IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 13, 2020, accepted April 4, 2020, date of publication April 8, 2020, date of current version May 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2986546

Mammographic Classification Based on XGBoost and DCNN With Multi Features

RUNYU SONG, TAOYING LI, AND YAN WANG
School of Shipping Economics and Management, Dalian Maritime University, Dalian 116026, China

Corresponding author: Taoying Li (litaoying@dlmu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 51939001 and 61976033, in part by the Liaoning Revitalization Talents Program under Grant XLYC1907084, in part by the Natural Science Foundation of Liaoning Province under Grant 20180550307, in part by the Double First-Class Construction Special Items (''Innovative Project'') under Grant CXXM2019SS016, and in part by the Fundamental Research Funds for the Central Universities under Grant 3132019353 and 3132020233.

ABSTRACT The classification of benign and malignant masses in mammograms by Computer-Aided Diagnosis (CAD) is one of the most difficult and important tasks in the development of CAD systems. This classification has commonly been automated by extracting a set of handcrafted features from mammograms and relating the responses to breast cancer. Recently, the application of Deep Learning (DL) technology in medical imaging informatics has been attracting extensive research interest. However, limited medical image datasets and feature expression often reduce the performance of DL-based schemes. Therefore, this study aims to develop a new combined feature CAD method based on DL for classifying mammographic masses into three classes: normal, benign and cancer (malignant) masses. Three kinds of breast masses were scored by using Deep Convolution Neural Network (DCNN) as a feature extractor. Then the scoring features are combined with the image texture features as input to the classifier. This features including the scoring features, Gray-Level Co-occurrence Matrix (GLCM) and Histogram of Oriented Gradient (HOT) were employed to extract the breast mass information in mammograms and the classifier of Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost) were trained for the classification task. Accuracy (ACC), Precision (Pre), Recall (Rec), F_1-score (F_1), and Overall Accuracy (Overall ACC) are used to evaluate the performance of the proposed system and the results show that the proposed multi-features combination model performs the best results. The performance of the XGBoost classifier has proved to be better in comparison to the SVM classification algorithms. As a result, when XGBoost was used as a classifier, the correct identification rate of the Overall ACC was 92.80% and that of malignant tumors was 84%, with reasonable and best results. These results indicate that the proposed method may help in more accurately diagnosing cases that are difficult to classify on images.

INDEX TERMS Deep learning, computer-aided diagnosis, deep convolution neural network, mammograms classification.

I. INTRODUCTION

Breast cancer is a huge health threat [1], presenting an increasing incidence and mortality rate in all age groups in the past decades [2]. And it is one of the most common causes of cancer deaths in women worldwide and it is responsible for 23% of all cancer cases and 14% of cancer-related deaths amongst women [3]. Early detection is the key to improve the prognosis of breast cancer [4]. Early diagnosis can significantly improve the chances of recovery: the 5-year relative survival rate increased from 24% when breast cancer is diagnosed at a distant stage to 99% if it is diagnosed at a localized stage.

Currently, mammography is the most effective and widely accepted method among all the imaging techniques for breast examination, and it is also the world recognized gold standard tool for breast cancer detection. Mammogram based diagnosis enables the radiologists to diagnose the breast cancer precisely as compared with symptoms based diagnosis. The improvement of breast cancer treatment methods and the wide application of breast cancer screening technology. Especially the wide use of mammography technology can early detect the occult breast cancer in asymptomatic women, which greatly promotes this favorable trend of effectively reducing mortality [2]. When breast cancer is identified as

The associate editor coordinating the review of this manuscript and approving it for publication was Quan Zou.

an early stage, it is more likely to respond to treatment and increases the survival opportunities for patients. During the screening process, radiologists examine mammograms and look for several important signs of breast cancers, such as clusters of microcalcifications, masses, and architectural distortions. All these findings may indicate the presence of cancer [5].

The clinical research report pointed out that in various types of breast abnormalities, breast masses are the most important findings since they may indicate the presence of malignant tumors [6], and also an important indicator of the development of early breast cancer [7]–[9]. Mass detection is a more complex task because it is often unable to distinguish from adjacent tissues. Moreover, humans are prone to make mistakes and the wrong diagnosis may lead to the incurable stage of breast cancer. Mass detection poses more difficult tasks in locating and identifying quality boundaries because it is often: (a) very pronounced in size, shape, and density; (b) low in image contrast and signal-to-noise ratio [10]; (c) high similarity with the surrounding healthy parenchymal tissue density, particularly for speculated lesions and (d) surrounded by no uniform tissue background with similar characteristics [11]. Besides, the morphological information of tumor shape (irregular, lobular, oval and round) and margin type (circumscribed, ill-defined, tapered and obscured) also play crucial roles in the diagnosis of tumor malignancy [12]. That makes progress to be considerably slow for mass detection. As a result, detection sensitivity and specificity of screening mammography is not optimal [13]. Generally speaking, an inspection of the generated large quantities of mammograms by experienced radiologists is tedious and subjective, which suffers from poor inter and intraobserver reproducibility [14], [15]. Therefore, a large part of the mass was missed by the radiologist.

Because of the clinical significance and great challenge of mammographic mass detection, since the 1960s [16], numerous computer-aided diagnosis (CAD) systems and quantitative image (QI) analysis technologies provide an assistive tool to the radiologist to reduce the false diagnosis rate and increase the accuracy of diagnosis. A majority of these approaches extract some certain features by hand-engineered, which employs a combination of heuristic and mathematical descriptors. Subsequently, the extracted features and pretrained classifiers are used to classify these masses or normal tissues. This feature extraction step mostly depends on the features extracted from the data and requires effort and sufficient interpreting knowledge due to the various geometrical and morphological structures. All these processing steps are equally important for efficient diagnosis. These systems are aimed at improved identification of subtle suspicious masses, calcifications, micro-calcifications and other abnormalities in mammograms [17], [18]. Using conventional machine learning methods, various hand-designed descriptors (i.e., morphological, topological and textural features) based on prior knowledge and expert guidance have been developed for these CAD systems. Previous publications have described and compared such approaches for automatic detection and segmentation of abnormalities in mammographic images [18]–[20]. First segmented adaptive regions of interest (ROIs) as suspicious areas, and then, classified each ROI using complex texture features and stepwise linear discriminant analysis. However, due to the low signal-to-noise ratio and variability in shape, size, appearance, texture, and location, breast tumor segmentation and classification is still a problem.

In recent years, to improve the performance of breast mass classification, many researches based on deep representation of breast images and combined features have been proposed [21]. A crucial step towards a new generation of machine learning approaches is enabling computers to learn the features as data representatives. These are expressed as low-level features such as margin and edge; middle-level features such as edge junctions and high-level object parts [22]. Texture is characterized by a set of local statistical properties of pixel intensities. In mammographic image processing, these features have been used to distinguish patterns that indicate different levels of risk to develop lesions. Texture has shown to be a promising technique in analyzing mammographic lesions caused by masses [23].

