A Deep Learning Method for Breast Cancer Classification in the Pathology Images

Min Liu, Lanlan Hu, Ying Tang, Chu Wang, Yu He, Chunyan Zeng, Kun Lin, Zhizi He, and Wujie Huo

Abstract*—*Objective: **Breast cancer is the most common female cancer in the world, and it poses a huge threat to women's health. There is currently promising research concerning its early diagnosis using deep learning methodologies. However, some commonly used Convolutional Neural Network (CNN) and their variations, such as AlexNet, VGGNet, GoogleNet and so on, are prone to overfitting in breast cancer classification, due to both small-scale breast pathology image datasets and overconfident softmax-cross-entropy loss. To alleviate the overfitting issue for better classification accuracy, we propose a novel framework for breast pathology classification, called the AlexNet-BC model. The model is pre-trained using the ImageNet dataset and fine-tuned using an augmented dataset. We also devise an improved cross-entropy loss function to penalize overconfident low-entropy output distributions and make the predictions suitable for uniform distributions. The proposed approach is then validated through a series of comparative experiments on BreaKHis, IDC and UCSB datasets. The experimental results show that the proposed method outperforms the state-of-the-art methods at different magnifications. Its strong robustness and generalization capabilities make it suitable for histopathology clinical computer-aided diagnosis systems.**

***Index Terms*—Deep learning, breast cancer, transfer learning, loss function, AlexNet.**

I. Introduction

According to a report recently published by the American Cancer Society [1], breast cancer alone accounts for 31% of estimated new cancer cases in females. It is the most common female cancer in the world with a very high mortality rate. Early detection and treatment are currently the most effective means to reduce breast cancer mortality. Pathologists often use the conventional diagnostic analysis method to determine histological grade and hormone receptor status through immunohistochemistry (IHC). With the accumulation of digital pathological images that are commonly in the form of whole slide images (WSIs), the diagnostic process is extremely tedious and subject to observer variants [2].

Machine learning (ML) methodologies, on the other hand, have received an increasing interest in moving this process to an automated region to avoid human errors with higher accuracy. Many ML-based approaches [3], [4] have been introduced to assist early diagnosis in computer-aided diagnosis systems (CADs) and proven to be reliable and effective [3]. ML-based approaches for the task of breast cancer classification typically follow three steps [4]: a) preprocess, b) feature extraction, and c) classification. The preprocess technique is performed to obtain mini local patches of size 256×256, sampled from large WSIs [5]–[8]. Note that both types of patches with and without tumor cells present are sent for feature extraction and classification. In the feature extraction step, hand-crafted features, such as a local binary pattern [9], a gray-level co-occurrence matrix [10], and shape-based features [11], etc., are often used in common classifiers, such as support vector machines [12], [13], [44], AdaBoost [43], random forest [14], and decision trees [15], etc. The feature extractor used is subject to human professional knowledge. Therefore, the reliability of ML-based approaches for breast cancer classification is relatively low due to the imperfection of manual features and the sensitivity of feature extractors.

Compared to the above-mentioned ML methods, deep learning approaches that adopt the end-to-end network [26] can realize automatic feature learning as well as merge feature extraction and classification together. In spite of the prominent advantages of deep learning and the noticeable improvement reported in [16]–[18] for breast cancer classification, its applications in the area of clinical medicine remain limited. Given that it is challenging to label pathological images manually, most of the public breast pathological images datasets for breast cancer classification are of a small scale, with which under-constraint deep

Manuscript received 29 January 2022; revised 27 May 2022; accepted 26 June 2022. Date of publication 1 July 2022; date of current version 5 October 2022. This work was supported in part by the China National Science Foundation under Grant 61901165, in part by the Special Project of Central Government for Local Science and Technology Development of Hubei Province under Grant 2019ZYYD020, in part by the Science and Technology Research Program of Hubei Provincial Department of Education under Grant T201805, in part by the Hubei University of Technology Ph.D. Research Startup Fund Project under Grants BSQD2020015 and BSQD2020014, and in part by the Key Research and Development Plan of Hubei Province under Grant 2021BGD013. *(Corresponding author: Ying Tang.)*

Min Liu, Lanlan Hu, Chu Wang, Yu He, Chunyan Zeng, Kun Lin, Zhizi He, and Wujie Huo are with the Hubei Key Laboratory for High-Efficiency Utilization of Solar Energy and Operation Control of Energy Storage System, Hubei University of Technology, Wuhan 430068, China (e-mail: liu_min@hbut.edu.cn; h_lanlan@hbut.edu.cn; 514944028@qq.com; heyu@hbut.edu.cn; cyzeng@hbut.edu.cn; 1004860611@qq.com; 1097930015@qq.com; 1401416803@qq.com).

Ying Tang is with the Department of Electrical and Computer Engineering, Rowan University, Glassboro, NJ 08028 USA, also with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the Institute of Smart Education, Qingdao Academy of Intelligent Industries, Qingdao, China (e-mail: tang@rowan.edu).

Digital Object Identifier 10.1109/JBHI.2022.3187765

learning models likely overfit. Ground-truth labels for breast cancer classification are typically in the form of one-hot, where the label set contains only one and zero, with one indicating the class probability of a hard target, and zero for the class probability of a non-hard target. The softmax-cross-entropy loss, commonly used to train deep learning models tend to be overconfident on probability predictions for training samples with such one-hot labels [24]. Therefore, these classical deep learning models that have been widely used in pathological image analysis, such as AlexNet [19], VGGNet [20], GoogleNet [21], Inception [22], and ResNet [23], are prone to overfitting.

