ORIGINAL ARTICLE

Distinction between benign and malignant breast masses at breast ultrasound using deep learning method with convolutional neural network

Tomoyuki Fujioka[1] · Kazunori Kubota[1] · Mio Mori[1] · Yuka Kikuchi[1] · Leona Katsuta[1] · Mai Kasahara[2] · Goshi Oda[2] · Toshiyuki Ishiba[2] · Tsuyoshi Nakagawa[2] · Ukihide Tateishi[1]

Abstract

Purpose We aimed to use deep learning with convolutional neural network (CNN) to discriminate between benign and malignant breast mass images from ultrasound.

Materials and Methods We retrospectively gathered 480 images of 96 benign masses and 467 images of 144 malignant masses for training data. Deep learning model was constructed using CNN architecture GoogLeNet and analyzed test data: 48 benign masses, 72 malignant masses. Three radiologists interpreted these test data. Sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (AUC) were calculated.

Results The CNN model and radiologists had a sensitivity of 0.958 and 0.583–0.917, specificity of 0.925 and 0.604–0.771, and accuracy of 0.925 and 0.658–0.792, respectively. The CNN model had equal or better diagnostic performance compared to radiologists (AUC = 0.913 and 0.728–0.845, p = 0.01–0.14).

Conclusion Deep learning with CNN shows high diagnostic performance to discriminate between benign and malignant breast masses on ultrasound.

Purpose

Breast cancer is the most commonly diagnosed cancer and is the second leading cause of cancer-related death in women [1].

When a solid or cystic mass abnormality is detected on mammography or by clinical examination, ultrasound is used as a complementary tool [2]. Ultrasound also helps to guide the biopsy needle to the target area such that the tissue can be sampled and examined [3].

Breast ultrasound has been proven to be useful for differentiating between benign and malignant masses. The Breast Imaging Reporting and Data System (BI-RADS) lexicon [4] was developed by the American College of Radiology to standardize the terminology for describing and classifying breast lesions; good diagnostic performance has been reported [5]. Nevertheless, the morphologic features of benign and malignant lesions overlap substantially. Furthermore, the diagnosis of images is greatly dependent on the experience of the radiologist and, thus, significant interobserver variability may occur [6].

Recently, a deep learning method using a convolution neural network (CNN) has attracted attention for its image pattern recognition and artificial intelligence strategy. Neural networks are one method used for deep machine learning, inspired by the structure and function of the brain. It mimics the mammalian visual cortex when processing data using an artificial neural network containing hidden layers. A convolution layer in which images are processed with several types of filters is effective for pattern recognition of images [7, 8].

Although conventional machine learning algorithms require features from images that are extracted prior to learning, deep learning learns to extract meaningful features

✉ Mio Mori
  m_mori_116@yahoo.co.jp

1 Department of Radiology, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan

2 Department of Surgery, Breast Surgery, Tokyo Medical and Dental University, Tokyo, Japan

from images and compute inferences and decisions autonomously. Therefore, a deep learning method with CNN uses all the information contained within the image. This learning method may have the potential to diagnose images without relying on the experience of a radiologist.

A deep learning method with CNN has been reported previously to achieve good performance in image pattern recognition. It has recently been applied to radiologic images to detect tuberculosis on chest X-rays, detect and diagnose breast tumors with mammography, and distinguish liver masses on dynamic contrast agent-enhanced CT images [9–11].

In this study, we aimed to investigate the diagnostic performance of a deep learning method with CNN to discriminate between benign and malignant breast masses on ultrasound examination.

Materials and methods

Patients

The medical ethics committee of our hospital approved this retrospective study and waived the requirement for written informed consent.

In the present study, the inclusion criteria for enrolling patients were: (a) those who had breast masses and underwent breast ultrasound examination at our hospital from January 2010 to December 2017 and (b) those whose masses were diagnosed as benign or malignant by pathology or more than a 2-year follow-up. The following patients were excluded: (a) those who were treated with hormone therapy or chemotherapy and (b) those who were younger than 20 years of age. After reviewing the database of radiology reports and clinical records at our institute, a breast

radiologist (T.F. with 8 years of imaging experience) and a medical student (M.K) randomly selected breast masses of patients for this study.

