


ORIGINAL ARTICLE

Establishment of a deep-learning system to diagnose BI-RADS4a or higher using breast ultrasound for clinical application

Tetsu Hayashida¹  | Erina Odani¹ | Masayuki Kikuchi¹ | Aiko Nagayama¹ | Tomoko Seki¹ | Maiko Takahashi¹ | Noriyuki Futatsugi² | Akiko Matsumoto³ | Takeshi Murata⁴ | Rurina Watanuki⁵ | Takamichi Yokoe⁵ | Ayako Nakashoji⁶ | Hinako Maeda⁷ | Tatsuya Onishi⁵ | Sota Asaga⁸ | Takashi Hojo⁹ | Hiromitsu Jinno³ | Keiichi Sotome⁷ | Akira Matsui⁶ | Akihiko Suto⁴ | Shigeru Imoto⁸ | Yuko Kitagawa¹

¹Department of Surgery, Keio University School of Medicine, Tokyo, Japan

²Fixstars Corporation, Tokyo, Japan

³Department of Surgery, Teikyo University School of Medicine, Tokyo, Japan

⁴Department of Breast Surgery, National Cancer Center Hospital, Tokyo, Japan

⁵Department of Breast Surgery, National Cancer Center Hospital East, Chiba, Japan

⁶Department of Breast Surgery, National Hospital Organization Tokyo Medical Center, Tokyo, Japan

⁷Department of Breast and Thyroid Surgery, Kitasato University Kitasato Institute Hospital, Tokyo, Japan

⁸Department of Breast Surgery, Kyorin University School of Medicine, Tokyo, Japan

⁹Department of Breast Oncology, Saitama Medical University International Medical Center, Saitama, Japan

Correspondence

Tetsu Hayashida, Department of Surgery, Keio University School of Medicine, 35 Shinanomachi, Shinjuku, Tokyo 160-8582, Japan.

Email: tetsu@keio.jp

Funding information

Fixstars Corporation; JSPS KAKENHI Grant-in-Aid for Scientific Research (C), Grant/Award Number: 20K08993

Abstract

Although the categorization of ultrasound using the Breast Imaging Reporting and Data System (BI-RADS) has become widespread worldwide, the problem of inter-observer variability remains. To maintain uniformity in diagnostic accuracy, we have developed a system in which artificial intelligence (AI) can distinguish whether a static image obtained using a breast ultrasound represents BI-RADS3 or lower or BI-RADS4a or higher to determine the medical management that should be performed on a patient whose breast ultrasound shows abnormalities. To establish and validate the AI system, a training dataset consisting of 4028 images containing 5014 lesions and a test dataset consisting of 3166 images containing 3656 lesions were collected and annotated. We selected a setting that maximized the area under the curve (AUC) and minimized the difference in sensitivity and specificity by adjusting the internal parameters of the AI system, achieving an AUC, sensitivity, and specificity of 0.95, 91.2%, and 90.7%, respectively. Furthermore, based on 30 images extracted from the test data, the diagnostic accuracy of 20 clinicians and the AI system was compared, and the AI system was found to be significantly superior to the clinicians (McNemar test, $p < 0.001$). Although deep-learning methods to categorize benign and malignant tumors using breast ultrasound have been extensively reported, our work represents the first attempt to establish an AI system to classify BI-RADS3 or lower and BI-RADS4a or higher successfully, providing important implications for clinical actions. These results suggest that the AI diagnostic system is sufficient to proceed to the next stage of clinical application.

KEYWORDS

AI diagnosis, artificial intelligence, BI-RADS, breast ultrasound, deep learning

Abbreviations: AI, artificial intelligence; AUC, area under the curve; BI-RADS, Breast Imaging Reporting and Data System; CNN, convolutional neural network; IoU, intersection over union; ROC curve, receiver-operating characteristic curve.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Cancer Science* published by John Wiley & Sons Australia, Ltd on behalf of Japanese Cancer Association.

