Breast Cancer Classification Based on Various CNNs and Classifiers

Yuchen Ge[1, †]
[1]Shandong University
Jinan, China
gycdwwd@gmail.com

Longxin Wang[3,†]
[3] Xidian University
Xi'an, China
wlx2096prince@gmail.com

Kejia Liu[2, †]
[2]University of British Columbia
Vancouver, Canada
kej19@student.ubc.ca

Qianyi Xue[4,*, †]
[4]Sichuan University
Chengdu, China
[*]Corresponding author's e-mail: xueqianyi@stu.scu.edu.cn
[†]These authors contributed equally.

Abstract*—Breast cancer is the second leading cause of death from cancer in women around the world. The CAD system utilizing machine learning and deep learning techniques facilitates the early detection of breast cancers. However, few recent studies focused on utilizing multiple feature extractors to compare and analyze the performances of various architectures. This paper analyzes the performances of architectures which are combinations of different feature extractors and classifiers in breast cancer diagnosis. Firstly, we collected histopathological breast cancer images from the BreakHis dataset. Secondly, the normalized data were converted to one-hot encoding for training, validating, and testing. Thirdly, we used VGG-16, VGG-19, Xception, ResNet50, Inception-V3, and Inception-Resnet-V2 to extract features. Next, fully connected layer (FCL), logistic regression (LR), and SVM were employed to classify breast cancers on the BreaKHis dataset. The experimental result shows that with the cyclical learning rate (CLR) policy, the ResNet50-SVM model obtained the optimal accuracy rate of 93.9% on eight-classification. The result shows that our proposed method could diagnose breast cancer with high accuracy.

***Keywords*—ResNet50, Inception-V3, Inception-Resnet-V2, SVM, Medical Imaging, Breast Cancer, Cyclical Learning Rate**

I. INTRODUCTION

In women worldwide, breast cancer (BC) is the top common cancer and the second leading cause of cancer death [1]. The significant number of victims of this fatal cancer constantly reminds individuals that methods capable of improving the patient recovery rate are desperately needed. Improving BC patient survival requires early detection, which is feasible with an efficient diagnosis system. Computer-aided diagnosis (CAD) systems are such systems, which can efficiently detect and diagnose cancer from histopathological images compared with the time-consuming visual examination of the cancerous cell by an expert pathologist [2]. The CAD system can identify whether a tumor is benign or malignant faster via machine learning (ML) and deep learning (DL) techniques with high accuracy, thus benefiting BC patients from receiving corresponding treatment on time to reduce complications and forestall aggravation.

Common tasks with CAD systems include feature extraction and classification [3]. Feature extraction transforms the input data into a set of the most relevant and informative features from the original data. In addition, feature extraction is crucial to the achievement of reliable classification [4]. Selecting inapplicable features might lead to the degradation of the performance of classification. Classification of images is a process of comprehending each image as a unit and categorizing the images by assigning them to a specific label.

There are numerous research works related to the classification of BC images with various feature extraction and classification approaches. Senan et al. [5] estimated the malignancy of the BC images from the BreaKHis dataset at different magnifications. They used AlexNet as the feature extractor and classifier, achieving an accuracy of 95% for binary classification [5]. However, they did not carry out multi-class classification on this dataset. The study by Nguyen et al. [6] bridging this gap constructed a CNN as both feature extractor and classifier for multi-class classification based on the BreakHis dataset. For this study, a multi-class classification accuracy of 73.68% was obtained [6]. However, these two studies were limited to using the same model as the feature extractor and classifier. On the contrary, Kate & Shukla [7] employed VGG-16 as a feature extractor and SVM as a classifier, rather than using only one model for both feature extraction and classification tasks. They achieved an accuracy of 98.2% on binary classification and 93.6% on multi-class classification [7]. Regarding binary classification, the performance of the architecture of Kate & Shukla [7] outshone that of Senan et al. [5]. In addition, regarding multi-classification, Kate & Shukla [7] obtained a higher accuracy than Nguyen et al. [6]. According to these comparisons, feeding the features to a different classifier will produce a result with higher accuracy. Consequently, the present study sought to classify the BC images with higher accuracy by feeding the features to a model different from what is used for feature extraction to conduct classification.

