Convolutional Feature Descriptor Selection for Mammogram Classification

Dong Li, Lei Zhang, *Senior Member, IEEE*, Jianwei Zhang, and Xingyu Xie

***Abstract*—Breast cancer was the most commonly diagnosed cancer among women worldwide in 2020. Recently, several deep learning-based classification approaches have been proposed to screen breast cancer in mammograms. However, most of these approaches require additional detection or segmentation annotations. Meanwhile, some other image-level label-based methods often pay insufficient attention to lesion areas, which are critical for diagnosis. This study designs a novel deep-learning method for automatically diagnosing breast cancer in mammography, which focuses on the local lesion areas and only utilizes image-level classification labels. In this study, we propose to select discriminative feature descriptors from feature maps instead of identifying lesion areas using precise annotations. And we design a novel adaptive convolutional feature descriptor selection (AFDS) structure based on the distribution of the deep activation map. Specifically, we adopt the triangle threshold strategy to calculate a specific threshold for guiding the activation map to determine which feature descriptors (local areas) are discriminative. Ablation experiments and visualization analysis indicate that the AFDS structure makes the model easier to learn the difference between malignant and benign/normal lesions. Furthermore, since the AFDS structure can be regarded as a highly efficient pooling structure, it can be easily plugged into most existing convolutional neural networks with negligible effort and time consumption. Experimental results on two publicly available INbreast and CBIS-DDSM datasets indicate that the proposed method performs satisfactorily compared with state-of-the-art methods.**

***Index Terms*—Discriminative representation, feature descriptor, activation map, mammogram classification.**

I. Introduction

BREAST cancer was the most commonly diagnosed cancer (24.5%) and the leading cause of cancer-related mortality (15.5%) among women worldwide in 2020 [1]. Incidence rates of breast cancer have increased rapidly in South America, Africa [2], and Asia [3] in recent years. Early-stage screening with mammography can significantly reduce breast cancer mortality in the long term [4]. In standard screening mammography, the craniocaudal (CC) view and the mediolateral oblique (MLO) view are acquired for each breast [5]. Radiologists assign a standardized assessment to each screening mammogram according to BI-RADS (Breast Imaging Reporting and Data System) [6]. However, accurately reading a screening mammogram is a challenging task that relies on the experience of radiologists [7]. In some complicated cases, experienced radiologists may come to different conclusions due to the presence of confounding and inconspicuous lesions.

A computer-aided diagnosis (CAD) system, which recognizes patterns in mammograms associated with breast cancer, may assist radiologists with screening mammography [8], [9]. In the last decade, various methods have been proposed to diagnose breast cancer based on mammography automatically. Traditional machine learning methods that rely on hand-crafted features from lesions must be designed according to specific data meticulously [10]. Furthermore, these methods may have poor portability because hand-crafted features are not data-driven. Other works have employed deep neural networks (DNNs) that can automatically learn features from raw data to detect lesions in mammograms. However, most of these methods require precise annotations, such as bounding boxes or segmentation ground truths. Establishing such annotations requires expert domain knowledge and considerable effort [11]. Recently, several approaches have been proposed to diagnose whole mammograms using image-level classification labels. In these approaches, the task of mammogram-based diagnosis is treated as a binary classification task, which helps radiologists determine the probability of malignancy for a given mammogram [12], [13], [14], [15]. However, these approaches usually do not pay enough attention to discriminative areas (local lesion areas).

The areas that contain lesions are discriminative and play a crucial role in the diagnosis result. Usually, only a few small lesions are contained in a mammogram. The lesions account for only 2% to 4% of the total areas on a mammogram [16]. But even if only one very small lesion is malignant, the whole mammogram should be classified as malignant [17]. The background or noise, which contributes little to diagnosis, accounts for a large portion of a mammogram. Hence, identifying discriminative lesion areas has potential advantages for mammogram-based diagnosis.

Recent studies have shown the potential of using deep convolutional feature descriptors to represent local targets [18], [19]. Most of those methods select discriminative descriptors according to the activation map. However, these methods simply use the maximum or mean value with a ratio as the threshold [19],

Manuscript received 2 March 2022; revised 27 October 2022 and 30 November 2022; accepted 28 December 2022. Date of publication 4 January 2023; date of current version 7 March 2023. This work was supported by the National Science Fund for Distinguished Young Scholars under Grant 62025601. *(Corresponding author: Lei Zhang.)*

The authors are with the College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: lidong@stu.scu.edu.cn; leizhang@scu.edu.cn; zhangjianwei@stu.scu.edu.cn; xiexingyu@stu.scu.edu.cn).

Digital Object Identifier 10.1109/JBHI.2022.3233535

[20], which could not change adaptively for different data distributions. Furthermore, these methods introduce additional hyperparameters. In this study, we design a deep-learning method for diagnosing breast cancer in mammography according to image-level classification labels. For this method, a novel adaptive convolutional feature descriptor selection (AFDS) structure is designed based on the histogram of the activation map.

Specifically, after feature maps are generated by convolutional neural networks (CNNs), the proposed method performs channel-wise average pooling on the feature maps to obtain the activation map. The values in the activation map are related to the probability of malignancy for each corresponding area in mammograms. So it is feasible to select suspicious lesion areas according to the activation map. In this study, we observe that the activation map of a mammogram usually exhibits a highly skewed distribution with an extremely high peak. So we adopt the triangle threshold strategy [21], which could deal with this condition better, to calculate a specific threshold for guiding the activation map to determine which feature descriptors are discriminative. As the triangle threshold can be calculated adaptively, the proposed AFDS structure does not require any additional hyperparameters or training parameters. In addition, the AFDS structure can be regarded as a pooling structure. It could be easily plugged into existing CNN models to replace average-pooling or max-pooling with negligible consumption.

To evaluate the performance of the proposed method, we conducted several experiments on two commonly used publicly available datasets — the INbreast dataset and the CBIS-DDSM dataset. The results demonstrate that the proposed AFDS method could achieve state-of-the-art classification performance on both datasets. The main contributions of this study can be summarized as follows:

1) This study designs a deep learning method for diagnosing breast cancer in mammography, which focuses on the discriminative lesion areas and only requires image-level classification labels.
2) This study proposes a novel AFDS structure based on the histogram of the activation map, which calculates the triangle threshold for guiding the activation map to determine which feature descriptors (local areas) are discriminative.

II. Related Work

In this section, we briefly review the related work in two areas: deep learning approaches for mammogram-based breast cancer diagnosis and methods for convolutional feature descriptor selection.

A. Deep Learning Approach for Mammogram-Based Diagnosis

CNNs have achieved breakthrough performance in object detection, image segmentation, and image classification [22], [23], [24]. Recently, some approaches have been proposed to apply CNNs to diagnose breast cancer from mammograms [25], [26].