A perusal of the current literature provided several techniques to overcome the above-mentioned overfitting issue for breast cancer classification. For instance, data augmentation [29], [30], dropout [33] and transfer learning [17], are often employed to improve the generalization ability of deep learning models by operating on hidden activations or weight [25]. Label smoothing [24], on the other hand, is an output regularization technique. By introducing noise to mix one-hot training labels with uniform label vectors, it alleviates the overconfidence of the softmax-cross-entropy learning, and consequently suppresses its overfitting. However, injecting the same noise into all training samples [35] inevitably causes bias in model training. In fact, a high-confidence probability prediction in softmax-cross-entropy learning means a low-entropy output distribution. Thus, another possible way of regularizing deep learning models is to penalize confident output distributions.

Motivated by these remarks, this paper proposes a breast cancer classification framework based on the original AlexNet model and makes the following contributions:

1. We propose a network model framework, called AlexNet-BC. Considering that the performance of AlexNet-BC will be affected by small-scale datasets, we utilize data augmentation techniques to expand the original datasets and introduce transfer learning strategy to fine-tune the parameters of the improved AlexNet-BC.
2. We devise a new loss function, where a penalty term is added to the original cross-entropy to make the prediction results suitable for uniform distributions, when a predicted probability is more than a preset threshold.
3. The proposed method is first trained and verified using the BreaKHis dataset. Its generalization ability is further demonstrated through the IDC and UCSB datasets.

The rest of the paper is structured as follows. Section II is the related works. Section III introduces our proposed data augmentation method, the AlexNet-BC network model, and the newly devised loss function. Section IV is the performance evaluation of the proposed model, followed by our conclusion and future directions in Section V.

II. Related Works

A. Alleviating Overfitting Caused By Undersized Datasets

Neural networks are data-driven models that need a large amount of data to train. However, many application domains, such as medical image analysis [30], do not have the access to big data. In such cases, data augmentation [31], [32], including geometric transformation, color change, random crop, scale jittering, flip, random erasing, noise injection, kernel filters, and mixing images, etc., is often used to make up for the small amount of data. Such a technique presents some limitations, particularly on the quality of augmented datasets.

Transfer learning is another way to alleviate the overfitting problem caused by small-scale data by pre-training CNNs on ImageNet. Related studies [17], [27], [28] have shown that models pre-trained on ImageNet can learn similar features from other tasks owing to the similarities of shallow features. The experiments conducted in [28] proved that a pre-trained CNN, after proper fine-tuning, can perform as good as training from scratch in the worst case. Vesal *et al.* (2018) [17] pre-trained Inception-V3 and ResNet-50, respectively, on ImageNet, and then fine-tuned them on breast pathology dataset with positive outcomes. Mohamed *et al.* (2018) [27] tested the effectiveness of migration learning using the AlexNet model, which was pre-trained on ImageNet and then fine-tuned using mammography data. They concluded that such off-the-shelf features extracted from an off-the-shelf deep learning model trained on natural images show competitive performance in certain medical tasks.

In this paper, we take good advantages of both data augmentation and transfer learning to alleviate the overfitting issue caused by the size and quality of breast pathology datasets. The former is used to expand the original breast pathology dataset; while the latter to fine-tune the parameters of the proposed network framework.

B. Reducing the Overfitting of Softmax-Cross-Entropy Learning

Label smoothing has been widely used to address the overfitting issue due to the over-confidence of the softmax-cross-entropy in deep learning models. The performance of label smoothing in image classification, speech recognition, and machine translation can be found in [34]. Lukasik *et al.* (2020) [35] proved that label smoothing is competitive in comparison to other loss correction techniques with more powerful noise reduction capability. The integration of label smoothing with other strategies were further explored by several researchers. For instance, Xu *et al.* (2020) [36] put forward a simple and effective strategy, so-called Two-Stage LAbel smoothing algorithm (TSLA), based on the idea of switching training from a smooth label to a hot label. Doing so improved the convergence of the model. Similar works can be seen in [24], [25]. Although the above methods demonstrate good performance to certain extent, the learning bias caused by the uniformly distributed noise in all training samples has not yet undertaken.

III. Methodology

In this paper, we propose an automatic classification modeling framework for breast pathology images. Fig. 1 shows the proposed modeling framework, including the data augmentation and transfer learning strategies. At first, our proposed AlexNet-BC model is pre-trained on the ImageNet dataset, and then coarsely tuned using the preprocessed images while the

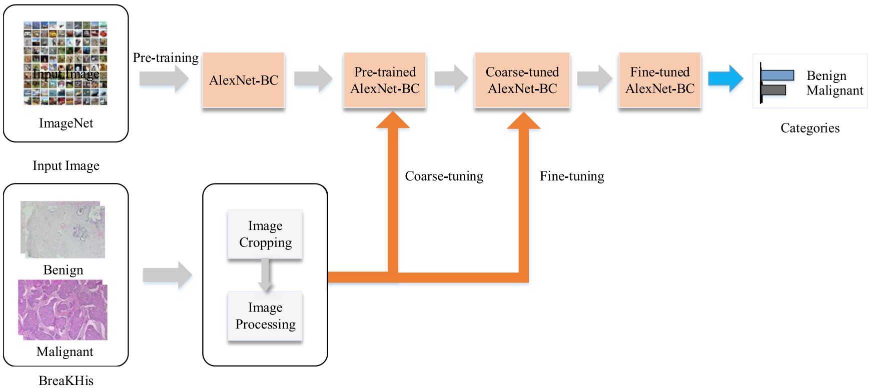


Fig. 1. The flowchart of the proposed modeling framework for breast pathological classification.

