Applied Intelligence (2022) 52:12943–12956
https://doi.org/10.1007/s10489-021-03140-5

Automated assessment of BI-RADS categories for ultrasound images using multi-scale neural networks with an order-constrained loss function

Yong Pi[1] · Qian Li[2] · Xiaofeng Qi[1] · Dan Deng[2] · Zhang Yi[1]

Accepted: 26 December 2021 / Published online: 16 February 2022

Abstract
Ultrasound imaging is one of the most frequently used diagnostic tools for detecting and analyzing abnormalities of the breast. Recently proposed methods for the automated analysis of breast ultrasound images have shown great success, especially in the classification of breast abnormalities into malignant or benign lesions. In this study, we explore the use of a deep convolutional neural network with a multi-scale module for the automated assessment of BI-RADS category on breast ultrasound images. We propose a multi-input region of interest extraction method to extract the breast region from ultrasound images, with this method avoiding the deformation of region of interest images and providing high classification performance. Moreover, we also propose an order-constrained loss function that fully considers the continuity between BI-RADS categories and shows higher performance than traditional loss functions. A large annotated dataset containing 8246 breast ultrasound images was collected to train and evaluate the proposed methods. Ablation experiments were performed to validate the effectiveness of the proposed methods. Experimental results indicate that our method can be used to mimic experienced radiologists in the assessment of the BI-RADS category of breast ultrasound images and that the automated interpretations could be acceptable in routine clinical breast ultrasound examination reports.

Keywords Deep neural networks · Breast ultrasound image · BI-RADS · Multi-scale · Order-constrained loss

Introduction Globally, breast cancer is the most common cancer among women, accounting for 11.6% of all new cancer cases, and remains the second leading cause of cancer death, accounting for 6.6% of the total cancer-related deaths in 2018 [4]. Breast cancer can usually be cured if diagnosed and treated early [42]; therefore, the early detection of breast cancer and its timely treatment are associated with lower breast cancer death rates in the long term.

Currently, the most effective modality for the early detection and diagnosis of breast cancer was mammography [42]. However, mammography presents several limitations in breast cancer detection: it is less effective for subjects under 40 years old and dense breasts, has low sensitivity for small tumors [19], and the ionizing radiation can present a health risk for both patients and radiographers. Magnetic resonance imaging (MRI) has particularly high sensitivity which can detect small lesions that cannot be detected by mammography; however, it is also expensive and has low specificity, which can lead to overdiagnosis [32]. Ultrasonography is an indispensable tool in breast imaging. It is complementary to both mammography and MRI of the breast, offering non-invasiveness, higher sensitivity and accuracy, lower cost, and faster imaging acquisition [20].

The American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) lexicon [35] is designed to standardize the reporting of breast imaging and reduce confusion in breast imaging interpretations, and also facilitates outcome monitoring and quality assessment. Generally, radiologists manually analyze ultrasonography images and classify them into one of seven BI-RADS grades, as defined in the BI-RADS lexicon. Table 1

✉ Dan Deng
dengdan@boe.com.cn

✉ Zhang Yi
zhangyi@scu.edu.cn

[1] Machine Intelligence Laboratory, College of Computer Science, Sichuan University, 610065 Chengdu, People's Republic of China

[2] Department of Ultrasonography, Chengdu BOE Hospital, Chengdu, People's Republic of China

Table 1 The BI-RADS assessment categories and the corresponding definition

Category	Definition	Likelihood of cancer
	Need additional imaging or prior examinations	n/a
1	Negative	Essentially 0%
2	Benign	Essentially 0%
3	Probably Benign	>0% but <2%
4A	Suspicious	low suspicion for malignancy (>2% to ≤ 10%)
4B	Suspicious	moderate suspicion for malignancy (>10% to ≤ 50%)
4C	Suspicious	high suspicion for malignancy (>50% to <95%)
5	highly suspicious of malignancy	≥ 95%
6	known biopsy with proven malignancy	n/a

shows the definition of the seven BI-RADS assessment categories, with category 4 being further subdivided into 4A, 4B, and 4C, according to the probability of malignancy (Fig. 1 shows example images of the seven BI-RADS categories).

However, the manual analysis of breast ultrasound images is highly observer-dependent and requires well-trained and experienced radiologists, which leads to longer examination and imaging review times [1]. Thus, the development of automated analysis tools for breast ultrasound imaging would be helpful for improving the reliability and efficiency of diagnostic examinations. Many previous computer-aided breast ultrasound image diagnostic systems have been developed [11], with these being mostly focused on the detection of breast cancer. These previous systems have often used texture analysis methods with manual assistance to provide the estimated diagnosis. For example, using fractal analysis and k-means classification methods, reference [7] achieved an accuracy of 88.8% on a dataset consisting of 250 breast ultrasound examinations. Recently, the performance of deep neural networks for diagnosing breast cancer has been explored, and some great achievements have been made [29]; however, few of the previous studies focused on the automated assessment of BI-RADS grades, which may be a more difficult task than the binary assessment of malignancy or benignity.

In this study, we describe our novel proposed deep neural network-based methodology for the automatic assessment of the seven BI-RADS grades on breast ultrasound images. Taking a breast ultrasound image as input, the diagnostic model generates a corresponding BI-RADS assessment for radiologists and breast surgeons in a fully automated manner. The proposed model has strong consistency and good diagnostic accuracy, which suggests that it would be useful in medical institutions with ultrasonic equipment.

The creation of an automated BI-RADS assessment system for breast ultrasound images faces several challenges, which include the following.

1. The BI-RADS assessment of ultrasound images is indicated by a series of breast lesion findings including shape, orientation, margin, echo pattern, posterior features, and calcification [37]. These findings present with various appearances in different ultrasound images, rendering them difficult to extract for automated assessment.
2. To learn the features of BI-RADS grades from ultrasound images, a large number of annotated data items are required. However, the datasets used in the current related studies generally do not meet this requirement.
3. Traditional deep convolutional neural networks (CNNs) use square images as inputs [17]. The regions of interest (ROIs) on breast ultrasound images are of various sizes and aspect ratios, and directly resizing these ROI images to a fixed size could lead to deformation. The BI-RADS category assessment is very sensitive to the textural features of breast tissues, and images with unexpected deformations could result in deviations from the original images.
4. Breast lesions are presented as various sizes and shapes in ultrasound images. Extracting the features of breast lesions at various scales should increase diagnostic performance.

We overcome the above stated challenges through the following methods. (1) CNNs with a powerful abstract ability are employed to extract the features from the data. Similar CNNs have been widely applied in computer vision-related problems, and have been shown to be able to learn high-level features directly from data. (2) We constructed a large-scale dataset of ultrasound images annotated by an experienced radiologist. This large dataset contributes to the CNNs' learning of features related to the BI-RADS assessment and reduces overfitting. (3) To avoid the deformation of breast tissues and achieve higher recognition performance, a multi-input image processing method was developed. The details of this proposed multi-input method are depicted in Section 3.2. (4) We propose a multi-scale module for the assessment of breast lesions at different scales. The architecture of this proposed model is shown in Fig. 5.

