852
p	0.566	0.409	<0.001	0.001
Inexperienced vs. Experienced	0.722 vs. 0.812	0.765 vs. 0.833	0.794 vs. 0.858	0.834 vs. 0.866
p	0.008	0.025	0.011	0.123

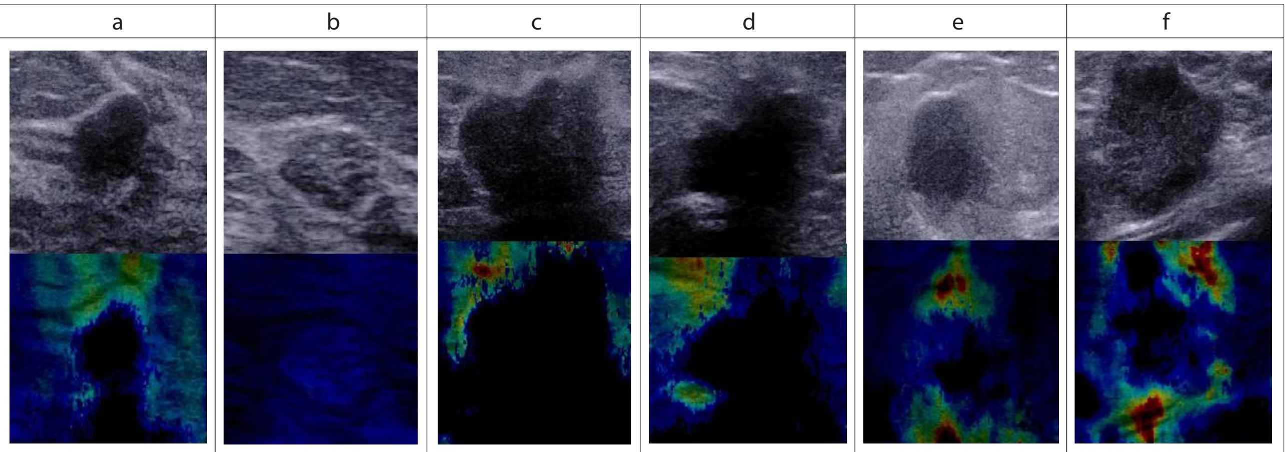
AI, artificial intelligence; Independent-D, independent-diagnosis; Secondary-D: secondary-diagnosis; AUC, area under the receiver-operator characteristics curve.

Results

As shown in Fig. 2, the classification results based on dual-mode image pairs can suppress the classification errors of US images (Fig. 2a, 2b) or SWE images (Fig. 2c, 2d) via fully using the inter-mode knowledge. Due to medical instrument operation conditions, some information is often missing on SWE images and sometimes the missing information will occupy the main part of the image (As the black part shown in Fig. 2c, 2d), which makes the diagnosis unreliable. In such a case, US images can otherwise provide a reliable basis for diagnosis. In a few cases, the classification results based on two modes may provide misdiagnosis even if both single-mode images are correctly classified by US-only image and SWE-only image classifiers (Fig. 2e, 2f). The AUCs of each radiologist, each reader group and all the readers with and without AI system are shown in Tables 1–4.

The experienced radiologist group showed higher AUC than the inexperienced radiologist group on both US-mode (0.812 vs. 0.722, $p = 0.008$) and dual-mode (0.858 vs. 0.794, $p = 0.011$) at independent-diagnosis mode. AI system showed comparable AUC to the inexperienced radiologist group (0.781 vs. 0.722, $p = 0.116$) and experienced radiologist group (0.781 vs. 0.812, $p = 0.264$) and all the readers (0.793 vs. 0.781, $p = 0.566$) on US-mode breast lesion image analysis. On dual-mode, AI system obtained higher AUC than the inexperienced radiologist group (0.892 vs. 0.794, $p < 0.001$) and was equivalent to the experienced radiologist group (0.892 vs. 0.858, $p = 0.062$). On average, AI system showed higher AUC than all the radiologists on dual-mode image analysis (0.892 vs. 0.835, $p < 0.001$). Results are shown in Tables 1 and 2.

