Medical Image Analysis 52 (2019) 185–198

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Automated diagnosis of breast ultrasonography images using deep neural networks

Xiaofeng Qi[a], Lei Zhang[a], Yao Chen[b], Yong Pi[a], Yi Chen[a], Qing Lv[b,*], Zhang Yi[a,*]

[a] Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, 610065, PR China
[b] Department of Galactophore Surgery, West China Hospital, Sichuan University, Chengdu, 610041, PR China

ARTICLE INFO

Article history:
Received 1 September 2018
Revised 26 November 2018
Accepted 19 December 2018
Available online 20 December 2018

Keywords:
Breast cancer
Ultrasonography
Deep neural networks

ABSTRACT

Ultrasonography images of breast mass aid in the detection and diagnosis of breast cancer. Manually analyzing ultrasonography images is time-consuming, exhausting and subjective. Automated analyzing such images is desired. In this study, we develop an automated breast cancer diagnosis model for ultrasonography images. Traditional methods of automated ultrasonography images analysis employ hand-crafted features to classify images, and lack robustness to the variation in the shapes, size and texture of breast lesions, leading to low sensitivity in clinical applications. To overcome these shortcomings, we propose a method to diagnose breast ultrasonography images using deep convolutional neural networks with multi-scale kernels and skip connections. Our method consists of two components: the first one is to determine whether there are malignant tumors in the image, and the second one is to recognize solid nodules. In order to let the two networks work in a collaborative way, a region enhance mechanism based on class activation maps is proposed. The mechanism helps to improve classification accuracy and sensitivity for both networks. A cross training algorithm is introduced to train the networks. We construct a large annotated dataset containing a total of 8145 breast ultrasonography images to train and evaluate the models. All of the annotations are proven by pathological records. The proposed method is compared with two state-of-the-art approaches, and outperforms both of them by a large margin. Experimental results show that our approach achieves a performance comparable to human sonographers and can be applied to clinical scenarios.

1. Introduction

Breast cancer is the most common cancer among women worldwide. There has been a general uptrend in the morbidity of breast cancer since the 1990s. (Fitzmaurice et al., 2015) According to the World Health Organization, breast cancer is responsible for over 500,000 deaths each year and 1.7 million new cases are diagnosed every year. Breast cancer is the cancer with the highest incidence for women in 161 countries and the most common cause for cancer deaths in women in 98 countries. Breast cancer can be cured if diagnosed and treated early. Therefore, early detection is crucial in increasing survivability, and periodic examinations are essential to potential risk groups (Lee, 2002).

Nowadays, diagnosis by pathological examination is considered the gold standard for almost all types of cancer, and breast cancer diagnosis usually requires a biopsy (Leong and Zhuang, 2011). However, pathological examination is inefficient and inconvenient, requiring the participation of experienced surgeons and pathologists. In histopathology image analyses, pathologists classify the tissues by cytology and histology descriptions, such as cell nuclei grade, comedo necrosis, and micro infiltration. The identification of immunohistochemical markers is also required. This type of manual feature extraction is tedious and exhausting, and may result in misdiagnosis or inconsistent results between different pathologists.

The breasts are superficial organs of the human body, which means that anomalies in breast mass are detectable by imaging techniques. Mammography is one of the most popular imaging modalities in detecting breast cancer. However, mammography is inapplicable to dense breasts, because of its sensitivity and ionization limitations. Thus, other modalities such as ultrasonography are often suggested. Ultrasonography is capable of detecting and classifying nodules in the breast mass and is widely used because of its convenience, speed, non-invasiveness and low cost. Usually, ultrasonography images are analyzed manually by sonographers, this is time-consuming and subjective. Thus, the development of an automated breast cancer imaging analysis model is essential for improving the efficiency and reliability of ultrasound examinations.

* Corresponding author.
E-mail addresses: lqlq1963@163.com (Q. Lv), zhangyi@scu.edu.cn (Z. Yi).

https://doi.org/10.1016/j.media.2018.12.006
1361-8415/

An ultrasound examination involves taking several images of the breast mass using ultrasonic equipment. The sonographer interacts with the patient, examines tissues of interest from various directions and captures images containing features of the artifacts. The sonographer then provides descriptions of each image. If there are abnormalities detected in ultrasonography, it is recommended that the patient submit to further analyses such as mammography, biopsy, or frozen-section examination. The results of these examinations lead to a final diagnosis accompanied by a pathological report. Although ultrasonic equipment is widely used in medical institutions, professional sonographers and breast surgeons are in short supply to primary hospitals and clinics. Clinical diagnoses often take a long time following ultrasonography examinations, which may impede successful recovery. Furthermore, manual analysis of ultrasonography images is highly subjective: the specificity and sensitivity of manual diagnoses are 91% and 84% in the classification of breast cancer (Giger et al., 2013).

Automated breast cancer diagnosis is helpful to make the diagnosis of breast cancer more reliable and efficient. Many computer-aided diagnosis systems have been proposed to assist sonographers. However, most of them focus on the detection of breast nodules, followed by manually designed feature extractors.

In this study, we propose a new methodology to diagnose breast ultrasonography images in a fully automated manner. Taking an ultrasonography image as input, the automated breast cancer diagnosis model generates a diagnosis and clinical advice for sonographers and breast surgeons. Our model could be used in some medical institutions with ultrasonic equipment. After the ultrasound examination, sonographers and breast surgeons could use our model to perform instant diagnosis, and the diagnosis results can assist the doctors to provide advice for the patient. The model is also easily accessible to patients everywhere through the internet. The diagnosis is accomplished in two steps. The first step is designed to recognize whether the input image contains malignant tissues, which is the most urgent concern. The second step is to recognize solid nodules in the image because solid nodules have a strong correlation with malignant tissues and should be treated carefully. If malignant tumors or solid nodules are recognized, the patient are suggested to get further examinations and medical treatments immediately, otherwise periodic inspections are recommended. Early and reliable diagnoses allow timely treatments, leading to the reduction of the morbidity and mortality of breast cancer.

Creating an automated breast cancer diagnosis model based on ultrasonography images is a challenging task because of the following obstacles. (1) In an ultrasonography image, breast cancer and disease are indicated by a large variety of features including the texture of breast mass, the existence of calcification, and the thickness of vessels and ducts. These features are of different shapes and sizes, rendering them difficult to extract appropriately for classification. Traditional methods use carefully hand-crafted low-level operators and algorithms to extract features from the images. These methods are sensitive to image quality and may have poor generalization ability under different scenarios. Moreover, the image acquisition is often performed under varying conditions, increasing the variability of tissue appearance. (2) To learn the features of breast ailments from ultrasonography images, a large number of annotated data is required. The collection and annotation of such data are laborious. (3) The assessment of ultrasonography images depends on the experience of the annotators, and inconsistencies between different annotators for the same image are likely to occur.

We overcome the above stated challenges through the following methods. (1) We use convolutional neural networks with multi-scale filters and skip connections to extract the features from the data. Moreover, we propose a region enhance mechanism, which helps to improve the diagnosis accuracy. (2) We construct a large-scale dataset of ultrasonography images annotated by professional sonographers and surgeons. (3) Each image is annotated by two different experts, referred to the histopathology diagnosis by pathologists.

The proposed study makes several contributions:

(1) We construct a large-scale ultrasonography image dataset annotated by professional sonographers and surgeons. The dataset is used to build fully automated models of breast cancer diagnosis. To the best of our knowledge, our dataset is larger in size than any ultrasonography image datasets publicly available.
(2) We propose a novel breast cancer diagnosis model. We use convolutional neural networks to construct two sub-networks to analyze and classify breast ultrasonography images. A novel region enhance mechanism is proposed to improve the performance of our model. To train the networks, a cross training algorithm is introduced.
(3) The proposed model is capable of recognizing malignant tumors and solid nodules in breast mass. As evaluated by professional sonographers and breast surgeons with over 20 years of experience, the classification results are comparable to those of human experts, indicating that the neural networks have learned the correct features for breast cancer diagnosis.

2. Related works

Computer-aided Diagnosis System applied to breast cancer imaging have been studied for decades (Jalalian et al., 2012). In this section, an overview of previous studies in breast cancer diagnosis and deep neural networks is presented.

2.1. Traditional methods for breast cancer diagnosis

In breast ultrasonography images, the majority of the existing publications focus on segmentation and classification using traditional image processing techniques. Four main steps are involved: (1) Lesion detection; (2) image preprocessing; (3) feature extraction; and (4) classification. The lesion detection step aims at detecting the location of breast nodules and the following two steps focus on extracting graphical features. Based on these features, the last step is to build a classifier for breast cancer diagnosis.

For example, Chen et al. (2005) used the differences of gray value of neighboring pixels to estimate the fractal dimension of a breast ultrasonography image by using the fractal Brownian motion. Furthermore, a CAD system based on the fractal analysis and the k-means classification was proposed to classify the breast lesions into two classes: Malignant and benign. Ultrasonography images of 110 malignant and 140 benign tumors were included in the dataset. The AUC was 0.922 and the accuracy was up to 88.80%. Kuo et al. (2008) used the Virtual Organ Computer-aided AnaLysis (VOCAL^{TM}) imaging program to draw contour around the breast lesion and obtain vascularization histogram indices for the tumor. 102 benign and 93 malignant breast tumor images are collected. A multi-layer perceptron was used to build a classifier and achieved a sensitivity for classifying malignancy of 90.3%, with a specificity of 79.4%, a positive predictive value (PPV) of 80% and a negative predictive value (NPV) of 90%. Tianur et al. (2017) employed region growing method along with manually conducted region of interest (RoI) to determine the area of breast lesions in ultrasonography images. Six histogram features were extracted, followed by a multilayer perceptron to classify the breast cancer lesion into posterior acoustic enhancement or no posterior acoustic. The dataset consisted of 69 lesions with posterior acoustic enhancement and 29 no

posterior acoustic. Performance of the proposed method achieved accuracy of 87.79%, sensitivity of 92.75% and specificity of 82.75%.