Textural information is important to outline the performance of CAD system, is required for the classification that distinguishes masses from normal tissues [24]. Researchers apply traditional feature engineering methods to deal with the breast masses classification task, which generally involves ROIs segmentation, feature extraction, and classification. Oliver *et al.* extracted morphological and texture features from breast tissue regions which were segmented using a fuzzy C-means clustering technique, and these features were then treated as inputs for the classifier [7]. Chen *et al.* evaluated different local features using texture representation algorithms. After that, they modeled mammographic tissue patterns based on the local tissue appearances in mammograms [25].

On the other hand, the fast development of the Deep Learning (DL) field offers a promising CAD method for medical image analysis [26]–[29]. The challenge is to define a classification algorithm that can provide the most appropriate response to the problem, i.e., statistical, artificial intelligence, support vector machines, or polynomial methods commonly consider the prediction of the breast cancer pattern [30]. DL approaches termed one of the significant breakthrough technologies of recent years by the MIT Technology Review has made headlines in producing semantic information due to its nature of adaptive learning from input data. Convolutional Neural Networks (CNNs) 8 [31] have become the technique of choice for supervised approaches. In recent years, a noticeable shift from conventional machine learning methods to DL based methods is seen in a wide variety of real-world, especially medical, applications and several review papers have been published [26], [31], [32]. DL methods have multi-levels of representation learning which use raw data and discover the essential representations for detection or

classification [31]. At the same time as the DL concepts were developed, a step-change in processing power through high-performance GPUs and open source frameworks/libraries developed on CUDA or OpenCL platforms have made significant progress for the implementation of DL based methods. These open-source frameworks and libraries provide the chance to optimize the implementation of convolutions and other related functions. Also, they facilitate the ability to perform a high number of computations at a relatively low cost through their massive parallel architectures.

CNNs are one type of these deep networks that have already shown excellent performance in image classification [33], detection, and segmentation. CNNs can learn highly nonlinear relationships between the inputs and outputs without human intervention and was used to classify masses using texture features extracted from mammography based descriptors of image crops of mass area. These textural features could be interpreted as inputs to a classifier. The classifier is a tool that receives the extracted feature values as input and provides the classes related to available tissue as output. However, depending only on texture feature is not sufficient to classify the breast cancer masses from mammograms. Thus, some studies attempt to use morphological information of tumor shape in classifying breast cancer masses.

In this paper, we propose a Deep Convolution Neural Network (DCNN) based method for automatic extraction of the scoring feature of three kinds of breast masses. Subsequently, multi-features of Gray Level Co-occurrence Matrix (GLCM) and Histogram of Oriented Gradient (HOG) [34], [35] are extracted to focus the texture points of ROI. Subsequently, a set of texture features and scoring features corresponding to each breast image are combined into multi-features. Then, the extracted features are introduced into different classifiers and classified into the desired classes. For classifier, we use Support Vector Machine (SVM) [36] or eXtreme Gradient Boosting (XGBoost) [37], [38] classification algorithm to define the ROI as normal, benign or malignant. Therefore, the proposed system consists of three main stages: scoring feature extraction, texture extraction, and lesion classification. The classification accuracy (ACC), Precision (Pre), Recall (Rec), F_1-score (F_1), and Overall Accuracy (Overall Acc) is used to evaluate the performance of the proposed system.

II. MATERIAL AND METHODS

A. DATASETS

The Digital Database for Screening Mammography (DDSM) is the largest public mammography database and is a joint effort of professional researchers from the Massachusetts General Hospital (D. Kopans, R. Moore), the University of South Florida (K. Bowyer), and the Sandia National Laboratories (P. Kegelmeyer). The DDSM database has been widely used as a benchmark for numerous research on the field of mammography, for being free of charge and having a large number of and diverse quantity of cases. It has 2604 cases, and each case consists of four views which contained a mixture of Mediolateral oblique (MLO) and Craniocaudal (CC) view from the left and right breast (i.e., LEFT-CC, LEFT-MLO, RIGHT-CC, and RIGHT-MLO) [39], [40]. Each mammographic images have its corresponding technical and clinical information, including diverse shapes, margins, sizes, breast densities of the masses, which annotations labeled by experienced radiologists, as well as exam dates, digitalization equipment, lesion types (according to BI-RADS) [41], [42], and existent pathologies, patients' ages and races of mammography.

TABLE 1. Experimental data.

Samples	Benign	Cancer	Normal	**Total**
Train	677	764	5530	**6971**
Validation	225	254	1844	**2323**
Test	217	248	1803	**2268**
Total	**1119**	**1266**	**9177**	**11562**

The experimental dataset used in this study encompasses the ROIs of mammographic screen digitized images retrieved from DDSM by Jiang *et al.* [42], Heath *et al.* [43]. Because the significant part of the whole mammographic image comprises the pectoral muscle and the background with a lot of artifacts, the classification decomposition was performed on a limited ROIs that contains the prospective abnormality. The ROIs are just the rectangular area around the lesion. The size of the ROIs is chosen to ensure that the ROIs covers the entire lesion without including too much normal tissue surrounding the lesion. Therefore, mammograms were first cropped to remove the parts that affect classification. Image cropping was performed according to the comments of professional radiologists provided in the DDSM dataset.

To simulate practical scenarios, a series of ROIs [42] depicting benign, cancer, and normal masses are extracted following the conventions in [44]–[46]. To make our experimental setup more consistent with practice and more challenging, we divide these data into training set, verification set and test set by 60%, 20% and 20% (as shown in Table 1), which is composed of 6971 images, 2323 images, and 2268 images, respectively. In these datasets, the three classes are realistic, which contains 1119 benign ROIs, 1266 cancer ROIs, and 9177 normal ROIs. 677 benign ROIs, 764 cancer ROIs, and 5530 normal ROIs are randomly selected as the training dataset; 225 benign ROIs, 254 cancer ROIs, and 1844 normal ROIs are selected as the validation dataset; the remaining 217 benign ROIs, 248 cancer ROIs, and 1803 normal ROIs is used as the test dataset, 11562 ROIs in total, form a large database. The database ROIs are selected from different cases to avoid positive bias.

B. DEEP CONVOLUTIONAL NEURAL NETWORK (DCNN)

GoogLeNet [47] is the first implementation using the Inception module. The main idea of this module is based on the authors' finding out how an optimal local sparse structure in