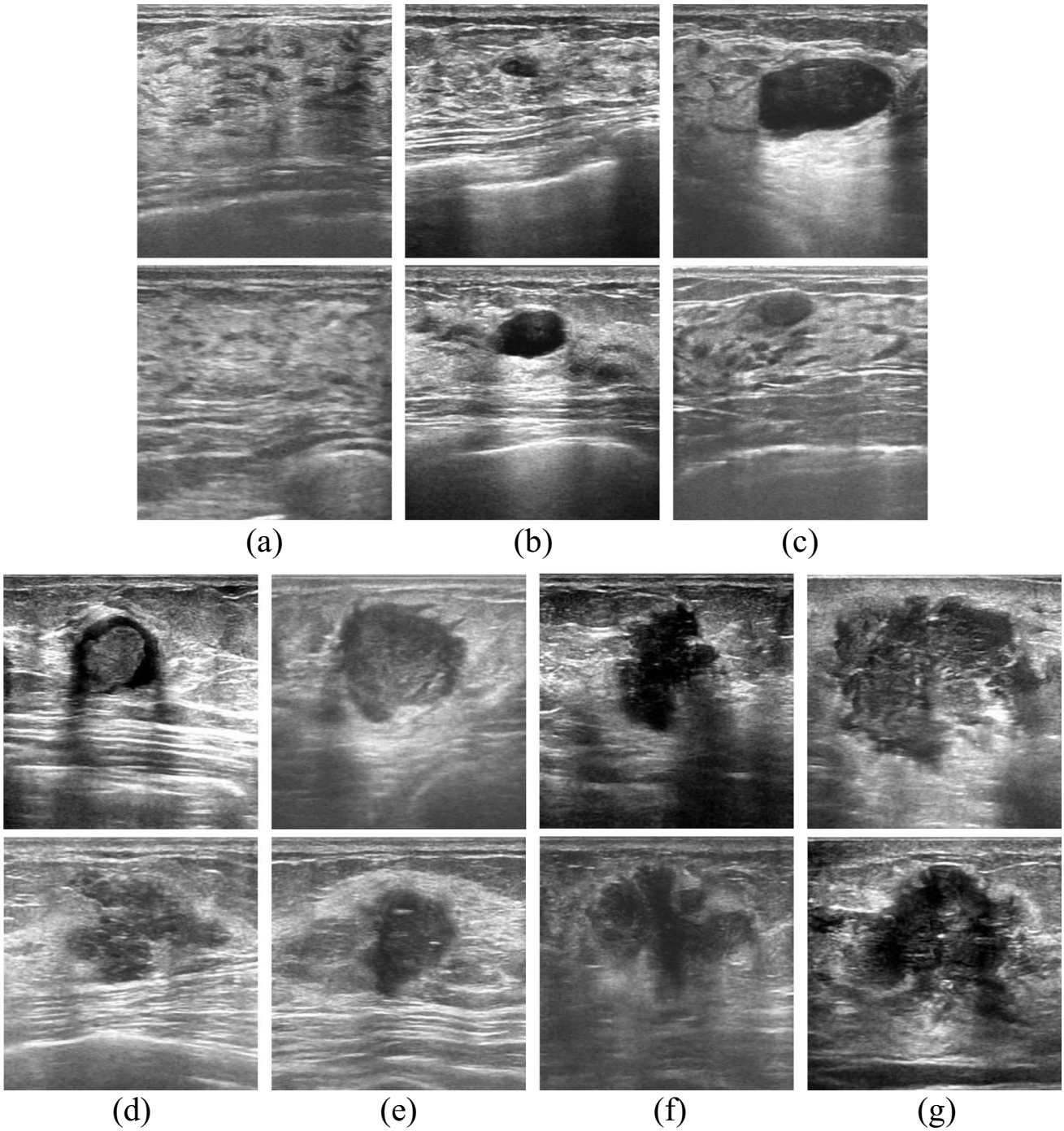


Fig. 1 Examples of breast ultrasound images from seven BI-RADS categories. Each class contains two images from different patients. **a** BI-RADS 1 ultrasound images. **b** BI-RADS 2 ultrasound images. **c** BI-RADS 3 ultrasound images. **d** BI-RADS 4A ultrasound images. **e** BI-RADS 4B ultrasound images. **f** BI-RADS 4C ultrasound images. **g** BI-RADS 5 ultrasound images. There is no lesion in the BI-RADS 1 images. It's also clear that the lesions present various appearances in different BI-RADS categories

The main contributions of this study can be summarized as follows:

1. We propose a novel BI-RADS category assessment model for breast ultrasonography images. To the best of our knowledge, this is the first report on the automated classification of the seven BI-RADS grades on breast ultrasound images.
2. We developed an ROI extraction method to extract the breast region from ultrasound images. Then, a multi-

input image pre-processing method is proposed to adapt variously sized input images.
3. A multi-scale module to capture various scale features of breast lesions. Moreover, we propose an order-constrained loss function that fully considers the continuity between categories, rendering a higher performance than traditional loss functions.
4. A large breast ultrasonography image dataset labeled by an experienced radiologist was constructed for the automatic assessment of BI-RADS grade. This dataset is larger than any previously described breast ultrasonography image dataset with BI-RADS assessment labels.

1 Related works

In this section, we present an overview of previous studies on the automated analysis of breast ultrasonography images, followed by a brief introduction to our multi-scale strategy.

1.1 Breast ultrasonography image analysis

Ultrasound imaging-based computer-aided diagnostic systems for breast lesions have been a subject of study for several decades [22], with the vast majority of publications on breast ultrasonography image diagnosis being focused on the recognition of breast cancer.

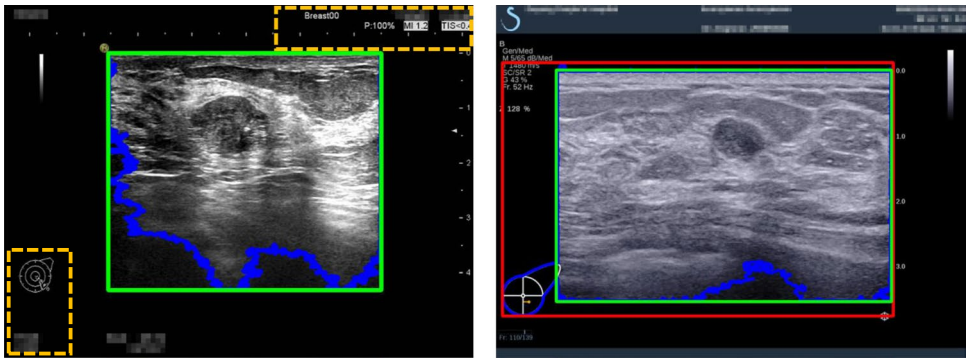
Most of the conventional publications for breast cancer diagnostic systems contain four main steps: image preprocessing, lesion detection or segmentation, feature extraction, and classification [23, 33, 43]. The lesion detection step aims to locate the areas of breast nodule within the breast ultrasonography image, following which several graphical features of the breast nodule are extracted. A classifier is then built to analyze these features and predict the breast lesion diagnosis. For example, reference [23] employed a relatively simple segmentation algorithm to segment breast nodules from manually preselected ROIs. Five morphological features were extracted from the breast nodule region and a multi-layer perceptron (MLP) neural network was then used to distinguish malignant nodules according to these features. A database of 584 histologically confirmed cases containing 300 benign and 284 malignant breast nodules was constructed. The performance of the proposed method achieved an accuracy of 91.4%, sensitivity of 92.3%, specificity of 90.7%, and an area under the curve (AUC) of 0.95. Besides conventional image features, there are some studies employ BI-RADS features in breast ultrasound images to differentiate benign and malignant tumors [33, 34].