Although traditional methods have made great achievements in aiding the diagnosis of breast cancer, the drawbacks are obvious. The classification of these methods heavily relies on the lesion segmentation, meaning that errors in segmentation may cause failures in subsequent classification. In real-world applications, there is a lot of noise in ultrasonography images, making it harder to recognize. Traditional methods lack robustness because of their dependency on hand-crafted features. Moreover, choosing features manually is exhausting and subjective, the performances of these approaches are difficult to improve.

2.2. Deep neural networks for breast cancer diagnosis

Deep neural networks, powered by advances in compute capability and very large annotated datasets, have achieved revolutionary breakthroughs in computer vision. In recent years, very deep convolutional neural networks (CNNs) became the mainstream in image classification tasks since AlexNet (Krizhevsky et al., 2012) won the 2012 ImageNet competition (Russakovsky et al., 2015). Numerous novel architectures of CNNs have been proposed in a range of image classification and detection applications (Simonyan and Zisserman, 2014; Girshick et al., 2014; Shelhamer et al., 2017). Compared with traditional image processing methods, deep CNNs extract features of different levels in a data-driven manner and there is no need for hand-crafted features, reducing the workload of doctors.

Yap et al. (2018) proposed the use of deep learning approaches for breast lesion detection in ultrasonography images and investigated three different methods: Patch-based LeNet (LeCun et al., 1998), U-Net (Ronneberger et al., 2015), and transfer learning approach with a pretrained FCN-AlexNet. Their methods are compared with four state-of-the-art lesion detection algorithms and outperformed the latter. The FCN-AlexNet based model achieved a true positive rate over 0.93, averaged false positives per image is around 0.16.

Byra et al. (2017) built a neural network with three convolutional layers and two fully-connected layers for breast classification. The dataset used consisted of 166 malignant tumors and 292 benign masses. Five-fold cross-validation was performed, the averaged AUC was 0.912, with an accuracy of 83.0% and a sensitivity of 82.4%.

Cheng et al. (2016) used stacked denoising autoencoders (Vincent et al., 2010) to build a CADx for the classification of malignant and benign tissues in breast ultrasonography images and pulmonary nodules in CT scans. The method was compared with two state-of-the-art traditional algorithms, outperformed them on both tasks. The proposed model achieved an accuracy of 94.4% and an AUC of 0.984 on a dataset involving 275 benign and 245 malignant lesions.

Han et al. (2017) exploited the deep learning framework to differentiate malignant and benign lesions and nodules in breast ultrasonography images. A biopsy-proven dataset containing 7408 images was built, with target regions of interest selected by radiologists. The images were cropped to fix the distance between the lesion boundary and the boundary of the cropped image. GoogLeNet (Szegedy et al., 2015) was employed to build the classification model, showing an accuracy of about 90%, a sensitivity of 86% and a specificity of 96%.

According to previous studies, deep neural networks have shown better performance than traditional methods for breast cancer diagnosis based on ultrasonography images. However, most works are based on breast lesion detection, requiring manually annotated RoIs. Moreover, missed diagnosis is the most severe situation in breast cancer examinations, false negative rate (FNR) should be as low as possible. The FNRs of previous studies are over 10%, making it unsuitable for real-world clinical applications. In the current study, we propose a fully automated manner for breast ultrasonography images analysis. The input image is classified into malignant or non-malignant in an end-to-end manner, with a false negative rate lower than 5%. Further investigation is performed on non-malignant cases to recognize solid nodules, which is helpful to decrease missed diagnosis and allows doctors to focus on high risk groups. Experimental results show that our approach is applicable to real-world scenarios.

3. Methods

In this section, we describe the proposed method used in the diagnosis of breast ultrasonography images in detail. The diagnosis is accomplished by two neural networks in a cascade manner. Firstly, a network is constructed to classify images according to whether they contain malignant tumors since malignant tumors are the most severe among all breast ailments, we call this network the Mt-Net. Secondly, the images are further classified by another network according to whether they include solid nodules, as solid nodules are closely related to cancer and should be treated carefully. The second network is called the Sn-Net. We describe the basic architectures of the two networks in the following subsection, then a novel and powerful region enhance mechanism is presented.

3.1. Basic architectures of the proposed neural networks

Deep neural networks are a powerful tool in machine learning domains such as pattern recognition, audio signal recognition, and natural language processing. This kind of model is formally denoted by a highly nonlinear function $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ represents the input space and $\mathcal{Y}$ represents the output space. Deep neural networks learn to create a more abstract representation of the data, and have been shown to be universal function approximators.

Deep convolutional neural networks have achieved remarkable performance on pattern recognition tasks because of their powerful feature extraction capabilities. A deep CNN is a feedforward network constructed of convolutional, pooling and fully connected layers. The main components of deep CNNs are convolution kernels, resembling the manually designed and calibrated filters in traditional image processing methods. The operation in each convolutional layer of the CNNs is defined by:

$$\boldsymbol{a}_n^{l+1} = f\left(\sum_{m=1}^{M} \boldsymbol{W}_{nm}^{l} * \boldsymbol{a}_m^{l} + \boldsymbol{b}_n^{l+1}\right) \tag{1}$$

where $\boldsymbol{a}_n^{l+1}$ and $\boldsymbol{a}_m^{l}$ represent the nth feature map in the $(l+1)$th layer and the mth feature map in the lth layer, respectively, and M denotes the number of feature maps in layer l. $\boldsymbol{W}_{nm}^{l} \in \mathbf{R}^{I \times J}$ represents the 2D convolution kernel of size $I \times J$ from the mth feature map in the lth layer ($\boldsymbol{a}_m^{l}$) to the nth feature map in the posterior layer ($\boldsymbol{a}_n^{l+1}$) and $*$ denotes the 2D convolution operation. If the previous layer contains more than one feature map, the results of the corresponding convolution operations are summed up as shown in Eq. (1), then passed through a nonlinear activation function f. There are many activation functions in neural networks, the most common used in CNNs is *relu* function, as defined in Eq. (2). Convolutional layers are typically followed by pooling layers where the most frequently used methods are max pooling and average pooling. Fully connected layers usually appear in the bottom of the architecture, unlike convolutional layers, which are locally connected, all neurons in a fully connected layer are connected with neurons

Table 1
The outline of the Mt-Net used in our work.

Type	Kernel size/stride(padding)				Output size
conv	$3 \times 3/2(0)$				$149 \times 149 \times 32$
conv	$3 \times 3/1(0)$				$147 \times 147 \times 32$
conv	$3 \times 3/1(1)$				$147 \times 147 \times 64$
pool	$3 \times 3/2(0)$				$73 \times 73 \times 64$
conv	$1 \times 1/1(0)$				$73 \times 73 \times 80$
conv	$3 \times 3/1(0)$				$71 \times 71 \times 192$
pool	$3 \times 3/2(0)$				$35 \times 35 \times 192$
inception	$1 \times 1/1(0)$ $3 \times 3/1(1)$ $3 \times 3/1(1)$	$1 \times 1/1(0)$ $5 \times 5/1(2)$	pool$3 \times 3/1(1)$ $1 \times 1/1(0)$	$1 \times 1/1(0)$	$35 \times 35 \times 320$
	Feature map concatenation				
block (35) $\times 10$	$1 \times 1/1(0)$ $3 \times 3/1(1)$ $3 \times 3/1(1)$	$1 \times 1/1(0)$ $3 \times 3/1(1)$	$1 \times 1/1(0)$		$35 \times 35 \times 320$
	Feature map concatenation				
	$1 \times 1/1(0)$				
	Scale down by factor $0.17 +$ input				
	relu activation				
inception	$1 \times 1/1(0)$ $3 \times 3/1(1)$ $3 \times 3/2(0)$	$3 \times 3/2(0)$	pool$3 \times 3/2(0)$		$17 \times 17 \times 1088$
	Feature map concatenation				
block (17) $\times 20$	$1 \times 1/1(0)$ $1 \times 7/1(1,2)$ $7 \times 1/1(2,1)$		$1 \times 1/1(0)$		$17 \times 17 \times 1088$
	Feature map concatenation				
	$1 \times 1/1(0)$				
	Scale down by factor $0.10 +$ input				
	relu activation				
inception	$1 \times 1/1(0)$ $3 \times 3/1(1)$ $3 \times 3/2(0)$	$1 \times 1/1(0)$ $3 \times 3/2(0)$	$1 \times 1/1(0)$ $3 \times 3/2(0)$	pool$3 \times 3/2(0)$	$8 \times 8 \times 2080$
	Feature map concatenation				
block (8) $\times 9$	$1 \times 1/1(0)$ $1 \times 3/1(0,1)$ $3 \times 1/1(1,0)$		$1 \times 1/1(0)$		$8 \times 8 \times 2080$
	Feature map concatenation				
	$1 \times 1/1(0)$				
	Scale down by factor $0.20 +$ input				
	relu activation				
block (8) without the last relu activation	$1 \times 1/1(0)$ $1 \times 3/1(0,1)$ $3 \times 1/1(1,0)$		$1 \times 1/1(0)$		$8 \times 8 \times 2080$
	Feature map concatenation				
	$1 \times 1/1(0)$				
	Scale down by factor $0.20 +$ input				
conv	$1 \times 1/1(0)$				$8 \times 8 \times 1536$
average pool	$8 \times 8/1(0)$				1536
dropout (0.2)					1536
linear					2
softmax					2

in the subsequent layer.

$$f(x) = \begin{cases} 0 & x \leq 0, \\ x & x > 0. \end{cases} \tag{2}$$

Breast cancer and ailments are characterized by a large variety of features, such as the texture of breast mass, existences of calcification, thicknesses of vessels and ducts, comedo necrosis and micro infiltration. These features are of different shapes and sizes which makes it difficult to extract appropriately for classification. Furthermore, the ultrasonography image has a high dimension and covers a wide range of breast mass, and the features indicating breast cancer and ailments lie in different local regions of the image. To overcome these challenges, we employ convolution kernels with different sizes to extract features of various scales. The application of multi-scale convolution kernels is known as the Inception module, which is firstly introduced by the GoogLeNet. Feature maps from the previous layer are passed through several convolutional layers and pooling layers with different kernel sizes like 1×1, 3×3, and 5×5, the output feature maps are then concatenated, as shown in Fig. 2. The intuition behind the use of Inception modules is to let the network learn features with different types, shapes, and sizes. To overcome the difficulties in the optimization of deep networks, residual module has been proposed. The module can be formulated as $\boldsymbol{a}^{l+1} = F^l(\boldsymbol{a}^l) + \boldsymbol{a}^l$, where $\boldsymbol{a}^l$ represents the input feature maps and $F^l(\boldsymbol{a}^l)$ denotes the residual connections.

