IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Breast Cancer Classification using Deep Learning Approaches and Histopathology Image: A Comparison Study

FAEZEHSADAT SHAHIDI, SALWANI MOHD DAUD, Member, IEEE, HAFIZA ABAS, NOOR AZURATI AHMAD, Member, IEEE, AND NURAZEAN MAAROP

Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia.

Corresponding author: Salwani Mohd Daud (salwani.kl@utm.my)

This work was supported in part by the Ministry of Education, Malaysia, and Universiti Teknologi Malaysia under Grant Q.K130000.3556.05G05.

ABSTRACT Convolutional Neural Network (CNN) models are a type of deep learning architecture introduced to achieve the correct classification of breast cancer. This paper has a two-fold purpose. The first aim is to investigate the various deep learning models in classifying breast cancer histopathology images. This study identified the most accurate models in terms of the binary, four, and eight classifications of breast cancer histopathology image databases. The different accuracy scores obtained for the deep learning models on the same database showed that other factors such as pre-processing, data augmentation, and transfer learning methods can impact the ability of the models to achieve higher accuracy. The second purpose of our manuscript is to investigate the latest models that have no or limited examination done in previous studies. The models like ResNeXt, Dual Path Net, SENet, and NASNet had been identified with the most cutting-edge results for the ImageNet database. These models were examined for the binary, and eight classifications on BreakHis, a breast cancer histopathology image database. Furthermore, the BACH database was used to investigate these models for four classifications. Then, these models were compared with the previous studies to find and propose the most state-of-the-art models for each classification. Since the Inception-ResNet-V2 architecture achieved the best results for binary and eight classifications, we have examined this model in our study as well to provide a better comparison result. In short, this paper provides an extensive evaluation and discussion about the experimental settings for each study that had been conducted on the breast cancer histopathology images.

INDEX TERMS Breast cancer, histopathology medical images, deep learning, transfer learning, data augmentation, pre-processing, classification.

I. INTRODUCTION

According to the World Health Organization (WHO) [1], breast cancer is the most common cancer among women globally. Every one out of three affected women will die. Several methods including mammography, magnetic resonance imaging (MRI), and pathological tests are the current investigation modalities of breast cancer. Among those methods, the histopathology images are considered as the gold standard to improve the accuracy of the diagnosis for patients who have already undergone other investigations such as mammograms [2]. Moreover, the histopathological examination can provide more comprehensive and reliable information to diagnose cancer and assess its effects on the surrounding tissues [3]–[5].

In order to obtain the histopathological slides from breast cancer tissues of the patients, the laboratory technicians first apply hematoxylin to stain the cell nuclei blue before counter-staining the cytoplasmic and non-nuclear components with eosin in different shades to highlight the different parts of transparent tissue structures and cellular features [6], [7]. Then, digital histopathological images are obtained from the microscopic examination of the stained biopsy tissues of breast cancer [3], [8], [9].

Although these images provide pathologists (human) with an all-inclusive view, mistakes can still happen when the diagnosis becomes too time-consuming due to the large-sized slides [6], [10]–[15].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3029881, IEEE Access

To overcome this problem, more researches are focusing on utilizing deep learning approaches to examine the histopathological images to improve the accuracy of the cancer diagnosis [14], [16]. The methods of breast cancer diagnosis using digital histopathology images can be categorized as detection, classification, and segmentation [17].

This study aimed to show deep learning techniques in the field of breast cancer histopathological image classification. The challenges for the breast cancer pathology image classification were identified and the solutions to these challenges were discussed.

The paper is structured as follows: Section II provides an overview of breast cancer and its subcategories. Subsequently, section III includes several parts, such as public databases, data augmentation methods, and pre-processing techniques. Section IV covers two main parts: deep learning models and transfer learning methods. Next, Section V delivers a comparison analysis based on previous reviews. Moreover, in this part, the most current deep learning models that have been identified and examined will be compared with previous studies and discussed. Finally, a conclusion that consists of critical discussions and an overview of the future works are outlined.

II. BREAST CANCER TYPES AND SUBTYPES

Although there are about 20 major types of breast cancer, the majority can be classified into two main histopathological classes: Invasive Ductal Carcinoma (IDC) and Invasive Lobular Carcinoma (ILC) [16], [18]. Among these two types of breast cancer, IDC is given more focus by the researchers. Fig. 1 shows the progression of breast cancer from hyperplasia to invasive carcinoma.

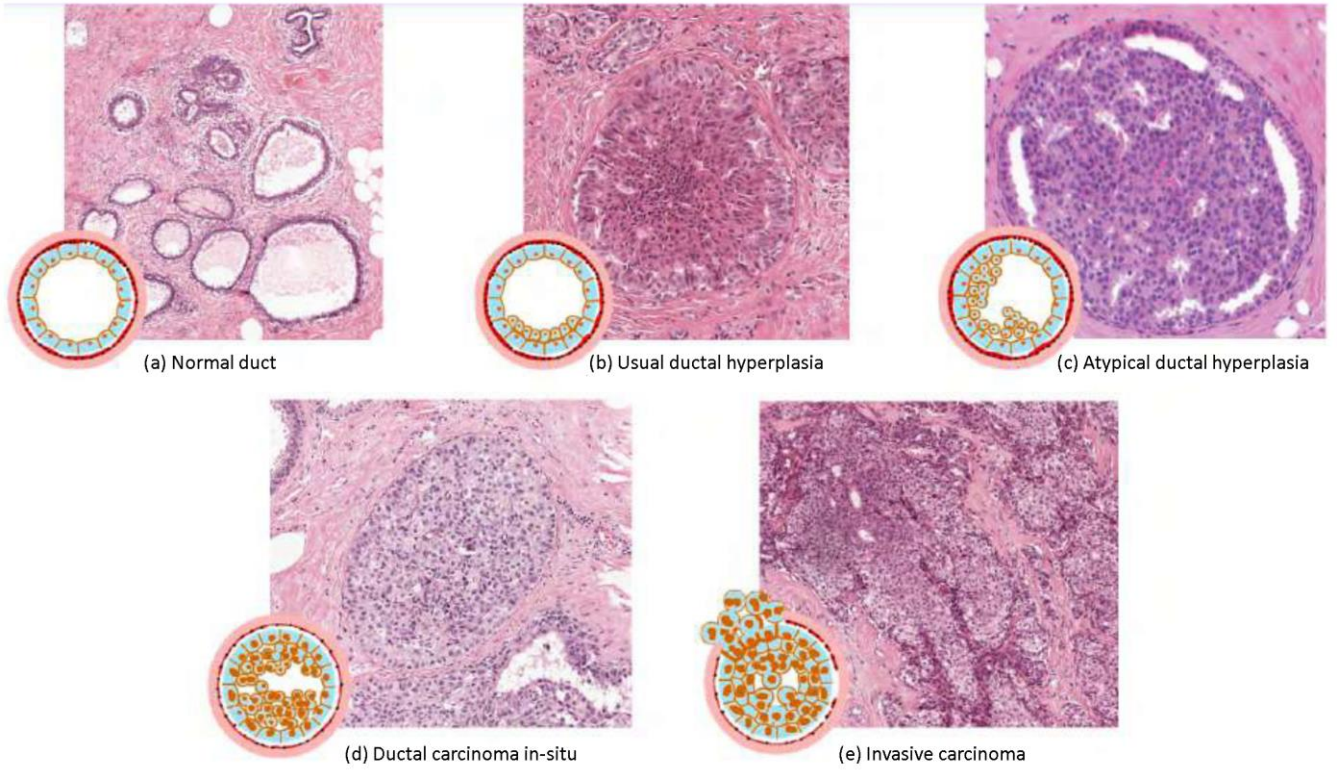


FIGURE 1. The progression of breast cancer disease at different stages. The first two pictures on the top left depict (a) Normal duct and (b) Usual ductal hyperplasia. However, breast cancer encompasses the progress from (c) Atypical hyperplasia, (d) Ductal carcinoma in-situ (DCIS), to (e) Invasive cancer [6].

