Physics in Medicine & Biology

Computer-aided diagnosis system for breast ultrasound images using deep learning

To cite this article before publication: Hiroki Tanaka *et al* 2019 *Phys. Med. Biol.* in press https://doi.org/10.1088/1361-6560/ab5093

Computer-aided diagnosis system for breast ultrasound images using deep learning

Hiroki Tanaka[1], Shih-Wei Chiu[1], Takanori Watanabe[2], Setsuko Kaoku[3] and Takuhiro Yamaguchi[1]

[1] Division of Biostatistics, Tohoku University Graduate School of Medicine, Sendai, Miyagi, Japan

[2] Department of Breast Surgery, National Hospital Organization, Sendai Medical Center, Sendai, Miyagi, Japan

[3] Department of Ultrasonics, National Hospital Organization, Osaka National Hospital, Osaka, Osaka, Japan

E-mail: hirotaro-tnk@med.tohoku.ac.jp

Abstract

The purpose of this study was to develop a computer-aided diagnosis (CAD) system for the classification of malignant and benign masses in the breast using ultrasonography based on a convolutional neural network (CNN), a state-of-the-art deep learning technique. We explored the regions for the correct classification by generating a heat map that presented the important regions used by the CNN for human malignancy/benign classification. Clinical data was obtained from a large-scale clinical trial previously conducted by the Japan Association of Breast and Thyroid Sonology. Images of 1536 breast masses (897 malignant and 639 benign) confirmed by pathological examinations were collected, with each breast mass captured from various angles using an ultrasound imaging probe. We constructed an ensemble network by combining two CNN models (VGG19 and ResNet152) fine-tuned on balanced training data with augmentation and used the mass-level

classification method to enable the CNN to classify a given mass using all views. For an independent test set consisting of 154 masses (77 malignant and 77 benign), our network showed outstanding classification performance with a sensitivity of 90.9% (95% confidence interval 84.5–97.3), a specificity of 87.0% (79.5–94.5), and area under the curve of 0.951 (0.916–0.987) compared to that of the two CNN models. In addition, our study indicated that the breast masses themselves were not detected by the CNN as important regions for correct mass classification. Collectively, this CNN-based CAD system is expected to assist doctors by improving the diagnosis of breast cancer in clinical practice.

Key words: computer-aided diagnosis, deep learning, convolutional neural network, ultrasonography, breast cancer

1. Introduction

Ultrasonography has been recommended as an adjunctive modality to mammography (Crystal 2003, Uematsu 2016), which is insufficient for accurate diagnosis of breast cancer in women with dense breast tissue (Shankar et al 2005, Suzuki et al 2008, Freer et al, 2015). However, ultrasonography has the disadvantage of being operator dependent (Baker et al 1999), requiring proficiency in reading ultrasound (US) images, and increasing the false positive rate (Berg et al 2008, Ohuchi et al 2016).

Thus, computer-aided diagnosis (CAD) for breast US images was developed to assist doctors to obtain a second opinion (Giger et al 2013, Lee et al 2015, Takahashi 2017). In general, a CAD system automatically classifies the breast lesions in US images into malignant or benign, which helps doctors in providing a more accurate diagnosis (Chabi et al 2012, Cho et al 2018). A CAD system is designed in three steps: segmentation, feature extraction, and classification based on machine learning (Cheng et al 2010). Among these steps, segmentation and feature extraction pose a bottleneck in developing an efficient CAD system. In the segmentation step, the region of interest (ROI), i.e., the candidate region of a suspicious breast lesion, is separated from the whole image. Because US images are of low quality, it is difficult to effectively perform segmentation (Noble et al 2006, Ding et al 2012). In the feature extraction step, the features, mainly categorized into texture, morphology, model-based, and descriptor features, are computed using ROI information (Cheng et al 2010). Because this computation is an extremely complex task requiring medical

expertise, there are no commercially available CAD systems with both high sensitivity and specificity.

Recently, convolutional neural networks (CNNs), a deep learning technique, have attracted considerable attention as a powerful tool to extract and learn efficient features directly from a given data set (Suzuki 2017). In numerous studies, a CNN was used for the automatic classification of lesions in various medical images, including ultrasonographic scans, mammograms, and fundus and skin images, with outstanding performance results (Byra M et al 2019, Gulshan et al 2016, Esteva et al 2017, Fujioka T et al 2019, Han et al 2017, Huang Y et al 2019, Kooi et al 2017, Shinchijo 2017, Wu JY et al 2019, Xiao et al 2018). However, a CNN has two main limitations. First, it has the "black box" problem; it is so complex that humans cannot understand it. This problem must be solved to apply a CNN in clinical practice, in which accountability is typically imposed (Miotto et al 2017). Second, significant data is needed to train a CNN, which is difficult in the medical filed. Few studies have been conducted on overcoming these limitations.

To our best knowledge, three previous studies using CNNs have been undertaken to classify breast lesions as benign or malignant on US images (Han et al 2017, Xiao et al 2018, Byra M et al 2019). Although the classification performance was outstanding in these studies, no studies have been performed to overcome the described limitations. Moreover, the CNNs developed by these studies could only classify a given mass using one US image. Similar to doctors, CNN should desirably evaluate multiple US images to make diagnoses per mass.

Accordingly, this study aimed to develop a CNN-based CAD system to automatically classify breast masses using all related US images by using a large-scale dataset and visualize the regions detected by the CNN for the correct classification.

2. Materials and Methods

2.1 Image collection and datasets

The B-mode images of breast masses used in this study were collected from 17 facilities in Japan that had participated in the BC-04 study (UMIN000007605), a large-scale clinical trial previously conducted by the Japan Association of Breast and Thyroid Sonology (JABTS). This study population included women with breast masses who were referred for further examination after their initial screening examination of breast cancer and then underwent ultrasonography and pathological examination during their secondary examination from November 2011 to December 2015. Women with breast masses with the following characteristics were excluded: 1) typical cysts and 2) mass lesions ≥ 4.5-cm diameter. We first identified 1543 eligible breast masses, one or two masses per patient, and built a dataset for analysis after removing seven specific non-labeled masses. The final dataset consisted of 1536 breast masses (897 malignant and 639 benign), labeled during pathological examinations or in the subsequent two-year follow-ups of patients diagnosed with benign masses by ultrasonography in the initial screening examination. As ultrasonography

is performed using an US imaging probe that is moved along the surface of the breast skin and angled to capture various views, there were multiple US images for each mass. Therefore, each image was labeled with the same label as the mass.