Most of these approaches treat the automatic diagnosis as an object detection or segmentation task [27], [28]. Hagos et al. [27] proposed a patch-based model that learned symmetrical differences to detect masses on breasts. Dhungel et al. [28] proposed a multi-stage detection network to perform mammogram classification. However, the detection or segmentation annotations required in these methods are expensive and time-consuming. In order to eliminate this problem, several approaches have also been proposed to perform mammogram-based breast cancer diagnosis using image-level classification labels [12], [13], [14], [15]. Shu et al. [12] proposed two different pooling structures and demonstrated they were more suitable for analyzing mammograms than the commonly used max-pooling and average-pooling structures. However, the number of regions in their method needs to be fixed and pre-defined, which limits the model's effectiveness and flexibility. Zhu et al. [13] proposed a multi-instance network for whole mammogram classification, which divided mammograms into regions based on the feature map layer to transform the task into a multi-instance learning problem. However, this method requires a sparsity factor μ, which is difficult to evaluate. Shen et al. [14] proposed a globally-aware multiple instance classifier, which first applied a low-capacity global model on the whole image to roughly locate possible lesions, and then utilized a high-capacity module to extract visual details. However, the coarse localization precision of possible lesions provided by the low-capacity model could not be guaranteed. Xie et al. [15] proposed a multi-scale structure for mammogram classification, enabling the screening of unique features in lesions of various sizes. However, their method did not distinctly treat the lesion areas and the background. Generally, most previous approaches neglect the fact that small local lesion areas usually have an essential role in mammogram-based breast cancer diagnosis. In this study, we pay more attention to those discriminative feature descriptors, which are strongly associated with small lesion areas.

B. Convolutional Feature Descriptor Selection

Deep convolutional descriptors have been used in many studies in weakly supervised object localization (WSOL), fine-grained image recognition, etc. [18], [19], [20]. In most of these studies, discriminative descriptors are selected according to the activation map. Zhou et al. [29] proposed a class activation maps (CAM) approach to obtain the activation map using only image-level annotations. Selvaraju et al. [30] extended such object localization abilities by computing gradients of the logits with respect to intermediate feature maps in Grad-CAM. However, both CAM and Grad-CAM methods utilize post-processing since they require weights associated with the target-class logit [31]. As a result, they are not flexible in cases when the target image label is unknown. Some other methods perform channel-wise average pooling on the input feature maps to obtain the activation map and then select descriptors based on the activation map, which does not require any trainable parameters. Wei et al. [18] proposed the Mask-CNN model, which consisted of a convolutional architecture to locate the discriminative parts and generated object/part masks. Tian et al. [32] proposed an iterative

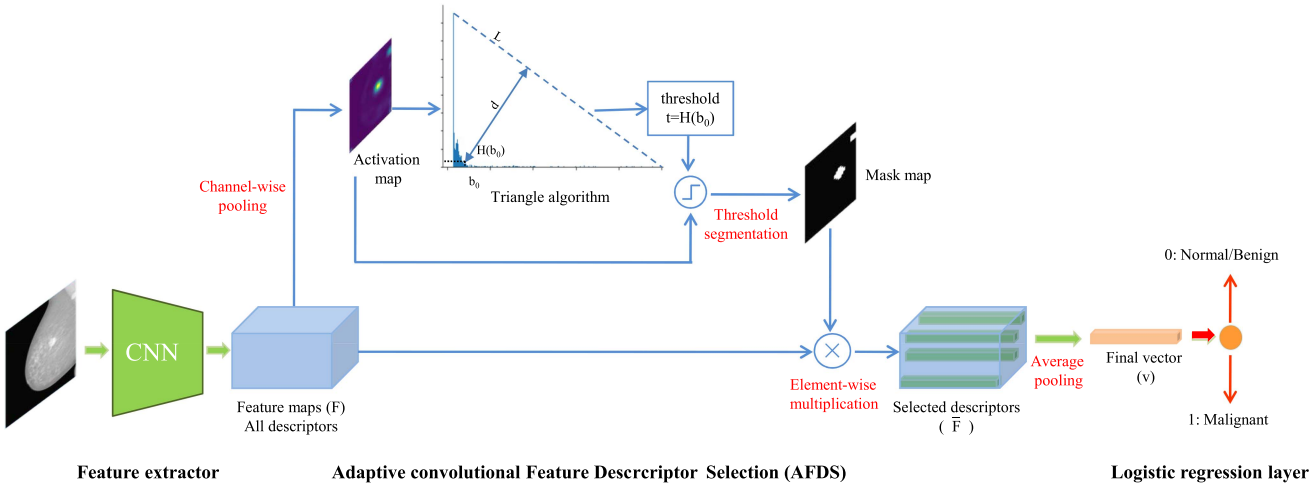

Fig. 1. Overall architecture of the proposed method, which is composed of a feature extractor, an AFDS structure, and a logistic regression layer. F represents the extracted deep feature maps, which include all feature descriptors. $\bar{F}$ stands for the discriminative representation, which includes only the discriminative feature descriptors. v is a fixed-length vector, which is obtained by applying average pooling on $\bar{F}$.

discrimination CNN, in which region features were extracted from the original image with higher activation responses. Choe et al. [20] proposed an attention-based dropout layer (ADL) for WSOL, which erased the most discriminative regions to prevent the model from only relying on the most discriminative parts. They generated the activation map by compressing the feature maps using channel-wise average pooling. Then they set a threshold by prefixing a ratio γ of the maximum value to find discriminative areas. Wei et al. [19] proposed an approach for the fine-grained image retrieval task. They calculated the mean value of the activation map as a threshold to decide which positions localized main objects. However, these methods simply use the maximum or mean value with a ratio as the threshold, which could not change adaptively for different data distributions. Besides, these methods could introduce additional hyper-parameters. In our study, we propose a novel structure to automatically extract the discriminative representation for mammograms based on the distribution of the activation map without any hyper-parameters.

III. Methods

A. Overall Architecture

This study designs a deep learning method to automatically diagnose breast cancer in mammography using image-level classification labels. Fig. 1 presents the overall architecture of the proposed method. The feature extractor captures the deep convolutional feature map from a given mammogram. The AFDS structure adaptively selects discriminative deep feature descriptors. And the logistic regression layer is used to compute the malignancy probability of an input mammogram.

