Ultrasound in Med. & Biol., Vol. 46, No. 5, pp. 1119−1132, 2020

0301-5629/$ - see front matter

https://doi.org/10.1016/j.ultrasmedbio.2020.01.001

● *Original Contribution*

# BREAST CANCER CLASSIFICATION IN AUTOMATED BREAST ULTRASOUND USING MULTIVIEW CONVOLUTIONAL NEURAL NETWORK WITH TRANSFER LEARNING

YI WANG,*[,1] EUN JUNG CHOI,†[,1] YOUNHEE CHOI,* HAO ZHANG,* GONG YONG JIN,† and SEOK-BUM KO*

* Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon, Canada; and † Department of Radiology, Research Institute of Clinical Medicine of Jeonbuk National University−Biomedical Research Institute of Jeonbuk National University Hospital, Jeonbuk National University Medical School, Jeonju City, Jeollabuk-Do, South Korea

(*Received* 11 *July* 2019; *revised* 12 *December* 2019; *in final from* 2 *January* 2020)

**Abstract—To assist radiologists in breast cancer classification in automated breast ultrasound (ABUS) imaging, we propose a computer-aided diagnosis based on a convolutional neural network (CNN) that classifies breast lesions as benign and malignant. The proposed CNN adopts a modified Inception-v3 architecture to provide efficient feature extraction in ABUS imaging. Because the ABUS images can be visualized in transverse and coronal views, the proposed CNN provides an efficient way to extract multiview features from both views. The proposed CNN was trained and evaluated on 316 breast lesions (135 malignant and 181 benign). An observer performance test was conducted to compare five human reviewers' diagnostic performance before and after referring to the predicting outcomes of the proposed CNN. Our method achieved an area under the curve (AUC) value of 0.9468 with five-folder cross-validation, for which the sensitivity and specificity were 0.886 and 0.876, respectively. Compared with conventional machine learning-based feature extraction schemes, particularly principal component analysis (PCA) and histogram of oriented gradients (HOG), our method achieved a significant improvement in classification performance. The proposed CNN achieved a >10% increased AUC value compared with PCA and HOG. During the observer performance test, the diagnostic results of all human reviewers had increased AUC values and sensitivities after referring to the classification results of the proposed CNN, and four of the five human reviewers' AUCs were significantly improved. The proposed CNN employing a multiview strategy showed promise for the diagnosis of breast cancer, and could be used as a second reviewer for increasing diagnostic reliability. (E-mail: seokbum.ko@usask.ca)  © 2020 World Federation for Ultrasound in Medicine & Biology. All rights reserved.**

***Key Words:*** **Convolutional neural network, Multiview convolutional neural network, Transfer learning, Breast cancer classification, Automated breast ultrasound.**

## INTRODUCTION

Breast cancer is the second leading cause of cancer death in women (Siegel et al. 2019). To reduce the mortality rate of breast cancer, early detection and treatment are important. Mammography and ultrasound (US) are common screening modalities used to diagnose breast cancer. Recent studies have reported improved diagnostic performance in the dense breast by interpreting handheld US imaging in addition to mammography (Kolb et al. 2007; Brem et al. 2015; Thigpen et al. 2018).

Nevertheless, interpreting the handheld US is operator dependent since the visualization of one handheld US image is limited in one certain orientation (Brem et al. 2015; Thigpen et al. 2018). To minimize the operator dependence, the automated breast ultrasound (ABUS) has been introduced to provide the capability of breast visualization in a 3-D volume where it is helpful to locate breast lesion prior to diagnose breast cancer. One example of ABUS is shown in Figure 1. The ABUS imaging reconstructs the breast in transverse view, which is similar to the handheld US. Besides, the ABUS imaging offers a coronal view of the breast. In practice, although ABUS demonstrates better reproducibility and less time-consuming than the hand-held US (Shin et al., 2015), screening ABUS still takes a significantly longer time than mammography (Rella et al., 2018).

Address correspondence to: Seok-Bum Ko, Department of Electrical and Computer Engineering, University of Saskatchewan, 57 Campus Drive, Saskatoon, SK S7N 5A9, Canada. E-mail: seokbum.ko@usask.ca

[1] Yi Wang and Eun Jung Choi contributed equally to this article.

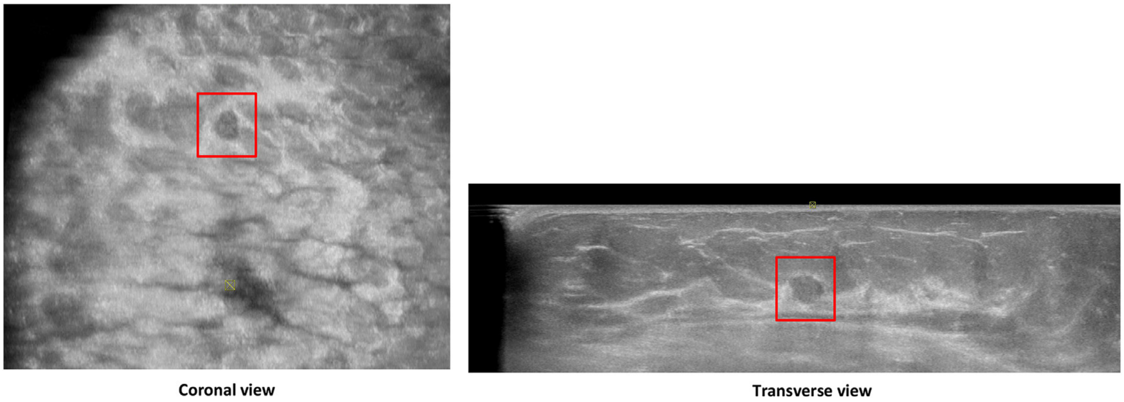

Fig. 1. Examples of automated breast ultrasound images acquired during screening in a 50-y-old woman. A benign lesion located in both coronal and transverse views is enclosed by *red rectangular boxes*.

The usefulness of a computer-aided diagnosis (CADx) system as a second reviewer to support radiologists in decision making has also been reported. Furthermore, the screening time, especially for interobserver variation, is reduced (Wang et al. 2010; Marcomini et al. 2016; Van Zelst et al. 2017). To capture unique characteristics of lesion patches, machine learning-based feature extractors are commonly applied in CADx systems. The common feature extractors include histogram of oriented gradients (HOG) (Moura and Guevara Lopez 2013), local binary pattern (Iakovidis et al. 2008) and principal component analysis (PCA) (Huang et al. 2005). As an alternative, morphologic operations are also used to extract breast lesion features (Cheng et al. 2010; Tan et al. 2012). After feature extraction, a classifier is followed, such as support vector machine (SVM) (Tan et al. 2012; Huang et al. 2017b), *k*-nearest neighbor (Ahmed Medjahed et al. 2013), logistic regression (Moon et al. 2011) and linear discriminant analysis (Rajaguru and Kumar Prabhakar 2017; Sadeghi-Naini et al. 2017). However, both machine learning-based and morphologic feature extraction schemes consist of several manual processing steps. For this reason, it is complicated to select optimal feature combinations, which require numerous trials and errors before a convincing result is obtained at each intermediate step (Han et al. 2017; Chiang et al. 2019).

The convolutional neural network (CNN), one of the deep learning approaches, has achieved promising results in natural image classification tasks (Russakovsky et al. 2015). Instead of using the aforementioned feature extractors, the CNN extracts imaging features directly from input data without designing explicit feature extractors or tuning intermediate results at each intermediate step. In recent years, CNN has been effectively applied to the field of breast cancer classification (Han et al. 2017; Byra 2018; Byra et al. 2018b). In particular, transfer learning is a good solution for a CNN-based CADx system if the size of the database is limited. With transfer learning, the pre-trained models, explicitly trained by a large-scale natural image database such as ImageNet (Russakovsky et al. 2015), are tuned to be applied for solving classification problems in the medical imaging domain. In Han et al. (2017), a pre-trained GoogLeNet (Szegedy et al. 2015) was adopted for discrimination between malignant and benign lesions. A pre-trained VGG (Simonyan and Zisserman 2014) was used for lesion feature extraction (Byra 2018), followed by Fisher discriminant analysis to classify breast lesions. In the work of Byra et al. (2018b), the Inception-v3 (Szegedy et al. 2016) was used to classify breast lesions in US imaging.