The dataset was randomly divided according to mass in an 8:1:1 ratio into a training set (743 malignant from a total of 1228 masses), validation set (77 malignant of 154 masses), and test set (77 malignant of 154 masses). This was done so that the number of masses was equal across each class in the validation and test sets, and the remaining were in the training set. Consequently, it was possible to calculate non-biased validation and test accuracy of the CNN. The training set was used for training the CNN, the validation set for selecting the CNN with high validation accuracy during training, and the test set for evaluating classification performance.

The study protocol was approved by the Institutional Review Board of Tohoku University Graduate School of Medicine (2017-1-954, approved on 29 January 2018). The need for informed consent was waived because all data from the image database was anonymous, and an opt-out regarding the secondary use of the data was provided to BC-04 study patients.

2.2 Preprocessing

The breast US images collected in this study were captured using different US equipment within each facility.

Therefore, there was significant variation across views captured by the probe in complete US images, such as in the size and shape, speckle noise, and color (e.g., contrast, saturation, and brightness). To suppress some of these variations, preprocessing was performed in the following four steps:

1. Square views were cropped from each complete US image to remove the black margin, as shown in Figure 1. However, if the view was not a square or two views were compounded into one US image, cropping of a rectangle including the view or separate cropping of each view was performed, respectively (Fig. 1).
2. Three patches were cropped from the left, center, and right or the top, center, and bottom square area of the cropped view so that the size of the square including the view was adapted to the smallest dimension (width or height), as shown in Figure 2.
3. The cropped patches were resized to 256 × 256 pixels.
4. A 3 × 3 median filter was applied on the resized patches to reduce speckle noise.

It is notable that by this approach view patches having the same size and shape can be generated without changing the aspect ratio of the breast mass and subjective cropping. Table 1 lists the number of patients, masses, views, and patches in each set. Color variety was not suppressed. Instead, we performed data augmentation to make the CNN model robust to color variation (discussed subsequently in section 2.3).

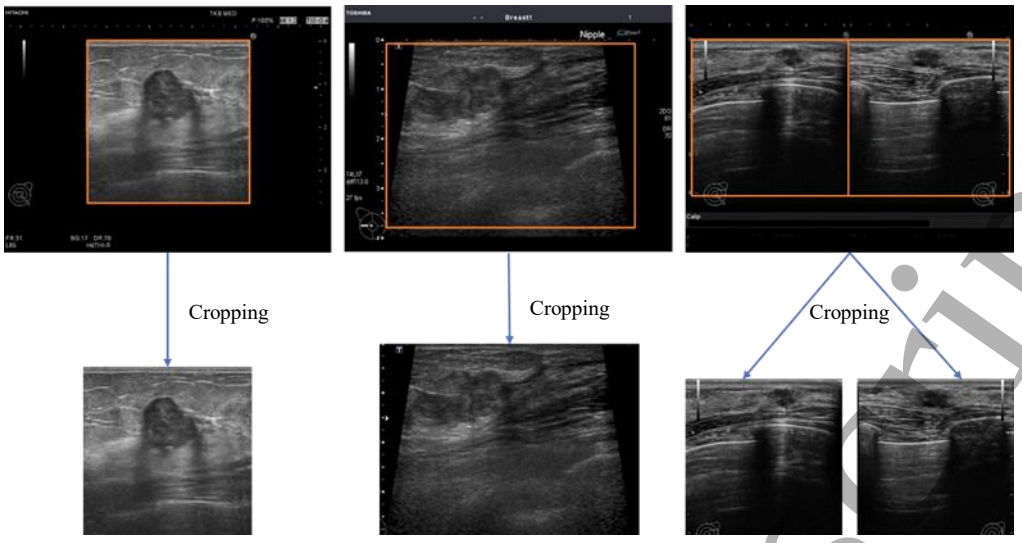
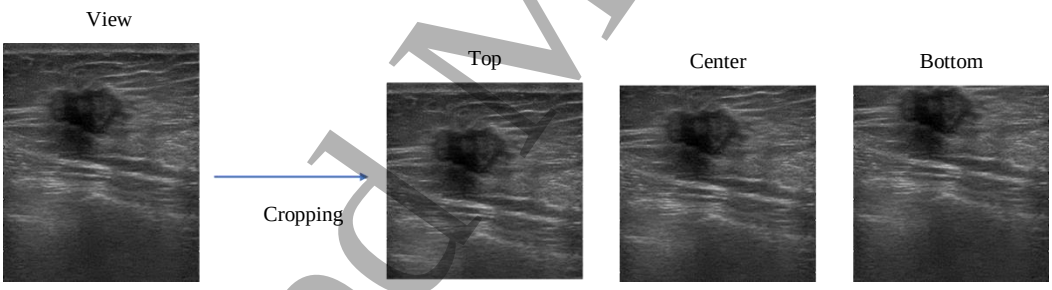
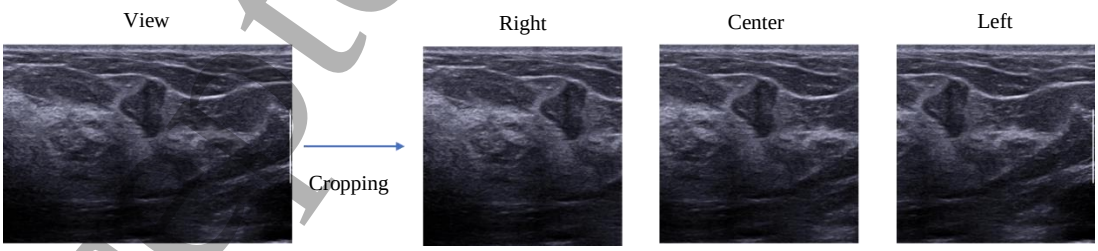


Figure 1. Three ways of cropping the views into a square shape from a complete US image. The orange frame shows the cropped area. The top row shows different complete US images, and the bottom row shows the corresponding views. (a) Cropping for a square-shaped view. (b) Cropping for a non-square-shaped view. (c) Cropping for compounded views.



