Development of a Deep Learning–Based Model for Diagnosing Breast Nodules With Ultrasound

Jianming Li, MD 🔵 *, Yunyun Bu, MD, Shuqiang Lu, PhD, Hao Pang, PhD, Chang Luo, PhD, Yujiang Liu, MD, Linxue Qian, MD*

Abbreviations
ACP, associate chief physician; AI, artificial intelligence; AP, attending physician; AUC, area under the curve; BUS, breast ultrasound; CI, confidence interval; CNN, convolutional neural network; CP, chief physician; DL, deep learning; F1$_{measure}$, harmonic mean of recall and precision; mAP, mean average precision; RP, resident physician; US, ultrasound

Objectives—Artificial intelligence (AI) has been an important addition to medicine. We aimed to explore the use of deep learning (DL) to distinguish benign from malignant lesions with breast ultrasound (BUS).

Methods—The DL model was trained with BUS nodule data using a standard protocol (1271 malignant nodules, 1053 benign nodules, and 2144 images of the contralateral normal breast). The model was tested with 692 images of 256 breast nodules. We used the accuracy, precision, recall, harmonic mean of recall and precision, and mean average precision as the indices to assess the DL model. We used 100 BUS images to evaluate differences in diagnostic accuracy among the AI system, experts (>25 years of experience), and physicians with varying levels of experience. A receiver operating characteristic curve was generated to evaluate the accuracy for distinguishing between benign and malignant breast nodules.

Results—The DL model showed 73.3% sensitivity and 94.9% specificity for the diagnosis of benign versus malignant breast nodules (area under the curve, 0.943). No significant difference in diagnostic ability was found between the AI system and the expert group (*P* = .951), although the physicians with lower levels of experience showed significant differences from the AI and expert groups (*P* = .01 and .03, respectively).

Conclusions—Deep learning could distinguish between benign and malignant breast nodules with BUS. On BUS images, DL achieved diagnostic accuracy equivalent to that of expert physicians.

Key Words—artificial intelligence; breast; deep learning; radiologist; ultrasound

In 2017, the American Cancer Society reported that breast cancer is the most common cancer in women and that it is expected to account for 29% of all new malignancies in women.[1] Globally, breast cancer is the primary cause of cancer death in women[2]; however, early detection of breast cancer can increase the effectiveness of treatment and improve the survival rate.[3]

Ultrasound (US) is playing an increasingly important role in the screening for breast cancer and its differential diagnosis. Ultrasound involves no ionizing radiation and is noninvasive and cost-effective, especially for detecting dense breast cancers, in which mammography may present some deficiencies.[4] However, breast ultrasound (BUS) examinations are highly operator dependent, and medical treatment resources are unevenly distributed. Physicians' different experience levels may affect patients' diagnoses and prognoses.[5] Under the

pressure of high-intensity work, even experienced physicians may make misdiagnoses.[6]

Many early studies on computer-aided diagnosis reported that it could improve the detection and diagnosis of breast nodules,[7,8] and artificial intelligence (AI) is being increasingly used to supplement human interpretation. There have now been many clinical trials of AI, which have shown that it can achieve an accurate analysis and diagnosis of medical images.[9,10] We therefore aimed to assess the value of deep learning (DL) for improving the diagnosis of breast nodules.

Materials and Methods

This study was approved by our Institutional Review Board, and informed consent was obtained from all patients.

Breast US Data Acquisition, Training, and Test Sets

From January 2015 to February 2018, we enrolled 1970 patients with 2071 nodules, including 1271 malignant nodules (2785 images), 1053 benign nodules (3040 images), and 2144 images of the contralateral normal breast. The data sets thus contained a total of 7969 images. Each benign and malignant nodule underwent surgical resection or core needle biopsy under US guidance to obtain a pathologic diagnosis. The malignant nodules included invasive ductal carcinomas, invasive lobular carcinomas, and breast adenocarcinomas.[5] The benign breast nodules included fibroadenomas and hamartomas.