To analyse breast ultrasonography images, we propose two networks in a cascade manner. The input image is first fed into the Mt-Net to identify malignant tumors, and then into the Sn-Net to recognize solid nodules. Latest studies indicated that a large convolutional kernel can be replaced by several small kernels, we follow the design and use two 3×3 kernels to replace a 5×5 kernel (Ioffe and Szegedy, 2015). Based on the Inception-Resnet-v2 architecture (Szegedy et al., 2016a), we equip the Mt-Net with Inception modules and skip connections, as shown in Fig. 2. We utilize 1×1 kernels to reduce the number of feature maps, and scale down the residuals with a factor between 0.1 and 0.2 before adding them to

Table 2
The outline of the Sn-Net used in our work.

Type	Kernel size/stride(padding)				Output size
conv		3 × 3/2(0)			149 × 149 × 32
conv		3 × 3/1(0)			147 × 147x32
conv		3 × 3/1(1)			147 × 147x64
pool		3 × 3/2(0)			73 × 73 × 64
conv		1 × 1/1(0)			73 × 73 × 80
conv		3 × 3/1(0)			71 × 71 × 192
pool		3 × 3/2(0)			35x35 × 192
block (3a)	1 × 1/1(0)	1 × 1/1(0)	pool3 × 3/1(1)	1 × 1/1(0)	35 × 35 × 256
block (3b)	3 × 3/1(1)	5 × 5/1(2)	1 × 1/1(0)		35 × 35 × 288
block (3c)	3 × 3/1(1)				35 × 35 × 288
		Feature map concatenation			
block (3d)	1 × 1/1(0)	3 × 3/2(0)	pool3 × 3/2(0)		17 × 17 × 768
	3 × 3/1(1)				
	3 × 3/0(2)				
		Feature map concatenation			
block (4a)	1 × 1/1(0)	1 × 1/1(0)	pool3 × 3/1(1)	1 × 1/1(0)	17 × 17 × 768
block (4b)	1 × 7/1(3)	1 × 7/1(3)	1 × 1/1(0)		17 × 17 × 768
block (4c)	7 × 1/1(3)	7 × 1/1(3)			17 × 17 × 768
block (4d)	1 × 7/1(3)				17 × 17 × 768
	7 × 1/1(3)				
		Feature map concatenation			
block (4e)	1 × 1/1(0)	1 × 1/1(0)	pool3 × 3/2(0)		8 × 8 × 1280
	1 × 7/1(3)	3 × 3/2(0)			
	7 × 1/1(3)				
	3 × 3/2(0)				
		Feature map concatenation			
block (5a)	1 × 1/1(0)	1 × 1/1(0)	pool3 × 3/1(1)	1 × 1/1(0)	8 × 8 × 2048
block (5b)	3 × 3/1(1)	1 × 3/1(1) 3 × 1/1(1)	1 × 1/1(0)		8 × 8 × 2048
	1 × 3/1(1) 3 × 1/1(1)				
		Feature map concatenation			
average pool		8x8/1(0)			2048
dropout (0.2)					2048
linear					2
softmax					2

the input feature map. The architecture of the Sn-Net is based on the Inception-v3 architecture (Szegedy et al., 2016b), and the final classification of the default network is replaced by a convolutional layer with kernel size of 1 × 1. The number of output feature maps are set to 2 for binary classification.

The basic architectures of the Mt-Net and the Sn-Net are presented in detail in Tables 1 and 2. In the Mt-Net, there are three types of inception blocks with skip connections, we name them block (35), block (17) and block (8), which are repeated 10, 20 and 9 times, respectively. In the Sn-Net, there are 11 inception blocks in total. To be clear, the blocks are named from block (3a) to block (5b).

3.2. Region enhance mechanism

In addition to the basic networks described above, we propose a novel region enhance mechanism. The motivation is to let the Mt-Net and the Sn-Net work collaboratively. Inspired by the attention mechanism, we use the feature visualization results as inputs of the proposed networks. The feature visualization result indicates the region of input image that the network focuses on, and we use it as an additional input to subsequent networks. For example, the visualization result of the Mt-Net represents the region responsible for the identification of malignant tumors, since malignant tumors are very close to solid nodules, this kind of information may be helpful for the recognition of solid nodules. Similarly, the visualization of the Sn-Net is beneficial for the classification of the Mt-Net.

Fig. 1 shows the architecture of our proposed method. The Mt-Net and the Sn-Net work in a collaborative way. The Mt-Net takes two images as input, the breast ultrasonography image, and the visualization result from the Sn-Net. Two inputs go through two pathways separately, high level features are aggregated in the middle of the network. The network outputs the classification result of malignant or non-malignant, as well as the visualization result. The Sn-Net takes the visualization and the ultrasonography image as input, and works in a similar way.

The visualization results are calculated based on the Class Activation Mapping (CAM) algorithm (Zhou et al., 2016). The CAM algorithm uses the weights for a particular category in the classification layer along with the last group of feature maps before the global average pooling, that is, 1536 feature maps of 8 × 8 size in the Mt-Net and 2048 feature maps of 8 × 8 size in the Sn-Net, to generate a heat map of the same size as the feature map. The heat map is then normalized and resized to match the size of input image. The computation of the class activation maps is defined as follows:

$$CAM_c = \sum_{n=1}^{N} \boldsymbol{W}_{cn}^{L-1} \cdot \boldsymbol{a}_n^{L-1}, \tag{3}$$

where c denotes the classification result of the network. For example, for the Mt-Net, $c = 1$ means non-malignant and $c = 2$ means malignant. $\boldsymbol{W}^{L-1}$ represents the convolutional kernel in the layer after global pooling, and $\boldsymbol{a}_n^{L-1}$ represents the nth feature map before global pooling.

We use the class activation map of the category with larger probability as the visualization result, meaning that $c = \arg\max_c \boldsymbol{a}^L$, where $\boldsymbol{a}^L$ is the *softmax* output of the last layer. The map acts as input of another network, and the high level feature aggregation are defined as follows:

$$\boldsymbol{a}^l = F_{img}\left(x_{img}\right) + \lambda * F_{cam}\left(x_{cam}\right) \tag{4}$$

where x_{img} and x_{cam} represent the breast ultrasonography image and the class activation map respectively. F_{img} and F_{cam} represent the convolutional layers that the two inputs go through, and λ de-

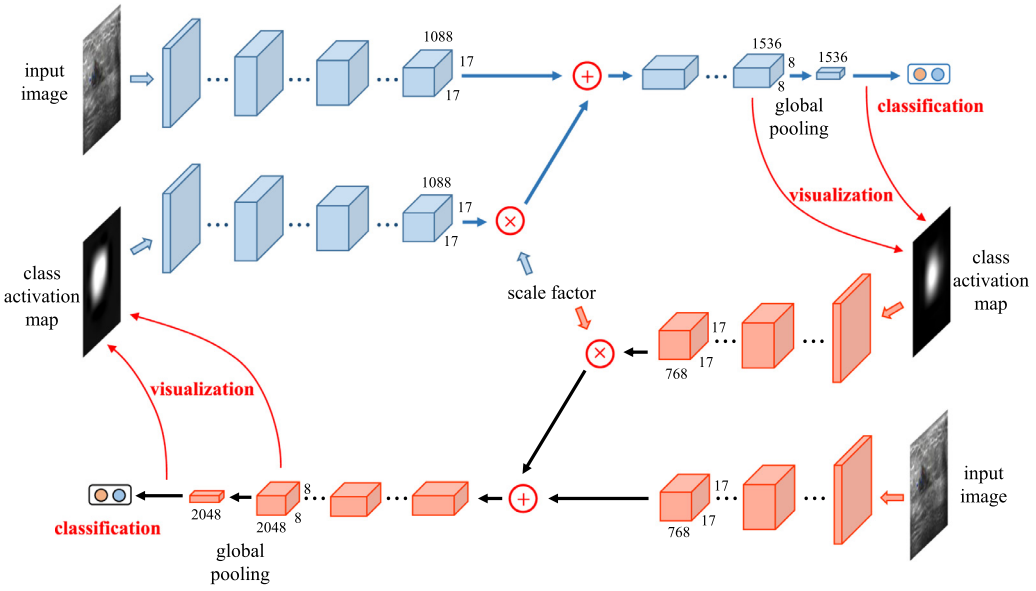

Fig. 1. The main architecture of the proposed model. Two networks, the Mt-Net and the Sn-Net work collaboratively in a cascade manner. We use class activation maps as additional inputs to construct a region enhance mechanism, experiment results show that the mechanism makes improvements to model performance.

notes the scale factor of the features from x_{cam}. Since the activation function in the basic network architectures are *relu* function, defined in Eq. (2), there may be many zeros in $F_{img}(x_{img})$ and $F_{cam}(x_{cam})$. We use add operations here to act as an "enhance" mechanism, the focused region is enhanced, and the left region is unchanged. Scale factor λ is used to control the strength of the enhancement.

For the Mt-Net, there are 40 inception blocks with skip connections in the basic network architecture. We perform the feature aggregation after 30 blocks, on feature maps of size $1088 \times 17 \times 17$, as presented in Fig. 1. For the Sn-Net, which is based on the Inception-v3 architecture, the feature aggregation is performed after block (3d), which is introduced in Table 2.