1 | INTRODUCTION

Breast ultrasound has progressed enormously over the last decade with markedly improved resolution and rapid image processing.¹ It is classified as a physiological function test in which a patient's body is directly examined by an operator. Thus, its accuracy depends on the quality of the equipment and environment, as well as on the observer's technique, experience, knowledge of disease, and how the findings are obtained. Therefore, the reliability of handheld ultrasound remains controversial owing to operator dependence.² In 2003, the American College of Radiology developed the Breast Imaging Reporting and Data System (BI-RADS) classification to standardize the terminology for reporting breast ultrasound findings and to ensure a consistent diagnosis.³ The latest version, the 5th edition, was revised in 2013 to provide comprehensive guidelines for breast imaging diagnosis by integrating mammography, ultrasonography, and MRI.⁴ Although this categorization of ultrasound using BI-RADS has become widespread worldwide, the problem of inter-observer variability remains.^{5,6} Moreover, real-time scanning permits detailed lesion evaluation compared with analysis of static images on a workstation.⁷ Therefore, the expertise of the diagnostician is required to make an accurate diagnosis from the static images taken after the scan.

To maintain uniformity in the diagnostic accuracy of breast ultrasound images, we developed a system for applying artificial intelligence (AI) to breast ultrasound diagnosis using deep-learning technology, which has been advanced remarkably in recent years. The BI-RADS classification primarily examines ultrasound findings based on "shape," "orientation," "margin," "echo pattern," and "posterior features";⁴ however, the system is designed to enable AI to detect features of ultrasound images that cannot be distinguished by the human eye and to determine BI-RADS categorization. Several studies have shown that the negative predictive value of the BI-RADS rating system was >99%, and the presence of malignancy was substantially less likely if the patient was rated BI-RADS3 or lower.^{6,8,9} Therefore, the purpose of this study was to construct a system in which AI can distinguish whether a static image from a breast ultrasound is BI-RADS3 or lower or BI-RADS4a or higher and to verify its accuracy. Thus, the development and operation of the AI diagnosis system are expected to enable diagnosticians to judge the findings with a deep sense of confidence and to reduce incorrect judgments and missed malignancies in both diagnostic breast imaging and initial screening situations.

2 | MATERIALS AND METHODS

2.1 | Study design

This was a multicenter exploratory study aimed to establish an AI system for breast ultrasound diagnosis using a deep-learning technology and verify its accuracy. The AI diagnostic system determined whether the test image was BI-RADS3 or lower or BI-RADS4a or higher. These results were compared with the predetermined diagnoses made by

human experts, and the sensitivity, specificity, and the area under the curve (AUC) were calculated and used for evaluation.

2.2 | Collection of ultrasound images

Breast ultrasound images for evaluation were collected using opt-out recruitment methods from Keio University Hospital, Teikyo University Hospital, National Cancer Center Hospital, National Cancer Center Hospital East, Kitasato University Kitasato Institute Hospital, Kyorin University Hospital, Saitama Medical University International Medical Center, and Tokyo Medical Center. The study was conducted in accordance with the Declaration of Helsinki, and the study protocol was approved by the Institutional Review Board of Keio University School of Medicine (Approval No. 20170146). The local ethics committees approved the study in the participating facilities. Breast ultrasound images were collected either from women who had histologically determined benign or malignant breast tumors or from women who were clinically diagnosed with benign tumors after 6 months of follow-up or longer. One or more images were selected from the ultrasound images of a single case that met the criteria of a breast cancer specialist in each institution certified by the Japanese Breast Cancer Society. At the time of collection, the following information was assigned to each image: the institution where the ultrasound was performed, benign or malignant as judged histologically or clinically, histological type, and ultrasound machine manufacturer type. At each facility, the ultrasound images were anonymized by removing all personally identifiable information, and the encrypted image information was stored at Keio University. The quality of the breast ultrasound images collected from each facility varied, with the lowest pixel count at 208,318 pixels and the highest at 1,757,668 pixels. The median number of pixels in the images from each institution was as follows: Keio University Hospital, 691,200 pixels; Teikyo University Hospital, 1,228,800 pixels; National Cancer Center Hospital, 661,287 pixels; National Cancer Center Hospital East, 659,476 pixels; Kitasato University Kitasato Institute Hospital, 786,432 pixels; Kyorin University Hospital, 691,200 pixels; Saitama Medical University International Medical Center, 487,808 pixels; and Tokyo Medical Center, 229,542 pixels. The collected images were carefully examined for the initial defective state, and images containing Doppler or elastography, or those deemed technically inappropriate for evaluation, were eliminated.

