Automated detection of masses on whole breast volume ultrasound scanner: false positive reduction using deep convolutional neural network

Yuya Hiramatsu[a], Chisako Muramatsu*[a], Hironobu Kobayashi[b], Takeshi Hara[a], Hiroshi Fujita[a]
[a]Department of Intelligent Image Information, Graduate School of Medicine, Gifu University, 1-1 Yanagido, Gifu 501-1194, Japan
[b]Department of Breast & Endocrine Surgery, Nagoya Central Hospital, 3-7-7 Taiko, Nakamura-ku, Nagoya, Aichi 453-0801, Japan

ABSTRACT

Breast cancer screening with mammography and ultrasonography is expected to improve sensitivity compared with mammography alone, especially for women with dense breast. An automated breast volume scanner (ABVS) provides the operator-independent whole breast data which facilitate double reading and comparison with past exams, contralateral breast, and multimodality images. However, large volumetric data in screening practice increase radiologists' workload. Therefore, our goal is to develop a computer-aided detection scheme of breast masses in ABVS data for assisting radiologists' diagnosis and comparison with mammographic findings. In this study, false positive (FP) reduction scheme using deep convolutional neural network (DCNN) was investigated. For training DCNN, true positive and FP samples were obtained from the result of our initial mass detection scheme using the vector convergence filter. Regions of interest including the detected regions were extracted from the multiplanar reconstraction slices. We investigated methods to select effective FP samples for training the DCNN. Based on the free response receiver operating characteristic analysis, simple random sampling from the entire candidates was most effective in this study. Using DCNN, the number of FPs could be reduced by 60%, while retaining 90% of true masses. The result indicates the potential usefulness of DCNN for FP reduction in automated mass detection on ABVS images.

Keywords: Breast cancer, automated ultrasound breast volume scanner, computer-aided detection, deep convolutional neural network, false positive reduction

1. INTRODUCTION

Breast ultrasound is considered an effective screening exam in adjunct to mammography screening. Studies have shown improved sensitivities of cancer detection with combined mammography and ultrasonography compared with mammography alone, especially for young women and women with dense breasts [1, 2].

Conventional hand-held ultrasound systems have an advantage of being a real-time exam, enabling examiners to obtain additional information such as elasticity and blood flows. However, in the screening setting, the exam takes time and is operator dependent. An automatic breast volume scanner (ABVS), on the other hand, is generally operator independent, and the whole breast data facilitates double reading, longitudinal and contralateral comparisons, and comparison with the findings on mammography and breast MRI.

Despite these advantages, volumetric data with a large number of axial, sagittal and coronal slices increase radiologists' reading time. The positional difference from mammography, i.e., supine vs compressed cranio-caudal or mediolateral oblique positions, may further increase the reading time. Computer-aided detection of suspicious regions can assist radiologists in diagnosis and efficient reporting.

Lo et al. [3] proposed an automated mass detection scheme on automated whole breast ultrasound images. They applied watershed transform for segmenting the candidate regions. The likelihood of being tumors was estimated using the

*chisa@fjt.info.gifu-u.ac.jp; phone +81-58-230-6518; fax +81-58-230-6514

Medical Imaging 2017: Computer-Aided Diagnosis, edited by Samuel G. Armato III, Nicholas A. Petrick,
Proc. of SPIE Vol. 10134, 101342S · © 2017 SPIE · CCC code: 1605-7422/17/$18 · doi: 10.1117/12.2254581

quantitative morphology, intensity and texture features. Ye et al. [4] reported a mass detection algorithm that consists of three major steps: active contour based segmentation, feature extraction, and classification using a support vector machine. They found that multi-view features play an important role in reduction of false positives (FPs).

Our goal is to develop a computerized scheme for detection of breast masses in ABVS volumes. In our previous study, suspicious areas were detected on the basis of the convergence of gradient vectors. For detecting small lesions and low contrast lesions, a large number of FP regions were also detected. In this study, we investigated the use of a deep convolutional neural network (DCNN) for reducing the number of FPs.