3.3. Multi-input prediction

In an ultrasound examination, the sonographer examines tissues of interest from various directions and using various modes, with various pressures of the probe. With different directions and pressures of the probe, the captured images of the same tissue show great difference. As presented in Fig. 3, different pressures cause deformation of the same tumor, generating different images. The examined tumor is round under low pressure, while under high pressure, it appears to be oval and flat. In order to encompass the main characteristics of the tissues and reduce the side effects of misoperation, the sonographer needs to be trained rigorously for many years.

Conventional CNNs takes square images as inputs, for example, the input size is 224×224 for Inception-v2, and 299×299 for Inception-v3. Confronted with non-square images, the most frequently used image preprocessing method is resizing the original image to match the input size of CNNs. In this study, because the images are captured with different kinds of instruments, the collected breast ultrasonography images are of various sizes and aspect ratios. However, resizing the ultrasonography images to square images causes deformation of the breast tissues, resulting in deviation from the original images. As presented above, images with unexpected deformation would have little diagnostic value, which may have side effects on the training and prediction of neural networks.

In this study, we propose a multi-input prediction method, as shown in Fig. 4. The original images are resized with aspect ratios maintained. To generate a square input image, we first scale the smallest side of the original image to the input size of the neural networks, that is, 299 pixels for both the Mt-Net and the Sn-Net. Then, a random crop of 299×299 pixels is selected for training. For the validation and testing set, five uniformly-spaced image areas of 299×299 pixels are selected. We then average the prediction results over all five cropped images.

3.4. Experimental setups and training

Employing the proposed region enhance mechanism, the Mt-Net and the Sn-Net work collaboratively, which means performance of one network relies partially on another network. We propose a cross training procedure, as presented in Algorithm 1. The Mt-Net and the Sn-Net are trained alternately, one at a time, and we use the cross-entropy function as the cost function in the training process:

$$J = -\sum_{i=1}^{N} y_i \log(a_i) \tag{5}$$

where N denotes the number of elements in the model prediction. a_i denotes the ith element of the model prediction and y_i represents the ith element of the label.

In the training of the Mt-Net, since the two networks are in a cascade manner, the Mt-Net has no visualization inputs at the beginning. To solve the problem, two ways of training are proposed. In each step, one of them is picked randomly to train the Mt-Net.

We first sample a probability p from uniform distribution to pick one way. If $p < 0.5$, the Mt-Net is trained without visualization inputs. x_{cam}^{Mt} is set to zeros and x_{img}^{Mt} is set to one mini-batch of breast ultrasonography images to train the Mt-Net. The corresponding labels are employed to calculate cross-entropy cost, and parameters of the Mt-Net are updated by gradient descending. If $p \geq 0.5$, the Mt-Net is trained with visualization inputs. The first step is to set x_{cam}^{Mt} to zeros and x_{img}^{Mt} to one mini-batch of breast ultrasonography images, and perform forward computation of the Mt-Net, generating visualization results as inputs for the Sn-Net. In

Algorithm 1: Cross training.

begin
  for *number of training iterations* **do**
    for *number of mini-batches* **do**
      Sample p from uniform distribution.
      if $0 < p < 0.5$ **then**
        $x_{cam}^{Mt} \leftarrow 0, x_{img}^{Mt} \leftarrow$ mini-batch;
        Update parameters of the Mt-Net by gradient descending.
      end
      else
        $x_{cam}^{Mt} \leftarrow 0, x_{img}^{Mt} \leftarrow$ mini-batch;
        Forward computation of the Mt-Net, get CAM result CAM^{Mt} to be used as the input of the Sn-Net;
        $x_{cam}^{Sn} \leftarrow CAM^{Mt}, x_{img}^{Sn} \leftarrow$ mini-batch;
        Forward computation of the Sn-Net, get CAM result CAM^{Sn} to be used as the input of the Mt-Net;
        $x_{cam}^{Mt} \leftarrow CAM^{Sn}$, $x_{img}^{Mt} \leftarrow$ mini-batch;
        Update parameters of the Mt-Net by gradient descending.
      end
      $x_{cam}^{Mt} \leftarrow 0, x_{img}^{Mt} \leftarrow$ mini-batch;
      Forward computation of the Mt-Net, get CAM result CAM^{Mt} to be used as the input of the Sn-Net;
      $x_{cam}^{Sn} \leftarrow CAM^{Mt}, x_{img}^{Sn} \leftarrow$ mini-batch;
      Update parameters of the Sn-Net by gradient descending.
    end
  end
end

the second step, the Sn-Net takes the ultrasonography images and visualization results of the Mt-Net as inputs x_{img}^{Sn} and x_{cam}^{Sn}, respectively, generating visualization results as inputs for the Mt-Net. In the third step, we use the visualizations as x_{cam}^{Mt}, along with the ultrasonography images x_{img}^{Mt} to train the Mt-Net.

In the training of the Sn-Net, we set the x_{cam}^{Mt} for the Mt-Net to zeros, the visualization results are then used as x_{cam}^{Sn}, which is fed into the Sn-Net. The class activation maps x_{cam}^{Sn} and breast ultrasonography images x_{img}^{Sn} are then used as inputs to train the Sn-Net.

The cross training algorithm is designed to satisfy the needs of real-world applications. Since the Mt-Net is trained with two ways randomly, it works properly even with no visualization inputs, resulting in a significant reduction in computation time cost. In real-world applications, the region enhance mechanism could be turned off to save computing resources. While in applications requiring high accuracy and sensitivity, we could turn the region enhance mechanism on to get better performance.

Previous studies indicate that the parameters obtained by pre-training on a large dataset can be transferred to another application trained on a different dataset (Yosinski et al., 2014). Both the networks are pre-trained on the ImageNet dataset which contains 1.2×10^6 training images, 5×10^4 validation images and 10^5 testing images. After pre-training, the networks are fine-tuned on our own dataset with ADADELTA (Zeiler, 2012) as the optimizer. Dropout (Hinton et al., 2012) is applied to the last convolutional layers of the Mt-Net and the Sn-Net with a keep probability of 0.8. To reduce the side effects of over-fitting, we apply L2 regularization with λ of 10^{-4}. The mini-batch size is fixed at 10. The

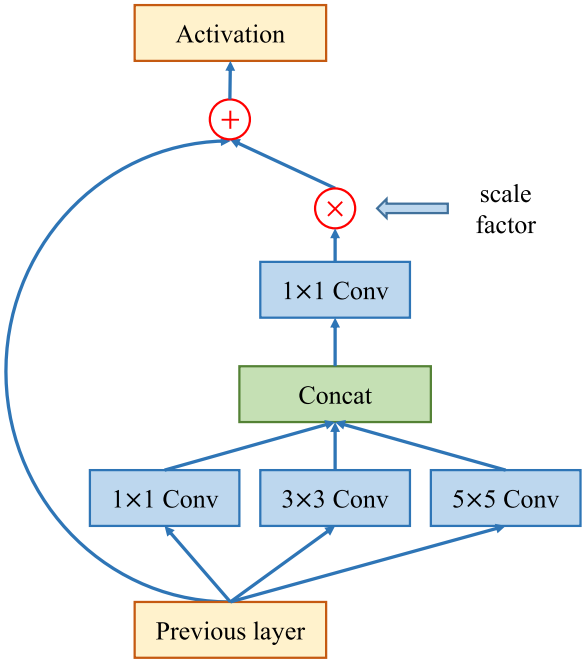

Fig. 2. An Inception module used in the Mt-Net, with a skip connection. There are three sizes of convolutional kernels in the module, 1×1, 3×3, and 5×5. The residual connection is scaled down with a factor between 0.1 and 0.2.

experiments are implemented using MXNet (Chen et al., 2015), an open-source scalable deep learning framework. It takes around 30 hours to fine-tune the Mt-Net and 10 hours to fine-tune the Nt-Net for each experiment, using a workstation with a NVIDIA Tesla K40m GPU.

4. Materials

To develop the automated breast cancer diagnosis model, we construct a large dataset of breast ultrasonography images. All of the images are labeled referring to the reports by sonographers and pathological records under the supervision of breast cancer surgeons. After data annotation, the labeled images are partitioned into training set, validation set and testing set.

4.1. Image collection

The images of breast mass are screened by different kinds of color Doppler instruments including Philips iU22, ATL HDI5000 and GE LOGIQ E9. All of the ultrasonography images are obtained from the Department of Galactophore Surgery and Department of Oncology of West China Hospital, Sichuan University. Over 8000 images from 2047 patients from October 2014 to August 2017 are collected. For each image, we collect the ultrasound examination record and pathological report, images without pathological reports are not used in this study. If the patient took other examinations or surgical operations, corresponding reports such as immunohistochemical analysis report and operation note are also collected. Fig. 5 shows examples of ultrasonography images.

4.2. Data annotation

The ultrasonography images are labelled in two phrases. First, a malignant dataset (Mt-Set) is constructed to determine whether an examined image contains malignant changes, which is the most meaningful in clinical applications. All images are classified into

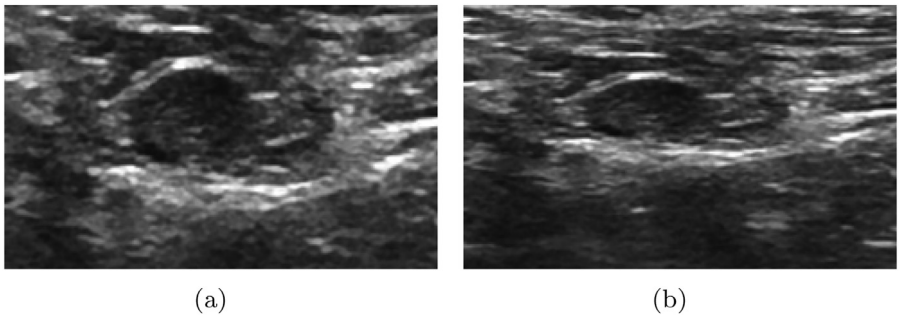

Fig. 3. Examples of ultrasonography images captured with different level of pressure. (a) Low pressure. (b) High pressure.