(a) Cropping three patches for a view with a width less than the height



(b) Cropping three patches for a view with a height less than the width

Figure 2. Cropping to the smallest dimension for views with an uneven size

Table 1. Breakdown of training, validation, and test sets within the dataset

Dataset		No. of patients	No. of masses	No. of views	No. of patches	No. of augmented patches
Training set	Malignancy	707	743	4255	12,765	204,240
	Benign	468	485	2487	7461	204,240
Validation set	Malignancy	74	77	475	1425	-
	Benign	70	77	405	1215	-
Test set	Malignancy	76	77	448	1344	-
	Benign	74	77	402	1206	-

2.3 Data augmentation

Data augmentation is a well-known technique for synthetically generating new samples from an original set of training data (Krizhevsky et al 2012, Hussain et al 2018). In this study, we applied data augmentation for training patches to counteract the imbalanced training set and make the CNN model more robust to the variation in view color.

2.3.1 Augmentation for patches of a minority class.

If a binary-class training set has an unbalanced class distribution, where the number of samples in the majority class outweighs that of the minority class, the classifier becomes excessively sensitive to the majority class. This bias causes a reduction in the performance of the CNN model (Ali et al 2015).

In this study, because our training set was unbalanced, we ensured that the number of training patches in the minority class (benign) was equivalent to that of the patches in the majority class (malignant) using data

augmentation to overcome the above-mentioned problem. First, to augment the patches labeled as benign, we randomly selected 5304 patches from the newly generated patches that were obtained by cropping 256 × 256 centers from the training patches resized to 288 × 288 pixels. Then, we added them to the training patches. Therefore, there were 7461 + 5304 = 12,765 benign patches, equivalent to the number of malignant ones.

2.3.2 Augmentation for improving robustness against variation in US images.

Additional augmented patches were generated from the balanced training set (12,765 malignant and 12,765 benign patches); here, data was augmented by eight transformations: 1) horizontal flip, 2) sharpening, 3) color shifting, 4) color jittering, 5) γ correction with $\gamma > 1$, 6) γ correction with $\gamma < 1$, 7) pyramid down, and 8) pyramid up. An example of these transformations is illustrated in Figure 3. In particular, color transformation, including color shifting, color jittering, and γ-correction, were applied to improve the robustness to color variation. Color jittering randomly changes the contrast, brightness, and saturation; color shifting randomly changes the color intensity values; and γ-correction randomly adjusts the brightness of a view based on γ being larger or smaller than 1. In contrast, "pyramid down," which decreases view resolution, and "pyramid up," which increases resolution, were performed to improve the robustness of resolution variation in the original views. After this operation, training patches were augmented to 408,480 patches in total by applying the horizontal flip transformation followed by the other remaining transformations (12,765

× 2 × 8 = 204,240 patches per class; sixth column in Table 1).

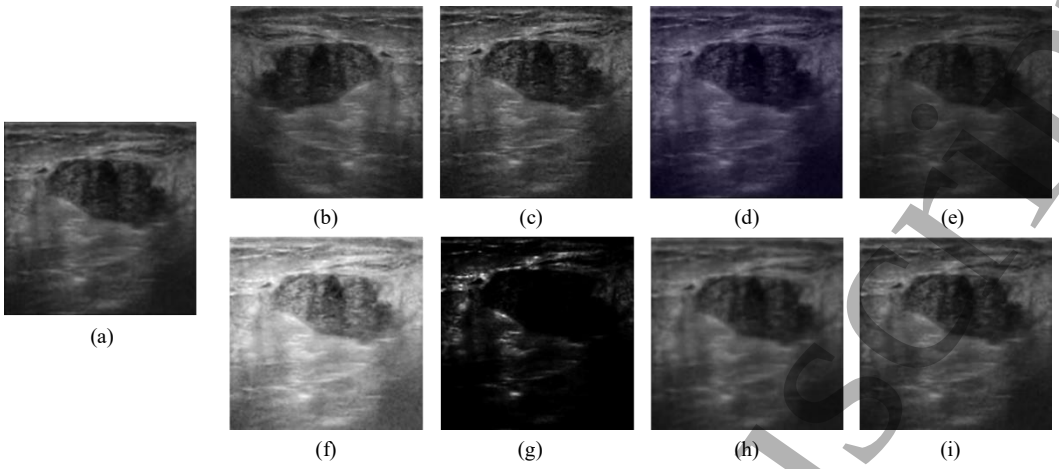


Figure 3. Examples of augmented patches in the training set. (a) An original patch. (b) The transformed patch after a horizontal flip. (c), The transformed patch after sharpening. (d) The transformed patch after color shifting. (e) The transformed patch after color jittering. (f) The transformed patch that is γ-corrected ($\gamma > 1$). (g) The transformed patch that is γ-corrected ($\gamma \leq 1$). (h) The transformed patch after pyramid down transformation. (i) The transformed patch after pyramid up transformation.

2.4 CNNs

In this study, we employed three CNN models, VGG19 (Simonyan et al 2015), ResNet152 (He et al 2016), and an ensemble network. VGG19 consists of 16 convolutional layers with 3 × 3 kernels and three fully connected layers with 4096, 4096, and 1000 units, respectively. Dropout, which deactivates a fraction of the units with a probability of 0.5, was also applied to the fully connected layers to avoid overfitting (Srivastava et al 2014). ResNet152 adopts 50 residual blocks consisting of three convolutional layers and a shortcut connection that skips the convolutional layers. Introducing this shortcut enables the model to dramatically (as deep as 152 layers) overcome the gradient vanishing problem. Regarding this architecture, it is also

notable that batch normalization occurs after each convolutional layer to improve generalization and reduce the training process (Ioffe et al 2015).