Recently, the multiview CNN has been introduced and found to improve breast lesion classification in mammography (Geras et al. 2017). In this method, mammography visualizes the same breast lesion in different mammographic views such as cranial-caudal and mediolateral oblique views. To provide more useful features, multiple CNNs were used to extract features from different views independently. In terms of US imaging, the multiview strategy has also been applied to CNN by Han et al. (2017). For each breast lesion, multiple lesion patches are cropped by multiple scales. Through combination of lesion patches with multiple scales, classification performance was improved compared with that using a lesion patch with a single scale. Nevertheless, there are some limitations to the work of Han et al. (2017). First, the sizes of the lesions are required before generating the multiview lesion patches, which involves manual intervention. Second, extraction of features from multiple scaled lesion patches may be less generic compared with that in Geras et al. (2017), in which the overall extracted features are redundant.

In this article, a multiview CNN is proposed to classify breast lesions between malignant and benign. Without any manual pre-processing step, the proposed CNN extracts features from the lesion patch directly. To provide an efficient feature extraction scheme, more lesion features are extracted by the proposed CNN from different views of the automated breast ultrasound (ABUS) imaging. On the basis of our results, the proposed multiview CNN outperformed conventional machine learning approaches and single-view CNNs. Furthermore, an observer performance test was conducted. The test results revealed that the proposed CNN could be used as a second reviewer to improve human reviewers' diagnostic performance in breast cancer classification. The main contributions of the proposed CNN are as follows:

- The proposed CNN is the first CNN model that has been successfully applied to breast cancer classification in ABUS imaging, to the best of our knowledge.
- Multiview strategies have been utilized in the proposed CNN to provide robust and effective feature extraction.
- Comprehensive evaluations are performed to ensure the effectiveness of the proposed CNN model.
- An observer performance test is performed to justify the usefulness of the proposed CNN model from the clinical perspective.

## METHODS

### *Clinical data set*

The data set used in this study was collected between March 2012 and March 2018 at Jeonbuk National University Hospital (JNUH). An ACUSON S2000 (Siemens Medical Solutions, Mountain View, CA, USA) automated breast volume scanner (ABVS) in combination with a 15-cm-wide linear array transducer was used to acquire ABUS images by a single technologist who had more than 3 y of experience in operating ABVS. Depending on breast size, acquisition frequencies vary from 9−11 MHz. Each ABUS scan produces 15.4 × 16.8 × maximum 6 cm volume data with a slice thickness of 1 mm. The volume data obtained were reconstructed to 2-D slice images through multiplanar reconstruction. For our data set, slice images of each breast lesion were collected from both coronal and transverse views. For our retrospective study, the informed consent for data usage was approved by the institutional review board of JNUH.

A total of 316 breast lesions in 263 patients (age range: 28−76 y, mean age: 51.4 ± 9.8 y) were included in our data set, which consists of 135 malignant and 181 benign lesions. Mean lesion size was 13.23 mm with a standard deviation of 4.29 mm. A detailed distribution of the number of lesions by size is provided in Table 1. All lesions were pathologically confirmed after surgery or biopsy. For each benign lesion, a 2-y follow-up examination was done to ensure the lesion was unchanged. The ground truth of each lesion was annotated by the physicians using the bounding box to cover the lesion. After annotation, the ground truth of each lesion was verified by a radiologist with 8 y of experience in US imaging. The size of the bounding box varied with the size of the lesion. The sample of bounding boxes is illustrated in Figure 2. Lesion patches were cropped along the bounding boxes, and the proposed CNN extracted the lesion features from the cropped lesion patches. To generate more effective lesion patches, multiple lesion patches were cropped from different slices of each lesion. For example, as illustrated in Figure 2, the same lesion at different slices has different visual representations. Thereafter, 743 malignant patches (359 and 384 patches from coronal and transverse views, respectively) and 419 benign patches (233 and 186 patches from coronal and transverse views, respectively) were generated in this study.

Table 1. Distribution of the number of lesions by size

| Size (mm) | Malignant (n = 135) | Benign (n = 181) |
|---|---|---|
| 1−5 | 4 (2.96%) | 32 (17.68%) |
| 5−10 | 37 (27.41%) | 91 (50.28%) |
| 10−20 | 94 (69.63%) | 58 (30.04%) |
| Mean ± SD | 13.23 ± 4.29 | 9.28 ± 3.99 |

SD = standard deviation.

### *CNN-based lesion feature extraction and classification*

Lesion features were extracted and classified by employing a modified Inception-v3 CNN (Szegedy et al. 2016), the third generation of GoogLeNet (Szegedy et al. 2015). One novel aspect of Inception-v3 is the Inception modules used to replace the conventional convolutional layer. The Inception module uses multiple conventional convolutional layers for feature extraction and concatenates extracted features as the output. The main difference between the Inception module and a conventional convolutional layer is that the Inception module allows extraction of features with different kernel sizes. Therefore, the extracted features are not limited to the fixed-scale local regions. Intuitively, local regions of varying sizes are covered. With respect to the usefulness of the Inception module in lesion feature extraction, various kernel sizes help the model to generalize lesions of different sizes effectively. In terms of the architecture of the Inception module, there are three types of Inception modules: Inception A, Inception B and Inception C. Inception A (Fig. 3a) is equivalent to the inception module used in GoogLeNet (Szegedy et al. 2016); however, the 5 × 5 convolution is factored to two 3 × 3 convolutions. For Inception B (Fig. 3b), the 7 × 7

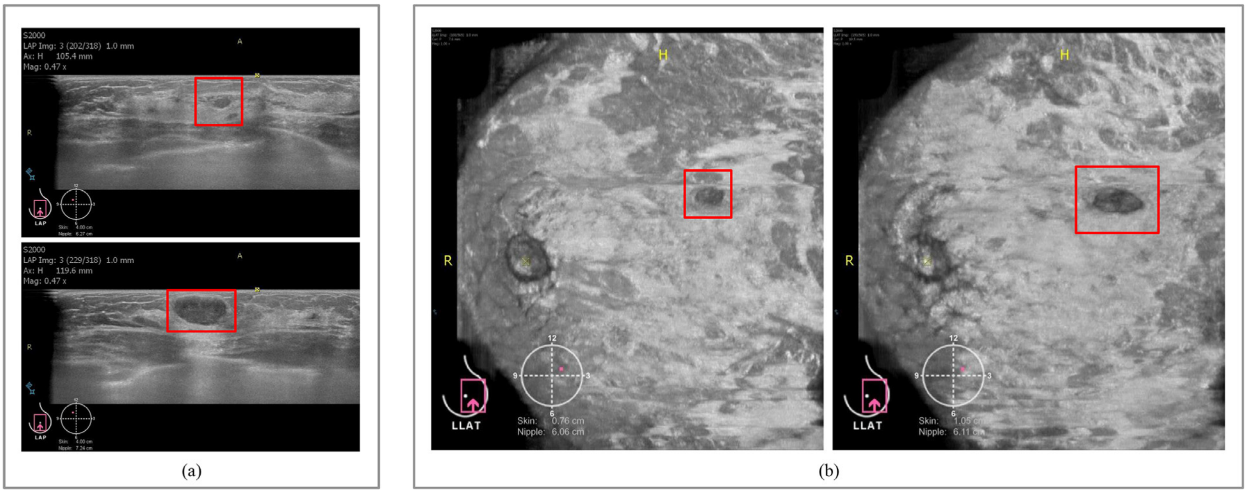

Fig. 2. Lesion patches around bounding boxes in *red*. (a) Two lesion patches obtained from two slices of the same benign lesion in transverse view. (b) Two lesion patches obtained from two slices of the same benign lesion in coronal view.

convolution is factored to two asymmetric convolutions with kernel sizes of 1 × 7 and 7 × 1.

With factorization, the total number of learnable parameters is reduced; on the other hand, it reduces the risk of overfitting. Inception C (Fig. 3c) adopts multiple kernels of different sizes to promote high-dimensional representations (Szegedy et al. 2016). The kernel sizes of Inception C are 1 × 1, 1 × 3, 3 × 1 and 3 × 3.