CNNs are employed as the feature extractor, which could be defined as f_e. For the input image x, $F = f_e(x|\theta_e)$. $F \in \mathbb{R}^{C\times H\times W}$ represents the extracted deep feature maps; C, H, and W represent the channel, height, and weight dimensions, respectively; θ_e denotes the parameter of the feature extractor. The feature maps F will be fed into the AFDS structure to generate a discriminative representation for the input mammogram. The AFDS structure can be formulated as $\bar{F} = f_a(F)$, where f_a is the mapping of the AFDS structure, and $\bar{F}$ stands for the discriminative representation, which includes only the discriminative feature descriptors. The details of the AFDS structure will be introduced in the following subsection.

We apply average pooling on $\bar{F}$ to get a fixed-length vector $v \in \mathbb{R}^{C\times 1\times 1}$. And then, we fed v into the linear regression layer to calculate the probability of malignancy. We treat mammogram classification as a binary classification task to predict whether a mammogram contains a malignant lesion or not. The linear regression layer can be formulated as $p = \delta(f_l(v|\theta_l))$, where p is the probability of malignancy, δ is the sigmoid active function, and θ_l denotes the parameter of the linear regression layer.

B. AFDS Structure

This subsection introduces the details of the AFDS structure, which generates the discriminative representation for a given mammogram. The principle of the AFDS structure is illustrated in the middle part of Fig. 1. As previously mentioned, identifying discriminative lesion areas in mammograms is essential to diagnosis. However, selecting lesion areas from a mammogram without detection bounding box or segmentation ground truth is difficult. In this study, we treat the problem as another task: selecting discriminative feature descriptors from deep feature maps. Specifically, after obtaining the feature maps F of mammogram x, we generate the activation map A by compressing the feature maps F using channel-wise average pooling, which could be computed as (1).

$$A = \sum_{i=1}^{C} F_i, \tag{1}$$

where F_i is the $i-th$ feature map in the obtained feature maps, and $A \in \mathbb{R}^{H\times W}$ is the activation map that contains $H \times W$

activation responses. Each activation response in A corresponds to an area in the input mammogram. Because the model is trained for the classification task, the intensity of each pixel (i, j) in A is proportional to the discriminative power (probability of malignant) of the corresponding area in the input mammogram. Thus, the spatial discriminative power of the input mammogram can be effectively approximated by the distribution of the activation map.

Next, we calculate a specific threshold t to decide which feature descriptors are discriminative and could localize targets: the position (i, j) whose activation response is higher than t is considered as a discriminative feature descriptor. Based on the threshold t, a mask map $M \in \mathbb{R}^{H \times W}$ with the same size as A can be obtained using (2).

$$M_{i,j} = \begin{cases} 1, & \text{if } A_{i,j} \geq t, \\ 0, & otherwise, \end{cases} \tag{2}$$

where $A_{i,j}$ is the activation response value at position (i, j) in A, and $M_{i,j}$ is the value at position (i, j) in M. The descriptor $F_{i,j} \in \mathbb{R}^{C \times 1 \times 1}$ should be kept when $M_{i,j} = 1$, which means the position (i, j) is discriminative and might contain lesions. And it should be ignored when $M_{i,j} = 0$, which means the position (i, j) might be a background or noisy part. In this way, we obtain the discriminative representation $\bar{F}$, which is aggregated with the selected feature descriptors. It could be formulated as (3):

$$\bar{F} = \left\{ F_{(i,j)} \mid M_{i,j} = 1 \right\}. \tag{3}$$

Then, we apply average pooling on $\bar{F}$ to get a fixed-length vector v. In the implementation, we perform element-wise multiplication on the feature maps F and the mask map M. Then we apply global average pooling on the selected matrix $F \circ M$. It can be formulated as (4).

$$\begin{aligned} v &= \frac{1}{S} \sum_{i=1}^{S} \bar{F}_i = \frac{1}{S} \sum_{i=1}^{W} \sum_{j=1}^{H} (F \circ M)_{i,j} \\ &= \frac{W \times H}{S} GAP(F \circ M), \end{aligned} \tag{4}$$

where S is the number of the selected descriptors, $\circ$ denotes the element-wise multiplication of two matrices, and GAP is the global average pooling operation.

Then, the main problem is how to choose an appropriate value for the threshold t for the purpose of adaptive selecting feature descriptors (local areas). As previously mentioned, Choe et al. and Wei et al. [19] simply used the maximum or mean value with a ratio as the threshold, which could not change adaptively for different data distributions. In this study, we observe that the histogram of a mammogram activation map usually presents one extremely high peak value due to the lesion areas. Since the triangle threshold [21] is more suitable under such conditions [33], it is adopted to calculate a specific threshold in this study. The triangle algorithm was first proposed in a chromosome study and used to find the optimal threshold on a histogram using the geometry method [21]. As illustrated in Fig. 2, when using the triangle algorithm, a line L is constructed between the maximum and minimum values on the histogram,

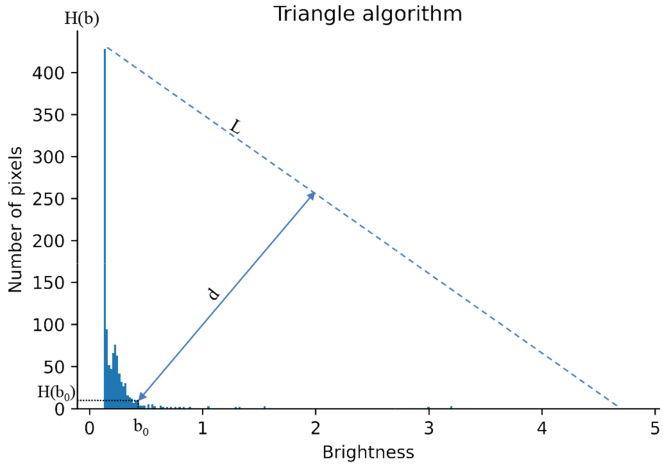

Fig. 2. Triangle algorithm used to select the adaptive threshold t.

which correspond to the brightest and darkest regions on the image. Then, the value b_0, where the distance between $H[b_0]$ and the straight line L is maximal, is selected as the triangle threshold.

Fig. 3 compares several threshold methods through visualization. The isodata threshold is an iterative technique for choosing a threshold t to separate foreground and background. As shown in the second row, the activation maps generated with this method can roughly localize lesions on malignant samples. Still, the entire breast area is highlighted for normal/benign mammograms. This illustrates that although this model is trained without any annotated lesion information, the model still could learn discriminative lesion information.

