Multimedia Tools and Applications
https://doi.org/10.1007/s11042-021-10505-y

1182: DEEP PROCESSING OF MULTIMEDIA DATA

A computer-aided diagnostic system for mammograms based on YOLOv3

Jianhui Zhao[1] · Tianquan Chen[1] · Bo Cai[2]

Received: 3 July 2020 / Revised: 2 November 2020 / Accepted: 5 January 2021 /
Published online: 22 February 2021

Abstract
Due to a large amount of noise in medical images, the task of detecting and classifying the lesions of mammograms remains a huge challenge. Based on the existing deep learning methods, focusing on the diversity of breast cancer lesion types, this paper proposes a computer-aided diagnosis system based on YOLOv3 (You Only Look Once version 3) convolutional neural network for mammograms. In this system, we integrate detection and multi-classification problems of breast lesions into a regression problem, thereby simultaneously accomplish the two tasks in one framework. The proposed computer-aided diagnosis system is mainly divided into three components: preprocessing part of the original mammograms, deep convolutional neural network based on YOLOv3, processing and evaluation of the network output. We use the dataset from CBIS-DDSM to train three models: general model, mass model and microcalcification model. These trained models can detect the position of the input mammograms in different situations, and then classify them into mass, microcalcification, benign, malignant, and other categories. After evaluating the performance by using test set images, the accuracy rates of the general model, mass model, and microcalcification model trained by our system reach 93.667 %, 97.767 %, 96.870 % in the detection task, and 93.927 %, 98.121 %, 97.045 % in the classification task. The computer-aided diagnosis system performs well in lesion detection and classification tasks with high-noise mammograms, reflecting well robustness.

Keywords Breast cancer · Deep learning · Convolutional neural network · Object detection and classification · YOLOv3

1 Introduction

The breast cancer is the most common cancer among women in many countries or regions [7]. Also, breast cancer is the leading cause of death among women in 103 countries. The number of new breast cancer patients in US women will account for 30 % of all women's

✉ Bo Cai
caib@whu.edu.cn

[1] School of Computer Science, Wuhan University, Wuhan, Hubei, China

[2] School of Cyber Science and Engineering, Wuhan University, Wuhan, Hubei, China

new malignant tumor patients in 2020 [29]. To diagnose breast cancer, many screening methods have been presented. Mammography, breast ultrasound, and breast magnetic resonance imaging (MRI) examinations are currently the main screening methods for breast cancer [17]. In particular, mammography can detect abnormal areas that are not clinically accessible in the early stage of breast cancer. Therefore, mammography plays an indispensable role in improving the diagnosis rate of breast cancer. Usually, the radiologists browse the mammograms in a subjective visual way to find out the position of the lesion and classify it. However, errors in human eye and other factors have caused misdiagnosis and missed diagnosis, which brings challenges to radiologists and pressures to patients [11]. Therefore, computer-aided diagnosis technology is indispensable, and it can give the second opinion to doctors for reference.

In recent years, with the breakthrough in computer hardware performance and the rapid development of deep learning algorithms, the application of deep learning for mammograms classifying has become more and more widely used. However, due to the complexity of human tissues and the insufficient availability of the information which is collected by medical imaging equipment, the breast tissue image information often contains a lot of noise, which results in a relatively low signal-to-noise ratio (SNR). Using noisy mammograms can cause wrong learning direction in deep learning models' training stage and misjudgment in the testing stage. Therefore, reducing the effect of noise in mammograms will bring accurate learning direction and proper judgment, thereby improving the robustness of the model. Inspired by this insight, our research aims to further reduce the impact of mammography image noise, which can promote the performance of diagnostic task.

Based on state-of-the-art methods, this article focuses on the issues including multi-scale feature maps, lesion detection, and classification. We further propose a computer-aided diagnosis system for mammograms based on YOLOv3 [24]. Figure 1 shows the work flow chart of our computer-aided diagnosis system. It is mainly divided into three stages: preprocessing part, YOLOv3-based convolutional neural network part, and evaluation part.

The following narrative structure of this article is as follows: Section 2 mainly introduces the existing related research work; Section 3 will describe in detail each structure of the entire computer-aided diagnosis system, including dataset, image preprocessing method, neural networks' architecture and processing method for network output; in Section 4, we will give our evaluation method, and quantitatively evaluate the performance of our system through various metrics and compare with the previous methods; finally, in Section 5, we will summarize this study and propose measures that can be further improved afterward.

2 Related work

With the achievement of convolutional neural network (CNN) technology, in the field of mammogram recognition, CNN-based deep learning models have attracted the attention of many researchers, and various efficient algorithms have been continuously proposed. For the classification of mass lesions, Arevalo et al. [4], Kooi et al. [16], Sun et al. [31], Sun et al. [30] and Suzuki et al. [32] used artificially labeled suspicious lesion areas and used CNNs with different structures for feature extraction and recognition. In particular, Suzuki et al. presented the deep convolutional neural network (DCNN) with the transfer learning strategy for mass detection in mammographic images, and achieved a recall rate of 89.90 % on the DDSM [12] dataset [32]. Differently, Arfan et al. adopted CNN to extract the features of the entire image, and then used support vector machine (SVM) for classification,

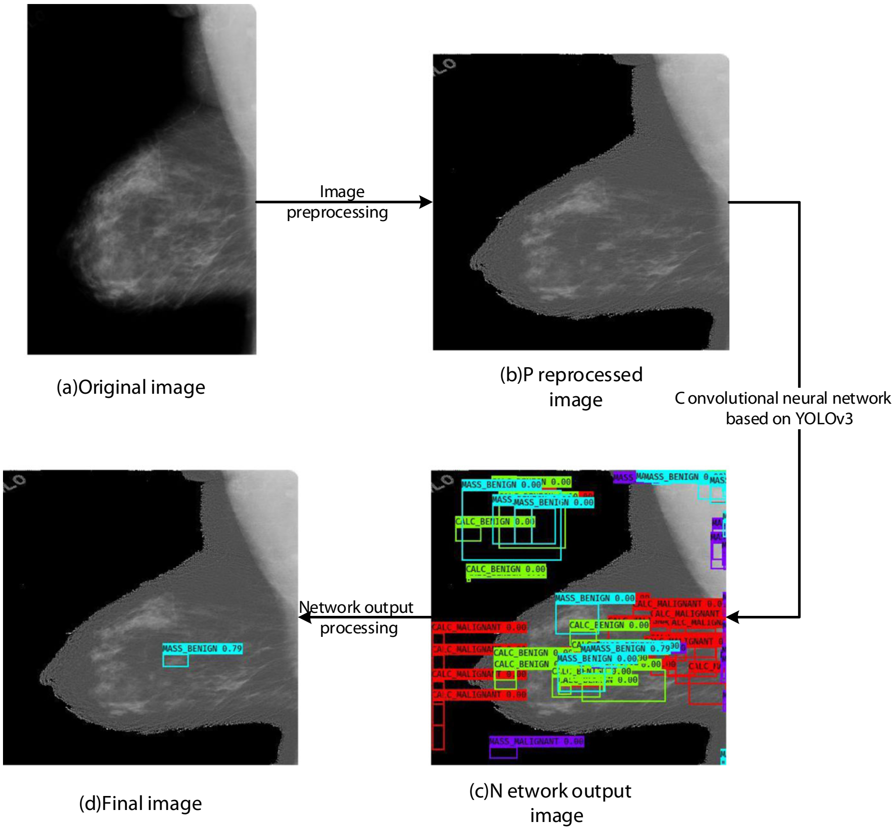

Fig. 1 Flow chart of our computer-aided diagnosis system

achieving 93 % AUC on MIAS [12] and DDSM dataset [5]. Mordang et al. [20] and Bria et al. [8] both paid attention to the classification of microcalcification lesions. For that, they used different preprocessing methods and CNN structures leading to different results. Mordang et al. applied a hard negative mining strategy, which helps overcome the large class imbalance between pixels belonging to microcalcifications and other breast tissue [20]. Bria et al. proposed a preprocessing algorithm for defogging images, achieving a recall rate of 76.26 % [8]. There are also some researchers who ignored the type of lesion and focused on whether the lesion is benign or malignant. For example, Omonigho et al. [22] utilized a DCNN based on AlexNet [15] to extract and classify the mammograms of the MIAS dataset into two classes of benign (normal) and malignant (abnormal) tumors. With augmentation techniques for improving classification accuracy, the system finally obtained an accuracy rate of 95.70 %.