2.3 | Image evaluation and annotation

Ultrasound images were evaluated separately by two independent evaluators certified by the Japan Central Organization on Quality Assurance of Breast Cancer Screening to conduct breast cancer ultrasound. The evaluator reviewed an image without any additional information, marked all lesions observed, and provided assessments based on the 5th edition of BI-RADS, which was

revised in 2013.⁴ If multiple lesions were present in a single image, each was marked and evaluated individually. After all images were evaluated, the results were disclosed, and if any lesions were assessed differently by the two evaluators, the final evaluation was determined by discussion. If no consensus on the evaluation was reached through discussion, one of the opinions was chosen by a third evaluator. In this manner, lesion-by-lesion assessment was collected and analyzed.

The annotation process was performed using the Labelme ver. 4.5.9 (Wada, K. Labelme: Image polygonal annotation with Python [Computer software] <https://doi.org/10.5281/zenodo.5711226>). Tumors, skin, muscle, adipose tissue, and mammary tissue on the image were surrounded by polygons, and each was tagged with the corresponding tissue. Tumors were tagged based on BI-RADS classification. If multiple tumors were present in a single image, a tag with the BI-RADS classification was assigned. All statistical calculations were performed using Python 3.6 (Python Software Foundation) with NumPy 1.18.1 and the scikit-learn libraries 0.22.1.

3 | RESULTS

3.1 | Establishment of the AI diagnosis system

The annotated images were randomly classified into two datasets: a training dataset of 4028 images, including 5014 lesions, and a test dataset of 3166 images, including 3656 lesions; the former dataset was used to train the AI system (Table 1). As images determined as BI-RADS1 have no lesions, the number of images was assessed instead of the number of lesions. The ultrasound systems used in the hospitals participating in this study were manufactured and distributed by GE Healthcare Systems, FUJIFILM Health Care (HITACHI Aloka), and Canon Medical Systems. The ratio of the number of lesions in the training dataset by the manufacturer was 849, 1720, and 2445, respectively (Table 1). The AI system for image identification based on convolutional neural network (CNN) in the deep-learning technology was provided by the IoT company Fixstars Corporation. The number of lesions for each BI-RADS classification in the training dataset was 437, 579, 2794, and 1204 lesions for BI-RADS2, BI-RADS3, BI-RADS4, and BI-RADS5, respectively (Table 2). Images diagnosed as BI-RADS1 were excluded from the training dataset. The percentage of malignancies in each category of the overall data was 0%, 5.57%, 32.7%, 88.9%, and 98.7% for BI-RADS3 and below, BI-RADS4a, BI-RADS4b, BI-RADS4c, and BI-RADS5, respectively (Table 2).

3.2 | Validation of the diagnostic accuracy by AI

The validation of the test dataset, which included 3656 lesions, was conducted using AI trained with the training dataset. The number of lesions for each BI-RADS classification in the test dataset was 1470, 176, 278, 1200, and 532 for BI-RADS1, BI-RADS2, BI-RADS3, BI-RADS4, and BI-RADS5, respectively (Table 2).

TABLE 1 Number of images and lesions in the training and test datasets obtained using ultrasound devices by different manufacturers

Manufacturers of the ultrasound devices	Training data			Test data			Total
	GE Healthcare Systems	FUJIFILM Health care (HITACHI Aloka)	Canon Medical Systems	GE Healthcare Systems	FUJIFILM Health care (HITACHI Aloka)	Canon Medical Systems	
No. of images	747	1104	2177	650	531	1985	7194
No. of lesions	849	1720	2445	695	871	2090	8670
			Total No. of training data			Total No. of test data	
			4028			3166	
			5014			3656	

TABLE 2 Properties of the ultrasound image dataset. The table shows the number of lesions determined to be in each Breast Imaging Reporting and Data System (BI-RADS) category and the number of benign or malignant tumors included in these categories

BI-RADS	Training data			Test data			Total No. of lesions	% of malignant tumor
	No. of lesions diagnosed as malignant	No. of lesions diagnosed as benign	No. of lesions in training data	No. of lesions diagnosed as malignant	No. of lesions diagnosed as benign	No. of lesions in test data		
1	0	0	0	0	1470	1470	1470	0
2	0	437	437	0	176	176	613	0
3	0	579	579	0	278	278	857	0
4a	44	701	745	16	317	333	1078	5.57
4b	291	653	944	148	251	399	1343	32.7
4c	978	127	1105	420	48	468	1573	88.9
5	1189	15	1204	524	8	532	1736	98.7