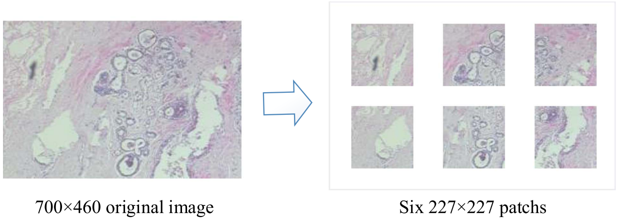


Fig. 2. Patch Cropping.

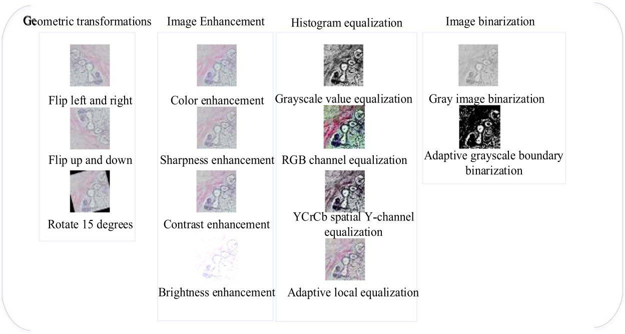


Fig. 3. Image Processing.

convolutional layers of the pre-trained model are frozen. Finally, the model is fine-tuned by unfreezing the convolutional layers.

A. Data Augmentation

The accuracy and robustness of deep learning methods rely heavily on sufficient training samples. However, current publicly available breast pathology image datasets are too small, or their image quality is too poor to be used to train models with high accuracy and robustness. To that end, this paper adopts a series of data augmentation methods to solve this problem.

Firstly, six patches are randomly extracted from each image, as shown in Fig. 2. Each cropped image uses four different types of image preprocessing methods, which are geometric transformation, image enhancement, histogram equalization, and image binarization, for data expansion, as shown in Fig. 3. The two geometric transformations used in the paper are flipping and rotating. The four different image enhancement methods include color enhancement, sharpness enhancement, contrast enhancement, and brightness enhancement. Additionally, the histogram equalization method is applied to both grayscale and RGB images to enhance their contrast. Furthermore, the image binarization method is used to make the features more prominent by setting different grayscale boundaries.

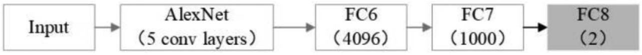


Fig. 4. The network structure of the original AlexNet.

Through data augmentation, the dataset is expanded to 20 times of the original. All the images after data augmentation are normalized as the input of the network.

B. AlexNet-BC Model

AlexNet [19] was the champion model of the ImageNet competition in 2012. Compared to other CNN models, such as VGGNet, GoogleNet, and ResNet etc., the structure of the original AlexNet as shown in Fig. 4, with five convolutional layers and three fully-connected (FC) layers, is much smaller with fewer parameters. For clear presentation, the five convolutional layers in AlexNet are simplified into a single box, and the fully-connected layers are depicted individually using FCi, where the subscription i indicates the layer level and the number of neurons in FCi is specified. Apparently, such structure is more advantageous in terms of computational complexity. In addition, AlexNet uses the Rectified Linear Unit (ReLU) to improve its nonlinearity and the dropout technique [39], [40] to alleviate overfitting caused by small-scale datasets. The layer without ReLU and dropout is highlighted in gray in the network model.

Dropout is an effective way to deal with overfitting in AlexNet. Typically, an input neuron is dropped from the network with a probability parameter d during the training, resulting in a rather "thinned" network [39]. Without the loss of generality, the dropout process as well as its impact on each output neuron are presented in (1)–(4).

$$r_j^{(l)} \sim \text{Bernoulli}(d) \qquad (1)$$

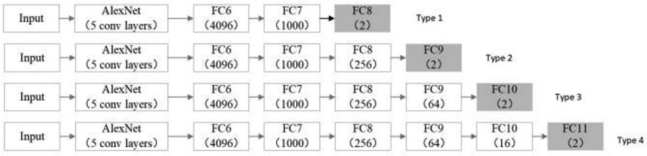


Fig. 5. Comparison of the model structure of the modified full connection layer.

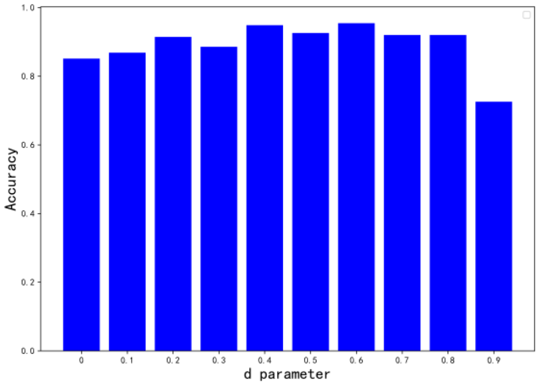


Fig. 6. Model accuracy changes in different dropout strategies.