The above methods either directly input the preprocessed whole image, which will be doped with a lot of noise and affect the performance of the classifiers, or simply use the cropped lesion area, which is extremely dependent on manual annotation information. To reduce the effects of manual annotation information, Ben-ari et al. paid attention to the detection of the lesions. They provided a new R-CNN method by using a pretrained network on a candidate region guided by clinical observations, to detect and classify lesions in the DDSM dataset [6]. Based on the characteristics of mammograms with multiple views, Ma et al. used faster-RCNN [26] based method, termed Cross-View Relation Region-based Convolutional Neural Networks (CVR-RCNN), to detect and classify the lesions from two paired views, and achieved an F1 score of 73 % [19]. Sarath et al. proposed a two-stage Multi-Instance Learning (MIL) framework with the first stage for extracting local candidate patches in the mammograms and the second stage for classifying an image level benign vs. malignant mass, and achieved an accuracy of 76 % in the detection task and an AUC of 0.91 in the classification task on the INbreast dataset [27]. Jung et al. adopted the FaceBook AI team's RetinaNet [18] as the deep learning network to train for mammogram lesions detection and classification tasks, and achieved comparable or better performance [13] on the INbreast dataset [21]. Al-Masni et al. [1] and Platania et al. [23] also paid attention to the detection of mammogram lesions. They used YOLOv1 [25] algorithm to achieve the two tasks of detecting and classifying breast mass lesions in the same framework. In particular, Al-Masni et al. achieved 96.33 % detection accuracy rate and 85.52 % classification accuracy rate on the subset of CBIS-DDSM [1]. Lately, they utilized data augmentation to further improve the detection accuracy rate of breast lesions to 99.7 % and the classification accuracy rate to 97 % [2]. The framework of Platania et al. achieved a detection accuracy of up to 90 % and a classification accuracy of 93.5 % (AUC of 92.315 %) [23]. However, the methods still have the following shortcomings. First, the recognition accuracy of potential small lesions is relatively low. Second, these methods only identify breast mass lesions, but there are other types of mammogram lesions such as microcalcification. These problems have thus determined the limitation of this kind of method.

In order to solve the limitations mentioned above, while paying attention to the detection and classification of mammograms, we further notice the problems of various types of lesions and small-sized lesions and propose a YOLOv3-based computer-aided diagnosis system for mammograms. Our detailed contributions are summarized as follows. According to the types of lesions in the mammograms (mass and microcalcification), we train three models using the mammograms from CBIS-DDSM dataset [28]: the general model trained using all images, the mass model trained by mass images only and the microcalcification model trained using microcalcification images only. The computer-aided diagnosis system we proposed can learn the entire image in one network architecture to achieve two tasks including detecting the positions of the lesions and classifying the lesions simultaneously. Compared with other state-of-the-art methods, we have enhanced the ability to detect small-sized lesions and paid attention to the diversity of lesion types, so that our system has better performance and can handle more tasks.

3 Computer-aided diagnosis system based on YOLOv3

3.1 Dataset

DDSM is a mammogram dataset maintained by the University of South Florida in 1997 [12]. It contains mammograms from 2620 patients. Each patient generally has four images from two views including the mediolateral oblique (MLO) view and the craniocaudal (CC) view of the left and right breasts. To use the DDSM dataset in a standardized manner, the TCIA website [9] collated and obtained the CBIS-DDSM dataset [28]. This is an updated standardized version of a subset of DDSM, including images of two lesion categories, mass, and microcalcification. Each mammogram is marked with a label (benign and malignant)

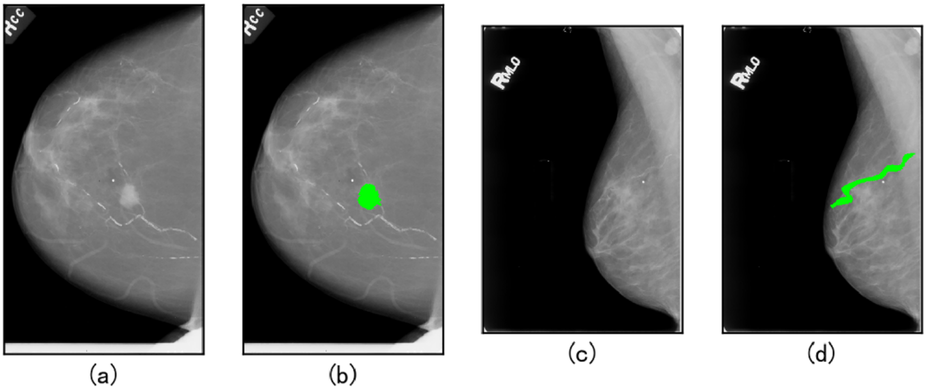

Fig. 2 Mammograms from the CBIS-DDSM dataset. **a** and **b** respectively show the images of mass lesion; **c** and **d** show the image of microcalcification lesion. The green part is the lesion

and provides an accurate bounding box of the lesion area. Figure 2 shows the original mammograms and lesion outline from CBIS-DDSM. We can find the mass lesions are small and dense, while the microcalcification lesions are large and banded. These differentiated features bring great challenges to our computer-aided diagnosis system.

As shown in the Table 1, after removing the duplicate images, we use all the mass images and microcalcification images in the CBIS-DDSM dataset. As the deep learning algorithms often have better performance on large amounts of data, we use data augmentation to increase the number of images. Based on the original images, we rotate each image clockwise by 90°, 180°, and 270° to expand the dataset by 4 times and randomly mix these images. As shown in Table 2, we use a total of 12,040 images to train and test our computer-aided diagnosis system. Since the lesion appears as a complex curve on the image, in order to conveniently express the position of the lesion, as shown in the Fig. 3, we use a rectangle with the center point coordinates and the length and width information to replace the complex curve.

3.2 Image preprocessing method

In the process of mammography scanning, the breast is often deformed due to compression, which has a certain impact on the gray value of the generated image. In order to reduce the impact of breast compression for correct diagnosis, we refer to the multi-threshold peripheral equalization technique from the article [3] and make some adjustments according to the actual situation. Our algorithm mainly creates multiple images by using multiple thresholds and then averages these images to produce a smooth transition between the central and the edge areas of the mammogram, while enhancing the surrounding and lesion areas of the

Table 1 Original dataset after removing the duplicate images

original dataset	benign	malignant	total
mass	839	731	1570
microcalcification	856	584	1440
total	1695	1315	3010

Table 2 Expanded dataset

expanded dataset	benign	malignant	total
mass	3356	2924	6280
microcalcification	3424	2336	5760
total	6780	5260	12040

mammogram. As shown in Fig. 4, our algorithm mainly includes the following consecutive steps: First, we apply a gaussian low-pass filter (GLPF) to the original mammography image I_{orig} as shown in Fig. 4a, thereby generating a blurred image I_{blur} as shown in Fig. 4b. Second, we take the gray average value ave of all non-zero gray pixels on I_{blur}, and use five thresholds T_k as the average of the non-zero gray pixels of the five threshold images to estimate normalized thickness profile (NTP) of the mammogram. The calculation method of each threshold T_k is as formula (1):

$$T_k = ave \times F_k; \quad k = 1, 2, ..., 5 \tag{1}$$

Here, the value of F_k is [0.8, 0.9, 1.0, 1.1, 1.2], which means that it will be used to adjust the proportional parameters around ave. By using this method, the surrounding area of the mammogram can be enhanced, and at the same time, the effect of the breast boundary being too obvious when using a single threshold is eliminated. Then we create five $\hat{I}_{blur}$ images based on these five thresholds T_k, where each pixel (i, j) of the k-th image is represented as $\hat{I}_{blur}(k, i, j)$, their calculation method is formula (2):

$$\hat{I}_{blur}(k, i, j) = \begin{cases} \frac{I_{blur}(i,j)}{T_k} \ ; & I_{blur(i,j)} \le T_k \\ 1 \ ; & \text{otherwise} \end{cases} \tag{2}$$

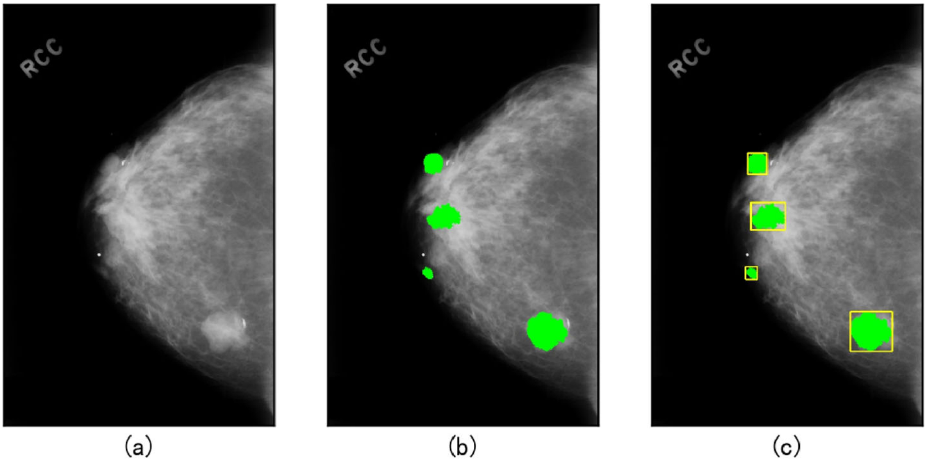

Fig. 3 Position label diagram. **a** is original image; **b** is the image with a precise label; **c** is the image with a label frame