The experienced radiologist group showed higher AUC than the inexperienced



	a	b	c	d	e	f
Ground truth	Benign	Benign	Malignant	Malignant	Benign	Malignant
Classification	US-only image:					
	Malignant	Malignant	Malignant	Malignant	Benign	Malignant
	SWE-only image:					
	Benign	Benign	Benign	Benign	Benign	Malignant
	US-SWE image pair:					
	Benign	Benign	Malignant	Malignant	Malignant	Benign

Figure 2, a–f. Exemplary images of some cases that were wrongly classified by at least one classification task **(a–d)**. US-mode images and SWE-mode images are a set of model-aligned images of the same lesion. In panels **(e, f)**, correct diagnoses were obtained from US-mode and SWE-mode images when analyzed separately; however, US and SWE images together yielded incorrect diagnoses.

Table 2. Comparison of AUCs of different reading modes

	US-mode AUC (mean±SE)			Dual-mode AUC (mean±SE)		
	Independent-D	Secondary-D	*p*	Independent-D	Secondary-D	*p*
Reader						
1	0.672±0.05	0.727±0.05	0.118	0.806±0.05	0.837±0.04	0.339
2	0.687±0.05	0.760±0.05	0.012	0.730±0.05	0.782±0.05	0.092
3	0.807±0.04	0.807±0.04	1.000	0.848±0.04	0.882±0.04	0.113
Inexperienced group (1–3)	0.722±0.03	0.765±0.03	0.050	0.794±0.03	0.834±0.03	0.019
4	0.824±0.04	0.848±0.04	0.079	0.824±0.04	0.840±0.04	0.317
5	0.779±0.05	0.816±0.05	0.149	0.845±0.04	0.863±0.04	0.317
6	0.834±0.05	0.834±0.05	1.000	0.895±0.04	0.905±0.04	0.594
Experienced group (4–6)	0.812±0.03	0.833±0.03	0.039	0.858±0.02	0.866±0.02	0.458
All readers*	0.793 (0.672–0.834)	0.812 (0.706–0.848)	0.001	0.835 (0.730–0.895)	0.852 (0.782–0.905)	0.060
AI	0.781±0.03			0.892±0.02		0.026

US, ultrasonography; AUC, area under the receiver-operator characteristics curve; SE, standard error; Independent-D, independent diagnosis; Secondary-D, secondary diagnosis; AI, artificial intelligence.
*Data was shown as median (min-max).

radiologist group on US-mode (0.833 vs. 0.765, *p* = 0.025) and no statistical difference on dual-mode (0.866 vs. 0.834, *p* = 0.123) at secondary-diagnosis mode (Table 2). With the aid of AI system, the average AUC for the inexperienced radiologists increased from 0.722 to 0.765 (*p* = 0.050) while the one for the experienced radiologist group increased from 0.812 to 0.833 (*p* = 0.039) on US-mode. Finally, the average AUC of all the radiologists showed a significant increase from 0.793 to 0.812 with AI system (*p* < 0.001) on US-mode. On dual-mode, the average AUC of the inexperienced radiologists was significantly improved by AI system at secondary-diagnosis mode (0.794 vs. 0.834, *p* = 0.019), but not for the experienced radiologist group (0.858 vs. 0.866, *p* = 0.458) and all the radiologists (0.835 vs. 0.852, *p* = 0.060).

The accuracy, sensitivity and specificity of AI system of US-mode and dual-mode were 85.9%, 59.3%, 96.9% and 92.4%, 81.5%, 96.9%, respectively (Table 3). The sensitivi-

Table 3. Comparison of diagnostic performance on different reading modes (independent diagnosis vs. secondary diagnosis)