2. METHODS

2.1 Overview of the initial mass detection scheme

The initial candidates of masses were detected on the basis of the concentration of gradient vectors. First, images were preprocessed by applying median filter, hysteresis smoothing, and anisotropic diffusion filter for decreasing image noise. Subsequently, three dimensional (3D) vector convergence filter [5] was applied to emphasize the voxel with strong concentration of incoming vectors. After applying an empirical threshold, region growing based on the mean pixel value was performed. For the stopping criterion, 3D vector convergence filter was applied to the binary output. If the filter output of the centroid of the grown region was above the prespecified value, the region growing process was repeated. Figure 1 shows an example of filter output and the candidate regions. The detected region was considered as a true detection candidate, if the centroid of the candidate region is inside the gold standard region or the candidate region is overlapped with the gold standard region by more than 30%. The voxels of interest (VOIs) including these candidates were used for training and testing DCNN for FP reduction.

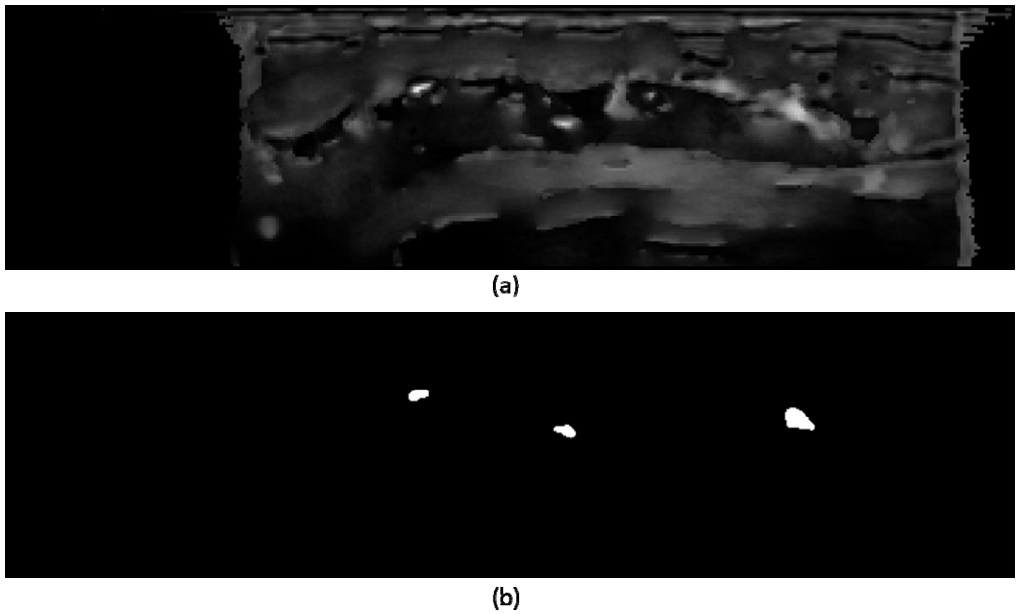

Figure 1. Initial mass detection. (a) Vector convergence filter output, and (b) mass candidates by thresholding and region growing.

2.2 Network architecture and data preparation

In this study, AlexNet network architecture [6], which includes 5 sets of convolution layers, 3 max pooling layers, 2 fully connected layers with dropout, and a fully connected layer with softmax output, was employed. As input data, multiplanar reconstructed slices were prepared by rotating the plane in axial-sagittal direction with respect to the centroid of a detected candidate region. For non-mass training data, 9 slices by step of 20 degrees were obtained from an FP

candidate. To increase positive data, 36 slices by step of 5 degrees were obtained from each true positive (TP) candidate. Such an image set was created by shifting the rotation axis to the previous and next axial slices.

The sizes of square regions of interest (ROIs) were set to 2 times the largest diameters of the candidate regions. As AlexNet takes input data of 256 x 256 pixels, ROIs were resized to 256 x 256 pixels which were then cropped to 227 x 227 pixels before input to the first layer for reducing the overtraining effect.

2.3 Training data sampling

The usefulness of DCNN was evaluated by a 2-fold cross validation. First, cases were divided into two groups semi-randomly so that the numbers of TP candidates in sets A and B were equivalent. The initial candidate regions included a large number of FPs. For balancing the numbers of positive and negative samples for training, FP samples were selected. In this study, three sampling methods were investigated: (1) random sampling, (2) high FP score sampling, and (3) low FP score sampling.

In method (1), when network was trained using set A, all TP images and randomly sampled FP images in set A were employed. The trained model was tested using set B. The process was repeated by switching the set A and set B.