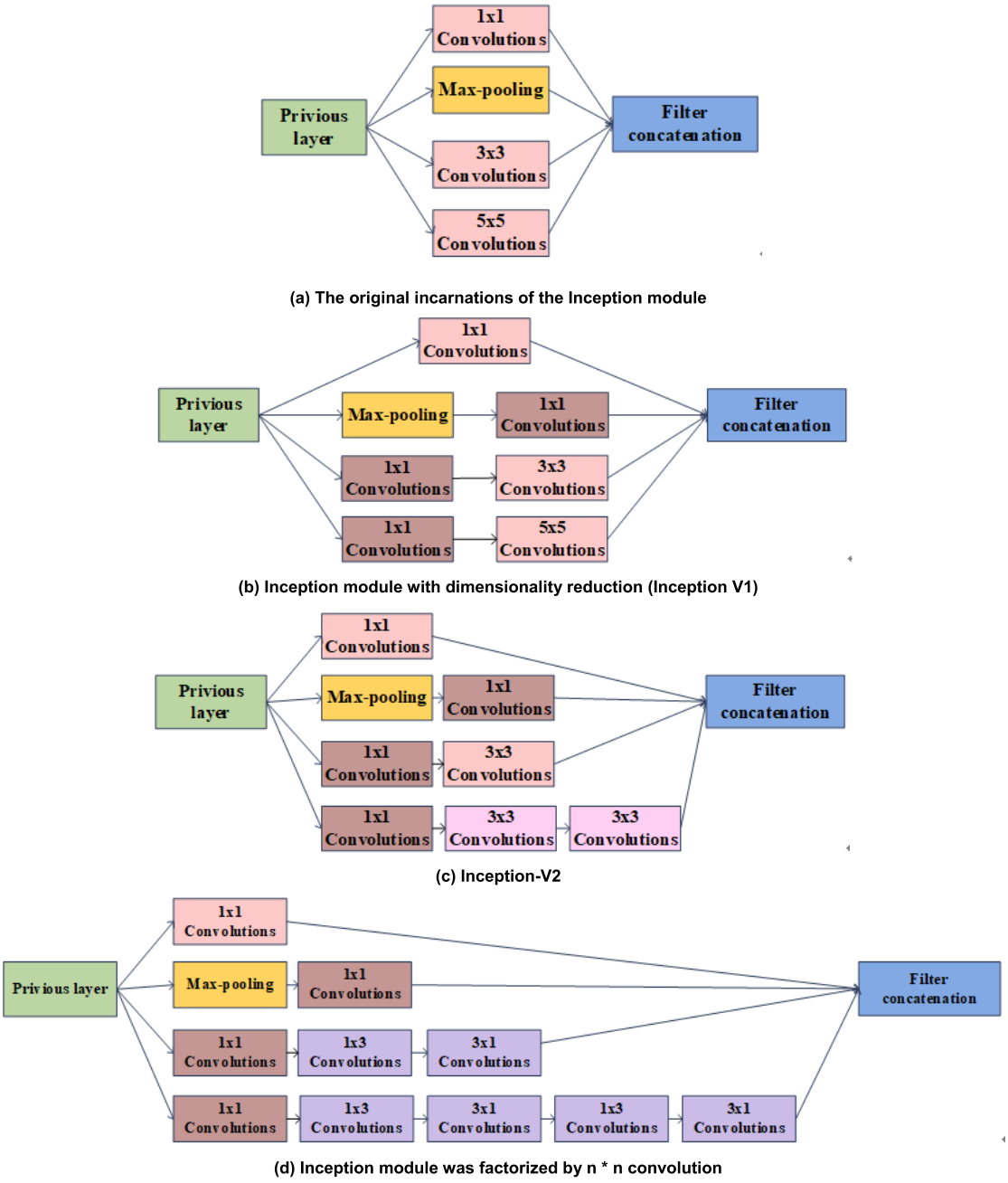


FIGURE 1. Evolution of inception module.

a convolutional network can be approximated and covered by dense components [48]. They aimed to find the optimal local structure and repeat it, constructing a multi-layer network, which assuming translation invariance means that our network will be built from convolutional building blocks.

According to Arora *et al.* [49], the author assumes that each unit from the earlier layer corresponds to some region of the input image, and these units are grouped into filter banks. In the lower layers (the ones close to the input) correlated units would concentrate in local regions, which means they can be covered by a layer of 1*1 convolutions in the next layer, as suggested in [50]. However, one can also expect that there will be a smaller number of more spatially spread out clusters that can be covered by convolutions over larger patches, and there will be a decreasing number of patches over larger and larger regions. To avoid patch-alignment issues, the original incarnations of the Inception architecture are restricted to filter sizes 1∗1, 3*3 and 5*5, but this decision was based more on convenience rather than necessity. Besides, since pooling operations are essential for the success in the convolutional networks, it suggests that adding an alternative parallel pooling path in each such stage should also have additional beneficial effect (see Figure 1(a) the original incarnations of the Inception module).

A big problem of the above modules is that even a modest amount of 5*5 convolutions can be prohibitively expensive on top of a convolutional layer with a large number of filters. Once pool units are added to the mix, the problem becomes more prominent. Although this architecture might cover the optimal sparse structure, it would do it very inefficiently, resulting in a computational blow up within a few stages. The author applies the dimension reductions and projections where the amount of computational cost would increases. Before expensive 3*3 and 5*5 convolutions, 1*1 convolutions are used to reduce computation [47]. As shown in Figure 1(b), the Inception module consists of four branches that get the same input. The first branch uses 1*1 convolution to filter the input, and this convolution plays as a linear transformation on the input channels. The second and third branches perform 1*1 kernels convolutions for dimensionality reduction performed by 3*3 and 5*5 kernels convolution layers, respectively. The fourth branch performs max-pooling followed by convolution with 1*1 kernels. Finally, the outputs of each branch are concatenated and fed as input to the next block [51].

In [52], a revised version of the initial module along with a slightly modified network architecture have been proposed. Batch normalization (BN) [52] was proposed by authors and incorporated it into the Inception network. BN is a technology that takes normalization part of the model architecture and performs normalization for each training mini-batch. In the author's opinion, BN has higher learning rates and simpler initialization techniques without experiencing adverse effects. The network used in [52], namely Inception-V2, was a slight modification of GoogLeNet. In terms of computation, convolutions with larger spatial filters tend to be disproportionally expensive. Apart from the incorporation of BN, the most important change is that the 5*5 convolutional layers of the Inception module were replaced by two consecutive 3*3 layers (Figure 1(c)). The author constructs a vision network by using translation invariance and replace the fully connected components by a two-layer convolutional architecture: the first layer is a 3*3 convolution, and the second layer is the fully connected layer above the 3*3 output grid of the first layer. Sliding this small network over the input activation grid boils down to replacing the 5*5 convolution with two layers of 3*3 convolution (compare Figure 1(b) with (c)).

Whether one should factorize larger filters into smaller, for example, 2*2 convolutions. However, by using asymmetric convolutions n*1, it can even do better than 2*2. For example, using a 3*1 convolution followed by a 1*3 convolution is equivalent to sliding a two-layer network with the same receptive field as in a 3*3 convolution. Still, the two-layer solution is 33% cheaper for the same number of output filters, if the number of input and output filters is equal. By comparison, factorizing a 3*3 convolution into a two 2*2 convolution represents only 11% saving of computation. It is further demonstrated that a 1*n convolution followed by an n*1 convolution can be used to replace any n*n convolution and, as n grows, the computational cost savings increase significantly (see Figure 1(d)). Inception-v3 [53] also uses the auxiliary classifier as the regularizer based on Inception-V2. If the side branch is batch normalized [52] or has a dropout layer, the performance of the main classifier of the network performs better will be better.

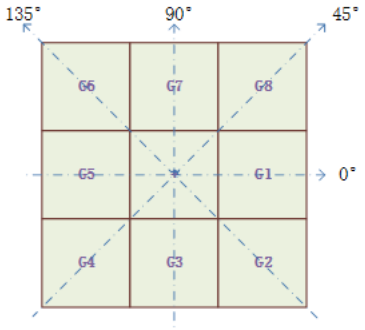


FIGURE 2. Eight nearest neighbor resolution cells in four different directions.