The ensemble network consisted of VGG19 and ResNet152, as shown in Figure 4. Ensemble learning is a method that combines the predictions of several trained models to enhance classification performance (Jin et al 2016). A network constructed by this method can output the class probability values of malignant and benign masses with a simple averaging method, in which each probability value predicted by VGG19 and ResNet152 is averaged per class (Jin et al 2016).

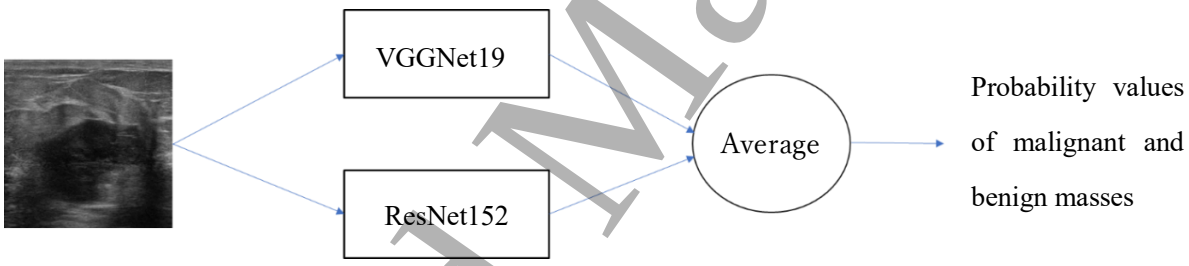


Figure 4. Illustration of the ensemble network

2.5 Training algorithm

We employed transfer learning with pre-trained VGG19 and ResNet152 on 1.2 million natural images from the ImageNet dataset (Deng et al 2009). First, we modified the unit of pre-trained models of the last fully connected layer to match the number of classes in the dataset. Second, we fine-tuned the modified models

by retraining all layers of the augmented training patches resized to 224 × 224 pixels, with a batch size of 64 patches, and using adaptive moment estimation (Adam) (Kingma et al 2014). The latter was designed to adjust the learning rate adaptively in response to the learning process. The training hyperparameters, i.e., the learning rate and weight decay, were set as 0.00001 and 0.0005, respectively. The maximum learning epoch was set as 50. During training, we snapshot the CNN model every 1 epoch and selected the one with the highest validation accuracy as the final model.

2.6 View-level classification and mass-level classification

We proposed a view and mass-level classification method to predict the class per view and mass, respectively. Here, the view level was calculated as the baseline for the comparison with the mass level. In practice, doctors evaluate some views in US images and make a diagnosis per mass (patient) and not per view. Therefore, it is desirable for a CNN to perform its diagnosis accordingly.

The procedure to classify the view and mass was as follows. First, the class probability values of the patches were computed by the fine-tuned CNN models. Then, as shown in Figure 5, the view-level classification was obtained by combining the class probability values of the three patches cropped from a view using the proposed method. For this, each value was averaged per class, and the class with largest value was selected as the classified class. In contrast, as shown in Figure 6, the mass-level classification was

obtained by combining each class' probability value of the patches cropped from all views of a mass using the above-mentioned method. These classifications were performed using test patches resized to 224 × 224 pixels.

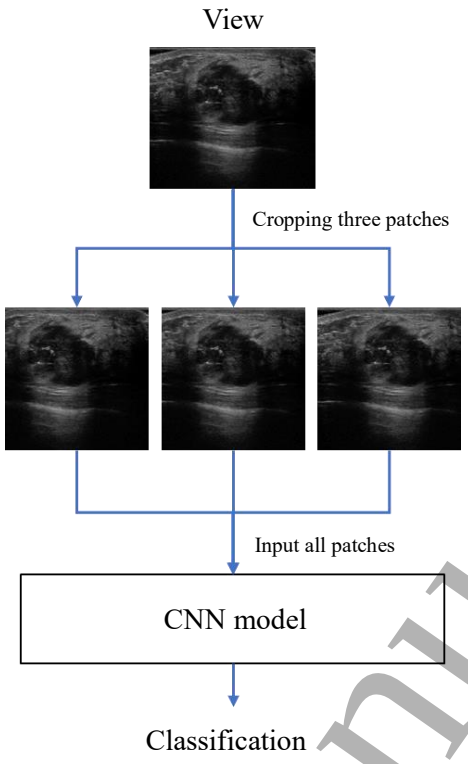


Figure 5. Framework for view-level classification

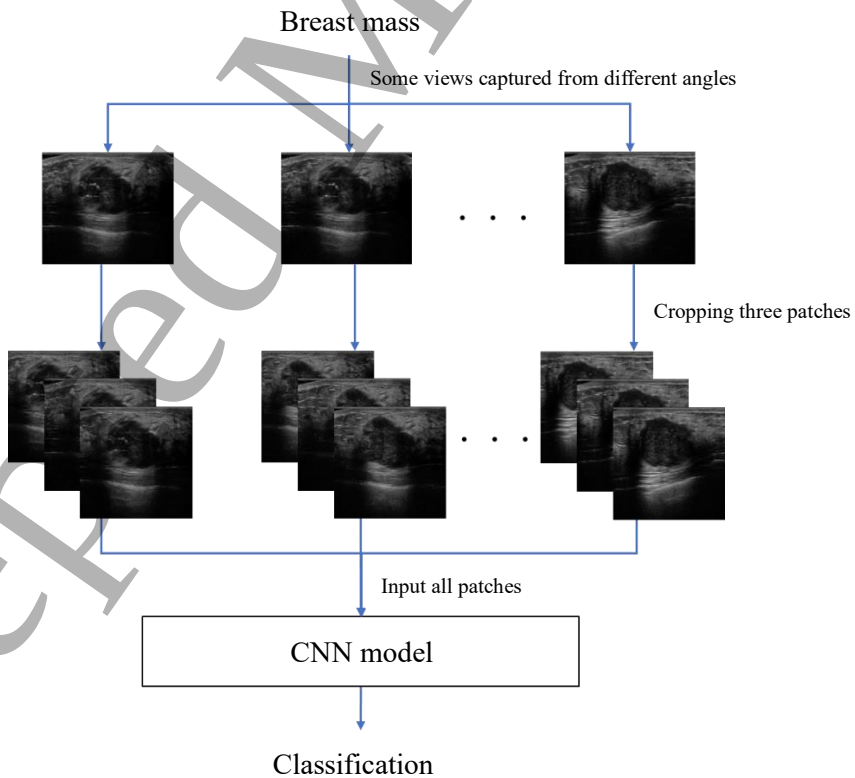


Figure 6. Framework for mass-level classification

2.7 Heat map analysis

To visualize the regions detected by the CNN models, we generated heat maps (Zeiler et al 2014) based on the probability of the correct class using test patches. The heat map indicated the important regions used in the CNN model to correctly classify the patches.