With regard to classification techniques, plenty of studies investigated the performance of ML approaches and DL methods. Said Boumaraf et al. [8] compared the effectiveness of ML and DL techniques on the classification of the BreakHis dataset and came to some conclusions: (1) The optimal result of eight classifications using KNN, RF, MLP, AdaBoost, and SVM for the extracted Zernike

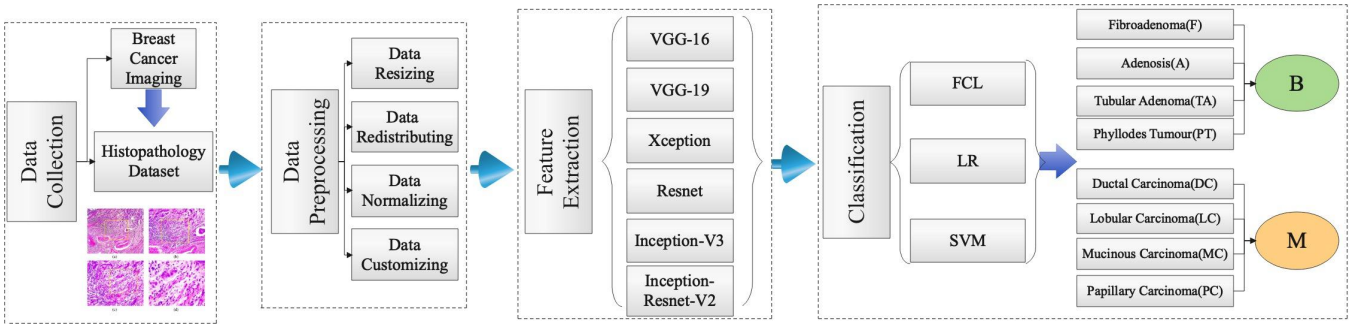

Fig. 1. System architecture

Moments Features, Haralick Features, and Color Histogram Features can only reach 69.69%. (2) The eight-classification accuracy of 88.95% was obtained using different fine-tuned blocks of VGG-19. However, they did not employ the combined use of ML and DL techniques for medical image classification. Saleh et al. [9] constructed a compound of CNN and SVM to classify lung CT images with the help of the fully connected layer using a modified SVM architecture. They obtained a higher accuracy than using CNN alone, at 97.91% [9]. The example of combining ML and DL has been reflected in the classification of lung cancer image data. Additionally, Jasti et al. [10] used AlexNet to extract features and LS-SVM, KNN, Random Forest, and Naive Bayes as classifiers, concluding that the accuracy of LS-SVM was higher than the other classifiers. Moreover, Liew et al. [11] used a pre-trained DenseNet201 model for feature extraction of images. They replaced the Fully Connected Layer (FCL) with an XGBoost classifier, which achieved an average accuracy of 97% for both binary and multi-classification [11]. The above-proposed models perform well in different situations, and the comparisons performed in the same experimental environment are the most effective. The main contributions of this paper are as follows.

We constructed a more comprehensive CAD system utilizing combinations of various feature extractors and classifiers to assist breast cancer diagnosis based on the BreaKHis dataset [12]. Some well-performing models (VGG-16, VGG-19, Xception, ResNet50, Inception-V3, and Inception-Resnet-V2) are used as feature extractors, Fully Connected Layer (FCL), LR, and SVM models are used as classifiers. We maximized the use of the BreaKHis dataset [12] and performed a comparison of the accuracy between all models on a binary classification task and an eight-classification task. The binary classification task is to distinguish the benign and the malignant in the dataset, and the 8-classification task is to distinguish eight specific cancer types. Finally, we selected the best model for each sub-task. We conducted various fine-tuning tests on the model, such as adding a fully connected layer to test whether the model's performance improved. We modified the training method, using the CLR policy. The results demonstrated that this method could improve the performance of our proposed methods.