C. TEXTURE FEATURE

1) GREY-LEVEL CO-OCCURRENCE MATRIX (GLCM)

One of the important characteristics is the texture feature used in identifying objects or ROI in an image. Because of the difference and diversity of messes, the shape of breast masses in mammography is different. The edge of the mass can reflect the type and degree of the cancer. In the large mass area, the edge is long, and the more irregular and different protrusions become a texture feature of a certain type of mass. In this paper, some easily computable textural features based on the gray tone spatial dependence are used in the classification of mammography image data [54]. GLCM describes the texture of an image by calculating the frequency of pixel pairs with specific values and spatial relationships. Suppose that the image to extract features is rectangular, with n resolution cells in the horizontal direction and m resolution cells in the vertical direction.

A resolution cell has eight nearest neighbor resolution cells in four different directions, as shown in Figure 2. The texture context information in the image I is adequately specified by the matrix of relative frequencies $P_{i,j}$, in which two neighboring resolution cells are separated by distance d on the image, one is gray tone i and the other is gray tone j. The gray tone spatial dependence frequencies matrices are a function of the angular relationship between the neighboring resolution cells and a function of the distance between them ($P(i, j, d, 0)$; $P(i, j, d, 45)$; $P(i, j, d, 90)$; $P(i, j, d, 135)$) [54].

We extract the eighty of the GLCM features from each image. These features are the mean and variance of four different angles of contrast, correlation, entropy and inverse different moment (homogeneity), which are measured at ten different distances. These various features are all functions of distance and angle. It is assumed that image I has features A, B, C and D of 0°, 450°, 900°, and 135° angles, respectively (as shown in Figure 2). We take the functions of a, b,

c, and d, their average and variance, be used as inputs to the classifier.

Contrast: The contrast feature is the contrast or local variation in an image. It shows the clarity of the image and the depth of the texture groove.

$$f_{Con} = \sum_i \sum_j (i-j)^2 P(i,j) \tag{1}$$

Correlation: The correlation feature is the gray tone linear dependencies in the image.

$$f_{Corr} = \frac{\sum_i \sum_j (ij) P(i,j) - \mu_i \mu_j}{\sigma_i \sigma_j} \tag{2}$$

where μ_i, μ_j, σ_i and σ_j are the means and standard deviations of P(i) and P(j).

Entropy: Entropy in physics is the regularity of objects. The entropy also is the amount of information in the image. It shows the degree of complexity in the texture of the image.

$$f_{Ent} = -\sum_i \sum_j P(i,j) \log P(i,j) \tag{3}$$

Homogeneity: Homogeneity is the inverse distance difference of image texture, which measures the local change of image texture.

$$f_{IDE} = \sum_i \sum_j \frac{P(i,j)}{1+(i-j)^2} \tag{4}$$

2) HISTOGRAM OF ORIENTED GRADIENT (HOG)

As a feature descriptor, HOG has the force to describe the structure of the object and has a strong identification effect on the description of the local area. It is very sensitive to the gradient and direction and can describe the appearance edge and structure characteristics of the mass quickly and accurately. HOG has different types of spatial organization, gradient calculation and normalization methods. By calculating the horizontal and vertical gradients of the image, the image is divided into equal cell units, and the histogram within the unit is counted to better express the relationship between pixels. The dense representation of the image at a specific resolution is defined according to the structure in [55]. First, the image is divided into 16*16 nonoverlapping pixel regions or cells. For each cell, we accumulate a one-dimensional histogram of the gradient directions on the pixel. The gradient at each pixel is dispersed to one of the nine orientation bins. Each pixel ''votes'' the orientation of its gradient, and its strength depends on the gradient magnitude at that pixel. Finally, the histogram of each cell is normalized to the gradient energy of the neighborhood around it. We use the four 2*2 block cells with specific cells, and normalize the histogram of a given cell with the total energy in each of these blocks. This produces a 9*4 dimensional vector that represents the local gradient information in the cell.

The gradient direction value of each pixel position is calculated according to the gradient of image abscissa and ordinate direction. The derivative can not only obtain the contour and some texture information but also further weaken the influence of brightness.

The gradient of pixels (i, j) in an image is defined as:

$$G_i(i,j) = H(i+1,j) - H(i-1,j) \tag{5}$$

$$G_j(i,j) = H(i,j+1) - H(i,j-1) \tag{6}$$

where the horizontal gradient, vertical gradient and pixel value of the pixel (i, j) in the input image is $G\_i(i, j)$, $G\_j(i, j)$ and $H(i, j)$, respectively. The gradient and gradient direction at pixels (i, j) is defined as:

$$G(i,j) = \sqrt{G_i(i,j)^2 + G_j(i,j)^2} \tag{7}$$

$$G(i,j) = tan^{-1}\left(\frac{G_i(i,j)}{G_j(i,j)}\right) \tag{8}$$

III. EXPERIMENTAL MODEL

Four strategies were conducted to assess the abilities of ROIs of the DDSM in mass classification within the DCNN framework equipped with or without transfer learning, as well as to explore an eligible combination strategy of DCNN with the SVM and XGboost framework in enhancing classification performance. This mass classification task was to categorize a target into cancer, benign or normal.

Strategy 1: The DCNN can be either equipped with or without transfer learning, i.e., the network was first initialized with the trained parameters from ImageNet and then trained with the ROIs datasets (i.e., transfer learning), or directly trained by the ROIs datasets (i.e., no transfer learning), and their classification performance were examined (termed as 'DCNN-ROIs-TL'', and 'DCNN-ROIs' as shown in Figure 3(a)). After an image is input into the DCNN network, through a series of convolution and pooling operations, the final prediction classification layer at the end converts the information output from the full connection layer into the corresponding classification score, which plays a classification role. This experiment was to compare the classification capabilities of with or without transfer learning, as well as to confirm the role of transfer learning in DCNN training.

Strategy 2: the strategy of combining DCNN and SVM or XGboost training on ROIs data set is discussed. The classification score of the DCNN network is taken as the feature vector (i.e., Feature-Class Score in Figure 3(b)), which is input into SVM or XGboost to train the model. Based on two DCNN models, two combined classification models will be obtained, including two models using two DCNN training modes (Figure 3(b), Score-Classifier model based on transfer learning or non-transfer learning).

Strategy 3: Texture is one of the important characteristics used in identifying objects or regions of interest in an image. This strategy describes some easily computable textural features based on Gray-level Co-occurrence Matrix (GLCM) and Histograms of Oriented Gradient (HOG) (as shown in Figure 3(c), Feature-GLCM and Feature-HOG). The classification ability of image texture features is also evaluated on SVM or XGboost. These two image texture features are extracted from ROIs for training (Figure 3(c), Feature-Classifier).