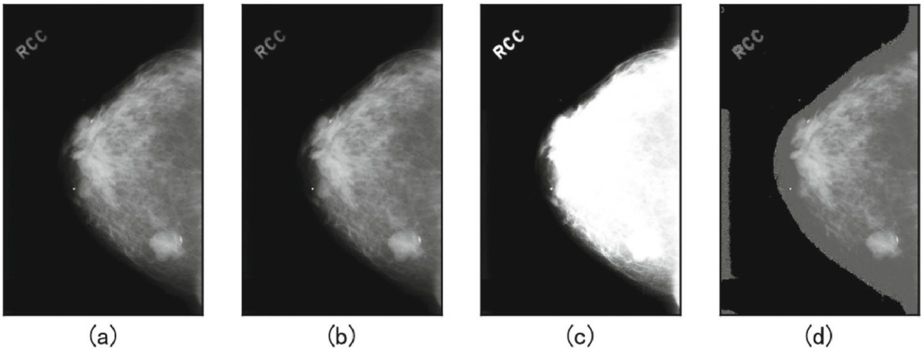

Fig. 4 Image preprocessing. **a** is I_{orig}; **b** is I_{blur}; **c** is I_{ntp}; **d** is I_{peq}

Here, $k = 1, 2, ..., 5$; $i = 1, 2, ..., M$; $j = 1, 2, ..., N$; $M \times N$ is the size of the mammogram. According to these five images $\hat{I}_{blur}(k)$, use formula (3) to calculate the I_{ntp} image as shown in Fig. 4c.

$$I_{ntp} = \frac{1}{5}\sum_{k=1}^{5} \hat{I}_{blur}(k) \tag{3}$$

And then, based on the I_{ntp} image and using the formula (4), we calculate and obtain the peripheral equalized image I_{peq} as shown in Fig. 4d.

$$I_{peq} = \frac{I_{orig}}{I_{ntp}} \tag{4}$$

Finally, since the input size of our YOLOv3-based convolutional neural network is (416, 416, 3), we scale all the images to (416, 416, 3) to match the network input size.

3.3 Cluster anchor boxes

As shown in Fig. 5, in our computer-aided diagnosis system, each image is divided into $N \times N$ non-overlapping grid cells. YOLOv3 is inspired by the multi-scale feature maps. The convolutional network divides the entire image into 13×13, 26×26, 52×52 grid cells on average. Each grid cell is preset with three bounding box vectors, and each bounding box vector is responsible for predicting the potential lesion whose geometric center is located in the grid. Geometric center, width and height of the potential lesion are described by x, y, w, h, confidence of the potential lesion is represented by $conf$, and conditional probability of each category is represented by $class$ ($class_i$ represents the conditional

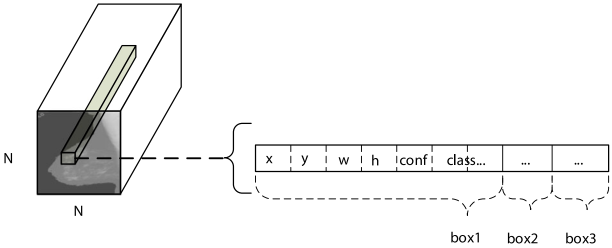

Fig. 5 Sketch map of predicted bounding box vectors

probability of the i-th class if the object is known to be a lesion). Here, $class$ is a multidimensional vector. Therefore, if the length of the class is $len(class)$, the length of the bounding box vector is $len(p) = 4 + 1 + len(class)$, the output shape of the network is a list of $(m, 13, 13, 3, len(p))$, $(m, 26, 26, 3, len(p))$, $(m, 52, 52, 3, len(p))$.

To improve accuracy, we make the large-scale grid cells responsible for predicting large lesions, small-scale grid cells responsible for predicting small lesions. We set different widths and heights for each bounding box of different scales. The preset box with different widths and heights is called anchor box. To make the preset anchor boxes as close as possible to the widths and heights of the label border of more images, an algorithm similar to the k-means clustering algorithm is used to calculate the 9 anchor boxes needed by the three bounding boxes of three scales. Unlike the k-means clustering algorithm, the algorithm we adopted defines distance as formula (5).

$$dist(a, b) = 1 - \frac{\min(w_a, w_b) \cdot \min(h_a \cdot h_b)}{w_a \cdot h_a + w_b \cdot h_b - \min(w_a, w_b) \cdot \min(h_a \cdot h_b)} \tag{5}$$

Here, a, b represent two anchor box vectors. We can easily find that the method defined by formula (5) can express the degree of fitting between two anchor boxes, the higher the degree of fitting of anchor boxes a and b, the smaller the $dist(a, b)$. Except for the different ways of defining distance, our algorithm and k-means clustering algorithm are generally similar in other parts. After the algorithm is completed, we allocate the calculated anchor boxes, so that the large-scale grid cell presets the large-scale anchor box, while the small-scale grid cell presets the small-scale anchor box. Figure 6 shows the distribution of all anchor boxes and the position of nine central anchor boxes calculated by the algorithm.

3.4 Structure and implementation of YOLOv3 based network

YOLOv3 network is a unified structure. In order to facilitate the discussion of the network structure, we use only one image but not m images in each batch for illustration. Therefore, the input shape of the network is $(416, 416, 3)$, and the output shape of the network is a list of $(13, 13, 3, len(p))$, $(26, 26, 3, len(p))$, $(52, 52, 3, len(p))$. Figure 7 shows our

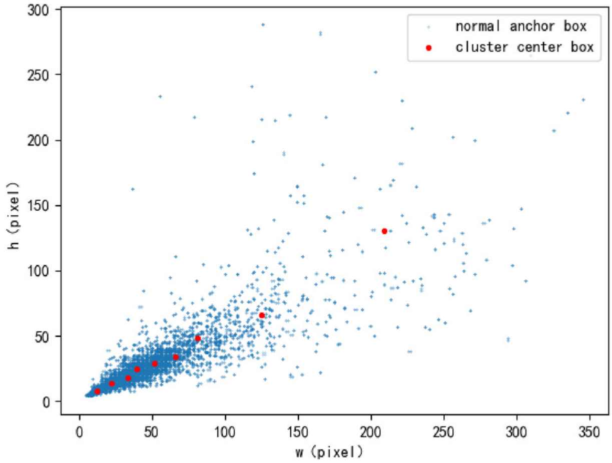

Fig. 6 Clustering scatterplot of anchor boxes. The blue dots are the anchor boxes of input; the red dots are the nine central anchor boxes of output

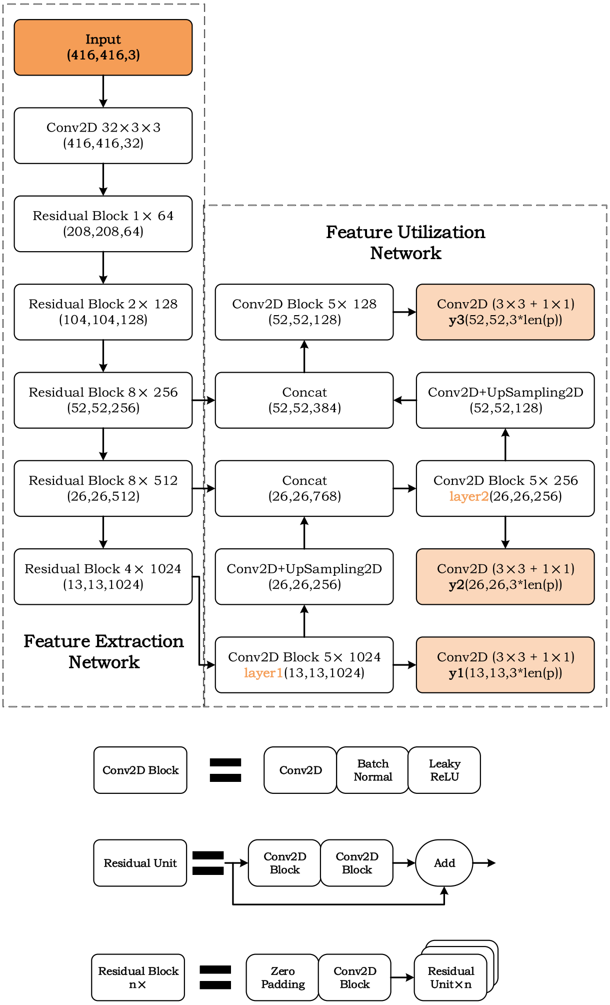

Fig. 7 Convolutional neural network structure based on YOLOv3

network architecture based on YOLOv3. The network is mainly divided into two parts. The first part is the feature extraction network, it is a Darknet-53 without the fully connected layer. It is mainly composed of two structures: Conv2D Block, and Residual Block $n\times$. The second part is the feature utilization network that uses the multiple feature layers from the feature extraction network. We extract a total of three feature maps, and their shapes are $(52, 52, 256)$, $(26, 26, 512)$, $(13, 13, 1024)$. The feature utilization network is mainly composed of Conv2D Block, up-sampling layer, and convolution layer. Through the operation of upsampling, the large-scale feature map can also have convolution features from the small-scale feature map.