Although conventional methods have made great achievements in the automated diagnosis of breast ultrasonography images, several obvious drawbacks remain. These conventional methods rely heavily on handed-crafted features and the segmentation/detection of the ROI. Deep neural networks, especially CNNs, which extract features in a data-driven manner, provide powerful support to alleviate the aforementioned drawbacks [2, 5]. Recently, many studies have been conducted on the automated diagnosis of breast ultrasonography images using CNNs [15, 28, 29]. Reference [29] introduced a cross-training algorithm to train two sub-networks: Mt-Net and Sn-Net. A dataset containing 2759 malignant and 5386 benign lesions was constructed to train and evaluate the model, which showed final accuracy of 94.48% and sensitivity of 95.65%.

According to these recent studies, deep convolutional neural networks (DCNNs) have shown better performance than traditional methods for the automated analysis of breast ultrasonography images. However, previous studies mainly focused on the recognition of breast cancer, with few of them focusing on the automated assessment of BI-RADS grade on breast ultrasonography images. Reference [31] developed a CAD system to classify breast ultrasound images into three classes: benign, BI-RADS 2 in benign lesions; probably benign, BI-RADS 3 and 4 in benign lesions; and malignant, BI-RADS 4 and 5 in malignant lesions. A dataset that contains 1128 images was constructed. Random forest was employed to perform prediction and showing a sensitivity of 82.60%. Reference [12] employed a DCNN with a sliding window method to categorize breast lesions into three classes: BI-RADS 2 (cysts, lymph nodes), BI-RADS 3 (non-cystic mass), and BI-RADS 4–5 (suspicious). Their method reached an average classification accuracy of 87.1% on an internal test dataset (101 images). Reference [6] proposed a CAD system to perform BI-RADS category prediction, which contains four stages: preprocessing, breast tumor segmentation, feature extraction, and classification. Four BI-RADS classes (B2, B3, B4, and B5) were considered, and the proposed method showed an accuracy of 85.42% on 48 testing images. Reference [30] developed a dual-modal neural network model to classify breast masses into four BI-RADS classes (B2, B3, B4, and B5) and achieved an accuracy of 80% on a dataset containing 2748 cases. Our previous study proposed a novel method to classify a breast nodule into benign or malignant by jointly analyzing the nodule on multiple planes [28]. Based on the previous work experience, in this study, we further explored automated BI-RADS categories assessment.

1.2 Multi-scale strategy

Objects usually present as different shapes and sizes in different images, especially in medical images. Traditional



(a) (b)

Fig. 2 Results of the ROI extraction method. The content in the yellow dashed rectangle is the metadata and the blue line is the contour of the region extracted using the basic thresholding method. **a** The green rectangle represents the ROI extracted using the basic thresholding method. **b** The red rectangle represents the ROI extracted using the basic threshold method. Some metadata is close to the ROI, causing a failure in ROI extraction. By adopting the line detecting method, the ROI is successfully extracted (green rectangle)

methods lack a combination of coarse-scale and fine-scale features, and are therefore likely to yield unsatisfactorily high error rates. In recent years, research has increasingly focused on multi-scale strategies [26, 47]. For example, to handle the problem of segmenting objects at multiple scales, Atrous Spatial Pyramid Pooling (ASPP) [8, 9] was proposed. This employs atrous convolution in cascade or in parallel to capture the multi-scale context by adopting multiple atrous rates. Several methods [10, 14, 48] have been used to apply CNNs to an image pyramid to extract features for input at each scale, with objects at different scales becoming prominent in different feature maps. GoogLeNet [39] introduced the Inception module, which employs multiple-scale convolution kernels to extract features from various scales. Inspired by these methods and the features of breast ultrasonography images, we developed a multi-scale module to capture the features of breast lesions at various scales, and describe the details of this multi-scale module in Section 3.3.

2 Methods

In this section, we describe in detail the methods developed and used in our automated BI-RADS category assessment system using breast ultrasonography images. The ROI extraction and multi-input methods are first described in detail. The image processing method provides a better understanding of the task, and we then describe and illustrate our proposed model, including the basic neural network architecture, the multi-scale module, and the order-constrained loss function. The source code of the proposed methods is available at https://github.com/yongpi-scu/BRNet.

2.1 ROI extraction

The original breast ultrasound images may contain metadata (e.g., black borders of the image, pictograms, parameter text, as shown by the yellow dashed rectangle in Fig. 2a), which may disturb the model's recognition ability. In this study, we developed a thresholding method to extract the breast region and eliminate metadata interference. The basic thresholding method can be described as follows: conversion of the original image to a binary image using a fixed constant threshold, obtaining all contours of the binary image, and finding the largest contour representing the ROI. For partial ultrasound images, this basic thresholding method is sufficient to extract the region of interest (ROI), as shown in the example in Fig. 2a, where the ROI is extracted perfectly.

However, there are also many images where the metadata intersects the ROI leading to the extracted region being bigger than the ROI, as shown by the red rectangle in Fig. 2b. In this case, it can be noticed that the contour of the ROI is mostly straight, and a straight-line detection method was therefore developed to find a better ROI contour. The total ROI extraction method developed is presented in Algorithm 1. The *findContours*, *ContourArea*, and *boundingRect* functions used in the algorithm were built-in functions of the Open Source Computer Vision Library [3]. Algorithm 2 presents the straight-line detection method for finding the top coordinate value of the ROI. The methods to find the left, right, and bottom coordinate values are omitted, as they are similar to the method for finding the top coordinate value. The *unique* function in this algorithm is the built-in function of the NumPy library [16]. The extracted result is shown in Fig. 2b (green rectangle), and using this improved method, the ROI can be extracted perfectly.

Algorithm 1: Thresholding method for breast region extraction in ultrasound images.

Input : Original breast ultrasound image I, constant threshold T.
Output: Extracted ROI image I_{roi}.

```
begin
    for i in height of I do
        for j in width of I do
            if I[i, j] > T then
                I[i, j] = 255
            end
            else
                I[i, j] = 0
            end
        end
    end
    L_cnts = findContours(I)
    area_max = 0, cnt_max = None
    for cnt in L_cnts do
        area = ContourArea(cnt)
        if area > area_max then
            area_max = area
            cnt_max = cnt
        end
    end
    left, top, right, bottom = DetectLine(cnt_max)
    I_roi = I[left : right, top : bottom]
    return I_roi.
end
```

Algorithm 2: Straight line detection method.

Input : Contour of ROI, cnt_{max}.
Output: left, top, right and bottom coordinate values of I_{roi}.

```
begin
    x, y, w, h = boundingRect(cnt_max)
    center_x = x + w/2, center_y = y + h/2
    points = [], values = []
    // find the top coordinate value.
    for point in cnt_max do
        if point.y > center_y then
            points.append(point.y)
        end
    end
    // Statistic frequency of y values.
    unique, counts = unique(points)
    for u, c in zip(unique, counts) do
        if c > 100 then
            values.append(u)
        end
    end
    if max(values) − min(values) < 5 then
        top = mean(values)
    end
    else
        top = y
    end
    ··· // find the left, right, and bottom coordinate values.
    return left, right, top, bottom.
end
```

2.2 Multi-input

In this study, the collected breast ultrasonography images were of a resolution of 1024×768 pixels. As the images were captured with different kinds of instruments and parameters, the extracted ROI images were of various sizes and aspect ratios. The resolution of ROI images in the constructed dataset was in the range of 334×256 to 738×597 pixels, with aspect ratios in the range of [0.74, 1.78].