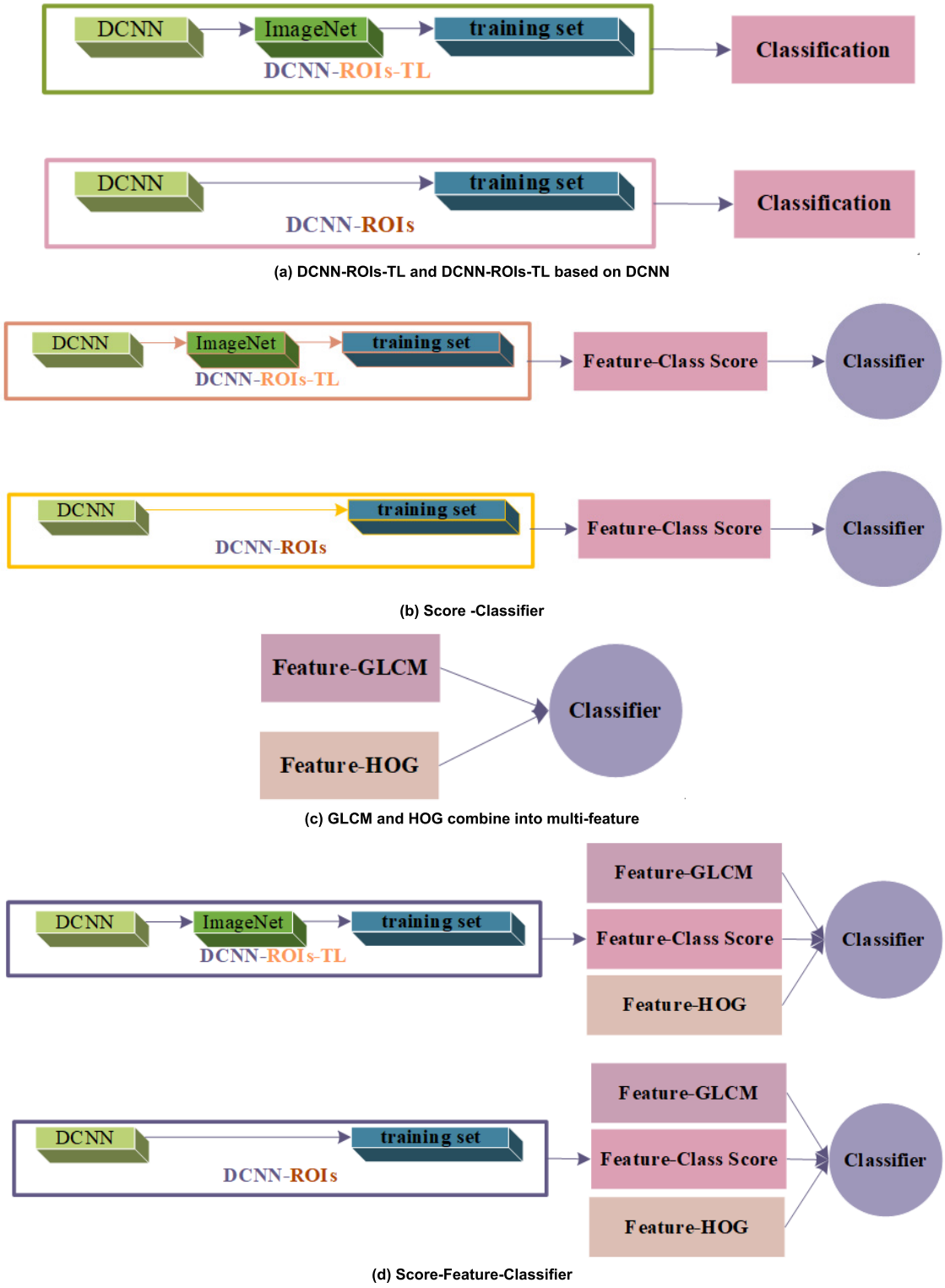
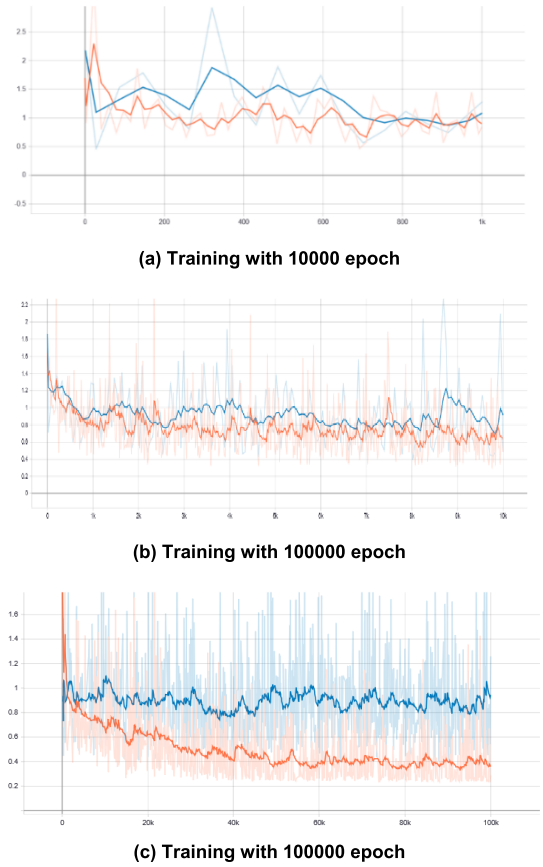


FIGURE 3. The model configurations of DCNN combination with multi-features for strategy 1-4.

Strategy 4: This strategy combines strategy 2 and strategy 3 to integrate the classification score of DCNN network with the GLCM and HOG texture features of images, and input the integrated feature vectors into SVM or XGboost for model training (Figure 3(d), Score-Feature-Classifier).

IV. EVALUATION METRICS

The DCNN models outputted the prediction score of each of the three classes (cancer, benign, and normal) from the softmax layer. Currently, most measures to evaluate the performance of classification algorithms focus on the ability of

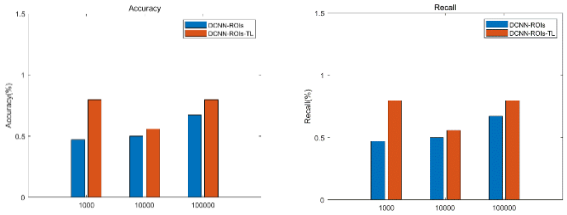


(a) Training with 10000 epoch

(b) Training with 100000 epoch

(c) Training with 100000 epoch

FIGURE 4. Comparison of classification loss of DCNN-ROIs-TL and DCNN-ROIs-TL.



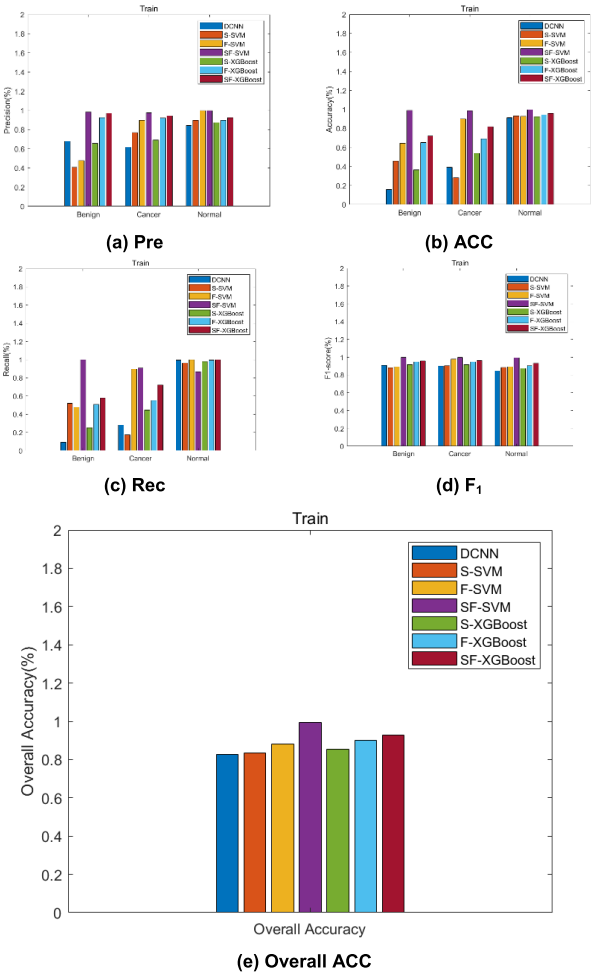
(a) ACC (b) Rec

FIGURE 5. Comparison of ACC and rec of these two models.

the classifier to correctly identify the class [56]. In a two-class problem, accuracy is a commonly used metric [57].