The IDC type of breast cancer can either be benign or malignant. There are five malignant or carcinoma subtypes under IDC: tubular, medullary, papillary, mucinous, and cribriform carcinomas. Benign IDC includes adenosis, fibroadenoma, phyllodes tumor, and tubular adenoma [16]. Fig. 1 shows the different stages of breast cancer disease. Each stage has its distinct features that can be helpful in the diagnosis process. Moreover, due to the heterogeneity in breast cancer and certain restrictions concerning the histopathological classifications, a proper evaluation approach based on cell morphological features such as shape, size, and object-counting must be conducted to detect any abnormalities [16]. As for any cancer, an accurate detection of the type and proliferation of the disease is vital for the physicians to treat breast cancer disease optimally [6], [19]–[21]. For instance, the treatment for early IDC usually involves a combination of surgery, radiation therapy, chemotherapy, hormone therapy, and/or HER2-targeted therapy [22], [23].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3029881, IEEE Access

III. ANALYZING

A. DATABASES

1) NATURAL IMAGE DATABASES

Based on the previous studies, the majority of the researchers applied pre-trained models with weights initiated from the natural image databases, including ImageNet, an object-centric database [24] with more than 14 million (M) labeled images.

2) PATHOLOGY DATABASES

Contrary to other deep learning research fields, one of the challenges in the medical field is the lack of annotated data such as GoogLeNet to train the models with deep layers.

Several public pathology databases, including Cancer Metastases in Lymph Nodes (Camelyon), have been introduced to provide a large amount of annotated data to overcome this issue [25]. In this paper, we discussed several breast cancer histopathology image databases that have been examined by other papers. All of the discussed pathology databases were similar since they consisted of whole-slide images generated from breast tissue biopsy samples stained with hematoxylin and eosin (HE). Thus, the generated images were all colorful with three channels. However, each database had different magnification factors as different hardware equipment was used [26]. Thus, the images were different in terms of the resolution, possibly impacting the diagnosis. The pathology databases highlighted by other researchers are discussed in the following paragraphs.

First, the Breast Cancer Histopathological Image Classification (BreakHis) is a pathology dataset that consists of 7,909 breast cancer histopathology images from 82 patients with different magnification factors, including 40×, 100×, 200×, and 400× [8]. The 7,909 images include 2,480 benign and 5,429 malignant sample images with all the subtypes mentioned above [16].

The second database is the Stanford Tissue Microarray (TMA) database, a public resource with an access to 205,161 images [27]. All the whole-slide images have been scanned by a 20× magnification factor for the tissue and 40× for the cells [28].

Third, the Cancer Metastases in Lymph Nodes (Camelyon) was established based on a research challenge dataset competition in 2016. This database comprises of 400 whole-slide images with a size of 218,000 × 95,000 pixels [15]. The whole-slide images are stored in a multi-resolution structure, including 1×, 10×, 40× magnifying factors. It also has both benign and malignant images [30]. The training dataset in this database has 270 whole-slide images, 160 of which are normal slides and 110 slides containing metastases [29]. As this database has been published in a whole-slide image format, the size of the patches can be defined by the individual researchers using this database [31].

The fourth database contains Breast Cancer Histopathology (BACH) images obtained from ICIAR 2018 Grand Challenge. This database includes 400 images with equal distribution of the four different classes, including normal (100), benign (100), in situ carcinoma (100), and invasive carcinoma (100). The high-resolution images are digitized with the same conditions and magnification factor of 200× [32]. All the images in this database have a fixed size of 2048 × 1536 pixels [33].

Lastly, the fifth database, the Bio-Image Semantic Query User Environment (BISQUE), contains only a small number of histopathological images of breast cancer. The size of the photos is 896 × 768 pixels. It contains both benign (32 images) and malignant images (26 images) [33].

B. DATA AUGMENTATION

The main challenges encountered while training deep learning models are insufficiently labeled data and imbalanced number of classes [25]. The lack of labeled data causes the models to generate biased results or also known as “overfitting problem” [34]. Meanwhile, the imbalanced classes can prevent efficient classification performance [35]. In order to address these two challenges to produce an efficient classification, data augmentation methods are necessary [25], [36].

Most of the data augmentation methods applied in breast cancer histopathology are listed in Table I, II, and III. The methods include:

1) Random cropping: Randomly select several valid corner points before cutting one image into multiple images. This method ensures no duplication in the cropped images.
2) Rotation: An image is rotated based on an angle, e.g. 45 degrees, and the rotation is repeated continuously [37].
3) Color Shifting: This method adds or subtracts numbers to the three channels of red, green, blue (RGB). It can help to create different color distortions to become more purplish, yellowish, or bluish.
4) Flipping: The images can be flipped horizontally or vertically.
5) Intensity variation: The intensity of the images can be varied between −0.1 to 0.1 or 2.0 to make them brighter or darker.
6) Translation: The image pixels can be adjusted with ±20 pixels.

C. PRE-PROCESSING

This section discusses the pre-processing methods that have been applied to breast cancer histopathology images to address challenges such as low resolution and noisy pictures [25]. Since these issues affect the performance of the models during the classification, pre-processing techniques are needed to eliminate the noise in the histopathological images resulted from the staining procedures [25], [38].

1) RESIZING

Resizing is applied by researchers to change the input pictures into specific sizes tailored to the deep learning models [39]. Although this method is often applied before feeding the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3029881, IEEE Access

images into the models, Chen *et al.* [30] performed the resizing process with a fully convolutional layer as part of the network to conduct a real-time experiment.

2) RE-BALANCING THE CLASSES

According to some studies, the application of datasets with unidentical classes may create a bias that leans towards the majority and produce false values in the model [36]. Methods such as data augmentation, under-sampling, or over-sampling are proposed by previous researchers to rebalance the number of classes.

3) NORMALIZATION

As the color in the histopathological images is very intense, researchers can apply standardization or normalization to map the numbers to a range between zero and one to decrease the distribution and the intensity of the colors [40].

4) IMAGE CONTRAST ENHANCEMENT

In the following sections, the image enhancement method for grayscale images will be explained based on a previous study [41]. While the breast cancer histopathology images follow the RGB color model with three channels, this operation will be conducted for the red, green, and blue color scales separately. Additionally, the image contrast enhancement algorithm can improve the brightness of the image uniformly by mapping the lowest gray level to 0 and the highest value to 255. By using this method, the values of the gray level would spread in the histogram. However, the overall shape of the histogram would be unchanged, except for becoming wider to fill the range (0, 255).