Although in the theory, CNNs can be trained with standard back-propagation, regardless of the input image size, in practice, graphic processing unit (GPU) implementations tend to be run on fixed-size input images to accelerate the training speed [17]. Conventional classification CNNs typically take square images as input, such as the input size of 299×299 pixels used by DenseNet. Confronted with images of various sizes and aspect ratios, the most common image processing method is to resize the original image to a fixed size, causing deformation of the original images. However, the assessment of BI-RADS grade is very sensitive to the textural features of breast tissues, and images with unexpected deformation would result in deviations from the original images.

To avoid the deformation of breast issues and achieve higher recognition performance, we propose a multi-input image processing method. The smallest edge of the ROI image is first resized to 299 with the aspect ratio maintained. Then, four square images are cropped from the image, with these being the left, center, right, and resized whole images. The proposed method is illustrated in Fig. 3. After image processing, the four square images are stacked at the channel of the model inputs.

2.3 Multi-scale module

Breast lesions in ultrasonography images may present with various sizes and shapes, and a good combination of features at different scales is likely to render a higher classification performance. Some studies have developed multi-scale strategy to handle the problem of segmenting objects at multiple scales [8, 9, 48]. Motivated by these works, we here propose a multi-scale module for assessing breast lesions at different scales. Different from the strategy employed by reference [48], which upon the final-layer-feature-map and using adaptive average pooling. In this study, we design our multi-scale module upon the input layer and using global average pooling to extract features at specified scales. The overall architecture of the proposed multi-scale module is presented in Fig. 4.

As shown in Fig. 4, suppose the input is x ($x \in \Re^{4\times h\times w}$), where the 4, h, and w denote the channels, height, and width of the input, respectively. The module extracts features from the input at three scales, which can be formulated as follows:

$$\begin{aligned} F_1^l &= f(x), \\ F_1^m &= f(g(x,2)), \\ F_1^s &= f(g(x,4)), \end{aligned} \tag{1}$$

where the l, m, s indicate large, medium, and small feature maps, respectively. g denotes the global average pooling layer, and f denotes the convolutional operation. The second parameter of g is stride size. The extracted feature maps are then upsampled to the original resolution. This process can be formulated as follows:

$$\begin{aligned} F_2^l &= f(F_1^l) \\ F_2^m &= u(F_1^m, 2) \\ F_2^s &= u(F_1^s, 4) \end{aligned} \tag{2}$$

where u denotes the upsample operation and the second parameter of the upsample function is the upscale factor. Recent studies have indicated that the bilinear interpolation operation shows better performance than deconvolution for the upsample operation [8, 45]. Thus, we employed bilinear interpolation as the upsample method in this study, instead of deconvolution, because it is faster and parameter-free. Finally, the feature maps are concatenated along the channel, and a convolution layer with a kernel size of 3 and stride of 2 is employed to convert the channel from $7c$ to $2c$. This step can be defined according to the following:

$$G = f(z(F_2^l, F_2^m, F_2^s)) \tag{3}$$

where z denotes the concatenation operation and G denotes the final output feature maps of the multi-scale module, $G \in \Re^{2c\times 0.5h\times 0.5w}$.

2.4 Basic neural network architecture

The overall architecture of the proposed method is depicted in Fig. 5. After image preprocessing, the input is first fed to a multi-scale module. Then, a basic classifier is employed to classify the extracted features. Deep neural networks have received much interest in the field of medical image analysis, and the model explored in this study is based on CNNs. Currently, fine-tuning with pre-trained networks has been proved as an effective method for training CNNs [24]. Pre-trained networks are especially useful in medical image analysis, because the number of labeled medical images is usually limited [41]. In this study, several standout pre-trained networks were explored, including InceptionV3 [40], InceptionResNetV2 [38], DenseNet169 [21], and ResNet50 [18]. These networks

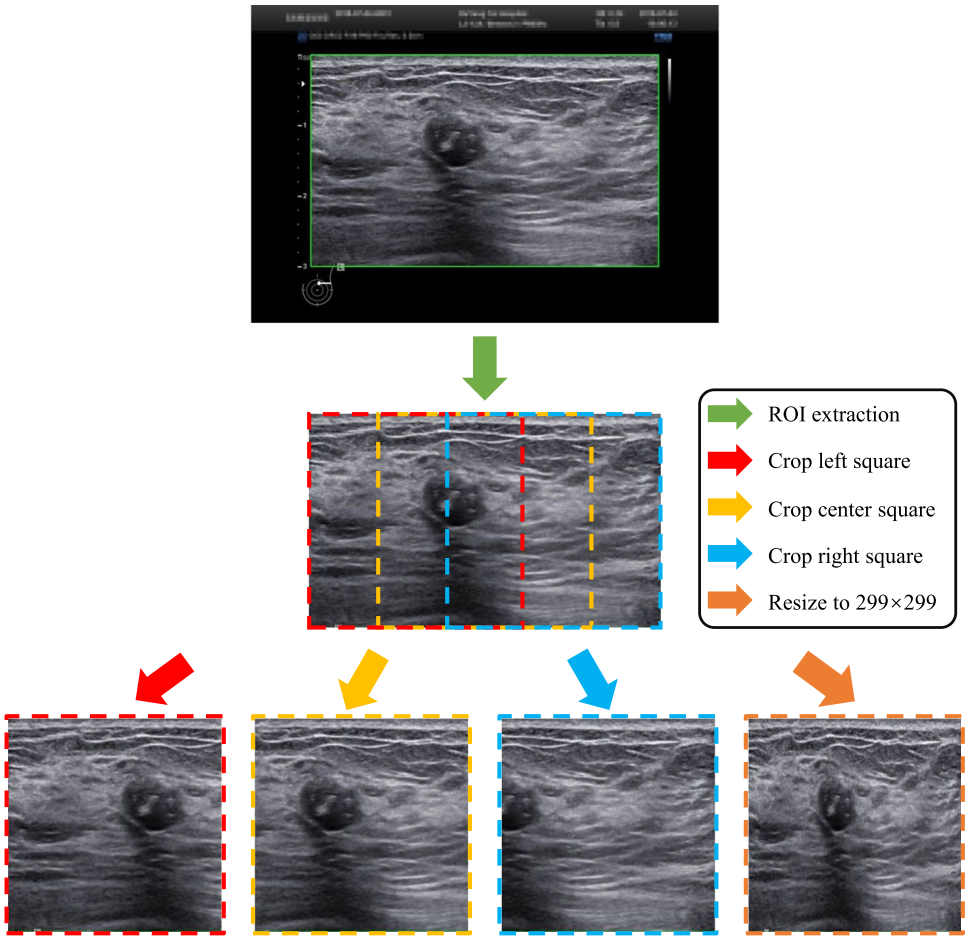


Fig. 3 Developed image processing method. The ROI is first extracted from the original image and the smaller edge is resized to 299 pixels with the aspect ratio maintained. Then, four square images are cropped from the ROI image

were first trained on ImageNet [13] and the parameters of the pre-trained models were employed to initialize the models in this study. The first convolution layer of these networks was removed and the cascade of the proposed multi-scale module for BI-RADS assessment was used instead.