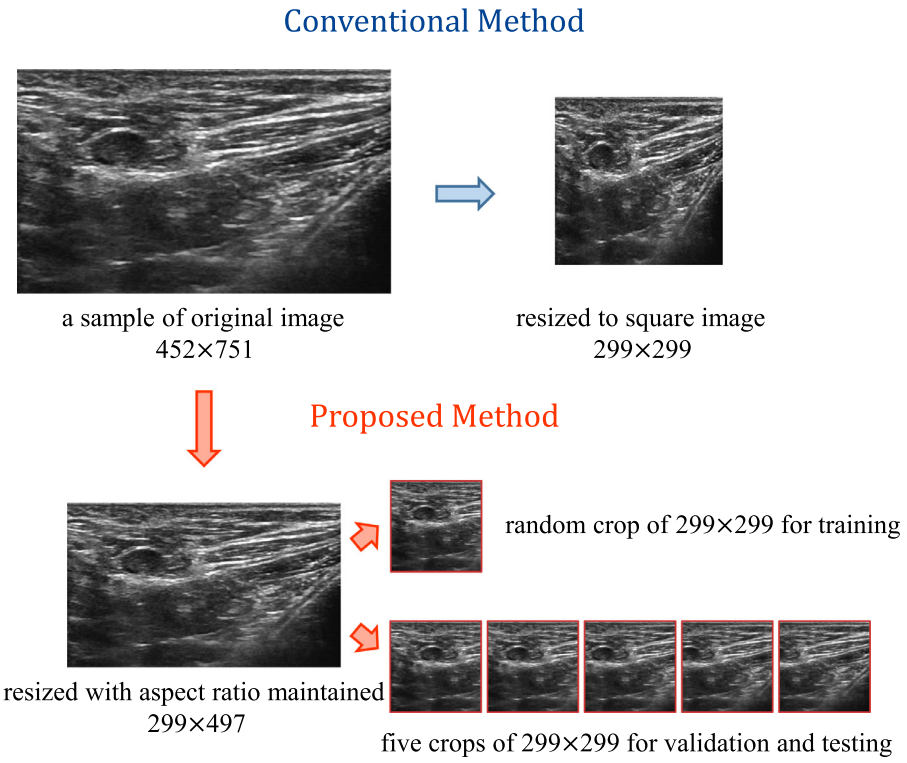

Fig. 4. The proposed multi-input prediction method. During the validation and testing steps, five crops are selected and the predictions are averaged.

Table 3
The clinical descriptions of labels.

Labels	Clinical descriptions
Malignant	Breast cancer detected
Non-malignant	No Breast cancer detected
With solid nodules	Tumors or calcification detected
Without solid nodules	No solid nodules are detected

two classes: non-malignant or malignant. The labels are determined on the basis of the Breast Imaging Reporting and Data System (BI-RADS) (Liberman et al., 1998), which is widely used in breast cancer imaging examinations. In BI-RADS, breast ailments can be divided into six grades: BI-RADS I, II, and III represent non-malignant, while BI-RADS IV, V, and VI indicate different degrees of malignant. Second, another dataset (Sn-Set) is employed to train a model recognizing solid nodules. All samples in the Mt-Set are further labeled as including or not including solid nodules. For example, images of tumors such as fibroadenoma and lipoma are considered as including solid nodules, while images of cysts and ductal ectasia are considered as not including solid nodules.

Table 3 shows the clinical features of the labels. Each ultrasonography image is annotated by two annotators individually, according to the pathological diagnosis and available reports. If there are inconsistencies between the two annotators, the annotations are then evaluated by a professional clinical breast cancer surgeon.

For the Mt-Set, there are 8145 images in total, 2759 images are malignant and 5386 are non-malignant. The Mt-Set shows a data imbalance where most of the images are non-malignant. This is consistent with clinical scenarios because there are fewer patients suffering from breast cancer than those with other ailments. For the Sn-Set, all malignant images and 2713 of the 5386 non-malignant images are labeled as including solid nodules, the remaining 2673 images are not.

4.3. Data partition

Each dataset is split into training set, validation set and testing set. 4/6 samples of the dataset are used for training, 1/6 for validation and 1/6 for testing. Images of each class are uniformly distributed in each subset. In each dataset, 5429, 1357, and 1359 samples are used for training, validation and testing, respectively.

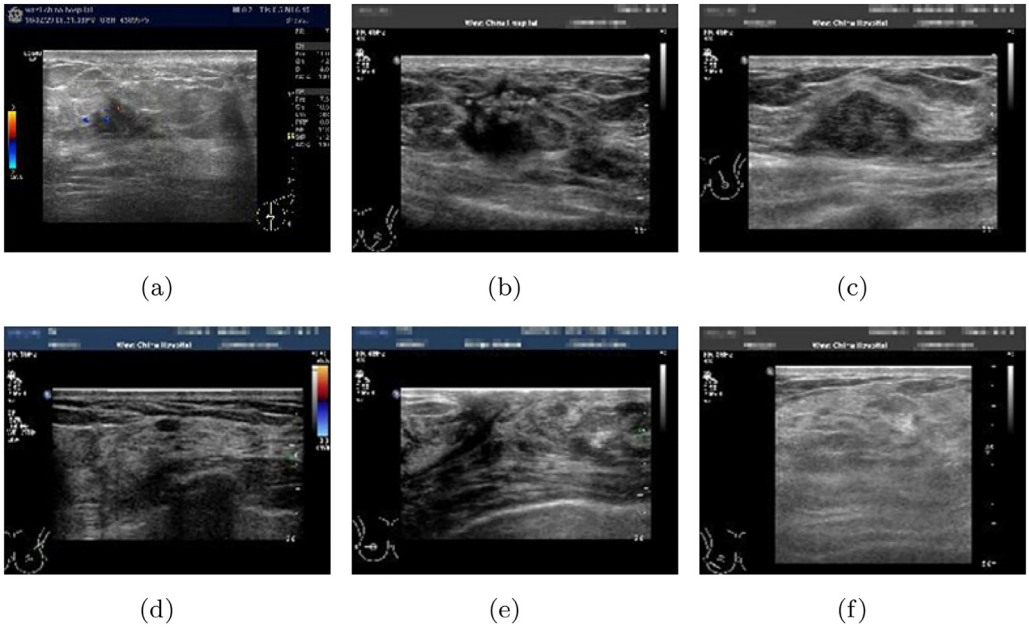

Fig. 5. Examples of ultrasonography images captured by colour Doppler instruments. (a) Malignant tumor. (b) Malignant tumor, note that the shape and size of this tumor are far away from those in (a). (c) Benign tumor. (d) Cyst, which is usually a small, ellipse sac filled with liquid. (e) Ductal ectasia. (f) Normal, means no breast sicknesses are detected.

5. Results

5.1. Evaluation criteria

In this study, we use the accuracy, the F-β score (Yates and Neto, 1999), and the area under the receiver operating characteristic curve (henceforth referred to as AUC) as evaluation criteria. We also determine the sensitivity (also called the true positive rate or recall) and the specificity (also called the true negative rate).

In breast cancer examinations, the most important criteria is the sensitivity, meaning that the top priority is reducing the number of missed diagnoses. We thus use the F-β score instead of the F-1 score to assign greater importance to sensitivity than to precision. We set β to 2.0 for both the Mt-Set and the Sn-Set according to the data distribution. All experimental results are presented using the model which achieves the highest F-β score on the validation set.

The criteria are defined as follows:

$$\begin{aligned} \text{accuracy} &= \frac{TP+TN}{TP+FP+FN+TN} \\ \text{sensitivity} &= \frac{TP}{TP+FN} \\ \text{precision} &= \frac{TP}{TP+FP} \\ \text{specificity} &= \frac{TN}{TN+FP} \\ \text{F-}\beta\ \text{score} &= \frac{\beta^2+1}{\frac{\beta^2}{\text{sensitivity}}+\frac{1}{\text{precision}}} \end{aligned} \tag{6}$$

where *TP, FP, FN, TN* are the number of true positives, false positives, false negatives and true negatives, respectively.

Fig. 6 shows the confusion matrix for the subtask to distinguish images containing malignant tumors from non-malignant images. We define malignant as “positive” and non-malignant as “negative”. For the other subtask, which recognizes solid nodules, the samples with solid nodules are defined as “positive” and those without are defined as “negative”.

		Predicted class	
		Malignant	Non-malignant
True class	Malignant	True Positive	False Negative
	Non-malignant	False Positive	True Negative

Fig. 6. Confusion matrix of the subtask to distinguish images containing malignant tumors from non-malignant images.

Table 4
Experimental results on the Mt-Set, validation set.

Method	F-β	Accuracy	Sensitivity	Specificity
Mt-Net(BASIC)	0.878	93.07%	86.74%	**96.32%**
Mt-Net(MIP)	0.916	92.85%	92.83%	92.87%
Mt-Net(REM)	**0.931**	**94.47%**	**93.91%**	94.76%
Han et al.	0.888	92.26%	88.91%	93.98%
Cheng et al.	0.639	78.92%	62.17%	87.51%

Table 5
Experimental results on the Mt-Set, test set.