II. METHOD

This section first introduces the architecture of our model (Sec. A). Then, we show the data pre-processing steps(Sec. B). Finally, the models used in this paper are given (Sec. C).

A. Approach Overview

This section describes our proposed system architecture composed of data collection, data pre-processing, feature extraction, and classification modules. We used the BreaKHis dataset [12] in this paper. We first resized, redistributed, normalized, and customized the data. In the following module, six models, VGG-16, VGG-19, Xception, ResNet50, Inception-V3, and Inception-Resnet-V2 were used for feature extraction. In the classification module, FCL, LR, and SVM were used to classify images into eight classes that correspond to benign and malignant types. The overview of our system architecture is displayed in Fig. 1.

B. Data Pre-processing

The Breast Cancer Histopathological Image Classification (BreaKHis) [12] was used in this paper. The BreaKHis consists of 7909 histopathology images of breast tumor tissue from 82 patients using four magnifying factors, 40X, 100X, 200X, and 400X. For the binary classification purpose, the tumors in the BreaKHis are categorized into two types: benign and malignant. Out of 7909 images, there are 2480 images of benign cancers, while 5429 images belong to the malignant class. All images in this dataset are in three-channel RGB of size 700 × 460 pixels. To further conduct multi-class classification, both benign cancers and malignant cancers are divided into four subtypes respectively. The four benign breast tumors are adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenoma (TA), while the four malignant are ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC). Table I gives detailed information about the distribution of these eight sub-types of BC corresponding to each magnification factor.

To make the data suitable for our model, we resized, redistributed, normalized, and customized the data. Firstly, we resized the images to 115×175, reducing the size of the image dimension by a factor of four. Using this size, we could keep the images readable and improve computational efficiency. Secondly, we redistributed the data samples of the training dataset, which is used in the feature extraction module. Thirdly, we normalized the batch data and encoded it using one-hot so that it can be used for training, validation, and testing. Finally, we customized the input layers of feature extraction models to accommodate the images with the input shape 115×175×3.

