Deep Neural Networks With Region-Based Pooling Structures for Mammographic Image Classification

Xin Shu, Lei Zhang, *Member, IEEE*, Zizhou Wang, Qing Lv, and Zhang Yi, *Fellow, IEEE*

***Abstract*— Breast cancer is one of the most frequently diagnosed solid cancers. Mammography is the most commonly used screening technology for detecting breast cancer. Traditional machine learning methods of mammographic image classification or segmentation using manual features require a great quantity of manual segmentation annotation data to train the model and test the results. But manual labeling is expensive, time-consuming, and laborious, and greatly increases the cost of system construction. To reduce this cost and the workload of radiologists, an end-to-end full-image mammogram classification method based on deep neural networks was proposed for classifier building, which can be constructed without bounding boxes or mask ground truth label of training data. The only label required in this method is the classification of mammographic images, which can be relatively easy to collect from diagnostic reports. Because breast lesions usually take up a fraction of the total area visualized in the mammographic image, we propose different pooling structures for convolutional neural networks(CNNs) instead of the common pooling methods, which divide the image into regions and select the few with high probability of malignancy as the representation of the whole mammographic image. The proposed pooling structures can be applied on most CNN-based models, which may greatly improve the models' performance on mammographic image data with the same input. Experimental results on the publicly available INbreast dataset and CBIS dataset indicate that the proposed pooling structures perform satisfactorily on mammographic image data compared with previous state-of-the-art mammographic image classifiers and detection algorithm using segmentation annotations.**

***Index Terms*— Mammographic image, breast cancer, deep neural networks.**

I. Introduction

ACCORDING to cancer statistics [1], breast cancer is one of the most frequently diagnosed cancers and has the highest incidence among women in the world. The research of WHO [2] points out that early breast detection is still the basis of breast cancer control, which can improve the prognosis and survival rate of breast cancer.

Mammography, the most common tool used in early detection and diagnosis, has been proven to be effective in reducing breast cancer mortality [3]. In a standard mammographic screening examination, two views are acquired of each breast. Radiologists check these images and characterize them, following the standard Breast Imaging Reporting and Data System (BI-RADS). However, inspecting progress is tedious and tiring. More importantly, the diagnostic accuracy is related to the technical level, experience, and even the mental state of the radiologist [4]. For these realistic limitations, a computer aided diagnosis (CAD) system was suggested as an adjunct reader to help radiologists improve the performance of the mammographic screening process. [5]

A considerable amount of effort has been put into the development of CAD systems on mammographic image screening for detecting abnormalities, and some systems have been put into practice [6]–[9].

Several studies in the last decade have suggested that current CAD technologies cannot help improve radiologists' work as expected in everyday practice in the United States [8], [10], [11]. These traditional CAD approaches for mammography typically focus on describing the raw image using hand-crafted features (e.g. texture, color, shape, etc.), followed by a machine learning classifier taking these features as input [7], [9], [12]. In this process, features have to be meticulously designed according to the specific data, which considerably depends upon, and is limited by, the designer's experience. Hand-crafted annotations mean the cost of time to process the image and additional experiments to verify the suitability of these features. With the great success of deep neural networks(DNNs) in the field of computer vision, several deep learning approaches have been explored to address the automatic classification of lesions

Manuscript received December 19, 2019; accepted January 15, 2020. Date of publication January 21, 2020; date of current version June 1, 2020. This work was supported by in part by the National Natural Science Foundation of China under Grant 61772353, in part by the Foundation for Youth Science and Technology Innovation Research Team of Sichuan Province under Grant 2016TD0018, and in part by the Sichuan University Innovation Spark Project Library under Grant 2018SCUH0040. *(Corresponding author: Lei Zhang.)*

Xin Shu, Lei Zhang, Zizhou Wang, and Zhang Yi are with the College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: shuxin@stu.scu.edu.cn; leizhang@scu.edu.cn; wangzizhou@stu.scu.edu.cn; zhangyi@scu.edu.cn).

Qing Lv is with the Department of Galactophore Surgery, West China Hospital, Sichuan University, Chengdu 610041, China (e-mail: lqlq1963@163.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2020.2968397

in mammography [9], [13], [14]. Many of these proposed approaches treat the problem as a segmentation or detection task, for instance, detecting a region of interest (ROI) and defining lesion boundaries in different stages [15]. Training these models requires a large number of annotations, including bounding boxes or segmentation ground truths in the training set. Unfortunately, annotating mammogram also need experienced radiologists with expert domain knowledge put a significant amount of effort to ensure the accuracy, which will greatly increase the workload of radiologists, and when the amount of data is large, it will be difficult to gather enough experienced radiologists to complete the work. The need for reduction of requirement for annotation in deep learning approaches to the field of mammographic image analysis is a significant problem to be overcome.

Another key challenge in mammographic image analysis is the substantial difference between mammographic images and RGB images, which makes it difficult to apply classification models with good performance on RGB images to mammographic images. Masses in breasts are typically dense or isodense, thus it has the characteristics of pixel intensity from gray to white. Geometrically, they can be oval, irregular, or round in shape with spiculated, circumscribed, obscured or ill-defined margins. [12], [16]. Due to the high variability of breast lesions, the complex appearance of the lesions makes it difficult for even professional radiologists to detect and classify them. On the other hand, the cancerous status of a full large mammographic image is determined by attributes of lesions, but the area of lesions only makes up a tiny percentage of this large image. Statistics in several studies show that a mammographic image in the size of 5500 (height by width) pixels often contains a mass or calcification clusters in the size of 100 pixels [17]–[19]. Therefore, the analysis of mammographic images is much more difficult. Recent widely used CNN methods in the field of computer vision are mainly designed for classification tasks on common image data sets [20]–[22]. They typically take a standard small input of size 224×224 or 299×299, and most deep networks on mammographic image analysis are designed using the same size input [9], [18]. Resizing a large mammographic image to a small size will likely make the hardly detected lesions harder to detect, seriously hindering the performance of these models. Therefore, improving the architecture of DNNs according to the special characteristics of mammographic images is also an important task.

To address the aforementioned challenges, a different classification methodology based on DNNs is presented for automatically diagnosing breast cancer in mammographic images. In this study, the full problem of mammogram analysis is treated as a binary cancer status classification task to help radiologists determine the possibility that a given mammographic image contains malignant lesions. Therefore, the annotations required by this method are only the labels of the mammographic image's cancer status as normal/benign or malignant, which are more easily obtained from pathological diagnosis than other labels.