All US examinations were performed by radiologists using commercial US scanners (Hi Vision Ascendus [Hitachi Medical Systems, Tokyo, Japan], EPIQ 5 and iU22 [Philips Healthcare, Bothell, WA], LOGIQ E9 [GE Healthcare, Chicago, IL], and Resona 7 [Mindray, Shenzhen, China]) equipped with 5–12-MHz linear array transducers. Transverse and sagittal images of breast tissue and nodules were acquired by a standard protocol,[11] and each breast nodule was measured according to the Breast Imaging Reporting and Data System of the American College of Radiology. Doppler US techniques were not performed. An experienced radiologist annotated the locations of benign and malignant nodules on the BUS images and added the information to the training data (Figure 1).

The first test image set included a total of 344 images of 130 breast nodules from centers 1 and 2: 172 images of 68 benign nodules and 172 images of 62 malignant nodules. The second test image set included a total of 348 images of 126 breast nodules from centers 1 and 3: 137 images of 96 benign nodules, 37 images of 30 malignant nodules, and 174 images of normal breasts (Figure 2).

Deep Learning Model Building

We used an open-source convolutional neural network (CNN; RetinaNet, www.retina-net.uni-koeln.de) to combine the speed of 1-stage detectors with high accuracy. This CNN can detect and classify lesions by setting an anchor and bounding box without separating the foreground from the background. It has good feature extraction abilities under natural conditions.[12] With a feature pyramid network, features extracted from RetinaNet were used to produce a multiscale convolutional pyramid.

We updated ResNet50 to enhance its detection sensitivity, including an input size of 256×256. With feature maps of 3 sizes (64×64, 32×32, and 16×16), a feature pyramid network was constructed to assist in predicting small objects in shallow layers, thereby facilitating improved detection and diagnosis with the DL model. Bounding boxes for object detection were generated with only 3 feature map sizes (64×64, 32×32, and 16×16; Figure 2).

Data augmentation operations were performed on the images, including random horizontal/vertical flips, random scaling, and random 90° rotations. Additionally, to adapt to the needs of the feature pyramid network structure, the backbone network reserved an output interface in all 3 key feature-mapping layers.

As many bounding boxes are generated on medical BUS images, there may be a foreground-background class imbalance, which could introduce difficulties during training. The focal loss function introduced by RetinaNet can alleviate this problem to some extent. It was implemented by the following equation[12]:

$$\text{Focal loss} = -\alpha \left(1 - p_j\right)^{\gamma} log\left(p_j\right). \qquad (1)$$

The focal loss is visualized for several values of $\gamma \in [0, 5]$, and when $\gamma = 0$, the focal loss is equivalent

to the cross-entropy loss. As nodule objects are relatively sparse in BUS images, the data imbalance is more serious, and following Equation 1,[12] the γ value was set to 3 instead of 2 to increase compensation.

Data Processing and Annotation

We randomly selected 7969 BUS images of malignant nodules, benign nodules, and normal breast tissue (1:1:1) and marked the locations of the nodules on each BUS image, including (1) the position of the upper-left corner of the nodule on the x-axis; (2) the position of the upper left-corner of the nodule on the y-axis; (3) the maximum transverse diameter of the nodule; and (4) the maximum length offset of the nodule.

Evaluation Indices Based on DL

The evaluation indices used for target detection were the accuracy, precision, recall, harmonic mean of recall and precision ($F1_{measure}$), and mean average precision (mAP). The accuracy, precision, recall, and $F1_{measure}$ were used to quantitatively assess image segmentation performance. In the test set, accuracy was the ratio of correct diagnoses in the sample to the total sample size; precision was the ratio of correctly detected samples to actually detected samples; recall was the ratio of the