Method	F-β	Accuracy	Sensitivity	Specificity
Mt-Net(BASIC)	0.885	93.52%	87.39%	**96.66%**
Mt-Net(MIP)	0.920	93.89%	92.61%	94.55%
Mt-Net(REM)	**0.942**	**94.48%**	**95.65%**	93.88%
Han et al.	0.905	93.08%	90.87%	94.22%
Cheng et al.	0.645	79.54%	62.61%	88.21%

5.2. Experimental results

5.2.1. Results of the basic networks

We introduce the experimental results of the basic networks first, region enhance mechanism and multi-input prediction are not employed. For convenience, the basic networks are called the Mt-Net(BASIC) and the Sn-Net(BASIC). For the subtask to distinguish malignant samples from non-malignant samples, the Mt-Net(BASIC) achieves comparable performance to human sonographers. As presented in the first lines of Tables 4 and 5, the pro-

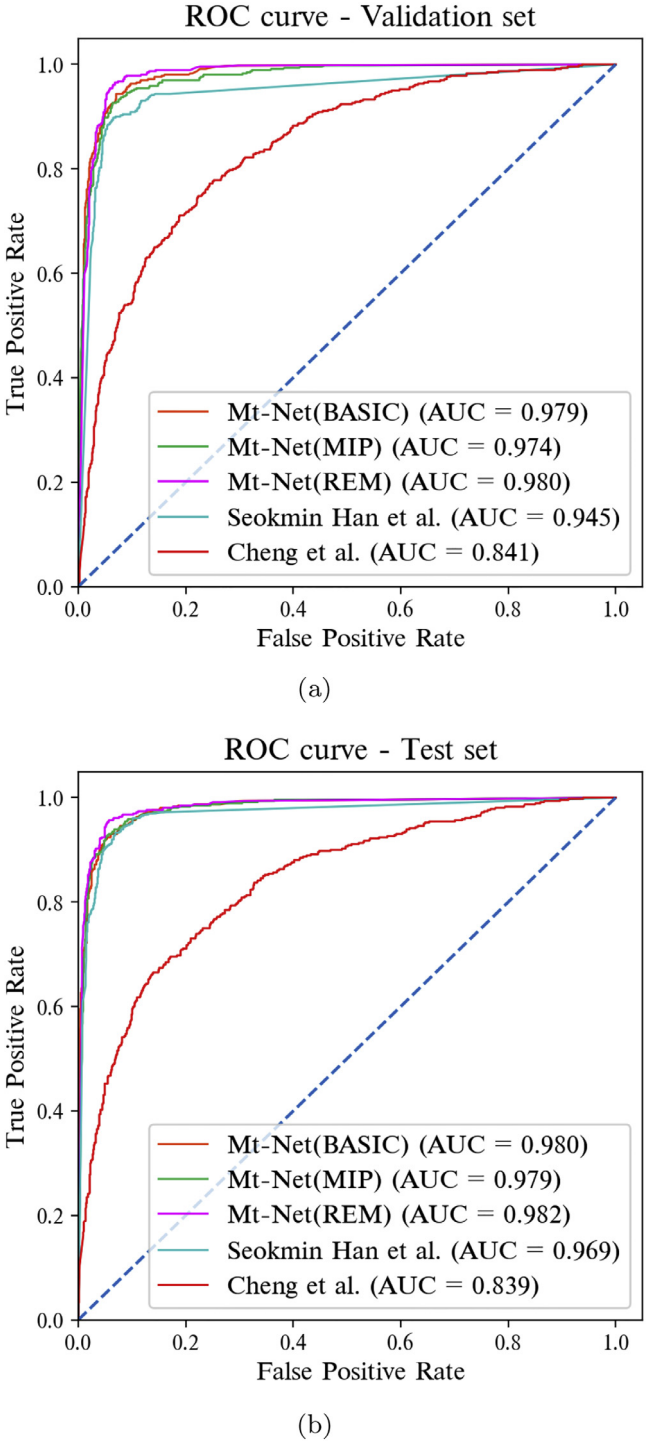

Fig. 7. The receiver operating characteristic curve for the Mt-Net. (a) ROC curve on the validation set. (b) ROC curve on the test set.

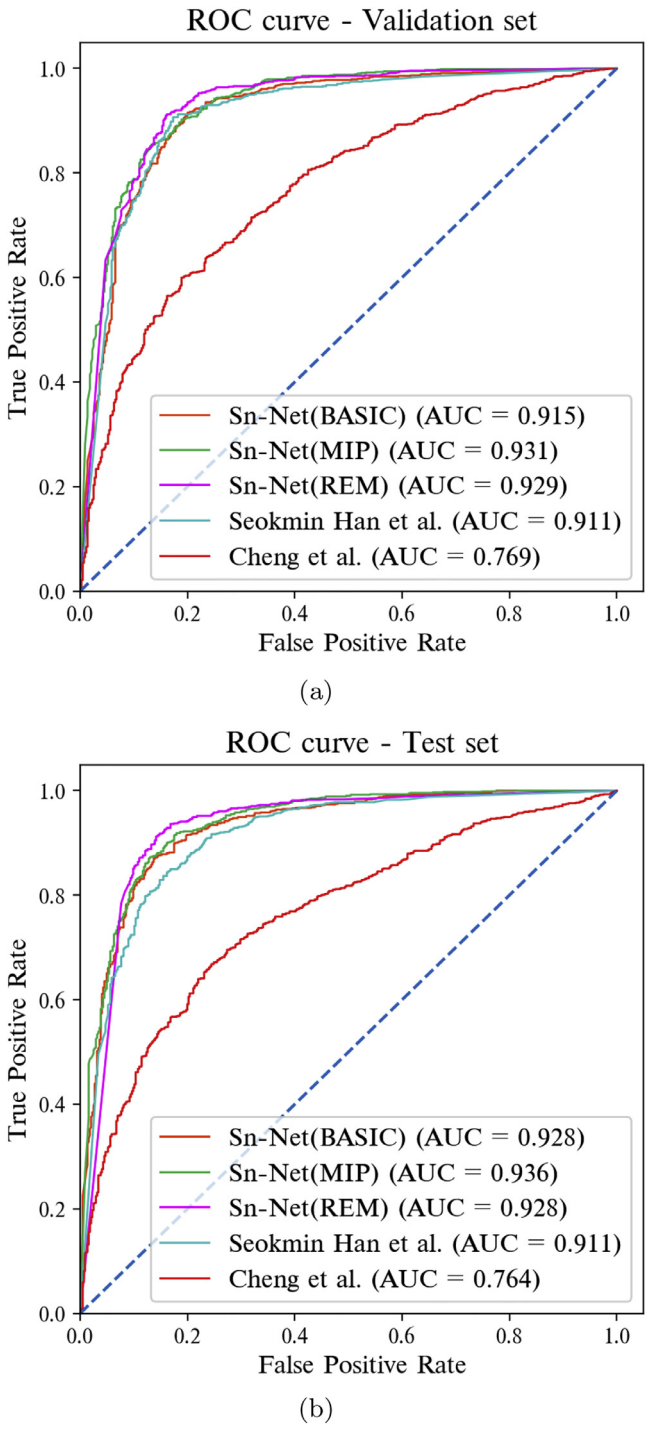

Fig. 8. The receiver operating characteristic curve for the Sn-Net. (a) ROC curve on the validation set. (b) ROC curve on the test set.

posed network achieved a F-β score of 0.878, an accuracy of 93.07%, with a sensitivity of 86.74% and a specificity of 96.32% on the validation set. The model with the highest F-β score is tested on the test set, and the Mt-Net(BASIC) achieved a F-β score of 0.885 and an accuracy of 93.52%. The Sn-Net(BASIC) achieved a F-β score of 0.916 and an accuracy of 87.34% on the test set, as presented in Tables 6 and 7. The sensitivity and the specificity are 92.22% and 77.35% respectively.

The proposed Mt-Net(BASIC) achieves a high accuracy and specificity, which is applicable in clinical scenarios. The sensitiv-

Table 6
Experimental results on the Sn-Set, validation set.

Method	F-β	Accuracy	Sensitivity	Specificity
Sn-Net(BASIC)	0.920	87.62%	92.76%	77.08%
Sn-Net(MIP)	0.930	87.91%	94.19%	75.06%
Sn-Net(REM)	**0.941**	**89.39%**	**95.29%**	77.30%
Han et al.	0.907	87.99%	90.57%	**82.70%**
Cheng et al.	0.819	73.03%	82.79%	53.03%

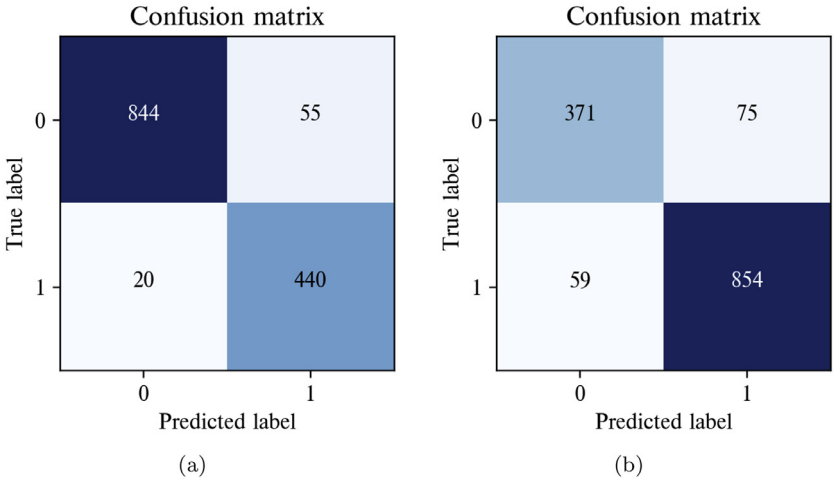

Fig. 9. Confusion matrices of two networks. (a) Mt-Net trained on the Mt-Set. (b) Sn-Net trained on the Sn-Set.