Lesion detection was performed by the one-stage detector for “BI-RADS 4a or higher” and “BI-RADS 3 or lower,” respectively. In cases of overlapping detection of a lesion with both BI-RADS 4a or higher and BI-RADS 3 or lower, priority was given to the higher confidence score. However, AUC representing multiple confidence scores could not be simply calculated. Therefore, the AUC by all possible thresholds of its confidence score for the detection in each image of BI-RADS 4a or higher, which is the most important indicator for diagnosis, was obtained, and its value was 0.95 (Figure 1A). The values with the smallest difference in sensitivity and specificity in this ROC curve were 91.2% and 90.7%, respectively. (Figure 1B).

The test data consisted of 695, 871, and 2090 lesions from GE Healthcare Systems, FUJIFILM Health care (HITACHI Aloka), and Canon Medical Systems equipment, respectively, and diagnostic validation of the AI system was performed for each ultrasound equipment manufacturer. The sensitivity and specificity determined for images from the GE Healthcare Systems were 80.8% and 94.0%, respectively. Similarly, the sensitivity and specificity were 90.9% and 75.5%, respectively, for FUJIFILM Health Care (HITACHI Aloka) and 92.2% and 88.7%, respectively, for Canon Medical Systems.

The intersection over union (IoU) was calculated as the percentage of overlap between human-annotated and AI-detected regions. The value of mIoU, which is the mean IoU obtained by AI analysis of all test data, was 0.6453. Considering that the IoU was $64/98 \approx 0.653$ when the length of one side of the square was shifted diagonally by one-ninth, the lesions were detected with a high degree of accuracy.

3.3 | Comparison of diagnostic performance between clinicians and AI

From the test dataset, excluding images with BI-RADS1 category, 30 images from GE Healthcare Systems were selected completely at random and diagnosed by AI as well as by 20 clinicians. Of the 30 images, 19 were classified as BI-RADS 4a or higher and 11 were classified as BI-RADS 3 or lower. All clinicians had 5-8 years of clinical experience, and 10 out of 20 were board-certified surgeons, certified by the Japan Surgical Society. The mean sensitivity and specificity of the diagnosis made by the clinicians were 67.1% (31.6%-84.2%) and 81.4% (47.4%-90.9%), respectively. The sensitivity and specificity of the diagnosis made by the clinician who provided the best responses were 84.2% and 90.9%, respectively. In contrast, AI had a sensitivity of 100% and a specificity of 90.9%, and a McNemar test demonstrated that AI was significantly superior to the clinicians in both sensitivity and specificity ($p < 0.001$, Figure 2).

4 | DISCUSSION

Although there have been many reports on the AI-based diagnosis of breast ultrasound, most of the studies have focused on technical aspects, such as the algorithm used for deep learning depending on