To retain the powerful feature extraction of Inception-v3 from natural images to ABUS images, we have adopted transfer learning, which has been widely applied in medical imaging analysis (Shin et al. 2016; Byra et al. 2018a; Xie et al. 2019). With transfer learning, the CNN model pre-trained from a very large scale database, such as ImageNet (Russakovsky et al. 2015), can efficiently apply the pre-trained knowledge to the specific task. By re-training the pre-trained model with a small amount of data, the retrained model can achieve a promising result on the specific task domain. In our case, all fully-connected (FC) layers proposed in Inception-v3 were redesigned, leaving the convolutional structure as the backbone for lesion feature extraction. The architecture of the backbone is illustrated in Figure 4a. The input size of the backbone is 299 × 299 × 3. The first several layers of the backbone consist of six convolutional layers with kernel sizes of 3 × 3 and an average pooling layer with a kernel size of 3 × 3, followed by five Inception A, four Inception B and two Inception C modules. The backbone outputs 2048 feature maps, and each feature map has a size of 8 × 8. A global average pooling layer is added on the top of the backbone's output feature maps, each of which is averaged to a single vector. The global average

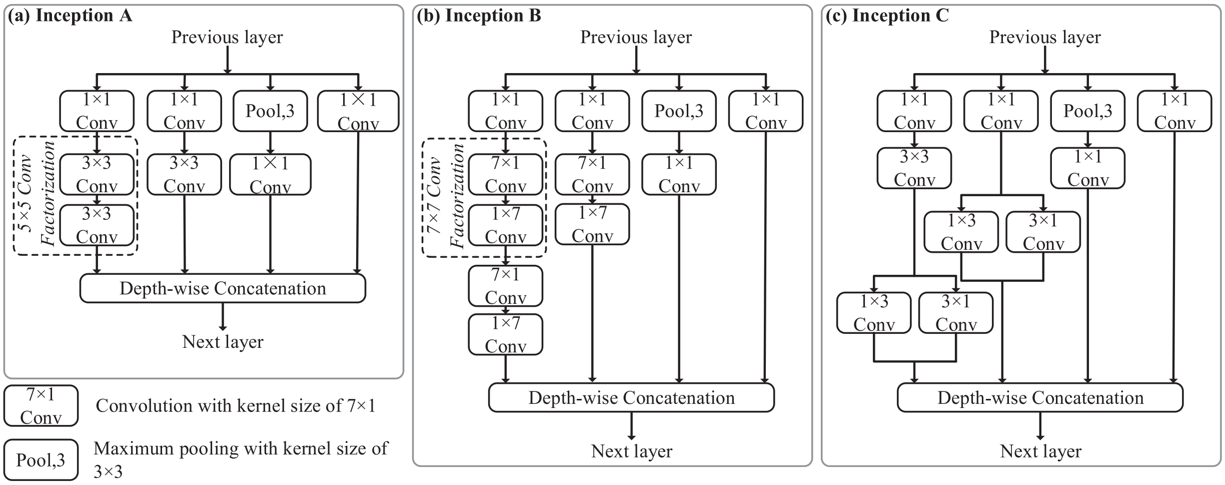

Fig. 3. Architectures of (a) Inception module A, (b) Inception module B and (c) Inception module C.

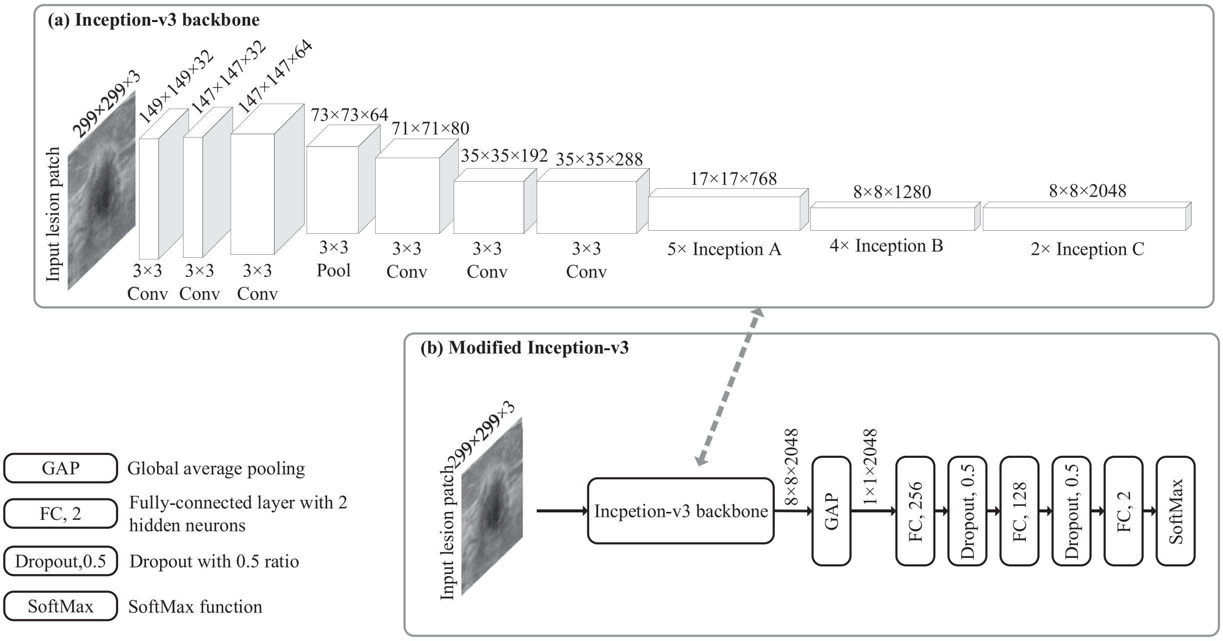

Fig. 4. Architectures of (a) Inception-v3 backbone and (b) modified Inception-v3 convolutional neural network (CNN).

pooling transforms the features that can be more robust for interpretation as categories (Lin et al. 2013). Then, three FC layers are added to bridge the convolutional features with a neural network classifier. The numbers of hidden neurons in the three FC layers are 256, 128 and 2. The dropout scheme is applied after the first and the second FC layers to relieve overfitting. Thereafter, the output of the last FC layer is normalized by the SoftMax function. The modified version of Inception-v3 is illustrated in Figure 4b. The modified version of Inception-v3 takes lesion patch as the input; then, the lesion features are extracted *via* the Inception-v3 backbone. Finally, it outputs the probabilities of malignant and benign.

*Multiview CNNs*

Two multiview CNNs, as illustrated in Figure 5, have been explored in this study. For the multiview CNN A (Fig. 5a), two lesion patches are cropped from the same lesion over transverse and coronal views independently. Then, the multiview lesion patch is generated by concatenating the cropped lesion patches into different image layers. For example, one breast lesion generates one lesion patch from the coronal view (CA) and two lesion patches from the transverse view (TA and TB). As a result, there are two multiview lesion patches with dual image layers combined, which are (CA-TA, CA-TB). By feeding the multiview lesion patch into the modified Inception-v3 model (Fig. 4b), the type of the lesion is classified. As the acquired ABUS images use grayscale reconstruction, one multiview lesion patch requires two image layers to encode both transverse and coronal views. However, the modified Inception-v3 model requires three input channels because the pre-trained model is trained by RGB images. To fulfill the input requirement of the modified Inception-v3 model, the lesion patch obtained from the transverse view is used as the third image layer. For each combination of dual image layers, all lesion patches obtained from the transverse view are used. In the aforementioned example, TA *and* TB are attached to generate two dual-layered multiview patches, and thus four three-layered multiview patches can be generated (CA-TA-TA, CA-TA-TB, CA-TB-TA, CA-TB-TB). Therefore, more effective multiview lesion patches are generated and used for training purposes. In total, we obtained 3085 multiview patches (1767 malignant and 1318 benign) that can be used to train the network.

Instead of extracting lesion features from the multiview lesion patch, the multiview CNN B (Fig. 5b) adopts two Inception-v3 backbones to enable multiview learning. Each Inception-v3 backbone corresponds to extract lesion features from a certain view. Then, the extracted features from the two Inception-v3 backbones are concatenated on top of the first FC layer. To match the input channels of the Inception-v3 backbone, the input lesion patch is transformed to three channels by duplicating the pixel value of the lesion patch. Thereafter, we obtained 1525 lesion samples (549 malignant and 468 benign) to train the multiview CNN B.