In method (2), each set was again divided into two groups as set A1, set A2, set B1, and set B2. First, using set A1, network was trained with all TP samples and randomly selected FP samples. Subsequently, the FP samples in set A2 were input to the trained model A1 to obtain scores. Based on the scores, most FP-like samples, i.e., those from the lowest TP scores or the highest FP scores, were selected as training set A2-s. Similarly, the network was trained with all TP samples and randomly selected FP samples in set A2. After the training, the FP samples in set A1 were tested by the model A2. As a result, most FP-like samples were selected as training set A1-s. Finally, network was retrained using two selected sets A1-s and A2-s and all TP samples in combined set A. The trained model was tested using set B, and the whole process was repeated by swapping the sets. The procedure is illustrated in Fig. 2.

In method (3), the procedure of method (2) was repeated only in the way that most TP-like samples with the highest TP scores were selected.

2.4 Evaluation

For testing, 9 multiplanar reconstruction slices with an interval of 20 degrees for all candidates in test set were input to the trained network. The result was evaluated using a free response receiver operating characteristic (FROC) curve. Based on the 9 outputs, FROC curves were determined using the average, maximum, and minimum values as shown in Fig. 3. For the masses depicted in multiple views, the lesion was considered as TP, if it was detected in at least one view.

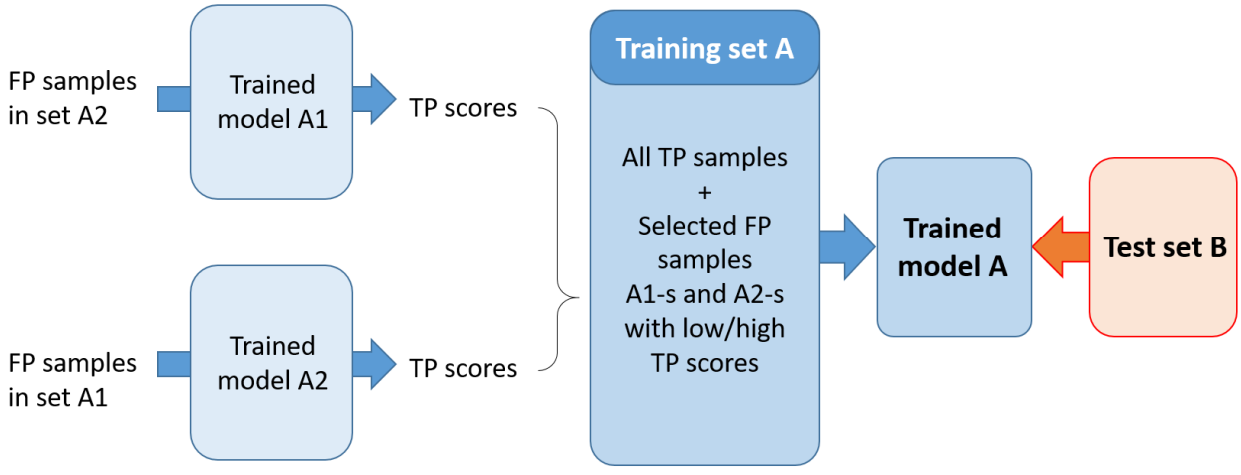

Figure 2. Procedure of selecting FP training samples in 2-fold cross validation

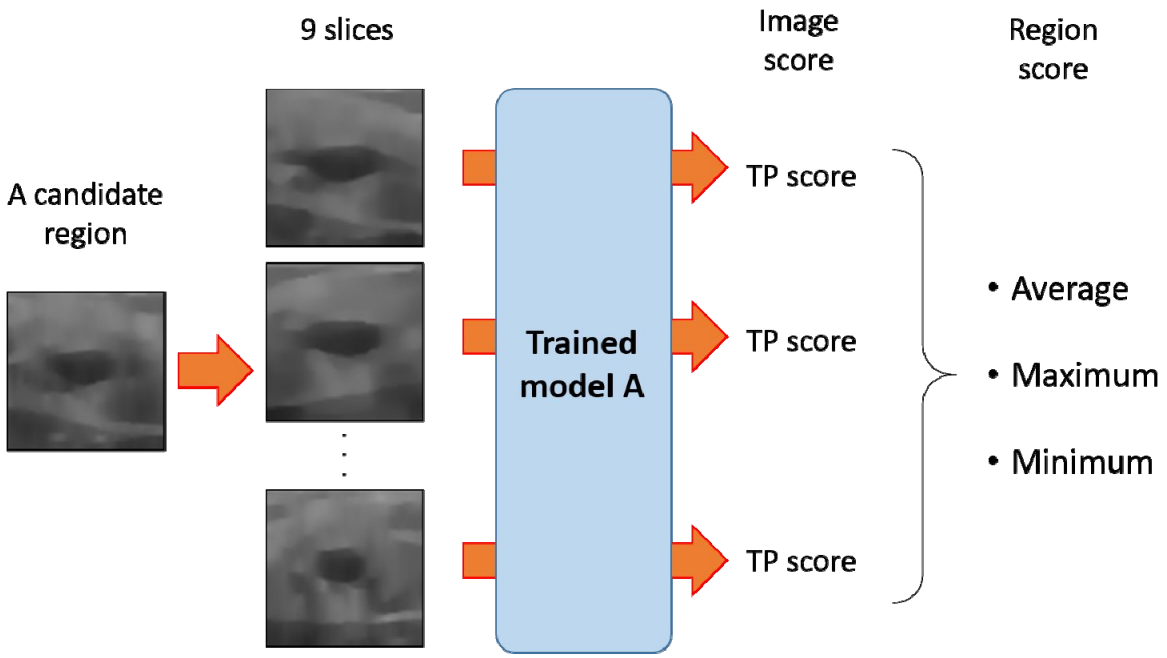

Figure 3. Test procedure and score summarization for a candidate region

3. MATERIALS

Volumetric breast ultrasound images were obtained by ACUSON S2000 ABVS (Siemens Healthcare) at Nagoya Central Hospital as a part of screening program. Breasts were scanned with a patient in supine position with soft compression. Generally three or four volumetric scans (views) per breast were obtained depending on the size of the breast. Additional views were obtained by ultrasound technologists if needed. The presences of breast masses are based on the diagnostic reports by breast surgeons with the certificate of breast ultrasound reading who, in general, read ABVS exams without mammograms.

The original volume data have different matrix sizes between different scans. The horizontal and vertical matrix sizes of axial slices range from 702 to 729 pixels and from 420 to 573 pixels, respectively. The pixel size in the horizontal direction is 0.21 mm, whereas the pixel size in the vertical direction ranges from 0.07 to 0.11 mm. The number of slices is 318 for all scans with 0.53 mm slice thickness. These volumes were converted to isotropic data with the voxel size of 0.21 mm by linear interpolation. Figure 4 shows the original axial slice and the converted isotropic slice.

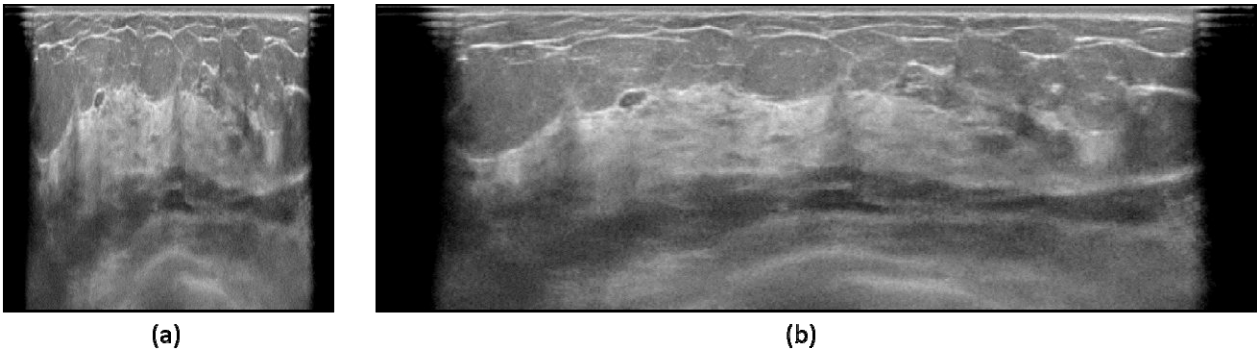

Figure 4. Image size normalization. (a) Original axial slice, and (b) converted isotropic slice.