Ultrasound imaging

Ultrasound examinations were performed by one of the five radiologists with 4- to 20-year experiences in breast ultrasound using an EUB-7500 scanner with a EUP-L54MA 9.75-MHz linear probe (Hitachi Medical Systems, Tokyo, Japan), an Aplio XG scanner with a PLT-805AT 8.0-MHz linear probe (Toshiba Medical Systems, Tochigi, Japan), and an Aplio 500 scanner with a PLT-805AT 8.0-MHz linear probe (Toshiba Medical Systems, Tochigi, Japan). The radiologists obtained transverse and longitudinal static images and measured the maximum diameter of the masses.

Data set

All solid and cystic masses, including simple cysts, were evaluated in this study. The ultrasound DICOM images were converted into jpeg using viewing software TFS-01 (Toshiba Medical Systems, Tochigi, Japan) and then trimmed to include skin to chest wall of size 256×256 pixels using Microsoft Paint (Microsoft, Redmond, Washington, USA) for analysis.

Table 1 shows the number of the patients, masses and images, patient's age, and maximum diameter of the mass. For the training phase, we extracted the maximum 6 different cross-sectional images of one mass and the maximum 2 masses in one patient. For our deep learning training, we randomly distributed and used a total set of 947 images of 240 masses in 237 patients (480 images of 97 masses in 95 patients for benign masses, 467 images of 143 malignant masses in 142 patients). For the test phase, we extracted

Table 1 Characteristics of patients and masses

	Training data				Test data				All data
	Benign	Malignant	All	p value benign versus malignant	Benign	Malignant	All	p value benign versus malignant	Training versus test data
Patients (n)	93	142	235		48	72	120		
Masses (n)	96	144	240		48	72	120		
Images (n)	480	467	947		48	72	120		
Age									
Mean (y)	48.6 ± 11.8	59.3 ± 12.3	55.0 ± 13.1	p < 0.001	49.2 ± 12.8	62.3 ± 13.3	57.0 ± 14.5	p < 0.001	p = 0.241
Range (y)	21–84	27–84	12–84		25–78	35–92	25–92		
Maximum diameter									
Mean (mm)	13.9 ± 8.4	16.6 ± 7.1	15.5 ± 7.7	p < 0.001	12.8 ± 7.4	18.2 ± 9.2	16.1 ± 8.9	p < 0.001	p = 0.924
Range (mm)	4–50	5–41	4–50		5–39	5–41	5–41		

Comparison was performed with the Mann–Whitney U test

only one image of one mass in one patient. A set total of 120 images in patients (48 images of benign and 72 images of malignant masses) were used for test purposes. Table 2 shows the histopathological results of the masses.

Deep learning with CNN

The computer used in our deep learning contained a graphics processing unit GeForce GTX 1080 (NVIDIA, Calif, USA), a central processing unit Core i7-8700 (Intel, Calif, USA), and a commercially available deep learning application DEEPstation DK-1000 (UEI, Tokyo, Japan) with a graphical user interface-based deep learning tool CSLAIER (Sony Computer Science Laboratories, Tokyo, Japan) that works on the framework Tensorflow (https://www.tensorflow.org) and Chainer (http://chainer.org). The deep learning model was constructed using the framework Chainer Ver 1.23.0, programing language Python 2.7.6 (https://www.python.org), and CNN architectures GoogleNet Inception v2 [12, 13]. GoogleNet is a 22-layer deep net having 9 inception units after three convolutional layers and a fully connected layer before the final output. GoogleNet Inception v2 is an improvement over GoogLeNet and has an essential deep learning technique called Batch Normalization, which is used for normalizing the value distribution before going into the next layer. There are two classes of problems in this study: benign and malignant lesions. Since GoogLeNet had 1000 class outputs, we reduced the output to 2 class outputs and calculated the possibility for malignancy in the mass. An architecture without fine-tuning was used in supervised learning, with a batch size of 32 and 50 epochs, because of high accuracy and low loss of data (Fig. 1). After building the models, we examined the accuracy of the trained models in distinguishing benign from malignant masses using the test image sets.