	US-mode			Dual-mode		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Reader 1	65.9 vs. 73.6	70.4 vs. 70.4	64.1 vs. 75.0	80.2 vs. 84.6	81.5 vs. 81.5	79.7 vs. 85.9
p	<0.001	1.000	<0.001	<0.001	1.000	<0.001
Reader 2	72.5 vs. 81.3	59.3 vs. 63.0	78.1 vs. 89.1	72.5 vs. 81.3	74.1 vs. 70.4	71.9 vs. 85.9
p	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Reader 3	75.8 vs. 75.8	92.6 vs. 92.6	68.7 vs. 68.7	84.6 vs. 87.9	85.2 vs. 88.9	84.4 vs. 87.5
p	1.000	1.000	1.000	<0.001	0.001	<0.001
Reader 4	81.3 vs. 84.6	85.2 vs. 85.2	79.7 vs. 84.4	81.3 vs. 83.5	85.2 vs. 85.2	79.7 vs. 82.8
p	<0.001	1.000	<0.001	<0.001	1.000	<0.001
Reader 5	82.4 vs. 84.6	66.7 vs. 74.1	89.2 vs. 89.2	86.8 vs. 86.8	81.5 vs. 85.2	87.5 vs. 87.5
p	<0.001	<0.001	1.000	1.000	<0.001	1.000
Reader 6	85.9 vs. 85.9	77.8 vs. 77.8	89.2 vs. 89.2	91.3 vs. 91.3	88.9 vs. 85.2	92.2 vs. 93.7
p	1.000	1.000	1.000	1.000	0.001	<0.001
AI	85.9	59.3	96.9	92.4	81.5	96.9

US, ultrasonography; AI, artificial intelligence.

Table 4. Interobserver agreements of experienced and inexperienced radiologists using different diagnosis modes

	Independent-D		Secondary-D		Improvement	
	κ (95% CI)	*p*	κ (95% CI)	*p*	κ (95% CI)	*p*
US-mode						
Inexperienced	0.50 (0.39–0.61)	0.003	0.53 (0.44–0.61)	0.001	0.03 (-0.16 to 0.22)	0.562
Experienced	0.62 (0.54–0.70)	0.001	0.65 (0.51–0.79)	0.003	0.03 (-0.03 to 0.09)	0.187
All readers	0.39 (0.29–0.50)	<0.001	0.53 (0.47–0.59)	<0.001	0.14 (0.06–0.22)	0.003
Dual-mode						
Inexperienced	0.61 (0.42–0.81)	0.005	0.74 (0.70–0.77)	<0.001	0.12 (-0.10 to 0.34)	0.141
Experienced	0.65 (0.46–0.83)	0.004	0.74 (0.60–0.88)	0.002	0.09 (0.04–0.14)	0.017
All readers	0.62 (0.54–0.70)	<0.001	0.73 (0.68–0.79)	<0.001	0.11 (0.06–0.16)	0.001

Independent-D, independent diagnosis; Secondary-D, secondary diagnosis; 95% CI, 95% confidence interval; US, ultrasonography.

ty and accuracy are significantly improved ($p = 0.013$ and $p < 0.001$) by integrating US-mode and SWE-mode images of breast lesions. For reader 1, although the sensitivity was not improved by AI system, the specificity was greatly promoted from 64.1% to 75.0% ($p < 0.001$) and from 79.7% to 85.9% ($p < 0.001$), respectively, under two modes. For reader 2, although the sensitivity was slightly decreased from 74.1% to 70.4% ($p < 0.001$), the other parameters became statistically higher after using the AI system ($p < 0.001$ for all). For reader 3, the sensitivity and specificity of dual-mode benefitted from AI ($p = 0.001$ and $p < 0.001$). In the experienced radiologist group, the specificity of two different modes of reader 4 and the specificity of dual-mode of reader 6 increased significantly ($p < 0.001$, for all). The sensitivity of reader 5 for two modes ($p < 0.001$ and $p < 0.001$) and reader 6 for dual-mode ($p = 0.001$) were all improved with the AI system. Regarding the accuracy, except reader 3 on US-mode and reader 5 on dual-mode and reader 6 on two modes, the accuracy of secondary-diagnose mode was significantly improved for all radiologists with the aid of AI system ($p < 0.001$, for all). Overall, we observed that the accuracy, sensitivity and specificity among the most of inexperienced and experienced radiologists were improved with our AI schema.