In this study, 59 volumetric data including at least one mass were employed for training and testing of the proposed method. A total number of breast masses included in 59 scans was 90. When the same lesions on different views were counted individually, the number of TP regions is 106. These gold standard regions were used as training TP samples.

Using our initial mass detection and simple rule-based FP reduction schemes, 79 masses (88% sensitivity) were detected with 109 detected regions. These candidate regions were employed as the test TP samples. The number of detected FPs per view was 132, totaling 7781 FP candidate regions. These FP regions were employed as training and testing FP samples. After FP selection, the number of training regions in both sets include 53 TP and 212 FP regions. After slice sampling, the numbers of ROIs became 1908 for both TP and FP sets.

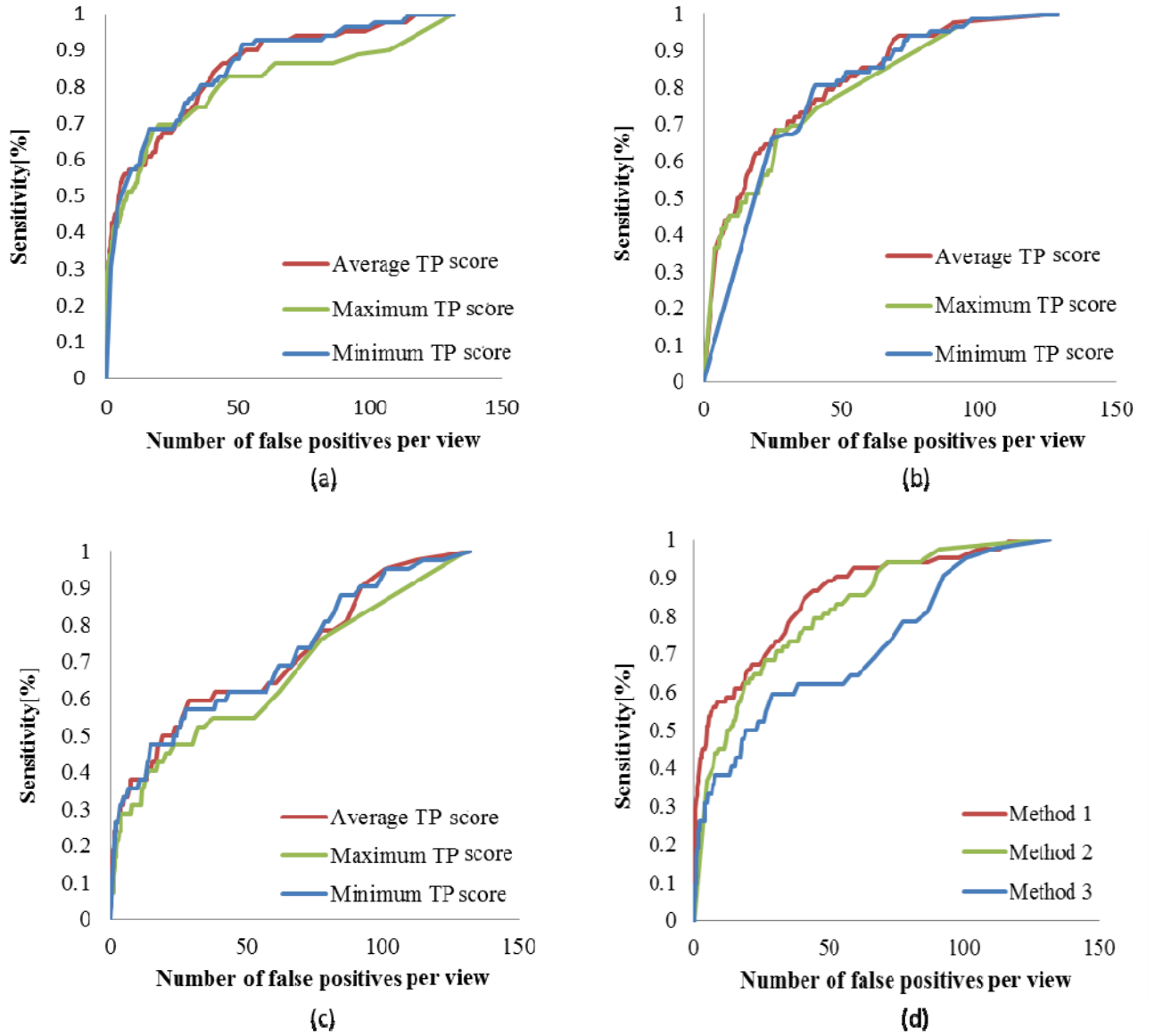

Figure 5. FROC curves for different FP sampling methods. Results for (a) method 1: random sampling, (b) method 2: high FP score sampling, (c) method 3: low FP score sampling, and (d) summary of 3 methods using average TP scores.