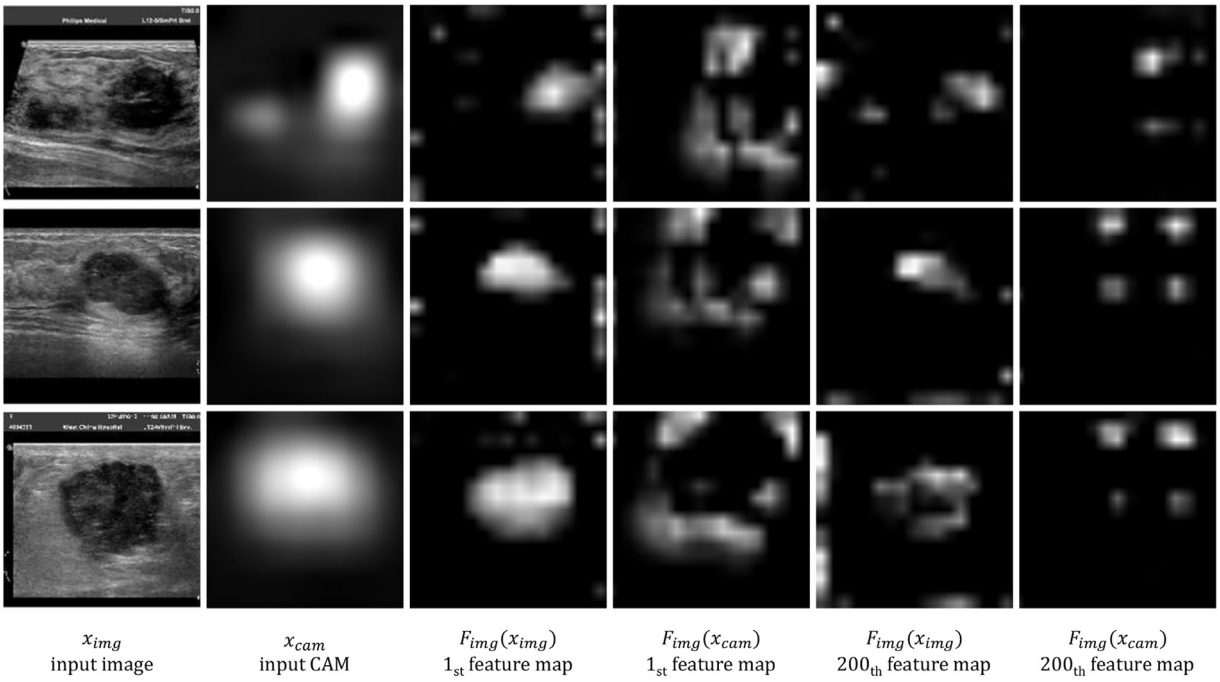

Fig. 10. The high level features of the Mt-Net to perform the region enhance mechanism. 1st column: The input ultrasonography images; 2nd column: The input CAM visualization results; 3rd column: Features extracted from the ultrasonography images, the 1st feature map; 4th column: Features extracted from the CAM visualizations, the 1st feature map; 5th column: Features extracted from the ultrasonography images, the 200th feature map; 6th column: features extracted from the CAM visualizations, the 200th feature map.

Table 7
Experimental results on the Sn-Set, test set.

Method	F-β	Accuracy	Sensitivity	Specificity
Sn-Net(BASIC)	0.916	87.34%	92.22%	77.35%
Sn-Net(MIP)	0.924	87.78%	93.31%	76.45%
Sn-Net(REM)	**0.932**	**90.13%**	**93.54%**	**83.18%**
Han et al.	0.887	85.28%	88.61%	78.47%
Cheng et al.	0.805	71.74%	81.16%	52.47%

ity is relatively low, as the samples of different classes are highly imbalanced in the dataset. Accuracy of the Sn-Net(BASIC) is a bit lower, misclassified samples show that a few images with giant cysts and inflammatory tissues are classified as solid nodules. The receiver operating characteristic curves of the basic networks are shown in Figs. 7 and 8. The AUC is 0.980 for the Mt-Net(BASIC) and 0.928 for the Sn-Net(BASIC).

5.2.2. The effectiveness of multi-input prediction

We than evaluate the effectiveness of the proposed multi-input prediction method. These models are called the Mt-Net(MIP) and the Sn-Net(MIP). As presented in the second lines of Tables 4–7, experimental results show a boost in sensitivity, which results in a significant progress in the F-β score. The accuracy and AUC are almost unchanged, however, the improvement in sensitivity is most meaningful in breast cancer diagnosis.

5.2.3. The effectiveness of region enhance mechanism

The proposed region enhance mechanism is employed based on the networks with multi-input prediction. The model is trained

following the algorithm presented as Algorithm 1, we call them the Mt-Net(REM) and the Sn-Net(REM). Both the networks achieve better performance than other approaches. The third lines of Tables 4–7 show the highest F-β scores, accuracies and sensitivities. The Mt-Net(REM) achieves an accuracy of 94.48% and a sensitivity of 95.65% on the test set, and the accuracy and sensitivity of the Sn-Net(REM) are 90.13% and 93.54%. The ROC curves are shown in Figs. 7 and 8, the AUC is 0.982 for the Mt-Net(REM) and 0.928 for the Sn-Net(REM). The confusion matrices on the test set are presented in Fig. 9.

For the Mt-Net, high level features of the two inputs are aggregated after 30 Inception blocks, on feature maps of size $1088 \times 17 \times 17$. To analyse the effectiveness of region enhance mechanism, we show the 1st and 200th feature maps in Fig. 10. The feature maps are resized to 299×299 pixels, same as the input images, and normalized to [0, 255]. On one hand, as the 3rd and 5th columns show, the features extracted from input images $F_{img}(x_{img})$ focus on the lesions in ultrasonography images. On the other hand, as the 4th and 6th columns show, the features from CAM visualizations $F_{cam}(x_{cam})$ focus on the surroundings of the tumors. Since the shape, orientation and margin are important characteristics in breast cancer diagnosis, a possible explanation is that this kind of features could work as supplementary features to enhance the regions around lesions.

5.2.4. Comparison with other methods

Tables 4–7 show a comparison of the proposed model with two recent studies proposed by Cheng et al. (2016) and Han et al. (2017), which achieved better performance than traditional hand-crafted features. Cheng et al. proposed a CADx based on stacked denoising autoencoders for the classification of breast ultrasonography images. The method was compared with the RANK algorithm (Yang et al., 2013) with well-known GLCM-based texture features, which is the state-of-the-art conventional algorithm, and showed significant performance boost. Han et al. employed GoogLeNet to classify malignant and benign lesions, following a standard deep learning procedure. The trained model has been embedded in Samsung medison ultrasound instruments, registered as S-Detect™, which aims at helping standardize reporting and classification of suspicious breast lesions.

To implement the method proposed by Cheng et al., we train a stacked denoising autoencoder with input size 28×28, as presented in the paper. Since the input size is smaller than the image size, for each image in the Mt-Set and the Sn-Set, a 60% patch containing the lesions is cropped from the original image and resized to 28×28 pixels. The SDAE is then trained with the patches layer by layer. Optimal parameters are chosen based on 10 individual experiments, and we find that a 7 layer SDAE performs best. After the training of the autoencoder, two extra output neurons are added on the top of the network, and the scaling factors in the *height* and *width* dimensions and aspect ratio are used as extra inputs, which follows the method proposed in the paper exactly.

To implement the method proposed by Han et al., we trained a deep neural network based on the GoogLeNet architecture on our datasets. The input ultrasonography images are resized to 255×255 and are converted to gray images. Two auxiliary classifiers of the default GoogLeNet are removed, and the last layer is replaced by a fully connected layer with two output neurons.

Experimental results show that our model outperforms both the methods described above. For both the subtasks to identify malignant tumors and recognize solid nodules, there is a large margin between the method proposed by Cheng et al. and our models. The test accuracies of our models are 15% higher than those of Cheng et al., and the F-β scores and AUC values are also considerably higher. There are several possible explanations accounting for the superior performance of our method: (1) The method of Cheng et al. takes a region of interest of size 28×28 as input, while the input image size of our method is 299×299 pixels. There may be severe information lost in Cheng et al.'s approach; (2) Deep convolution neural networks are more powerful than fully connected autoencoders in computer vision for feature extraction; (3) Instead of pre-training as an autoencoder, the parameters of our models

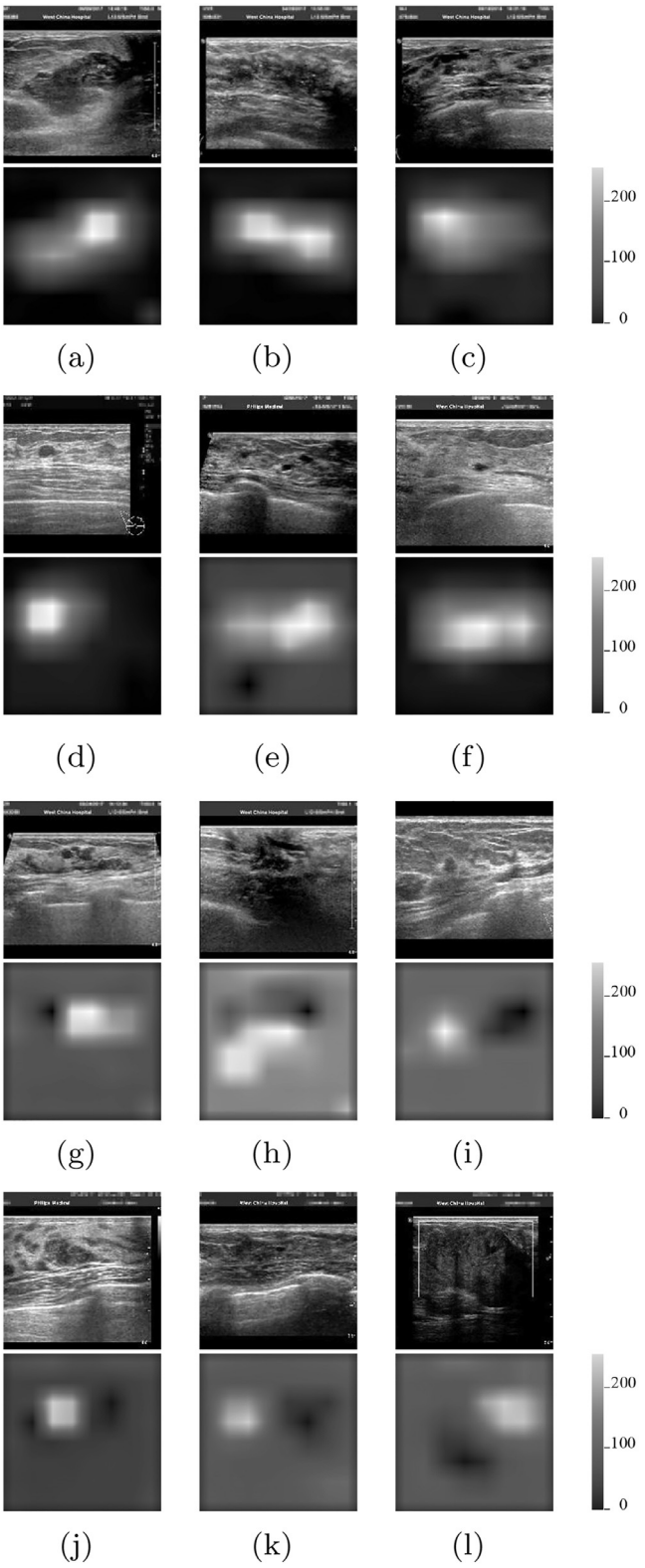

Fig. 11. Visualization examples of the Mt-Net using the CAM algorithm, each row contains three pairs of input images and the correlated regions that the network focuses on.

are pre-trained on the large ImageNet dataset. The transfer learning method is helpful in training very deep networks with inadequate data.