However, accuracy is particularly prone to bias in multi-class classifications that are sensitive to training data. Thus, to evaluate the performance of the classification models, the classification ACC, Pre, Rec, and F_1 statistical measures are used. For each class, the confusion matrix was generated which reported the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN). The classification ACC, Pre, Rec, and F_1 of each class were computed from the confusion matrix according to the following statistical formulas:

$$Acc = \frac{\sum_{i=1}^{K} tp_i}{\sum_{i=1}^{K} tp_i + fp_i} \tag{9}$$



(a) Pre (b) ACC

(c) Rec (d) F_1

(e) Overall ACC

FIGURE 6. Comparison performance of these all models in the training dataset.

$$Pre = \frac{1}{K} \sum_{i=1}^{K} \frac{tp_i}{tp_i + fp_i} \tag{10}$$

$$Rec = \frac{1}{K} \sum_{i=1}^{K} \frac{tp_i}{tp_i + fn_i} \tag{11}$$

$$F_1 = \frac{2Pre * Rec}{Pre + Rec} \tag{12}$$

where K is the number of classes, and tp_i represents the number of data objects correctly grouped into the i class, fp_i represents the number of objects that do not belong to the i class but have been partitioned into the i class, and fn_i represents the number of objects that belong to the i class but have not been partitioned into the i class.

V. RESULTS AND DISCUSSION

A. IMPACT OF TRANSFER LEARNING AND NO TRANSFER LEARNING

As shown in Figure 4 and Figure 5, DCNN-ROIs-TL significantly outperformed DCNN-ROIs for ROIs in terms of classification loss, ACC and Rec. Especially when the epoch was 100000, i.e., Figure 4 (c), DCNN-ROIs-TL was significantly better than DCNN ROIs in terms of the classification

TABLE 2. Comparison of different feature combination strategies in three categories.

Category	Model	Pre (%)	Rec (%)	ACC (%)	F_1 (%)
Caring	S-SVM	40.81	52.14	45.78	88.01
	F-SVM	47.68	99.99	64.57	89.34
	SF-SVM	98.26	99.99	99.12	99.83
	S-XGBoost	65.61	25.23	36.44	91.69
	F-XGBoost	92.24	50.61	65.36	94.94
	SF-XGBoost	96.93	57.60	72.26	95.83
Cancer	S-SVM	76.88	17.41	28.39	90.37
	F-SVM	89.77	90.71	90.23	97.85
	SF-SVM	97.57	99.74	98.64	99.70
	S-XGBoost	68.90	44.43	54.02	91.72
	F-XGBoost	92.12	55.18	69.02	94.58
	SF-XGBoost	94.20	72.35	81.84	96.49
Normal	S-SVM	89.85	96.40	93.01	88.51
	F-SVM	99.99	86.42	92.72	89.23
	SF-SVM	99.99	99.48	99.74	99.58
	S-XGBoost	87.39	98.04	92.41	87.18
	F-XGBoost	89.81	99.57	94.44	90.66
	SF-XGBoost	92.39	99.78	95.95	93.29

loss. Furthermore, DCNN-ROIs-TL achieved significantly higher classification ACC and Rec than DCNN-ROIs. As shown in Figure 5 (a) and (b), considering the stability and rationality, DCNN-ROIs-TL with the epoch of 100000 has good performance.

B. MASSES CLASSIFICATION WITH DIFFERENT FEATURE COMBINATION STRATEGIES

As shown in Figure 6, the Score-Feature-Classifier (i.e., SF-SVM and SF-XGBoost) models achieved better classification performances (especially in terms of Pre, ACC, and Rec) for the benign and cancer. However, no statistical differences were observed in F_1. Furthermore, it should be noted that there was an overfitting on the SF-SVM models, **which achieved abnormally high classification performances** (in terms of all metrics) for all classes.

Similar observations were made by using SVM and XGBoost as classifiers, evidenced by the superior performance (in terms of Pre, ACC, and Rec) of the XGBoost model in the classification of the benign and cancer, although this superiority was not evident for the normal. Again, the SF-XGBoost was comparable to the DCNN in terms of F_1. Moreover, for all models, using the score feature and texture feature as the multi-features (i.e., Score-Feature-Classifier) achieved significantly better performance than only using the score feature or texture feature (i.e., Score-Classifier or

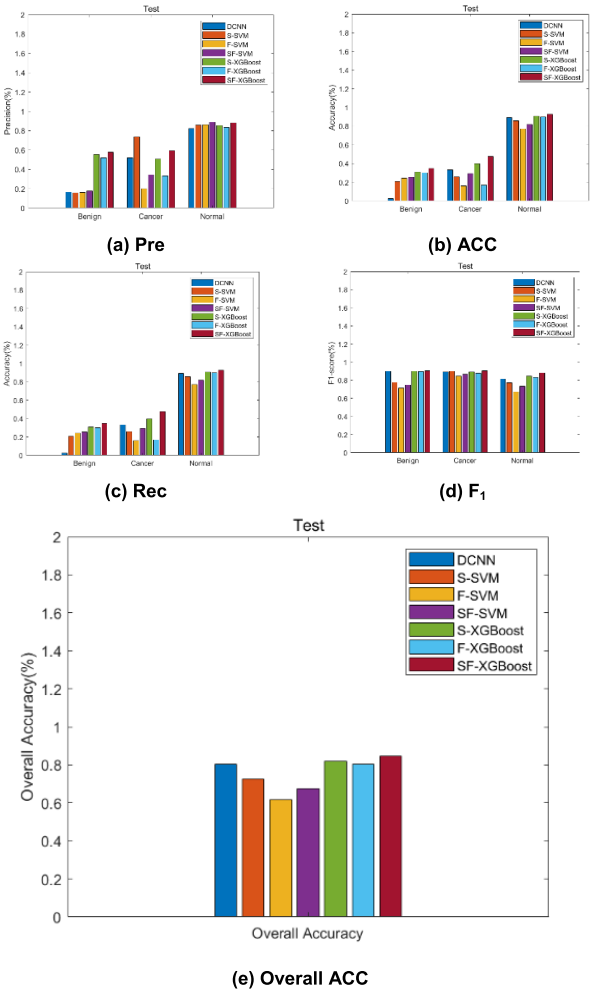


FIGURE 7. Comparison performance of these all models in the testing dataset.