In this study, two pooling structures and an end-to-end architecture based on deep CNNs are proposed. This architecture is composed of three components: the feature extraction network, the special pooling method, and the classification network. The feature extraction network is designed to capture features from mammographic images instead of traditional manual features. A region-based group-max pooling structure and a global group-max pooling structure are explored to solve the challenge of the difficulty of classifying large mammographic images with small lesions. In this process, the mammographic image will be divided into regions according to the feature maps extracted by the feature extraction network and the regions with a high probability of containing malignant will be selected to obtain the final feature. Region-based group-max pooling(RGP) and global group-max pooling(GGP) are designed according to the characteristics of the mammographic images, which are more suitable for this task than other pooling methods. Finally, the aggregated features are fed into the classification network to calculate the diagnostic result. The major contributions of this work include:

1) A new mammographic image classification model is proposed, which is trained on raw mammographic images with only classification labels (without detection annotations). Fewer annotations make it easier to build and train a model.

2) A region-based group-max pooling structure is proposed, which divides the whole mammographic image into several regions, classifies each region, and finally selects some of the most suspicious foci as the final feature according to the classification results.

3) On the basis of the aforementioned pooling structure, the global group-max pooling structure is further explored. In this method, the step of classifying each region is abandoned, the feature selection is carried out in each channel separately, and the selected sections are finally combined as the final feature.

II. Related Work

Neural networks have been studied for many years [23]–[27], and recently they have achieved important breakthroughs in the field of computer vision [20], [28], [29]. Especially in the analysis of large annotated datasets, the performance of DNN models is far better than traditional machine learning models [30]–[34]. Compared to traditional machine learning models, DNNs can automatically extract deep features from the original input images, which are abstract but contain rich information.

Inspired by this work and Dense Convolutional Network [20], a deep CNN based model with 121 layers was introduced by Rajpurkar *et al.* [35] The model was trained on the ChestX-ray 14 dataset. Roza *et al.* [36] proposed a DNN structure with feature extraction networks for identifying two kinds of arrhythmias in ECG signals. These DNN approaches achieve state-of-the-art performance in medical image analysis.

In mammographic image analysis, most of the approaches based on DNNs can be executed by a two-stage procedure: segmentation of or extracting the Region of Interests (ROIs), and classification of the ROIs [37], [38]. Beura *et al.* [39] designed co-occurrence features and used the wavelet transform for breast cancer detection. Several other works have

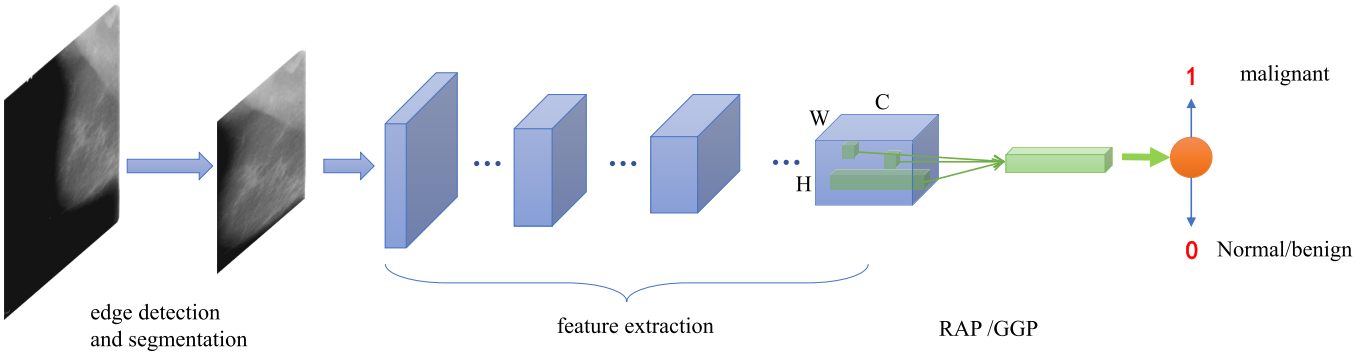

Fig. 1. Diagram of the overall architecture in mammographic image classification. Edge detection method is firstly used to segment the mammographic image, retain the breast area and resize the image to 800 × 800. Then the deep CNN accepts the resized breast image as input to extract the feature and capture the feature map. Next, the special pooling structures(RGP or GGP) designed for mammographic image are used to pooling the feature map. Finally, logistic regression is employed for the malignant probability of the image.

also used DNNs to perform mammographic images mass classification [9], [40]. However, those methodologies required annotated mass ROIs or segmentation ground truths to train the model.

Several recent studies have noted that multiple-stage learning cannot fully explore the potential of the DNN and there are not enough annotations, usually when the data set is very large. Chougrad *et al.* [41] obtained lesion patch from the original mammogram using the ROI annotations and built a lesion classification model. The study in [18] addressed the problem of full-image mammographic image classification by converting the problem into a multiple instance learning (MIL) situation. Pre-trained AlexNet [42] were used as their initial CNN architecture and the large mammogram images were re-scaled to the size of 224 × 224. But re-scaling the mammographic images brings a significant information loss in raw images.

Different from the above works, the proposed architecture does not use segmentation and ROI labels, nor does it transform the image classification problem into a MIL problem. According to the characteristics that the area containing lesions often comprise a tiny percentage of the area of the whole large mammographic image, two pooling methods are designed to obtain the final mammographic image feature representation, which can roughly locate the lesions and select features of the corresponding regions. Also, the proposed pooling method can be easily built and applied to any CNN models with any size of the input. In overall architecture, only class labels are required to help train the model, which can be easily obtained from pathological diagnosis. It is particularly useful in medical image analysis, where the detailed annotations often require time-consuming clinical expertise.

III. Method

A. Overall Architecture

The overall architecture of this method is shown in Fig. 1. It consists of several stages:

1) Edge detection and Otsus segmentation is used to locate the breast and remove the background for the reason that the black background takes up a large proportion of raw mammographic images, which are not necessary for classifying mammographic images. The specific implementation is using the functions provided by OpenCV library to detect the breast edge, and then generate multiple selection boxes according to the edge, finally select the largest box, which contains the breast.

2) A feature extraction network is built to capture deep features from the mammographic images.

3) Two pooling methods are employed to obtain vector representation from the extracted deep features. The details of these two pooling methods will be detailed in Sec III-B and Sec III-C, respectively.

4) Logistic regression is used to compute the probability of malignancy in the mammographic image.

Formulaically, the annotated dataset used in our proposed framework contains processed mammographic images and classification labels $D = \{(x_n, t_n); n = 1, 2, \cdots, N\}$. This dataset includes N samples of mammographic images and its corresponding classification labels. x_n is the n-th image in the data set D and t_n is its corresponding classification label. The proposed architecture aims at learning a robust model that extracts features from the mammographic images and computes the probability of containing malignant lesions.