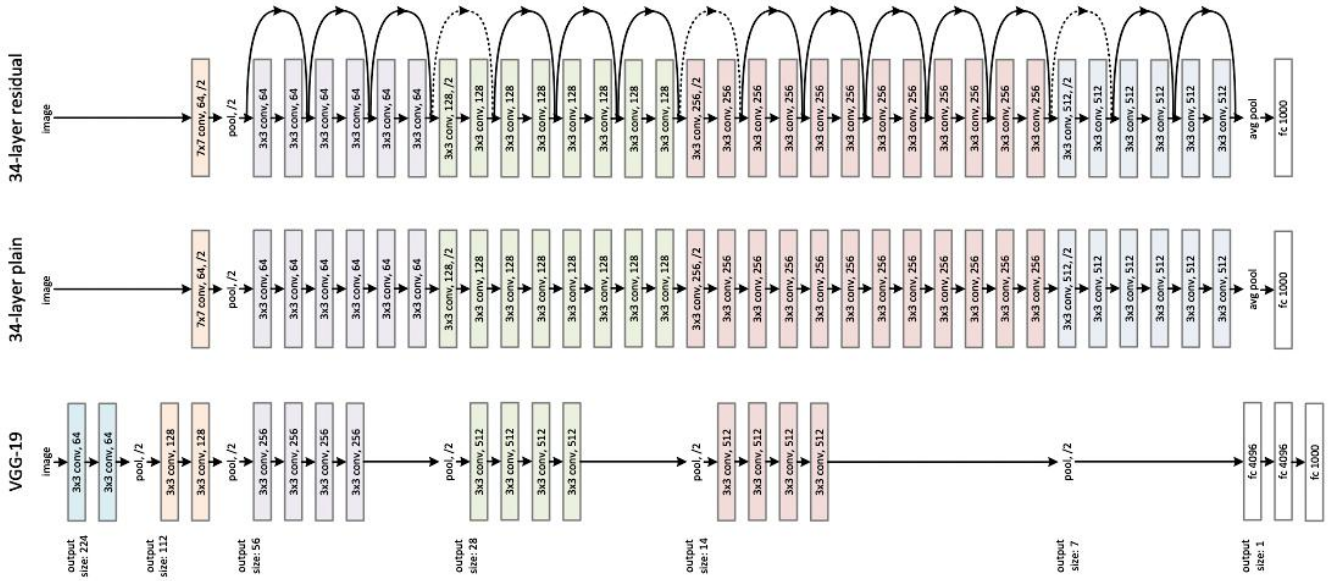

Fig. 2. The architecture of different modern CNNs [16]

TABLE I. BC IMAGES DISTRIBUTION BY MAGNIFICATION FACTOR AND HISTOLOGICAL SUBTYPES

Type	Subtype	40X	100X	200X	400X	Total
Benign	A	114	113	111	106	444
	F	253	260	264	237	1014
	PT	109	121	108	115	453
	TA	149	150	140	130	569
Malignant	DC	864	903	896	788	3451
	LC	156	170	163	137	626
	MC	205	222	196	169	792
	PC	145	142	135	138	560
Total		1995	2081	2013	1820	7909

According to Table I, the eight subtypes of BC are not evenly distributed. Fibroadenomas (F) and ductal carcinomas (DC) dominate all the other subtypes in the dataset. Fibroadenoma images together with ductal carcinoma images are roughly 56.5% of the entire dataset. To be more specific, fibroadenoma tumors account for approximately 40.9% of the benign cancers in the BreaKHis, whereas the proportion of ductal carcinoma tumors to all malignant tumors in the dataset is about 63.6%.

C. Models

LeNet [13] is designed to process the mist handwritten digit database. After ResNet, different teams have improved ResNet and proposed models such as VGG-16 [14], VGG-19 [14], Xception [15], ResNet [16], Inception-V3 [17], and Inception-Resnet-V2 [18]. In the paper, we used VGG-16, VGG-19, Xception, ResNet50, Inception-V3, and Inception-Resnet-V2 to extract features, FCL, LR, and SVM as classifiers. VGG-16 and VGG-19 consist of 16 and 19 convolutional layers, respectively. One of the crucial downsides of the VGG16 network is that it is a huge network, which means that it takes more time to train its parameters. ResNet consists of a stack of basic blocks, as shown in Fig. 3. A basic block is to add a shortcut connection to convolutional layers so that the network can contain both nonlinear and linear components.

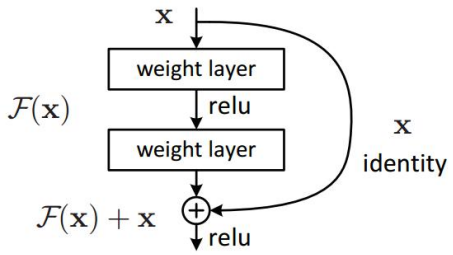

Fig. 3. Basic block in ResNet [16]

As shown in (1), ResNet is not to fit the initial distribution H(x) but to fit the residual function $F(x)$, which is defined in (1):

$$F(x) = H(x) - x \quad (1)$$

After solving the network degradation problem, we increase the network depth by stacking more basic blocks to achieve better performance. As shown in Fig. 2, it is the architecture of 34-layer ResNet and other networks.

Instead of adding layers to make the networks deeper with higher computation costs, Inception-V3 [17] filters with multiple sizes on the same level by applying the naive inception module. The different sizes of filters are 1×1, 3×3, and 5×5 convolutions. Then we concatenated the outputs and sent them to the next layer.

The second main idea of the proposed architecture is to reduce dimensions. Accordingly, the number of input channels is constrained to adding an extra 1×1 convolution before the expensive 3×3, and 5×5 convolutions. It is the basic module that builds up the Inception-V3 model. The entire structure of Inception-V3 is shown in Fig. 4. It is 27 layers deep (counting pooling layers) with nine basic inception modules (black boxes). A global average pooling layer (red block) and a softmax layer (yellow block) are at the end of the whole structure.