Figure 1. Locations of a malignant nodule (**A1** and **A2**) and a benign nodule (**B1** and **B2**) were annotated on transverse and sagittal BUS images included in the training data.
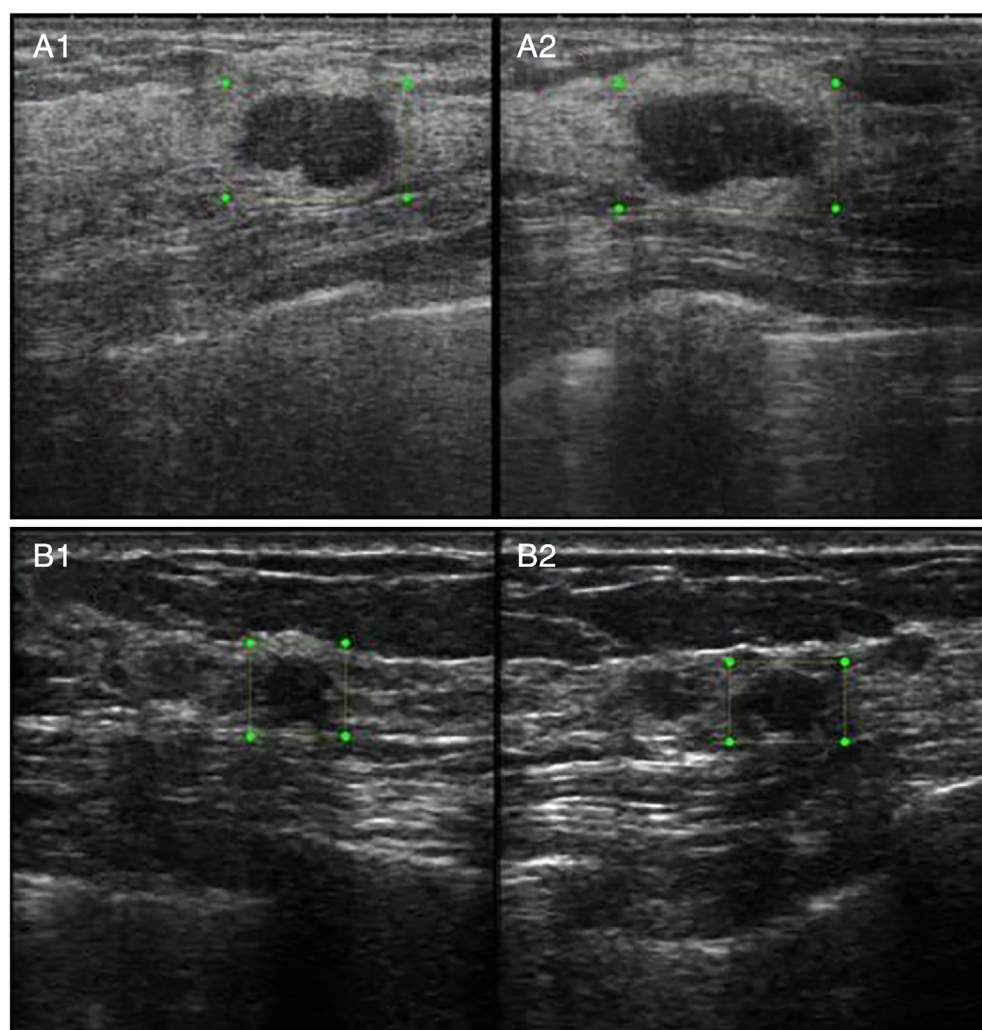
number of correctly detected samples to the total number of samples; AP was the area under the curve (AUC); and mAP was the average of the multiple categories of AP, which reflected the DL model's performance at distinguishing between the various categories:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}; \quad (2)$$

$$\text{Precision} = \frac{TP}{TP + FP}; \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN}; \quad (4)$$

$$\text{F1measure} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}; \quad (5)$$

$$AP = \frac{1}{11} \sum \text{Pinterp}^{(\gamma)} \gamma \in \{0, 0.1, \cdots, 1\}. \quad (6)$$

where FN was false-negative; FP, false-positive; TN, true-negative; and TP, true-positive.

Evaluations of Experts and Physicians
The evaluations of the experts, chief physicians (CPs), associate chief physicians (ACPs), attending physicians (APs), and resident physicians (RPs) were placed into competitions with those of the AI model. The verified physicians came from different areas of China. The 343 radiologists who enrolled in the competition were divided into different levels according to their experience, including (from the most to least experienced) experts (>25 years of experience), CPs (15–25 years of experience), ACPs (10–15 years of experience), APs (5–10 years of experience), and RPs (<5 years of experience). The 6 experts in this study were divided into 2 expert groups (3 in each group). We randomly selected 50 benign and 50 malignant breast nodules from center 1 to compare and evaluate differences in diagnostic performance between the AI system and radiologists. Correct and incorrect answers were counted as 1 and 0 points, respectively.