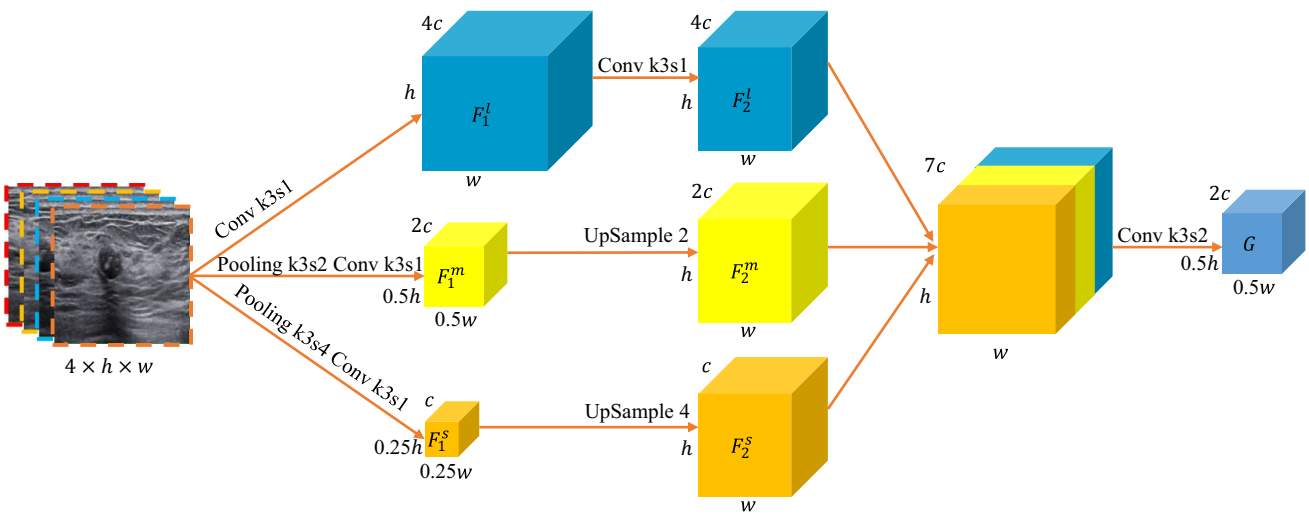


Fig. 4 Overall architecture of the proposed multi-scale module. The module extracts features from the input at three scales and these features are concatenated at the channel. *c*, *h* and *w* denote the channels, height, and width of the feature maps, respectively

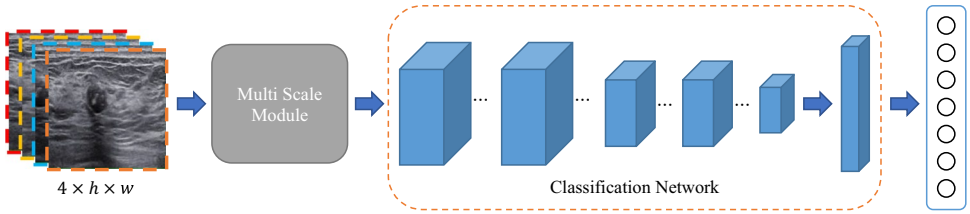


Fig. 5 Proposed architecture for the automated assessment of BI-RADS categories. A breast ultrasound image is first preprocessed with the ROI extraction and multi-input methods. A multi-scale module is used to extract features from the input and these extracted features are then fed to a pre-trained classification neural network that outputs the final predictions

2.5 Order-constrained loss function

Conventional classification tasks typically employ cross-entropy loss as the loss function, which can be formulated as:

$$L_1 = -\sum_{i=1}^{N} y_i log(p_i), \tag{4}$$

where N denotes the number of elements in the model prediction vector P. p_i denotes the ith element of the model prediction P and y_i represents the ith element of the ground truth Y. However, the assessed BI-RADS grades are sequential, and an image may be classified into its neighboring BI-RADS grades. For example, in routine clinical scenarios, a breast ultrasound image belonging to B3 may be assessed as B2 to B4a, but is most unlikely to be classified as B1 or B5. In this respect, the cross-entropy function does not well reflect the characteristics of the BI-RADS grade evaluation. Therefore, we propose an order-constrained loss function that can boost the classification performance. This function is defined as:

$$L_2 = -(\arg\max_i P - \arg\max_j Y)^2 log(1 - max(P)). \tag{5}$$

The coefficient $(\arg\max_i P - \arg\max_j Y)^2$ in this function can well depict the distance between the prediction class and the ground truth. The farther between the prediction class and the ground truth, the bigger is the loss value. The final loss function used in this study is a combination of cross entropy-loss and the proposed order-constrained loss, and can be defined as follows:

$$L = L_1 + \alpha L_2, \tag{6}$$

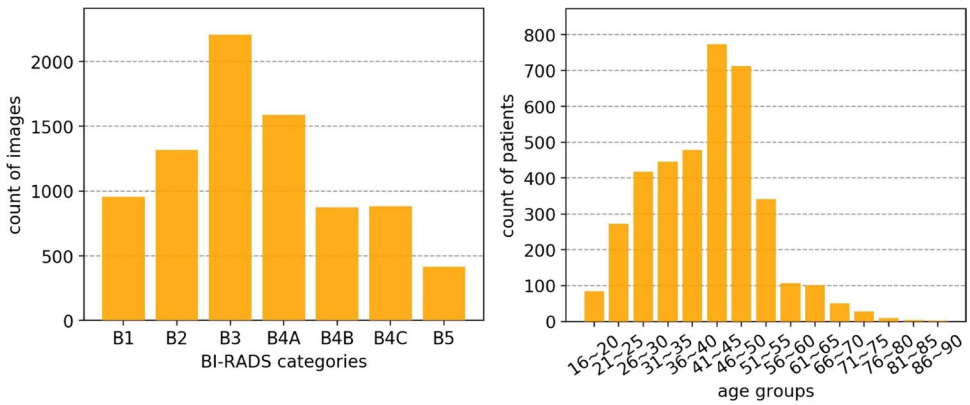
where α is a weight parameter for the order-constrained loss and is set to 0.1 for all experiments in this study. Section 5.2 analyzes the effectiveness of the proposed order-constrained term and Section 5.3 presents the selections of α.

3 Experimental setup

In this section, the experimental setup is presented in detail, including the dataset, implementation details, and evaluation metrics.

3.1 Materials

The proposed model was evaluated on an in-house collected dataset. The images used in this study were captured by several sonographers during ultrasound examinations performed using different kinds of color doppler instruments, including GE LOGIQ E9, Philips EPIQ 7c, Philips iU22, and Hitachi



(a) (b)

Fig. 6 Histograms of the distribution of patient age and the number of different BI-RADS categories in the dataset. **a** Distribution of BI-RADS categories. **b** Distribution of age groups

Table 2 Dataset constructed for training, validation, and testing the proposed models

Subsets	B1	B2	B3	B4A	B4B	B4C	B5	Total
Training set	669	922	1548	1113	614	620	293	5779
Validation set	95	131	220	158	87	88	41	820
Testing set	191	263	441	317	175	177	83	1647
Total	955	1316	2209	1588	876	885	417	8246

Ascendus. All images were stored in JPEG format with a resolution of 1024×768 pixels. Each examination contained a variable number of images, and all the images in each breast examination were collected. Over 40,000 images from 8217 examinations performed between January 2017 and March 2019 were collected. For each examination, both the corresponding ultrasound examination and the pathological report were collected.