CNN-based models usually include three kinds of operator: convolutional, pooling, and the fully-connected layer. We retain parts of the architecture containing all convolutional layers and pooling layers as an extractor to efficiently obtain features from the input mammographic image. In this study, the feature extractor is defined as f_e. Given the image x, $z = f_e(x|\phi)$ where $z \in \mathcal{R}^{W \times H \times C}$ denotes its feature map, W and H represent the row and column index of the feature map, respectively, and C is the channel dimension, ϕ is the parameter of this network which is initialized on ImageNET [43].

Then the feature map will be fed into the RGP or GGP structure to obtain a fixed-length vector representation. The special pooling method can be generally formulated as $v = f_p(z|\beta)$ where f_p is the mapping of the pooling model, v is the fixed-length vector, and β is the parameter. The details of the two pooling structures are introduced in the next two subsections. The linear regression layer at the end of the model is designed to compute the benign or malignant probability of

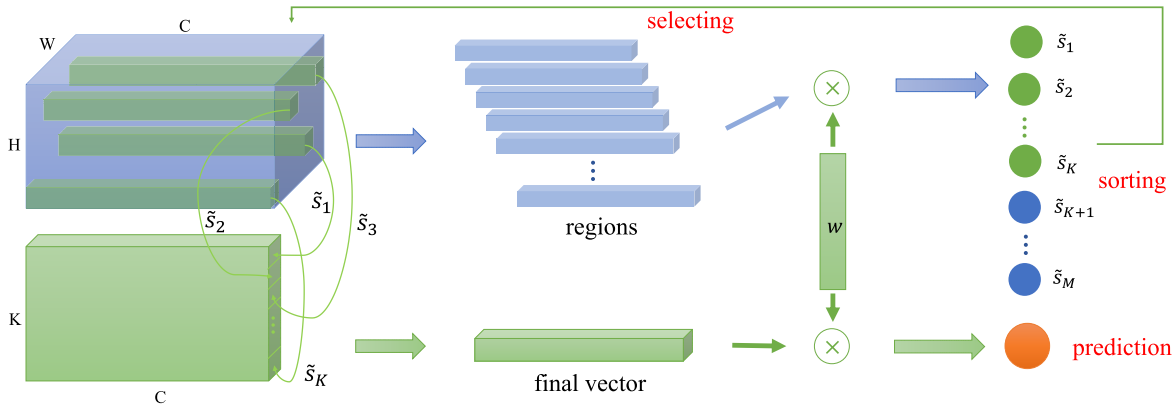

Fig. 2. Region group-max pooling. The last feature map will be divided into *M* regions according to its height and width. Each region corresponds to a *C*-dimensional vector, which will be fed into the logistic regression layer to calculate the malignant probability of the region. These probabilities will be sorted and the first k regions with high probability will selected to calculate a new feature. Finally, the feature will be classified by the same logistic regression.

the mammographic image taking a fixed-length vector v as input. In particular, we treat mammographic image classification as a binary class classification task to predict whether a mammographic image contains malignant lesions and use a single neuron node to calculate the probability, which can be computed as:

$$p = \sigma(w \cdot v + b), \tag{1}$$

where σ is the sigmoid active function, w is the weights of the logistic regression layer, b is the bias, and $\cdot$ is the inner product of the matrix w and vector v. The loss used in this model is defined as:

$$\mathcal{J} = -\frac{1}{N}\sum_{n=1}^{N}(w_n[t_n \log p_n + (1-t_n)\log(1-p_n)]) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2, \tag{2}$$

where $t_n \in \{0, 1\}$ is the true label (0 denotes the normal and benign label, 1 denotes the malignant label) of malignancy for mammographic image x_n, $\boldsymbol{\theta}$ is the parameter of the deep networks, λ is the regularizer that controls model complexity, and w_n is a manual re-scaling weight given to the loss.

Different from many existing detection models that require segmentation or ROI labels, the proposed model can be trained with only classification labels. Therefore, the key challenge in this architecture is achieving performance equivalent to that of other detection models using additional annotations. To improve performance, two pooling structures are designed to locate the lesions roughly, and obtain the final mammographic image feature. These methods make full use of the advantages of DNN and achieve comparable performance in classification results.

B. Region Based Group-Max Pooling

This section will introduce the details of the RGP structure designed for f_p. The benignity and malignancy of mammographic images are mainly determined by the attribute of the lesions, thus the main idea in this method is to locate and select these lesion areas from the image, then to classify them. Typically, in a mammographic image, the area containing lesion (mass or calcification) just comprise a tiny percentage of the total area, which means that most of the area in the mammographic image contributes little to identify the whole image as malignant or benign. It is hard to select lesion regions from the mammographic image without ROIs using traditional methods, but in this work, the problem is regarded as another task: dividing the mammographic image into regions and then computing scores for every region, with the regions with qualified scores selected as candidate regions. Characteristically, there are always few and small lesions contained in a sample. But even if only one of the lesions is malignant, the whole image must be classified as malignant. In other words, the area of interest in a mammographic image is the few regions suspected to contain malignant lesions. Therefore, the scoring system must be able to measure the probability of each region to malignancy. From this point of view, this problem can be converted into calculating the malignancy probability of each region in the mammographic image and then screening those with the highest probabilities.

In this method, the mammographic image is divided into regions according to the size of the feature map z extracted by the feature extraction network f_e. As the size of the feature map is $W \times H \times C$, the mammographic image is divided into $W \times H$ patches. The representation vector z_{ij} corresponds to the patch Q_{ij} in the mammographic image x_n, where $i \in \{1, 2, \cdots, W\}$ and $j \in \{1, 2, \cdots, H\}$ denote the specific row and column dimension in the feature map, respectively. If all of these regions are selected to compute the final malignant probability, the procedure of pooling and logistic regression can be formulated as follows:

$$\begin{cases} v = \frac{1}{M}\sum_{i=1}^{W}\sum_{j=1}^{H} z_{ij}, \\ p = \sigma\left(\frac{1}{M}\sum_{i=1}^{W}\sum_{j=1}^{H} w z_{ij} + b\right), \end{cases} \tag{3}$$

where $M = W \times H$. It can be clearly seen from the formula that each region of the feature map will contribute to the output probability, and the contribution is affected by the matrix w. Assuming that the model is well trained, for a mammographic image with malignant lesions, the average