TABLE I
TEST RESULT OF DIFFERENT MODELS

Type	40×	100×	200×	400×
Type 1	93.71±3.5%	92.71±2.8%	93.85±1.9%	89.38±1.5%
Type 2	**96.57±3.8%**	**94.27±2.9%**	**94.41±2.5%**	**91.87±1.3%**
Type 3	95.43±3.4%	93.75±3.9%	92.74±2.6%	91.25±2.3%
Type 4	65.71±6.5%	90.62±6.8%	91.62±7.2%	90.62±5.9%

TABLE II
PARAMETERS OF ALEXNET-BC MODEL

Type	Filter size	Channels(Strides)
Input size	227×227×3	
Conv1	11×11	64(4)
Max Pool1	3×3	(2)
Conv2	5×5	192
Max Pool2	3×3	(2)
Conv3	3×3	384
Conv4	3×3	256
Conv5	3×3	256
Max Pool5	3×3	(2)
Fc6	4096	
Dropout6	0.6	
Fc7	1000	
Dropout7	0.6	
Fc8	256	
Dropout8	0.6	
Fc9+ Softmax	2	

$$\tilde{\mathbf{y}}^{(l)} = \mathbf{r}^{(l)} * \mathbf{y}^{(l)} \tag{2}$$

$$\mathbf{z}_j^{(l+1)} = \mathbf{w}_j^{(l+1)} \tilde{\mathbf{y}}^l + \mathbf{b}_j^{(l+1)} \tag{3}$$

$$\mathbf{y}_j^{(l+1)} = f\left(\mathbf{z}_j^{(l+1)}\right) \tag{4}$$

where l is the index of hidden layers of a network and j the index of neuron nodes in each layer. $\mathbf{z}^{(l)}$ and $\mathbf{y}^{(l)}$ denote the input/output vector at the layer l, respectively. $\mathbf{w}^{(l)}$ is the weight of the layer l and $\mathbf{b}^{(l)}$ the bias of the layer l. $\mathbf{r}^{(l)}$ is an independent Bernoulli random variable that takes the value 0 with probability d (the dropout ratio) and the value 1 with probability 1-d. "thinned" outputs $\tilde{\mathbf{y}}^{(l)}$ at the layer l are then used as the input to the next layer l+1.

At the test time, the neuron obtained after the dropout operation is the output weight multiplied by the probability, 1-d (the retention ratio), which is shown in formula (5):

$$W_{\text{test}}^{(l)} = (1-d)W^{(l)} \tag{5}$$

Although there is a rule of thumb to choose the retention probability, the choice of d matters in terms of the network performance. To fully understand such effects, an experiment is conducted, where the original AlexNet is trained using the BreaKHis dataset for 200 iterations. With d changes from 0 to 0.9, the accuracy of the model is measured and compared as shown in Fig. 6. Apparently, the performance of the model is slightly improved with the increase of d when the value of is less than 0.6; and then begins to decline when d is greater than 0.6. Randomly disabling neurons and their corresponding connections prevents the network from relying too much on individual nodes, and in turn forces all nodes to learn with better generalization ability. However, if too many neurons are discarded, the model would not even perform.

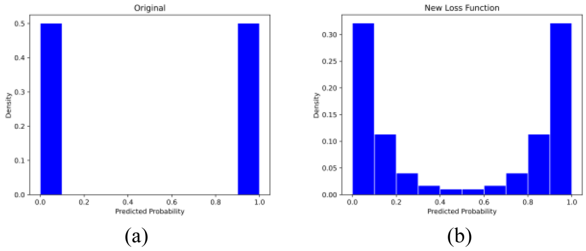
Pathological images are more complex than the ones in ImageNet, requiring a network with stronger nonlinear learning ability. This observation motivates us to consider the benefits of additional fully-connected layers to AlexNet. To that end, an experimental evaluation is carried out, where four types of networks are constructed by adding a different number of FC layers to the original AlexNet as shown in Fig. 5. As the magnification increases from 40× to 400×, the performance of the four networks is measured and compared using the BreaKHis dataset, where the dropout value is set as 0.6. It is clear to see in Table I that the type 2 network outperforms others. Inserting an additional FC layer to the original AlexNet does increase the nonlinearity of the network, consequently improving its accuracy. However, as more additional FC layers added, the overfitting issue becomes more severe.

Through the aforementioned exploration, our proposed model, AlexNet-BC, is constructed by adding one fully-connected layer to the original AlexNet and setting the dropout parameter as 0.6. All necessary parameters of AlexNet-BC are then given in Table II.

C. New Loss Function

The softmax classifier, as shown in (6), is usually used in the last layer of the network. Its purpose is to calculate the logits and output the values in the form of a probability vector, representing the probability of a sample belonging to each output class.

$$\mathbf{q}^i = \frac{\mathrm{e}^{\mathbf{y}^i}}{\sum_{i=1}^{k} \mathrm{e}^{\mathbf{y}^i}} \tag{6}$$



(a)

(b)

Fig. 7. Distribution of magnitude of softmax probabilities on BreaKHis validation set.

where k is the number of neuron nodes in the softmax layer; $\mathbf{y} = [\mathbf{y}^1; \mathbf{y}^2; \ldots; \mathbf{y}^k] \in R^{k\times 1}$ the original output matrix of the layer; and $\mathbf{q} = [\mathbf{q}^1; \mathbf{q}^2; \ldots; \mathbf{q}^k] \in R^{k\times 1}$ the probability distribution matrix, $\sum_{i=1}^{k} \mathbf{q}^i = 1$.