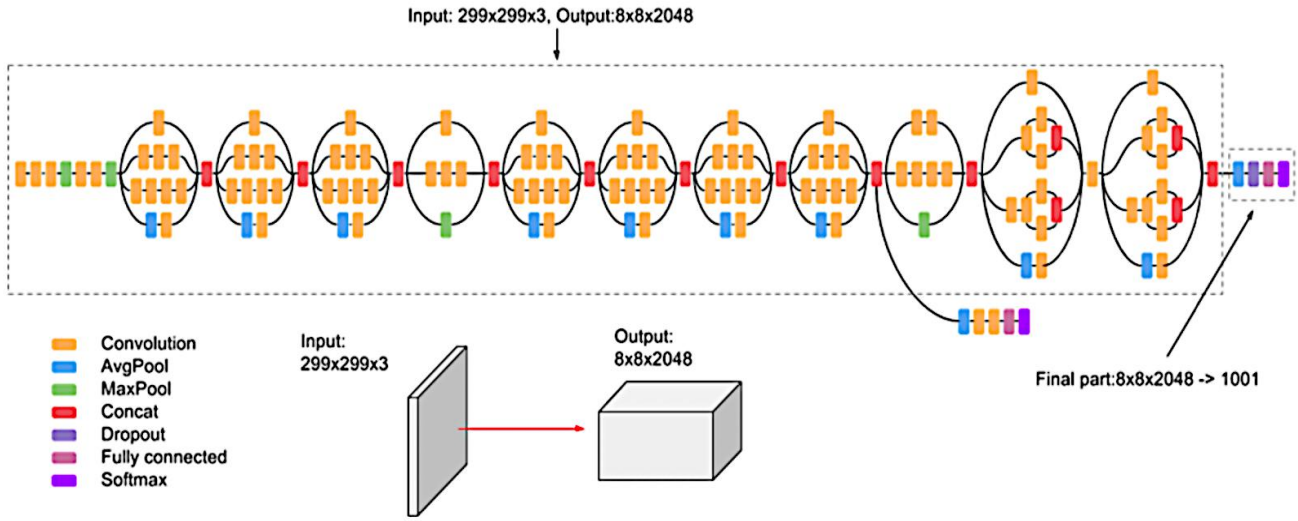

Fig. 4. Inception-V3 architecture [17]

In this paper, we used Inception-ResNet-V2 [18] to reduce dimensions. These two models not only reduce computation time but also prevent degradation problems resulting from deep structures. Inception-V3 contains fewer parameters (7 million) than VGGNet (about 200 million) and AlexNet [19] (60 million). The only difference between Inception-ResNet-V1 and V2 is the hyper-parameter settings. The Inception-ResNet-V2 model has a relatively lower computation cost but higher accuracy. Consequently, Inception-ResNet-V2 became a widespread network for classification tasks in practice.

D. Cyclical Learning Rate

Cyclical Learning Rate is a method letting the learning rate vary in a reasonable range cyclically. CLR was demonstrated to improve the performance of various architectures substantially, such as ResNet, Stochastic Depth network, DenseNet, AlexNet, and Inception on the datasets CIFAR-10 and CIFAR-100 [20]. Inspired by Smith [20], we used the CLR method in our work to enhance the performance of the proposed architectures.

E. Training and Evaluation

We trained the feature extraction models and classifiers separately. When training the feature extraction models (VGG-16, VGG-19, Xception, ResNet50, Inception-V3, Inception-ResNet-V2), the parameters are batch_size=64, epoch=100, lr=0.00006. We used the early stopping strategy to prevent overfitting, saving CheckPoints, and used ReduceLROnPlateau to adjust the learning rate dynamically. After training the classifiers (FCL, LR, SVM), the accuracy, precision, and recall of test sets are calculated. For each combined model at each magnification (40X, 100X, 200X, 400X), the training was repeated five times (i.e., CLR) from which the best performing model was finally selected.