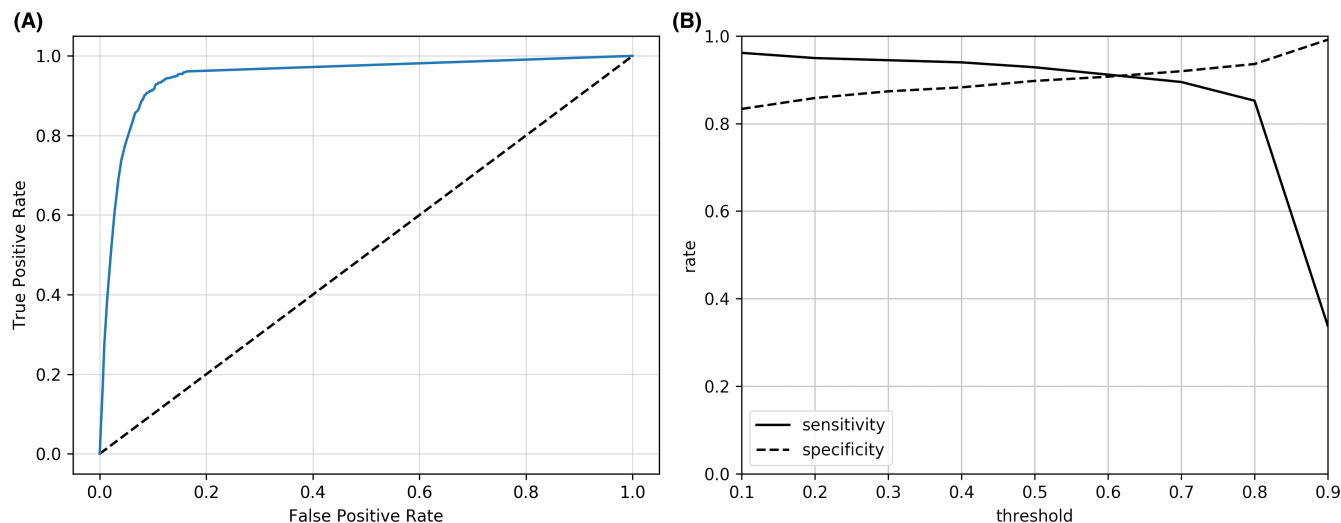


FIGURE 1 Receiver-operating characteristic (ROC) curve by possible thresholds of the confidence score for the detection in each image of Breast Imaging Reporting and Data System (BI-RADS) 4a or higher. A, ROC curve with an area under the curve (AUC) of 0.95. B, Sensitivity and specificity with variations in thresholds of the confidence score

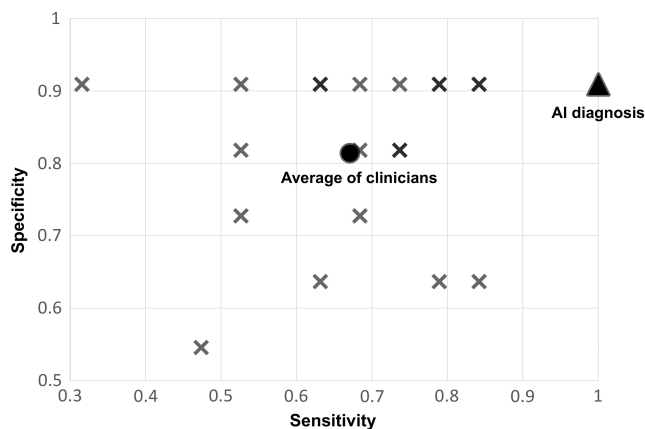


FIGURE 2 Sensitivity and specificity of diagnosis by artificial Intelligence (AI) and 20 clinicians for 30 images. X: diagnosis by each clinician, ▲: AI diagnosis, ●: average of clinicians

the purpose, while only a few studies have focused on clinical applications and utility of AI.¹⁰ Among these technical considerations, the applications of deep-learning techniques are mainly categorized into object detection,^{11,12} segmentation,^{13,14} image classification,^{15,16} and image synthesis.¹⁷ Regarding image classification, there are many reports on the distinction between benign and malignant lesions in static images.^{18–22} In contrast, although the identification of the tumor as benign or malignant is an important clinical element, the most critical issue is to determine what medical management should be performed on a patient whose breast ultrasound shows abnormalities. The BI-RADS system recommends action for every diagnosed category, which is consistent across imaging modalities.⁴ Therefore, we conducted this study to improve the accuracy of the BI-RADS assessment using the deep-learning technology. Huang et al. reported the results of the classification using a two-stage grading system for BI-RADS categories 3, 4a, 4b, 4c, and 5 in the

evaluation of 2238 cases. The results showed that their proposed scheme could extract effective features from breast ultrasound images for final classification with high accuracy.²³ If a patient is diagnosed with BI-RADS category 4a or higher, the clinical necessity for subclassification of the patient is low because tissue diagnosis is recommended in the BI-RADS system.⁴ Given this background, our AI system was developed to classify patients into two groups based on BI-RADS categories.