5) MULTIRESOLUTION SEGMENTATION

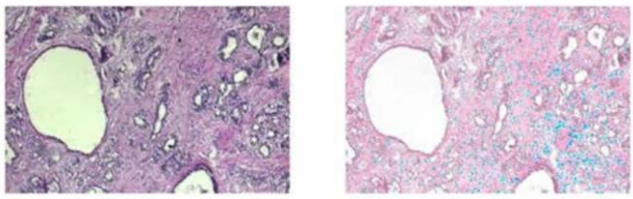
In a previous study [42], multiresolution segmentation was used to convert pixels into superpixels. This method can be used as an optimization approach to deal with large-scale breast cancer histopathological images. The creation of superpixels begins by growing the pixels of the image objects before merging them by classifying similar pixels to be adjacent to one another. This method is carried out based on several similarities, such as scale, compactness, shape (correlated with color), and layer weights of the images.

6) STAIN NORMALIZATION

This method was utilized by Nawaz *et al.* [43] to normalize breast cancer histopathological images. The algorithm of stain normalization was proposed in another study [44] as explained below:

For RGB slides, the output is stained normalized images as shown in Fig. 2.

1) Convert RGB to optical density (OD) [where OD = $-\log_{10}$(RGB image)].
2) Remove transparent pixels or data with lower than threshold β.
3) Compute singular value decomposition (SVD) for OD tuples.
4) Create a plane from the SVD directions corresponding to the two largest singular values (higher variance).
5) Project the values onto the plane spanned by the eigenvectors corresponding to the two largest eigenvalues.
6) Calculate the angle of each point for the first SVD direction.
7) Determine the stain concentrations by finding robust extremes, α^{th} and $(100-\alpha)^{th}$ percentiles of the angle.
8) Convert the extreme values back to OD space (normalize stain concentrations).
9) Deconvolve the image using the determined stain vectors according to Ruifrok *et al.* [45].



(a) Original image (b) Image after stain normalization

FIGURE 2. Histological image stain normalization. Left: (a) Original image. Right: (b) Image after normalization [43].

7) STAIN NORMALIZATION WITH COLOR TRANSFER BETWEEN IMAGES

The method of color transfer between images was first introduced by Reinhard *et al.* [46]. This method was later applied by Vesal *et al.* [47] to normalize the histopathological images by matching the statistical color histogram of one image as the source image for another one. This method first converts the RGB to CIE LAB color space to decorrelate the channels, i.e. L* (white), a* (red/magenta), and b* (blue). Then, the mean and standard deviation for each channel are calculated separately to correct the colors. The following steps should be taken to normalize the stain pictures by using a target image; where κ shows the channel as in (1), I_κ is the normalized image as in (2), (3), (4), σ is the standard deviation, S^κ is the stained source image, and T^κ is the target image.

$$\kappa = (l, \alpha, B) \tag{1}$$

$$I_\kappa = \frac{\sigma_t^\kappa}{\sigma_S^\kappa} \tag{2}$$

$$I_\kappa = I_\kappa \ (S^\kappa - mean\ (S^\kappa)) \tag{3}$$

$$I_\kappa = I_\kappa + mean\ (T^\kappa) \tag{4}$$

V. DEEP LEARNING

Deep learning dates back to the 1980s. However, training such models was not applicable because of the lack of data and the limited power of the hardware equipment back then. Nowadays, with a large amount of data and sophisticated hardware, deep learning models can be applied easily.

Due to the variation and complexity of image data in the medical field, the features should be extracted manually. Thus, the traditional learning models used in other fields are not as suitable and reliable. Furthermore, the learning models are not able to learn fast as the raw data cannot be fed efficiently [48].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3029881, IEEE Access

Thus, the Convolutional Neural Network (CNN) models, a type of deep learning architecture, are introduced to solve the enormous number of parameters in the traditional neural network while working with the images. As images have highly correlated pixels, the CNN models can extract the most significant features that play the most fundamental roles in the image classification.

According to a previous study [49], the most fundamental parts of the CNN models are convolutional layers, pooling layers (for subsampling), and fully connected layers. For the convolutional layers, the critical part is the filter (or kernel) by which the features in the input images can be extracted. Each kernel has a certain width, height, and the number of channels. Different width and height of kernels in the convolutional layers create the different spatial sizes of the output image.

Moreover, the number of channels in the kernel corresponds to the number of feature maps. In each step of the CNN model, as the size of the image decreases, the size of the feature maps increases. This trend can be observed in most of the CNN models. After the convolutional layers, subsampling is conducted using pooling layers, for example, maximum (max) or average pooling.

Similarly, pooling layers have kernels and they can also extract the max or average features. With the aid of subsampling, the spatial size (width and height) of the images can be decreased with no computational complexes since the kernels related to the pooling layers do not need to be trained. Furthermore, the number of stride and padding can help the layers to preserve or change the size of the image. Finally, the fully connected layers, along with the SoftMax functions, are utilized to perform the classifications.

According to Litjens *et al.* [25], it is challenging to choose an architecture of the deep learning models based on the input formats. In the following sections, the models of deep learning suitable for breast cancer histopathological images are outlined.

A. DEEP LEARNING MODELS

1) CLASS STRUCTURE-BASED DEEP CONVOLUTIONAL NEURAL NETWORK (CSDCNN)

This model was applied by Han *et al.* [50]. It consists of three convolution layers with kernel sizes of 3 × 3, 5 × 5, and 7 × 7. Moreover, the stride in each step is two, and before the fully connected layers, Han *et al.* [50] applied the mean-pooling strategy with a 7 × 7 receptive fields and a stride of one to flatten the layers. The final input of this model is 256 × 256 × 3.

2) ALEXNET

The AlexNet model was proposed by Krizhevsky *et al.* [51] after being inspired by LeNet-5 in another study [49]. This model was implemented to input RGB images with a size of 227 × 227.

Furthermore, this architecture has eight layers with five convolutional layers. Three of the layers are followed by max-pooling layers while another three are fully connected layers. The 1000 SoftMax activations were used to classify the outputs.

In this model, the size of the input image in each layer decreases while the number of channels increases. This enables the model to extract more features. After the extraction, the fully connected layers will provide the feature weight so that the final SoftMax layer can classify the output. However, AlexNet has 60 M parameters. A significant number of trainable parameters in this model may affect the computational operations negatively.

3) VISUAL GEOMETRY GROUP NETWORK(VGGNET)

This network was introduced by Simonyan and Zisserman [52]. According to He *et al.* [53], the 19 layers of VGG architecture consist of six parts. The first two parts have two convolutional layers and the next three parts have four convolutional layers. The final part consists of three fully connected layers for classification. The trend of the layers in this architecture is simple. The number of channels increases by a factor of two whereas the spatial resolution (width and height) decreases by half in each step. Furthermore, the filter in all convolutional layers is 3 × 3, making the learning faster than previous models. Moreover, the simplicity of this model makes it attractive to researchers. As VGG-16 has 138 M parameters compared to VGG-19 with 144 M parameters, most of the researchers usually prefer to work with the VGG-16.

4) VGG-M

According to Mahmoud [54], VGG-M consists of eight layers and uses five convolutional layers. This model is different from the VGG models discussed earlier. The kernel size for the first and second layers does not follow the rule of 3 × 3. Furthermore, the numbers of the width and height do not halve by the factor of two like the other VGG models. Furthermore, the input size suitable for this model is 224 × 224 with three channels.

5) DEEP RESIDUAL LEARNING (RESNET)

As explained previously, the number of the convolutional layers in the VGG models can reach up to 19 layers. In practice, increasing the number of layers hinders the training tasks as it increases the error rate [53]. In order to have deeper layers with no complexity, ResNet blocks with shortcut connections are proposed by He *et al.* [53]. Furthermore, Rectified Linear Units (ReLU) is applied as the activation function in this block.