To evaluate different combined models, we employed precision-recall analysis. Due to the imbalanced data in each class (DC contains 3451 samples and PT contains 453 samples), we compared the model's precision and recall simultaneously.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2)$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (3)$$

III. Result

This section first shows our results and analysis of all the related models (Sec. A). Then, we give a further discussion about our work in this paper (Sec. B).

A. Results and Analysis

TABLE II. THE ACCURACY OF THE FCL COMBINED MODEL (8-CLASS)

FCL (Fully Connected Layer) ACCURACY (%)				
	40X	**100X**	**200X**	**400X**
VGG16	90.8	89.6	89.9	83.3
VGG19	82.4	85.8	83.4	80.7
Xception	81.7	79.0	78.4	75.4
ResNet50	89.2	91.2	**89.9**	87.3
InceptionV3	81.0	78.3	79.7	76.9
InceptionResNetV2	87.3	86.4	85.2	79.6

TABLE III. THE ACCURACY OF THE LR COMBINED MODEL (8-CLASS)

LR (Logistic Regression) ACCURACY (%)				
	40X	**100X**	**200X**	**400X**
VGG16	**92.2**	89.6	90.5	83.7
VGG19	86.8	87.1	86.8	84.8
Xception	86.1	**85.8**	**85.5**	**77.3**
ResNet50	92.5	89.3	86.8	87.5
InceptionV3	87.8	82.8	**84.1**	**76.9**
InceptionResNetV2	89.2	86.4	88.2	82.6

TABLE IV. THE ACCURACY OF THE SVM COMBINED MODEL (8-CLASS)

SVM ACCURACY (%)				
	40X	**100X**	**200X**	**400X**
VGG16	91.9	**91.3**	**90.9**	**85.2**
VGG19	**90.8**	**89.0**	**90.2**	**88.2**
Xception	**87.5**	85.1	83.1	73.9
ResNet50	**93.9**	**91.3**	89.2	**87.9**
InceptionV3	**88.1**	**85.4**	83.4	76.5
InceptionResNetV2	**90.5**	**86.7**	**89.5**	**84.8**

Tables Ⅱ-Ⅳ present the accuracy rates of all the related models. The best model is ResNet50, which has an average accuracy of 90.8%. The SVM is the best classifier for improving accuracy rates. It has 17 of the 24 configurations

(4 magnification levels ×6 feature extractors) that make the model reach the optimal solution (bold parts in the table). The LR classifier performs the best in 6 cases. FCL is the weakest among them.

TABLE V. ACCURACY AND RECALL OF THE SVM ON 100X SAMPLES

Accuracy and recall of SVM on 100X (3 out of 5 sets)						
	1		2		3	
	Accuracy (%)	Recall (%)	Accuracy (%)	Recall (%)	Accuracy (%)	Recall (%)
VGG16	87.1	83.5	87.7	83.3	**91.3**	**88.3**
VGG19	87.1	83.9	**89.0**	**86.1**	87.7	83.6
Xception	81.2	75.4	**85.1**	**80.1**	84.5	78.6
ResNet50	90.3	**86.9**	88.7	84.5	**91.3**	86.7
Inception V3	79.9	71.6	82.5	76.8	**85.4**	**81.1**
Inception-ResNetV2	84.8	79.6	83.2	77.5	**86.7**	**82.0**

The performance of the SVM combination model at 100X is shown in Table Ⅴ (2 sets of experiments are omitted). The precision and recall rates do not reach the maximum value at the same time. In medical diagnosis, a False Negative (FN) is a patient with cancer who has not been diagnosed. The model should minimize FN, although this may sacrifice precision. The global maximum recall under this condition is 88.3%. Combining precision and recall, VGG-16 is the optimal model under this condition.