The third row shows that the lesions localized by the isodata threshold strategy contain the slightest noise or background on malignant mammograms. However, it may mistake confused areas (like nipples) as lesions in normal/benign samples. It can be seen in the fourth row that the mask map generated by the triangle thresholding strategy locates lesions mixed with a small amount of noise in malignant mammograms. In addition, it selects the entire breast area in normal/benign mammograms. As shown in the last row, the lesions localized by the mean-threshold strategy contain a great deal of noise and background in malignant mammograms. Moreover, the mean-threshold strategy localizes part of the breast without any rules in normal/benign mammograms. Synthetically, it seems that the isodata threshold strategy has potential advantages for malignant mammograms but great limitations for benign mammograms. The triangle threshold strategy offers a more balanced performance on benign and malignant mammograms. The mask map generated by the mean threshold strategy seems to lack regularity. We will carry out systematic experiments and evaluations on these threshold strategies in Section IV.

In addition, the proposed AFDS structure can be considered as a pooling structure. But unlike average pooling, which selects all features to represent the image, and max-pooling, which selects the largest value in each channel to represent the image, the AFDS method only selects discriminative feature descriptors to represent the entire image.

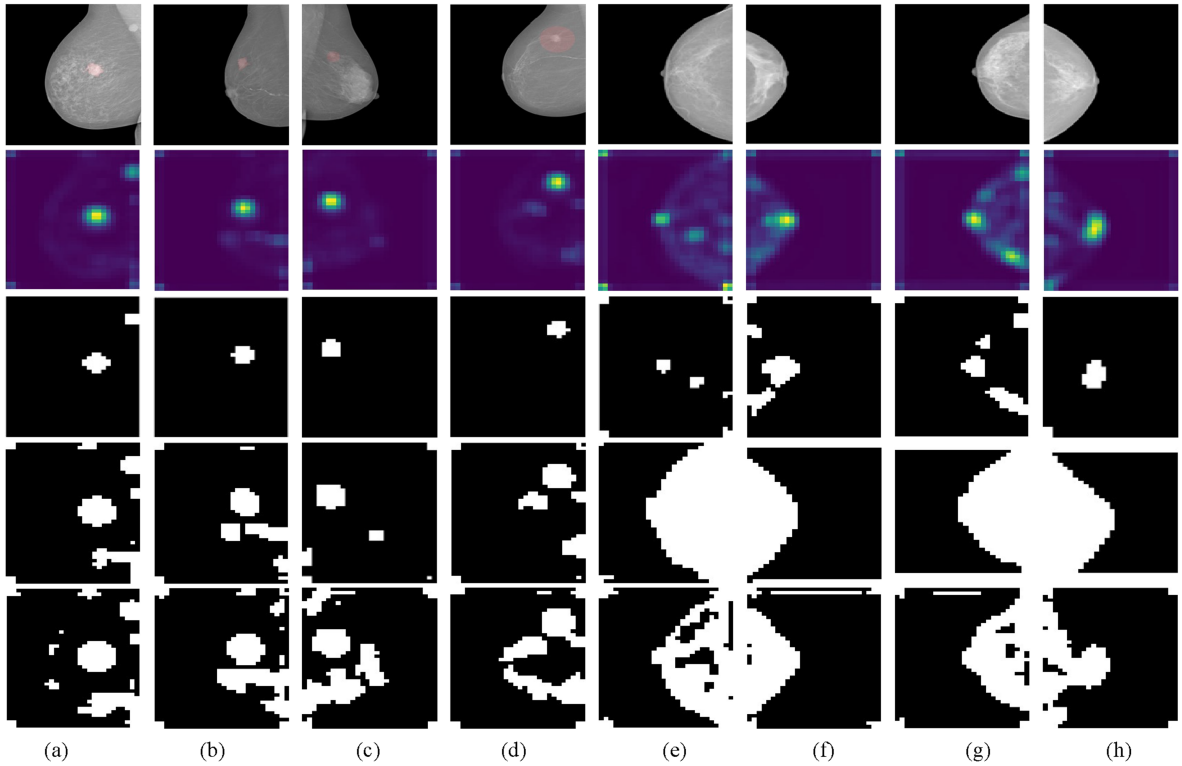

Fig. 3. Mammogram samples, activation maps, and mask maps. The first row shows some samples of input mammograms. The lesion area is masked by red color on the input mammograms. The second row displays corresponding activation maps. The last three rows are the corresponding mask maps generated by isodata threshold, triangle threshold, and mean threshold. Mammogram samples (a)–(d) are positive (i.e., malignant), and mammogram samples (e)–(h) are negative (i.e., normal/benign).

C. Loss Function

In this study, we use focal loss [35] as the loss function instead of the cross-entropy loss, which assigns larger weights to hard samples. The focal loss could be defined as (5) in this binary classification problem.

$$\mathcal{L} = \alpha(1-p)^{\gamma} y log(p) - (1-\alpha)p^{\gamma}(1-y)log(1-p), \quad (5)$$

where $p \in [0, 1]$ represents the model's estimated probability of being positive, and $y = [0, 1]$ is the true label of the given sample. Weight α is added to different terms to control the ratios between different categories of errors. The parameter γ indicates the level of attention paid to hard samples. Lin et al. get the best performance when $\alpha = 0.25$ and $\gamma = 2$. In order to continuously increase the weight of hard samples during training, this study dynamically increases the value of γ from 1 to 2.5 during training. Following Lin et al., α is set to 0.25 in this study.

The process of the proposed AFDS method is shown in Algorithm 1.

IV. Experiments

A. Datasets

We conduct experiments on two commonly used, publicly available datasets—the INbreast dataset [36] and the CBIS-DDSM dataset [37].

1) INbreast Dataset: The INbreast dataset contains 410 full-field digital mammograms from 115 cases. Based on the suspicion level, mammograms are placed into 6 BI-RADS categories. In this study, we treat categories 0 to 3 as normal/benign and categories 4 to 6 as malignant. We divide the INbreast dataset into training and test sets according to the proportion of 4:1. Each subset keeps the consistency of the original data distribution as much as possible. Then we utilize 5-fold cross-validation to evaluate different models, which improves the stability and reliability of the evaluation results. The statistic of this dataset is shown in Table I.

2) CBIS-DDSM Dataset: The CBIS-DDSM (Curated Breast Imaging Subset of DDSM) dataset consists of decompressed images and precise annotations, containing 3,103 images from 1,645 patients. We divide the CBIS-DDSM dataset according to its original database. The statistic of this dataset is shown in Table I.

In the pre-processing procedure, we resize the mammogram to the size of 1350 × 950 pixels and then randomly crop it to 1280 × 896 pixels in the training set to reduce the influence of over-fitting. In the test set, it is center cropped to 1280 × 896 pixels after resizing. We utilize augmentation methods, such as random rotation from -30 to +30 degrees, random vertical flip, and random horizontal flip, to train the model sufficiently. We normalize the augmented data to [0,1] by the mean and variance. We do not use any technology to locate the breast and remove the background from mammograms before feeding the mammograms into the model, because the proposed method

Algorithm 1: The Process of the Proposed AFDS Method.