Compared with the method proposed by Han et al., our BASIC models show similar performance. For the Mt-Net(BASIC), the accuracy is about 0.5% higher, but the sensitivity is a bit lower. For the Sn-Net(BASIC), the accuracy and sensitivity are higher, but the specificity is 0.9% lower. Due to the powerful convolutional layers, both Han et al.'s and our methods outperform the method proposed by Cheng et al. Our REM models exhibit superior performance on most metrics, except the specificity. This is attributed to the multi-input prediction and region enhance mechanism proposed in this study.

5.2.5. Visualization

Visualization results of the Mt-Net are presented in Fig. 11. Each row consists of three pairs of input image and corresponding visualization.

The input images in the first and second row of Fig. 11 are from true positive and true negative examinations. Comparing the input images and the corresponding visualization results, the maximum output neuron in the softmax classifier is highly correlated with the lesion area in input image. This is consistent with the most concerned areas of radiologists and pathologists, indicating that the proposed model is capable of extracting the essential features from the input for the diagnosis of breast cancer. In illegible cases such as Fig. 11(f), cyst and ductal ectasia are presented simultaneously, and the proposed model focuses on the areas covering both of the ailments.

The false negative cases can be seen in the third row. As shown in Fig. 11(h), the correlated region by visualization deviates from the actual lesion area, leading to an incorrect prediction. Compared with the results of the true positive and true negative examinations, the model focuses more on regions without disease, which may interfere with the diagnosis of breast cancer.

The false positive cases are shown in the forth row. In the validation set and the test set, most false positive cases are atypical samples which present similar features to malignant samples. Although the predictions are wrong, the model focuses on valuable features in the input images. In Fig. 11(j), the correlated region matches the tumor even though the tumor is benign. In Fig. 11(k), the model focuses on the disordered echo beside the cyst, which is important in breast cancer diagnosis. In Fig. 11(l), the model focuses on the blood flow signals, and the blood flow signals are highly related to malignant changes.

6. Conclusion

Automated ultrasonography image diagnosis can improve the efficiency and reliability of breast cancer screening and guide pathological examination. In this paper, we demonstrate that deep neural networks can be used in ultrasonography image classification for both malignant tumors and solid nodules. We propose a method to diagnose breast ultrasonography images using deep convolutional neural networks with multi-scale kernels and skip connections. Two networks, the Mt-Net and the Sn-Net, are proposed to identify malignant tumors and recognize solid nodules in a cascade manner. In order to let the two networks work in a collaborative way, a region enhance mechanism based on class activation maps is proposed, along with a cross training algorithm. The mechanism helps to improve classification accuracy and sensitivity for both networks. We construct a large-scale annotated ultrasonography breast image dataset to train and evaluate the models. The proposed method is compared with two state-of-the-art approaches, and outperforms both of them by a large margin.

Visualization results demonstrate that the model learned to locate and recognize lesion areas. Our proposed method showed robustness to the real-world dataset, rendering it feasible for clinical applications. Our future research will focus on the following: (1) Utilizing deep neural networks to determine whether ultrasonography screening is helpful in guiding clinical operations, such as the prediction of infiltration, necrosis, and nuclear grade; (2) further enlarging the dataset during the application in real-world scenarios; and (3) combining different image modalities, potentially rendering the diagnosis more accurate.

Acknowledgment

This work was supported by the National Natural Science Foundation of China [Grant 61432012 and U1435213].

References

Byra, M., Piotrzkowska-Wrblewska, H., Dobruch-Sobczak, K., Nowicki, A., 2017. Combining Nakagami imaging and convolutional neural network for breast lesion classification. In: 2017 IEEE International Ultrasonics Symposium (IUS), pp. 1–4. doi:10.1109/ultsym.2017.8092154.

Chen, D.R., Chang, R.F., Chen, C.J., Ho, M.F., Kuo, S.J., Chen, S.T., Hung, S.J., Moon, W.K., 2005. Classification of breast ultrasound images using fractal feature. Clin. Imag. 29 (4), 235–245. doi:10.1016/j.clinimag.2004.11.024.

Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., Zhang, Z., 2015. Mxnet: a flexible and efficient machine learning library for heterogeneous distributed systems. arXiv:1512.01274.

Cheng, J.Z., Ni, D., Chou, Y.H., Qin, J., Tiu, C.M., Chang, Y.C., Huang, C.S., Shen, D., Chen, C.M., 2016. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans. Sci. Rep. 6, 24454. doi:10.1038/srep24454.

Fitzmaurice, C., Dicker, D., Pain, A., Hamavid, H., Moradilakeh, M., Macintyre, M.F., Allen, C., Hansen, G., Woodbrook, R., Wolfe, C., 2015. The global burden of cancer 2013. JAMA Oncol. 1 (4), 505–527. doi:10.1001/jamaoncol.2015.0735.

Giger, M.L., Karssemeijer, N., Schnabel, J.A., 2013. Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. Annu. Rev. Biomed. Eng. 15, 327–357. doi:10.1146/annurev-bioeng-071812-152416.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587. doi:10.1109/cvpr.2014.81.

Han, S., Kang, H.-K., Jeong, J.-Y., Park, M.-H., Kim, W., Bang, W.-C., Seong, Y.-K., 2017. A deep learning framework for supporting the classification of breast lesions in ultrasound images. Phys. Med. Biol. 62, 7714–7728. doi:10.1088/1361-6560/aa82ec.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. Comput. Sci. 3 (4), 212–223.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456.

Jalalian, A., Mashohor, S., Mahmud, R., Saripan, M.I., Ramli, A., Karasfi, B., 2012. Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. Clin. Imag. 37 (3), 420–426. doi:10.1016/j.clinimag.2012.09.024.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: International Conference on Neural Information Processing Systems, pp. 1097–1105. doi:10.1145/3065386.

Kuo, S.J., Hsiao, Y.H., Huang, Y.L., Chen, D.R., 2008. Classification of benign and malignant breast tumors using neural networks and three-dimensional power doppler ultrasound. Ultrasound Obstet. Gynecol. 32 (1), 97–102. doi:10.1002/uog.4103.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324. doi:10.1109/5.726791.

Lee, C.H., 2002. Screening mammography: proven benefit, continued controversy. Radiol. Clin. North Am. 40 (3), 395–407. doi:10.1016/s0033-8389(01)00015-x.

Leong, A.S.Y., Zhuang, Z., 2011. The changing role of pathology in breast cancer diagnosis and treatment. Pathobiology 78 (2), 99–114. doi:10.1159/000292644.

Liberman, L., Abramson, A.F., Squires, F.B., Glassman, J., Morris, E., Dershaw, D., 1998. The breast imaging reporting and data system: positive predictive value of mammographic features and final assessment categories. AJR Am. J. Roentgenol. 171 (1), 35–40. doi:10.2214/ajr.171.1.9648759.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241. doi:10.1007/978-3-319-24574-4_28.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252. doi:10.1007/s11263-015-0816-y.

Shelhamer, E., Long, J., Darrell, T., 2017. Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (4), 640–651. doi:10.1109/tpami.2016.2572683.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Szegedy, C., Ioffe, S., Vanhoucke, V., 2016. Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI Conference on Artificial Intelligence.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9. doi:10.1109/cvpr.2015.7298594.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition doi:10.1109/cvpr.2016.308.

Tianur, Nugroho, H.A., Sahar, M., Ardiyanto, I., Indrastuti, R., Choridah, L., 2017. Classification of breast ultrasound images based on posterior feature. In: International Conference on Information Technology Systems and Innovation, pp. 1–4. doi:10.1109/icitsi.2016.7858239.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., 2010. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. 11, 3371–3408. http://dl.acm.org/citation.cfm?id=1756006.1953039.

Yang, M.C., Moon, W.K., Wang, Y.C.F., Min, S.B., Huang, C.S., Chen, J.H., Chang, R.F., 2013. Robust texture analysis using multi-resolution gray-scale invariant features for breast sonographic tumor diagnosis. IEEE Trans. Med. Imag. 32 (12), 2262–2273.

Yap, M.H., Pons, G., MartÃ, J., Ganau, S., SentÃs, M., Zwiggelaar, R., Davison, A.K., MartÃ, R., 2018. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J. Biomed. Health Inf. doi:10.1109/jbhi.2017.2731873.

Yates, R.A.B., Neto, B.A.R., 1999. Modern Information Retrieval. ACM Press Book, Addison Wesley.

Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? In: In Proceedings of the Advances in Neural Information Processing Systems, pp. 3320–3328.

Zeiler, M.D., 2012. Adadelta: an adaptive learning rate method. Comput. Sci.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: Computer Vision and Pattern Recognition, pp. 2921–2929. doi:10.1109/cvpr.2016.319.