IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY SECTION

Received 24 August 2023, accepted 10 September 2023, date of publication 13 September 2023,
date of current version 20 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3314978

RESEARCH ARTICLE

Brea-Net: An Interpretable Dual-Attention Network for Imbalanced Breast Cancer Classification

YI LIANG AND ZUQIANG MENG
School of Computer and Electronic Information, Guangxi University, Nanning 530005, China
Corresponding author: Zuqiang Meng (171313540@qq.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 62266004.

ABSTRACT Breast cancer is a prevalent disease worldwide, and early diagnosis plays a vital role in improving patient outcomes. Recent advancements in deep learning have shown great potential for accurate and efficient breast cancer classification. However, the existing methods still suffer from low accuracy and lack of interpretability. To overcome these limitations, we propose a novel and interpretable network to improve the performance of breast cancer classification tasks. By employing a dual-attention module called Convolutional Block Attention Module (CBAM) and a flexible and efficient classifier named Convolutional Multi-Layer Perceptron (ConvMLP), our model is able to effectively capture and exploit the discriminative spatial and channel features within histopathological images and learn complex patterns and relationships between features, leading to improved classification performance. The proposed model outperforms previous state-of-the-art works in terms of accuracy, precision, recall, and F1-score, yielding the highest accuracy of binary and eight-class classification on the BreaKHis dataset. Further, the generalization ability of our proposed network is tested on a different dataset called ICIAR 2018, scoring an outstanding accuracy of 95.5%.

INDEX TERMS Breast cancer, interpretability, imbalance, CBAM, classification.

I. INTRODUCTION

Breast cancer is a grave public health issue that affects nations at all levels of development [1], with breast cancer being a common occurrence in women. However, diagnosing cancer manually from biomedical images like histopathological images presents challenges due to the potential for inaccuracies caused by human error. A computer-aided diagnosis system is necessary to optimize the diagnosis process and reduce manual effort. Automated processes are already available to differentiate between malignant and benign cancer images [2], and deep learning has further enhanced the efficiency and accuracy of diagnosis, particularly in the early stages. With recent advancements in computational hardware accelerators and parallel processing, the role of deep learning has significantly expanded in various fields, including healthcare and bio-medicine, where significant progress has been achieved [3].

The Convolutional Neural Network (CNN) [4] is arguably one of the most popular deep learning architectures. A typical CNN is made up of convolutional building blocks that are combined to automate feature extraction and classification. Unlike in machine learning, where these steps are carried out separately, CNNs perform both tasks simultaneously. CNN networks can be broadly divided into two categories: those trained from scratch, and those pre-trained on a large dataset, which enables the transfer of knowledge from one domain to another. However, CNNs are often referred to as black boxes because they can be challenging to interpret and understand how they arrive at their decisions. Another reason why CNNs are sometimes referred to as black boxes is that they are often used in situations where the focus is on achieving high accuracy rather than understanding how the network works. In other words, the primary goal is to produce accurate

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

predictions rather than to gain insight into how the network arrived at those predictions.

In the field of medicine, we are facing complex challenges, particularly in integrating, fusing, and mapping distributed and heterogeneous data in high-dimensional spaces. Therefore, explainable AI in the medical domain must consider the potential contribution of diverse data to generate a significant result [5]. This requires that medical experts must understand how and why a machine-made decision was reached. Another problem is those real-world datasets often suffer from the class imbalance problem [6], where there is a disproportionate distribution of samples across different classes. This can make it difficult to accurately predict the less represented class(es). One example of an imbalanced dataset is the BreaKHis dataset [7], which contains significantly more Malignant samples than Benign ones. The class imbalance problem in this dataset presents many challenges, as it can lead to incorrect predictions. Therefore, it is essential to develop models that can effectively handle class imbalance and accurately identify cancerous patterns in biomedical data.

Deep learning methods are considered to lack interpretability, however, interpretability is crucial in healthcare to build trust, enable transparency, and improve accuracy for these important systems that impact people's health and lives [8]. The complex and high-stakes nature of healthcare makes interpretability especially vital. Doctors and patients need to trust the results and predictions of healthcare models; they need to understand how the conclusions were reached to have confidence in them. This requires the models to be interpretable and explainable. In terms of transparency, patients have a right to know how decisions about their health and treatment were made. Furthermore, interpretable models can help catch errors and correct inaccuracies. When doctors can understand how a model reached a conclusion, they can more easily identify mistakes or biases in the data or model, which improves the overall accuracy.

To overcome the challenge of class imbalance and detect cancerous patterns from bio-medical data, there is a need for a model that can handle these limitations and, more importantly, produce an interpretable prediction. This paper proposes a novel VGG-based architecture that is capable of extracting and exploiting both spatial and channel features within histopathological images and can learn complex patterns and relationships between features. Additionally, we use transfer learning to improve training efficiency by reducing the training time and computational cost required to train the model from scratch. The contributions of our work can be summarized as:

- We successfully propose a novel model to solve the dataset imbalance problem and produce a reasonable prediction.
- Using transfer learning and fine-tuning, our model yields the highest accuracy of binary and multi-class classification in comparison to state-of-the-art approaches for the BreaKHis dataset.
- After a series of evaluation experiments conducted, our model has demonstrated robust generalization ability and can be used for other classification tasks.

II. RELATED WORKS

Breast cancer detection using deep learning has been a topic of active research in recent years. In [9], Wang et al. proposed a CNN architecture combined with support vector machines (SVMs) for breast cancer detection. Spanhol et al. used the sliding window to extract patches for efficient training [10]. Kamyar et al. proposed a two-stage network and used a pre-trained patch-wise network to extract local information and an image-wise network to obtain global information [11]. Bayramoglu et al. introduced single-task and multi-task CNN models for the classification of the BreaKHis Histopathological dataset [12]. Bardou et al. examined several configurations of convolutional neural networks (CNN) along with other conventional methods. They emphasized the significance of data augmentation operations in improving the model's performance. The analysis of the results confirmed that the features extracted by CNN outperformed the handcrafted features [13].