Radiologist readout

In the present study, 3 breast radiologists (M.M. with 4-year, T.F. with 8-year and K.K. with 20-year experience in breast imaging) also evaluated ultrasound images of the test data. Aside from knowing that the patients had a breast mass, they were blind to other pathological and clinical findings of this study. The assessed BI-RADS categories were reported as 2, 3, 4a, 4b, 4c or 5 by the readers.

Statistical analysis

All statistical analyses in this study were performed using the EZR software package version 1.31 (Saitama Medical Center, Jichi Medical University, Saitama, Japan) [14] and the Visualizing Categorical Data (VCR) package version 1.4-4 with graphical user interface for R software package (version 3.5.1; R Development Core Team, Vienna, Austria).

The data are presented as mean ± standard deviation. We performed a Mann–Whitney U test to compare patients' characteristics (age and maximum diameter of the mass). Interobserver agreement was assessed using weighted k statistics and was interpreted as follows: < 0.20, poor, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 good, and 0.81–1.0 very good agreement [15].

The test data set was used to calculate the sensitivity, specificity, and accuracy for differentiating benign masses from malignant masses. Receiver operating characteristic (ROC) analyses were performed to calculate the area under the receiver operating characteristic curve (AUC) for performance of the CNN models and radiologists readers in distinguishing benign masses from malignant masses. We considered p value of < 0.05 as statistically significant.

Table 2 Histopathology of masses

Training data		Test data	
Benign ($n = 96$)	Malignant ($n = 144$)	Benign ($n = 48$)	Malignant ($n = 72$)
Fibroadenoma 30	Ductal carcinoma in situ 14	Fibroadenoma 17	Ductal carcinoma in situ 3
Intraductal papilloma 9	Invasive ductal cancer 111	Intraductal papilloma 8	Invasive ductal cancer 57
Mastopathy 14	Mucinous carcinoma 5	Mastopathy 5	Mucinous carcinoma 3
Phyllodes tumor (benign) 2	Invasive lobular carcinoma 5	Adenosis 1	Invasive lobular carcinoma 4
Fibrous disease 1	Apocrine carcinoma 5	Pseudoangiomatous stromal hyperplasia 1	Apocrine carcinoma 2
Lactating adenoma 1	Medullary carcinoma 1	Radial scar/complex sclerosing lesion 1	Invasive micropapillary carcinoma 2
Abscess 1	Adenoid cystic carcinoma 1	No malignancy 2	Malignant lymphoma 1
No malignancy 3	Phyllodes tumor (malignant) 1		
Not known 35 (diagnosed by follow-up)	Adenomyoepithelioma with carcinoma 1	Not known 13 (diagnosed by follow-up)	

Fig. 1 Accuracy and loss of the training data
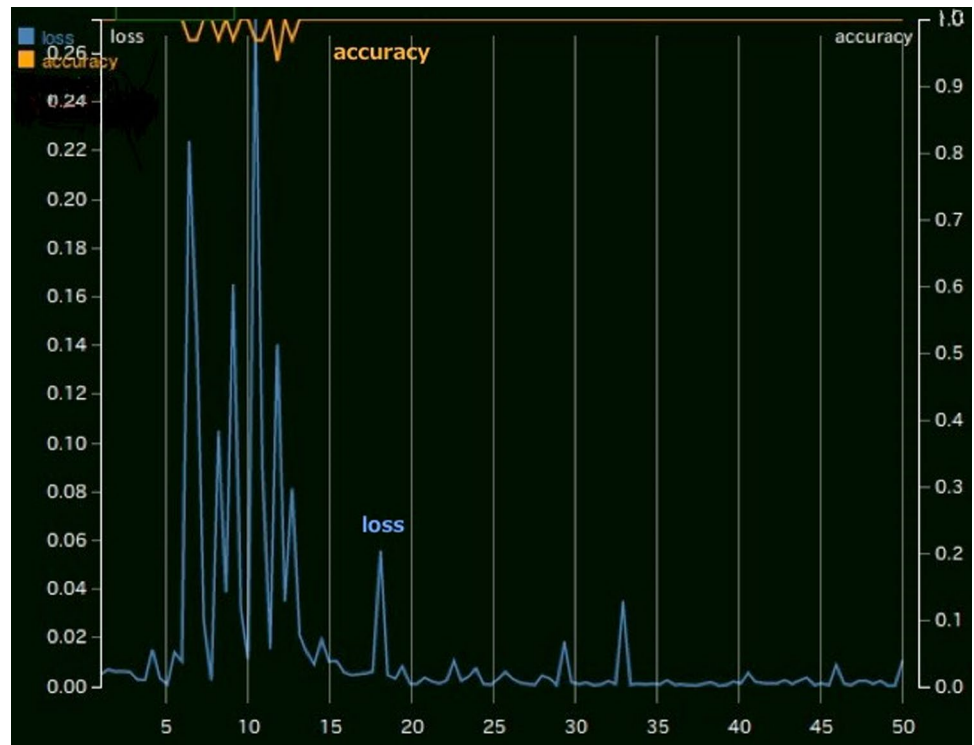
Table 3 BI-RADS category diagnosed by readers and CNN model