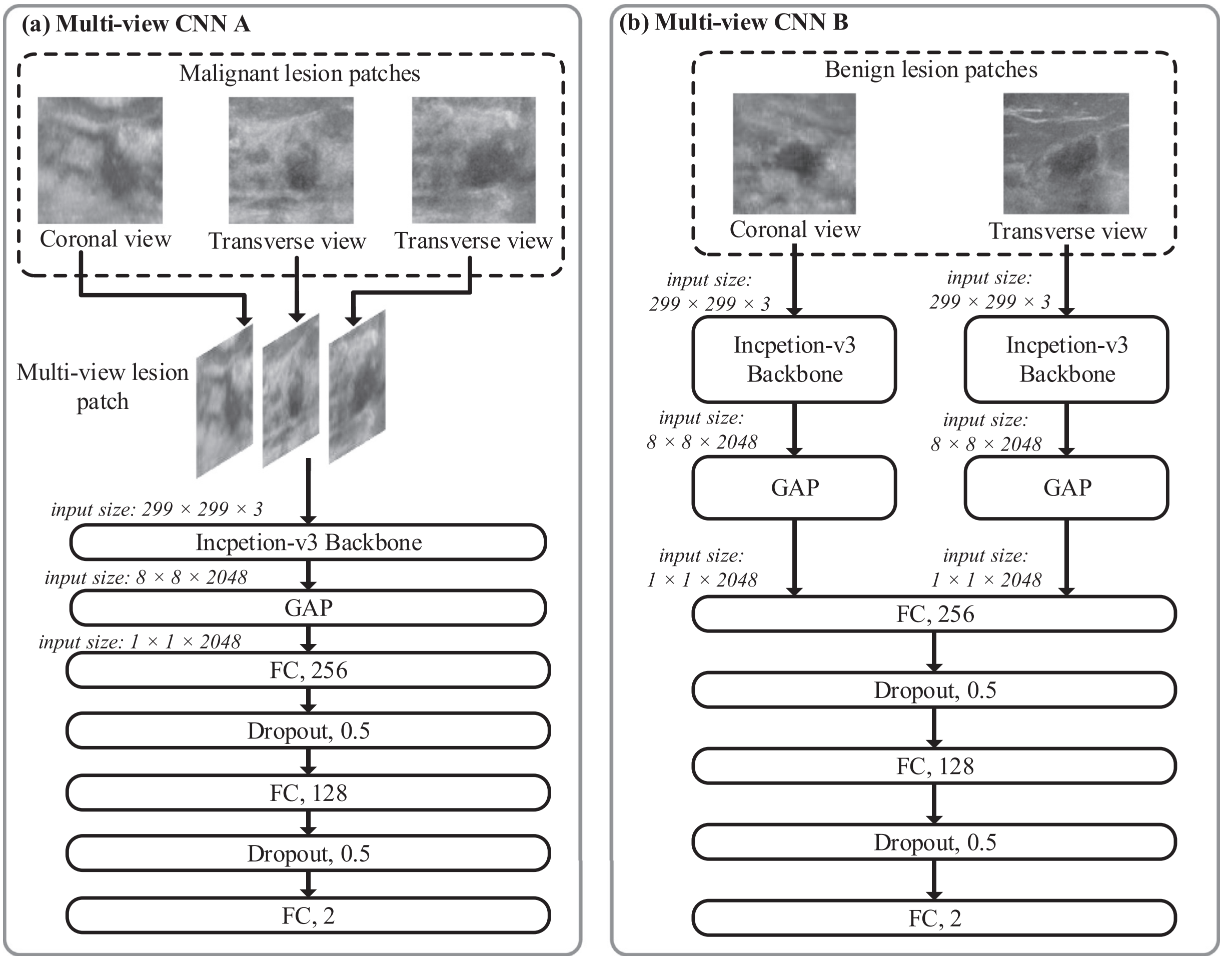

Fig. 5. Architectures of proposed multiview convolutional neural networks (CNNs). FC = fully connected layer; GAP = global average pooling.

*Network training and evaluation*

A five-folder cross-validation was used to train, validate and test the multiview CNNs. Specifically, each folder contains 20% of the breast lesions in our ABUS data set, including 31 malignant lesions and 33 benign lesions. In each round of cross-validation, one folder is reserved as the testing data set. The remaining four folders are split into training and validation data sets; three of four folders belong to the training data set and the remaining folder is used as the validation data set. After each epoch in training, the trained model is evaluated by the validation data set. To select the final trained model, an early stopping strategy is applied based on the classification performance on the validation data set. Practically, the final trained model is selected when the area under the curve (AUC) value does not increase for five epochs. Then, the final trained model is evaluated by the testing data set. Thereafter, the classification performance for each round of cross-validation is obtained. In this study, the reported results were obtained by averaging the classification performance across five rounds of cross-validation, including sensitivities, specificities and AUC values.

To enhance the learning process within limited training samples, the training data set is augmented. For multiview CNN A, the multiview lesion patches used for training are rotated three times with rotation angles of 90˚, 180˚ and 270˚, followed by flipping operation horizontally and vertically. For multiview CNN B, the same augmentation method is applied to the lesion patches obtained from transverse and coronal views.

The weights of the FC layers in multiview CNNs are initialized by following the Xavier uniform initializer (Glorot and Bengio 2010). In addition, the weights of the Inception-v3 backbone are initialized by applying pre-trained weights optimized for ImageNet database. For transfer learning, we followed the approach of Azizpour et al. (2015), in which the entire layers undergo fine-tuning during the training phase. Furthermore, the multiview CNNs are trained with a batch size of 32. The

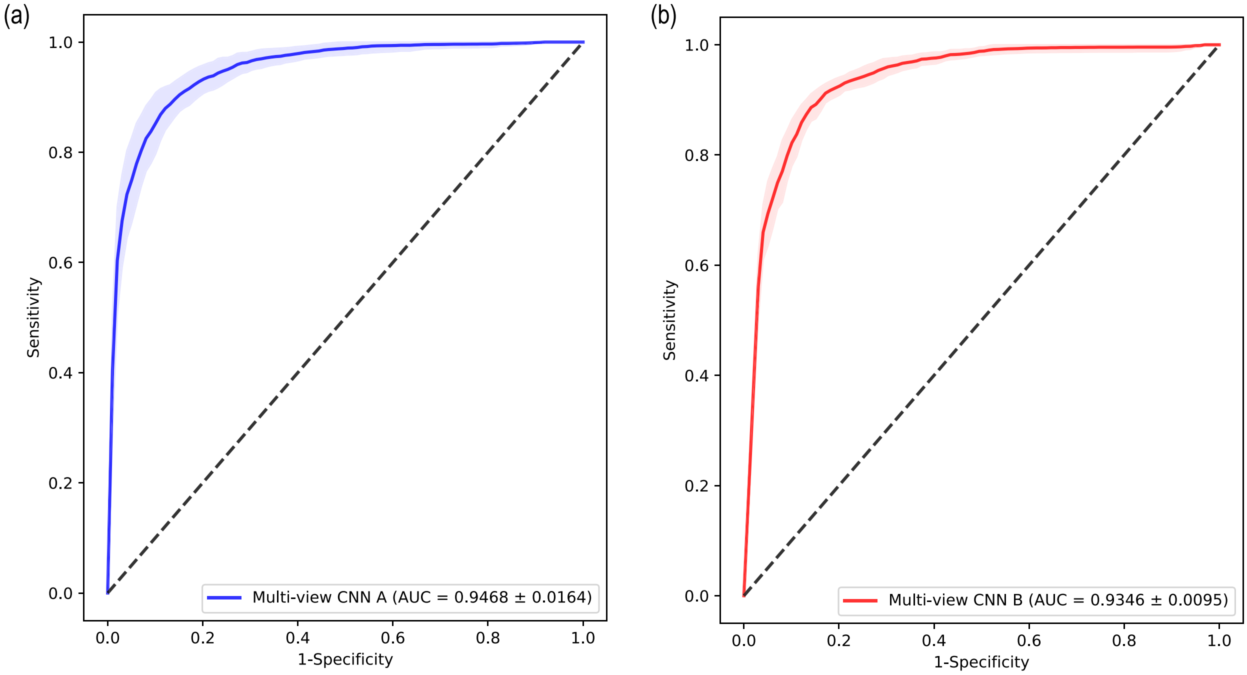

Fig. 6. (a) Mean receiver operating characteristic (ROC) curve of multiview convolutional neural network (CNN A) (area under the ROC curve [AUC] = 0.9468 with standard deviation of 0.0164). (b) Mean ROC curve of multiview CNN B (AUC value = 0.9346 with standard deviation of 0.0095). The shadow of each ROC curve illustrates the variance of the ROC during five-folder cross-validation.

losses are optimized by Adadelta with an adaptive learning rate (Zeiler 2012). The CNNs were implemented with Keras and were trained by using an Nvidia P6000 GPU on an Ubuntu 18.04 system.

## RESULTS

### *Classification performance of multiview CNNs*

To select the effective multiview strategies as described under Multiview CNNs, the classification performance of the multiview CNNs with two different configurations are compared. Multiview CNN A achieved a sensitivity of 0.886, specificity of 0.876 and mean AUC value of 0.9468 with a standard deviation of 0.0164. Compared with multiview CNN A, multiview CNN B had roughly 2% reduced sensitivity (0.865) and specificity (0.848), and the mean AUC value dropped to 0.9346 with a standard deviation of 0.0095. The mean receiver operating characteristic curves for the multiview CNNs are illustrated in Figure 6.