It is worth considering whether the threshold for classification into two groups by an AI system should be BI-RADS3 or lower or BI-RADS4a or lower. In BI-RADS, complicated cyst and solid mass are classified as category 4a or higher; therefore, identifying the echo pattern is especially important for this classification. The presence of malignancy was considerably less likely if the patient was rated BI-RADS3 or lower.^{6,8,9} In a combined assessment of multiple reports, the incidence of malignancy was only 0.2% at 6 months in lesions determined to be BI-RADS3.²⁴ BI-RADS4a lesions are defined as having malignancy rates of 2%-10%, and the actual training and test data used in this study showed a malignancy rate of 5.57%. Setting the threshold category should be determined based on the specific application of the AI system. As BI-RADS is intended to be evaluated with other modalities, such as mammograms, and recommended clinical action changes above BI-RADS4a, the threshold for this study was set at BI-RADS3 or lower to prioritize sensitivity at the expense of specificity for the diagnosis of benign and malignant tumors. In this setting, the problem of a lower rate of malignancy in BI-RADS4a arises, resulting in a higher frequency of unnecessary invasive testing; however, this is not a specific problem in AI diagnosis. In this study, clinicians, unlike AI, frequently underestimated the cases with BI-RADS 4a and misjudged them as BI-RADS3 or lower, causing lower sensitivity. To solve this problem, downgrading of BI-RADS 4a lesions with elastography and clinical nomograms has been attempted with excellent performance.^{25,26} Moreover, there

are many reports of AUCs of >0.9 for benign and malignant tumor diagnoses by AI-based image classification.¹⁷ The use of such AI for downgrading BI-RADS4a lesions should be considered in the future.

Breast ultrasound is rarely used as a primary diagnosis but is mostly used to differentiate between benign and malignant disease in most clinically symptomatic patients or as an adjunct tool to further analyze abnormalities on screening mammograms. However, Asians, such as Japanese and Chinese, are considered to have a high incidence of dense breasts,^{27,28} and there are several reports on the effectiveness of using breast ultrasound for screening tests.^{29,30} The results of a large randomized controlled trial (J-START), including women in their 40s, showed significantly better detection rates of early-stage breast cancer for screening with both mammography and ultrasound than for screening with mammography alone.³¹ Therefore, the development of a comprehensive evaluation system using both mammograms and ultrasound in breast cancer screening has been attempted.²⁹ However, the difference between the prevalence in the test data and the actual prevalence in screening tests is considered a significant bias. For example, in the J-START trial, 36,752 women were assigned to the intervention group that was subjected to both ultrasound and mammograms, but the recall rate was 12.6%, and the diagnosis rate of breast cancer was less than 0.5%.³¹ In contrast, our test data showed that 47.4% of the 3656 lesions had a BI-RADS4a or higher, and 30.3% contained lesions with a histological diagnosis of breast cancer. As the J-START study is diagnostic on a per-patient basis and our test study is diagnostic on a per-lesion basis, a comparison could not be made in principle. However, notably, in order to validate the accuracy of AI in this study, it was necessary to examine a larger prevalence of malignancy than the prevalence that would be detected by an actual screening test. The sensitivity and specificity of the AI system in this study appear to be sufficiently practical for screening tests when BI-RADS4a or higher is judged to be a recall. However, owing to the above bias, verification from a clinical aspect, such as examining whether the accuracy increases when a clinician gives a final diagnosis after considering the results of the AI system, is necessary for clinical application.

This study is the first to examine the diagnostic accuracy of AI by the manufacturer of the ultrasound device from which images were acquired. Because each manufacturer has a different image-rendering engine, it is possible for a human observer to perceive differences in image quality, and it is unclear how these differences affect the accuracy of the AI system. The sensitivity for images obtained from GE Healthcare Systems was lower than that from the other systems, possibly because of the low number of training datasets. In contrast, the diagnosis of BI-RADS1 images with no lesions or images that can be judged as benign immediately, such as simple cysts, is less affected by the differences between devices from different manufacturers. The low specificity of diagnosis of images from FUJIFILM Healthcare could be attributed to the small amount of BI-RADS1-3 test data. Because this is a bias related to the distribution of the test dataset, it is not considered an essential accuracy issue. It is also pointed out that overfitting is a possible reason for the loss of accuracy in constructing a deep-learning system. Data

augmentation may decrease the risk of overfitting to the training data by introducing some variability to the images but cannot fill in the missing information if the original small training set does not contain samples covering the wide range of disease characteristics in the real-world population.¹⁷ However, data augmentation was not applied to this study, but rather images from multiple facilities and different manufacturers were collected to ensure diversity of data and reduce the risk of overlearning. This data diversity is also a solution to the domain shift problem, the loss of accuracy for domains that the AI is not expecting. Because of the relatively large number of images from Canon Medical Systems in our dataset, it is necessary to add more images from the other two manufacturers to ensure this diversity and to resolve the domain shift problem. Moreover, image rendering technology in ultrasound equipment is remarkable. When major technological improvements in resolution and/or image quality are accomplished, revalidation will need to be performed.