As the evaluation of the BI-RADS category is somewhat subjective, different annotators may give different results for the same image. In this study, all images were annotated by one radiologist with over 10 years of experience, to provide a high level of consistency in the labeling. In the annotation process, the image labels were determined by the radiologist according to the corresponding ultrasound examination. The annotator performed the annotation process for no more than 2 hours a day, to avoid the influence of annotator fatigue on labeling consistency. Images containing blood flow signals were excluded, because the signals could occlude features and influence the diagnosis. Finally, 8246 images from 3825 examinations originating from 3825 patients were labeled (age: 41.06 ± 11.20). Figure 6 illustrates the distributions of patient age and the number of different BI-RADS categories in the dataset constructed in this study.

To assess the performance of the proposed methods, the collected dataset was randomly split into three parts: training set, validation set, and testing set. The number of images in each class was uniformly distributed across the subsets. The partition of three subsets is patient-dependent to ensure that images from a patient belong to the same subset. Further detailed statistics of each subset are shown in Table 2.

3.2 Implementation details

For this study, all models were trained using Adam [25] as the optimizer with a learning rate of 10^{-3} and a weight decay rate of 10^{-4} for 200 epochs. The mini-batch size was fixed at 16. To reduce the side effect of overfitting, dropout [36] was applied to the last fully connected layer with a drop probability of 0.7. The algorithms were implemented using PyTorch [27]. All experiments were performed on a high-end workstation with a Linux operating system and the following hardware: an Intel Xeon E5-2620 CPU, three NVIDIA Tesla P100 GPUs, and 64 GB of RAM.

Data augmentation is an important trick, widely used to help deep convolutional neural networks learn invariant features, prevent them from overfitting, and improve their generalization ability. This study used two online data augmentation methods (random horizontal flipping and mixup [46]) on the proposed method and baseline networks.

3.3 Evaluation metrics

The training process is carried out on the training set, while the validation set is used to fine-tune the model. The overall performance of each methods is assessed on the testing set. In this study, we use accuracy, sensitivity, precision, and F_1 score as evaluation criteria. These metrics are defined as follow:

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + FP + TN + FN} \\ Sensitivity &= \frac{TP}{TP + FN} \\ Precision &= \frac{TP}{TP + FP} \\ F_1 &= \frac{2TP}{2TP + FP + FN} \end{aligned} \tag{7}$$

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

4 Results and discussion

In this section, the experimental results of the proposed methods are compared with four basic classification networks, and two related methods are first described and discussed. Then, the results of ablation experiments are given and analyzed in detail. Finally, parameter selections experiment is discussed in detail.

4.1 Comparison results

We now introduce the experimental results of our proposed method for the classification of the seven BI-RADS categories and compare them with two related studies (Han et al. [15] and Xie et al. [44]) and four well-known classifiers: InceptionV3, InceptionResNetV2, DenseNet169, and ResNet50. For convenience, we use the term

Table 3 Experimental results for the classification of seven BI-RADS categories. All metrics are macro-averaged on seven categories

Methods	Accuracy	Sensitivity	Precision	F_1
InceptionV3	74.20%	72.15%	72.00%	71.88%
IncepResV2	73.95%	73.48%	72.88%	72.73%
DenseNet169	64.12%	57.39%	64.61%	58.13%
ResNet50	68.79%	65.38%	67.90%	66.03%
InceptionV3-BRNet	**77.35%**	**75.37%**	**77.75%**	**76.03%**
IncepResV2-BRNet	75.41%	72.90%	73.07%	72.88%
DenseNet169-BRNet	66.55%	62.35%	65.23%	61.22%
ResNet50-BRNet	70.86%	66.85%	71.04%	68.18%
Han et al. [15]	73.71%	69.47%	71.82%	70.10%
Xie et al. [44]	74.92%	72.99%	74.98%	72.79%

The top results are highlighted in bold

"InceptionV3-BRNet" for our proposed method using InceptionV3 as the feature extraction network. The experimental results are depicted in Table 3 and Fig. 7. First, we compared the performance between the four basic networks. As shown in the table, InceptionV3 exhibited the highest performance of the four basic networks, achieving an accuracy of 74.20%. The high classification performance of InceptionV3 benefited from the multi-size kernels used in the Inception block. As the size and shape of breast lesions changes between different images, networks with multi-size kernels can efficiently extract such features. However, the basic classifiers showed only limited ability to assess the seven BI-RADS categories. Our proposed method employing InceptionV3 as the feature extraction network achieved the highest performance in all four metrics, with an accuracy of 77.35%, sensitivity of 75.37%, precision of 77.75%, and an F_1 score of 76.03%. The accuracy of InceptionV3-BRNet was 3.15% higher than that of InceptionV3, showing the effectiveness of the proposed methods. Moreover, all proposed methods showed higher performance than the corresponding basic networks, thereby demonstrating their effectiveness.

Compared with the methods in Han et al. [15] and Xie et al. [44], our method also exhibited superior performance in all metrics. The accuracy of our proposed method (InceptionV3-BRNet) was 2.43% higher than that of [44], which further demonstrate the effectiveness and robustness of our proposed method.

Table 4 demonstrates the detailed performance of InceptionV3-BRNet for each BI-RADS category. It can be seen that our method shows the highest recall for the B3 category and the lowest recall for the B4B category. These results are also similar to human radiologists. B4B category images may be easily misclassified to B4A or B4C, and we will explore ways to improve the performance of the B4B category classifications in a future study.

4.2 Ablation experiments

In this study, three operations were proposed to improve the classification performance, with these being multi-inputs, a multi-scale module, and order-constrained loss. We performed ablation experiments to validate the comparative effectiveness of these three operations, with the baseline network used being InceptionV3. Table 5 demonstrates the results produced by InceptionV3-BRNet without a