To confront the class imbalance problem which happens due to an unequal distribution of classes in the dataset, several solutions have been suggested such as oversampling and undersampling [14]. Data augmentation can be applied to the minority class to increase the number of samples and make it equivalent to the majority one [15]. In [16], Saini et al. increased the number of minority samples using the Deep Convolutional Generative Adversarial Network (DCGAN) by synthetically generating the fake samples, and features were extracted via a modified VGG16-based architecture. Ding et al. demonstrated that a deep network (more than 10 layers) can improve the training process and achieve a better rate of convergence for imbalanced datasets [17]. The authors performed experiments to verify the assertion using deep neural networks trained over 100 epochs, as compared to shallower networks. Other researchers have proposed relevant techniques to address the issue of imbalanced datasets, such as Abbas's Decompose, transfer and compose (DeTraC) model which leverages the concept of class decomposition within CNN approach to effectively learn class boundaries [18]. They also applied error correction criteria to the softmax layers of the network, which resulted in improved classification performance. Although these methods have to some extent mitigated the impact of imbalanced datasets on classification results, there are still some shortcomings. First, many of the existing methods rely on re-sampling techniques such as oversampling or undersampling to balance the class distribution. However, these methods can lead to overfitting or loss of valuable information. Second, although deeper networks can achieve better classification performance, if the dataset is small, it may lead to overfitting.

There are several techniques available for visualizing networks in a way that can be easily interpreted. One such method is feature visualization, which involves

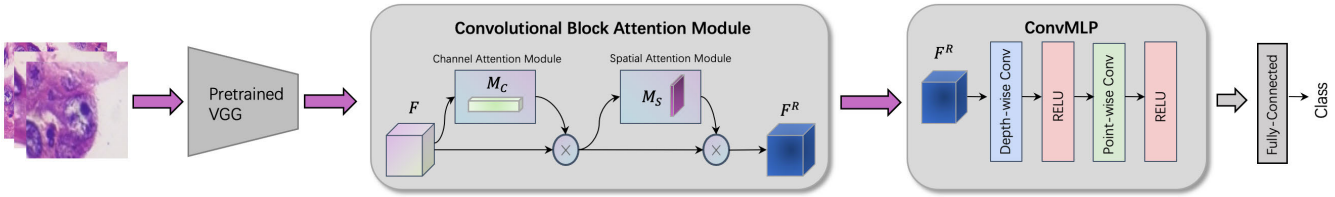

FIGURE 1. The overview of the proposed network. The network has three main components: pre-trained VGG network, CBAM, and ConvMLP.

reconstructing an image from the feature space to show how a particular feature map corresponds to the appearance of the image. While this approach, such as DeConvNet [19], can be useful for analyzing hidden layer features, the resulting visualizations can often be abstract and difficult for non-specialists to understand [5]. Gradient-based methods are another common approach to visual interpretation in computer vision. These methods, including saliency mapping, guided backpropagation (GBP) [20], and layer-wise relevance propagation (LRP) [21], rely on gradients to identify discriminative features in an image. Deep Taylor decomposition [22], integrated gradient [23], and pattern attribution [24] are examples of gradient-based methods that offer superior visual performance. The class activation map (CAM) [25] is a popular method that generates a heatmap of discriminative features for classification, using heuristics that take into account the interaction between weights and activation intensity of certain units in the model. To improve the reliability of CAM, several techniques such as GradCAM [26] have been developed to smooth or regularize the visualization. These methods assign scores to input components to explain the model's decision, making the interpretation more intuitive.

In this paper, we have proposed an inspired network by concatenating pre-trained layers at the lower level with trainable layers at the higher level. This approach learns higher-level features specific to the breast cancer dataset while transferring general features from the lower layers of pre-trained convolutional models. Hybrid sampling and focal loss are applied to eliminate dataset imbalance. Our model consists of a dual-attention mechanism that incorporates both spatial-wise and channel-wise attention. The spatial attention module focuses on the most relevant regions of the histopathology image, while the channel attention module emphasizes the most discriminative features for classification. This sets our work apart from previous ones in this area.

III. METHODS

A novel deep-learning network has been proposed, as shown in Figure 1. We take the $3 \times 224 \times 224$ images as input, using pre-trained VGG16 feature layers (till block5 pool layer) as image encoder, where we can extract the most relevant bottleneck feature F. Also, the use of a pre-trained encoder can improve the performance of few-shot learning. Further, CBAM takes the obtained feature as input, sequentially infers a 1-D channel attention map M_c and a 2-D spatial attention map M_s, then generates refined feature F^R and transfers it to the next stage for classification. The final prediction is produced via a ConvMLP.

A. DATASETS

1) BreaKHis DATASET

Collected through clinical studies in 2014, the BreaKHis dataset contains 7909 high-resolution breast cancer histopathological images that are divided into two classes, benign and malignant. There are four subclasses within each type. The benign type includes adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenoma (TA), while the malignant type includes ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC). The images are captured using a three-channel RGB true color space, with varying magnifications of 40X, 100X, 200X, and 400X. The size of each image is 700×460, and Table 1 provides an overview of the distribution of these images. As we can observe, the number of samples belonging to one class is significantly higher than another, which means it is an imbalanced dataset.

TABLE 1. Distribution statistics of the BreaKHis dataset.