Moreover, if the output of the two convolutional layers or F(x) within the block is zero, the ResNet block output is equal to x. In other words, the shortcut connection helps the model learn the identity function in a short time with low complexity. However, in certain best-case scenarios when F(x) is not zero, the model can enhance.

To build a ResNet model, He *et al.* [53] applied VGG-19 as the reference network and added more layers to eventually create a plain 34-layer model. Following that, the author applied shortcut connections after every two blocks of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3029881, IEEE Access

convolutional layer in the 34-layer plain network to improve it into a deep residual network. The dotted shortcuts are applied in a 34-layer residual, indicating that the channels increase by a factor of two.

By comparison, the number of the parameters for ResNet-34 is 21.8 M, ResNet-50 is 25.6 M, ResNet-101 is 44.5 M, and ResNet-152 is 60.2 M. Therefore, the number of the trainable parameters increases as more layers are added to the architectures. Thus, the following network is applied to build a deeper network with fewer parameters.

6) NETWORK IN NETWORK (1×1 CONVOLUTION)

A one-by-one convolutional layer was introduced by Lin *et al.* [55] to address the operational complexes. This convolutional layer applies a 1×1 kernel size to change the number of dimensions. This network is embraced by almost all the inception models and the DenseNet model to build deep and accurate networks as explained below.

7) INCEPTION-V1 (GOOGLENET)

GoogLeNet network was proposed by Zeng *et al.* [56] to build a model with more layers and fewer parameters to increase the accuracy as discussed in previous studies [52], [53], [56]. This model is built by the inception block. By using a one-by-one network as a bottleneck in this block, the number of the channels (or dimensions) is then reduced before passing the input to the next layers with a 5 × 5 or 3 × 3 filter size.

In this model, nine inception blocks are tightened together to build a 22-layer GoogLeNet architecture. Moreover, this model consists of max pooling, average pooling, convolutional layers, fully connected layers, and SoftMax layers.

In a recent study, Zeng *et al.* [56] utilized two auxiliary classifiers in which the prediction has been done by the GoogLeNet model beforehand. The classifiers were then compared with the ultimate result. The auxiliary classifier is designed to help the GoogLeNet model by providing regularization to tackle the vanishing gradient problem [57].

According to Szegedy *et al.* [58], at the end of the training, the network with auxiliary branches outperformed the network without any auxiliary branch in terms of accuracy. Additionally, Zeng *et al.* [56] also attempted to overcome the congestion problem in the fully connected layers by applying the 70% dropout techniques. In this model, the number of the parameters is 5 M. It is considerably smaller than that of the ResNet models, including the 34-layered ResNet. Consequently, the development of a deeper network with less trainable parameters actually started from this model.

8) INCEPTION-V2 / BATCH NORMALIZATION (BN)-INCEPTION

Inception-V2 was proposed by Szegedy *et al.* [58] with three kinds of inception modules. This model is explained below.

First, the inception module with two 3 × 3 convolutions is adopted instead of using one block of 5 × 5. Szegedy *et al.* [58] used two blocks of 3 × 3 in order to decrease the size of kernels but ended up with computational complexes.

After the factorization into smaller convolutions, Inception-V2 blocks have fewer parameters than Inception-V1. For instance, instead of using convolutional layers with a kernel size of 7 × 7 (49 parameters), the author used 1 × 7 or 7 × 1 (14 parameters).

Third, Szegedy *et al.* [58] improved the dimensional representation by using an expanded filter block. With more activations per tile, the model can be trained faster.

Finally, the author employed a batch normalization algorithm in this model. This algorithm is based on two fundamental concepts: normalization and distributions [59], [60].

In every batch, the following steps are computed. The values of x_i are the input for the mini-batch B of size m, and μ_B is the mean of the specific mini-batch (5). Meanwhile, σ is the mini-batch variance (6) and y_i is the output corresponding to each input in the batch (8). Gamma (γ) and beta (β) are known as scale and shift parameters respectively and they have to be trained.

$$\mu_B \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \tag{5}$$

$$\sigma_B^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2 \tag{6}$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \tag{7}$$

$$y_i \leftarrow \gamma \hat{x}_i + \beta \equiv BN_{\gamma,\beta}(x_i) \tag{8}$$

Batch normalization completes.

9) INCEPTION-V3

This model is the same as Inception-V2 except that it applies batch normalization in auxiliary classifiers [58], [59]. This model is very popular among the researchers for medical imaging [25]

10) INCEPTION-V3 – FULLY CONVOLUTIONAL NETWORK (FCN)

To expand Inception-V3, Chen *et al.* [30] proposed a new and cost-effective real-time method composed of a deep learning model with an augmented reality microscope (ARM). The deep learning architecture was set in a computer connected to the ARM through a software pipeline to produce real-time results.

This deep learning model consists of two parts: FCN and Inception-V3. The first part of this model is fully comprised of convolutional networks. This part converts the architecture to an application-agnostic platform so that large-scale images can be resized before being fed into the next section. The second part is the Inception-V3 with modified blocks. The same padding replaces the context of the image with zeros. Thus, instead of using the same padding in the inception blocks, the author did not apply any padding (or valid padding) to keep to the context. Furthermore, with the advantage of the cropping layer added to the inception blocks, the size of the images can be reduced before concatenation.

11) INCEPTION-V4

Inception-V4 was introduced by Szegedy *et al.* [61]. This model originated from Inception-V2 and Inception-V3 with batch normalization. In this model, the input image size is 299 × 299 × 3, and it has three types of inception modules. After passing through the stem, the image size decreases but the number of channels increases. Each inception block is followed by a block of reduction to decrease the image size and increase the channels. Finally, the dropout keeps 80% of the weights to 1000 SoftMax to classify the output.

12) INCEPTION-RESNET-V2

Inception-ResNet-V2 schema originated from the ResNet module and Inception-V4. The overall network is explained as follows:

First, the spatial input image size for this model is 299 × 299 and the number of the RGB channels is three.

Next, the stem part comprises six layers. The input for this part is 299 × 299 × 3 and the output is 35 × 35 × 384. The stride of two along with valid or no padding (V padding) is adopted to reduce the size of the image. On the contrary, the convolutional layers without V padding preserve the size of the image (same padding).

Thirdly, the block of 5× Inception-Resnet-A can change the number of channels with ease with the aid of one-by-one convolutional layers [61]. Thus, compared with the previous inception blocks without identity connection, this block can be learned faster. The output of this block is 35 × 35 × 384.

Fourthly, the Reduction-A block decreases the size of the image from 35 × 35 to 17 × 17 with the aid of a stride of 2 and V padding while increasing the number of channels. The output of this reduction block is 17 × 17 × 1154.

With that, in the 10× Inception-Resnet-B, a factorization along with identity connection is applied in this block to achieve an output of 17 × 17 × 1154.

In the sixth step, for Reduction-B, with the stride of two and V padding, the size of the image decreases from 17 × 17 to 8 × 8 and the output size is 8 × 8 × 2048.

Following that, the 5× Inception-Resnet-C module has an output of 8 × 8 × 2048 and it is also equipped with factorization and identity connection.

In the eighth step of average pooling, with the kernel size of 8 × 8, the image size is flattened to 1 × 1 [61] since the average pooling selects one average number out of 8 × 8 (64). The output of this layer is 1 × 1 × 2048.

That step is followed by dropout in which 0.2 of the connections will be removed and the model keeps 80% of the weights.

Lastly, the SoftMax layer is used for the classification of 1000 classes.