BI-RADS category	Possibility of malignancy (%)	Reader 1		Reader 2		Reader 3		CNN model	
		Benign (*n*)	Malignant (*n*)	Benign (*n*)	Malignant (*n*)	Benign (*n*)	Malignant (*n*)	Benign (*n*)	Malignant (*n*)
2	0	25	14	15	1	3	0	21	1
3	–2	12	16	16	9	26	6	4	1
4a	2–10	6	15	12	20	9	11	6	0
4b	10–50	0	0	3	13	7	19	8	1
4c	50–95	3	10	2	19	1	16	5	23
5	95–	2	17	0	10	2	20	4	46

BI-RADS breast imaging reporting and data system, *CNN* convolutional neural network

Results

Malignant masses are larger than benign masses and patients with malignant masses are significantly older than those with benign masses. There was no significant difference between training data and test data regarding patient's age and tumor size (Table 1).

Table 3 summarizes the BI-RADS categories of the CNN model and three radiologists. Reader 1, reader 2, reader 3, and the CNN model required total of 29.0 min (14.5 s per case), 34.0 min (17.0 s per case), 19.0 min (9.5 s per case), and within 2.0 min (within 1.0 s per case) for reading the images. So, the reading time by the CNN is shorter than by each reader.

Table 4 Interobserver agreement

	Reader 1	Reader 2	Reader 3	CNN model
Reader 1	1	0.3735	0.3147	0.2157
Reader 2	0.3735	1	0.4695	0.2796
Reader 3	0.3147	0.4695	1	0.2738
CNN model	0.2157	0.2796	0.2738	1

Comparison was performed with the weighted *k* statistics

CNN convolutional neural network

The CNN model tended to classify category 2, 3, or 5 to radiologists (CNN models; 87.5%, reader 1; 80.2%, reader 2; 62.1% and reader 3; 61.7%). Table 4 summarizes the interobserver agreement of the CNN models and radiologists.

Table 5 Sensitivity, specificity, accuracy and AUC of readers and CNN model

	Sensitivity	Specificity	Accuracy	AUC	p value (vs CNN model)
Reader 1	0.583	0.771	0.658	0.728	0.01
Reader 2	0.861	0.646	0.775	0.841	0.08
Reader 3	0.917	0.604	0.792	0.845	0.14
CNN model	0.958	0.875	0.925	0.913	

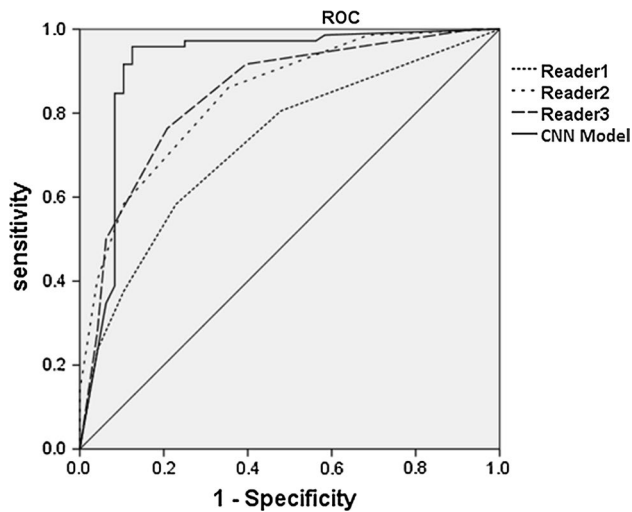*AUC* area under the receiver operating characteristic curve, *CNN* convolutional neural network