4. RESULT

The FROC curves for methods (1), (2) and (3) are shown in Figs. 5(a), 5(b), and 5(c), respectively. In all methods, the effect of output summarization was small whether by taking average, maximum or minimum scores, although the average TP score seems to provide the most stable results. The results of three FP selection methods using the average scores are shown in Fig. 5(d). Based on the results, the random sampling provided the best performance. The result indicates that training the network using samples with a large variation is more preferable than using difficult samples or relatively obvious samples. Using the DCNN, the number of FPs could be reduced by 60% while keeping 90% of TP candidates.

5. CONCLUSION

We have investigated the use of DCNN for FP reduction in automated detection of masses on ABVS images. In balancing the numbers of TP and FP samples for training, three methods for selecting the FP samples were tested. The result showed that random sampling provided better performance in terms of FROC curves than sampling FP-like samples or TP-like samples based on the TP scores. The proposed FP reduction method successfully reduced the number of FPs by 60% at the cost of 10% sensitivity reduction. The result indicates the potential utility of DCNN for FP reduction in automated mass detection on ABVS. Further optimization of network architecture and investigation of input data arrangement are needed in the future.

ACKNOWLEDGEMENTS

This study was partly supported by a research grant by the Koshiyama Science and Technology Foundation, Japan, a Grant-in-Aid for Scientific Research for Young Scientists (No. 26860399) by Japan Society for the Promotion of Science, and a Grant-in-Aid for Scientific Research on Innovative Areas (Multidisciplinary Computational Anatomy, No. 26108005) by Ministry of Education, Culture, Sports, Science and Technology, Japan.

REFERENCES

[1] Berg, W. A, Zhang, Z., Lehrer, D., Jong, R. A., Pisano, E. D., Barr, R. G., Bohm-Velez, M., Mahoney, M. C., Evans III, W. P., Larsen, L. H., Morton, M. J., Mendelson, E. B., Farria, D. M., Cormack, J. B., Marques, H. S., Adams, A., Yeh, N. M., Gabrielli, G. G., and ACRIN 6666 investigators, "Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk," JAMA 307, 1394-1404 (2012).

[2] Ohuchi, N., Suzuki, A., Sobue, T., Kawai, M., Yamamoto, S., Zhang, Y. F., Shiono, Y. M., Saito, H., Kuriyama, S., Tohno, E., Endo, T., Fukao, A., Tsuji, I., Yamaguchi, T., Ohashi, Y., Fukuda, M., Ishida, T., and J-START investigator groups, "Sensitivity and specificity of mammography and adjunctive ultrasonography to screen for breast cancer in the Japan Strategic Anti-cancer Randomized Trial (J-START); a randomized controlled trial," Lancet 387, 341-348 (2016).

[3] Lo, C. M., Chen, R. T., Chang, Y. C., Yang, Y. W., Hung, M. J., Huang, C. S., Chang, R. F., "Multi-dimensional tumar detection in automated whole breast ultrasound using topographic watershed," IEEE Trans Med Imaging 33, 1503-1511 (2014).

[4] Ye, C., Vaidya, V., Zhao, F., "Improved mass detection in 3D automated breast ultrasound using region based features and multi-view information," IEEE Eng Med Bio Soc 2014, 2865-2868 (2014).

[5] Okura, J., Uchiyama, Y., Yamauchi, M., Yokoyama R., Hara, T., Yamakawa, H., Ando, H., Iwama, T., Hoshi, H., and Fuita, H., "Computerized detection of aneurysms in MRA images based on gradient concentration filter," Jpn. J Imaging and Information Science in Medicine 24, 84-89 (2007).

[6] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "ImageNet classification with deep convolutional neural network," In: Advances in Neural Information processing Systems 25, 1106-1114 (2012).