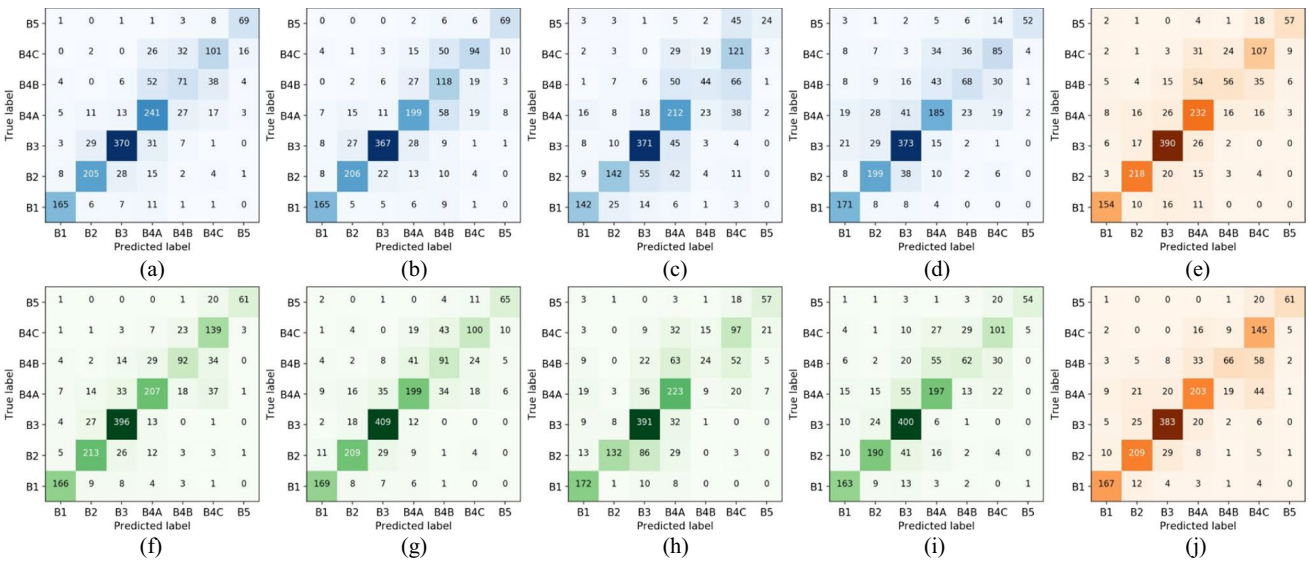


Fig. 7 The confusion matrices of the proposed and compared methods on the seven BI-RADS categories classification task. **a** InceptionV3. **b** IncepResV2. **c** DenseNet169. **d** ResNet50. **e** Han et al. [15]. **f** InceptionV3-BRNet. **g** IncepResV2-BRNet. **h** DenseNet169-BRNet. **i** ResNet50-BRNet. **j** Xie et al. [44]

Table 4 The performance of InceptionV3-BRNet for each BI-RADS category

Metrics	B1	B2	B3	B4A	B4B	B4C	B5
Recall	86.91%	80.99%	89.80%	65.30%	52.57%	78.53%	73.49%
Precision	88.30%	80.08%	82.50%	76.10%	65.71%	59.15%	92.42%
Specificity	98.49%	96.17%	93.03%	95.11%	96.74%	93.47%	99.68%
Accuracy	97.15%	93.75%	92.17%	89.37%	92.05%	91.86%	98.36%
F_1	87.60%	80.53%	85.99%	70.29%	58.41%	67.48%	81.88%

Table 5 Effects of the multi-input method, multi-scale module, and order-constrained loss. The baseline network is InceptionV3. Case 4 is the complete InceptionV3-BRNet

Cases	Multi-input	Multi-scale	Order-constrained loss	Accuracy
Baseline	-	-	-	74.20%
1	×	√	√	76.14%
2	√	×	√	76.02%
3	√	√	×	75.41%
4	√	√	√	77.35%

specific operation. For convenience, we refer to the InceptionV3-BRNet lacking one operation as the incomplete InceptionV3-BRNet.

The following conclusions were evident from the experimental results. First, although the incomplete InceptionV3-BRNets had one of the three operations removed, they still showed better performance than the baseline network. Second, all three of the incomplete InceptionV3-BRNets showed lower performance than the complete InceptionV3-BRNet. Third, the order-constrained loss provided the largest performance improvement of the three operations. Finally, all three operations were effective, and each of them provided an improvement in the final performance.

Fixed resize ROI to the square image will cause deformation of breast lesions, which may cause a degradation of performance. It can be seen from Table 5 that by replacing the multi-input strategy with fixed resize strategy, the accuracy dropped from 77.35% to 76.14%. Random crop is also a popular strategy that can avoid ROI deformation, but may lose some lesion area after crop. The proposed multi-input strategy aims to retain all features of the ROI without causing deformation. We conducted an experiment by replacing the multi-input strategy with random crop. Random crop strategy showing an accuracy of 76.50%, which is outperformed fixed resize but still worse than our developed multi-input strategy.

4.3 Parameter selections

In this subsection, extensive hyper-parameter experiments are empirically conducted to analysis the effect of hyper-parameter α in the proposed order-constrained loss, which assists to determine the optimal parameter. As the common practices, we determine α through roughly searching in the

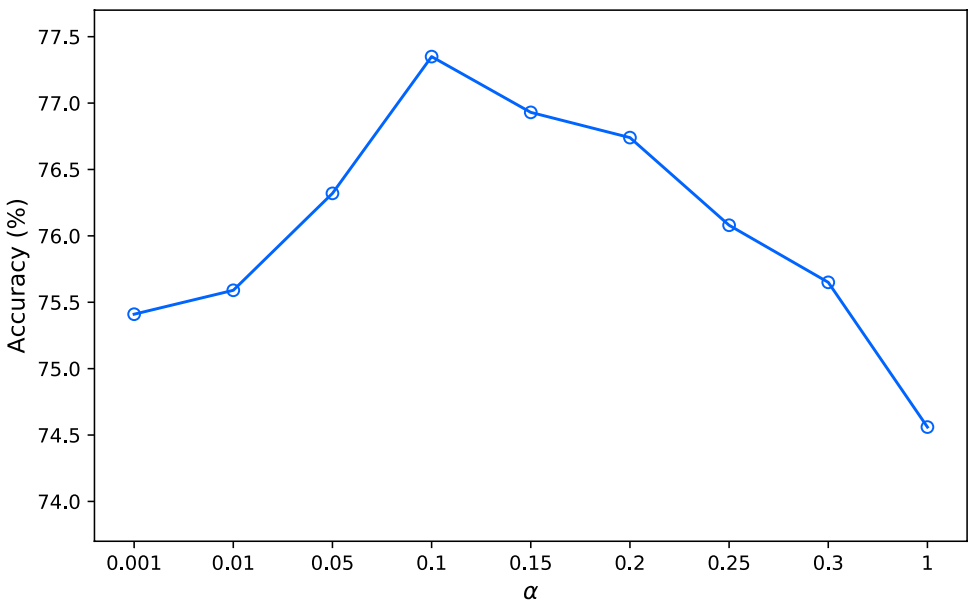


Fig. 8 The parameter selections about the hyper-parameter α in the proposed order-constrained loss

range with a specific step. The InceptionV3-BRNet is utilized to perform different values of the hyper-parameter α. Notice that our model produces the highest accuracy 77.35% when $\alpha = 0.1$, as shown in Fig. 8.

5 Conclusion

In this paper, we propose a novel architecture for a CNN for the automated assessment of BI-RADS category on breast ultrasound images. This CNN could improve the efficiency and reliability of breast ultrasound examinations. An ROI extraction method was first developed to extract the breast region from ultrasound images, and then a multi-input method was proposed to avoid the deformation of ROI images. Additionally, a multi-scale module was developed to capture features of breast lesions at various scales. Finally, we proposed an order-constrained loss function that fully considers the continuity between BI-RADS categories, and that showed higher performance than traditional loss functions. A large-scale dataset containing 8246 labeled breast ultrasound images was constructed to verify the proposed methods. To the best of our knowledge, this is the first such dataset to be annotated with the seven BI-RADS categories.