When paired with the cross-entropy loss function formulated in Eq. (7), the softmax classifier presents a high risk of overfitting. Typically, the softmax function maps the outputs of multiple neurons in the last fully-connected layer to interval (0, 1). Doing so, the predicted probability of a correct sample is always forced to be close to 1 in the binary classification of breast cancer, making the cross-entropy loss value close to zero.

$$C(\mathbf{p},\mathbf{q}) = -\sum_{i}^{k} \mathbf{p}^i \log(\mathbf{q}^i)$$
$$= -\sum_{i}^{k} \mathbf{p}^i \log\left(\frac{e^{yi}}{\sum_{i}^{k} e^{yi}}\right) \tag{7}$$

where $\mathbf{p} = [\mathbf{p}^1; \mathbf{p}^2; \ldots; \mathbf{p}^k] \in R^{k\times 1}$ is the label matrix.

With this in mind, we devise a new loss function, M(p, q), with the aim to regularize the model in a way of being less confident and more adaptable. In particular, the new function penalizes the cross-entropy losses when they exceed a preset threshold *v*. (8) presents its mathematical formula.

$$M(\mathbf{p},\mathbf{q}) = \begin{cases} C(\mathbf{p},\mathbf{q}), \mathbf{q}^i \le v \\ (1-\lambda)C(\mathbf{p},\mathbf{q}) + \lambda \cdot C(\mathbf{\Delta},\mathbf{q}), \mathbf{q}^i > v \end{cases} \tag{8}$$

where, v is the threshold value and λ is the weight coefficient of the penalty term, both of which are determined based on experiences. $\mathbf{\Delta} = [\frac{1}{k}; \frac{1}{k}; \ldots; \frac{1}{k}] \in R^{k\times 1}$.

When $\mathbf{q}^i > v$, it can be further simplified into formula (9).

$$M(\mathbf{p},\mathbf{q}) = (1-\lambda)C(\mathbf{p},\mathbf{q}) + \lambda \cdot C(\mathbf{\Delta},\mathbf{q})$$
$$= C(\mathbf{p},\mathbf{q}) + \lambda(C(\mathbf{\Delta},\mathbf{q}) - C(\mathbf{p},\mathbf{q})) \tag{9}$$

where $\lambda(C(\mathbf{\Delta},\mathbf{q}) - C(\mathbf{p},\mathbf{q}))$ is the added penalty term. Its essence is to make the network fitting uniformly distributed to some extent, which can relieve the overfitting. The effect of the new loss function is verified and analyzed in Section IV.

From Fig. 7(a), it can be seen that the output probability of softmax is prone to be in the interval (0.9, 1) and interval (0, 0.1). When the probability output is close to one, it causes the output entropy of the original loss function be near to zero, which tends to cause the model to be overconfident. As can be seen from Fig. 7(b), when the penalty term is added to the original loss function, the model's predicted probability values become more uniformly distributed. It is indicated that the new loss function is effective in reducing the overconfidence and alleviating the overfitting in the deep learning model.

TABLE III
DISTRIBUTION OF BENIGN AND MALIGNANT DATA SETS WITH DIFFERENT MAGNIFICATION FACTORS

Magnification	Benign	Malignant	Total
40×	625	1370	1995
100×	644	1437	2081
200×	623	1390	2013
400×	588	1232	1820
Total	2480	5429	7909
#Patients	24	58	82

IV. PERFORMANCE EVALUATION

A series of experiments is carried out in this section to validate our proposed model and showcase its performance. The experimental environment, including the datasets and evaluation metrics, is first given in Section IV.A. The experimental results on the network structure are presented in Section IV.B, followed by the evaluation on the proposed loss function in Section IV.C. Note that all data presented here is the average of the results from the same experiment repeated ten times.

A. Experimental Preparation

1) Experimental Environment: The whole series of experiments is performed on a computer with an Intel i7-9700 3.0 GHz CPU and an NVIDIA Quadro RTX 4000 GPU. It is trained under the PyCharm IDE with Python 3.6.4, Tensorflow 1.8.0 and Keras 2.2.4.

2) Datasets: The experiments are conducted on three standard benchmark datasets to test the performance of our method, including the BreaKHis dataset [11], the invasive ductal carcinoma (IDC) dataset [41], and the UCSB Bio-Segmentation Benchmark dataset [42]. The BreaKHis dataset is divided into three parts: 60% is the training set for the network, 20% is the validation set, and the rest is used as the test set. In addition, the IDC dataset and USCB dataset are both used as the testing sets to verify the model's generalization capability. The details of individual datasets are given below.

The BreaKHis database was established in 2014 by a P&D medical laboratory in Brazil. The laboratory obtained breast cancer pathological living tissue from 82 patients, used hematoxylin and eosin (H&E) staining to make tissue sections, and finally generated corresponding digital images with the help of an electron microscope. The final diagnosis of each case is provided by an experienced pathologist and confirmed by additional examinations such as immunohistochemical analysis. The image is saved in a three-channel RGB uncompressed portable network graphics format, with a size of 700×460 pixels, which is the original image without normalization and color standardization. The image distribution of the dataset is shown in Table III.

The (IDC) dataset is a commercially available dataset consisting of 162 whole mount slide images of Breast Cancer (BC)

TABLE IV
THE TEST RESULT OF TRANSFER LEARNING AND TRAINING FROM SCRATCH

Magnification	40×	100×	200×	400×
AlexNet+tf	90.9±2.9%	90.2±3.6%	91.2±1.5%	85.4±0.9%
From scratch [38]	89.6±6.5%	85.0±4.8%	84.0±3.2%	80.8±3.1%

specimens scanned at 40×. From it, we extract 277,524 patches of size 50×50, with 198,738 IDC negative and 78,786 IDC positive.