Finally, the network output is consolidated into a predicted tensor with the shapes of $(13, 13, 3, len(p))$, $(26, 26, 3, len(p))$, $(52, 52, 3, len(p))$. Each predicted bounding box vector is $p = (x, y, w, h, conf, class)$, while (x, y, w, h) describes the position of the lesion, $conf$ contains the confidences of two sources. One is network's confidence in the presence of lesion in the predicted position, the other is network's confidence about the degree of fitting of its predicted position and label position. As formula (6) shows, we multiply these two confidences to get the final confidence $conf$.

$$conf = P(Lesion) * IOU_{pred}^{truth} \tag{6}$$

The $class$ in predicted bounding box vector p is a conditional probability vector for the categories, as shown in formula (7). It is expressed as the conditional probability that the lesion belongs to various categories $Class_i$ after determining that the object is a lesion.

$$class_i = P(Class_i|Lesion) \tag{7}$$

The loss function of each batch will be calculated by the average loss function of all the images in the batch, as shown in the formula (8).

$$Loss = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{3} Loss_{ij} \tag{8}$$

Here, m is the number of images in each batch, $Loss_{ij}$ is the loss function value of the j-th feature map of the i-th image of the batch, and $Loss_{ij}$ is calculated by formula (9). It is important to note that different feature maps j will correspond to different numbers of grid cells $N \times N$.

$$\begin{aligned} Loss_{ij} = {} & \lambda_{coord}\sum_{i=1}^{N\times N}\sum_{j=1}^{3} 1_{ij}^{obj} * [BCE(x_{ij}, \hat{x}_{ij}) + BCE(y_{ij}, \hat{y}_{ij})] \\ & +\lambda_{coord}\sum_{i=1}^{N\times N}\sum_{j=1}^{3} 1_{ij}^{obj} * [MSE(w_{ij}, \hat{w}_{ij}) + MSE(h_{ij}, \hat{h}_{ij})] \\ & +\sum_{i=1}^{N\times N}\sum_{j=1}^{3} 1_{ij}^{obj} * BCE(conf_{ij}, \hat{conf}_{ij}) \\ & +\lambda_{noobj}\sum_{i=1}^{N\times N}\sum_{j=1}^{3} 1_{ij}^{noobj} * BCE(conf_{ij}, \hat{conf}_{ij}) \\ & +\sum_{i=1}^{N\times N}\sum_{j=1}^{3}\sum_{k=1}^{n_{class}} 1_{ij}^{obj} * BCE(class_{ijk}, \hat{class}_{ijk}) \end{aligned} \tag{9}$$

Among them, function BCE is defined as the form shown in formula (10), which is considered to be the binary cross-entropy loss function in the form of a single sample. And function MSE is the square loss function in the form of a single sample, which is defined as the formula (11).

$$BCE(x, \hat{x}) = (-1) * [x \cdot \log \hat{x} + (1 - x)\log(1 - \hat{x})] \tag{10}$$

$$MSE(x, \hat{x}) = \frac{1}{2}(x - \hat{x})^2 \tag{11}$$

In formula (9), 1_{ij}^{obj} means that the j-th predicted bounding box vector of the i-th grid is a positive sample, and 1_{ij}^{noobj} means the j-th predicted bounding box vector of the i-th grid is a negative sample. The selection strategy of positive and negative samples is determined by the following rules:

- The predicted bounding box with the largest IOU with a certain label bounding box is a positive sample.
- The predicted bounding box whose IOU with a certain label bounding box exceeds the ignoring threshold (we set it as 0.5 in experiments) is a positive sample.
- The predicted bounding box that does not exceed the ignoring threshold and does not have the largest IOU with a certain label bounding box is a negative sample.

In formula (9), λ_{coord} is the coefficient used to balance the coordinate loss, which is defined by formula (12) used to measure the size of the lesion, where $w^{uncoded}$ and $h^{uncoded}$ are the proportions of the length and width of the label bounding box on the entire image. When the lesion is smaller, the value of λ_{coord} is larger, thereby improving the detection ability for small lesions. The coefficient λ_{noobj} is used to balance the loss function of negative samples. Due to a large number of negative samples in the actual situation, in practical application, we let $\lambda_{noobj} = 1$.

$$\lambda_{coord} = 2 - w_{ij}^{uncoded} \cdot h_{ij}^{uncoded} \tag{12}$$

3.5 Network output processing method

Since there are still a lot of predicted bounding box vectors with low confidence or even 0 in the final output tensor of the network, we need to process the output tensor and get the real bounding box and the score of each category.

First, we calculate the confidence $score_i$ of each category in each predicted bounding box vector according to formula (13).

$$\begin{aligned} score_i &= conf * class_i \\ &= P(Lesion) * IOU_{pred}^{truth} * P(Class_i|Lesion) \\ &= P(Class_i) * IOU_{pred}^{truth} \end{aligned} \tag{13}$$

Then we copy the position information in predicted bounding box vectors $len(class)$ times as box vectors, so that each box vector is assigned unique $score$ and $class$. Then we filter the boxes with low score. We set a score threshold, all boxes that do not reach the score threshold and have no maximum score are filtered out.

In the actual situation, as the grid cells near the geometric center of the label bounding box may generate many boxes with higher scores for the same lesion, there are often a large number of boxes after threshold filtering. They always overlap each other and interweave in the core area of the lesion. In order to eliminate these redundant boxes, we use the non-maximum suppression algorithm to find the best box for the lesion. The specific process of the non-maximum suppression algorithm is as follows:

1) Sort all boxes according to their scores.
2) Select the box with the highest score and add to the output list, then delete the box from the box list.
3) Calculate all IOUs of the box with the highest score and other boxes, and delete the boxes whose IOU is greater than the IOU threshold.
4) Repeat the above process until the box list is empty, and exit the algorithm.

Figure 8 shows the effect of the non-maximum suppression algorithm. We use this box list after the final step of algorithm to draw the output mammogram.

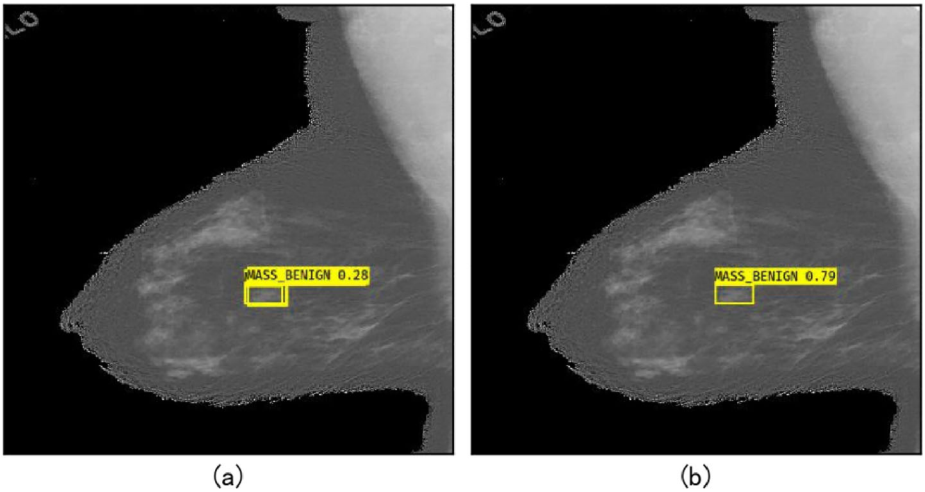

Fig. 8 The effect of non-maximum suppression. **a** is the image before using the non-maximum suppression algorithm, **b** is the image after using the non-maximum suppression algorithm

4 Experiment and analysis

4.1 Experiment setup

4.1.1 Operating environment and training strategy

In our dataset, there are two types of the breast tumor lesions: mass and microcalcification. At the same time, lesion is one of the two cases of being benign and malignant, thus we merge these types into four categories: mass-benign (MB), mass-malignant (MM), microcalcification-benign (CB), and microcalcification-malignant (CM). In some situations, if the lesion of the mammogram can be determined to be mass or microcalcification, our computer-aided diagnosis system will show better performance. Therefore we train three models under this system: general model, mass model and microcalcification model. Since the number of the dataset we use is medium-sized, we shuffle the dataset and divide it into 10 subsets on average, maintaining the proportion of each category of the original dataset. Then we take nine subsets of them as the training set and the remaining one subset as the test set (also as the verification set).