Fig. 2 Receiver operating characteristic curve (ROC) of readers and convolutional neural network (CNN) model

Interobserver agreement of the BI-RADS categorical assessment between the CNN model and radiologists (k = mean 0.2564, range 0.2157–0.2737) was lower than between only radiologists (k = mean 0.3859, range 0.3147–0.4695). Table 5 depicts diagnostic performance and Fig. 2 demonstrates the ROC of both the CNN model and radiologists' performance. The CNN model and the three radiologists (reader 1, reader 2 and reader 3) showed a sensitivity of 0.958, 0.583, 0.861, and 0.917, specificity of 0.875, 0.771, 0.644, and 0.604, and accuracy of 0.925, 0.658, 0.775, and 0.792 and AUC of 0.913, 0.728, 0.841, and 0.845, respectively. Cut-off values were 66.7% for the CNN model and 2% for radiologists. The CNN model showed good diagnostic performance. Reader 1, with 4-year experiences in breast imaging, showed a lower AUC of 0.728 than the CNN model (p = 0.01), and reader 2, with 8-year experience, and reader 3, with 20-year experiences in breast imaging, showed comparable AUCs of 0.841 and 0.845 to the CNN model, respectively (p = 0.08 and 0.14). Figure 3 summarizes the 3 cases of this study.

Discussion

Machine learning is a field of artificial intelligence where computers are not explicitly programmed, but can analyze relationships of existing data and perform tasks based on these new data [7, 8]. In our study, we demonstrated that breast masses could be differentiated using ultrasound using a deep learning method with a CNN model, using deep learning with multiple hidden layers. We found that the
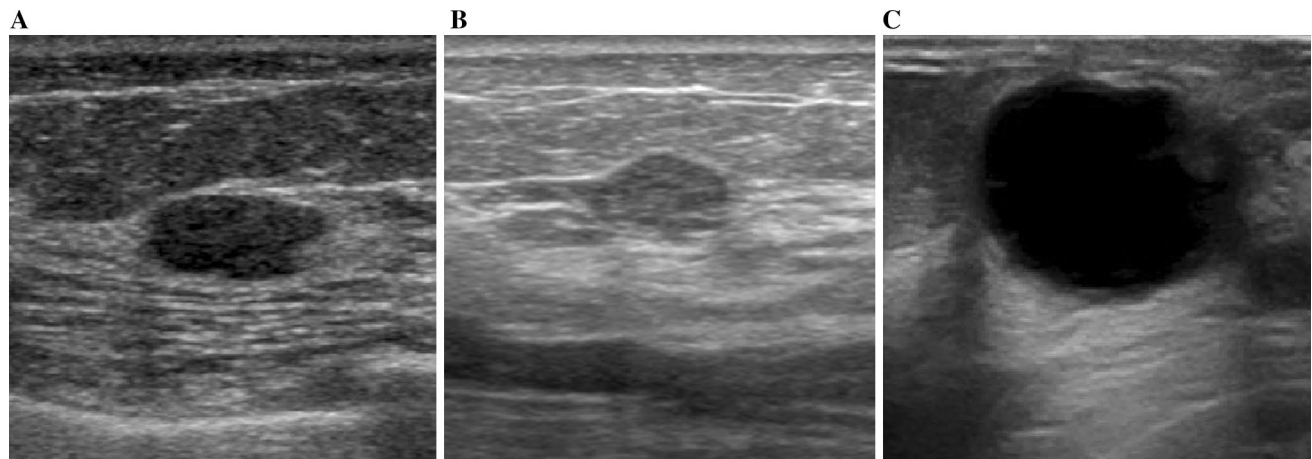
Fig. 3 Images of the 3 cases of this study. **a** Example of a true negative assessed by all readers and CNN model. This fibroadenoma was diagnosed as category 3 by reader 1, category 2 by reader 2, category 3 by reader 3, and category 1 by the CNN model. **b** Example of a false negative assessed by all readers and true positive evaluated by CNN model. This invasive ductal carcinoma was diagnosed as cat-

egory 2 by reader 1, category 3 by reader 2, category 3 by reader 3, and category 4c by the CNN model. **c** Example of a true negative assessed by readers and false positive evaluated by CNN model. This intraductal papilloma was diagnosed as category 2 by reader 1, category 2 by reader 2, category 3 by reader 3, and category 5 by the CNN model