In summary, we annotated 8670 lesions in breast ultrasound images, trained AI using 5014 lesions as the training dataset, and built an AI diagnosis system. To the best of our knowledge, this is the first attempt to establish an AI system that classifies tumor as BI-RADS3 or lower and BI-RADS4a or higher, instead of benign or malignant, to provide important recommendations for clinical action. The AUC, sensitivity, and specificity of the test dataset for the 3656 lesions in 3166 images were 0.95, 91.2%, and 90.7%, respectively. Although there was bias in both the training and test datasets that needed to be improved, the results were sufficient for us to proceed to the next stage of clinical application. Future research should be conducted to verify the safety of AI, determine whether any disadvantages arise from its use, and study its effectiveness in specific clinical applications.

ACKNOWLEDGEMENTS

We greatly acknowledge Ms. Naoko Maeda for her expert opinion and helpful discussions.

FUNDING INFORMATION

The AI algorithms and research funds provided by Fixstars Corporation were used to establish the deep-learning system in this study. This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (C), Grant Number 20K08993.

DISCLOSURE

N. Futatsugi is an employee of Fixstars Corporation, the provider of the AI algorithms used in this study. T. Hayashida received funding from Fixstars Corporation. T. Hayashida and Y. Kitagawa are editorial board members of the *Cancer Science* Journal.

ETHICS STATEMENT

The study was conducted in accordance with the Declaration of Helsinki, and the study protocol was approved by the Institutional Review Board of Keio University School of Medicine (Approval No. 20170146). The local ethics committees approved the study in the participating facilities.

INFORMED CONSENT

Breast ultrasound images were collected using opt-out recruitment methods from each facility.

REGISTRY AND THE REGISTRATION NO. OF THE STUDY/TRIAL

N/A.

ANIMAL STUDIES

N/A.