Input:
Dataset: $D = \{(x_i, y_i); i = 1, 2, \ldots, N\}$;
The maximum epoch: T;
The learning rate: lr.
Output:
The learned parameters: $\theta = \{\theta_e, \theta_l\}$.
for *epoch* $e = 1 : T$ **do**
  for *sample* $i = 1 : N$ **do**
    Pre-processing and augmenting on image x_i;
    Extracting feature maps F_i of image x_i through $F_i = f_e(x_i|\theta_e)$;
    Calculating the activation map A_i using (1);
    Calculating the threshold t by the triangle algorithm;
    Calculating the mask map M_i using (2);
    Generating discriminative representation $\bar{F}_i$ using (3);
    Applying average pooling on $\bar{F}_i$ to get the fixed-length vector v_i;
    Inputting v_i into the linear regression layer to calculate the malignant probability p_i of x_i;
    Calculating the focal loss $\mathcal{L}$ using (5);
    Calculating gradients of $\mathcal{L}$ to parameters θ;
    Updating parameters $\theta : \theta \leftarrow \theta - lr \cdot \frac{\partial \mathcal{L}}{\partial \theta}$.
  end for
end for

TABLE I
CATEGORY DISTRIBUTIONS OF TWO DATASETS

Dataset	Malignant	Normal/Benign	Total
CBIS-DDSM dataset	1375	1728	3103
INbreast dataset	100	310	410

is intended to select the discriminative descriptors that could automatically remove the background and noise.

B. Metrics

Following most medical image classification work, we use ACC (accuracy) [38], ROC (receiver operating characteristic) [39], AUC (area under the ROC curve) [40], sensitivity [38], specificity [38], and ECE (expected calibration error) [41] as evaluation metrics in this study.

ACC is the ratio between the number of correctly classified samples and the total number of samples in the evaluation dataset, as shown in (6).

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \tag{6}$$

where TP (true positive) denotes the number of correctly classified positive samples, TN (true negative) denotes the number of correctly classified negative samples, FP (false positive) is the number of samples incorrectly classified as positive, FN (false negatives) indicates the number of samples incorrectly classified as negative.

The specificity is calculated as the number of correct negative predictions divided by the total number of negatives, as shown in (7).

$$specifcity = \frac{TN}{TN + FP}. \tag{7}$$

The sensitivity is calculated as the number of correct positive predictions divided by the total number of positives, as shown in (8).

$$sensitivity = \frac{TP}{TP + FN}, \tag{8}$$

ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

AUC is the area under the ROC curve. It is a classification function that expresses the probability that a randomly selected positive example gets a higher score than a randomly selected negative example.

To verify the confidence and uncertainty [42] of the proposed model, we also conduct ECE performance tests. ECE is a weighted average of the absolute difference between the accuracy and confidence of each bin. The calculation of ECE is shown in (9).

$$ECE = \sum_{b=1}^{B} \frac{|n_b|}{N}, \left| acc(b) - conf(b) \right| \tag{9}$$

where B is the number of bins, n_b is the number of predictions in bin b, and N is the total number of data points. $acc(b)$ and $conf(b)$ are the accuracy and confidence of bin b, respectively.

C. Experimental Design

We experimented with CUDA 10.1 on an Ubuntu server with two NVIDIA Tesla V100 GPUs. Pytorch with python is used to construct and train the proposed model. The model is trained for 100 epochs with the Adam optimizer [43] end to end. The momentum and weight decay for Adam are both set to 5×10^{-5} empirically. The learning rate for the feature extractor layer is initialized to 5×10^{-5} and decayed every 5 epochs with a decay rate of 0.9. For the logistic regression layer, the learning rate is initialized to 1×10^{-4} and decayed every 5 epochs with a decay rate of 0.85.

The batch size is 8. We choose the pre-trained DenseNet169 [44] with ImageNet to serve as the feature extractor. The shape of the feature maps is $1664 \times 40 \times 28$.

D. Ablation Experiments

We evaluate three different threshold strategies for the AFDS structure, including the isodata threshold strategy, the triangle threshold strategy, and the mean threshold strategy. As shown in Fig. 3, different threshold strategies could generate different mask maps, then make the selected feature descriptors different, which further affects the performance of the model.

TABLE II
ABLATION EXPERIMENT RESULTS OF AFDS STRUCTURE WITH THREE THRESHOLD STRATEGIES

Method	dataset	ACC (%)	AUC (%)
AFDS with triangle threshold	INbreast	**92.4**	**97.2**
AFDS with isodata threshold	INbreast	92.0	95.6
AFDS with mean threshold	INbreast	90.2	96.7
AFDS with triangle threshold	CBIS-DDSM	**79.7**	**86.2**
AFDS with isodata threshold	CBIS-DDSM	77.1	84.4
AFDS with mean threshold	CBIS-DDSM	75.6	84.7

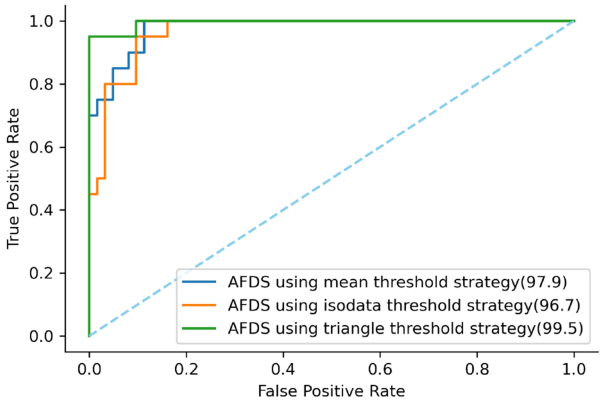

Fig. 4. ROC curve of the AFDS structure with three threshold strategies on one partition of the INbresast dataset.

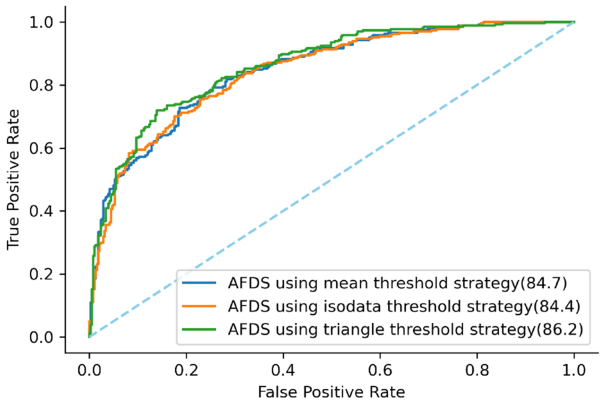

Fig. 5. ROC curve of the AFDS structure with three threshold strategies on the CBIS-DDSM dataset.