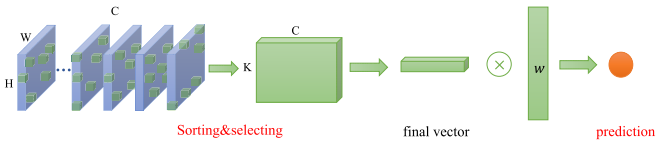

Fig. 3. Global group-max pooling. At the last feature map, all nodes in each channel will be sorted in descending order, and the first k values will be selected as the represent feature of this channel. Finally, all the features obtained from channels will be combined to compute the final feature vector, which will be fed into the logistic regression layer to predict the malignant probability of the mammogram.

malignant probability value of all regions must be higher than that of a normal mammographic image to make the final malignant probability value high. However, the main difference between these two kinds of mammographic images is whether there are lesions in an image or not. It means that for all mammographic images, the probabilities of malignancy in most regions are similar. The regions containing malignant lesions tend to have higher scores than other regions. In this way, the final probability is affected by both w and z_{ij}, and the probability of each region can be calculated according to them. Here the score s_{ij} of patch Q_{ij} will be computed using the representation z_{ij} at location (i, j) of the feature map by the classifier introduced in Eq(1):

$$s_{ij} = \sigma(W \cdot z_{ij} + b), \tag{4}$$

After malignant probabilities for all the regions in a mammographic image sample are calculated, features of the first few regions with higher malignant probabilities will be selected out to compute the final feature for the whole mammographic image. Here the set of scores $\{s_{11}, s_{12}, \cdots, s_{ij}, \cdots, s_{WH}\}$ for all regions in the mammographic image will be sorted in descending order as $\{\widetilde{s}_1 > \widetilde{s}_2 > \widetilde{s}_3, \cdots, \widetilde{s}_{M-1} > \widetilde{s}_M\}$. The first K higher scores $\{\widetilde{s}_1, \widetilde{s}_2, \cdots, \widetilde{s}_K\}$ and its corresponding regions will be selected. The features of these regions will be used to calculate the final feature, which can be formulated as:

$$v = \frac{1}{K}\sum_{k=1}^{K}\widetilde{z}_k, \tag{5}$$

where $\widetilde{z}_k$ is the feature corresponding to the selected region $\widetilde{s}_k$. The final feature v will be fed into the last layer of the network formulated in Eq(1). One advantage of RGP is that it scores all the regions in a mammographic image, so the sections with high probability can be selected out from the image. Abnormalities can be roughly located according to the scores computed by this pooling method. In particular, if the label of malignancy is set to be 0, K regions with lower probability should be selected to compute the malignant probabilities. Furthermore, due to the characteristic that the feature of one region extracted by convolutional layers always involves surrounding information, more regions should be selected.

C. Global Group-Max Pooling

In this section, another pooling structure named GGP is explored as an extension of the aforementioned introduced RGP. The main idea of GGP is to select the regions with a high probability of containing malignant lesions from a mammographic image. In RGP, it is required to score all the regions of the mammographic image, so the performance of the selection progress is limited by the performance of the scoring method. In GGP, the feature map will be fully utilized. Considering that the channel $z^c(c = 1, 2, \cdots, C)$ in the feature map z can be regarded as one view of the image, the values in this channel directly represent regions in this view, and the first few regions with high probabilities of being malignant can be selected from each channel. In this way, the representation that needs to be scored is itself, a single value z^c_{ij}, so the score can be easily represented using its identity value. To complete this procedure, K maximal values have been selected for each channel, and calculate C times to capture the final feature. For the c-th channel, the set of scores is $\{s^c_{11}, s^c_{12}, \cdots, s^c_{ij}, \cdots, s^c_{WH}\}$. Similar to RGP, all the scores will be sorted in descending order as $\{\widetilde{s}^c_1 > \widetilde{s}^c_2 > \widetilde{s}^c_3, \cdots, \widetilde{s}^c_{M-1} > \widetilde{s}^c_M\}$ and its corresponding first K regions $\{\widetilde{z}^c_1, \widetilde{z}^c_2, \cdots, \widetilde{z}^c_K\}$ with higher scores will be selected. The value of the neural node in the final feature at location c can be calculated as:

$$v_c = \frac{1}{K}\sum_{k=1}^{K}\widetilde{z}^c_k, \tag{6}$$

where v_c is the c-th component of the final feature v. It is the mean score of all the selected regions in channel c. Finally, v will be fed into the last layer of the network formulated in Eq(1). Different from RGP, the label of malignant utilize does not need to be considered. The scores of regions in every view will be learned adaptively no matter that the malignant label is 0 or 1. Same as RGP, due to the characteristics of convolution layers that regions in feature map always involve surrounding information, more regions have to be selected out.

IV. Experiment

A. Data Set

The proposed two methods in this work are validated on two public datasets, which are raw and scanned data of mammography, and are stored in DICOM format. The following is a detailed description of the two datasets.

CBIS-DDSM [44]: The CBIS-DDMS (Curated Breast Imaging Subset of DDSM), consists of decompressed images and precise annotations, formatted similarly to modern computer vision data sets. The CBIS-DDMS contains 753 calcification cases and 891 mass cases, with a total of 3071 images.

INbreast [45]: The INbreast database consists of full-field digital mammographic images and precise annotations. It has a total of 115 cases including multiple view images from each breast, with a total of 410 images. The INbreast database presents a wide variety of cases, including several types of lesions (masses, calcifications, asymmetries, and distortions). In this work, only the BI-RADS annotations(BI-RADS $\in \{4, 5, 6\}$ as malignant) are used.

The distribution of these two databases is shown in Table I. The INbreast dataset is divided into 80% training data and 20% test data. The CBIS-DDSM dataset is divided according to its original database (85% for training and 15% for testing).

TABLE I
DATA DISTRIBUTION IN CBIS-DDSM AND INBREST

Database	Benign/normal	Malignant	Total
CBIS-DDSM	1718(844mass/874cal)	1353(734mass/619cal)	3071
INbreast	310	100	410

In the preprocessing procedure, edge detection is firstly used to locate the breast, segment the mammographic image, and remove the background. Then the mammogram is resized to a standard size of 800 × 800. In this study, the original mammographic images were augmented. The resized images are divided by 225 to ensure the value is located in [0,1]. To reduce the influence of overfitting, for each training epoch, the mammographic images are randomly flipped horizontally, image contrast and saturation adjusted, rotated from -30 degrees to +30 degrees, and Gaussian noise is added. The image will be augmented before it is fed into the network, and there will be no variation of a same image in each epoch.