Feature-Classifier) in the classification of the benign and cancer, as listed in Table 2, but for normal it was not significantly superiority.

C. EXPERIMENT ON THE TEST DATASETS

As shown in Figure 7, When experimenting on test datasets, the XGBoost with multiple features (i.e., SF-XGBoost) models achieved better classification performances in terms of Pre, ACC, and Rec for the benign and cancer. The performance was not superior outstanding in distinguishing the normal for SF-XGBoost. Similar observations were evidenced by the superior performance of the SF-XGBoost model in the classification of the masses, although this superiority was not evident in comparison with DCNN. Figure 7(e) shows that our multi-feature (i.e., SF-XGBoost) model achieved the higher Overall ACC among other strategies of models. In Table 3 the experimental result shows that the Overall ACC of the SF-XGBoost model, owing to its ability to integrate multi-feature information, outperforms all other approaches, and achieves improvements of at least 4.1% for the CDNN.

TABLE 3. Comparison of Overall ACC of DCNN model and multi-features with XGBoost models.

Model	Train (%)	Test (%)
DCNN	82.84	80.38
S-XGBoost	85.3	81.92
F-XGBoost	90.09	80.38
SF-XGBoost	92.8	84.48

VI. CONCLUSION

To elevate the classification performance of the networks, in this study, we proposed a method to classify breast masses into benign, cancer and normal in mammography by using multi-feature and combine the classification results based on DCNN as features. The multi-feature should be selected if the amount of data is large and has multiple features. Meanwhile, the machine learning methods can be considered if the amount of data is small. This study showed that the multi-feature model generally outperformed the single feature model or only DCNN model when based on DCNN equipped with transfer learning and using the XGBoost model generally achieved higher ACC than that others. As a result, multi-features are extracted from mammographic and are then input XGBoost framework can perform better than conventional deep learning networks in problems of object classification. This also proves that the XGBoost classifier is more effective than deep learning networks when dealing with the problem of a limited number of available training samples and texture features. The key advantages of the proposed method are that it employs an integration of DCNN trained to extract classification score features from mammography images and integrate them with other texture features, and then uses XGBoost framework to achieve a better classification performance despite the limited number of breast cancer samples and imbalanced training data, which are the challenging problems. In the training model and testing model, the improvement rate of Overall ACC is 9.96% and 4.10% respectively. Besides, we plan to use a method that can classify all masses in mammography, and then consider the relevant features that can more fully express the mass features in mammography based on the existing texture features GLCM and HOT, to improve the performance of breast mass classification.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2017," *CA, Cancer J. Clinicians*, vol. 67, no. 1, pp. 7–30, Jan. 2017.
[2] E. A. Sickles, "Breast cancer screening outcomes in women ages 40-49: Clinical experience with service screening using modern mammography," *JNCI Monographs*, vol. 1997, no. 22, pp. 99–104, Jan. 1997.
[3] J. S. Whang, S. R. Baker, R. Patel, L. Luk, and A. Castro, "The causes of medical malpractice suits against radiologists in the United States," *Radiology*, vol. 266, no. 2, pp. 548–554, Feb. 2013.
[4] M. Elter and A. Horsch, "CADx of mammographic masses and clustered microcalcifications: A review," *Med. Phys.*, vol. 36, pp. 2052–2068, Jun. 2009.
[5] R. M. Rangayyan, S. Banik, and J. E. L. Desautels, "Computer-aided detection of architectural distortion in prior mammograms of interval cancer," *J. Digit. Imag.*, vol. 23, no. 5, pp. 611–631, Oct. 2010.
[6] L. T. Niklason, D. B. Kopans, and L. M. Hamberg, "Digital breast imaging: Tomosynthesis and digital subtraction mammography," *Breast Disease*, vol. 10, nos. 3–4, pp. 151–164, Aug. 1998.
[7] A. Oliver, J. Freixenet, R. Marti, J. Pont, E. Perez, E. R. E. Denton, and R. Zwiggelaar, "A novel breast tissue density classification methodology," *IEEE Trans. Inf. Technol. Biomed.*, vol. 12, no. 1, pp. 55–65, Jan. 2008.
[8] A. Oliver, M. Tortajada, X. Lladó, J. Freixenet, S. Ganau, L. Tortajada, M. Vilagran, M. Sentís, and R. Martí, "Breast density analysis using an automatic density segmentation algorithm," *J. Digit. Imag.*, vol. 28, no. 5, pp. 604–612, Oct. 2015.
[9] A. Rampun, B. Scotney, P. Morrow, H. Wang, and J. Winder, "Breast density classification using local quinary patterns with various neighbourhood topologies," *J. Imag.*, vol. 4, no. 1, p. 14, 2018.
[10] L. J. Warren, "Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy," *Breast Diseases: A Year Book Quart.*, vol. 21, no. 4, pp. 330–332, Jan. 2010.
[11] B.-W. Hong and B.-S. Sohn, "Segmentation of regions of interest in mammograms in a topographic approach," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 1, pp. 129–139, Jan. 2010.
[12] 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.
[13] J. J. Fenton, J. Egger, P. A. Carney, G. Cutter, C. D'Orsi, E. A. Sickles, J. Fosse, L. Abraham, S. H. Taplin, W. Barlow, R. E. Hendrick, and J. G. Elmore, "Reality check: Perceived versus actual performance of community mammographers," *Amer. J. Roentgenology*, vol. 187, no. 1, pp. 42–46, Jul. 2006.
[14] C. W. Huo, G. L. Chew, K. L. Britt, W. V. Ingman, M. A. Henderson, J. L. Hopper, and E. W. Thompson, "Mammographic density—A review on the current understanding of its association with breast cancer," *Breast Cancer Res. Treatment*, vol. 144, no. 3, pp. 479–502, Apr. 2014.
[15] W. A. Berg, C. Campassi, P. Langenberg, and M. J. Sexton, "Breast imaging reporting and data system: Inter- and intraobserver variability in feature analysis and final assessment," *Amer. J. Roentgenol.*, vol. 174, pp. 1769–1777, Jun. 2000.
[16] F. Winsberg, M. Elkin, J. Macy, V. Bordaz, and W. Weymouth, "Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis," *Radiology*, vol. 89, no. 2, pp. 211–215, Aug. 1967.
[17] W. He, A. Juette, E. R. E. Denton, A. Oliver, R. Martí, and R. Zwiggelaar, "A review on automatic mammographic density and parenchymal segmentation," *Int. J. Breast Cancer*, vol. 2015, Jun. 2015, Art. no. 276217.
[18] A. Oliver, J. Freixenet, J. Martí, E. Pérez, J. Pont, E. R. E. Denton, and R. Zwiggelaar, "A review of automatic mass detection and segmentation in mammographic images," *Med. Image Anal.*, vol. 14, no. 2, pp. 87–110, Apr. 2010.
[19] M. L. Giger, N. Karssemeijer, and J. A. Schnabel, "Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer," *Annu. Rev. Biomed. Eng.*, vol. 15, no. 1, pp. 327–357, Jul. 2013.
[20] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.
[21] Z. Jiao, X. Gao, Y. Wang, and J. Li, "A parasitic metric learning net for breast mass classification based on mammography," *Pattern Recognit.*, vol. 75, pp. 292–301, Mar. 2018.
[22] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2018–2025.
[23] S. Gupta and M. K. Markey, "Correspondence in texture features between two mammographic views," *Med. Phys.*, vol. 32, pp. 1598–1606, May 2005.
[24] W. Qian, X. Sun, D. Song, and R. A. Clark, "Digital mammography: Wavelet transform and Kalman-filtering neural network in mass segmentation and detection," *Academic Radiol.*, vol. 8, pp. 1074–1082, Nov. 2001.
[25] Z. Chen, E. Denton, and R. Zwiggelaar, "Local feature based mammographic tissue pattern modelling and breast density classification," in *Proc. 4th Int. Conf. Biomed. Eng. Informat. (BMEI)*, Oct. 2011, pp. 351–355.