The classification performance of the multiview CNNs were compared with that of the single-view CNNs. The single-view CNNs followed the architecture of the multiview CNN A; however, the inputs of the single-view CNNs take the lesion patches obtained from either a transverse view or a coronal view. To match the input size of the modified Inception-v3, the lesion patches were encoded by duplicating the gray-scale intensities. Table 2 summarized the classification performance of multiview and single-view CNNs. The single-

Table 2. Comparison of multiview and single-view CNNs

| Single view | | Multiview | | | | |
|---|---|---|---|---|---|---|
| Coronal | Transverse | A | B | Sensitivity | Specificity | Area under curve* |
| ✓ | | | | 0.831 | 0.800 | 0.8874 ± 0.0054 |
| | ✓ | | | 0.832 | 0.838 | 0.9076 ± 0.0233 |
| | | ✓ | | **0.886** | **0.876** | **0.9468 ± 0.0164** |
| | | | ✓ | 0.865 | 0.848 | 0.9346 ± 0.0095 |

CNN = convolutional neural network.
Significance of bold values indicate the best performance among all compared methods.
* Mean ± standard deviation.

Table 3. Classification performance of multiview CNN A using different backbones

| Backbone | Pre-train weight | Sensitivity | Specificity | Area under curve* |
|---|---|---|---|---|
| ResNet | ✓ | 0.809 | 0.830 | 0.9045 ± 0.0153 |
| DenseNet | ✓ | 0.847 | 0.848 | 0.9221 ± 0.0206 |
| Inception-v4 | ✓ | 0.805 | 0.823 | 0.8851 ± 0.0096 |
| Inception-ResNet-v2 | ✓ | 0.866 | 0.873 | 0.9303 ± 0.0156 |
| Inception-v3 | ✘ | 0.822 | 0.859 | 0.9043 ± 0.0173 |
| Inception-v3 | ✓ | **0.886** | **0.876** | **0.9468 ± 0.0164** |

CNN = convolutional neural network.
Significance of bold values indicate the best performance among all compared methods.
* Mean ± standard deviation.

view CNN using the coronal view achieved a sensitivity of 0.831, a specificity of 0.800 and a mean AUC value of 0.8874 with a standard deviation of 0.0054. By employing the transverse view only, the single-view CNN's classification performance increased (sensitivity = 0.831, specificity = 0.800 and mean AUC = 0.8874 ± 0.0233) compared with that using the coronal view, but it still could not reach the classification performance of the multiview CNNs.

Furthermore, the effectiveness of lesion feature extraction is justified by adopting different backbones. Specifically, the proposed backbone was replaced with the convolution structures of ResNet (He et al. 2016), DenseNet (Huang et al. 2017a), Inception-v4 (Szegedy et al. 2017) and Inception-ResNet-v2 (Szegedy et al. 2017). Table 3 summarizes the classification performance of multiview CNN A employing different backbones. Based on the results, the multiview CNN A with the Inception-v3 backbone outperformed those with ResNet, DenseNet, Inception-v4 or Inception-ResNet-v2 backbones. In addition, the effectiveness of the Inception-v3 backbone with the ImageNet pre-train weights was explored by training the multiview CNN A from scratch. To train the model from scratch, the weights of all convolutional layers were initialized by following the Xavier uniform initializer, and the biases were initialized to zeros. As outlined in Table 3, the multiview CNN A trained from scratch degraded classification performance compared with that using ImageNet pre-train weights.

### *Comparison with conventional machine learning feature extractors*

In this section, the classification performance of multiview CNN A is compared with that of conventional machine learning feature extractors, including PCA and HOG, which both revealed the usefulness of lesion feature extractions in US imaging (Huang et al. 2005; Moura and Guevara Lopez 2013). To set up the experiment for PCA, we followed the pre-processing steps as described in Huang et al. (2005). First, the multiview lesion patches were re-sized to 128 × 128 × 3; this was followed by histogram equalization for contrast enhancement. Furthermore, we added an additional pre-processing step, local contrast normalization, as described in Arevalo et al. (2016). Such normalization can improve the learning process of the classifier. After the pre-processing steps, PCA was applied to extract the lesion features. An exhaustive search was performed to find the optimal number of principal components. By following the same pre-processing steps used for PCA, the lesion features were also extracted by employing HOG. The optimal parameters of HOG were obtained through an exhaustive search on different numbers of orientation bins and pixels per block as suggested in Moura and Guevara Lopez (2013), which are(8, 16) and (3 × 3, 5 × 5), respectively. Thereafter, we used a SVM with a radial basis function kernel to classify extracted lesion features. To obtain optimal classification performance, the penalty parameter (*C*) was determined by performing an exhaustive search in (0.001, 0.01, 0.1, 1, 10, 100,

Table 4. Comparison of the multiview CNN A with conventional machine learning approaches

| Feature extractor | Classifier | Sensitivity | Specificity | Area under curve* |
|---|---|---|---|---|
| PCA | SVM | 0.719 | 0.764 | 0.8195 ± 0.0644 |
| HOG | SVM | 0.782 | 0.771 | 0.8537 ± 0.0201 |
| Multi-view CNN A | | **0.886** | **0.876** | **0.9468 ± 0.0164** |

CNN = convolutional neural network; PCA = principal component analysis; HOG = histogram of oriented gradients; SVM = support vector machine.
Significance of bold values indicate the best performance among all compared methods.
* Mean ± standard deviation.

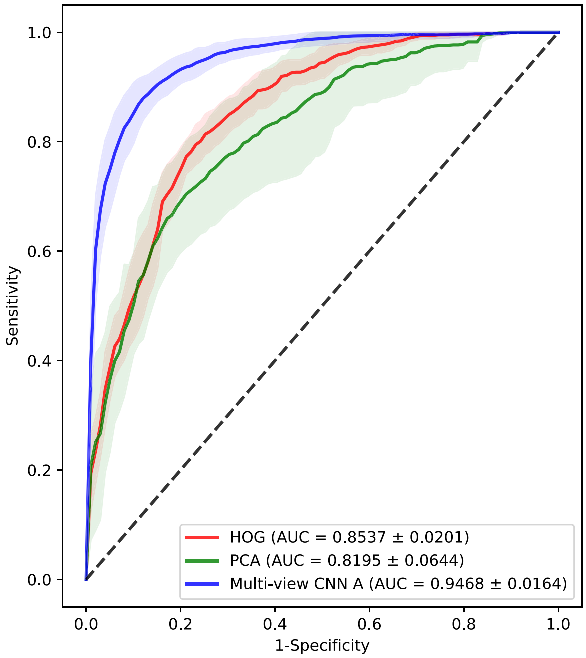

Fig. 7. Mean receiver operating characteristic (ROC) curves of histogram of oriented gradients (HOG) with support vector machine (SVM), principal component analysis (PCA) with SVM and multiview convolutional neural network (CNN) A. The shadow of each ROC curve illustrates the variance of the ROC curve during five-folder cross-validation.

1000), and the possible parameters for the kernel coefficient were (0.0001, 0.001, 0.1, 1 ,10). The SVM was trained and evaluated by five-folder cross-validation with the same data partitioning used for multiview CNN A. The PCA and SVM were implemented with Scikit-learn (Pedregosa et al. 2011), and HOG was implemented with Scipy (Jones et al. 2001).

As outlined in Table 4, the PCA with SVM had inferior classification performance, a sensitivity of 0.719, a specificity of 0.764 and a mean AUC value of 0.8195 with a standard deviation of 0.0644. HOG outperformed PCA (sensitivity = 0.782, specificity = 0.771, mean AUC = 0.8537 ± 0.0201), but still could not reach the classification performance of the multiview CNN. Figure 7 illustrates the mean ROCs of the multiview CNN, PCA and HOG approaches.