Statistical Analysis
Sklearn and SPSS version 19 software packages (IBM Corporation, Armonk, NY) were used for the statistical analysis. Matplotlib (matplotlib.org) was used to draw the receiver operating characteristic curves. Benign and malignant nodules were evaluated by receiver operating characteristic curves to calculate the accuracy, sensitivity, specificity, and AUC. A κ analysis was used to compare inter-rater reliability between AI and the 2 expert groups. Quantitative data were expressed as mean ± standard deviation and range. The groups were compared by the Mann–Whitney U test. Qualitative data from the groups were compared by the χ^2 test or Fisher exact test.

Results

General Data in Training and Test Sets
The training data set consisted of BUS images from 1963 women (age range, 18–91 years; mean, 49.12 ± 13.70 years) and 7 men (age range, 55–85 years; mean,

Figure 2. Schema for breast nodule detection. **A**, Image preprocessing. **B**, Backbone network. **C**, Feature pyramid for nodule detection. To enhance the detection sensitivity of the ResNet50 network, we used the backbone network of USNet, which uses 3 sizes of feature maps (64 × 64, 32 × 32, and 16 × 16). In these feature maps, a feature pyramid network was constructed to assist with predicting small objects in shallow layers, which allowed excellent detection and diagnosis using the DL model.
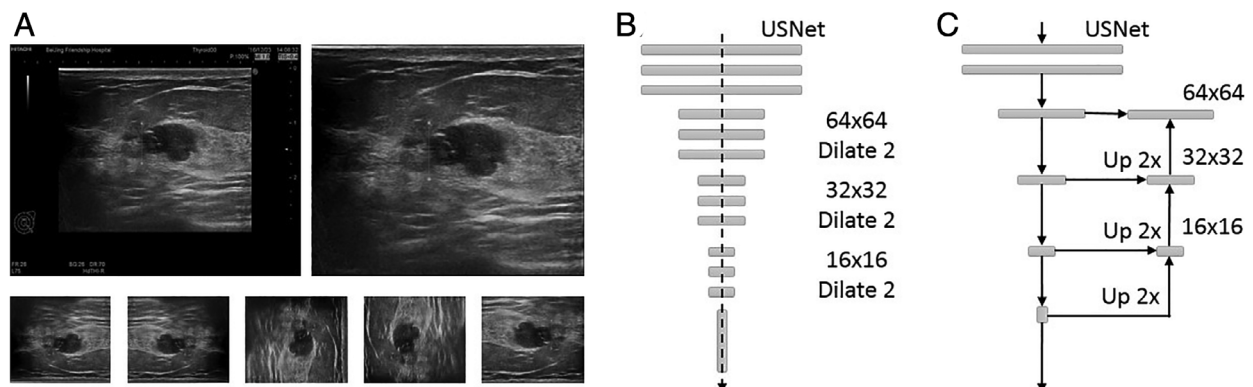
69.43 ± 10.99 years) who had nodules with maximum diameters of 0.37 to 7.40 cm (mean, 1.85 ± 0.85 cm). The 344 BUS images in the first test set were acquired

from 130 women (age range, 31–76 years; mean, 51.16 ± 11.18 years) between January 2015 and February 2018. The diameters of the lesions were 0.77

Figure 3. Receiver operating characteristic curves for the DL model's predictions of benign and malignant breast nodules and normal breast tissue.
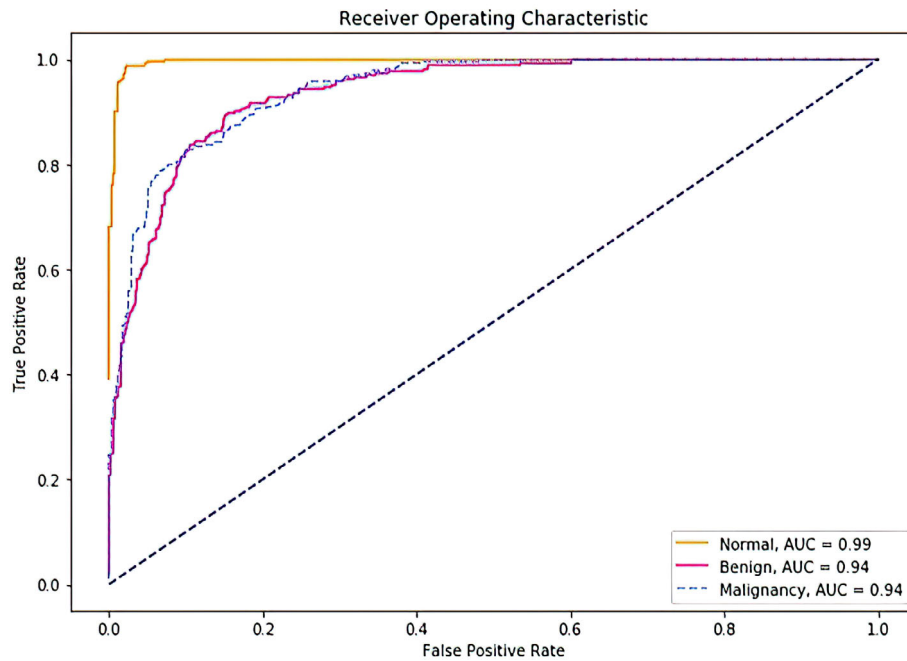