Our training process is completed on the NVIDIA GeForce GTX 1080Ti graphics card on the cloud server, which has a core frequency of 1480Mhz and 11GB of video memory. The test process is completed on a local laptop, the graphics card used is NVIDIA GeForce GTX 960m, and it has a core frequency of 1176Mhz and 2GB of video memory. It is worth noting that although we spend several days training the model on a high-performance graphics card, the overall system flow from input to output is achieved at an average of 0.144s per image on our low-performance mobile graphics card.

We mainly use the following training tricks in the training process.

- **Transfer learning.** Transfer learning has proven to be effective in the training of convolutional neural networks for mammograms [32]. The first few layers in the convolutional neural network are often used to extract shallow features such as upper and

Table 3 Number of mammograms used by general model

classes	MB	MM	CB	CM	total
train set	3021	2632	3082	2103	10838
test set	335	292	342	233	1202
train set rate	90.018 %	90.014 %	90.012 %	90.026 %	90.017 %

lower boundaries, which are also applicable to mammograms. The original YOLOv3 weights are trained on the ImageNet [10] dataset which has a large number of images. We only need to adjust them slightly based on our data, so we first load the original weight of YOLOv3.

- **Adam optimizer.** In the training process, we use Adam optimizer [14] instead of the stochastic gradient descent algorithm. At the same time, we use an adaptive learning rate adjustment strategy. After monitoring that the loss function value of the validation set has not decreased after multiple epochs of training, the learning rate will drop to one-tenth of the original to obtain more delicate optimization.

4.1.2 General, mass and microcalcification model

According to the type of breast lesions, we train three models in the computer-aided diagnosis system based on YOLOv3: general model, mass model and microcalcification model. For the training and evaluation of the general model, we use all mammograms of the expanded dataset. As shown in Table 3, we use a total of 10838 mammograms for general model training in the system. In the test phase, we use a total of 1202 mammograms for general model evaluation. There are four categories involved in the general model: mass-benign (MB), mass-malignant (MM), microcalcification-benign (CB), microcalcification-malignant (CM). Therefore, the number of classes $len(class) = 4$, the length of each prediction box vector $len(p) = 9$, the network output size of general model is $(m, 13, 13, 3, 9)$, $(m, 26, 26, 3, 9)$, $(m, 52, 52, 3, 9)$.

Although the number of these types of images is not strictly limited to $1 : 1$, their numbers are not much different. To enable the model to learn more image features as much as possible, as shown in Tables 4 and 5, we use all the mass images and microcalcification images in the expanded dataset to train the mass model and micro-microcalcification model respectively. Only two types of benign and malignant are involved in the mass model and microcalcification model. Therefore, the number of classes $len(class) = 2$, the length of each prediction box vector $len(p) = 7$, and the network output size of each model is $(m, 13, 13, 3, 7)$, $(m, 26, 26, 3, 7)$, $(m, 52, 52, 3, 7)$.

Table 4 Number of mammograms used by mass model

classes	MB	MM	total
train set	3021	2632	5653
test set	335	292	627
train set rate	90.018 %	90.014 %	90.016 %

Table 5 Number of mammograms used by microcalcification model

classes	CB	CM	total
train set	3082	2103	5185
test set	342	233	575
train set rate	90.012 %	90.026 %	90.017 %

4.2 Evaluation system

4.2.1 Evaluation logic

We evaluate the performance of our computer-aided diagnostic system based on objective quantitative methods. Figure 9 shows our evaluation logic during the testing phase. We input the test images into the trained model to obtain output tensors. The output tensor is first decoded and filtered by the score threshold, and then processed by the non-maximum suppression algorithm to obtain the final box list. Our evaluation is based on the output box list. It is important to note that if all the scores of all boxes do not exceed the score threshold, the box with the highest score will be retained.

Figure 10a is the image output through the final evaluation system when the score threshold is 0. At this time, the predicted boxes on the screen appear messy because of the large number of boxes. Figure 10b is the image with the score threshold of 0.01. Although the predicted boxes have been filtered to only two, the box with a low score is still wrong and unavailable. Figure 10c is the image with the score threshold of 0.1, and Fig. 10d is the image with the label bounding box. Obviously, the two bounding boxes are very close. After comparing the predicted box and the label box, we set the score threshold to 0.1 and the IOU threshold used in the non-maximum suppression algorithm to 0.4. Therefore, the criteria for setting thresholds is based on experimental results.

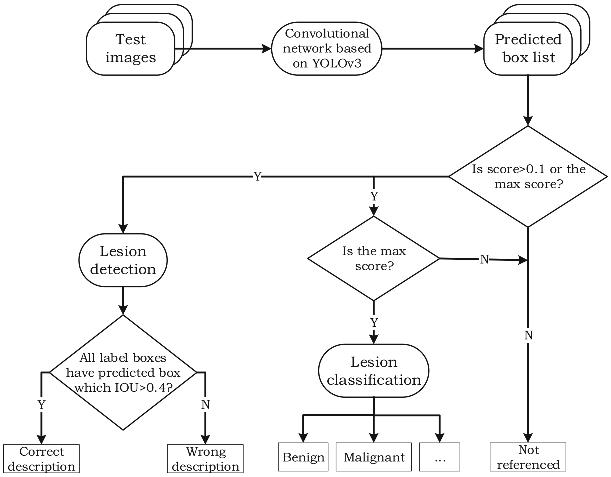

Fig. 9 Evaluation logic of the evaluation system

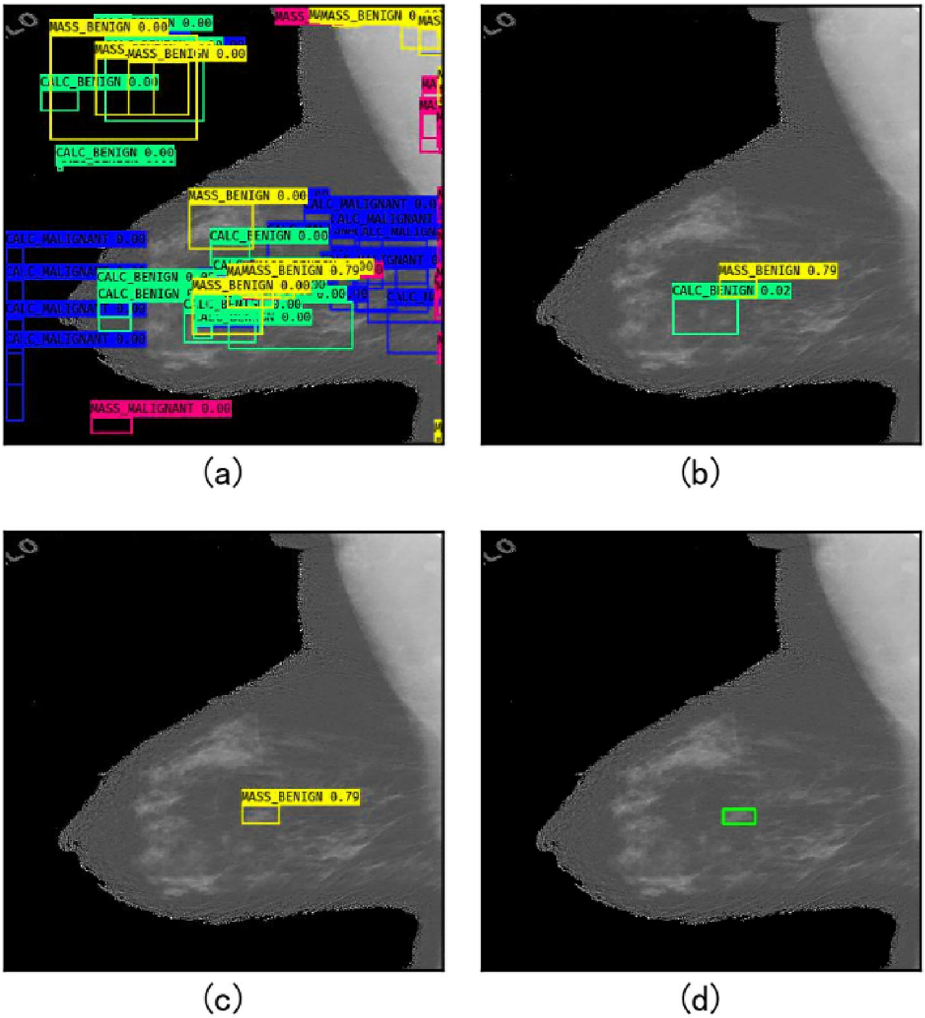

Fig. 10 The effect of different score thresholds on the system results

4.2.2 Evaluation methods and metrics of lesion detection

In the evaluation of system's lesion detection performance, we still use the IOU of predicted box and label box as reference. Similarly, if the IOU of the predicted box and the label box reaches 0.4, we believe that the predicted box correctly describes the label box. Therefore, the specific algorithm flow is as follows: traverse the label boxes of each lesion in each image, and traverse all the predicted boxes for each label box. If a predicted box is found and reaches the IOU threshold with this label box, we think that the label box is correctly described. If all label boxes of lesions in the entire image are correctly described, the model is judged as predicting the position of the lesions in this image correctly. If a certain label box is not described correctly, it is considered that the model is wrong in predicting the position of this image lesion. Therefore, we use the accuracy rate as the metric to measure the performance of the model trained by our system in the task of detecting the lesions.