*Observer performance test*

To justify the usefulness of the proposed multiview CNNs, an observer performance test was conducted by comparing the diagnostic performance of human reviewers before and after referring to the predicting outcomes of the multiview CNN A. There were five human reviewers, including one breast special radiologist, two senior radiology residents, one first-year radiology resident and one physician. It should be noted that the reviewers who participated in the observed performance test did not annotate the ground truth of each lesion. Without having the patients' information, the reviewers independently reviewed the ABUS images used in this study. Each lesion was interpreted by reporting the malignancy rating on a 7-point scale (1 = definitely not malignant, 2 = almost definitely not malignant, 3 = probably not malignant, 4 = may be malignant, 5 = probably malignant, 6 = almost definitely malignant and 7 = definitely malignant). After the first round of interpretation, the reviewers were able to change their previous decisions by considering predicted outcomes of the multiview CNN A. Table 5 summarizes the changes in diagnostic performance of the five human reviewers from round 1 to round 2 of the observer performance test. With the aid of the multiview CNN A, all human reviewers improved in diagnostic accuracy (mean accuracy improvement: 0.102; range: −0.083 to 0.170). Also, all human reviewers' diagnoses had increased sensitivity (mean sensitivity improvement: 0.052; range: −0.015 to 0.008). The physician, a non-radiologist, achieved significant improvement in diagnostic specificity (from 0.403−0.851). In contrast, the proposed multiview CNN was found to be less advantageous in improving the diagnostic specificity of the radiologists (mean specificity improvement: 0.015; range: −0.021 to 0.040). Practically, the special radiologist had a reduced diagnostic specificity

Table 5. Results of observer performance test

| | Accuracy | | Sensitivity | | Specificity | |
|---|---|---|---|---|---|---|
| | Round 1* | Round 2† | Round 1 | Round 2 | Round 1 | Round 2 |
| Special radiologist | 0.864 | 0.883 | 0.881 | 0.933 | 0.851 | 0.845 |
| Senior resident 1 | 0.797 | 0.823 | 0.867 | 0.926 | 0.746 | 0.746 |
| Senior resident 2 | 0.842 | 0.900 | 0.874 | 0.926 | 0.818 | 0.873 |
| First-year resident | 0.816 | 0.848 | 0.881 | 0.941 | 0.768 | 0.780 |
| Physician | 0.604 | 0.876 | 0.874 | 0.911 | 0.403 | 0.851 |
| Multiview CNN A | 0.880 | | 0.886 | | 0.876 | |

* Round 1: independent interpretation.

† Round 2: interpretation with aid of the multiview CNN A.

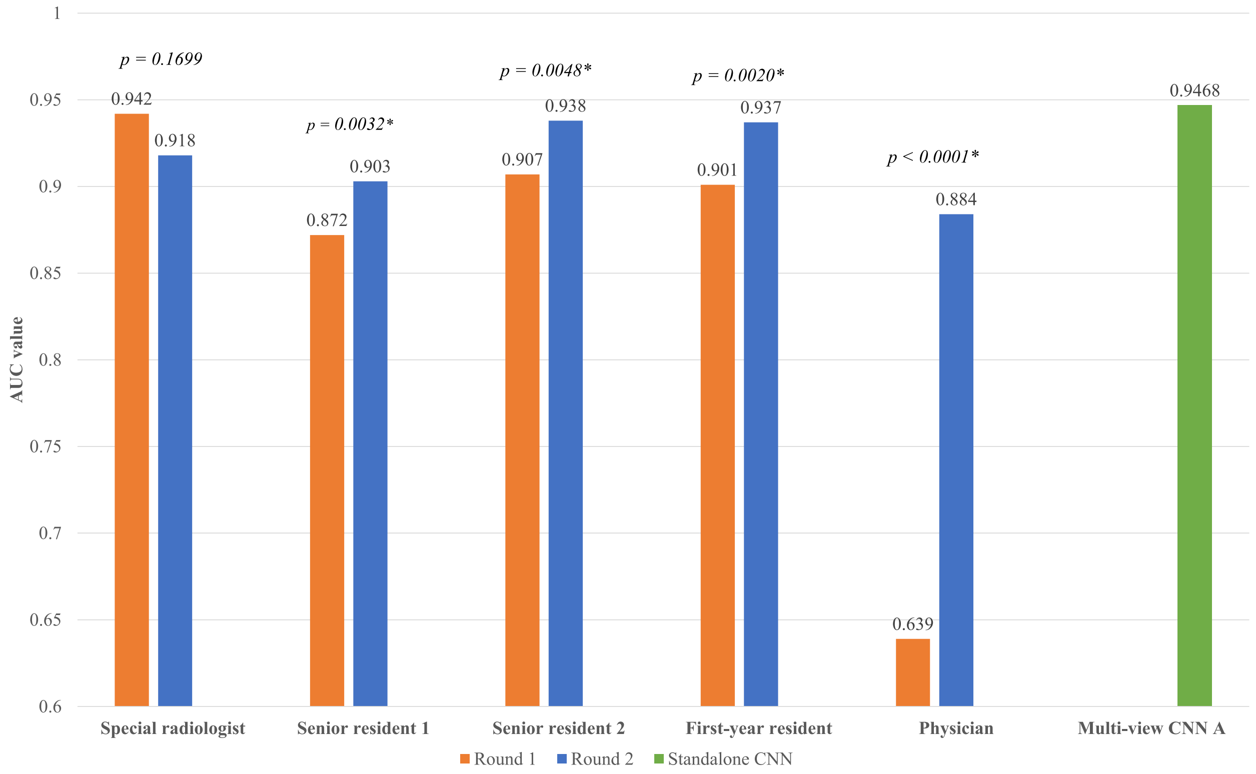

Fig. 8. Changes in the five human reviewers' areas under the curve (AUCs) in the observer performance test. Round 1: independent interpretation. Round 2: interpretation with aid of the multi-view convolutional neural network (CNN) A. *Significant difference in AUCs between rounds 1 and 2.

(from 0.851−0.845). Regardless of the support of the multiview CNN, senior resident 1 exhibited no difference in performance in terms of specificity.

The AUC values of the five human reviewers and the multiview CNN A are illustrated in Figure 8. AUC values of four of the five human reviewers improved from round 1 to round 2; the exception was the special radiologist. Specifically, the AUC value of the non-radiologist (physician) was significantly increased by 38% (AUC improvement: 0.245). To evaluate the statistical significance of the changes in AUC changes for each reviewer, we performed the Delong test under 1000 bootstrap samples. For the special radiologist, the AUC value decreased by 0.024 from round 1 to round 2 was decreased; however, the reduced AUC value for the special radiologist was not statistically significant. In turn, the remaining four human reviewers exhibited significant improvement statistically ($p < 0.05$) after referring to the classification results of the multiview CNN A.

## DISCUSSION

In this study, the proposed method employing a modified Inception-v3 CNN with multiview strategies achieved promising classification performance for discrimination of breast lesion patches between malignant and benign. We explored two different multiview CNN architectures (multiview CNNs A and B). Both multiview CNNs outperformed the single-view CNNs using either transverse or coronal lesion patches. The improvement over the single-view CNNs may have occurred because the multi-view strategies enable feature extraction over both transverse and coronal views, which increases the amount of extracted lesion features. Therefore, the features learned in the multiview CNNs had better discriminative power than those learned from the single-view CNNs.

### *Analysis of multiview CNNs*

It should be noted that the concept of the multiview CNN B was first used for mammography (Geras et al. 2017), and it also exhibited great discriminating performance in ABUS imaging based on our experimental results. However, the multiview CNN A would be preferred for two reasons. First, the multiview strategy applied in multiview CNN B requires two backbones; thus, the increased model complexity is not easily fine-tuned. Second, the inference time of multiview CNN A was faster than that of multi-view CNN B (34.34 ms/lesion vs. 73.79 ms/lesion), which indicated computational efficiency.

The Inception-v3 backbone, using Inception modules for convolution, exhibited improved lesion feature extraction compared with ResNet and DenseNet, which both employ conventional convolutional layers. Our experimental results confirmed that the Inception modules could extract features from lesions of different sizes; however, the conventional convolutional layer extracted fixed-scale features regardless of the size of the lesions. Compared with Inception-based CNNs, the proposed multiview CNN A employing the Inception-v3 backbone outperformed the proposed one using the Inception-v4 or Inception-ResNet-v2 backbone. The effectiveness of the Inception-v3 backbone justifies two aspects: (i) The main difference from Inception-v3 to Inception-v4 or Inception-ResNet-v2 is that Inception-v3 consists of fewer Inception modules. However, based on our experimental results, adding more Inception modules has no benefit in improving classification performance in ABUS imaging. (ii) When CNN goes deeper, adding more Inception modules, the performance of the deeper CNN could be affected by the vanishing gradient. To avoid the vanishing gradient, Inception-ResNet-v2 applies a residual connection from the input to the output of each Inception module. In contrast, there is no residual connection in Inception-v4, which could explain why the vanishing gradient may cause further performance degradation.