From Table II, Figs. 4 and 5, it can be seen that different threshold strategies have different effects on the performance of the model. The triangle threshold strategy yields the best ACC and AUC for both the INbeast and the CBIS-DDSM datasets. This result is consistent with the previous discussion in Section III. The AFDS structure with isodata threshold does not perform well because, although it localizes more precise lesions on malignant samples, the nipple is often incorrectly extracted as a lesion in normal/benign mammogram samples. Meanwhile, the AFDS structure with the mean threshold strategy does not perform well because the mask map generated by the mean threshold cannot localize precise lesions without noise and seems to lack regularity. The mask map generated using the triangle threshold localizes lesions mixed with little noise or background in malignant mammogram samples and selects the entire breast in normal/benign mammogram samples. This allows the model to perform well on both malignant and normal/benign mammogram samples, thereby improving the performance of the model.

TABLE III
COMPARISON OF AFDS STRUCTURE WITH COMMONLY USED POOLING STRUCTURES

Method	dataset	ACC (%)	AUC (%)
AFDS	INbreast	**92.4**	**97.2**
Max-pooling	INbreast	89.0	93.6
Average-pooling	INbreast	91.7	95.9
AFDS	CBIS-DDSM	**79.7**	**86.2**
Max-pooling	CBIS-DDSM	74.6	83.9
Average-pooling	CBIS-DDSM	75.0	82.7

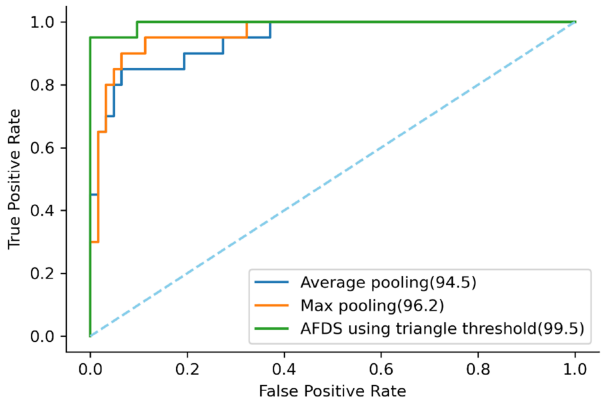

Fig. 6. ROC curve of the three pooling structures on one partition of the INbresast dataset.

E. Comparative Experiments

Because the proposed AFDS structure can be considered as a novel pooling structure, we compare the performance of the proposed AFDS structure with the max-pooling and the average-pooling structures on the INbreast and CBIS-DDSM datasets.

According to Table III, Figs. 6 and 7, it can be seen the model with the AFDS structure obtains better performance on both INbreast and CBIS-DDSM datasets. We believe it is because the proposed AFDS structure selects discriminative feature descriptors to represent the whole image, which makes the model more efficient and robust. In contrast, the average-pooling structure selects all features to represent the image. But the areas of interest only take a few parts suspected of containing malignant lesions. Therefore, the average-pooling structure that performs well on natural images may not suit the mammogram classification task. The max-pooling structure selects the unique maximum value in each channel to represent the image, which does not take full advantage of all lesion areas and increases the negative impact of confounding areas, such as nipples.

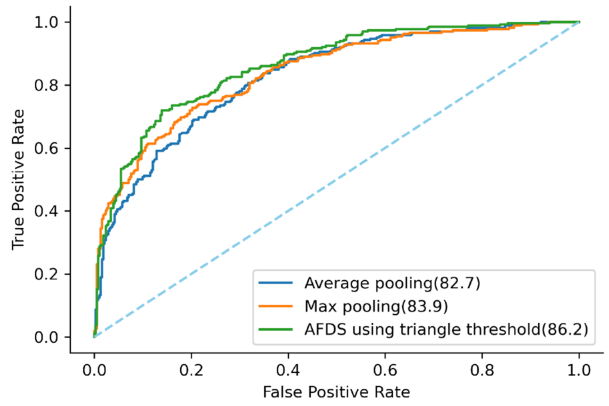

Fig. 7. ROC curve of the three pooling structures on the CBIS-DDSM dataset.

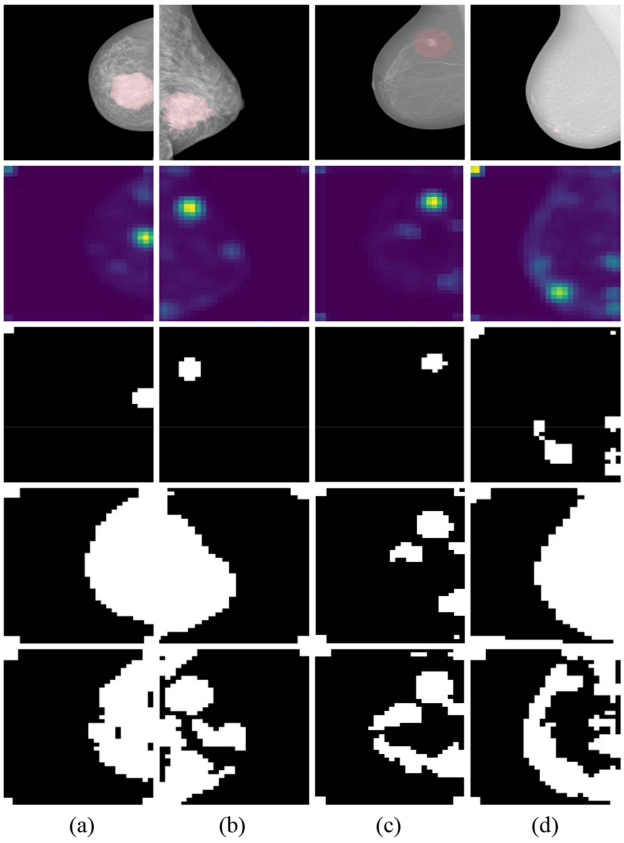

Fig. 8. Visualization results of some extreme cases. The first row is a sample of a mammogram, on which the lesion area is masked by red color. The second row shows the corresponding activation maps. The last three rows are the corresponding mask maps generated by the isodata threshold, triangle threshold, and mean threshold.

Next, we compare the proposed method with state-of-the-art mammogram classification methods, such as densenet169, RGP, and GGP proposed by Shu et al. [12], MIL proposed by Zhu et al. and MS proposed by Xie et al. [15], on the INbreast dataset. For a fair comparison, we implement these methods and conduct experiments under the same data pre-processing, dataset partition, backbone, and parameter settings. We use the symbol * to identify the re-implemented version, e.g., MIL*. The results reported in their paper are also listed.