Table 1. Comparison of Diagnostic Scores for AI, Experts, and Physicians With Different Levels of Experience

Examiners	Radiologists	Age, y	Female/Male	Experience, y	Score	P
RPs	69	30.7 ± 0.43	42/27	5.1 ± 0.41	65.9 ± 1.62	
APs	133	36.2 ± 0.42	78/55	10.5 ± 0.41	68.8 ± 1.26	
ACPs	80	43.9 ± 0.52	42/38	18.2 ± 0.57	70.5 ± 1.41	
CPs	61	48.3 ± 0.66	41/20	23.0 ± 0.83	69.7 ± 1.66	.051
Experts[a]	6	54.6 ± 1.41	3/3	28.9 ± 1.10	86.5 ± 0.50	.010
AI	NA	NA	NA	NA	91	.030
AI vs experts						.951

NA indicates not applicable.
[a]The 6 experts were assigned to 2 groups of 3 experts each.

Table 2. Diagnostic Performance of AI, Expert Group 1, and Expert Group 2

Group	Sensitivity, % (n)	Specificity, % (n)	PPV, % (n)	NPV, % (n)	Accuracy, % (n)	κ
AI	92 (46/50)	90 (45/50)	90.2 (46/51)	91.8 (45/49)	91 (91/100)	0.82
Expert group 1	88 (44/50)	84 (42/50)	84.6 (44/52)	87.5 (42/48)	86 (86/100)	0.72
Expert group 2	88 (44/50)	86 (43/50)	86.3 (44/51)	87.8 (43/49)	87 (87/100)	0.74

NPV indicates negative predictive value; and PPV, positive predictive value.

to 4.82 cm (mean, 2.25 ± 1.07 cm). The 348 BUS images in the second test set were acquired from 126 women (age range, 31–73 years; mean, 50.2 ± 15.47 years) between January 2015 and February 2018. The diameters of the lesions were 0.29 to 1.00 cm (mean, 0.76 ± 0.15 cm).

Detection and Diagnosis of Breast Nodules and Normal Breast Tissue

The DL model for detection and diagnosis of benign breast nodules showed the following values: accuracy, 86.82%; precision, 77.89%; recall, 84.41%; $F1_{measure}$, 0.8012; mAP, 0.76; sensitivity, 84.41%; specificity, 88.02%; and AUC, 0.9367. For malignant breast nodules, these values were as follows: accuracy, 87.73%; precision, 87.73%; recall, 73.34%; $F1_{measure}$, 0.7992; mAP, 0.75; sensitivity, 73.34%; specificity, 94.87%; and AUC, 0.9434. For normal breast tissue, these values were as follows: accuracy, 96.58%; precision, 91.55%; recall, 98.86%; $F1_{measure}$, 0.9506; sensitivity, 98.86%; specificity, 95.43%; and AUC, 0.9968 (Figure 3).

Diagnostic Performance of AI Versus Radiologists

The performance results did not show any significant differences between the female and male radiologists in any experience-level group ($P = .905$). The scores for the radiologists with different levels of experience were as follows: experts, 86.50 ± 0.50 (confidence interval [CI], 87–86); CPs, 69.67 ± 1.66 (CI, 89–37); ACPs, 70.49 ± 1.41 (CI, 85–38); APs, 68.8 ± 1.26 (CI, 88–33); and RPs, 65.88 ± 1.62 (CI, 89–33). However, comparisons of the average and individual scores revealed that the AI system had the highest overall score (ie, 91). There were no significant differences between the 4 physician groups (CPs, ACPs, APs, and RPs; $P = .051$), although there were statistically significant differences in the comparisons of all physicians versus experts and all physicians versus AI ($P = .01$ and $.03$, respectively). The comparisons of scores and diagnostic performance between AI and the experts at distinguishing between 46 malignant and 45 benign nodules are shown in Tables 1 and 2, respectively.