Type	Magnification Factor			
	40X	100X	200X	400X
Benign	625	644	623	588
Malignant	1370	1437	1390	1232
Total (7909)	1995	2081	2013	1820

2) ICIAR 2018 DATASET

The BACH 2018 Grand Challenge dataset (ICIAR 2018) [27] consists of 400 H&E-stained breast cancer histopathological images with large spatial dimensions (2048×1536 pixels). Each image is captured at 200X magnification, with a pixel size of 0.42 μm x 0.42 μm. A pair of expert pathologists carried out the annotation process, labeling each microscopic image as Normal, Benign, In-situ, or Invasive carcinoma, according to the predominant breast cancer subtype present in the image. Although smaller than the BreaKHis dataset, the ICIAR dataset is highly balanced.

B. DATA PREPROCESSING

The datasets are preprocessed before training. At first, images are normalized to eliminate color variability, improving robustness and accuracy. To avoid overfitting, data augmentation strategies are adopted, including random horizontal flipping, random vertical flipping, random rotation, and Gaussian blurring. Table 2 shows the selected parameters for these augmentation techniques. All images are ultimately resized to a dimension of 224 × 224. We shuffle each dataset and divide them into three groups: training set (70%), validation set (15%), and testing set (15%).

TABLE 2. Data augmentation metrics.

Technique	Value
RandomHorizontalFlip	True
RandomVerticalFlip	True
GaussianBlur	kernel_size=(5, 9), sigma=(0.1, 5)
RandomRotation	90

C. HYBRID SAMPLING TECHNIQUE

Many previous works have applied re-sampling techniques to improve the performance of imbalanced data classification. To balance the number of instances between classes, the sampling method employs oversampling and undersampling techniques, which involve generating new samples and deleting existing samples, respectively. While oversampling and undersampling can be effective, they also have some potential drawbacks. Oversampling can lead to overfitting, where the model becomes too specialized to the training data and does not generalize well to new, unseen data. On the other hand, undersampling can result in the loss of valuable information from the overrepresented class, leading to a biased or incomplete understanding of the data. In this work, we use a novel method called hybrid sampling, which combines oversampling and undersampling techniques. Firstly, we randomly remove some instances from the overrepresented class to match a certain level. Secondly, we generate instances of the underrepresented class to increase its size to the size of the overrepresented class.

D. FOCAL LOSS

Proposed by Lin et al., Focal loss is a more effective approach for dealing with class imbalance in dense object detection [28]. In this work, we creatively apply Focal loss to breast cancer histopathological image classification. In the case of imbalanced datasets, the model may tend to overfit on the majority class, resulting in poor performance on the minority class. Focal loss aims to address this problem by down-weighting the contribution of easy examples and emphasizing the contribution of hard examples. The focal loss function is defined as:

$$FL(p_t) = -\alpha_t(1-p_t)^{\lambda} log(p_t), \tag{1}$$

where p_t is the predicted probability for the true class, λ is a tunable parameter that controls the degree of down-weighting, and α_t is the weight assigned to the true class.

E. PRE-TRAINED IMAGE ENCODER

The original VGG Network was proposed by Simonyan and Zisserman [29], at the University of Oxford. A VGG block comprises 3 × 3 convolutional kernels and a 2 × 2 pooling layer. In this work, we use VGG16 feature layers, pre-trained in ImageNet-1k [30], to obtain the most relevant feature of breast cancer images. The obtained feature is then fed to the next stage, which is called CBAM.

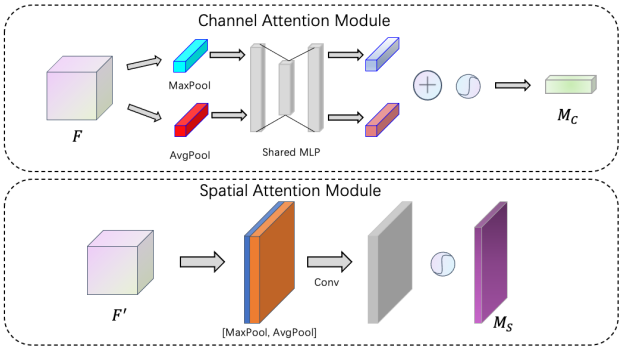

FIGURE 2. The pipeline of Channel Attention Module and Spatial Attention Module.

F. CBAM

CBAM is a simple yet effective attention module for the feed-forward convolutional neural network [31]. It contains two sequential sub-modules: Channel Attention Module (CAM) and Spatial Attention Module (SAM), both modules are given in Figure 2. CBAM takes obtained feature $F \in \mathbb{R}^{C\times H\times W}$ as input, sequentially infers channel attention map $M_c \in \mathbb{R}^{C\times 1\times 1}$ and spatial attention map $M_s \in \mathbb{R}^{1\times H\times W}$. The overall process can be summarized as:

$$F' = M_c(F) \otimes F, \tag{2}$$

$$F^R = M_s(F') \otimes F', \tag{3}$$

where $\otimes$ denotes element-wise multiplication, which means the attention values are broadcasted accordingly: channel attention values are broadcasted along the spatial dimension, and vice versa. F^R is the final refined feature.

1) CHANNEL ATTENTION MODULE

The channel-wise attention mechanism is designed to highlight important features in the feature maps. By selectively attending to important channels, the model can effectively extract more discriminative features, leading to improved performance. At first, we gather spatial context descriptors from a feature map by combining average-pooling and max-pooling operations, noted as F^c_{avg} and F^c_{max} respectively. Both descriptors are then fed to a multi-layer perceptron (MLP) to produce channel attention map M_c. In short, the channel attention can be calculated as follows:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F^c_{\text{avg}})) + W_1(W_0(F^c_{\text{max}}))), \end{aligned} \tag{4}$$

where σ denotes the sigmoid function, W_1 and W_0 are MLP weights.