B. Training Details

In this experiment, the pre-trained DenseNet169 [20] is used as feature extraction network to capture features from raw mammographic images. Its structure is based on the code published by their authors, only the last classification layer is replaced by the proposed pooling structure. Considering that the CNN model has been adequately trained, to avoid over-fitting, in the experiment of both INbreast dataset and CBIS-DDSM dataset, we initialize the CNN with ImageNet and update it with a relatively small learning rate. The parameters initialized on ImageNet are from the public package provided by the authors of DenseNet. Correspondingly, the learning rate of updating the logistic regression layer is higher. The dimension of the representation vector in the penultimate hidden layer of the CNN is 1664.

Additionally, the model is trained with the Adam update rule and the initialized learning rate of 10^{-4} for logistic regression layer, 2×10^{-5} for CNN. The λ for loss function is 1×10^{-5}. The learning rate is decayed every 8 epochs with a decay rate of 0.98. The parameters are updated in minibatches with a batch size of 32. Training stops at the 150th epoch. Pytorch [46] is used to construct and train the proposed model, and the GPU used in this experiment is Tesla P100. The proposed method is evaluated in terms of classification accuracy, the area under the receiver operating characteristic (ROC) curve (area under curve=AUC). The Wilcoxon rank-sum test, which is based on the rank of each measurement in each sample, is used for statistical analysis. The accuracy of the models is statistically analyzed and the statistical significance level is set to $\alpha = 0.05$. The null hypothesis is that there existed significant differences between the sample sets if $p < 0.05$, and no significant differences if $p > 0.05$. Many segmentation models classify the mammogram after segmentation. The comparison models in this experiment all include classification tasks, and the classification results are used to compare with the proposed model.

TABLE II
ACCURACY AND AUC COMPARISONS OF DIFFERENT K IN RGP STRUCTURE ON INBREST DATABASE

k value	Acc.	AUC
0.1	0.864 ± 0.0002	0.873 ± 0.0001
0.2	0.872 ± 0.0001	0.874 ± 0.0003
0.3	0.890 ± 0.0002	0.911 ± 0.0002
0.4	0.893 ± 0.0003	0.904 ± 0.0003
0.5	$\mathbf{0.923 \pm 0.0003}$	$\mathbf{0.934 \pm 0.0002}$
0.6	0.903 ± 0.0002	0.902 ± 0.0001
0.7	0.911 ± 0.0004	0.903 ± 0.0002

TABLE III
ACCURACY AND AUC COMPARISONS OF DIFFERENT K IN RGP STRUCTURE ON CBIS DATABASE

k value	Acc.	AUC
0.1	0.739 ± 0.0002	0.798 ± 0.0001
0.2	0.748 ± 0.0001	0.813 ± 0.0003
0.3	0.750 ± 0.0002	0.811 ± 0.0002
0.4	0.748 ± 0.0003	0.822 ± 0.0003
0.5	0.745 ± 0.0003	0.832 ± 0.0002
0.6	0.758 ± 0.0002	0.831 ± 0.0001
0.7	$\mathbf{0.762 \pm 0.0002}$	$\mathbf{0.838 \pm 0.0001}$
0.8	0.736 ± 0.0003	0.805 ± 0.0002

C. Results and Analysis

In this work, a new variable, K, the number of mammographic image sections to be selected, is introduced. The different values of K will directly affect the performance of the network on the data set. In order to select a suitable value of K as the basis for screening features, a number of values of K have been selected for experiments. Because the size of the last feature map in different CNNs is different, the mammographic image will be divided into different numbers of regions in different CNNs. Thus a new variable k is defined, which indicates the proportion of selected regions to the total regions, where $K = k \cdot W \cdot H$. In the experiment, the choice of k is 0.1 at each interval from 0.1 to 0.9. The model corresponding to each k value will be trained and tested on the dataset three times. Finally, the mean value of the three results will be taken as the final result. Table II is the test result of RGP structure on the INbreast dataset.

In the above experiments, it can be seen that the choice of different values of k values will have an impact on the experimental results; too small and too large values for k are not conducive to network performance. The reason is that when the k value is too small, the network only pays attention to regions that tend to contain obvious malignancy, so the model tends to classify the mammographic image as malignant as the calculated probability is large. In addition, the selected

TABLE IV
ACCURACY COMPARISONS OF THE PROPOSED FRAMEWORK AND RELATED MODELS ON INBREST DATABASE

Methodology	Set-up	Acc.	AUC	p-value
Moreira et al. [45]	Manual+feat.	0.89	N/A	N/A
Pretrained CNN+Random Forest [14]	Auto.+feat.	0.91 ± 0.0002	0.76 ± 0.0023	N/A
deep MIL [18]	Auto.	0.900 ± 0.0002	0.89 ± 0.0004	N/A
max pooling	Auto.	0.837 ± 0.0004	0.821 ± 0.0003	0.008
average Pooling	Auto.	0.862 ± 0.0003	0.857 ± 0.0003	0.009
RGP	Auto.	0.919 ± 0.0003	$\mathbf{0.934 \pm 0.0003}$	0.013
GGP	Auto.	$\mathbf{0.922 \pm 0.0002}$	0.924 ± 0.0003	0.011

TABLE V
ACCURACY COMPARISONS OF THE PROPOSED PROPOSED FRAMEWORK AND RELATED MODELS ON CBIS-DDSM DATABASE

Methodology	Acc.	AUC	p-value
deep MIL [18]	0.742 ± 0.0003	0.791 ± 0.0002	N/A
max pooling	0.675 ± 0.0001	0.741 ± 0.0003	0.006
average pooling	0.703 ± 0.0002	0.764 ± 0.0002	0.009
RGP	0.762 ± 0.0002	$\mathbf{0.838 \pm 0.0001}$	0.012
GGP	$\mathbf{0.767 \pm 0.0002}$	0.823 ± 0.0002	0.009

regions may not fully cover the lesion, thus the final diagnosis will probably be wrong. However, when k is too large, it will be approximately equivalent to the common average pooling, which will lead to the problem described at the beginning of this paper: there are many invalid regions that will interfere with the performance of the model. The final experimental results show that the model performs best when $k = 0.5$. With this value, the model not only removes part of the cluttered regional features, but also ensures that the lesions or focus is within the scope of the selected regions. When k is less than or greater than 0.5, the experimental results show a downward trend.

After completing the experiment on the value of k, we choose the value that makes the model perform best as the final value of k to build the model and compare it with other models.