Discussion

The diagnostic efficiency of US is operator dependent.[6,13] Artificial intelligence is expected to improve the diagnostic performance of BUS and decrease interobserver variability. The performance of various AI models for diagnosing breast nodules has been addressed in several studies. Nindrea et al[14] showed that a support vector machine algorithm could calculate breast cancer risk with better accuracy than other machine learning algorithms, including the decision tree, naïve Bayes, and k-nearest neighbor models. Machine learning learns from experience, and its performance improves as it learns. The work flow of machine learning teaches the system how to recognize important features. However, the machine learning training process is extremely time-consuming because it requires predictive features and has difficulty dealing with multidimensional data.[15] Xu et al[16] reported that the AUC, sensitivity, specificity, negative predictive value, and positive predictive value of an artificial neural network for judging breast cancer risk were 0.96, 0.82, 0.90, 0.90, and 0.80, respectively. Artificial neural networks can process data input similarly to the human brain. With developments in computational power, artificial neural networks have become increasingly complex, with as many as 100 layers of depth; hence, the term DL.[15] Xu et al[16] reported that DL CNNs could distinguish between functional tissue types in BUS images without manual intervention. That study's quantitative metrics for the evaluation of segmentation results all had values of greater than 80%.

Deep learning can automatically extract the essential and effective features from data, and CNNs can be optimized and trained with large amounts of training data.[17,18] Currently, there are 2 main target detection frameworks. One is a region-based CNN, which is based on regional nomination and has high recognition accuracy and location nomination ability. The major disadvantage of that method is its low computing efficiency and large resource occupancy. The other main framework is a series of you-only-look-once and single-shot multibox detectors, which have a fast recognition speed but no regional nomination ability. One disadvantage of the latter framework is that its recognition accuracy can be reduced by an imbalance between positive and negative samples.[19] Hence, we chose RetinaNet for this study. In a traditional ResNet50 network, the drawback of using a large stride is that small objects may be missed. Small objects may be easily missed after a decrease in the spatial resolution of the

feature map or during integration of large-context information. Extraction of image features in medical environments is more complicated than that from graphic images in natural situations. Thus, detailed resolution requires more attention. To overcome this shortcoming, we updated ResNet50 to enhance the network's detection sensitivity.

Artificial intelligence algorithms for work flow improvement and outcome analysis are being improved with the use of large quantities of high-quality imaging data.[20,21] In this study, high sensitivity and specificity for distinguishing benign from malignant breast nodules, as well as distinctions from normal tissue, were achieved by standardized data acquisition and training. The mAP values were not obtained from BUS images of normal tissues because no nodule frames could be labeled in such images. Our results were superior to those found with other algorithms such as computer-aided diagnosis and machine learning.[22]

Many studies have confirmed that the final assessments and diagnostic accuracy of BUS can vary considerably according to the radiologists' experience.[23,24] Our study verified this finding, with the expert group scoring 86.50 ± 0.50 points, which was higher than the scores of the CPs (69.67 ± 1.66 points), ACPs (70.49 ± 1.41 points), APs (68.8 ± 1.26 points), and RPs (65.88 ± 1.62 points). Overall, experience is the main factor for improving diagnostic performance.[23] For the final comparison, the AI system's κ value was 0.82, showing excellent diagnostic ability (sensitivity, 92%; specificity, 90%; positive predictive value, 90.2%; negative predictive value, 91.8%; and accuracy, 91%). There were no significant differences between AI and the experienced experts (P = .951). Thus, this DL model could assist radiologists in making decisions and diagnoses with the application of further data.

