AI Auxiliary Labeling and Classification of Breast Ultrasound Images

Lei Wang, Biao Liu, Shaohua Xu, and Ji Pan
Suzhou Kowloon Hospital Shanghai Jiao Tong University School of Medicine, Suzhou, China
Email: {thunderwanglei, LB19650125, hongyu771011}@163.com

Qi Zhou
Mitai Artificial Intelligent Institute, Suzhou, China
Email: 59074604@qq.com

***Abstract*—In this paper we developed a Deep Learning (DL) method to assistant radiologists quickly and accurately labeling and classifying the lesions of the breast ultrasound images. A faster R-CNN detector was trained to label and classify the lesions with the Breast Imaging Reporting and Data System (BI-RADS). The initial trained model used 2000 labeled images. From the testing results with 6000 images, we got poor accuracy. Therefore, we developed the second DL model with 4294-image set in which the images of BI-RADS 4 were removed. Then the second DL model was tested by 1000 images and used to classify 1836 images of BI-RADS 4. The results show that the classification accuracy, sensitivity and specificity are achieved as 92.37%, 98.34%, and 82.46%, respectively when it used to classify the BI-DADS 4 images into 4A and 4B, and 98.10%, 97.78% and 98.13%, respectively when it is used for breast cancer screening.**

***Index Terms*—deep learning, breast ultrasound image, DL labeling and classification**

I. Introduction

Deep learning (DL) has been gaining attraction in radiology [1]-[5]. For its excellent performance in image target recognition [6], DL tasks more and more applied to automatic diagnosis of medical images, especially for cancer diagnosis, such as breast cancer-assisted diagnosis [7]. It can greatly improve the efficiency and accuracy of lesions detection, increase the sensitivity and reduce the false detection rate [8]. The use of DL automatic diagnosis, while improving the consistency of reading, greatly saves the time of reading and analyzing for the experienced experts, remedy the gap of the lack of experience and skills of junior doctors, thereby saving a lot of medical costs.

DL has been a great success in the field of target identification. There are many effective DL recognition algorithms [9]-[12]. Similarly, DL has been widely used in the identification and classification of medical imaging lesions and has achieved great success. DL identification and classification systems reach or exceed expert levels [13]. Training a precise medical image DL recognition and classification model requires a large number of labeling and classification images. Therefore, the lack of well-labeled medical imaging data is the main obstacle to the success of a medical imaging DL model. However, labeling and classifying a large number of medical images is a time-consuming and tedious task for an experienced expert. The goal of this project is to develop an automatic labeling system for breast ultrasound images using the DL recognition model to automatically complete the annotation task of a large number of images and reach the level of the experts.

Manuscript received September 24, 2020; revised March 1, 2021.

Clinical Relevance: This DL labelling and classification method for breast ultrasound images is an efficient way to help the radiologists to annotate the lesions of the ultrasound breast images and relieve them of the heavy workload when they annotate a lot of images for DL model training.

II. Method

In this paper, an automatic DL labeling and classification system was developed for breast ultrasound images.

A. DL Model

The Faster R-CNN is used as the DL labeling and classification model for breast ultrasound images. By comparing efficiency and accuracy, it has good performance in detection the lesions of breast ultrasound images.

B. Training of the DL Model

- 2000 breast ultrasound images labeled and classified by experienced radiologists were used to train the initial DL model.
- 6000 breast ultrasound images were automatically labeling and classification by the initial DL model. The experienced experts to cross check them and then put the correct labeled and classified image into the training set.
- 4294-image set removed BI-RADS 4 were used to retraining the DL model.
- Then, the unlabeled BI-RADS 4 images were automatically labeled and classified into 4A and 4B by the retraining DL model.

doi: 10.18178/joig.9.2.45-49

- Finally, the performance of the automatically labeled and classified by the DL model was tested by a 1000-images testing set.

C. Testing

The testing show that:

(a) For automatically labeled and classified images of BI-RADS 4, the accuracy, sensitivity, and specificity were achieved as 92.37%, 98.34%, and 82.46%, respectively, and 98.10%, 97.78% and 98.13% for BI-RADS 5.