4.2.3 Evaluation methods and metrics of lesion classification

For evaluating the system's lesion classification performance, we directly use the category of the predicted box with the highest score as the reference. Because each image in the CBIS-DDSM dataset only has one type of lesion, the predicted box with the highest score will have the highest confidence. For all models, we use accuracy to assess the overall classification performance.

Our dataset can be divided into benign or malignant, mass or microcalcification categories which are semantically opposite. In order to show the model performance from different aspects, we also calculate the classification results of each opposing category to describe the results of the model. As our main task is to find malignant lesions as much as possible, we define malignant lesions as positive and benign lesions as negative. Besides, microcalcification lesions are more difficult to be detected than mass lesions, inspired by this, we define microcalcification lesions as positive and mass lesions as negative. The various evaluation metrics in binary classification, including $accuracy$, $precision$, $recall$ and $f1\ score$, will be applied to the reference standard of the classification performance of our model. The definition of each metric is shown in formula (14):

$$\begin{cases} accuracy = \frac{TP+TN}{TP+TN+FP+FN} \\ precision = \frac{TP}{TP+FP} \\ recall = \frac{TP}{TP+FN} \\ f1 = \frac{2*precision*recall}{precision+recall} \end{cases} \tag{14}$$

In formula (14), TP and TN respectively indicate the number of positive and negative categories that are correctly predicted in the test set, while FP and FN indicate the number of positive and negative classes that are incorrectly predicted in the test set.

The $accuracy$ rate measures the overall prediction. The $precision$ rate represents the proportion of the samples of the predicted positive class that are actually positive; the $recall$ rate represents the proportion of the samples of the actual positive class that are predicted to be positive; the $f1\ score$ integrates the two metrics of the precision rate and the recall rate. Besides, we also adopt the area under the ROC curve AUC as one of the evaluation metrics, in which the ROC curve is a curve with the axis of false positive rate and the axis of true positive rate.

In particular, the inference time of the test phase is also an important metric. The moment we get the image from the hard disk is set as the start time and the moment after the model infers the position and type of the lesion is set as the end time. The time between the two moments is the inference time of our model.

4.3 The detection performance of our models

Figure 11 shows the detection ability of our model trained by the proposed computer-aided diagnosis system. It can be seen that the prediction of the lesion position and category by our system is relatively accurate.

Table 6 records the performance on each category in the lesion detection task of the general model. In the table, the image with the IOU of the predicted box and the label box lower than 0.4 is considered as incorrectly describing the position of the lesions. The

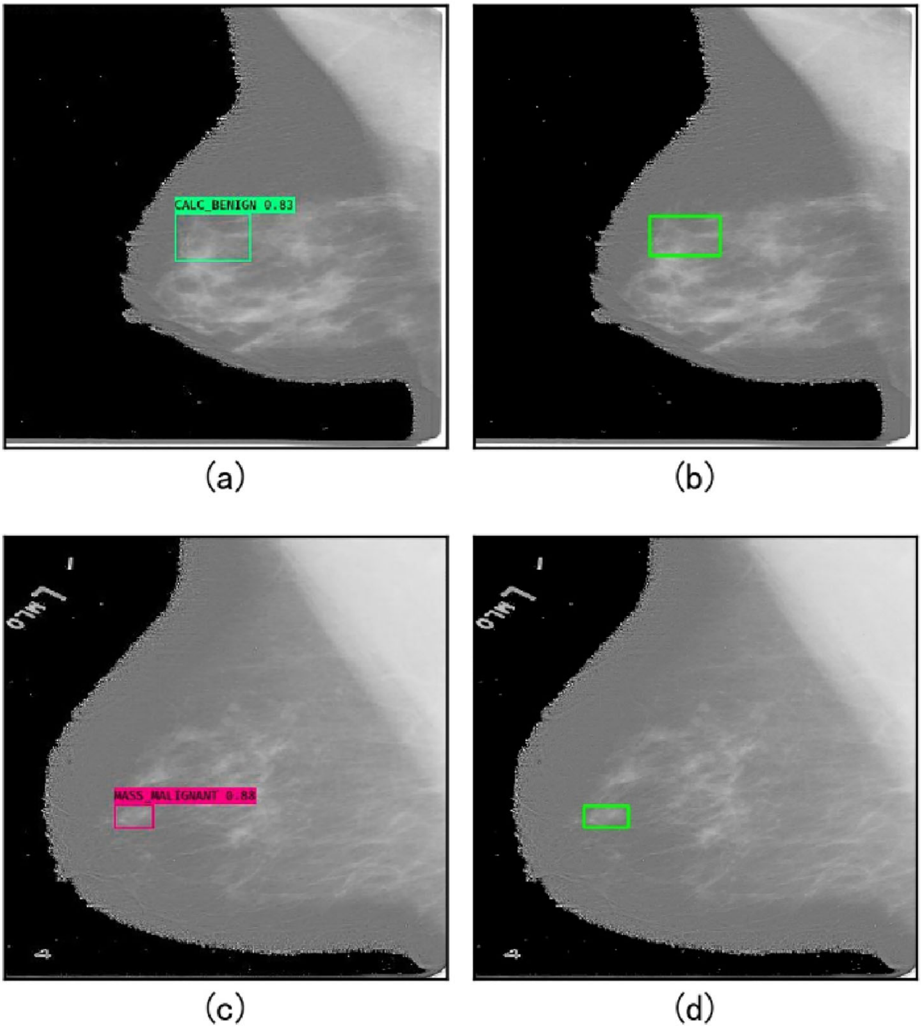

Fig. 11 Model test results. **a** and **b** show the prediction and label boxes of a microcalcification image respectively; **c** and **d** show the prediction and label boxes of a mass image respectively

Table 6 Table of the number of detections of each category of the general model

classes	MB	MM	CB	CM	total
true	318	280	316	212	1126
false	17	12	26	21	76
total	335	292	342	233	1202
accuracy	94.925 %	95.890 %	92.398 %	90.987 %	93.677 %

Table 7 Table of the number of detections of each opposite category of the general model

classes	benign	malignant	mass	calc
true	634	492	598	528
false	43	33	29	47
total	677	525	627	575
accuracy	93.648 %	93.714 %	95.375 %	91.826 %

result shows that the general model trained by our computer-aided diagnosis system has high robustness, 93.677 % of the mammogram lesions can be correctly described under the condition when the IOU threshold is 0.4. In addition, 95.890 % of the mammogram lesions whose category is the mass-malignant can be correctly described.

Table 7 lists the general model performance in detection task for two major categories with opposite semantic meaning (benign and malignant, mass and microcalcification). Among them, the benign category includes mass-benign and microcalcification-benign; the malignant category includes mass-malignant and microcalcification-malignant; the mass category includes mass-benign and mass-malignant; the microcalcification category includes microcalcification-benign and microcalcification-malignant. In the detection task, the general model trained by our system has little difference in the detection performance of benign and malignant lesions, but the detection performance of the mass lesions is significantly better than that of microcalcification lesions. This may be caused by the fact that the mass lesions are usually small and dense while the microcalcification lesions are comparatively large and band-shaped.

Due to the large difference between mass lesions and microcalcification lesions, if we can determine the type of lesions, the detection ability of the mass model and microcalcification model will be obviously improved than the general model. As shown in Tables 8 and 9, the lesion detection performance of the mass model can reach an accuracy rate of 97.767 %, compared with the detection accuracy rate of 95.375 % of general model, which is improved by 2.410 %. The detection ability of the microcalcification model has been significantly improved after the detection task is limited to the scope of microcalcification lesions. It can achieve an accuracy rate of 96.870 % in the position detection task, which is an increase of 5.044 % compared with 91.826 % of the detection accuracy rate of the microcalcification category of the general model. Therefore, if it can be determined that the type of lesion is microcalcification or mass, using the corresponding model for position detection will have better performance.