13) RESNEXT

After the introduction of the ResNet models by researchers [53], the residual connection was considered as an essential factor. Subsequently, wide residual networks were introduced by researchers [62] to increase the width of the network instead of its depth. Since parallelizing width operations in the wider network appeared to be computationally efficient, Xie *et al.* [63] built another cutting-edge model entitled ResNeXt. This model is made up of in-built wide residual networks along with identity connections. The width of the network becomes a new hyperparameter that has been added to this model. This is known as cardinality and it can improve the accuracy and ability to withstand the complexity. Unlike the Inception models with deep and complicated architectures, the ResNeXt model delivers simplicity with the aid of cardinality. Basically, the complexity of a ResNeXt with 50 layers is equal to a ResNet model with 101 layers. Besides that, the number of parameters changes due to the number of cardinalities. For instance, ResNeXt-101 with 32 cardinalities has 44.18 M parameters and ResNeXt-101 with 64 cardinalities has 83.46 M parameters. Moreover, the input size of this model is 224 × 224 × 3.

14) SQUEEZE AND EXCITATION NETWORK (SENET)

Squeeze and Excitation block (SE) was introduced by Hu *et al.* [64]. This block has three parts: squeezing, excitation, and scaling. First, the squeezing part of this block applies global average pooling to make use of the contextual information along with the local receptive field. This part changes the size of the image (U) from H x W x C (height, width, and channel) to 1 × 1 × C.

Next, the excitation part exploits the information collected in the first part via two fully connected (FC) layers. The first FC layer, with the size of 1 × 1 × C/r, (r = hyperparameter reduction ratio with the default value of 16 for SE-ResNet-50) is followed by the ReLU function to learn the nonlinear interactions between the channels. The second FC, with a size of 1 × 1 × C, is accompanied by a sigmoid function to capture the mutual relationships between channels.

Finally, the scaling part rescales the output of the excitation part (s_c) by channel-wise multiplication of s_c and feature maps of the input image (u_c). By simply piling a group of SE blocks, we can form an SE network with ease. The number of parameters of this model depends on the number of SE blocks applied in the model. For instance, SENet-154 has 115.09 M parameters. Furthermore, the size of the input image for this model is fixed to 224 × 224 × 3. The SE module can be added to a residual block for the block to transform into the SE-ResNet module. By employing the SE module within the blocks of ResNet-50 architecture, Hu *et al*. turned this model into SE-ResNet-50 and increased the accuracy by 0.26%. In addition to that, another study discussed the flexibility of this module as other networks can be added onto it, for example, VGG, Inception, Inception-ResNet, ResNeXt, and others [39].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3029881, IEEE Access

15) BREAST CANCER HISTOPATHOLOGY IMAGE CLASSIFICATION NETWORK (BHCNET-N)

Jiang *et al.* [65] applied the SE approach and introduced a small SE-ResNet module to build a BHCNet-N structure. Small SE-ResNet module uses factorizations by adopting two convolutional layers of 1 × 3 and 3 × 1 instead of two 3 × 3 convolutional layers, reducing the number of parameters to a great extent. This model consists of three parts. The first part is the single convolutional layer, the second part is N small SE-ResNet modules, and the third is a fully connected layer along with the output layer for classifications. The number of N after BHCNet indicates the number of small SE-ResNet modules applied in the model. For example, BHCNet-3 has three small SE-ResNet modules.

16) DENSE CONVOLUTIONAL NETWORK (DENSENET)

Introduced by a group of researchers [66], the Dense Convolutional Network uses the ResNet [53] identity connections for all the layers. The DenseNet network consists of three parts, including the dense blocks, transitional layers, and one classifier layer.

First, the dense block consists of several convolutional layers whereby each is connected to the successive layers. The transitional layer is followed by the dense block. Using the 1 × 1 convolutional layer and pooling layer decreases the feature maps to a fixed number [67]. Unlike the previous models, the feature maps are fixed in this model. Thus, the number of parameters and computational complexes decreases to a greater extent. For instance, the number of parameters of the DenseNet-121 model is 7.98 M. Moreover, the size of the input image for the DenseNet model is 244 × 244 with three channels.

17) DUAL PATH NETWORK (DPN)

This model was proposed by Chen *et al.* [68] by adopting both ResNet and DenseNet blocks [66] to form a new architecture. Since residual blocks reuse features and DenseNet blocks explore new features, the concatenation of the strengths of both modules leads to the introduction of a new macro-block. With this module, the accuracy can be improved and the computational complexes can be decreased. Each macro-block has three convolutional layers of 1 × 1, 3 × 3, and 1 × 1. The last convolutional layer is split into two paths. The first path is for the addition of the residual, like ResNet, while the second one is for the concatenation of densely connected parts, like DenseNet. In order to obtain more efficiency, the grouped convolution approach can be used to increase the width of the model that has been employed in the ResNeXt model [63] as introduced by researchers in a published study [51]. Interestingly, the lower number of parameters in the DPN model makes it more computationally efficient compared to the ResNeXt models. For example, the number of the parameters for DPN-131 is about 79.25 M. The suggested input size for this model is 224 × 224 × 3. The DPN-92 has 0.63% higher top-1 accuracy score than ResNeXt -101, and the DPN-98 consumes about 25% less FLOPs than ResNeXt-101(64 × 4d).

18) NEURAL ARCHITECTURE SEARCH NETWORK (NASNET)

Despite the cutting-edge results gained by the neural networks so far, designing a network suitable for a specific database is still time-consuming and prone to errors [69]. NASNet model is inspired by the automated neural architecture search (NAS) method introduced by Zoph *et al*. [70].

NASNet architecture was designed by Zoph *et al*. [71]. The author applied a recurrent neural network (RNN) and trained this network using a reinforcement learning approach to generate a network with the maximum expected accuracy on the validation dataset. Furthermore, in this model, the author benefitted from a new regularization method called ScheduledDropPath. Moreover, from the advantage of using Controller Recurrent Neural Network (CRNN), CNN, and reinforced evolutionary algorithm, this model is able to choose the best cell candidate to form the blocks and end up building the best architecture depending on the database [72]-[73]. In this model, the controller RNN generates sample architecture with a sample probability by using a set of operations. Then, the CNN model trains a child network with sample architecture to obtain a target accuracy result. Next, the controller RNN will update the sample architecture based on the gradient computed by using the sample probability and scale it by the target accuracy. There are three types of NASNet models: A, B, and C (Fig. 3). NASNet-A-Large has received the highest accuracy results. The input image size of this model is 331 × 331 × 3 and it has 8,89,49,818 parameters. Furthermore, the blocks are operational modules, including normal convolutions, separable-convolutions, max-pooling, average pooling, and identity mapping in the NASNet architecture, and the cell is the combination of these blocks. The number and types of these cells and blocks are optimized based on the selected database. The network is formed based on three factors, including the number of cells to be stacked (N), the number of filters in the first layer (F), and the combination of the N and F that are fixed during the search. The hidden layers are built through pairwise combinations and updated by a concatenation operation within each cell. Each block receives the hidden layers that are the output of the previous cells, and maps them into one output feature map that will be fed to the next cells as an input.