TABLE IV
COMPARISON OF PROPOSED METHOD WITH STATE-OF-THE-ART METHODS ON THE INBREAST DATASET

Method	ACC (%)	AUC (%)	Sensitivity (%)	Specificity (%)	ECE
MIL [13]	90.0	85.9	N/A	N/A	N/A
RGP [12]	91.9	93.4	N/A	N/A	N/A
GGP [12]	92.2	92.4	N/A	N/A	N/A
MS [15]	96.1	97.1	N/A	N/A	N/A
MIL* [13]	90.5	94.3	73.0	96.1	0.74
RGP* [12]	91.9	95.2	**78.0**	96.5	0.75
GGP* [12]	92.2	95.4	77.0	97.1	0.75
MS* [15]	90.0	95.4	74.0	95.1	0.77
Our method	**92.4**	**97.2**	75.0	**97.4**	**0.70**

From Table IV, it can be seen that the proposed method obtained the best performance on AUC (97.2%), ACC (92.4%), specificity (97.4%), and ECE (0.7), outperforming state-of-the-art methods. For Shu et al.'s method, the AUC performance in the re-implemented version is slightly higher than that reported in their paper, and the ACC performance is almost the same. We think this should be caused by the higher-resolution images used in our study. For the method proposed by Xie et al., the ACC and AUC performance in the re-implemented version are lower than their reported ones, which may be caused by different data pre-processing methods performed in their experiment. They used the breast region segmentation (BRS) module to pre-process input images. This study focused on the model's performance, so we did not use any additional data pre-processing method in our experiments.

Finally, we compare the proposed method with state-of-the-art mammogram classification methods, such as RGP and GGP proposed by Shu et al. [12], MIL proposed by Zhu et al., MS proposed by Xie et al. and GMIC proposed by Shen et al. [14], on the CBIS-DDSM dataset.

From Table V, it can be seen that the proposed AFDS method obtained the best performance on ACC (79.7%), AUC (86.2%), sensitivity (74.6%), and ECE (0.54). Because the model proposed by Shen et al. is not trained end to end, we did not implement it in this study. Shen et al. reported that the GMIC method achieved an AUC performance of 84.0% using GMIC-ResNet18. They also applied the model ensemble to further improved the AUC performance to 85.8%. However, our method outperforms both their single GMIC model and ensemble models. For Shu et al.'s method, the re-implemented versions perform better than they reported, which is the same as the experiments on the INbreast dataset. Xie et al. and Zhu et al. did not report the performances on the CBIS-DDSM dataset. In our re-implemented version, the proposed ADFS method still outperforms them on most of the metrics.

F. Discussion

This study compares the proposed method with state-of-the-art methods on two publicly available datasets. Experimental

TABLE V
COMPARISON OF PROPOSED METHOD WITH STATE-OF-THE-ART METHODS ON THE CBIS-DDSM DATASET

Method	ACC (%)	AUC (%)	Sensitivity (%)	Specificity (%)	ECE
RGP [12]	76.2	83.8	N/A	N/A	N/A
GGP [12]	76.7	82.3	N/A	N/A	N/A
GMIC-ResNet18 [14]	N/A	84.0	N/A	N/A	N/A
GMIC-ensemble [14]	N/A	85.8	N/A	N/A	N/A
MIL* *et al.* [13]	73.64	80.5	65.2	79.5	0.58
MS* [15]	78.14	85.16	72.0	82.4	0.59
RGP* [12]	75.0	84.3	68.9	79.3	0.58
GGP* [12]	77.2	84.0	67.8	**83.7**	0.60
Our method	**79.7**	**86.2**	**74.6**	80.0	**0.54**

TABLE VI
TIME CONSUMPTION OF DIFFERENT METHODS

Method	ms/image	image/s
MIL* [13]	22.75	43.95
RGP* [12]	22.75	43.95
MS* [15]	27.13	36.87
Densenet169	22.75	43.95
AFDS with mean threshold	22.88	43.72
AFDS with isodata threshold	22.88	43.72
AFDS with triangle threshold	23.13	43.24

results have demonstrated that the proposed method outperforms state-of-the-art methods on different metrics such as ACC, AUC, and ECE. In the experiments on the INbreast dataset, the proposed method outperforms other methods except for sensitivity. And on the CBIS-DDSM dataset, the AFDS method is superior to other methods except for specificity. Nevertheless, the proposed method outperforms state-of-the-art methods by 1-2 points in the AUC evaluation under the same conditions. AUC is a measure of the overall performance and is interpreted as the average value of sensitivity for all possible values of specificity [45]. These results indicate that the proposed method has a better overall ability for mammogram classification. In addition, the proposed model also performs best in the ECE evaluation, which proves the advantages of the proposed model in terms of confidence and uncertainty.

To evaluate the complexity and time consumption of the proposed method, we further compare the time consumption of our model with state-of-the-art methods, other threshold methods, and the vanilla DenseNet169 network. The hardware used in the experiments is a Tesla V100 GPU. The results are shown in Table VI. Compared with the DenseNet169 network, the proposed structure only increases the computation time by 0.38 ms. And it only increases by 0.25 ms compared with the mean threshold. It shows that the calculation of the triangle threshold is not expensive or even could be negligible.

To further discuss the capabilities and inadequacies of the model, some extreme samples are visualized in Fig. 8. Samples (a) and (b) contain very large lesions. Sample (c) contains a normal-sized lesion. The lesion in sample (d) is quite small. All those four samples are malignant.

As shown in Fig. 8, the triangle threshold method always produces a mask map covering the entire breast in cases with extremely large or small lesions. In contrast, the areas selected by the isodata method cannot cover the entire lesion area (Fig. 8(a)) and may even be misled to select some non-lesion areas (Fig. 8(b)). The areas selected by the mean threshold are always very irregular. These results indicate that the triangle method also has advantages in these extreme cases.

Although the proposed method has achieved convincing results, there also have some limitations. The triangle threshold tends to produce a more oversized mask map to cover the entire breast in cases with extremely large or small lesions. And it always generates a mask map mixed with noise in cases with normal-sized lesions. In contrast, the mask map generated by the isodata threshold seems to contain less noise for normal-sized malignant lesions, but the performance degrades in malignant cases with large/small lesions or normal/benign cases. This phenomenon provides some ideas for future research. In future work, we plan to find ways to improve the precision of mask maps for malignant lesions, and convert the classification problem into the weakly supervised object localization(WSOL) task.