In the UCSB dataset, there are about 58 H&E stained histopathology images used in breast cancer cell detection with associated ground truth data available.

3) Evaluation Metrics: This paper uses the recognition accuracy of breast cancer pathological images to measure the classification performance of the model expressed in (10), where N_r represents the number of correctly classified pictures and N_a represents the total number of pathological pictures in the dataset.

$$\text{Accuracy} = \frac{N_r}{N_a} \tag{10}$$

B. Model Structure Improvement

In this set of experiments, we augment the original BreaKHis dataset to train the AlexNet network using the transfer learning strategy presented in Section III and compare its performance with that of the network trained from scratch [38]. The comparative results at different magnifications are shown in Table IV.

The average accuracy of "AlexNet+tf" under four different magnifications, is consistently higher than that "from scratch", where "tf" represents transfer learning strategy, indicating that the transfer learning strategy can improve the network performance. Because of the pre-training of the transfer learning experience, the network learns some prior knowledge from ImageNet, which has a positive effect on the classification of the BreaKHis dataset.

Additionally, the augmented BreaKHis dataset and the transfer learning strategy are used to train both our proposed model (the AlexNet-BC model) and the original AlexNet model, after which we compare their performance. As shown in Fig. 8, "AlexNet-BC+tf" at four different magnifications outperforms "AlexNet+tf", especially at 400×. The combination of our proposed model and transfer learning method can better improve the performance of the model.

C. Loss Functions

We augment the original BreaKHis dataset to feed the AlexNet-BC network structure with the common cross-entropy loss function (formula (7)) and our proposed function (formula (8)) as AlexNet-BC's loss function, respectively. And we use the transfer learning strategy to train them. The penalty term λ and the threshold v are decided empirically. For our experiments, λ is set as 0.1, and the penalty term v is around 0.6 at 40×, 100× and 400×, and around 0.8 at 200×.

It can be seen from Table V that the average accuracy of our proposed loss function has been significantly improved under four different magnifications. This is because the proposed loss function alleviates the overconfidence of the model. Specifically, we still use the cross-entropy loss function when the output of softmax is below the threshold, while we add a penalty term to fit a uniform distribution when the output is greater than the threshold.

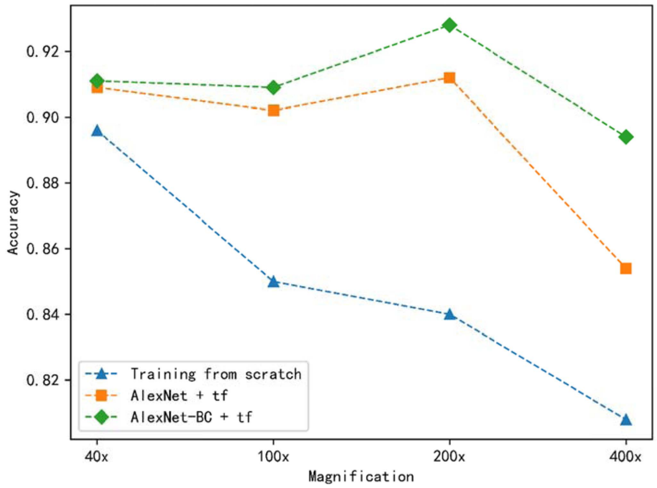


Fig. 8. The improvement of AlexNet-BC and transfer learning.

TABLE V
TEST RESULT OF DIFFERENT LOSS FUNCTIONS

Magnification	40×	100×	200×	400×
Formula (7)	91.1±2.7%	90.9±3.3%	92.8±2.5%	89.4±1.8%
Formula (8)	98.15±0.9%	97.71±1.9%	97.96±0.7%	98.48±1.1%

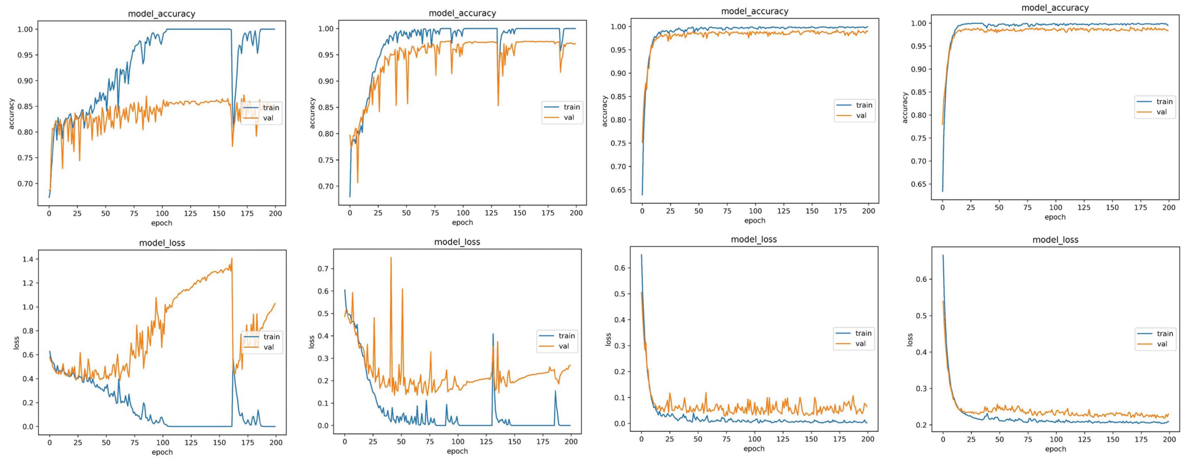
TABLE VI
COMPARISON WITH OTHER METHODS

Magnification	40×	100×	200×	400×
VGG-16 [37]	97.02%	97.23%	97.89%	97.50%
GoogleNet [16]	93.3±2.3%	94.6±2.2%	94.8±3.2%	88.4±4.1%
ResNet	97.74±0.8%	96.64±0.5%	95.03±1.9%	94.24±2.3%
This paper	98.15±0.9%	97.71±1.9%	97.96±0.7%	98.48±1.1%