Our study had some limitations. The training and test data sets were all collected from only 3 centers; thus, the algorithm may have limitations in terms of generalizability based on the shortcomings of the data. Deep learning generally requires substantial quantities of training data, and training often results in overfitting. Therefore, obtaining sufficient amounts of high-quality data is a great difficulty for training DL applications in the medical field. Our study used 7969 BUS images for model training. To a certain extent, the model's degree of effectiveness only reflects the detection and classification performance of DL itself. By comparison, a physician's comprehensive judgment is derived from more factors, including the physician's experience and the patient's medical history. Thus, DL mathematically parameterizes only information obtained from the images. Therefore, it is necessary to develop an AI algorithm that can incorporate other relevant nonimage information. Further study is necessary before AI can be introduced into the clinical setting.

In conclusion, DL was applied to perform lesion recognition and aid in BUS image analysis. The DL system could distinguish between benign and malignant breast nodules. Deep learning achieved diagnostic accuracy equivalent to that of experts on BUS images.

References

1. Siegel RL, Jemal A. Cancer statistics, 2017. *CA Cancer J Clin* 2017; 67:7–30.

2. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA Cancer J Clin* 2015; 65:87–108.

3. Lee CH. Screening mammography: proven benefit, continued controversy. *Radiol Clin North Am* 2002; 40:395–407.

4. Anderson BO, Shyyan R, Eniu A, et al. Breast cancer in limited-resource countries: an overview of the Breast Health Global Initiative 2005 guidelines. *Breast J* 2006; 12(suppl 1):S3–S15.

5. 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.

6. Giess CS, Frost EP, Birdwell RL. Difficulties and errors in diagnosis of breast neoplasms. *Semin Ultrasound CT MR* 2012; 33:288–299.

7. Ikedo Y, Morita T, Fukuoka D, et al. Automated analysis of breast parenchymal patterns in whole breast ultrasound images: preliminary experience. *Int J Comput Assist Radiol Surg* 2009; 4:299–306.

8. Huang YL, Chen DR, Jiang YR, Kuo SJ, Wu HK, Moon WK. Computer-aided diagnosis using morphological features for classifying breast lesions on ultrasound. *Ultrasound Obstet Gynecol* 2008; 32:565–572.

9. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 2018; 286:800–809.

10. Kahn CE Jr. From images to actions: opportunities for artificial intelligence in radiology. *Radiology* 2017; 285:719–720.

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

12. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell* 2020; 42: 318–327.

13. Park JM, Yang L, Laroia A, Franken EA Jr, Fajardo LL. Missed and/or misinterpreted lesions in breast ultrasound: reasons and solutions. *Can Assoc Radiol J* 2011; 62:41–49.

14. Nindrea RD, Aryandono T, Lazuardi L, Dwiprahasto I. Diagnostic accuracy of different machine learning algorithms for breast cancer risk calculation: a meta-analysis. *Asian Pac J Cancer Prev* 2018; 19: 1747–1752.

15. Bini SA. Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care? *J Arthroplasty* 2018; 33:2358–2361.

16. Xu Y, Wang Y, Yuan J, Cheng Q, Wang X, Carson PL. Medical breast ultrasound image segmentation by machine learning. *Ultrasonics* 2019; 91:1–9.

17. Wang W, Qin J, Chui YP, Heng PA. A multiresolution framework for ultrasound image segmentation by combinative active contours. *IEEE Int Conf Proc* 2013; 2013:1144–1147.

18. Del Fiol G, Michelson M, Iorio A, Cotoi C, Haynes RB. A deep learning method to automatically identify reports of scientifically rigorous clinical research from the biomedical literature: comparative analytic study. *J Med Internet Res* 2018; 20:e10281.

19. Girshick R, Donahue J, Darrell T, Malik J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans Pattern Anal Mach Intell* 2016; 38:142–158.

20. Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism* 2017; 69S:S36–S40.

21. Mendelson EB. Artificial intelligence in breast imaging: potentials and limitations. *AJR Am J Roentgenol* 2019; 212:293–299.

22. Nichols JA, Herbert Chan HW, Baker MAB. Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophys Rev* 2019; 11:111–118.

23. Feng T, Sun X, Niu W, Wu H, Jiang C. Evaluation of an intervention to improve skills in diagnostic radiology of rural physicians over one year in four rural hospitals. *PLoS One* 2014; 9: e93889.

24. Kim HG, Kwak JY, Kim EK, Choi SH, Moon HJ. Man to man training: can it help improve the diagnostic performances and interobserver variabilities of thyroid ultrasonography in residents? *Eur J Radiol* 2012; 81:e352–e356.