(b) For breast cancer screening, category as malignant and benign, the accuracy, sensitivity and specificity were achieved as 98.10%, 97.78% and 98.13%, respectively.

As the result, the DL method can be used to assistant the radiologists label and classify the lesions of the breast ultrasound images fast and accurately. Moreover, when the training data is increased, the accuracy of labeling and classification will be further improved.

III. Data and Process

Fig. 1 shows the methodology of proposed methods of the Faster R-CNN recognition model:

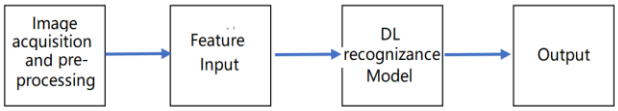

Figure 1. Medical imaging DL processing.

In the automatic labeling and classification system of medical imaging, we used the image target recognition algorithm to automatically classify and extract the lesions. In this study, the Faster R-CNN framework was selected after testing and comparing other frameworks.

A. The Procedures of DL Automatically Labeling and Classification

Fig. 2 shows the procedures of automatically labeling and classification breast ultrasound images by the DL model.

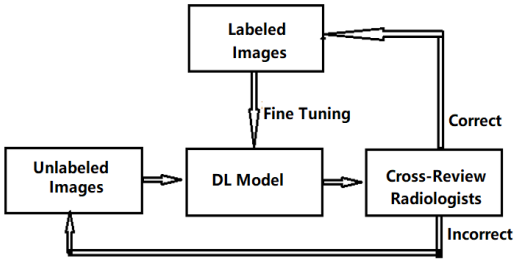

Figure 2. The procedures of automatically labeling and classification by DL model.

(a) The DL model trained by the original training set is used to automatically label and classify the unlabeled images.

(b) The correct labeled and classified images were put into the training (labeled image) set after they were cross-viewed and agreed by three radiology experts. The incorrect labeled and classified images were put into the unlabeled images set.

(c) The new training set was used to fine tune the DL model.

(d) Going back to (a) and using the new DL model labeling and classification the unlabeled image again.

Please pay attention to that the increasing images of the unlabeled image set can come from either the incorrect labeled images in the processing or from different clinical hospitals.

B. The Category of the Breast Ultrasound Images

The breast ultrasound images were labeled and classified following the American College of Radiology BI-RADS (Breast Imaging Reporting and Data System) [14]. They were categorized as follows:

- BI-RADS 1, Negative: The breasts are symmetric, and no masses, architectural distortion or suspicious calcifications are present.
- BI-RADS 2, Benign Finding: A benign is found in the ultrasound image. Fat-containing lesions such as oil cysts, lipomas, galactoceles and mixed-density hamartomas. They all have characteristically benign appearances and may be labeled with confidence.
- BI-RADS 3, Probably Benign Finding: A finding placed in this category should have less than a 2% risk of malignancy.
- BI-RADS 4, Suspicious Abnormality: It has a wide range of probability of malignancy (2 - 95%). By subdividing BI-RADS 4 into 4A, 4B and 4C, it is encouraged that relevant probabilities for malignancy be indicated within this category so the patient and her physician can make an informed decision on the ultimate course of action.
- BI-RADS 5, Highly Suggestive of Malignancy. The percutaneous tissue diagnosis is malignant.

First, the initial DL model is trained with 2000 images well labeled and classified by a group of radiology experts, and then more images are automatically labeled and classified by the DL model. The DL model will be fine-tuned again and again when the labeled image in the training set increased.

C. The Development of the DL Model

We use object recognition models to frame out the lesions area and extract the features. Here the Faster R-CNN model in which the CNN with ResNet-101, was used and given an 800×800 pixel view of the breast ultrasound image. During the model development, we augmented our training data ten times with up, down, left, right flips, random rotations, and bright processing of the original images, and experimented with various regularization strategies and model architectures, such as testing different CNN, ResNet-34 and ResNet-50, normalization in different stages. Finally, we chose this DL model because it is fast and has the best performance.