2) SPATIAL ATTENTION MODULE

The spatial-wise attention mechanism allows the model to selectively attend to the most informative regions while ignoring irrelevant regions. This helps to reduce noise and improve the accuracy of the model. To obtain spatial attention, we use average-pooling and max-pooling operations, noted as F^s_{avg} and F^s_{max}, and concatenate them to produce a post-pooling feature. Further, we apply a convolutional layer to generate the spatial attention map. The process of computation is as follows:

$$\begin{aligned} M_s(F) &= \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) \\ &= \sigma(f^{7\times7}([F^s_{\text{avg}}; F^s_{\text{max}}])), \end{aligned} \tag{5}$$

where σ denotes the sigmoid function, $f^{7\times7}$ means we use 7×7 kernel in convolutional layer.

The aim of the CBAM module used in this study is to highlight important features on both spatial and channel axes. These axes are applied sequentially to enable each branch to learn what to emphasize or suppress on the channel and spatial axes. Therefore, our model facilitates the flow of information within the network. The channel attention module utilizes both average-pooled and max-pooled features simultaneously, resulting in an improved representation power compared to using each feature independently. On the other hand, the spatial attention module first applies average pooling and then max pooling along the channel axis. The resulting efficient feature descriptor is then fed into a convolutional layer to generate a spatial attention map that identifies where to emphasize or suppress. We used the channel module first, followed by the spatial module, and found that this sequential arrangement yielded better results compared to the parallel arrangement.

G. ConvMLP

Traditional MLP is fundamental for image classification. However, it suffers from lacking flexibility since it requires a fixed size of inputs. Moreover, using large consecutive MLPs results in increased computational overhead and brings more parameters. To overcome these problems, we apply ConvMLP to replace the original MLP. Inspired by Li et al. [32], ConvMLP can accommodate inputs of various sizes, thus facilitating transfer learning. In terms of computational expense, ConvMLP is more efficient compared to MLP due to its lower parameter count. According to the original paper, ConvMLP has achieved competitive results in several computer vision tasks. At first, we use a depth-wise convolution layer for spatial mixing. Further, we add a 1×1 convolution layer called the point-wise convolution layer, which is used for channel mixing. It succeeded by a custom fully connected layer that is used as a classifier.

IV. EXPERIMENTS

Our model was developed using PyTorch and all experiments were carried out on a platform equipped with an Intel Xeon(R) Gold 6330 CPU and Nvidia RTX 3090 GPU.

TABLE 3. Hyperparameter setup during the training stage.

Training Options	
Parameter	Value
Loss function	FocalLoss(gamma = 2.0,alpha = 0.25)
Optimizer	AdamW
Learning rate	0.001
StepLR	(step_size = 4, gamma = 0.7)
Max epochs	80
Batch size	64
Dropout	0.3

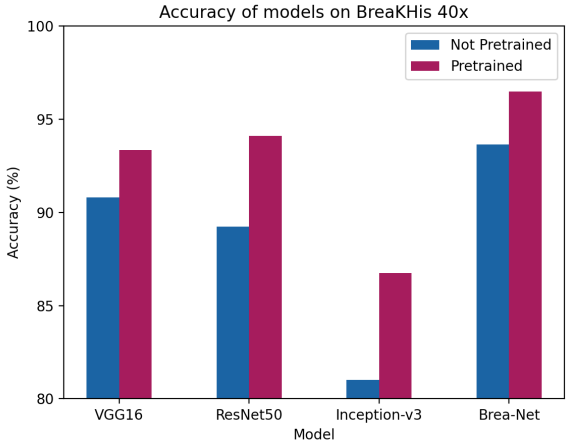

FIGURE 3. Classification performance of the proposed model and state-of-the-art networks.

In this study, we selected typical methods like VGG16 [29], ResNet50 [33], and InceptionV3 [34] as baselines. All models were pre-trained on ImageNet-1K. The reason for this was that our imbalanced classification dataset lacks sufficient samples when compared to the massive datasets, typically consisting of millions of images, needed to train a large ConvNet effectively from the beginning. The training setup is given in Table 3. The models were trained using AdamW as an optimizer. The Focal Loss was used as we were working on an imbalanced classification task. The models were trained for 80 epochs using an initial learning rate of 0.001 and a batch size of 64, and early stopping strategy was stepped in to prevent overfitting and improve model generalization. During the training phase, the model that demonstrated the highest performance on the validation set was selected for evaluation. The following metrics were used to evaluate the performance of models:

- Accuracy measures the proportion of correct predictions made by a model over the total number of predictions made and is defined as:

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

  where TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative respectively.
- Precision is a performance metric that measures the proportion of true positives over the total number of

TABLE 4. Performance comparison of VGG16, InceptionV3, ResNet50 and modified VGG16 with proposed modules, where A, P, R, and F stand for Accuracy, Precision, Recall, and F1-score respectively.

Models	40X				100X			
	A	P	R	F	A	P	R	F
VGG16	0.9335	0.91/0.95	0.89/0.96	0.9270	0.9296	0.90/0.93	0.85/0.96	0.9085
InceptionV3	0.8674	0.88/0.91	0.75/0.96	0.8701	0.8593	0.91/0.84	0.68/0.96	0.8368
ResNet50	0.9410	0.89/0.93	0.87/0.98	0.9323	0.9335	0.93/0.94	0.87/0.97	0.9257
Modified VGG16 w/ ConvMLP	0.9565	0.93/0.97	0.92/0.96	0.9457	0.9599	0.90/0.98	0.92/0.96	0.9559
Modified VGG16 w/ ConvMLP w/ CBAM	**0.9648**	**0.96/0.98**	**0.95/0.98**	**0.9650**	**0.9610**	**0.92/0.98**	**0.95/0.97**	**0.9635**
Models	200X				400X			
	A	P	R	F	A	P	R	F
VGG16	0.9296	0.91/0.93	0.88/0.95	0.9166	0.9023	0.87/0.92	0.81/0.95	0.8867
InceptionV3	0.8710	0.83/0.89	0.73/0.93	0.8416	0.9140	0.88/0.92	0.78/0.96	0.8824
ResNet50	0.8984	0.92/0.88	0.73/0.97	0.8677	0.8632	0.84/0.87	0.70/0.94	0.8320
Modified VGG16 w/ ConvMLP	0.9409	0.91/0.97	0.92/0.96	0.9417	0.9335	0.94/0.94	0.86/0.98	0.9294
Modified VGG16 w/ ConvMLP w/ CBAM	**0.9570**	**0.95/0.96**	**0.95/0.96**	**0.9596**	**0.9414**	**0.90/0.96**	**0.91/0.96**	**0.9344**

positive predictions made by a model and is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