*Comparison with previous works*

With five-folder cross-validation of 316 breast lesions, the proposed method employing multiview CNN A achieved a mean AUC value of 0.9468. In previous studies using conventional machine learning feature extraction schemes, Moon et al. (2011) evaluated 147 breast lesions and reported the best AUC value of 0.9466, and Tan et al. (2012) reported the best AUC value of 0.93 based on 201 breast lesions. However, we cannot make a direct comparison. Because we use a different private database, it is impossible to estimate the complexity of each database based on the number of lesions. Nevertheless, the previous studies could be less generic than the proposed method because the performance of the previous methods could be affected by the quality of the input data, in which the lesion has to be segmented precisely. In contrast, the proposed method takes a raw lesion patch as input.

As mentioned earlier, this is the first study using CNNs for breast cancer classification in ABUS imaging. For US imaging, some previous studies explored the usefulness of CNNs. One previous study (Byra 2018) used pre-trained VGG and obtained an AUC value of 0.847 in evaluating 100 breast lesions. In a previous method used on 251 breast lesions (Byra et al. 2018b), the AUC value was 0.857 by employing pre-trained Inception-v3.

Because the database is different, we still cannot make a direct comparison. However, considering the AUCs reported from the two previous studies, the proposed method using the multiview strategy could perform better without sacrificing efficiency of computation. Because the multiview strategy was applied to the input of the CNN, there was no additional cost of computation compared with previous studies. Furthermore, one previous study adopted a multiview strategy to classify 829 breast lesions and achieved a best AUC of 0.9601 (Han et al. 2017). Similar to the multiview CNN A, the multiview patches were generated by combining multiple lesion patches with different scales. However, to generate these patches, radiologists must measure the distance between the boundary of the lesion and the boundary of the patch itself. When the amount of cases is large, this is a significant time-consuming task. In contrast, the proposed method directly crops the lesion patch from ABUS imaging; thus, it is expected to reduce the workflow of the radiologist. On the other hand, on comparison of AUCs, the best AUC value we obtained was 0.9704, which is similar to that in the previous study. Moreover, similar to the multiview CNN B, a previous study by Xiao et al. (2018) combined lesion features extracted from three different CNNs, including an Inception-v3, an Xception and a ResNet. By evaluating 2058 breast lesions, the previous method achieved the best AUC of 0.93. Nevertheless, using multiple CNNs for feature extraction may produce redundant features as all CNNs used the same lesion patch as the input. In turn, the proposed method generates more features by performing feature extraction over lesion patches of different views.

*Analysis of observer performance test*

In our observer performance test, the results confirmed that human reviewer increases diagnostic accuracy after referring to the proposed CNN. In addition, our results suggest that the proposed CNN is more effective in improving the diagnostic performance of reviewer compared with a reviewer who has more experience in diagnostic ABUS imaging. In the real clinical situation, follow-up medical examination or treatment is required before interpretation. Therefore, the proposed CNN could be used as a second reviewer to shorten the follow-up interval. Furthermore, our results confirmed that the proposed method could help reviewers make the correct diagnostic decision when there is a difference between transverse and coronal views. For example, as illustrated in Figure 9a, four of the five reviewers decided to interpret the malignant lesion as benign because the lesion shape on the transverse view appeared benign, and reviewers can see the suspicious shape of the lesion on the coronal view only. Another example is illustrated in Figure 9b. A benign lesion was misclassified by four of the five reviewers because the lesion shape on the transverse view was a suspicious finding, and the lesion shape appeared benign on the coronal view. In fact, ABUS stores the volume data digitally and

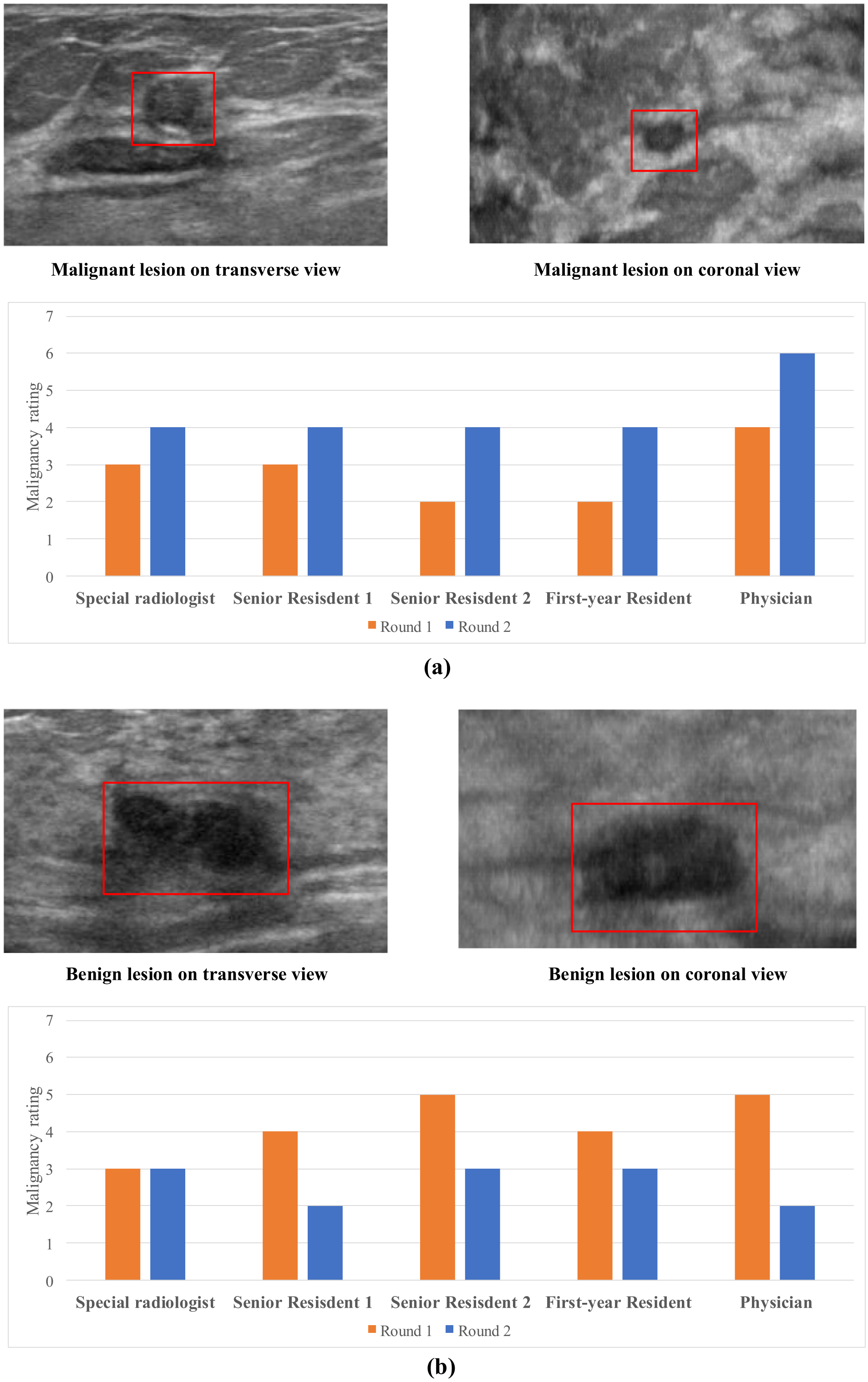

Fig. 9. Samples of changes in decisions of the five human reviewers. Round 1: independent interpretation. Round 2: interpretation with aid of the multiview convolutional neural network (CNN) A. (a) A malignant lesion was classified as malignant by the multiview CNN A with a malignancy rating of 7. (b) A benign lesion was classified as benign by the multiview CNN A with a malignancy rating of 1.

reconstructs multiplanar images, which allows radiologists to simultaneously evaluate breast lesions using reconstructed coronal planes as well as transverse planes. Previous studies have reported that retraction phenomenon, an especially useful feature of coronal planes, had high diagnostic accuracy for breast malignancy (Chen et al. 2013; Zheng et al. 2015). However, the reviewers in this study were less familiar with coronal planes than transverse planes because ABUS has become popular only recently. In contrast, the proposed CNN evaluated breast lesions with transverse and coronal planes equally.