CNN model was helpful in distinguishing between benign and malignant masses and showed good diagnostic performance. Moreover, the reading time by the CNN model is clearly shorter than by each reader.

Breast ultrasound has been proven to be useful in differentiating between benign and malignant masses and is widely used in daily practice [2, 3]. The BI-RADS lexicon standardizes how to describe terms and classify breast masses accordingly [5]. However, the morphologic features of benign and malignant masses sometimes overlap. Furthermore, diagnostic results depend greatly on the experience of the radiologist and, consequently, a large number of unnecessary biopsies and follow-ups are performed yearly [16].

There are some studies which report the usefulness of deep learning for diagnostic imaging of breast masses with ultrasound [17]. As far as we know, Han et al. employed the greatest number of samples to train a deep neural network with 4254 benign samples and 3154 malignant samples, and used GoogLeNet to distinguish the malignancy of breast masses on ultrasound, reporting an accuracy of 0.91 (91%), a sensitivity of 0.86, a specificity of 0.93, and an AUC over 0.9 [18]. It is said that fine-tuning has a better performance than learning from scratch in many cases [19], and the fine-tuning process was conducted by the ImageNet pretraining model in their study. On the other hand, our study, without the fine-tuning process, showed a high diagnostic performance consistent with this previous work. In their research, radiologists were required to define the region of interest for the breast masses; however, in our study, it was not required as the images were only trimmed to include the skin to chest wall. Therefore, our research is simpler and more reproducible than their research.

Shi et al. used a deep polynomial network to extract textural features from 100 malignant and 100 benign masses on ultrasound images. Although the number of cases in their study is smaller than ours, it showed the same diagnostic accuracy of 92.4% as our study [20]. Stoffel et al. showed that deep learning software is able to differentiate between phyllodes tumors and fibroadenomas with good diagnostic accuracy (AUC = 0.73) and high negative predictive value (NPV = 100%). Radiologists showed comparable accuracy (AUC 0.60–0.77) but at lower NPV (64–80%) [21].

In this research, we did not examine the ability of segmentation. However, a computer-aided tool for detection of breast masses using CNN has been developed and reported. Kumar et al. used Multi U-net algorithm which is based on CNN and segments the suspicious breast masses. It achieved a mean Dice coefficient of 0.82, a true-positive fraction of 0.84, and a false-positive fraction of 0.01 [22].

To the best of our knowledge, our study was the first to directly compare diagnostic performance between the CNN model and radiologists to discriminate between benign and malignant breast masses using ultrasound. Reader 1, with 4-year experiences in breast imaging had a lower AUC than the CNN, and reader 2, with 8-year experiences and reader 3 with 20-year experiences in breast imaging showed comparable AUC to the CNN. The difference in radiologist performance compared to CNN appears to be related to the number of years' experience of each radiologist, as greater experience has led to a more accurate diagnosis of masses in our study. Our results imply that this CNN model could be helpful to radiologists to diagnose breast masses, especially those with only a few years of clinical experience.

In this present study, the CNN model had a tendency to diagnose masses as BI-RADS category 2, 3 or 5 compared to the radiologist. Regarding interobserver agreement, there is a higher correlation between only radiologists than between the CNN model and radiologists. We must assume that the CNN models and radiologists find and evaluate completely different aspects of the images. Human thought levels are limited to several dimensions. Conversely, existing deep learning technology can refer "thoughts" up to hundreds of dimensions [23]. This is known as "black box problem", in which it is impossible for us to understand the process of how deep learning has reached any particular answer and the cause of false positive or negative. We require further research to understand how the CNN model detects and diagnoses these images, and the research may contribute to improvement of the diagnostic performance of radiologists.