2.8 Evaluation

2.8.1 Classification performance.

In this study, view and mass-level classification performance were evaluated in terms of sensitivity, specificity, precision, f-score, and area under the curve (AUC). This evaluation was performed using test patches. The 95% confidence interval (CI) for AUC was estimated using Hanley's method (Hanley et al 1982).

2.8.2 Mass detection rate.

We calculated the mass detection rates of VGG19 and ResNet152 using heat maps to explore the extent of the breast mass localization in the patches for correct classification by the CNN models. The mass detection rate was calculated via the following four steps:

1. Fifty patches (25 malignant and 25 benign) were randomly selected from the test patches used by both

VGG19 and ResNet152 to classify the correct class.

2. For each selected patch, two heat maps were generated; one from VGG19 and one from ResNet152.

3. The CNN was considered to have detected the mass if the red region of the heat map and mass overlapped partially or if the red region covered the mass. If the red region covered an extremely broad range of the region containing the mass or the mass did not appear clearly in a patch, the corresponding patches were excluded from the calculation. These evaluations were carefully performed by a breast cancer expert who was a JABTS board member.

4. Based on the results of step 3, malignant, benign, and overall mass detection rates were calculated using the following formula:

$$mass\ detection\ rate\ = \frac{number\ of\ detected\ masses}{number\ of\ randmly\ sellcted\ masses} \times 100$$

2.9 Software and hardware

The image augmentation, CNN models, and heat map analysis were implemented in SAS software (SAS® Visual Data Mining and Machine Learning 8.3 / SAS® Viya® 3.4, Copyright ©2018 SAS Institute Inc.) through Python interface (Smith et al 2017) and DLPy 0.7, the high-level Python APIs designed to efficiently apply the deep learning methods in SAS Visual Data Mining and Machine Learning. To accelerate the training, we used NVIDIA Tesla P100 GPU×2 and 14C XeonE5-2580v4 × 2 with 64 GB RAM on CentOS7 Operating

System.

3 Results

3.1 Classification performance

The view and mass-level classification performance of VGG19, ResNet152, and the ensemble network were evaluated using the test set, and the results are summarized in Table 3. In the view-level classification of 850 test views, the ensemble network achieves the best performance with the highest specificity (83.1%; 95% CI 79.4–86.7) and AUC of 0.938 (0.921–0.954).

The performance of each CNN model was better at mass-level classification of 154 test masses than at view-level classification. In the case of ResNet152, AUC was enhanced from 0.919 to 0.935, sensitivity from 0.875 to 0.922, and specificity from 0.811 to 0.844. The ensemble network achieved the best performance with an accuracy of 89.0% (84.0–93.9), a sensitivity of 90.9% (84.5–97.3), specificity of 87.0% (79.5–94.5), precision of 87.5% (76.9–92.7), F-score of 0.892, and AUC of 0.951 (0.916–0.987). In comparison, the accuracy, sensitivity, specificity, precision, F-score, and AUC were 85.7% (80.2–91.2), 87.0% (79.5–94.5), 84.4% (76.3–92.5), 84.8% (76.9–92.7), 0.859, and 0.945 (0.907–0.983), respectively, for VCC19. They were 88.3% (83.2–93.4), 92.2% (86.2–98.2), 84.4% (76.3–92.5), 85.5% (78.0–93.1), 0.888, and 0.935 (0.894–0.976), respectively, for ResNet152.

Table 2. Classification performance of the three CNN models in view-level and mass-level classifications performed on an independent test set

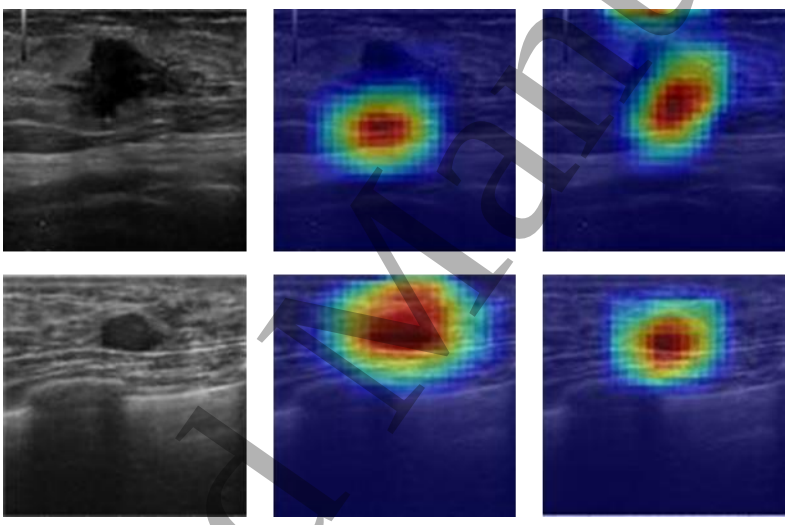
	Accuracy (95%CI)	Sensitivity (95%CI)	Specificity (95%CI)	Precision (95%CI)	F-score	AUC (95%CI)
View-level classification						
VGG19	86.4% (84.0–88.7)	90.0% (87.2–92.7)	82.3% (78.6–86.1)	85.0% (81.8–88.2)	0.874	0.926 (0.908–0.944)
ResNet152	84.5% (82.9–85.8)	87.5% (84.4–90.6)	81.1% (77.3–84.9)	83.8% (80.4–87.1)	0.856	0.919 (0.900–0.938)
Ensemble network	86.0% (83.7–88.3)	88.6% (85.7–91.6)	83.1% (79.4–86.7)	85.4% (82.2–88.6)	0.870	0.938 (0.921-0.954)
Mass-level classification						
VGG19	85.7% (80.2–91.2)	87.0% (79.5–94.5)	84.4% (76.3–92.5)	84.8% (76.9–92.7)	0.859	0.945 (0.907–0.983)
ResNet152	88.3% (83.2–93.4)	92.2% (86.2–98.2)	84.4% (76.3–92.5)	85.5% (78.0–93.1)	0.888	0.935 (0.894–0.976)
Ensemble network	89.0% (84.0–93.9)	90.9% (84.5–97.3)	87.0% (79.5–94.5)	87.5% (80.3–94.7)	0.892	0.951 (0.916–0.987)