ORCID

Tetsu Hayashida  <https://orcid.org/0000-0002-1657-803X>

REFERENCES

- Shahan CL, Layne GP. Advances in breast imaging with current screening recommendations and controversies. *Obstet Gynecol Clin North Am.* 2022;49:1-33.
- Berg WA, Blume JD, Cormack JB, et al. Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. *JAMA.* 2008;299:2151-2163.
- Lazarus E, Mainiero MB, Schepps B, Koelliker SL, Livingston LS. BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value. *Radiology.* 2006;239:385-391.
- Spak DA, Plaxco JS, Santiago L, Dryden MJ, Dogan BE. BI-RADS([R]) fifth edition: a summary of changes. *Diagn Interv Imaging.* 2017;98:179-190.
- Abdullah N, Mesurole B, El-Khoury M, Kao E. Breast imaging reporting and data system lexicon for US: interobserver agreement for assessment of breast masses. *Radiology.* 2009;252:665-672.
- Raza S, Chikarmane SA, Neilsen SS, Zorn LM, Birdwell RL. BI-RADS 3, 4, and 5 lesions: value of US in management—follow-up and outcome. *Radiology.* 2008;248:773-781.
- Hooley RJ, Scoutt LM, Philpotts LE. Breast ultrasonography: state of the art. *Radiology.* 2013;268:642-659.
- Barr RG, DeSivestri A, Golatta M. Outcomes of return to routine screening for BI-RADS 3 lesions detected at supplemental automated whole-breast ultrasound in women with dense breasts: a prospective study. *AJR Am J Roentgenol.* 2021;217:1313-1321.
- Janu E, Krikavova L, Little J, et al. Prospective evaluation of contrast-enhanced ultrasound of breast BI-RADS 3-5 lesions. *BMC Med Imaging.* 2020;20:66.
- Fujioka T, Mori M, Kubota K, et al. The utility of deep learning in breast ultrasonic imaging: a review. *Diagnostics (Basel).* 2020;10:1055.
- Li J, Bu Y, Lu S, et al. Development of a deep learning-based model for diagnosing breast nodules with ultrasound. *J Ultrasound Med.* 2021;40:513-520.
- Gao Y, Liu B, Zhu Y, et al. Detection and recognition of ultrasound breast nodules based on semi-supervised deep learning: a powerful alternative strategy. *Quant Imaging Med Surg.* 2021;11:2265-2278.
- Badawy SM, Mohamed AEA, Hefnawy AA, Zidan HE, GadAllah MT, El-Banby GM. Automatic semantic segmentation of breast tumors in ultrasound images based on combining fuzzy logic and deep learning—a feasibility study. *PLoS ONE.* 2021;16:e0251899.
- Vakanski A, Xian M, Freer PE. Attention-enriched deep learning model for breast tumor segmentation in ultrasound images. *Ultrasound Med Biol.* 2020;46:2819-2833.
- Wu T, Sultan LR, Tian J, Cary TW, Sehgal CM. Machine learning for diagnostic ultrasound of triple-negative breast cancer. *Breast Cancer Res Treat.* 2019;173:365-373.
- Tanaka H, Chiu SW, Watanabe T, Kaoku S, Yamaguchi T. Computer-aided diagnosis system for breast ultrasound images using deep learning. *Phys Med Biol.* 2019;64:235013.
- Chan HP, Samala RK, Hadjiiski LM. CAD and AI for breast cancer—recent development and challenges. *Br J Radiol.* 2020;93:20190580.
- Fujioka T, Kubota K, Mori M, et al. Distinction between benign and malignant breast masses at breast ultrasound using deep learning method with convolutional neural network. *Jpn J Radiol.* 2019;37:466-472.
- 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.
- Xiao T, Liu L, Li K, Qin W, Yu S, Li Z. Comparison of transferred deep neural networks in ultrasonic breast masses discrimination. *Biomed Res Int.* 2018;2018:4605191.
- Han S, Kang HK, Jeong JY, et al. A deep learning framework for supporting the classification of breast lesions in ultrasound images. *Phys Med Biol.* 2017;62:7714-7728.
- Antropova N, Huynh BQ, Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Med Phys.* 2017;44:5162-5171.
- Huang Y, Han L, Dou H, et al. Two-stage CNNs for computerized BI-RADS categorization in breast ultrasound images. *Biomed Eng Online.* 2019;18:8.
- Berg WA. BI-RADS 3 on screening breast ultrasound: what is it and what is the appropriate management? *J Breast Imaging.* 2021;3:527-538.
- Han J, Li F, Peng C, et al. Reducing unnecessary biopsy of breast lesions: preliminary results with combination of strain and shear-wave elastography. *Ultrasound Med Biol.* 2019;45:2317-2327.
- Niu Z, Tian JW, Ran HT, et al. Risk-predicted dual nomograms consisting of clinical and ultrasound factors for downgrading BI-RADS category 4a breast lesions – a multiple Centre study. *J Cancer.* 2021;12:292-304.
- Tan SM, Evans AJ, Lam TP, Cheung KL. How relevant is breast cancer screening in the Asia/Pacific region? *Breast.* 2007;16:113-119.
- Maskarinec G, Nagata C, Shimizu H, Kashiki Y. Comparison of mammographic densities and their determinants in women from Japan and Hawaii. *Int J Cancer.* 2002;102:29-33.
- Ohnuki K, Tohno E, Tsunoda H, Uematsu T, Nakajima Y. Overall assessment system of combined mammography and ultrasound for breast cancer screening in Japan. *Breast Cancer.* 2021;28:254-262.
- Sun L, Legood R, Sadique Z, Dos-Santos-Silva I, Yang L. Cost-effectiveness of risk-based breast cancer screening programme, China. *Bull World Health Organ.* 2018;96:568-577.
- Ohuchi N, Suzuki A, Sobue T, et al. Sensitivity and specificity of mammography and adjunctive ultrasonography to screen for breast cancer in the Japan strategic anti-cancer randomized trial (J-START): a randomised controlled trial. *Lancet.* 2016;387:341-348.

How to cite this article: Hayashida T, Odani E, Kikuchi M, et al. Establishment of a deep-learning system to diagnose BI-RADS4a or higher using breast ultrasound for clinical application. *Cancer Sci.* 2022;113:3528-3534. doi: [10.1111/cas.15511](https://doi.org/10.1111/cas.15511)