This study has several limitations. First, this retrospective study was conducted at a single institution. Therefore, more extensive, multicenter studies are warranted to validate the findings of this study. Second, not all recurrent lesions were diagnosed using cytological or histological diagnosis. Third, we performed this study using images that were converted into 256 × 256 pixels. This image processing might result in a loss of information and, thus, influence the performance of the models. Fourth, because we used three ultrasound systems of two companies, there may have been a problem in the adaptability of learning outcome in testing using other ultrasound system images. So, further study is required to verify whether other ultrasound systems can also show a good performance.

Conclusion

Deep learning image analysis with CNN showed high diagnostic performance to distinguish between benign and malignant masses on breast ultrasound. This CNN model could help radiologists to diagnose the malignancy of breast masses.

Compliance with ethical standards

Conflict of interest All the authors and their institution have no conflicts of interest.

Ethical approval All procedures used in this research were approved by the Ethical Committee of Tokyo Medical and Dental University, Medical Hospital.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin. 2018;68:7–30.
2. Kornecki A. Current status of breast ultrasound. Can Assoc Radiol J. 2011;62:31–40.
3. Newell MS, Mahoney MC. Ultrasound-guided percutaneous breast biopsy. Tech Vasc Interv Radiol. 2014;17:23–31.
4. D'Orsi C, Sickles E, Mendelson E, Morris E. Breast imaging reporting and data system. 5th ed. Reston: American College of Radiology; 2013.
5. Rao AA, Feneis J, Lalonde C, Ojeda-Fournier H. A pictorial review of changes in the BI-RADS fifth edition. Radiographics. 2016;36:623–39.
6. Youk JH, Son EJ, Gweon HM, Kim H, Park YJ, Kim JA. Comparison of strain and shear wave elastography for the differentiation of benign from malignant breast lesions, combined with B-mode ultrasonography: qualitative and quantitative assessments. Ultrasound Med Biol. 2014;40:2336–44.
7. Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O. Deep learning with convolutional neural network in radiology. Jpn J Radiol. 2018;36:257–72.
8. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, et al. Deep learning: a primer for radiologists. Radiographics. 2017;37:2113–31.
9. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 2017;284:574–82.
10. Hamidinekoo A, Denton E, Rampun A, Honnor K, Zwiggelaar R. Deep learning in mammography and breast histology, an overview and future trends. Med Image Anal. 2018;47:45–67.
11. Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O. Deep learning with convolutional neural network in radiology. Jpn J Radiol. 2018;36:257–72.
12. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceeding of IEEE international conference on computer vision pattern recognition; 2015. p. 1–9.
13. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167 (2015).
14. Kanda Y. Investigation of the freely available easy-to-use software 'EZR' for medical statistics. Bone Marrow Transplant. 2013;48:452–8.
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.
16. Demircioğlu Ö, Uluer M, Arıbal E. How many of the biopsy decisions taken at inexperienced breast radiology units were correct? J Breast Health. 2017;13:23–6.
17. Han S, Kang HK, Jeong JY, Park MH, Kim W, Bang WC, et al. A deep learning framework for supporting the classification of breast lesions in ultrasound images. Phys Med Biol. 2017;62:7714–28.
18. Huang Q, Zhang F, Li X. Machine learning in ultrasound computer-aided diagnostic systems: a survey. Biomed Res Int. 2018;4(2018):5137904.
19. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging. 2016;35:1285–98.
20. Shi J, Zhou S, Liu X, Zhang Q, Lu M, Wang T. Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset. Neurocomputing. 2016;194:87–94.
21. Stoffel E, Becker AS, Wurnig MC, Marcon M, Ghafoor S, Berger N, et al. Distinction between phyllodes tumor and fibroadenoma in breast ultrasound using deep learning image analysis. Eur J Radiol Open. 2018;24:165–70.
22. Kumar V, Webb JM, Gregory A, Denis M, Meixner DD, Bayat M, et al. Automated and real-time segmentation of suspicious breast masses using convolutional neural network. PLoS ONE. 2018;16(13):e0195816.
23. M Claesen, B De Moor. Hyperparameter search in machine learning. arXiv:1502.02127 (2015).