- Recall is a performance metric that measures the proportion of true positives over the total number of actual positive instances and is defined as:

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

- Macro-F1 is the harmonic mean of precision and recall across all classes in a multi-class classification problem and can better reflect the performance of the classifier for imbalanced datasets. It is defined as:

$$\text{Macro-F1} = \frac{1}{k}\sum_{i=1}^{k} F1_i \tag{9}$$

where k is the number of classes, and $F1_i$ is the F1 score for the i^{th} class, defined as:

$$F1_i = 2 \cdot \frac{\text{precision}_i \cdot \text{recall}_i}{\text{precision}_i + \text{recall}_i} \tag{10}$$

where precision_i and recall_i are the precision and recall for the i^{th} class, respectively.

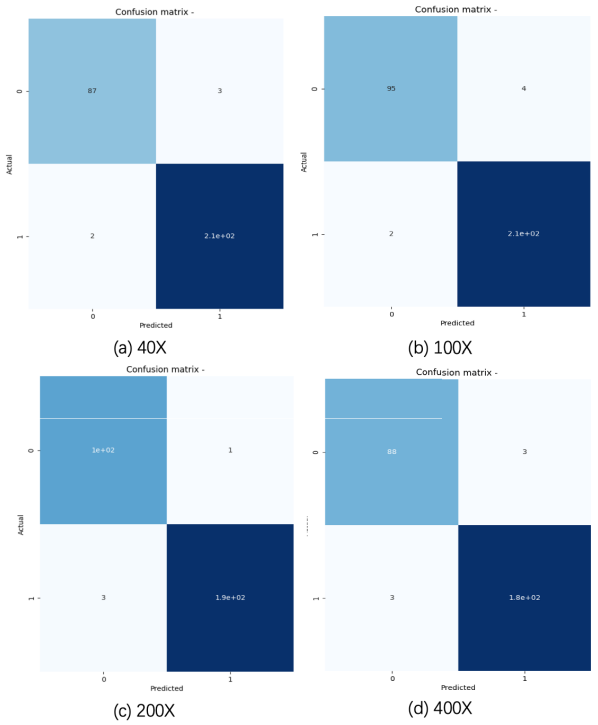

FIGURE 4. Confusion matrices of the proposed model for binary classification on the BreaKHis dataset.

V. RESULTS AND DISCUSSION

A. ABLATION STUDY

To construct the breast cancer classification model as proposed, a series of experiments were performed using an ablation study approach while taking into consideration the three main components of the model. Firstly, we sequentially added ConvMLP and CBAM to a modified VGG16 to

TABLE 5. Binary classification performance comparison of the proposed model with state-of-the-art networks on the BreaKHis dataset.

Technique	40X	100X	200X	400X
Spanhol et al. [35]	0.8960 ± 0.0650	0.8500 ± 0.0480	0.8400 ± 0.0320	0.8080 ± 0.0310
Spanhol et al. [10]	0.8460 ± 0.0290	0.8480 ± 0.0420	0.8420 ± 0.0170	0.8160 ± 0.0370
Bayramoglu et al. [12]	0.8300 ± 0.0300	0.8310 ± 0.0350	0.8460 ± 0.0270	0.8210 ± 0.0440
Zhu et al. [36]	0.8570 ± 0.0190	0.8420 ± 0.0320	0.8490 ± 0.0220	0.8010 ± 0.0440
Gupta et al. [37]	0.8674 ± 0.0237	0.8856 ± 0.0273	0.9031 ± 0.0376	0.8831 ± 0.0301
Deniz et al. [38]	0.9096 ± 0.0159	0.9058 ± 0.0196	0.9137 ± 0.0172	0.9130 ± 0.0740
Song et al. [39]	0.9002 ± 0.0302	0.9120 ± 0.0440	0.8780 ± 0.0530	0.8740 ± 0.0720
Gupta et al. [40]	0.9471 ± 0.0088	0.9590 ± 0.0420	0.9676 ± 0.0109	0.8911 ± 0.0012
Ours	0.9648 ± 0.0122	0.9610 ± 0.0033	0.9570 ± 0.0096	0.9414 ± 0.0101
Ours (fine-tuning)	$\mathbf{0.9844 \pm 0.0174}$	$\mathbf{0.9804 \pm 0.0098}$	$\mathbf{0.9702 \pm 0.0057}$	$\mathbf{0.9765 \pm 0.0044}$

examine whether the two modules have an impact on the classification results. We then compared the experimental results with the baselines, as shown in Table 4. It was observed that our model was performing better than the original version of VGG16 and other baselines.

As our dataset is small and imbalanced, we used transfer learning to improve classification accuracy, which means we used pre-trained weights to initialize the network's weights. The results of using pre-trained weights and training from scratch are given in Figure 3. We can observe a significant improvement in classification accuracy after using transfer learning, the accuracy improvements on VGG16, ResNet-50, Inception-v3, and Brea-Net are 2.54%, 4.86%, 5.72%, and 2.82%, respectively.