Our experimental results demonstrate that our method could be used to mimic experienced radiologists in the assessment of the BI-RADS category of breast ultrasound images, and that these interpretations should be acceptable for routine clinical breast US examination reports. Moreover, the proposed model outperformed two comparison methods, verifying the effectiveness of our proposed methods. Most of the images collected for this study were not annotated because the annotation is exhausting and time-consuming; in the future, we will explore a better semi-supervised learning method to use the rest of the images without labels and render improved performance.

Acknowledgements This work was financially supported by the National Major Science and Technology Projects of China (2018AAA0100201) and the Sichuan Science and Technology Program of China (2020JDRC0042).

References

1. Abdullah N, Mesurolle B, El-Khoury M, Kao E (2009) Breast imaging reporting and data system lexicon for us: interobserver agreement for assessment of breast masses. Radiology 252(3):665–672
2. Albarqouni S, Baur C, Achilles F, Belagiannis V, Demirci S, Navab N (2016) Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. IEEE transactions on medical imaging 35(5):1313–1321
3. Bradski G (2000) The OpenCV Library. Dr. Dobb's Journal of Software Tools
4. 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 journal for clinicians 68(6), 394–424
5. Carneiro G, Nascimento J, Bradley AP (2017) Automated analysis of unregistered multi-view mammograms with deep learning. IEEE transactions on medical imaging 36(11):2355–2365
6. Chang YW, Chen YR, Ko CC, Lin WY, Lin KP (2020) A novel computer-aided-diagnosis system for breast ultrasound images based on bi-rads categories. Applied Sciences 10(5):1830
7. Chen DR, Chang RF, Chen CJ, Ho MF, Kuo SJ, Chen ST, Hung SJ, Moon WK (2005) Classification of breast ultrasound images using fractal feature. Clinical imaging 29(4):235–245
8. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence 40(4):834–848
9. Chen, LC, Papandreou, G, Schroff, F, Adam, H (2017) Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587
10. Chen LC, Yang Y, Wang J, Xu W, Yuille AL (2016) Attention to scale: Scale-aware semantic image segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3640–3649
11. Cheng HD, Shan J, Ju W, Guo Y, Zhang L (2010) Automated breast cancer detection and classification using ultrasound images: A survey. Pattern recognition 43(1):299–317
12. Ciritsis A, Rossi C, Eberhard M, Marcon M, Becker AS, Boss A (2019) Automatic classification of ultrasound breast lesions using a deep convolutional neural network mimicking human decision-making. European radiology 29(10):5458–5468
13. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee, pp 248–255
14. Farabet C, Couprie C, Najman L, LeCun Y (2012) Learning hierarchical features for scene labeling. IEEE transactions on pattern analysis and machine intelligence 35(8):1915–1929
15. Han S, Kang HK, Jeong JY, Park MH, Kim W, Bang WC, Seong YK (2017) A deep learning framework for supporting the classification of breast lesions in ultrasound images. Physics in Medicine & Biology 62(19):7714
16. ...Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del R'1o JF, Wiebe M, Peterson P, G'erard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE (2020) Array programming with NumPy. Nature 585(7825):357–362
17. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence 37(9):1904–1916
18. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
19. Hellquist BN, Duffy SW, Abdsaleh S, Björneld L, Bordás P, Tabár L, Viták B, Zackrisson S, Nyström L, Jonsson H (2011) Effectiveness of population-based service screening with mammography for women ages 40 to 49 years: evaluation of the swedish mammography screening in young women (scry) cohort. Cancer 117(4):714–722

20. Hooley RJ, Scoutt LM, Philpotts LE (2013) Breast ultrasonography: state of the art. Radiology 268(3):642–659
21. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
22. Jalalian A, Mashohor SB, Mahmud HR, Saripan MIB, Ramli ARB, Karasfi B (2013) Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. Clinical imaging 37(3):420–426
23. Joo S, Yang YS, Moon WK, Kim HC (2004) Computer-aided diagnosis of solid breast nodules: use of an artificial neural network based on multiple sonographic features. IEEE transactions on medical imaging 23(10):1292–1300
24. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F et al (2018) Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172(5):1122–1131
25. Kingma, DP, Ba, J (2014) Adam: A method for stochastic optimization. arXiv:1412.6980
26. Nah S, Hyun Kim T, Mu Lee K (2017) Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3883–3891
27. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A (2017) Automatic differentiation in pytorch. In: NIPS-W
28. Pi Y, Chen Y, Deng D, Qi X, Li J, Lv Q, Yi Z (2020) Automated diagnosis of multi-plane breast ultrasonography images using deep neural networks. Neurocomputing
29. Qi X, Zhang L, Chen Y, Pi Y, Chen Y, Lv Q, Yi Z (2019) Automated diagnosis of breast ultrasonography images using deep neural networks. Medical image analysis 52:185–198
30. Qian X, Zhang B, Liu S, Wang Y, Chen X, Liu J, Yang Y, Chen X, Wei Y, Xiao Q, et al (2020) A combined ultrasonic b-mode and color doppler system for the classification of breast masses using neural network. European Radiology pp. 1–11
31. Rodríguez-Cristerna A, Gómez-Flores W, de Albuquerque Pereira WC (2018) A computer-aided diagnosis system for breast ultrasound based on weighted bi-rads classes. Computer Methods and Programs in Biomedicine 153:33–40
32. Roganovic D, Djilas D, Vujnovic S, Pavic D, Stojanov D (2015) Breast mri, digital mammography and breast tomosynthesis: comparison of three methods for early detection of breast cancer. Bosnian journal of basic medical sciences 15(4):64
33. Shan J, Alam SK, Garra B, Zhang Y, Ahmed T (2016) Computer-aided diagnosis for breast ultrasound using computerized bi-rads features and machine learning methods. Ultrasound in medicine & biology 42(4):980–988
34. Shen WC, Chang RF, Moon WK, Chou YH, Huang CS (2007) Breast ultrasound computer-aided diagnosis using bi-rads features. Academic radiology 14(8):928–939
35. Sickles EA, D'Orsi CJ, Bassett LW, Appleton CM, Berg WA, Burnside ES, et al (2013) Acr bi-rads® atlas, breast imaging reporting and data system. Reston, VA: American College of Radiology pp. 39–48
36. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15(1):1929–1958
37. Stavros AT (2004) Breast ultrasound. Lippincott Williams & Wilkins
38. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence
39. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
40. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
41. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J (2016) Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging 35(5):1299–1312
42. Wang L (2017) Early diagnosis of breast cancer. Sensors 17(7):1572
43. Wei M, Wu X, Zhu J, Liu P, Luo Y, Zheng L, Du Y (2019) Multi-feature fusion for ultrasound breast image classification of benign and malignant. In: 2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC). IEEE, pp 474–478
44. Xie J, Song X, Zhang W, Dong Q, Wang Y, Li F, Wan C (2020) A novel approach with dual-sampling convolutional neural network for ultrasound image classification of breast tumors. Physics in Medicine & Biology 65(24):245001
45. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: European conference on computer vision. Springer, pp 818–833
46. Zhang, H, Cisse, M, Dauphin, YN, Lopez-Paz, D (2017) mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412
47. Zhang R, Shen J, Wei F, Li X, Sangaiah AK (2017) Medical image classification based on multi-scale non-negative sparse coding. Artificial intelligence in medicine 83:44–51
48. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2881–2890

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com