TABLE VII
COMPARISON WITH IDC DATASETS

Model	Acc
VGG-16	81.06±3.2%
GoogleNet	85.83±1.2%
ResNet	86.05±0.8%
CNN [41]	84.23%
This paper	**86.31±1.7%**

Furthermore, it can be seen from Fig. 9 that our proposed method is more effective in alleviating the overfitting. We observe inconsistencies in the predictions shown by the model on the training and validation sets. If the model is tested on both the training and validation sets, the closer the predictions obtained, the better the fit. The larger interval between the training and validation curves of the model is, the more severe



(a) AlexNet (b) AlexNet-BC (c) AlexNet-BC+tf (d) AlexNet-BC+tf+NL

Fig. 9. Accuracy and loss results plots for the model training set and validation set with different training stages, where ‘tf’ represents the transfer learning method, where ‘NL’ represents our proposed loss function.

TABLE VIII
TEST WITH UCSB DATASETS

Model	Acc
VGG-16	90.09±1.8%
GoogleNet	87.02±3.8%
ResNet	90.06±0.9%
This paper	**96.10±0.8%**

the overfitting is. It can be observed that our proposed modeling framework effectively alleviates the overfitting problem of the original AlexNet network on the BreaKHis dataset. Specially, our proposed new loss function makes the training and validation curves closer together and suppresses the overfitting effectively.

D. Comparison With Other Methods

In this experiment, the same BreaKHis dataset is used to train and test our proposed method together with three other existing ones, VGG-16-based model proposed by Wei *et al.* (2017) [37], GoolgeNet-based model proposed by Zhi *et al.* (2017) [16] and ResNet [23]. Their performance at different magnifications is then compared and shown in Table VI. Apparently, the average accuracy of our proposed method is higher than that of [37], [16] and the original structure of ResNet [23].

E. Validation of the Model’s Generalization Capability

This set of experiments uses IDC dataset [41] and the UCSB dataset to test the generalization capability of our proposed model. Given that the magnification of the IDC dataset is 40×, we choose the model weights under BreaKHis dataset at 40× only when using the IDC dataset for testing. Similarly, since the magnification of the UCSB dataset is unknown, we use the model weights on the entire BreaKHis dataset when testing using the UCSB dataset. The results shown in Tables VII and VIII demonstrate that the proposed approach outperforms the transfer learning-based the original structure of VGG-16 [20], GoogleNet [21], and ResNet [23], presenting a relatively good generalization ability.

V. CONCLUSION

In this paper, we propose an automatic classification modeling framework for breast pathological images, called AlexNet-BC. To prevent the overfitting caused by the small-scale datasets of breast pathological images, we first perform data augmentation on the BreaKHis and optimize the network structure of the original AlexNet. To alleviate the overfitting caused by the overconfidence of softmax-cross-entropy learning, we propose a new low-entropy output penalty in which the cross-entropy loss function will be penalized when the predicted likelihood probability is higher than a preset threshold. We adopt the transfer learning strategy to pre-train and fine-tune this improved neural network framework. The experimental results prove that the proposed method outperforms The state-of-the-art methods, which can be applied to CADs for breast histopathology clinical diagnosis. On the BreaKHis dataset, our results outperform the best currently published results by 0.41%, 0.48%, 0.07% and 0.92% at 40×, 100×, 200× and 400×, respectively. Moreover, the generalization ability of the network is verified via both the IDC dataset and UCSB dataset. While our method performs with high accuracy of the breast pathological images classification, the lesion area is not divided. Dividing the lesion area by combining the knowledge of deep learning semantic segmentation will be our future research.

REFERENCES

[1] A. N. Giaquinto et al., “Cancer statistics for African American/Black People 2022,” *CA Cancer J. Clin.*, vol. 72, no. 1, pp. 202–229, 2022.

[2] N. Kumar, R. Gupta, and S. Gupta, “Whole slide imaging (WSI) in pathology: Current perspectives and future directions,” *J. Digit. Imag.*, vol. 33, no. 4, pp. 1034–1040, 2020.

[3] N. Abdullah-Al and Y. Kong, "Involvement of machine learning for breast cancer image classification: A survey," *Comput. Math. Methods Med.*, vol. 2017, pp. 1–29, 2017.

[4] D. Komura and S. Ishikawa, "Machine learning methods for histopathological image analysis," *Comput. Struct. Biotechnol. J.*, vol. 16, pp. 34–42, 2018.

[5] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Comput.*, 2016, pp. 2424–2433.

[6] J. Wang et al., "Tumor detection for whole slide image of liver based on patch-based convolutional neural network," *Multimedia Tools Appl.*, vol. 80, no. 11, pp. 17429–17440, 2021.

[7] N. Dimitriou, O. Arandjelović, and P. D. Caie, "Deep learning for whole slide image analysis: An overview," *Front. Med.*, vol. 6, 2019, Art. no. 264.