TABLE 6. Eight-class classification performance comparison of the proposed model with state-of-the-art networks on the BreakHis dataset.

Technique	40X	100X	200X	400X
Bardou et al. [13]	0.8037	0.6384	0.7454	0.5470
Sharma et al. [41]	0.9028	0.9010	0.8743	0.8655
Boumaraf et al. [42]	0.9449	0.9327	0.9129	0.8956
Joseph et al. [43]	0.9087	0.8957	0.9158	0.8867
Ours	**0.9531**	**0.9492**	**0.9532**	**0.9453**

TABLE 7. Performance comparison of the proposed model with some state-of-the-art models on the ICIAR dataset.

Technique	Number of classes	Accuracy
Golatkar et al. [44]	2	0.9300
Golatkar et al. [44]	4	0.8500
Awan et al. [45]	2	0.8700
Awan et al. [45]	4	0.8300
Guo et al. [46]	4	0.8750
Jawad et al. [47]	4	0.8948
Ours	4	**0.9550**

B. PERFORMANCE ANALYSIS AND COMPARISON

This section compares the proposed model with state-of-the-art deep learning approaches. The methods chosen for comparison are recent approaches that are used for breast cancer classification. Specifically, these methods are Spanhol et al. [10], [35], Bayramoglu et al. [12], Zhu et al. [36], Gupta et al. [37], [40], Deniz [38], Song et al. [39], Bardou et al. [13], Sharma et al. [41], Boumaraf et al. [42], and Joseph et al. [43]. The quantitative results of different models on all magnifications of the dataset are given in Table 5 and Table 6. Our model achieves the best accuracy of binary and multi-class classification on 40X, 100X, 200X, and 400X magnification.

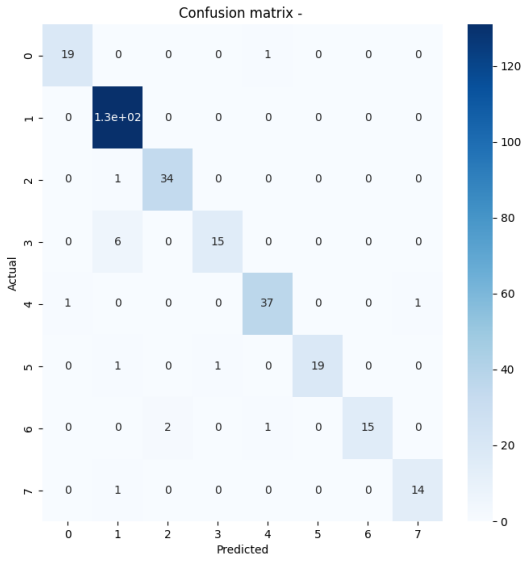

FIGURE 5. Confusion matrices of the proposed model for eight-class classification on the 40X BreaKHis dataset.

Our model generates confusion matrices for various magnifications of the dataset used in this experiment, as shown in Figure 4 and Figure 5. The misclassification rates of binary classification for 40X, 100X, 200X, and 400X magnification are 0.0005, 0.0082, 0.0082, and 0.0018, respectively. Figure 6 shows the changes in the loss and accuracy of the proposed model during the training phase. Notably, our model achieved the lowest false positive and false negative rates across all results. Figure 7 displays the ROC curves for all magnifications, which illustrate that the proposed model

has achieved a remarkable level of performance on the dataset.

To validate the generalization ability of the proposed network and to make its result more convincing, we tested our model on another histopathological dataset. In terms of the ICIAR 2018 dataset, the proposed network has obtained a competitive accuracy compared to some state-of-the-art methods in recent years, achieving 95.5% average accuracy. Table 7 presents a detailed performance of the proposed network and other methods on the ICIAR dataset.

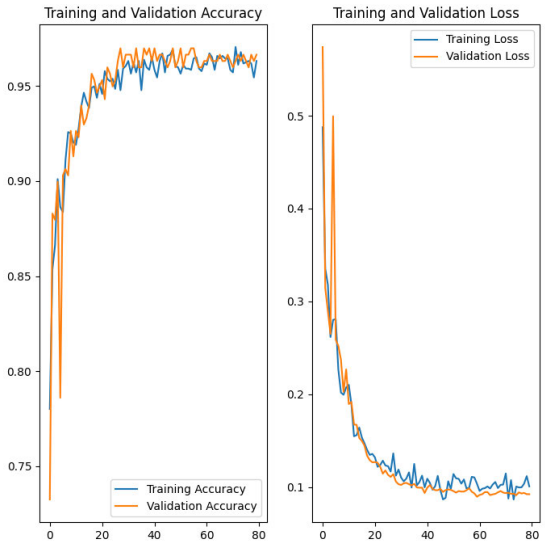

FIGURE 6. Accuracy and loss plot corresponding to the proposed network on the 40X BreaKHis dataset.

C. INTERPRETABILITY IN DEEP LEARNING

The proposed model, with its improved performance and interpretability, has the potential to enhance clinical practices in a number of ways. For example, it could aid in accurate diagnoses by providing more reliable and precise predictions, which in turn could lead to better treatment outcomes for patients. Additionally, the model could provide valuable insights to healthcare professionals, allowing them to better understand the underlying mechanisms of the disease or condition being studied. Furthermore, the improved interpretability of the model could also help to build trust and acceptance among healthcare professionals and patients, as it would be easier to understand and explain how the model arrived at its predictions.

In this section, we validate the interpretability by visualizing the weighted features using a CAM (Class Activation Mapping) based method, as shown in Figure 8, to understand which parts of the image the model relies on to make predictions. As we can observe, the Grad-CAM method has the ability to highlight all regions of a tumor, providing visual results that are more intuitive and easier to comprehend.