First, we compare our methods with previous models validated on the INbreast dataset in Table IV. Previous methods based on hand-crafted features required a large number of human-labeled detection bounding boxes or segmentation ROI, even in the process of testing, which costs a lot. These traditional machine learning methods either need manual labeling or multi-stage training. Recently, some new studies applied CNNs with good performance in other tasks to the classification of mammographic image, without considering the characteristics of mammography, the ability of neural network can not be fully utilized. The proposed method is fully automated, which can be trained without manual labeling or segmentation ROIs. Even classification labels can be obtained directly from pathological results. It fully realizes end-to-end training. In addition, we have also carried out experiments using average pooling and max pooling structures, which can be regarded as the special cases of RGP and GGP(when $k = 1$ in RGP, it is average pooling; when $K = 1$ in GGP, it is max pooling), respectively.

As shown in Tables IV and V, applying the group-max pooling to deep CNN leads to better performance than common pretrained CNNs, even several CNNs using detection/segmentation annotation during training. Pre-trained AlexNet and pre-trained R-CNN are used in random forest-based pretrained CNNs and deep MIL models, respectively. The parameters of all the pre-trained models are initialized on ImageNET. The deep MIL model was tested on CBIS dataset, using the learning algorithm and parameters they provided. The result shows the superiority of the end-to-end trained deep network for whole mammographic image classification, the features obtained by the pooling methods are helpful for training deep networks with mammogram as input. According to the accuracy metric, the GGP is better than the RGP with pre-trained DenseNet which is better than the max pooling-based or average pooling structure. This result is consistent with previous discussions that the sparsity lesions assumption benefited from the learning ability of the deep network itself, which selected most malignant regions and is more efficient than max pooling or average pooling. Our GGP structure achieves competitive accuracy to random forest-based pretrained CNNs and deep MIL models. Also, the AUC of the CNN with these pooling structures is higher than previous work, which shows our method is more robust. This shows the effectiveness and transferability of the features obtained by RGP or GGP to medical images. Our deep networks achieved the best AUC result, which proves the superior performance of the RGP and GGP structure. The p-value indicates the difference between the proposed methods and the MIL model. A significant difference existed in the two models if the p-value is less than 0.05. The Wilcoxon rank-sum test are shown in Tables IV and V. In terms of p-values, there are significant differences between the proposed models and the MIL model, the basic CNN models and the MIL model. A phenomenon can be found in the tables that the proposed RGP always achieves the better AUC, while GGP achieves the better Acc.. The reason is that the RGP structure divides the mammogram into several regions in space and the region features extracted by it accord with the characteristics of lesion

TABLE VI
PERFORMANCE OF THE PROPOSED MODEL ON DIFFERENT LESIONS IN CBIS-DDSM DATABASE

type	AUC	Acc.	FPR	FNR
mass	0.87	0.796	0.084	0.213
calcification	0.805	0.701	0.089	0.115

TABLE VII
COMPUTATION TIME OF FOUR POOLING STRUCTURES

method	ms/image	image/s
average pooling	2.95	339
max pooling	2.99	334
RGP	3.12	321
GGP	3.01	332

distribution in the mammogram. In the process of training, if the selected region does not contain lesion, the network will be updated using the gradient to reduce the score of this region otherwise the score of the region containing lesion will increase. This process enables the network to select the lesion region stably and has better robustness, so AUC is on the high side. For GGP structure, because such a process will be carried out in every channel of the feature map, the regional information in the feature map is abstract, so the selection of lesion region depends on the learning ability of the network itself. This gives full play to the potential of the deep neural network, but lacks the regional constraints, making the accuracy of the network higher, but the robustness is poor. This result also shows the importance of regional information. The accuracy in CBIS-DDSM is much lower than in INBREAST, which may be caused by different annotations. INbreast database only contains BI-RADS annotations, while CBIS-DDSM publishes benign/malignant annotations. In the field of deep learning, annotations are very important and have a great impact on the results. On the other hand, the inconsistency of data source devices in two datasets is also a major factor affecting the results.

Table VI shows the performance of the proposed model on different lesions. We selected the RGP model with k = 0.7. False positive rate(FPR) means the model predict benign as malignant, false negative rate(FNR) means the model predict malignant as benign. It can be seen that the performance of the model on the mass mammogram is much higher than that on the calcification mammogram, and the AUC in the mass data reaches 0.87. This is because the mass in the image is more obvious than calcification, and compared with calcification, the volume of the mass is generally larger. In addition, most prediction error of the model is to predict malignant samples as benign.

To evaluate the increase of the computational workload during the selection procedure as K times additional computation in the pooling structure, we further applied four structures of average pooling, max pooling, RGP and GGP in the model to compare the calculation time. The hardware used in the experiment is Tesla P100. The experimental results are shown in Table VII. Compared with average pooling, the proposed two structures bring 0.17ms and 0.06ms extra computing time respectively. Compared with max pooling, the proposed two structures bring about an increase in the calculation time of 0.13ms and 0.02ms respectively.

The major ingredients for the good performance of the proposed two structures are as follows: First, the end-to-end architecture makes full use of the DNNs' advantages; the deep features extracted in this architecture are much better than hand-crafted features. Second, the two pooling structures roughly locate the malignant regions, reducing the interference of nonrelevant information. Third, our models fully explore all the regions to select the most malignant regions, but they use the last logical layer as the scoring network, which has little effect on computing time. This is a distinct advantage over previous networks requiring additional annotations or employing several stages consisting of detection and segmentation networks.

To further demonstrate the ability of the proposed structure in roughly locating suspicious regions, as shown in Fig. 4, the scores of all regions in four mammographic image samples from the test set, which represent the tendency of each region to be malignant, were calculated and visualized. It can be clearly seen that the CNN model with the proposed structure learned not only the prediction of the whole mammographic image, but also the prediction of all regions within the whole mammographic image. Fig. 4(a)-(c) show that the proposed model is able to roughly locate the existing lesions in the whole mammographic image without any segmentation ground truth annotation or detecting ROIs. This is of great significance for assisting the analysis of huge quantity of data and satisfying the requirement for high-quality annotation. Among the three images, Fig. 4(a) locates the mass regions well, and has a high degree of coincidence with the ROI ground truth. The scores of the regions that contain lesions are higher than those of other regions. In Fig. 4(b) and (c), compared with the annotated ROI image, the model located not only the regions containing lesions, but also some other areas. These areas are often the nipples, mainly due to the large difference between the nipples and other areas, and the fact that the nipples are not obvious in many mammographic images, such as Fig. 4(a), resulting in some interference. The other reason is that some areas are obviously different from the surrounding areas, such as the high score area in Fig. 4(c), but these areas are usually smaller. However, these two problems can be alleviated by increasing the value of k. When k is large, more regions will be selected as candidate areas for final diagnosis, to a certain extent, avoiding the loss of areas containing lesions. In addition, there are also some mammographic images of healthy breasts, which do not contain lesions, as shown in Fig. 4(d), and the prediction scores of all regions in these images are low, which is also in line with common sense. To quantify the localization ability of the proposed model, the visualization results of all mammogram are counted. In the test set, 78.2% of the mammogram, the regions with high scores covered all lesions, and a considerable part of them also covered the nipple. It shows that the proposed method has the ability of learning location information and can roughly locate lesions in mammographic images. In general, the proposed