The ball chart in Fig. 3 shows a comprehensive comparison between almost all the deep learning models trained by the ImageNet dataset. The ball chart is based on the accuracy top-1 (Y-axis), operational complexes (X-axis), and trainable parameter size (the size of the balls shows the parameters in a million). The models proposed by the previous studies listed in Table I, II, and III for breast cancer histopathological images are all highlighted in Fig. 3 with red rectangular marks. Based on Fig.3, the most recent and accurate models are ResNeXt-101(32 × 4d), ResNeXt-101(64 × 4d), DualPathNet-131, SENet-154, and NASNet-A-Large, and they are illustrated with dark blue rectangular marks. Since there has been no or a limited number of studies investigating these models using the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3029881, IEEE Access

breast cancer histopathology image databases, they were examined in our study to determine the ones with the best results for these databases. Besides that, the Inception-ResNet-V2 was examined in our study as this model gained the best accuracy results for BreaKHis databases according to a study [14].

Some studies have shown a combination of the most accurate models as depicted in Fig. 3 for the breast cancer histopathological images. For instance, the BHCNET-N architecture was formed by N small (factorized) SE-ResNet modules. As illustrated in Fig. 3, the SENet module was added to the ResNet and ResNeXt skeletons to create SE-ResNet or SE-ResNeXt that produced better accuracy. In addition to that, ResNeXt-50, DPN-26, VGG-M, CSDCNN, and BHCNET-N models were not included in the comparison in Fig. 3. However, in our review, we will discuss them all.

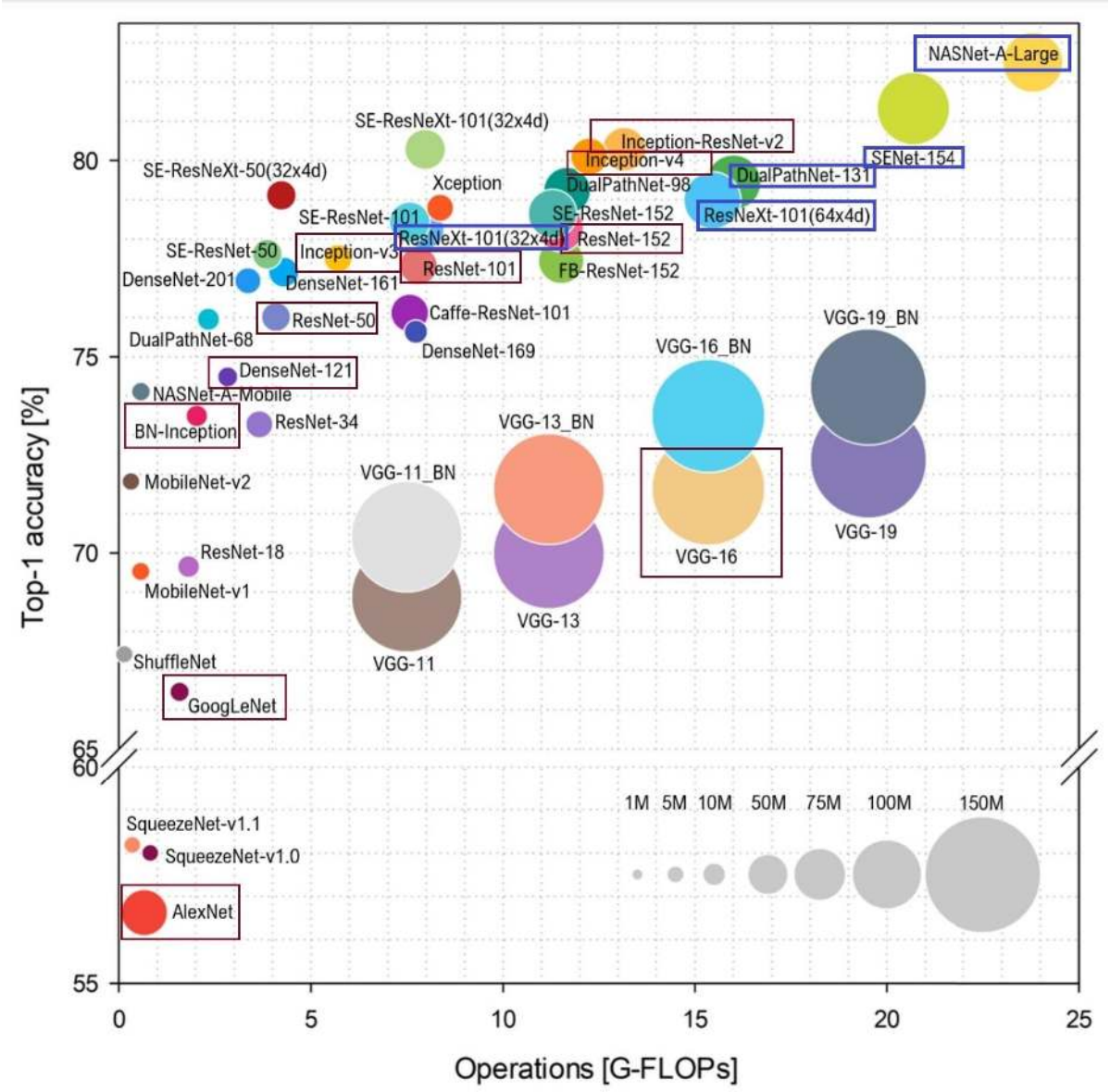


FIGURE 3. Ball chart reporting the Top-1 accuracy vs. computational complexity i.e. floating-point operations per second (FLOPS) in computing. The size of the balls shows the parameters. [74].

B. TRANSFER LEARNING

Without a doubt, the deep CNN models are more capable to gain significant results. However, training a model from scratch is not the most practical strategy due to the computational costs, convergence problems, and the insufficient number of high-quality labeled pathology images [67], [75]. Furthermore, for a large amount of data, the training of a model can be highly time-consuming due to the hardware limitations [76].

To solve these challenges, transfer learning methods are employed by researchers [77]. The methods of transfer learning are as follows:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3029881, IEEE Access

1) PRE-TRAINED

The community of deep learning believes in sharing. Therefore, other models that have been pre-trained by ImageNet can be applied. Nowadays, pre-trained models can be accessed via libraries such as Keras [62]. For example, we managed to implement a pre-trained model with the weights initiated from ImageNet [10].

2) FEATURE EXTRACTION

In this technique, all the layers are frozen except the SoftMax (last) layer. Moreover, the classification numbers in the last layer can be modified based on needs. For example, to classify the types of breast cancer, there can be two classes, either benign or malignant. Besides that, multiple classifications can be applied when classifying the subtypes of breast cancer.

3) FINE-TUNING

This method applies the pre-trained model, the same as feature extraction. However, unlike the feature extraction in which only the last layer is changed [75], this fine-tuning technique enables the researchers to retrain several layers based on the new data [75], [78].

V. COMPARISON ANALYSIS

A. TWO-CLASS CLASSIFICATION

1) PREVIOUSLY PUBLISHED LITERATURE REVIEWS

In this section, previously published literature related to deep learning models for histopathological images are compared and discussed. The review focuses on the examination of pre-trained deep learning models with weights initiated with the ImageNet database since models gained higher performance by using natural images than pathology ones [79].

The comparison conducted in this study is shown in three separate tables based on the number of classifications. Table I shows the experiments conducted for the binary classification of breast cancer, namely benign and malignant cancers on the BreakHis database. Table II illustrates the studies that include four classifications (normal, benign, in-situ carcinoma, and invasive carcinoma) on the BACH database. Besides that, Table III compares the experiments performed for eight classes, including all the breast cancer subcategories (adenosis, fibroadenoma, phyllodes-tumor, tubular-adenoma, ductal carcinoma, lobular-carcinoma, mucinous-carcinoma, and papillary-carcinoma) on the BreakHis database. As the BreakHis database supplies images in four different resolutions, multiple resolutions were used to assess the models. The models compared the average accuracy of all resolutions.