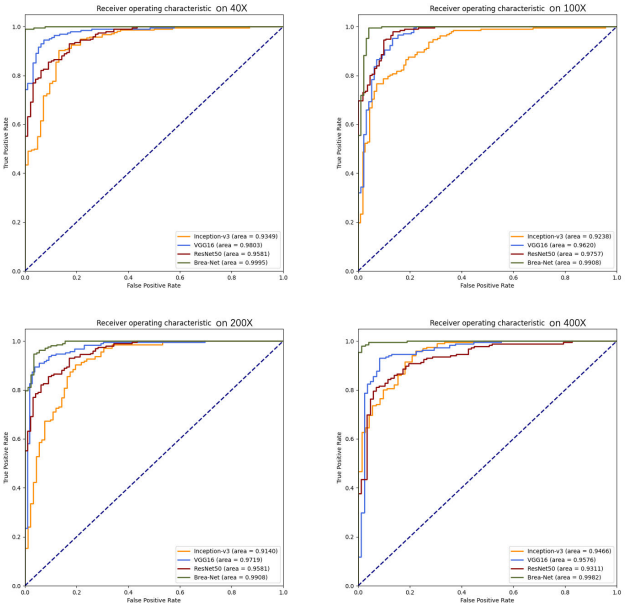

FIGURE 7. ROC curve comparison of the proposed approach with state-of-the-art networks in different magnifications.

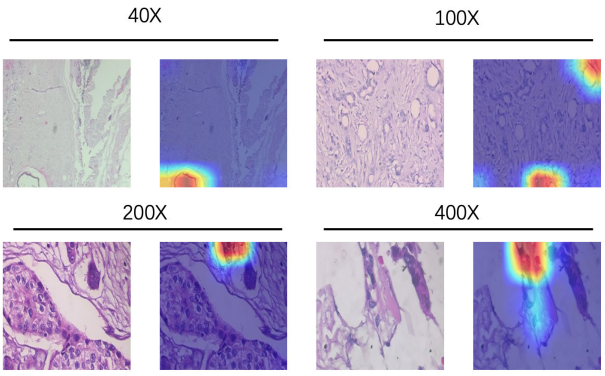

FIGURE 8. Visual interpretations of breast cancer classification prediction based on Grad-CAM.

VI. CONCLUSION AND FUTURE SCOPE

Our study introduces "Brea-Net," a network designed for breast cancer classification from imbalanced datasets that is both efficient and interpretable. The model consists of three main components. Initially, a pre-trained VGG16 image encoder is employed to extract the most significant features from breast cancer images. Next, CBAM is used to effectively capture and utilize the discriminative spatial and channel features within histopathological images. Finally, the refined feature is sent to a ConvMLP for the ultimate prediction.

A series of evaluation tests suggest that our model outperformed existing approaches in terms of accuracy, precision, recall, and F1-score, and achieved outstanding binary accuracy at magnifications of 40X, 100X, 200X, and 400X, with scores of 98.44%, 98.04%, 97.02%, and 97.65% respectively. In terms of multi-class classification, our model also demonstrated outstanding accuracy rates of 95.31%,

94.92%, 95.32%, and 94.53% for different magnifications. We validated the generalization ability of the model by obtaining a consistent and remarkable performance on a different histopathological dataset (ICIAR). The interpretability analysis result indicates that our model can produce reasonable and explainable predictive outcomes.

Most of the current interpretability research is still at a shallow level of visualizing features, and very few can explain results from a causal perspective. However, interpretability is of great significance in many industries, especially in special fields such as healthcare. Therefore, in future work, we will continue to explore methods to explain results from a causal perspective.

Explaining the prediction of a deep learning model from a causal perspective can be challenging, as deep learning models often involve complex and highly nonlinear relationships between the input and output variables. However, there are some methods that can be explored to explain the prediction of a deep learning model from a causal perspective, such as counterfactual analysis and attribution analysis. Counterfactual analysis involves examining how the output of a deep learning model would change if the input were altered in a specific way. By comparing the output of the model under different input conditions, researchers can identify the causal relationships between the input and output variables. On the other hand, attribution analysis involves identifying which input features or neurons in a deep learning model are most responsible for a given output. By understanding the specific features that contribute to the output, researchers can gain insight into the underlying causal relationships.

REFERENCES

[1] K. E. Lukong, "Understanding breast cancer—The long and winding road," *BBA Clin.*, vol. 7, pp. 64–66, Jan. 2017.

[2] H. D. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: A survey," *Pattern Recognit.*, vol. 43, no. 1, pp. 299–317, Jan. 2010.

[3] A. Eklund, P. Dufort, D. Forsberg, and S. M. LaConte, "Medical image processing on the GPU—Past, present and future," *Med. Image Anal.*, vol. 17, no. 8, pp. 1073–1094, Dec. 2013.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[5] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.

[6] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS Int. Trans. Comput. Sci. Eng.*, vol. 30, no. 1, pp. 25–36, 2006.

[7] 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.

[8] S. M. Mathews, "Explainable artificial intelligence applications in NLP, biomedical, and malware classification: A literature review," in *Proc. Comput. Conf.*, 2019, pp. 1269–1292.

[9] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," *Eur. J. Oper. Res.*, vol. 267, no. 2, p. 68799, 2008.

[10] F. A. Spanhol, L. S. Oliveira, P. R. Cavalin, C. Petitjean, and L. Heutte, "Deep features for breast cancer histopathological image classification," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2017, pp. 1868–1873.

[11] K. Nazeri, A. Aminpour, and M. Ebrahimi, "Two-stage convolutional neural network for breast cancer histology image classification," in *Proc. Int. Conf. Image Anal. Recognit.*, P de Varzim, Portugal, 2018, pp. 717–726.

[12] N. Bayramoglu, J. Kannala, and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Cancun, Mexico, Dec. 2016, pp. 2440–2445.