[8] N. Farahani, A. V. Parwani, and L. Pantanowitz, "Whole slide imaging in pathology: Advantages, limitations, and emerging perspectives," *Pathol. Lab. Med. Int.*, vol. 7, pp. 23–33, 2015.

[9] M. Kowal et al., "Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images," *Comput. Biol. Med.*, vol. 43, no. 10, pp. 1563–1572, 2013.

[10] A. Saito et al., "A novel method for morphological pleomorphism and heterogeneity quantitative measurement: Named cell feature level co-occurrence matrix," *J. Pathol. Inform.*, vol. 7, 2016, Art. no. 36.

[11] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE. Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1455–1462, Jul. 2016.

[12] Q. Kang, L. Shi, M. Zhou, X. Wang, Q. Wu, and Z. Wei, "A distance-based weighted undersampling scheme for support vector machines and its application to imbalanced classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 4152–4165, Sep. 2018.

[13] W. Yue et al., "Machine learning with applications in breast cancer diagnosis and prognosis," *Designs*, vol. 2, no. 2, 2018, Art. no. 13.

[14] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogrammetry Remote Sens.*, vol. 114, pp. 24–31, 2016.

[15] Y. Y. Song and L. U. Ying, "Decision tree methods: Applications for classification and prediction," *Shanghai Arch Psychiatry*, vol. 27, no. 2, pp. 130–135, 2015.

[16] B. Wei, Z. Han, X. He, and Y. Yin, "Deep learning model based breast cancer histopathological image classification," in *Proc. Int. Conf. Int. Conf. Cloud Comput. Big Data Anal.*, 2017, pp. 348–353.

[17] S. Vesal et al., "Classification of breast cancer histology images using transfer learning," in *Proc. Int. Conf. Image Anal. Recognit.*, 2018, pp. 812–819.

[18] S. H. Kassani et al., "Breast cancer diagnosis with transfer learning and global pooling," in *Proc. Int. Conf. Inf. Commun. Technol. Convergence*, 2019, pp. 519–524.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *J. Inf. Process. Syst.*, vol. 25, pp. 1097–1105, 2012.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[21] C. Szegedy et al., "Going deeper with convolutions," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[22] S. Arora et al., "Provable bounds for learning some deep representations," in *Proc. Int. Conf. PMLR*, 2014, pp. 584–592.

[23] K. He et al., "Deep residual learning for image recognition," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[24] G. Pereyra et al., "Regularizing neural networks by penalizing confident output distributions," 2017, *arXiv:1701.06548*.

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Comput.*, 2016, pp. 2818–2826.

[26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[27] A. A. Mohamed et al., "A deep learning method for classifying mammographic breast density categories," *J. Med. Phys.*, vol. 45, no. 1, pp. 314–321, 2018.

[28] N. Tajbakhsh et al., "Convolutional neural networks for medical image analysis: Full training or fine tuning?," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[29] Q. Kang, S. Yao, M. Zhou, K. Zhang, and A. Abusorrah, "Effective visual domain adaptation via generative adversarial distribution matching," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 3919–3929, Sep. 2021.

[30] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[31] A. Hernández-García and P. König, "Further advantages of data augmentation on convolutional neural networks," in *Proc. Int. J. Neural Netw.*, 2018, pp. 95–103.

[32] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, *arXiv:1712.04621*.

[33] L. Qian, L. Hu, L. Zhao, T. Wang, and R. Jiang, "Sequence-dropout block for reducing overfitting problem in image classification," *IEEE Access*, vol. 8, pp. 62830–62840, 2020.

[34] R. Müller, S. Kornblith, and G. Hinton, "When does label smoothing help?," 2019, *arXiv:1906.02629*.

[35] M. Lukasik et al., "Does label smoothing mitigate label noise?," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 6448–6458.

[36] Y. Xu et al., "Towards understanding label smoothing," 2020, *arXiv:2006.11653*.

[37] W. Zhi et al., "Using transfer learning with convolutional neural networks to diagnose breast cancer from histopathological images," in *Proc. Int. Conf. Neural Inf. Process.*, 2017, pp. 669–676.

[38] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw.*, 2016, pp. 2560–2567.

[39] N. Srivastava et al., "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[40] Z. Liu, "Soft-shell shrimp recognition based on an improved AlexNet for quality evaluations," *J. Food Eng.*, vol. 266, 2020, Art. no. 109698.

[41] A. Cruz-Roa et al., "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," *Proc. Med. Imag. 2014: Digit. Pathol.*, vol. 9041, 2014, Art. no. 904103.

[42] E. D. Gelasca et al., "A biosegmentation benchmark for evaluation of bioimage analysis methods," *BMC Bioinform.*, vol. 10, no. 1, pp. 1–12, 2009.

[43] Q. Kang, X. Chen, S. Li, and M. Zhou, "A noise-filtered under-sampling scheme for imbalanced classification," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4263–4274, Dec. 2017.

[44] Z. Tan, J. Chen, Q. Kang, M. Zhou, A. Abusorrah, and K. Sedraoui, "Dynamic embedding projection-gated convolutional neural networks for text classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 3, pp. 973–982, Mar. 2022.

[45] W. Li, G. Dasarathy, and V. Berisha, "Regularization via structural label smoothing," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 1453–1463.

[46] E. Yilmaz and M. Trocan, "Benign and malignant skin lesion classification comparison for three deep-learning architectures," in *Proc. Asian Conf. Intell. Inf. Database Syst.*, 2020, pp. 514–524.