IV. Results

In this paper, the initial DL model was trained by 2000 images which were categorized accordance with the BI-RADS. The 6000 images were automatically labeled

and classified by the initial DL model. The second model was trained by the training set in which the 4294 images were well labeled and classified and the images of BI-RADS 4 were removed. Then, the second DL model was used to category 1836 images of BI-RADS 4 into 4A, higher probability of benignity, and 4B, higher probability of malignancy. In addition, the 1000 images in the testing image set are automatically labeled and classified by the second DL model.

A. Methods of Evaluation of the Models

In this paper, the performance of different models is evaluated via three evaluation criteria, that is, accuracy (ACC), sensitivity (SEN), and specificity (SPE). More specifically, the accuracy means the proportion of images that are correctly predicted among all studied images, the sensitivity denotes the proportion of true positive that is correctly predicted, and the specificity represents the proportion of the true negative that are correctly predicted. The definitions and equations as follow [15]:

$$ACC = \frac{Correct\ predictions}{Total\ images} = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

$$SEN = \frac{Correctly\ Predicte\ Positive}{All\ Predicted\ Positive} = \frac{TP}{TP+FN} \quad (2)$$

$$SPE = \frac{Correctly\ Predicte\ Negtive}{All\ Predicted\ Negative} = \frac{TN}{FP+TN} \quad (3)$$

here, *TP* is all true positive, *FP* is all false positive, *TN* is all true negative, and *FN* is all false negative.

B. The Testing of the Initial DL Model

The initial DL model was trained by 2000 images which were categorized accordance with the five-way BI-RADS. The 6000 images were used to test it.

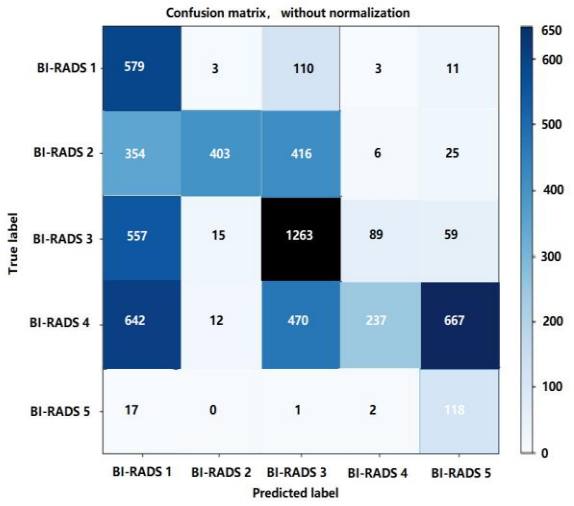

Figure 3. Confusion matrix without normalization for the testing of 6000 images with initial DL model.

From the confusion matrix (Fig. 3), we found that the images of BI-RADS 5 were labeled and classified well, sensitivity = 85.5%. The other images were not well categorized by the initial DL model. For example, in the True 4 category, only 11.69% were identified.

C. The Testing of the Second DL Model

The second DL model was trained by the 4294-image set in which the images of BI-RADS 4 were removed.

The 1000 images removed BI-RADS 4 were used to test the second DL model. The test result as show in the confusion matrix:

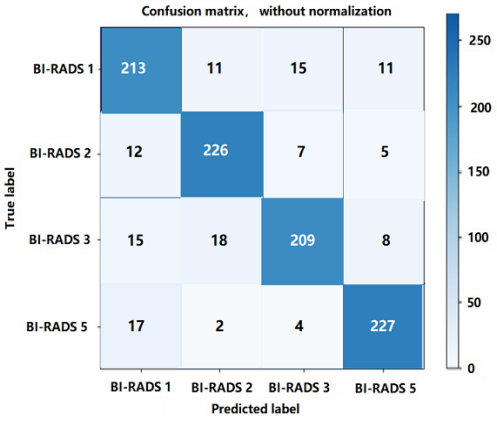

Figure 4. Confusion matrix without normalization for the 1000 test images set.

The evaluation of the performance of the second DL model by the Accuracy, Sensitivity, and Specificity are showed in Table I.