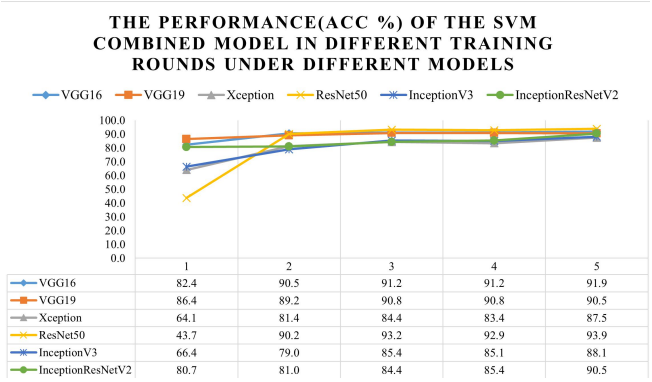

Fig. 5. The performance (Accuracy %) of the SVM combined model in different training rounds under different models

Fig. 5 presents the results of continuous training of the SVM combined model at 40X. The strategy adopted during training is the CLR. The initial learning rate is fixed for five rounds of circular training and the subsequent nonlinear decreasing learning rate. Fig. 5 confirms the necessity of continuous training. With the increase of training rounds, the accuracy of the SVM combination model on the validation set also increases and eventually approaches saturation.

TABLE VI. THE ACCURACY OF THE FCL COMBINED MODEL (2-CLASS)

FCL (Fully Connected Layer) ACCURACY (%)				
	40X	**100X**	**200X**	**400X**
VGG16	94.6	92.0	90.6	**98.6**
VGG19	88.6	93.	95.3	98.6
Xception	78.3	69.5	94.7	95.9
ResNet50	95.2	96.6	95.9	97.3
InceptionV3	99.4	97.1	98.2	**98.6**
InceptionResNetV2	97.0	97.7	95.9	97.3

We selected two of the eight types of breast cancer tumors, fibroadenoma (benign) and ductal carcinoma (malignant). The accuracy rates of the two breast cancer types are shown in Table Ⅵ. These two breast cancers were selected as they are representative. Compared with the results of the eight-class classification (Table Ⅱ), binary classification can produce higher accuracy rates.

B. Discussion

In this study, the six employed models, VGG-16, VGG-19, Xception, ResNet50, Inception-V3, and Inception-Resnet-V2, were trained from scratch. These models were initialized with new parameters, and thus required more computational resources and a large dataset to train. On the contrary, using pre-trained models is timesaving as features, weights and other parameters can be obtained from pre-trained models with transfer learning. In other words, pre-trained models are closer to converging as the weights have already been optimized. However, this approach requires the transferred information about the extracted features to be general in scope such that it works for both base and target tasks; otherwise, the performance may not be optimal. In comparison, when the dataset is large enough, the models training from scratch learn all parameters from training data, thus being expected to perform better.

Using XGBoost is another method that may improve the performance of our proposed architectures. There is a data imbalanced issue in the BreaKHis dataset, which is illustrated in the Table I. Tackling imbalanced data, XGBoost undersamples, oversamples, and adjusts the proportion of sampling data in each class, and thus helps avoid overfitting and bias in the models.

IV. CONCLUSION

This paper employed six feature extraction models (VGG-16, VGG-19, Xception, ResNet50, Inception-V2, and Inception-Resnet-V3) and three classifiers (FCL, LR, SVM) to build combined models. Among all proposed architectures, the ResNet50-SVM model obtained the best accuracy rate, 93.9%. The SVM classifier performed significantly better than the LR and FCL classifiers on the majority of the configurations. Compared with the eight-classification task, the binary classification task produced a significantly better accuracy. During training, CLR continuously enhanced the accuracy of each model, which implies the significance of CLR. Finally, we presented the classification results of every combined model at each magnification level in tables and the models with the best performance.

In the future, studies can focus on investigating whether using different methods of setting learning rates can enhance the performance of our proposed architectures on the BreaKHis dataset.

REFERENCES

[1] R. L. Siegel, K. D. Miller, S. A. Fedewa, D. J. Ahnen, R. G. Meester, A. Barzi, and A. Jemal, "Colorectal cancer statistics, 2017," CA: A Cancer Journal for Clinicians, vol. 67, no. 3, pp. 177–193, March 2017.