Table 8 Table of the number of detections of each category of the mass model

classes	MB	MM	total
true	326	287	613
false	9	5	14
total	335	292	627
accuracy	93.313 %	98.288 %	97.767 %

Table 9 Table of the number of detections of each category of the microcalcification model

classes	CB	CM	total
true	331	226	557
false	11	7	18
total	342	233	575
rate	96.784 %	96.996 %	96.870 %

4.4 The classification performance of our models

In the general model, we use all 1202 images in the test set to evaluate the performance of the general model. Table 10 shows the classification result of each category in the test set calculated by the general model. We can calculate the accuracy rate of general model classification, which reaches 93.927 % from the data in the table. Table 11 records the metrics of the general model in each classification task. We can observe these following results:

- The general model achieves the best performance on the task of classifying lesion types as mass or microcalcification.
- The accuracy rate of the general model in detecting the position of mass lesion is significantly higher than that of detecting the position of microcalcification lesion, while the accuracy in classifying mass lesion is slightly lower than that of classifying microcalcification lesion.
- In each classification task, the metrics of the general model are between 94 % and 98 %, which shows the relatively robust performance of our general model.

We test the mass model and the microcalcification model respectively with a total of 627 mass images and 575 microcalcification images in the test set. Table 12 shows the test results of classification performance of mass model and microcalcification model. In the Table 12, the mass model is on the upper left and the microcalcification model is on the lower right. Since the two models are independent and do not interfere with each other, the lower left and the upper right show 0. From Table 12, we can intuitively see that the number of correct predictions of the mass model and the microcalcification model is significantly higher than that of the general model. Table 13 lists the comparison between mass model and microcalcification model and the general model in terms of the classification of mass and microcalcification lesions. From the listed data, we can summarize these following results:

Table 10 Classification result of general model

pred label	MB	MM	CB	CM	total
MB	309	15	8	3	335
MM	9	274	3	6	292
CB	2	4	328	8	342
CM	1	6	8	218	233
total	321	299	347	235	1202

Table 11 Classification metrics of general model

tasks metrics	M or C	Mass	Calc	B or M
accuracy	97.255 %	95.215 %	96.348 %	95.757 %
precision	96.564 %	93.960 %	94.915 %	94.382 %
recall	97.739 %	95.890 %	96.137 %	96.000 %
f1 score	97.148 %	94.915 %	95.522 %	95.184 %
AUC	97.275 %	95.259 %	96.314 %	95.784 %

- The result metrics of the mass model and the microcalcification model are higher than those of the general model, while the metrics of the mass model increase most significantly.
- In general model, the classification result of mass lesions is weaker than that of microcalcification lesions, but the classification performance of the mass model is significantly better than that of microcalcification model.
- Performances of the mass model and the microcalcification model vary between 96 % and 99 %, showing better performance than the general model.

4.5 The test speed of our models

In particular, test speed is also an important metric of model performance. Our experiment is completed on a laptop with NVIDIA GeForce GTX 960m. The average test times for the three models are shown in Table 14, which conveys two very important messages. The first is that our model can also run on servers with poor performance, and the running time is extremely short, which makes it possible to use our model in hospitals. The second is that the overall operating speed does not increase significantly with the increasing of prediction categories, which shows that our computer-aided diagnostic system can maintain a high identification speed for a variety of complex breast lesion categories.

4.6 Comparison of our system with other methods

In this study, we have developed a computer-aided diagnosis system for mammograms. We use the system to train three models to detect mammogram lesions under certain conditions and classify them as mass, microcalcification, benign, malignant, and other categories. Figure 11 shows the ability of our system to correctly detect and classify various types of

Table 12 Classification results of mass and microcalcification models

pred label	MB	MM	CB	CM	total
MB	327	8	0	0	335
MM	4	288	0	0	292
CB	0	0	335	7	342
CM	0	0	9	224	233
total	331	296	344	231	1202

Table 13 Classification metrics of mass and microcalcification models

models metrics	mass	general(mass)	calc	general(calc)
accuracy	98.086 %	95.215 %	97.217 %	96.348 %
precision	97.297 %	93.960 %	96.970 %	94.915 %
recall	98.630 %	95.890 %	96.137 %	96.137 %
f1 score	97.959 %	94.915 %	96.552 %	95.522 %
AUC	98.121 %	95.259 %	97.045 %	96.314 %

potential breast lesions. Tables 6 to 14 show the overall performance of our system from various quantitative metrics. These results prove that our YOLOv3-based computer-aided diagnosis system has good performance for mammogram lesions.

To show the robustness of our computer-aided diagnosis system based on YOLOv3, we also compare this system with some of the state-of-the-art methods. Among them, Arevalo et al. [4] performed the CNN parameter exploration by training 25 models with random hyperparameter initializations and choosing the best according to validation performance. Sun et al. [31] used the batch size of 100, subsampling rate of 2, and the learning rate was set to 0.1 for 100 epochs. Sun et al. [30] randomly initialized the weights of network, and the learning rate was set as 0.0001 with 128 batch size. Suzuki et al. [32] utilized the DCNN containing 8 layers, with the first 5 convolution layers and the remaining 3 fully-connected layers. Arfan et al. [5] set batch size for stochastic gradient descent to 64 with momentum 0.8 and decay parameter of 1e-5. Mordang et al. [20] initially set the learning rate to 0.01 and linearly decreased to 0.0001 over the maximum number of epochs. Bria et al. [8] set momentum and weight decay to 0.9 and 0.0005 respectively, and the dropout probability to 0.5. Ben-ari et al. [6] used 0.4 as the threshold parameter on the overlap ratio. Sarath et al. [27] trained the localization network for 300 epochs with a batch size of 8 and 36 batch updates per epoch, trained the MIL network for 100 epochs with a learning rate of 0.001 and weight decay of 0.0005. Platania et al. [23] initialized the weights of CNN from pretraining, and stochastic gradient descent is utilized for the minimization.

Table 15 is a comparison of performance in mass classification task. Other state-of-the-art methods tend to crop images to improve accuracy. In particular, Arfan et al. [5] achieved 93 % AUC, but this result is still lower than the 98.121 % AUC of our mass model. Fewer methods paid attention to the classification of microcalcification lesions. As shown in Table 16, the performances of other researchers' methods in the task of classification of microcalcification lesions are relatively poor. Although the image preprocessing process by Bria et al. [8] adopted the defogging algorithm to effectively improve the recall rate to 76.26 %, it is still lower than the recall rate of 96.137 % that can be achieved by our microcalcification model. Because our model adds detection steps for the position of lesions, it further improves the accuracy of classification. Table 17 shows the comparison of the performances of our system's mass model and other researchers' methods in the detection and classification of mass lesions. On position detection task, Al-masni et al. [1] achieved the

Table 14 The results of model test speed

model	general model	mass model	calc model
time	0.144s	0.142s	0.143s

Table 15 Comparison of our system with other methods in mass lesion classification

paper	dataset	method	metric	result
L. Sun et al. [30]	DDSM	MVMDCNN	Accuracy	82.02 %
W. Sun et al. [31]	FFDM	SSL+ROI+CNN	AUC	82.43 %
Suzuki et al. [32]	DDSM	TL+CNN	Recall	89.90 %
Arevalo et al. [4]	BCDR	ROI+CNN	AUC	82.20 %
Arfan et al. [5]	MIAS, DDSM	CNN+SVM	AUC	93 %
This paper	DDSM	YOLOv3(mass)	Accuracy AUC Recall	98.086 % 98.121 % 98.630 %

Table 16 Comparison of our system with other methods in microcalcification lesion classification

paper	dataset	method	metric	result
Mordang et al. [20]	Multiple dataset	ROI+CNN	Recall	70 %
Bria et al. [8]	GE Sengraphe	Defog+CNN	Recall	76.26 %
This paper	DDSM	YOLOv3(calc)	Recall	96.137 %

Table 17 Comparison of our system with other methods in mass lesion detection and classification

paper	dataset	method	metric	result
Sarath et al. [27]	INbreast	Two-stage MIL	IOU=0.5	76 %
Platania et al. [23]	DDSM	YOLOv1	IOU=0.25	90 %
Bria et al. [8]	DDSM	YOLOv1	IOU=0.5	99.7 %
Ben-ari et al. [6]	DDSM	R-CNN	IOU=0.4	88 %(Recall)
This paper	DDSM	YOLOv3(mass)	IOU=0.4	97.767 %
Sarath et al. [27]	INbreast	Two-stage MIL	AUC	76 %
Platania et al. [23]	DDSM	YOLOv1	AUC	92.315 %
Bria et al. [8]	DDSM	YOLOv1	Accuracy	97 %
Ben-ari et al. [6]	DDSM	R-CNN	Recall	79 %
This paper	DDSM	YOLOv3(mass)	Recall Accuracy AUC	98.630 % 98.086 % 98.121 %

The top of the table is the comparison of detection performance, and the bottom of the table is the comparison of classification performance highest detection accuracy rate of 99.7 % when the IOU threshold was 0.5, but on classification task, our model shows a higher recognition effect, reaching the accuracy rate of 98.086 %.