Table 4 shows a summary of the interobserver variability among the six radiologists at independent-diagnosis mode and secondary-diagnosis mode. In all, the total interobserver agreements between inexperienced and experienced radiologists were improved from fair (κ=0.39) without AI to moderate (κ=0.53) with AI on US-mode ($p = 0.003$). On dual-mode, agreement among inexperienced and experienced radiologist groups without and with AI were all substantial. Moreover, the metric value of κ was significantly improved from 0.62 to 0.73 with AI ($p = 0.001$). It also showed significant improvement in the agreement among the experienced radiologists on dual-mode: the agreement was substantial for both independent-diagnosis mode and secondary-diagnosis mode (κ=0.65 and 0.74, $p = 0.017$).

Discussion

Many researches on US and SWE demonstrated the great potential of AI in breast ultrasound image analysis. Consistent with previous studies (14, 24), our results showed that the dual-mode utilization of our AI system was more helpful in differentiating breast lesions with AUC of 0.892 compared with US-mode with AUC of 0.781. Moreover, the AUC of our AI schema was comparable to our radiologists regardless of the experience level on US-mode. On dual-mode, the AUC of AI system was higher than the inexperienced radiologists group and no different from the experienced radiologists group. Practically, this manifests the great application potential of our AI architecture.

Choi et al. (25) found that with the application of CAD in B-mode breast ultrasound interpretation, the specificity was significantly improved (76.6% to 80.3%) without a change in the sensitivity (91.7%) for the experienced readers while both sensitivity and specificity were improved for the inexperienced readers (75.0% to 83.3% and 71.8% to 77.1%), which is in accordance with our study. As shown in Table 2, we observed that the accuracy, sensitivity and specificity among the less experienced and experienced radiologists were all improved with our AI schema. The most notable contrast was shown on reader 6 with excellent independent diagnostic performance and reader 2 at secondary-diagnosis mode with the help of AI system.

As for the results shown in Table 3, with the aid of AI system, the mean AUCs of all radiologists increased on US-mode breast ultrasound which is not in line with previous studies. This might be due to superiority of the AI system used. Lee et al. (26) reported that the diagnostic performance of the inexperienced group after combination with the CAD result was significantly improved, whereas that of the experienced group did not change after combination with the CAD result on breast ultrasound. Although the experienced radiologists showed no superiority to AI system on US-mode (0.812 vs. 0.781), with the excellent ability of excavating information, they achieved a statistically significantly higher AUC of 0.833 at the secondary-diagnosis mode. In terms of dual-mode, AI system explored much more fruitful information and achieved a comparable AUC to the experts (0.892 vs. 0.858), but it showed no substantial improvement for the experts. As we mentioned above, although experts are more likely to reach a higher diagnosis level based on sufficient information from dual-mode images, it is also clear that there is a limit to human ability. Whereas, with the help of the AI system, the mean AUC of dual-mode of the inexperienced readers was increased significantly (0.794 vs. 0.834) and was comparable to the experts (0.834 vs. 0.866), which was not observed on US-mode (0.765 vs. 0.833). This may also reflect the fact that, with multi-mode data, AI can extract more useful diagnostic information to assist less experienced doctors. As a whole, our AI system is more useful for the inexperienced readers which was also suggested by many studies (23, 25–27). More importantly, our observational study was the first to explore the role of AI in assisting the diagnosis of dual-mode breast ultrasound images.

Lee et al. (28) found that when SWE was added to B-mode breast US, better interobserver agreement was obtained. Park et al. (27) concluded that interobserver variability of breast lesion US assessment was significantly improved after CAD combination regardless of the radiologist's experience. In our experimental study, the interobserver variability was reduced by using AI for all the radiologists on US-mode and dual-mode. Moreover, the interobserver agreement was improved among the experienced radiologists on dual-mode. With regard to the inexperienced radiologists, the interobserver agreement was not significantly changed with AI assistance on US-mode or dual-mode. The possible explanation could be that some less experienced doctors with insufficient diagnostic capacity and low diagnostic confidence may not handle external disturbance (AI) appropriately. On the contrary, the experienced readers could accurately extract the additional information inspired by AI in addition to the limited information which they acquired visually. The final result was that the experienced radiologists drew near to the real results more consistently, while diagnostic variation among the inexperienced has increased. Therefore, the ideal objective of the current AI technique is to serve the inexperienced doctors better and could be on a par with the experts, which requires technological development and the gradual adaptation of the radiologists.