## CONCLUSIONS

In this work, a CNN is proposed to differentiate breast lesions between malignant and benign. To the best of our knowledge, this is the first study employing the deep learning approach to realize breast cancer classification in ABUS imaging. In addition, we implemented multiview strategies, which allow conventional CNNs to learn more lesion features from different image views of the ABUS imaging simultaneously. Furthermore, the proposed CNN takes a raw image as input and learns the image features directly; it does not require design of a series of manual processing steps such as segmenting breast lesion and selecting an optimal combination of features. With multiview strategies, our proposed CNN can effectively classify breast cancer with a sensitivity of 0.886 and specificity of 0.876. Compared with conventional machine learning methods, our proposed CNN is more accurate and robust, and has a >10% increased AUC. On the other hand, the clinical benefits of the proposed CNN are confirmed by our observer performance test. Our results suggest that the proposed CNN has advantages as a second reviewer to support diagnostic decisions of human reviewers and provide opportunities to compare decisions. With the aid of the proposed CNN, our results suggest that the diagnostic performance of human reviewers is improved. The proposed CNN can be integrated into existing CADx systems to provide effective lesion feature extraction and robust breast cancer classification. In future work, it would be worthwhile to evaluate our present work with dedicated detection algorithms for breast lesion detection and classification as an end-to-end CAD solution.

*Acknowledgments*—The authors thank the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Department of Electrical and Computer Engineering at the University of Saskatchewan for their financial support for this research work. This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) and the Ministry of Health & Welfare, Republic of Korea (Grant HI18 C2383). This paper was also supported by the Fund of Biomedical Research Institute, JNUH.

*Conflict of interest disclosure*—The authors declare that they have no conflicts of interest.

## REFERENCES

Ahmed Medjahed S, Ait Saadi T, Benyettou A. Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules. Int J Comput Appl 2013;62:1–5.

Arevalo J, Gonzalez FA, Ramos-Pollan R, Oliveira JL, Guevara Lopez MA. Representation learning for mammography mass lesion classification with convolutional neural networks. Comput Methods Programs Biomed 2016;127:248–257.

Azizpour H, Razavian AS, Sullivan J, Maki A, Carlsson S. From generic to specific deep representations for visual recognition. In: Proceedings, 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). vol. 10, Piscataway, NJ: IEEE; 2015. p. 36–45.

Brem RF, Lenihan MJ, Lieberman J, Torrente J. Screening breast ultrasound: Past, present, and future. AJR Am J Roentgenol 2015;204:234–240.

Byra M. Discriminant analysis of neural style representations for breast lesion classification in ultrasound. Biocybernet Biomed Eng 2018;38:684–690.

Byra M, Styczynski G, Szmigielski C, Kalinowski P, Michalowski L, Paluszkiewicz R, Ziarkiewicz-Wroblewska B, Zieniewicz K, Sobieraj P, Nowicki A. Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. Int J Comput Assist Radiol Surg 2018a;13:1895–1903.

Byra M, Sznajder T, Korzinek D, Piotrzkowska-Wroblewska H, Dobruch-Sobczak K, Nowicki A, Marasek K. Impact of ultrasound image reconstruction method on breast lesion classification with neural transfer learning. Pattern recognition and pattern analysis. Cham: Springer International; 2018.

Chen L, Chen Y, Diao XH, Fang L, Pang Y, Cheng AQ, Li WP, Wang Y. Comparative study of automated breast 3-D ultrasound and handheld b-mode ultrasound for differentiation of benign and malignant breast masses. Ultrasound Med Biol 2013;39:1735–1742.

Cheng JZ, Chou YH, Huang CS, Chang YC, Tiu CM, Chen KW, Chen CM. Computer-aided US diagnosis of breast lesions by using cell-based contour grouping. Radiology 2010;255:746–754.

Chiang TC, Huang YS, Chen RT, Huang CS, Chang RF. Tumor detection in automated breast ultrasound using 3-D CNN and prioritized candidate aggregation. IEEE Trans Med Imaging 2019;38:240–249.

Geras KJ, Wolfson S, Shen Y, Wu N, Kim SG, Kim E, Heacock L, Parikh U, Moy L, Cho K. High-resolution breast cancer screening with multi-view deep convolutional neural networks. In: Proceedings, Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE; 2017.

Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings, Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-10). 9: JMLR Proc; 2010. p. 249–256.

Han S, Kang HK, Jeong JY, Park MH, Kim W, Bang WC, Seong YK. A deep learning framework for supporting the classification of breast lesions in ultrasound images. Phys Med Biol 2017;62:7714–7728.

He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 12, Piscataway, NJ: IEEE; 2016. p. 770–778.

Huang YL, Kuo SJ, Chang CS, Liu YK, Moon WK, Chen DR. Image retrieval with principal component analysis for breast cancer diagnosis on various ultrasonic systems. Ultrasound Obstet Gynecol 2005;26:558–566.

Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE; 2017a. p. 2261–2269.

Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF. SVM and SVM ensembles in breast cancer prediction. PLoS One 2017b;12: 1–14.

Iakovidis DK, Keramidas EG, Maroulis D. Fuzzy local binary patterns for ultrasound texture characterization. Lecture Notes Comput Sci (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2008;5112:750–759.

Jones E, Oliphant T, Peterson P. Others. SciPy: Open source scientific tools for Python.

Kolb TM, Lichy J, Newhouse JH. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: An analysis of 27,825 patient evaluations. Radiology 2007;225:165–175.

Lin M, Chen Q, Yan S. Network in network. CoRR 2013; abs/ 1312.4400:1−10.

Marcomini KD, Fleury EFC, Schiabel H, Nishikawa RM. Proposal of semiautomatic classification of breast lesions for strain sonoelastography using a dedicated CAD system. In: Tingberg A, Lang K, Timberg P, (eds). Breast imaging. Cham: Springer International; 2016. p. 454–460.

Moon WK, Shen YW, Huang CS, Chiang LR, Chang RF. Computer-aided diagnosis for the classification of breast masses in automated whole breast ultrasound images. Ultrasound Med Biol 2011;37:539–548.

Moura DC, Guevara López MA. An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. Int J Comput Assist Radiol Surg 2013;8:561–574.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825–2830.

Rajaguru H, Kumar Prabhakar S. Bayesian linear discriminant analysis for breast cancer classification. In: Proceedings, 2nd International Conference on Communication and Electronics Systems (ICCES). Piscataway, NJ: IEEE; 2017. p. 266–269.

Rella R, Belli P, Giuliani M, Bufi E, Carlino G, Rinaldi P, Manfredi R. Automated breast ultrasonography (ABUS) in the screening and diagnostic setting: Indications and practical use. Acad Radiol 2018;25:1457–1470.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet large scale visual recognition challenge. Int J Comput Vision 2015;115:211–252.

Sadeghi-Naini A, Suraweera H, Tran WT, Hadizad F, Bruni G, Rastegar RF, Curpen B, Czarnota GJ. Breast-lesion characterization using textural features of quantitative ultrasound parametric maps. Sci Rep 2017;7:13638.

Shin HJ, Kim HH, Cha JH. Current status of automated breast ultrasonography. Ultrasonography 2015;34:165–172.

Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM. Deep Convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016;35:1285–1298.

Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019;69:7–34.

Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. CoRR 2014;1–14.

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Van- houcke V, Rabinovich A. Going deeper with convolutions. In: Proceedings, IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE; 2015. p. 1–9.

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings, IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE; 2016. p. 2818–2826.

Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Proceedings, 3rd AAAI Conference on Artificial Intelligence (AAAI). CA, USA: AAAI Press; 2017. p. 4278–4284.

Tan T, Platel B, Huisman H, Sanchez CI, Mus R, Karssemeijer N. Computer-aided lesion diagnosis in automated 3-D breast ultrasound using coronal spiculation. IEEE Trans Med Imaging 2012;31:1034–1042.

Thigpen D, Kappler A, Brem R. The role of ultrasound in screening dense breasts: A review of the literature and practical solutions for implementation. Diagnostics (Basel) 2018;8(1).

Van Zelst J, Tan T, Platel B, de Jong M, Steenbakkers A, Mourits M, Griveg-nee A, Borelli C, Karssemeijer N, Mann RM. Improved cancer detection in automated breast ultrasound by radiologists using computer aided detection. Eur J Radiol 2017;89:54–59.

Wang Y, Jiang S, Wang H, Guo YH, Liu B, Hou Y, Cheng H, Tian J. CAD algorithms for solid breast masses discrimination: Evaluation of the accuracy and interobserver variability. Ultrasound Med Biol 2010;36:1273–1281.

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.

Xie J, Liu R, Luttrell J, Zhang C. Deep learning based analysis of histopathological images of breast cancer. Front Genet 2019;10:80.

Zeiler MD. ADADELTA: An adaptive learning rate method. CoRR 2012; abs/1212.5.

Zheng FY, Yan LX, Huang BJ, Xia HS, Wang X, Lu Q, Li CX, Wang WP. Comparison of retraction phenomenon and BI-RADS-US descriptors in differentiating benign and malignant breast masses using an automated breast volume scanner. Eur J Radiol 2015;84:2123–2129.