[13] D. Bardou, K. Zhang, and S. M. Ahmad, "Classification of breast cancer based on histology images using convolutional neural networks," *IEEE Access*, vol. 6, pp. 24680–24693, 2018.

[14] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," in *Proc. 11th Int. Conf. Inf. Commun. Syst. (ICICS)*, Irbid, Jordan, Apr. 2020, pp. 243–248.

[15] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, p. 148, Dec. 2019.

[16] M. Saini and S. Susan, "Deep transfer with minority data augmentation for imbalanced breast cancer dataset," *Appl. Soft Comput.*, vol. 97, Dec. 2020, Art. no. 106759.

[17] W. Ding, D.-Y. Huang, Z. Chen, X. Yu, and W. Lin, "Facial action recognition using very deep networks for highly imbalanced class distribution," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Kuala Lumpur, Malaysia, Dec. 2017, pp. 1368–1372.

[18] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "DeTrac: Transfer learning of class decomposed medical images in convolutional neural networks," *IEEE Access*, vol. 8, pp. 74901–74913, 2020.

[19] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1520–1528.

[20] J. Tobias Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014, *arXiv:1412.6806*.

[21] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.

[22] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognit.*, vol. 65, pp. 211–222, May 2017.

[23] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.

[24] P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, and S. Dähne, "Learning how to explain neural networks: PatternNet and PatternAttribution," 2017, *arXiv:1705.05598*.

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

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

[27] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, and G. Fernandez, "BACH: Grand challenge on breast cancer histology images," *Med. Image Anal.*, vol. 56, pp. 122–139, Aug. 2019.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

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

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

[31] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, p. 319.

[32] J. Li, A. Hassani, S. Walton, and H. Shi, "ConvMLP: Hierarchical convolutional MLPs for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 6306–6315.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[34] 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. (CVPR)*, Jun. 2016, pp. 2818–2826.
[35] 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. (IJCNN)*, Jul. 2016, pp. 2560–2567.
[36] C. Zhu, F. Song, Y. Wang, H. Dong, Y. Guo, and J. Liu, "Breast cancer histopathology image classification through assembling multiple compact CNNs," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, pp. 1–17, Dec. 2019.
[37] V. Gupta and A. Bhavsar, "Breast cancer histopathological image classification: Is magnification important?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 769–776.
[38] E. Deniz, A. Şengür, Z. Kadiroğlu, Y. Guo, V. Bajaj, and Ü. Budak, "Transfer learning based histopathologic image classification for breast cancer detection," *Health Inf. Sci. Syst.*, vol. 6, no. 1, p. 17, Dec. 2018.
[39] Y. Song, H. Chang, H. Huang, and W. Cai, "Supervised intra-embedding of Fisher vectors for histopathology image classification," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Quebec City, QC, Canada, 2017, pp. 99–106.
[40] V. Gupta and A. Bhavsar, "Sequential modeling of deep features for breast cancer histopathological image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Salt Lake City, UT, USA, Jun. 2018, pp. 2254–2261.
[41] S. Sharma and R. Mehra, "Conventional machine learning and deep learning approach for multi-classification of breast cancer histopathology images—A comparative insight," *J. Digit. Imag.*, vol. 33, no. 3, pp. 632–654, Jun. 2020.
[42] S. Boumaraf, X. Liu, Z. Zheng, X. Ma, and C. Ferkous, "A new transfer learning based approach to magnification dependent and independent classification of breast cancer in histopathological images," *Biomed. Signal Process. Control*, vol. 63, Jan. 2021, Art. no. 102192.
[43] A. A. Joseph, M. Abdullahi, S. B. Junaidu, H. H. Ibrahim, and H. Chiroma, "Improved multi-classification of breast cancer histopathological images using handcrafted features and deep neural network (dense layer)," *Intell. Syst. Appl.*, vol. 14, May 2022, Art. no. 200066.
[44] A. Golatkar, D. Anand, and A. Sethi, "Classification of breast cancer histology using deep learning," in *Proc. Int. Conf. Image Anal. Recognit.*, Poa de Varzim, Portugal, 2018, pp. 837–844.
[45] R. Awan, N. A. Koohbanani, M. Shaban, A. Lisowska, and N. Rajpoot, "Context-aware learning using transferable features for classification of breast cancer histology images," in *Proc. Int. Conf. Image Anal. Recognit.*, Pvoa de Varzim, Portugal, 2018, pp. 788–795.
[46] Y. Guo, H. Dong, F. Song, C. Zhu, and J. Liu, "Breast cancer histology image classification based on deep neural networks," in *Proc. Int. Conf. Image Anal. Recognit.*, Pvoa de Varzim, Portugal, 2018, pp. 827–836.
[47] M. A. Jawad and F. Khursheed, "Deep and dense convolutional neural network for multi-category classification of magnification specific and magnification independent breast cancer histopathological images," *Biomed. Signal Process. Control*, vol. 78, Sep. 2022, Art. no. 103935.

YI LIANG received the B.S. degree in software engineering from South-Central Minzu University, Wuhan, China, in 2018. He is currently pursuing the Graduate degree with Guangxi University, Nanning, China. His primary research interests include computer-aided diagnosis, dense object detection, and multi-modal feature fusion.

ZUQIANG MENG received the Ph.D. degree in computer application technology from Central South University, Changsha, China, in 2004. He qualified as a Postdoctoral Fellow with the Institute of Computing Technology, Chinese Academy of Sciences, in 2009. From March 2015 to March 2016, he conducted a one-year visit and exchange program with the University of Kansas, USA. He is currently a Professor with the School of Computer and Electronic Information, Guangxi University. His research interests include interpretable artificial intelligence, fine-grained image classification, and multi-modal feature fusion.