3.2 Mass detection rate

Table 2 lists the malignant, benign, and overall mass detection rates in VGG19 and ResNet152. Of 2550 test patches, 1956 that were correctly classified both by VGG19 and ResNet152 were identified. Fifty of these were randomly selected for analysis. Six patches classified by VGG19 and four by ResNet152 were excluded from the analysis because of the reasons mentioned in section 2.8 (2.8.2 and 2.8.3). The overall detection rate did not reach 50% in either VGG19 (47.7%) or ResNet152 (37.0%), whereas the benign mass detection rate was higher in both VGG19 (78.3%) and ResNet152 (60.9%) than the malignant mass detection rate (14.3% and 13.0%, respectively). As shown in Figures 4 and 5, various important regions for correct classification in the patches, including the region with the mass (top and bottom row in Figure 4) and region without the mass (top and bottom row in Figure 5), were detected by VGG19 and ResNet152. One important region for the correct classification differs in the CNN models.

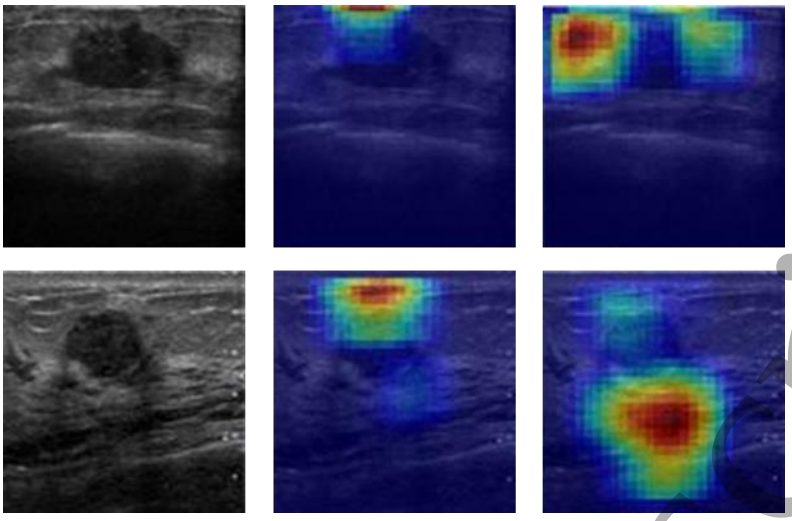
Table 3. Detection rate for malignant, benign, and overall masses in randomly selected test patches

	Number of detected masses/number of randomly selected masses (detection rate; %)	
	VGG19	ResNet152
Malignant mass	3/21 (14.3)	3/23 (13.0)
Benign mass	18/23 (78.3)	14/23 (60.9)
Overall mass	21/44 (47.7)	17/46 (37.0)



(a) (b) (c)

Figure 4. Visualization of an important region (highlighted area in patches) for correct classification via heat map analysis when both VGG19 and ResNet152 detect the mass. The top row shows a malignant mass, and the bottom a benign mass. (a) Original patches. (b) Patches overlaid with the heat map generated by VGG19. (c) Patches overlaid with the heat map generated by ResNet152.



(a) (b) (c)

Figure 5. Visualization of an important region (highlighted area in the patch) for correct classification via heat map analysis when both VGG19 and ResNet152 do not detect the mass. The top row shows a malignant mass, and the bottom a benign mass. (a) Original patches. (b) Patches overlaid with the heat map generated by VGG19. (c) Patches overlaid with the heat map generated by ResNet152.

4. Discussion

We developed three CNN-based CAD systems (VGG19, ResNet152, and an ensemble network) to classify breast masses in US images as benign or malignant using significant data (more than 1000 masses). We tested the view-level and mass-level classification methods; the mass-level outperformed the view-level classification. In the mass-level classification, the ensemble network yielded the best performance, with high sensitivity and specificity, and an AUC over 0.95. In addition, we visualized the basis for the classification to overcome the "black box" problem that a CNN suffers from. We revealed that most masses in the patches were not detected by the CNN as important regions for correct classification.

A large-scale randomized controlled trial conducted by the Japan Strategic Anti-cancer Randomized Trial (J-START) (Ohuchi et al 2016) in 2007 indicated that sensitivity significantly increased by combining mammography and ultrasonography (intervention group) screening than by mammography alone (control group) (91.1 vs 77.0; $p < 0.0004$). However, specificity was significantly lower in the intervention than in the control group (87.7 vs 91.4%; $p < 0.0001$). In comparison, mass-level classification by the ensemble network was comparable to that in the intervention group in terms of sensitivity and specificity. However, a reasonable comparison is difficult because each study used different populations; J-START focused on women with mass or non-mass lesions who underwent initial screening examinations, whereas our study focused on women with mass lesions who underwent secondary breast cancer screening examinations.

Nevertheless, we believe that our model is comparable to a human analysis and might assist doctors in diagnosis, enhance specificity, reduce the number of unnecessary biopsies in non-breast cancer cases, and solve problems concerning inadequate numbers of experienced doctors.