TABLE I. THE PERFORMANCE OF THE SECOND DL MODEL

Category	*ACC(%)*	*SEN(%)*	*SPE(%)*
BI-RADS 1	***91.14***	***85.23***	***94.37***
BI-RADS 2	***91.50***	***84.15***	***94.41***
BI-RADS 3	***92.40***	***87.38***	***94.56***
BI-RADS 5	***98.10***	***97.78***	***98.13***

D. The Accuracy of the Second DL Model for Binary Prediction

In the breast cancer screening, it is only necessary to determine the positive (malignant) or negative (benign) of the lesions. Therefore, the binary prediction (two categories) will be used. The results are showed in Table II.

TABLE II. BINARY PREDICTIONS

		Predicted Category	
		Positive	***Negative***
Actual Category	***Positive***	***88***	***2***
	Negative	***17***	***893***

Using the data from Table II, we calculate that the labeling and classification accuracy, sensitivity and specificity for the benign and malignant were achieved as 98.1%, 97.78% and 98.13%, respectively.

E. The Accuracy of the Second DL Model to Category of BI-RADS 4 Images

The 2nd DL model was used to label and classify the 1836 images of BI-RADS 4. The results are showed in Table III.

TABLE III. CATEGORY OF BI-RADS 4

		Predicted Category	
		4A	***4B***
Actual Category	***4A***	***1127***	***19***
	4B	***121***	***569***

From Table III, we calculated that the labeling and classification accuracy, sensitivity and specificity were achieved as 92.37%, 98.34% and 82.46%, respectively.

V. DISCUSSION

Training a highly sensitive and accurate breast cancer DL diagnostic model requires a large number of well-labeled and classified breast ultrasound images and a lot of hard work from the experienced radiologists. Using our DL model proposed in this paper, you can start training a DL model from a small number of images well labeled and classified by radiology experts, then use the DL model to label and classify the images quickly. Finally, the labeled and classified images were cross-checked and corrected by multiple radiology experts. With this way, you gradually increase the images of the well-labeled in your training image set while reducing the heavy workload of radiologists.

- The initial DL model trained by the 2000-image set with BI-RADS categories was not accurate enough to label and classify the breast ultrasound images from the test of 6000 images. This indicated that it is difficult to train a sufficiently accurate DL model if there are not enough training images well labeled and classified.
- The second DL model trained by 4294-image set which removed BI-RADS 4 images performed very well in the labeling and classification of BI-RADS 4 images into 4A which is more likely benign, and 4B which is more likely malignant. The DL model trained by the image set removed the BI-RADS 4 images is easier to label and classify the BI-RADS 4 images with more accuracy according to benign and malignant probability.
- Obviously, with the DL model and the method of DL automatically labeling and classification of breast ultrasound images in this paper, it is easy to quickly increase the image of the training set while reducing the heavy workload of radiology experts. It can be expected that when the number of images labeled and classified by this DL model is enough, the new model will reach the level of radiology experts. This DL labeling and classification method for breast ultrasound images is an efficient way to help the radiologists to annotate the lesions of the ultrasound breast images and relieve them of the heavy workload.

VI. CONCLUSION

In summary, we present a DL model based on Faster R-CNN. This model was trained with 4294-images training set in which the BI-RADS 4 images were removed and can be used for DL auxiliary labeling and classification of the breast ultrasound images acutely and efficiently. It showed excellent performance when it used to label and classify the BI-RADS 4 images into 4A and 4B. It also demonstrates high accuracy, sensitivity and specificity when used in breast cancer screening. Our tool provides a simple and cost-effective way to label and classify the images for the radiologists. This DL model will provide a way to get free the radiology experts from the heavy work of labeling and classification of the lesions of the images and solve the problem of insufficient radiologists and high cost in training the medical diagnostic DL model.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Lei Wang, ultrasound expert, Suzhou Kowloon Hospital Shanghai Jiao Tong University School of Medicine, project leader, wrote the paper as the main contributor.

Biao Liu, professor and ultrasound expert, Suzhou Kowloon Hospital Shanghai Jiao Tong University School of Medicine, led the Labelling of the training data.

Shaohua Xu and Pan Ji, ultrasound expert, Suzhou Kowloon Hospital Shanghai Jiao Tong University School of Medicine, collected and labeled the training data.