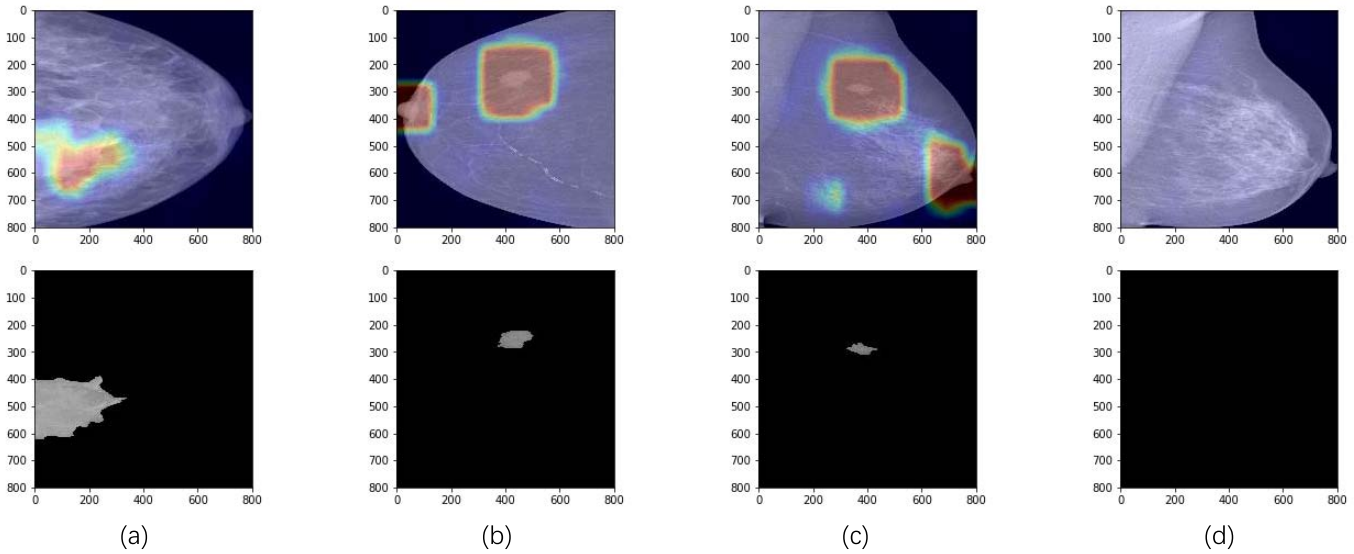

Fig. 4. Four samples visualizing the malignant probabilities of all regions in mammogram. The first row is the resized mammogram with predicted malignant probabilities from logistic regression layer for four images respectively, in which the matrix consisting of all probability values is resized to the same size as mammogram by interpolation and superimposed on the original image. The images in second line are ROI images of lesions from the annotations in the INbreast. The black-and-white images in the second row are the annotations in the INbreast, which were resized to the size of mammogram for comparison.

model can roughly locate suspicious regions to assist in the final prediction of the whole image.

V. Conclusion

In this study, an end-to-end mammographic image diagnostic architecture was proposed, in which RGP structure and GGP structure were proposed for screening features from the last feature map layer. The two pooling structures are designed for the mammographic characteristics of large images with small lesions. The mammographic image in this model will be divided into regions according to the size of the last feature map extracted by the extraction network, and these regions will be screened to detect the regions containing lesions for final diagnosis. The reason for this is to locate the lesions roughly. By discarding some features and choosing only those features of regions with lesions as the final diagnostic basis, a large number of nonrelevant regions can be excluded from interfering with the final diagnostic results. Compared with the typical max pooling and global average pooling, the proposed pooling structure is more suitable for analysis and diagnosis. In this process, the ability of DNN to extract hierarchical features and mine data is fully utilized, which is of great significance to improve the performance of the model. Finally, the results on two public datasets show that the performance of the proposed model is better than that of other models. The GGP structure achieves competitive accuracy to random forest-based pretrained CNNs and deep MIL models. Also, the AUC of the RGP is higher than previous work, which shows the proposed method is more robust. The results of visualization show that the proposed model can roughly locate suspicious regions. The capability of locating lesions is acquired through weak supervisory learning, which does not need corresponding labels at all. When the amount of training data is large enough, its ability to locate lesions will be improved accordingly. This weak supervised learning network can be applied to automatic annotating to reduce the cost of annotations and improve the performance of the model. But the k value needs to be set manually to get a better model on different datasets. This brings a new parameter to the training of the model. We will explore the rules and improve it in future work.

References

[1] M. J. Milroy, "Cancer statistics: Global and national," in *Quality Cancer Care*. Manhattan, NY, USA: Springer, 2018, pp. 29–35.
[2] (2018). *Breast Cancer: Prevention Control*. [Online]. Available: http://www.who.int/cancer/detection/breastcancer/en/
[3] S. Moss *et al.*, "The impact of mammographic screening on breast cancer mortality in Europe: A review of trend studies," *J. Med. Screening*, vol. 19, no. 1, pp. 26–32, Sep. 2012.
[4] M. S. Bae *et al.*, "Breast cancer detected with screening US: Reasons for nondetection at mammography," *Radiology*, vol. 270, no. 2, pp. 369–377, Feb. 2014.
[5] S. Yu and L. Guan, "A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram films," *IEEE Trans. Med. Imag.*, vol. 19, no. 2, pp. 115–126, Feb. 2000.
[6] I. Christoyianni, A. Koutras, E. Dermatas, and G. Kokkinakis, "Computer aided diagnosis of breast cancer in digitized mammograms," *Comput. Med. Imag. Graph.*, vol. 26, no. 5, pp. 309–319, Sep. 2002.
[7] R. Hupse and N. Karssemeijer, "Use of normal tissue context in computer-aided detection of masses in mammograms," *IEEE Trans. Med. Imag.*, vol. 28, no. 12, pp. 2033–2041, Aug. 2009.
[8] C. D. Lehman, R. D. Wellman, D. S. M. Buist, K. Kerlikowske, A. N. A. Tosteson, and D. L. Miglioretti, "Diagnostic accuracy of digital screening mammography with and without computer-aided detection," *JAMA Internal Med.*, vol. 175, no. 11, p. 1828, Nov. 2015.
[9] G. Carneiro, J. Nascimento, and A. P. Bradley, "Unregistered multiview mammogram analysis with pre-trained deep learning models," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 652–660.
[10] J. J. Fenton *et al.*, "Effectiveness of computer-aided detection in community mammography practice," *J. Nat. Cancer Inst.*, vol. 103, no. 15, pp. 1152–1161, Aug. 2011.
[11] D. Ribli, A. Horvath, Z. Unger, P. Pollner, and I. Csabai, "Detecting and classifying lesions in mammograms with deep learning," *Sci. Rep.*, vol. 8, no. 1, 2018, Art. no. 4165.