Here, mass-level classification outperformed view-level classification for all models. This observation suggests that mass-level classification may enable the robust classification of views by CNN that are otherwise difficult to classify correctly. In addition, this classification method is more practical in clinical practice because, analogous to doctors systematically evaluating multiple mass views to diagnose breast cancer, it allows CNN models to classify a breast mass using all of the related views.

Although all the patches correctly classified by both VGG19 and ResNet152 were used for the calculation of the mass detection rate, more than half of the breast masses were not detected. This suggests that in most patches, useful classification information did not involve masses, particularly in case of malignant masses. If the important regions are explicitly explained from a histological perspective and doctors can interpret them, the detection rate may not need to be high, and the prediction of the CNN will be more reliable. However, we have not examined the details of the important regions provided by the CNNs. This challenging task should be addressed in future studies to make the system more acceptable in clinical practice, as some doctors are likely hesitant to rely on a CNN prediction if the prediction process is not obvious.

Different CNN architectures can learn different features; shallow networks are suitable for learning low-

level features while deep networks can learn high-level features (Ma et al 2017). It is suggested that depending on CNN architecture, different level of features can be extracted from data. Thus, the ensemble network could classify breast masses using the information from multiple feature levels, leading to enhanced classification performance. Recently, five studies employing very deep CNNs to classify breast lesions as benign or malignant using US images have been undertaken. (Han et al 2017) employed a transferred GoogleNet CNN model architecture; (Xiao et al 2018) employed a model with combined features extracted by three transferred CNN models (ResNet50, Xception, and InceptionV3); (Byra M et al 2019) employed a transferred VGGNet19 to which a matching layer was introduced to rescale the pixel intensities of the grayscale US images; (Fujioka T et al 2019) employed GoogleNet Inception v2 which is an improvement over GoogleNet and has Batch Normalization; and (Wu JY et al 2019) investigated the performance of the commercial S-Detect software (designed by Samsung Medison) based on CNN. In the first model, the sensitivity, specificity, and AUC were 0.843, 0.961, and 0.960, respectively (Han et al 2017); in the second model, they were 0.887, 0.894, and 0.93, respectively (Xiao et al 2018); in the third model, they were 0.848, 0.897, and 0.936, respectively (Byra et al 2019) ; in the fourth model, they were 0.958, 0.875 and 0.913, respectively (Fujioka T et al 2019); in the fifth model, they were 0.876, 0.813, and 0.845, respectively (Wu JY et al 2019). Our ensemble network with the mass-level classification method produced results comparable to these in terms of AUC. Both models in the previous studies showed better specificity than sensitivity; in

contrast, our model showed a higher sensitivity than specificity. The difference is probably caused by the unbalanced training data and the presence of more malignant than benign masses in this study. This is opposite to the three previous studies (Byra et al Han et al 2017 2019 Xiao et al 2018), in which the CNN classified the majority class easily, and the sensitivity was high when the majority class was malignant. In addition, both CNNs in the previous studies performed the classification per view not per mass, which may have led to longer screening times and confused doctors by interpreting all prediction results per view. Thus, our model with mass-level classification may be more suited to clinical practice. Unlike our model, (Huang Y et al 2019) developed a CNN-based automatic grading system that can grade US-imaged breast tumors into five categories in accordance with the Breast Imaging Reporting and Data System (BI-RADS). Though direct comparison cannot be performed because the number of classes for classification is different from our model, the grading system shows a high classification performance with accuracy of 0.998 in Category "3", 0.940 in Category "4A", 0.734 in Category "4B", 0.922 in Category "4C", and 0.876 in Category "5" (Huang Y et al 2019).

In this study, before training the CNN, we generated a balanced training set by augmenting the benign patches by cropping the 256 × 256 centers from the training patches, which were then resized to 288 × 288 pixels. However, sensitivity was higher than specificity in all the CNNs. Thus, we consider that this method, generating nearly unchanged patches from the original, may not be a useful approach to balance the class

distribution.

There are several limitations in this study. First, the test set was too small to evaluate the classification performance of the CNN models. In this study, the training set size was increased and the test set size was decreased to train the CNN with maximum possible data to enhance its classification performance. Second, the dataset did not reflect the population of the patients who undergo breast cancer screening, which includes patients with both mass and non-mass lesions. Instead, we targeted only women with mass lesions diagnosed in the secondary examination, which resulted in having more malignant than benign masses in our dataset. This suggests that the developed CNN-based CAD system cannot be applied to women in the initial screening examination, but only to those with breast masses in the second examination. Third, mass detection rate was biased as it was evaluated subjectively by one doctor and not all test patches were used for the calculation.

5. Conclusion

In this study, we developed a CNN-based CAD system (ensemble network with mass-level classification) for classifying benign and malignant masses in US images, indicating a promising performance with high sensitivity and specificity. It is considered that this system is useful for doctors as a supplemental modality for screening women with breast masses. However, some doctors may be suspicious of the prediction potential of the system because it is not explicitly capable of explaining the important regions in the input

data that it uses to classify the masses from a histopathological perspective. Our study indicated that almost all breast masses themselves were not detected by the CNN as the important regions for correct classification. To overcome this challenge and further improve classification performance, we plan to conduct an additional study using more sophisticated heat map analysis as well as a larger dataset and an improved CNN architecture.

Acknowledgments

We would like to thank SAS Institute Japan Ltd. for their technical support and provision of the development environment for deep learning, which was founded by SAS Institute Inc.