[26] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[27] D. Shen, G. Wu, and H. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jun. 2017.

[28] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[29] L. Zhang, L. Lu, R. M. Summers, E. Kebebew, and J. Yao, "Convolutional invasion and expansion networks for tumor growth prediction," *IEEE Trans. Med. Imag.*, vol. 37, no. 2, pp. 638–648, Feb. 2018.

[30] J. F. Ramirez-Villegas and D. F. Ramirez-Moreno, "Wavelet packet energy, tsallis entropy and statistical parameterization for support vector-based and neural-based classification of mammographic regions," *Neurocomputing*, vol. 77, no. 1, pp. 82–100, Feb. 2012.

[31] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[32] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[33] S. A. Agnes, J. Anitha, S. I. A. Pandian, and J. D. Peter, "Classification of mammogram images using multiscale all convolutional neural network (MA-CNN)," *J. Med. Syst.*, vol. 44, no. 1, p. 30, Jan. 2020.

[34] A. Iqbal, N. A. Valous, F. Mendoza, D.-W. Sun, and P. Allen, "Classification of pre-sliced pork and Turkey ham qualities based on image colour and textural features and their relationships with consumer responses," *Meat Sci.*, vol. 84, no. 3, pp. 455–465, Mar. 2010.

[35] J. Erazo-Aux, H. Loaiza-Correa, and A. D. Restrepo-Giron, "Histograms of oriented gradients for automatic detection of defective regions in thermograms," *Appl. Opt.*, vol. 58, no. 13, p. 3620, May 2019.

[36] T. Li, R. Song, Q. Yin, M. Gao, and Y. Chen, "Identification of S-nitrosylation sites based on multiple features combination," *Sci. Rep.*, vol. 9, no. 1, p. 3098, Dec. 2019.

[37] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2016, pp. 785–794.

[38] P. Montero-Manso, G. Athanasopoulos, R. J. Hyndman, and T. S. Talagala, "FFORMA: Feature-based forecast model averaging," *Int. J. Forecasting*, vol. 36, no. 1, pp. 86–92, Jan. 2020.

[39] M. Z. D. Nascimento, A. S. Martins, L. A. Neves, R. P. Ramos, E. L. Flores, and G. A. Carrijo, "Classification of masses in mammographic image using wavelet domain features and polynomial classifier," *Expert Syst. Appl.*, vol. 40, no. 15, pp. 6213–6221, Nov. 2013.

[40] M. Jiang, S. Zhang, Y. Zheng, and D. N. Metaxas, "Mammographic mass segmentation with online learned shape and appearance priors," *Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, pp. 35–43, 2016.

[41] M. Kallenberg, K. Petersen, M. Nielsen, A. Y. Ng, P. Diao, C. Igel, C. M. Vachon, K. Holland, R. R. Winkel, N. Karssemeijer, and M. Lillholm, "Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1322–1331, May 2016.

[42] M. Jiang, S. Zhang, H. Li, and D. N. Metaxas, "Computer-aided diagnosis of mammographic masses using scalable image retrieval," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 2, pp. 783–792, Feb. 2015.

[43] M. Heath, K. Bowyer, D. Kopans, P. Kegelmeyer, R. Moore, K. Chang, and S. Munishkumaran, "Current status of the digital database for screening mammography," in *Digital Mammography* (Computational Imaging and Vision), vol. 13. Dordrecht, The Netherlands: Springer, 1998, pp. 457–460.

[44] G. D. Tourassi, B. Harrawood, S. Singh, J. Y. Lo, and C. E. Floyd, "Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms," *Med. Phys.*, vol. 34, no. 1, pp. 140–150, Dec. 2006.

[45] B. Zheng, A. Lu, L. A. Hardesty, J. H. Sumkin, C. M. Hakim, M. A. Ganott, and D. Gur, "A method to improve visual similarity of breast masses for an interactive computer-aided diagnosis environment," *Med. Phys.*, vol. 33, no. 1, pp. 111–117, Jan. 2006.

[46] S. C. Park, R. Sukthankar, L. Mummert, M. Satyanarayanan, and B. Zheng, "Optimization of reference library used in content-based medical image retrieval scheme," *Med. Phys.*, vol. 34, no. 11, pp. 4331–4339, Oct. 2007.

[47] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[48] R. K. Samala, H.-P. Chan, L. M. Hadjiiski, M. A. Helvie, K. H. Cha, and C. D. Richter, "Multi-task transfer learning deep convolutional neural network: Application to computer-aided diagnosis of breast cancer on mammograms," *Phys. Med. Biol.*, vol. 62, no. 23, pp. 8894–8908, Nov. 2017.

[49] S. Arora, A. Bhaskara, R. Ge, and T. Ma, "Provable bounds for learning some deep representations," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 584–592.

[50] Y. Zhao, M.-C.-A. Lin, M. Farajzadeh, and N. L. Wayne, "Early development of the gonadotropin-releasing hormone neuronal network in transgenic zebrafish," *Frontiers Endocrinol.*, vol. 4, p. 107, Aug. 2013.

[51] L. Tsochatzidis, L. Costaridou, and I. Pratikakis, "Deep learning for breast cancer diagnosis from mammograms—A comparative study," *J. Imag.*, vol. 5, no. 3, p. 37, 2019.

[52] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167

[53] Y. N. Fedorov and A. N. W. Hone, "Sigma-function solution to the general Somos-6 recurrence via hyperelliptic prym varieties," *J. Integrable Syst.*, vol. 1, no. 1, Sep. 2015, Art. no. xyw012.

[54] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[55] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[56] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation," in *Advances in Artificial Intelligence* (Lecture Notes in Computer Science), vol. 4304. 2006, pp. 1015–1021.

[57] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognit.*, vol. 48, no. 9, pp. 2839–2846, Sep. 2015.

RUNYU SONG received the B.S. degree from the City Institute, Dalian University of Technology. She is currently pursuing the M.S. degree with Dalian Maritime University, China. Her research interest is in deep learning.

TAOYING LI received the B.S. and Ph.D. degrees from Dalian Maritime University, China. She is currently an Associate Professor with Dalian Maritime University. Her research interests are in machine learning, big data, and data mining.

YAN WANG received the B.S. degree from Dalian Jiaotong University, China. She is currently pursuing the M.S. degree with Dalian Maritime University, China. Her research interest is in deep learning.