US and SWE images used in our study were all standard sections; there were problems regarding the reproducibility of the CAD results when examinations were performed by less-experienced operators. In daily potential use, the diagnostic decision should be made by experienced radiologist. From our results, this dual-mode AI system is on par with the experienced radiologist, therefore, could be used as a reference for clinical application, especially for inexperienced radiologist. In further investigations, cine loops rather than still images could be used for study which might reduce operator dependency. Second, a prospective research should be made in clinical practice; radiologists with different qualifications are required to collect pictures and input them into our AI system for a reproducibility and repeatability study. Third, in terms of model training, hard example mining and human-in-the-loop means could be utilized to further improve the final performance of our AI model by retraining the model to correct the misdiagnosed cases determined by the original model.

Our study has several limitations. First, as a retrospective study, there could be bias caused by repeated reading of the same dataset from radiologists. Second, as an internal information of our hospital, the possibility that the breast images have been seen by the readers cannot be excluded. Third, our AI modal was not trained and validated with external dataset from other hospitals. Making our AI model more robust with larger scale sample dataset and multicenter dataset in a prospective clinical trial is also our research direction.

In conclusion, our observational study demonstrated that AI showed radiologist-level performance in diagnosing both breast US and SWE images. In terms of diagnostic performance improvement, the inexperienced radiologists benefit more from the aid of AI system. In clinical practice, our AI architecture would be valuable for improving the overall diagnostic performance of breast dual-mode US images for all radiologists at different levels of experience. Regarding interobserver agreement, the variation among the experienced radiologists decreased more distinctively when using AI.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Siegel RL, Miller KD, Jemal A. Global cancer statistics. CA Cancer J Clin 2018; 68:7–30. [Crossref]
2. Wender RC, Brawley OW, Fedewa SA, Gansler T, Smith RA. A blueprint for cancer screening and early detection: advancing screening's contribution to cancer control. CA Cancer J Clin 2018; 0:1–31. [Crossref]