References

Baker J A, Kornguth P J, Soo M S, Walsh R and Mengoni P 1999 Sonography of solid breast lesions: Observer variability of lesion description and assessment *AJR Am. J. Roentgenol.* **172** 1621–1625

Berg W A et al 2008 Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer *JAMA* **299** 2151–2163

Byra M et al 2019 Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion *Med. Phys.* **46** 746–755

Chabi M L, Borget I, Ardiles R, Aboud G, Boussouar S, Vilar V, Dromain C and Balleyguier C 2012 Evaluation of the accuracy of a computer-aided diagnosis (CAD) system in breast ultrasound according to the radiologist's experience *Acad. Radiol.* **19** 311–319

Cheng H D, Shan J, Ju W, Guo Y and Zhang L 2010 Automated breast cancer detection and classification using ultrasound images: a survey *Pattern Recogn.* **43** 299–317

Cho E, Kim E K, Song M K and Yoon J H 2018 Application of Computer-Aided Diagnosis on Breast Ultrasonography: Evaluation of Diagnostic Performances and Agreement of Radiologists According to Different Levels of Experience *J. Ultrasound. Med.* **37** 209–216

Crystal P, Strano S D, Shcharynski S and Koretz M J 2003 Using sonography to screen women with mammographically dense breasts *AJR Am. J. Roentgenol*. **181** 177–182

Deng J, Dong W, Socher R, Li L J, Li K and Fei L F 2009 ImageNet: A Large-Scale Hierarchical Image Database *IEEE Conf. on Computer Vision and Pattern Recognition* 248–255

Ding J, Cheng H D, Huang J, Liu J and Zhang Y 2012 Breast ultrasound image classification based on multiple-instance learning *J. Digit. Imaging* **25** 620–627

Esteva A, Kuprel B, Novoa R A, Ko J, Swetter S M, Blau H M and Thrun S 2017 Dermatologist-level classification of skin cancer with deep neural networks *Nature* **542** 115–118

Fujioka T et al 2019 Distinction between benign and malignant breast masses at breast ultrasound using deep learning method with convolutional neural network *Jpn. J. Radiol.* **6** 466–472

Freer P E 2015 Mammographic breast density: impact on breast cancer risk and implications for screening *Radiographics* **35** 302–315

Giger M L, Karssemeijer N and Schnabel J A 2013 Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. *Annu. Rev. Biomed. Eng.* **15** 327–357

Gulshan V et al 2016 Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs *JAMA* **316** 2402–2410

Han S, Kang H K, Jeong J Y, Park M H, Kim W, Bang W C and Seong Y K 2017 A deep learning framework for supporting the classification of breast lesions in ultrasound Images *Phys. Med. Biol*. **62** 7714–7728

Hanley J A and McNeil B J 1982 The meaning and use of the area under a receiver operating characteristic

(ROC) curve *Radiology* **143** 29–36

He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* 770–778

Huang Y et al 2019 Two-stage CNNs for computerized BI-RADS categorization in breast ultrasound images *Biomed. Eng. Online.* **18** 8.

Hussain Z, Gimenez F, Yi D and Rubin D 2017 Differential Data Augmentation Techniques for Medical Imaging Classification Tasks *AMIA Annu. Symp. Proc.* **2017** 979–984

Ioffe S and Szegedy C 2015 Batch normalization: Accelerating deep network training by reducing internal covariate shift *Int. Conf. on Machine Learning* **37** 448–456.

Jin L P and Dong J 2016 Ensemble Deep Learning for Biomedical Time Series Classification *Comput. Intell. Neurosci.* **45** 1–13

Kingma D and Ba J 2015 Adam: a method for stochastic optimization *Int. Conf. for Learning Representation*

Kooi T, van Ginneken B, Karssemeijer N and den Heeten A 2017 Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network *Med. Phys.* **44** 1017–1027

Kotsiantis S, Kanellopoulos D and Pintelas P 2006 Handling imbalanced Datasets: A review *GESTS Int'l Trans. Computer Science and Eng.* **30** 25–36

Krizhevsky A, Sutskever I and Hinton G E 2012 Imagenet classification with deep convolutional neural networks Proc. Adv. Neural Inf. Process. Syst. **1** 1097–1105

Lee H and Chen Y P P 2015 Image based computer aided diagnosis system for cancer detection. *Expert Syst. Appl.* **42** 5356–5365

Ma J et al 2017 A pre-trained convolutional neural network based method for thyroid nodule diagnosis *Ultrasonics* **73** 221–230

Miotto R, Wang F, Wang S, Jiang X and Dudley J T 2018 Deep learning for healthcare: review, opportunities and challenges *Brief. Bioinform.* **19** 1236–1246

Noble J A and Boukerroui D 2006 Ultrasound Image segmentation: a survey *IEEE Trans. Med. Imag.* **25** 987–1010

Ohuchi N et al 2016 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* **387** 341–348

Shankar P M, Piccoli C W, Reid J M, Forsberg F and Goldberg B B 2005 Application of the compound probability density function for characterization of breast masses in ultrasound B scans *Phys. Med. Biol.* **50** 2241–2248

Shinchijo S et al 2017 Application of Convolutional Neural Networks in the Diagnosis of Helicobacter pylori

Infection Based on Endoscopic Images *EBioMedicine* **25** 106–111

Simonyan K and Zisserman A 2015 Very deep convolutional networks for large-scale image recognition *Int. Conf. on Learning Representation (ICLR)*

Smith KD, Meng X SAS® Viya®: The Python Perspective 2017 (Cary: SAS Institute Inc.)

Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: a simple way to prevent neural networks from overfitting *J. Mach. Learn. Res.* 15 1929–58

Suzuki A et al. 2008 Age-specific interval breast cancers in Japan: estimation of the proper sensitivity of screening using a population-based cancer registry *Cancer Sci.* **99** 2264–2267

Suzuki K 2017 Overview of deep learning in medical imaging *Radiol. Phys. Technol.* **10** 257–273

Takahashi R and Kajikawa Y 2017 Computer-aided diagnosis: A survey with bibliometric analysis *Int. J. Med. Inform.* **101** 58–67

Uematsu T 2016 The need for supplemental breast cancer screening modalities: a perspective of population-based breast cancer screening programs in Japan *Breast Cancer* **24** 26–31

Wu JY et al 2019 Computer-Aided Diagnosis of Solid Breast Lesions With Ultrasound: Factors Associated With False-negative and False-positive Results *J. Ultrasound. Med.* [Epub ahead of print]

Xiao T, Liu L, Li K, Qin W, Yu S and Li Z 2018 Comparison of Transferred Deep Neural Networks in Ultrasonic Breast Masses Discrimination *Biomed Res. Int.* **2018** 1–9

Zeiler M D and Fergus R 2014 Visualizing and understanding convolutional networks *European Conf. on Computer Vision* 818–33