[12] A. Oliver *et al.*, "A review of automatic mass detection and segmentation in mammographic images," *Med. Image Anal.*, vol. 14, no. 2, pp. 87–110, Apr. 2010.

[13] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. Guevara Lopez, "Representation learning for mammography mass lesion classification with convolutional neural networks," *Comput. Methods Programs Biomed.*, vol. 127, pp. 248–257, Apr. 2016.

[14] N. Dhungel, G. Carneiro, and A. P. Bradley, "Automated mass detection in mammograms using cascaded deep learning and random forests," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2015, pp. 1–8.

[15] W. Zhu, X. Xiang, T. D. Tran, and X. Xie, "Adversarial deep structural networks for mammographic mass segmentation," 2017, *arXiv:1612.05970*. [Online]. Available: https://arxiv.org/abs/1612.05970

[16] J. Tang, R. M. Rangayyan, J. Xu, I. E. Naqa, and Y. Yang, "Computer-aided detection and diagnosis of breast cancer with mammography: Recent advances," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 2, pp. 236–251, Mar. 2009.

[17] N. Dhungel, G. Carneiro, and A. P. Bradley, "Fully automated classification of mammograms using deep residual neural networks," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 310–314.

[18] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2017, pp. 603–611.

[19] W. Lotter, G. Sorensen, and D. Cox, "A multi-scale CNN and curriculum learning strategy for mammogram classification," in *Proc. Int. Workshop Multimodal Learn. Clin. Decis. Support*, 2017, pp. 169–177.

[20] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[22] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size," 2016, *arXiv:1602.07360*. [Online]. Available: https://arxiv.org/abs/1602.07360

[23] Y. Feng, L. Zhang, and J. Mo, "Deep manifold preserving autoencoder for classifying breast cancer histopathological images," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published.

[24] L. Zhang, Z. Yi, and J. Yu, "Multiperiodicity and attractivity of delayed recurrent neural networks with unsaturating piecewise linear transfer functions," *IEEE Trans. Neural Netw.*, vol. 19, no. 1, pp. 158–167, Jan. 2008.

[25] L. Zhang and Z. Yi, "Selectable and unselectable sets of neurons in recurrent neural networks with saturated piecewise linear transfer function," *IEEE Trans. Neural Netw.*, vol. 22, no. 7, pp. 1021–1031, Jul. 2011.

[26] L. Zhang, Z. Yi, and S.-I. Amari, "Theoretical study of oscillator neurons in recurrent neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5242–5248, Nov. 2018.

[27] L. Zhang, Z. Yi, S. Zhang, and P. Ann Heng, "Activity invariant sets and exponentially stable attractors of linear threshold discrete-time recurrent neural networks," *IEEE Trans. Autom. Control*, vol. 54, no. 6, pp. 1341–1347, Jun. 2009.

[28] L. Wang, L. Zhang, and Z. Yi, "Trajectory predictor by using recurrent neural networks in visual tracking," *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3172–3183, Oct. 2017.

[29] Y. Wang, L. Zhang, L. Wang, and Z. Wang, "Multitask learning for object localization with deep reinforcement learning," *IEEE Trans. Cogn. Develop. Syst.*, vol. 11, no. 4, pp. 573–580, Dec. 2019.

[30] C. M. Bishop, *Pattern Recognition And Machine Learning* (Information Science and Statistics), no. 4. New York, NY, USA: Springer, 2006, p. 049901.

[31] J. Hu, Y. Chen, J. Zhong, R. Ju, and Z. Yi, "Automated analysis for retinopathy of prematurity by deep neural networks," *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 269–279, Jan. 2019.

[32] J. Mo, L. Zhang, and Y. Feng, "Exudate-based diabetic macular edema recognition in retinal images using cascaded deep residual networks," *Neurocomputing*, vol. 290, pp. 161–171, May 2018.

[33] X. Qi *et al.*, "Automated diagnosis of breast ultrasonography images using deep neural networks," *Med. Image Anal.*, vol. 52, pp. 185–198, Feb. 2019.

[34] Y. Luo, J. Zou, C. Yao, X. Zhao, T. Li, and G. Bai, "HSI-CNN: A novel convolution neural network for hyperspectral image," in *Proc. Int. Conf. Audio, Lang. Image Process. (ICALIP)*, Jul. 2018, pp. 464–469.

[35] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning," 2017, *arXiv:1711.05225*. [Online]. Available: https://arxiv.org/abs/1711.05225

[36] V. C. C. Roza, A. M. De Almeida, and O. A. Postolache, "Design of an artificial neural network and feature extraction to identify arrhythmias from ECG," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, May 2017, pp. 391–396.

[37] H. Cheng, X. Shi, R. Min, L. Hu, X. Cai, and H. Du, "Approaches for automated detection and classification of masses in mammograms," *Pattern Recognit.*, vol. 39, no. 4, pp. 646–668, Apr. 2006.

[38] M. P. Sampat *et al.*, "Computer-aided detection and diagnosis in mammography," in *Handbook of Image and Video Processing*, vol. 2, no. 1. New York, NY, USA: Academic, 2005, pp. 1195–1217.

[39] S. Beura, B. Majhi, and R. Dash, "Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer," *Neurocomputing*, vol. 154, pp. 1–14, Apr. 2015.

[40] Z. Jiao, X. Gao, Y. Wang, and J. Li, "A deep feature based framework for breast masses classification," *Neurocomputing*, vol. 197, pp. 221–231, Jul. 2016.

[41] H. Chougrad, H. Zouaki, and O. Alheyane, "Multi-label transfer learning for the early diagnosis of breast cancer," *Neurocomputing*, to be published.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[44] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Sci. Data*, vol. 4, Dec. 2017, Art. no. 170177.

[45] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast: Toward a full-field digital mammographic database," *Acad. Radiol.*, vol. 19, no. 2, pp. 236–248, 2012.

[46] N. Ketkar, "Introduction to pytorch," in *Deep Learning With Python*. Berkeley, CA, USA: Apress, 2017, pp. 195–208.