The results based on each resolution are outlined in the table based on the average accuracy scores. Each table provides a comparison based on the models and their average accuracy scores. Moreover, the different methods, such as data augmentation, pre-processing, transfer learning, optimization, and regularization were compared. Additionally, the database column in each table shows the equality or inequality of the classes and the total number of patches applied in the study. Since the input images were resized at the pre-processing stage and batch normalization was implemented in almost all the studies, these two methods were excluded from the tables.

Subsequently, the papers that included the examination of the models through different databases, such as TMA, Camylon, and BISQUE, were discussed. In addition to that, different items in the experimental settings, such as hardware, software, train and test split, learning rate, batch size, epoch, and iteration, were explained according to the studies.

Table I shows the comparison between the previous studies and our examination of the most current models conducted on the binary classification for benign and malignant breast cancer images on the BreakHis database. In this section, the studies performed on the binary classifications will be discussed and then our examinations will be explained in the next section.

In this table, the highest level of accuracy for binary classification was achieved by the pre-trained Inception-ResNet-V2 and feature extraction method examined by them [14].

There were two experiments conducted by Xie *et al.* [14]. The first one was performed on the BreakHis database with imbalanced classes and 7,909 breast cancer histopathology images. In the second experiment, data augmentation techniques, including turning and clockwise rotation, were applied. The number of classes was balanced and the average accuracy was improved from 97.90% to 99.79%. The accuracy improvement in this model highlights the importance of data augmentation techniques in increasing the size of the database and balancing the classes. Furthermore, the highest accuracy score (99.79%) in this study was achieved by using the 40× resolution. Apart from that, other pre-processing techniques, including normalization between -1 and 1, cutting border, and saturation adjustment, were also applied by Xie *et al.* [14].

Similarly, the BHCNet-3 model proposed by Jiang *et al.* [65] with three small SE-ResNet modules gained an average accuracy score of $98.87 \pm 0.10\%$. The result for this model was achieved using the BreaKHis database for binary classifications with an imbalanced number of classes. Unlike other studies, the transfer learning method was not adopted [65]. In other words, BHCNet-3 was trained from scratch but it obtained a satisfactory result. This study shows the importance of the SE modules that can be embedded in the lightweight architectures to improve the accuracy score. Besides that, we have performed several experiments on SENet-154 model, and this model gained the highest level of accuracy for binary and four classifications in comparison with other studies and other examined models in our study. This model will be explained in the next part. Moreover, according to the study [80], applying a large number of parameters to pre-trained models offers little extra efficiency over smaller models in the medical field.

TABLE I
COMPARATIVE STUDY OF TWO-CLASS CLASSIFICATION USING BREAKHIS DATABASE

Study	Pre-trained Model	Database	Data Augmentation	Pre-processing	Transfer-learning	Optimization/ regularization	Results (Accuracy %)				
							40 ×	100 ×	200 ×	400 ×	Avg
Our method	SENet-154	Uneven/Total: 7,909	Rotation, flip	Normalization	Fine-tuned the last three FC Layers	Adam/Dropout 0.5	-	-	-	-	99. 87
Our method	DualPathNet-131	Uneven/Total: 7,909	Rotation, flip	Normalization	Fine-tuned all the layers of the network	Adam/Dropout 0.5	-	-	-	-	99.74
Our method	Inception-ResNet-V2	Uneven/Total: 7,909	Rotation, flip	Normalization	Fine-tuned the last three FC Layers	Adam/Dropout 0.5	-	-	-	-	99.74
Our method	ResNeXt-101(32×4d)	Uneven/Total: 7,909	Rotation, flip	Normalization	Fine-tuned the last three FC Layers	Adam/Dropout 0.5	-	-	-	-	99.49
[14]	Inception-ResNet-V2	Even /Total: 27,262	Turning/ clockwise rotation	Normalization [-1,1]/cutting border/ adjust saturation	Feature extraction	Adam/ exponential decay method	99.79	99.37	99.43	99.10	99.42
Our method	ResNeXt-101(64×4d)	Uneven/Total: 7,909	Rotation, flip	Normalization	Fine-tuned the last three FC Layers	Adam/Dropout 0.5	-	-	-	-	99.36
Our method	NASNet-A-Large	Uneven/Total: 7,909	Rotation, flip	Normalization	Fine-tuned the last three FC Layers	Adam/Dropout 0.5	-	-	-	-	99.24
[65]	BHCNet-3 (with 3 small SE-ResNet module)	Uneven/Total: 7,909	Height and width shift, horizontal flip, constant fill mode	Down sampling, zero-mean normalization	-	SGD + momentum 0.9/weight decay of 1e-4.	98.87± 0.10	99.04 ± 0.10	99.34 ± 0.06	98.99 ± 0.17	99.06± 0.11
[36]	Inception-V3	Even/ Total: 4,960	Rotation, shift, flip, zooming	under-sampling (EUS SVMs)	Fine-tuned the last three layers	RMSprop/ weight decay le-6	-	-	-	-	98.00
[14]	Inception-ResNet-V2	Uneven/Total: 7,909	-	Normalization [-1,1]/, cutting border, adjust saturation	Feature extraction	Adam/ exponential decay method	97.90	96.88	96.88	96.88	97.13
[36]	VGG-16	Even / Total: 4,960	Rotation, shift, flip, zooming	under-sampling (EUS SVMs)	Fine-tuned the last three layers	Adam/ weight decay (le-6)	-	-	-	-	97.00
[14]	Inception-V3	Uneven/ total: 7,909	-	Normalization [-1,1]/, cutting border, adjust saturation	Feature extraction	Adam/exponential decay method	96.84	96.76	96.49	94.71	96.2
[81]	AlexNet	Uneven/Total: 7,909	-	-	Fine-tuned the last three layers	SGD + momentum/L2-regularised	90.96 ± 1.59	90.58 ± 1.96	91.37 ± 1.72	91.30 ± 0.74	91.05± 1.5
[81]	AlexNet- fc6 + VGG16-fc6 + SVM	Uneven/Total: 7,909	-	-	Feature extraction	SGD + momentum/L2-regularised	84.87 ± 1.14	89.21 ± 1.44	88.65 ± 2.41	86.75 ± 4.21	87.37± 2.55
[81]	AlexNet- fc7 + VGG16-fc7 + SVM	Uneven/Total: 7,909	-	-	Feature extraction	SGD + momentum/L2-regularised	84.58 ± 1.49	89.03 ± 1.46	88.31 ± 3.20	86.00 ± 4.08	86.23± 2.55
[54]	VGG-M	Uneven/7,909	-	-	Feature extraction	SGD + momentum 0.5 / dropout 0.5	86.2±2 .7	85.9±0 .5	87.2±3 .6	86.3±1 .7	86.4±2 .12

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3029881, IEEE Access

Among all the studies conducted on Inception-V3 listed in Table I, the highest level of accuracy (98%) was achieved by Lim *et al*. [36]. They conducted fine-tuning on the last three layers of this model [36]. Furthermore, the total number of histopathological images used in this study was 4,960, and the number of classes was balanced by an under-sampling method. In this study, the author applied data augmentation techniques, such as rotation, shift, flip, and zooming. They also conducted another examination on VGG-16 with fine-tuning on the last three layers. This model achieved a 97% accuracy score using the BreaKHis database with balanced data.