3. Smith RA, Andrews KS, Brooks D et al. Cancer screening in the United States, 2018: a review of current American cancer society guidelines and current issues in cancer screening. CA Cancer J Clin 2018; 68:297–316. [Crossref]
4. Hou XY, Niu HY, Huang XL, Yu G. Correlation of breast ultrasound classifications with cancer in Chinese women. Ultrasound Med Biol 2016; 42:2616–2621. [Crossref]
5. da Costa Vieira RA, Biller G, Uemura G, Ruiz CA, Curado MP. Breast cancer screening in developing countries. Clinics 2017; 72:244–253. [Crossref]
6. D'Orsi C, Sickles E, Mendelson E, Morris E. Breast imaging reporting and data system 5th ed. Reston: American College of Radiology, 2013.
7. Schwab F, Redling K, Siebert M, Schotzau A, Schoenenberger CA, Zanetti-Dallenbach R. Inter- and intra-observer agreement in ultrasound BI-RADS classification and real-time elastography Tsukuba score assessment of breast lesions. Ultrasound Med Biol 2016; 42:2622–2629. [Crossref]
8. Ophir J, Cespedes I, Ponnekanti H, et al. Elastography: a quantitative method for imaging the elasticity of biological tissues. Ultrasonic Imaging 1991; 13:111–134. [Crossref]
9. Xue Y, Yao SX, Li XD, Zhang HR. Value of shear wave elastography in discriminating malignant and benign breast lesions A meta-analysis. Medicine 2017; 96:42. [Crossref]
10. Guo RR, Lu GL, Fei BW. Ultrasound imaging technologies for breast cancer detection and management- A review. Ultrasound Med Biol 2018; 44:37–70. [Crossref]
11. Lee SH, Chung J, Choi HY, et al. Evaluation of screening US–detected breast masses by combined use of elastography and color Doppler US with B-mode US in women with dense breasts: a multicenter prospective study. Radiology 2017; 285:660–669. [Crossref]
12. Imtiaz S, MBBS, FCPS. Breast elastography: A new paradigm in diagnostic breast imaging. Appl Radiol 2018; 3:14–19. [Crossref]
13. Lo CM, Lai YC, Chou YH, Chang RF. Quantitative breast lesion classification based on multichannel distributions in shear-wave imaging. Comput Methods Programs Biomed 2015; 122:354–361. [Crossref]
14. Zhang Q, Song S, Xiao Y, Chen S, Shi J, Zheng HR. Dual-mode artificially-intelligent diagnosis of breast tumours in shear-wave elastography and B-mode ultrasound using deep polynomial networks. Med Eng Phys 2019; 64:1–6. [Crossref]
15. Moon WK, Huang YS, Lee YW, et al. Computer-aided tumor diagnosis using shear wave breast elastography. Ultrasonics 2017; 78:125–133. [Crossref]
16. Shibusawa M, Nakayama R, Okanami Y, Kashikura Y, Imai N, Nakamura T. The usefulness of a computer-aided diagnosis scheme for improving the performance of clinicians to diagnose non-mass lesions on breast ultrasonographic images. J Med Ultrasonics 2016; 43:387–394. [Crossref]
17. Ciritsis A, Rossi C, Eberhard M, Marcon M, Becker AS, Boss A. Automatic classification of ultrasound breast lesions using a deep convolutional neural network mimicking human decision-making. Eur Radiol 2019; 29:5458–5468. [Crossref]
18. van Zelst JCM, Tan T, Clauser P, et al. Dedicated computer-aided detection software for automated 3D breast ultrasound; an efficient tool for the radiologist in supplemental screening of women with dense breasts. Eur Radiol 2018; 28:2996–3006. [Crossref]
19. Zhou Y, Xu J, Liu Q, et al. A radiomics approach with CNN for shear-wave elastography breast tumor classification. IEEE Trans Biomed Eng 2018; 65:1935–1942. [Crossref]
20. Yap MH, Pons G, Marti J, et al. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J Biomed Health Inform 2018; 22:1218–1225. [Crossref]
21. Xiao T, Liu L, Li K, Qin WJ, Yu SD, Li ZC. Comparison of transferred deep neural networks in ultrasonic breast masses discrimination. BioMed Res Int 2018; 2018:4605191. [Crossref]
22. Byra M, Galperin M, Ojeda-Fournier H, et al. Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. Med Phys 2019; 46:746–755. [Crossref]
23. Zhao CY, Xiao MS, Jiang YX, et al. Feasibility of computer-assisted diagnosis for breast ultrasound: the results of the diagnostic performance of s-detect from a single center in China. Cancer Manag Res 2019; 11:921–930. [Crossref]
24. Marcomini KD, Fleury EFC, Oliveira VM, et al. Evaluation of a computer-aided diagnosis system in the classification of lesions in breast strain elastography imaging. Bioengineering 2018; 5:62. [Crossref]
25. Choi JH, Kang BJ , Baek JE, Lee HS, Kim SH. Application of computer-aided diagnosis in breast ultrasound interpretation: improvements in diagnostic performance according to reader experience. Ultrasonography 2018; 37:217–225. [Crossref]
26. Lee J, Kim S, Kang BJ, Kim SH, Park GE. Evaluation of the effect of computer aided diagnosis system on breast ultrasound for inexperienced radiologists in describing and determining breast lesions. Med Ultrason 2019; 21:239–245. [Crossref]
27. Park HJ, Kim SM, Yun BL, et al. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of breast masses on ultrasound Added value for the inexperienced breast radiologist. Medicine 2019; 98:3(e14146). [Crossref]
28. Lee SH, Cho N, Chang JM, et al. Two-view versus single-view shear-wave elastography: comparison of observer performance in differentiating benign from malignant breast masses. Radiology 2014; 270:344–353. [Crossref]