[2] M. Gour, S. Jain, and T. S. Kurmar, "Residual learning based CNN for breast cancer histopathological image classification," International Journal of Imaging Systems and Technology, vol. 30, no. 3, pp. 621–635, February 2020.

[3] X. Li, C. Li, M. M. Rahaman, H. Sun, X. Li, J. Wu, Y. Yao, and M. Grzegorzek, "A comprehensive review of computer-aided whole-slide image analysis: From datasets to feature extraction, segmentation, classification and detection approaches," Artificial Intelligence Review, Jan. 2022.

[4] A. Subasi, "Feature Extraction and Dimension Reduction," in Practical Guide for Biomedical Signals Analysis Using Machine Learning Techniques, London: Academic Press, 2019, pp. 193–275.

[5] E. M. Senan, F. W. Alsaade, M. I. A. Al-mashhadani, T. H. H. aldhyani, and M. H. Al-Adhaileh, "Classification of histopathological images for early detection of breast cancer using Deep Learning," Journal of Applied Science and Engineering, Jan 2021.

[6] P. T. Nguyen, T. T. Nguyen, N. C. Nguyen, and T. T. Le, "Multiclass breast cancer classification using Convolutional Neural Network," 2019 International Symposium on Electrical and Electronics Engineering (ISEE), 2019.

[7] V. Kate and P. Shukla, "Breast Cancer Image MultiClassification Using Random Patch Aggregation and Depth-Wise Convolution based Deep-Net Model," 2021.

[8] S. Boumaraf, X. Liu, Y. Wan, Z. Zheng, C. Ferkous, X. Ma, Z. Li, and D. Bardou, "Conventional machine learning versus deep learning for magnification dependent histopathological breast cancer image classification: A comparative study with visual explanation," MDPI, 16-Mar-2021.

[9] A. Y. Saleh, C. K. Chin, V. Penshie, and H. R. H. Al-Absi, "Lung Cancer Medical Images classification using hybrid CNN-SVM," International Journal of Advances in Intelligent Informatics.

[10] V. D. P. Jasti, A. S. Zamani, K. Arumugam, M. Naved, H. Pallathadka, F. Sammy, A. Raghuvanshi, and K. Kaliyaperumal, "Computational technique based on machine learning and image processing for medical image analysis of breast cancer diagnosis," Security and Communication Networks, 09-Mar-2022.

[11] X. Y. Liew, N. Hameed, and J. Clos, "An investigation of XGBoost-based algorithm for Breast Cancer Classification," Machine Learning with Applications, 08-Sep-2021.

[12] F. A. Spanhol, L. S. Oliveira, C. Petitjean and L. Heutte, "A Dataset for Breast Cancer Histopathological Image Classification," in IEEE Transactions on Biomedical Engineering, vol. 63, no. 7, pp. 1455-1462, July 2016.

[13] LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R. Hubbard, W., Jackel, L., "Handwritten digit recognition with a back-propagation network", Advances in neural information processing systems 2, 1989, pp. 396–404.

[14] Simonyan, Karen, and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." ArXiv.org, 10 Apr. 2015, https://arxiv.org/abs/1409.1556.

[15] Chollet, François. "Xception: Deep Learning with Depthwise Separable Convolutions." ArXiv.org, 4 Apr. 2017, https://arxiv.org/abs/1610.02357.

[16] He, Kaiming, et al. "Deep Residual Learning for Image Recognition." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[17] Szegedy, Christian, et al. "Rethinking the Inception Architecture for Computer Vision." ArXiv.org, 11 Dec. 2015, https://arxiv.org/abs/1512.00567.

[18] Szegedy, Christian, et al. "Inception-V4, Inception-Resnet and the Impact of Residual Connections on Learning." ArXiv.org, 23 Aug. 2016, https://arxiv.org/abs/1602.07261.

[19] Inc, Alex Krizhevsky Google, et al. "ImageNet Classification with Deep Convolutional Neural Networks." Communications of the ACM, 1 June 2017.

[20] L. N. Smith, "Cyclical learning rates for training neural networks," 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017.