The results of the comparison show that the trained models based on our computer-aided diagnosis system proposed in this study have relatively high metrics and can accomplish more tasks, which are generally better than the experimental results of the methods proposed by other researchers.

5 Conclusion

In this article, we introduce a YOLOv3-based computer-aided diagnosis system for detection and classification of mammogram lesion. The system can overcome most of the medical image noise, and achieves two tasks of lesion detection and classification in the same neural network simultaneously. We train three models under this system: general model, mass model, microcalcification model. These three models have good performance in detecting and classifying lesions. Besides, the test speed of our system can reach 0.144s on a low-performance laptop, which makes it possible for the system to be applied in the hospitals.

References

1. Al-masni MA, Al-antari MA, Park JM, Gi G, Kim TY, Rivera P, Valarezo E, Han S-M, Kim TS (2017) Detection and classification of the breast abnormalities in digital mammograms via regional convolutional neural network. In: 2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, Seogwipo, pp 1230–1233. [Online]. Available: https://ieeexplore.ieee.org/document/8037053/
2. Al-masni MA, Al-antari MA, Park J-M, Gi G, Kim T-Y, Rivera P, Valarezo E, Choi M-T, Han S-M, Kim T-S (2018) Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. Comput Methods Prog Biomed 157:85–94. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0169260717314980
3. Al-antari MA, Al-masni MA, Park S-U, Park J, Metwally MK, Kadah YM, Han S-M, Kim T-S (2018) An automatic computer-aided diagnosis system for breast cancer in digital mammograms via deep belief network. J Med Biol Eng 38(3):443–456. [Online]. Available: http://link.springer.com/10.1007/s40846-017-0321-6
4. Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Guevara Lopez MA (2016) Representation learning for mammography mass lesion classification with convolutional neural networks. Comput Methods Prog Biomed 127:248–257. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0169260715300110
5. Arfan M (2017) Deep learning based computer aided diagnosis system for breast mammograms. Int J Adv Comput Sci App 8(7):286–290. [Online]. Available: http://thesai.org/Publications/ViewPaper?Volume=8&Issue=7&Code=ijacsa&SerialNo=38
6. Ben-Ari R, Akselrod-Ballin A, Karlinsky L, Hashoul S (2017) Domain specific convolutional neural nets for detection of architectural distortion in mammograms. In: 2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017). IEEE, Melbourne, pp 552–556. [Online]. Available: http://ieeexplore.ieee.org/document/7950581/
7. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer J Clin 68(6):394–424. [Online]. Available: http://doi.wiley.com/10.3322/caac.21492
8. Bria A, Marrocco C, Galdran A, Campilho A, Marchesi A, Mordang J-J, Karssemeijer N, Molinara M, Tortorella F (2017) Spatial enhancement by dehazing for detection of microcalcifications with convolutional nets. In: Battiato S, Gallo G, Schettini R, Stanco F (eds) Image analysis and processing - ICIAP

2017, vol 10485. Springer International Publishing, Cham, pp 288–298. series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-319-68548-9_27

9. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M et al (2013) The cancer imaging archive (TCIA): maintaining and operating a public information repository. J Digit Imag 26(6):1045–1057. publisher: Springer. [Online]. Available: http://link-springer-com-s.vpn.whu.edu.cn:8118/article/10.1007/s10278-013-9622-7
10. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database, 248–255. [Online]. Available: https://ieeexplore.ieee.org/document/5206848
11. Gøtzsche PC, Jørgensen KJ (2013) Screening for breast cancer with mammography. Cochrane Database Syst Rev. [Online]. Available: http://doi.wiley.com/10.1002/14651858.CD001877.pub5
12. Heath M, Bowyer K, Kopans D, Kegelmeyer P, Moore R, Chang K, Munishkumaran S (1998) Current status of the digital database for screening mammography. In: Viergever MA, Karssemeijer N, Thijssen M, Hendriks J, van Erning L (eds) Digital mammography, vol 13. Springer Netherlands, Dordrecht, pp 457–460. series Title: Computational Imaging and Vision. [Online]. Available: http://link.springer.com/10.1007/978-94-011-5318-8_75
13. Jung H, Kim B, Lee I, Yoo M, Lee J, Ham S, Woo O, Kang J (2018) Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network. PLOS ONE 13(9):e0203355. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0203355
14. Kingma DP, Ba J (2017) Adam: a method for stochastic optimization, arXiv:1412.6980 [cs]
15. Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60(6):84–90. [Online]. Available: https://dl.acm.org/doi/10.1145/3065386
16. Kooi T, Gubern-Merida A, Mordang J-J, Mann R, Pijnappel R, Schuur K, den Heeten A, Karssemeijer N (2016) A comparison between a deep convolutional neural network and radiologists for classifying regions of interest in mammography. In: Tingberg A, Lång K, Timberg P (eds) Breast imaging, vol 9699. Springer International Publishing, Cham, pp 51–56. series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-319-41546-8_7
17. Lee CH, Dershaw DD, Kopans D, Evans P, Monsees B, Monticciolo D, Brenner RJ, Bassett L, Berg W, Feig S, Hendrick E, Mendelson E, D'Orsi C, Sickles E, Burhenne LW (2010) Breast cancer screening with imaging: recommendations from the society of breast imaging and the ACR on the use of mammography, breast MRI, breast ultrasound, and other technologies for the detection of clinically occult breast cancer. J Amer College Radiol 7(1):18–27. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1546144009004803
18. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2018) Focal loss for dense object detection, arXiv:1708.02002 [cs]
19. Ma J, Liang S, Li X, Li H, Menze BH, Zhang R, Zheng W-S (2019) Cross-view relation networks for mammogram mass detection, arXiv:1907.00528 [cs]
20. Mordang J-J, Janssen T, Bria A, Kooi T, Gubern-Mérida A, Karssemeijer N (2016) Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks. In: Tingberg A, Lång K, Timberg P (eds) Breast imaging, vol 9699. Springer International Publishing, Cham, pp 35–42. series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-319-41546-8_5
21. Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS (2012) INbreast. Acad Radiol 19(2):236–248. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S107663321100451X
22. Omonigho EL, David M, Adejo A, Aliyu S (2020) Breast cancer: tumor detection in mammogram images using modified alexnet deep convolution neural network. In: 2020 international conference in mathematics, computer engineering and computer science (ICMCECS). IEEE, Ayobo, Ipaja, Lagos, Nigeria, pp 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9077659/
23. Platania R, Shams S, Yang S, Zhang J, Lee K, Park S-J (2017) Automated breast cancer diagnosis using deep learning and region of interest detection (BC-DROID). In: Proceedings of the 8th ACM international conference on bioinformatics, computational biology,and health informatics - ACM-BCB '17. ACM Press, Boston, pp 536–543. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3107411.3107484
24. Redmon J, Farhadi A (2018) YOLOv3: an incremental improvement, 1–6. [Online]. Available: https://pjreddie.com/media/files/papers/YOLOv3.pdf
25. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, Las Vegas, pp 779–788. [Online]. Available: http://ieeexplore.ieee.org/document/7780460/
26. Ren S, He K, Girshick R, Sun J (2016) Faster R-CNN: towards real-time object detection with region proposal networks, arXiv:1506.01497 [cs]

27. Sarath CK, Chakravarty A, Ghosh N, Sarkar T, Sethuraman R, Sheet D (2020) A two-stage multiple instance learning framework for the detection of breast cancer in mammograms, arXiv:2004.11726 [cs]
28. Sawyer-Lee R, Gimenez F, Hoogi A, Rubin D (2016) Curated breast imaging subset of DDSM, Cancer Imag Archive. [Online]. Available: https://wiki.cancerimagingarchive.net/x/lZNXAQ
29. Siegel RL, Miller KD, Jemal A (2020) Cancer statistics, 2020. CA: A Cancer J Clin 70(1):7–30. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21590
30. Sun L, Wang J, Hu Z, Xu Y, Cui Z (2019) Multi-view convolutional neural networks for mammographic image classification. IEEE Access 7:126273–126282. [Online]. Available: https://ieeexplore.ieee.org/document/8822935/
31. Sun W, Tseng T-LB, Zhang J, Qian W (2017) Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. Comput Med Imag Graph 57:4–9. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0895611116300696
32. Suzuki S, Zhang X, Homma N, Ichiji K, Sugita N, Kawasumi Y, Ishibashi T, Yoshizawa M (2016) Mass detection using deep convolutional neural network for mammographic computer-aided diagnosis. In: 2016 55th annual conference of the society of instrument and control engineers of Japan (SICE). IEEE, Tsukuba, pp 1382–1386. [Online]. Available: http://ieeexplore.ieee.org/document/7749265/

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.