V. CONCLUSION

This study proposes an image-level label-based deep-learning method for breast cancer diagnosis in mammography. For this method, this study designs a novel AFDS structure to select discriminative feature descriptors (lesion areas) adaptively. The AFDS structure employs the triangle threshold strategy to compute a threshold for guiding the activation map to determine which feature descriptors are discriminative. Visualization results demonstrate that the proposed AFDS structure makes it easier for models to learn the difference between malignant and normal/benign samples. Experiment results indicate the effectiveness and robustness of the proposed method. Furthermore, the AFDS structure can be easily plugged into most existing CNN models with negligible effort and time consumption. Nevertheless, the produced masks have some limitations, such as introducing noise in the mask map of malignant cases. We will explore the rules and improve them in future work.

REFERENCES

[1] H. Sung et al., "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] W. Y. Joko-Fru et al., "The evolving epidemic of breast cancer in sub-saharan africa: Results from the african cancer registry network," *Int. J. Cancer*, vol. 147, no. 8, pp. 2131–2141, 2020.

[3] F. Bray, P. McCarron, and D. M. Parkin, "The changing global patterns of female breast cancer incidence and mortality," *Breast Cancer Res.*, vol. 6, no. 6, pp. 1–11, 2004.

[4] L. Wang, "Early diagnosis of breast cancer," *Sensors*, vol. 17, no. 7, 2017, Art. no. 1572.

[5] B. Lauby-Secretan et al., "Breast-cancer screening–viewpoint of the IARC working group," *New England J. Med.*, vol. 372, no. 24, pp. 2353–2358, 2015.

[6] L. Liberman and J. H. Menell, "Breast imaging reporting and data system," *Radiol. Clin.*, vol. 40, no. 3, pp. 409–430, 2002.

[7] M. S. Bae et al., "Breast cancer detected with screening us: Reasons for nondetection at mammography," *Radiology*, vol. 270, no. 2, pp. 369–377, 2014.

[8] I. Christoyianni, A. Koutras, E. Dermatas, and G. Kokkinakis, "Computer aided diagnosis of breast cancer in digitized mammograms," *Computerized Med. Imag. Graph.*, vol. 26, no. 5, pp. 309–319, 2002.

[9] C. D. Lehman et al., "Diagnostic accuracy of digital screening mammography with and without computer-aided detection," *JAMA Intern. Med.*, vol. 175, no. 11, pp. 1828–1837, 2015.

[10] 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. Interv.*, Springer, 2015, pp. 652–660.

[11] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, *arXiv:1706.05098*.

[12] X. Shu, L. Zhang, Z. Wang, Q. Lv, and Z. Yi, "Deep neural networks with region-based pooling structures for mammographic image classification," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 2246–2255, Jun. 2020.

[13] 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. Interv.*, Springer, 2017, pp. 603–611.

[14] Y. Shen et al., "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization," *Med. Image Anal.*, vol. 68, 2021, Art. no. 101908.

[15] L. Xie, L. Zhang, T. Hu, H. Huang, and Z. Yi, "Neural networks model based on an automated multi-scale method for mammogram classification," *Knowl.-Based Syst.*, vol. 208, 2020, Art. no. 106465.

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

[17] J. Tang, R. M. Rangayyan, J. Xu, I. El 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.

[18] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, "Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization," *Pattern Recognit.*, vol. 76, pp. 704–714, 2018.

[19] X.-S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2868–2881, Jun. 2017.

[20] J. Choe and H. Shim, "Attention-based dropout layer for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2219–2228.

[21] G. W. Zack, W. E. Rogers, and S. A. Latt, "Automatic measurement of sister chromatid exchange frequency," *J. Histochemistry Cytochemistry*, vol. 25, no. 7, pp. 741–753, 1977.

[22] 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.

[23] Y. Wang et al., "Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images," *Med. Image Anal.*, vol. 81, 2022, Art. no. 102535.

[24] L. Zhen, P. Hu, X. Peng, R. S. M. Goh, and J. T. Zhou, "Deep multimodal transfer learning for cross-modal retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 798–810, Feb. 2022.

[25] N. Wu et al., "Deep neural networks improve radiologists' performance in breast cancer screening," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 1184–1194, Apr. 2020.

[26] D. Li, L. Wang, T. Hu, L. Zhang, and Q. Lv, "Deep multi-instance mammogram classification with region label assignment strategy and metric-based optimization," *IEEE Trans. Cogn. Devel. Syst.*, vol. 14, no. 4, pp. 1717–1728, Dec. 2022.

[27] Y. B. Hagos, A. G. Mérida, and J. Teuwen, "Improving breast cancer detection using symmetry information with deep learning," in *Proc. Image Anal. Moving Organ, Breast, Thoracic Images*, Springer, 2018, pp. 90–97.

[28] N. Dhungel, G. Carneiro, and A. P. Bradley, "The automated learning of deep features for breast mass classification from mammograms," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Springer, 2016, pp. 106–114.

[29] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[31] X. Gao, Y. Tian, and Z. Qi, "Multi-view feature augmentation with adaptive class activation mapping," 2022, *arXiv:2206.12943*.

[32] Y. Tian, W. Zhang, Q. Zhang, G. Lu, and X. Wu, "Selective multi-convolutional region feature extraction based iterative discrimination cnn for fine-grained vehicle model recognition," in *Proc. 24th Int. Conf. Pattern Recognit.*, 2018, pp. 3279–3284.

[33] M.-T. Le, T. R. Bretschneider, C. Kuss, and P. R. Preiser, "A novel semi-automatic image processing approach to determine plasmodium falciparum parasitemia in giemsa-stained thin blood smears," *BMC Cell Biol.*, vol. 9, no. 1, pp. 1–12, 2008.

[34] T. Ridler et al., "Picture thresholding using an iterative selection method," *IEEE Trans. Syst. Man Cybern.*, vol. 8, no. 8, pp. 630–632, Aug. 1978.

[35] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[36] 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, no. 1, pp. 1–9, 2017.

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

[38] S. A. Hicks et al., "On evaluation metrics for medical applications of artificial intelligence," *Sci. Rep.*, vol. 12, no. 1, pp. 1–9, 2022.

[39] J. P. Egan and J. P. Egan, *Signal Detection Theory and ROC-Analysis*. Cambridge, MA, USA: Academic press, 1975.

[40] J. A. Hanley and B. J. McNeil, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases," *Radiology*, vol. 148, no. 3, pp. 839–843, 1983.

[41] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330.

[42] S. Deshpande, J. Lengiewicz, and S. P. Bordas, "Probabilistic deep learning for real-time large deformation simulations," *Comput. Methods Appl. Mechanics Eng.*, vol. 398, 2022, Art. no. 115307.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[44] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[45] S. H. Park, J. M. Goo, and C.-H. Jo, "Receiver operating characteristic (ROC) curve: Practical review for radiologists," *Korean J. Radiol.*, vol. 5, no. 1, pp. 11–18, 2004.