Qi Zhou, AI expert, built and trained the AI model.

ACKNOWLEDGMENT

The author thanks the experts of the First Affiliated Hospital of Suzhou University, and Suzhou Kowloon Hospital of Shanghai Jiao Tong University School of Medicine, for providing a large number of breast ultrasound images. Also, thanks for support from Dr. Qiling Qin and Dr. Yaping Liu of Mitai Artificial Intelligence Institute.

REFERENCES

[1] N. Wu, K. J. Geras, Y. Shen, et al. (2017). Breast density classification with deep convolutional neural networks. arXiv preprint arXiv: 1711.03674. [Online]. Available: https://arxiv.org/abs/1711.03674

[2] M. Bahl, R. Barzilay, A. B. Yedidia, N. J. Locascio, L. Yu, and C. D. Lehman, "High-risk breast lesions: A machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision," *Radiology*, vol. 286, no. 3, pp. 810-818, 2018.

[3] M. Kohli, L. M. Prevedello, R. W. Filice, and J. R. Geis, "Implementing machine learning in radiology practice and research," *AJR Am. J. Roentgenol*., vol. 208, no. 4, pp. 754-760, 2017.

[4] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574-582, 2017.

[5] K. J. Geras, S. Wolfson, Y. Shen, *et al.* (2017). High-Resolution breast cancer screening with multi-view deep convolutional neural networks. arXiv preprint arXiv: 1703.07047. [Online]. Available: https://arxiv.org/abs/1703.07047

[6] K. Suzuki, "Overview of deep learning in medical imaging," *Radiol. Phys. Technol.*, vol. 10, pp. 257-273, 2017.

[7] A. Yala, C. D. Lehman, T. Schuster, T. Portnoi, and R. Barzilay, "A deep learning mammography-based model for improved breast cancer risk prediction," *Radiology*, 2019.

[8] C. D. Lehman, A. Yala, T. Schuster, *et al.*, "Mammographic breast density assessment using deep learning: clinical implementation," *Radiology*, vol. 290, no. 1, pp. 52-58, 2019.

[9] S. Han, H. K. Kang, J. Y. Jeong, M. H. Park, W. Kim, W. C. Bang, and Y. K. Seong, "A deep learning framework for supporting the classification of breast lesions in ultrasound images," *Phys. Med. Biol*., vol. 62, pp. 7714-7728, 2017.
[10] D. C. Ciresan, U. Meier, J. Masci, and J. Schmidhuber, "Multi-column deep neural network for traffic sign classification," *Neural. Netw.*, vol. 32, pp. 333-338, 2012.
[11] R. Girshick, "Fast R-CNN," in *Proc. IEEE International Conference on Computer Vision*, 2015.
[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017.
[13] M. Bahl, R. Barzilay, A. B. Yedidia, N. J. Locascio, L. Yu, and C. D. Lehman, 2017, "High-Risk breast lesions: A machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision," *Radiology*, 2017.
[14] American College of Radiology, *ACR BI-RADS Atlas—Mammography*, 5th ed., Reston, Va: American College of Radiology, 2013.
[15] C. R. Lamb, "Statistical briefing - Sensitivity and specificity," *Veterinary Radiology & Ultrasound*, vol. 48, no. 2, p. 189, March 2007.

Lei Wang is a ultrasound expert in Suzhou Kowloon Hospital, Shanghai Jiaotong University School of Medicine, Suzhou, China. Currently, his research interests are the ultrasound diagnosis for superficial organs and intervention treatment. He received his BS in Medical imaging science from Harbin Medical University, China in 2005.
Lei Wang, Ji Pan, Biao Liu, etc., The value of Ultrasound-guided percutaneous microwave ablation therapy for breast benign nodules, Journal of Qiqihar Medical College, 2018, 5:1024-1026.
Lei Wang, Biao Liu, Ji Pan, Lingyan Zhang, etc., The application value of real-time shear wave elastic imaging technique in thyroid malignant tumor, Imaging research and medical applications, 2020, 5:44-46.