Likewise, in a study by Xie *et al.*, Inception-V3 achieved a 96.84% level of accuracy using the BreaKHis database, with the total number of 7,909 imbalanced types of breast cancer histopathological images [14].

Additionally, the fine-tuned AlexNet was used by Deniz *et al*. [81]. According to them, it achieved 91.05±1.5% level of average accuracy on the BreakHis database, with uneven classes and 7,909 histopathological images. They followed up with a second experiment on the concatenation output of the sixth and seventh layers of the AlexNet and VGG-16 model before classifying the results with a support vector machine (SVM) model. By using the features extracted from the sixth layer of AlexNet and VGG-16, this architecture could gain an average score of 87.37 ± 2.55%. Furthermore, by using the features extracted from the seventh layer, this model was able to reach 86.23 ± 2.55% of accuracy.

In other words, this study has shown that the breast cancer classification for the AlexNet model had a better result than the concatenation and SVM of AlexNet with VGG-16. However, this study did not apply any pre-processing and data augmentation techniques.

The lowest accuracy score on the BreakHis database by VGG-M in Table I was obtained by Mahmoud [54]. The score was 86.4 ± 2.12%. In this study, the pre-processing and data augmentation techniques were eliminated and the numbers of classes were imbalanced.

In the subsequent sections, the deep learning models examining different databases containing breast cancer histopathological images will be discussed and reviewed.

First, Habibzadeh Motlagh *et al.* [16] combined BreaKHis and TMA databases with different resolutions. The number of types and subtypes were then balanced with data augmentation techniques [16]. The total number of input images in this study was 16,846. They also employed pre-processing techniques, such as normalization, and color-distortion before examining all the layers in the model with fine-tuning techniques. In the same study, several experiments on ResNet models, including ResNet-152, ResNet-101, and ResNet-50 were conducted, obtaining 98.70%, 98.40%, and 97.80% scores of accuracy, respectively. By using the same techniques and databases, this study illustrated the importance of the deep layers in which ResNet-152 achieved the highest accuracy among all ResNet models.

In another study, Inception-V4 and Inception-V3 were examined by Habibzadeh Motlagh *et al*. [16] using images from TMA and BreaKHis. Inception-V4 gained a lower accuracy score (77.70%) than Inception-V3 (82.20 %). By comparison, the accuracy scores gained by Inception-V3 for the BreakHis database were 96.2 and 98% in a study by Xie *et al.* [14] and Lim *et al*. [36], respectively. Thus, this model gained better accuracy scores when investigating the BreakHis database in comparison with the database collected from TMA and BreaKHis. The experiment by Habibzadeh applied techniques such as data augmentation and pre-processing. Moreover, the total number of histopathological input images was as high as 16,846, and all the types and subtypes of breast cancer were approximately balanced.

Besides that, Inception-V3 was implemented with FCN on the database that combined Camelyon and a private database [30]. The total number of patches applied for training the model was 216,000, which was gained from 27,000 whole-slide images. Furthermore, the total number of images for evaluating the model was 1,000, all originated from a private database. The FCN architecture acted as a catalyst to make the size of the images smaller for the Inception-V3 architecture. During the training phase, Inception-V3-FCN was trained using a pathology database that contained patches sized 911 × 911 × 3 pixels. After the input images were passed through the FCN, their sizes were reduced to 111 × 111 × 192. However, during the evaluation, the input size would change as the observer's field of view (FOV) of the microscope device captured 2,560 × 2,560 × 3 pixels of whole-slides each time. With regards to this, the FCN model is considered as highly flexible as it is also able to feed images with more than 2,560 × 2,560 × 3 pixels. The size of the photos can be changed based on the region of interests, and the context of significance to the pathologists and deep learning models to better distinguish the tumor. This model could gain up to 96% accuracy, thus indicating the positive impact of FCN on this architecture.

In another two studies, the researchers investigated Inception-V1 and obtained different accuracy scores while utilizing different databases [16], [79]. In the first study, Inception-V1 obtained an accuracy score of 93.60% with a combination of the BreaKHis and TMA databases [16]. As mentioned earlier, the number of classes was balanced and the total number of input images was 16,846. Moreover, the data augmentation and pre-processing techniques were used in this experiment. The researchers also managed to reach 91.80% accuracy by using Inception-V1 on the combination of TMA and OUHSC databases. The total number of the input images in this study was 36,192 and there was an even number of all types of breast cancer. However, in this study, data augmentation techniques and optimization algorithms were not applied, and the model was only fine-tuned on the last two fully connected layers. Moreover, in the pre-processing phase, multi-resolution methods and contrast enhancements were adopted for the input images. Based on the study by Du *et al.* [79], AlexNet reached an 88.70% level of accuracy with TMA

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3029881, IEEE Access

and OUHSC databases. Although Du *et al.* [79] applied pre-processing methods, even classes, and a higher number of data in their experiment, this model actually achieved lower accuracy with TMA and OUHSC databases compared to the BreakHis database.

Evidently, studies that applied data augmentation techniques and balanced classes gained better accuracy scores than those that did not, as shown by Xie *et al.* [14]. As mentioned earlier, according to Litjens *et al.* [25], factors that affected the performance of the deep learning models during practical applications included low resolution and the existence of noise. Furthermore, the models showed different accuracy scores with different databases or resolutions. Based on a study [82], models such as VGG-19 and DenseNet-201 achieved the highest accuracy scores on the BreakHis database compared to the other four publicly available databases: BreakHis, PatchCamelyon, ICIAR, and Bioimaging. Moreover, according to the investigation in this study, most models gained better accuracy scores on the BreaKHis database than other databases. Additionally, the models gained different results when images from the BreakHis database were examined with different resolutions.

The mini-batch or normal batch gradient descent can create several isolated movements, resulting in high noise for the path to reach the converging point to train larger data. Applying the momentum algorithm can prevent steep steps so that it can work faster [83].

In Table I, Xie *et al.* [14] achieved the highest accuracy with Adam optimization for Inception-ResNet-V2 (99.79%). Another study used Adam optimization for VGG-16 and RMSprop for Inception-V3 [36]. RMSprop is an unpublished adaptive learning rate method proposed by Geoff Hinton [83]. Besides that, the second rank of the accuracy in this table was gained by Jiang *et al.* [65] who employed Stochastic Gradient Descent (SGD) with a momentum of 0.9. Other studies also employed SGD with momentum while working on VGG-M and AlexNet [54], [81]. Since only one study [81] in Table I implemented L2, it can be deduced that most researchers preferred dropout and weight decay over other regularization methods.

The pre-processing method is likely to improve accuracy as these techniques were shown to improve the efficiency in the majority of the studies. Although transfer learning methods helped the models gain satisfactory results, a study managed to achieve a successful outcome without using this method [65]. Besides that, three other optimization methods, including Adam, RMSProp, and SGD were also employed with the momentum technique. Nevertheless, the most favorable ways for the regularization were the dropout method and weight decay.

As binary classifications by previous studies show cutting-edge results, we have applied several current models in order to compare our results with the other studies. To ensure the reliability and accuracy of some examinations done on Inception-ResNet-V2, we have performed several experiments on this model and compared this model with the most current models. Our investigation for binary classifications will be discussed in the following section.

2) OUR EXAMINATIONS

For the two classifications in our study, we have applied the BreaKHis dataset with two classes of benign and malignant. The number of classes in our experiments was uneven. In all of the experiments, we have normalized the input images between -1 and 1. Normalizing helps to keep the network weights near zero, in turn making backpropagation more stable. Without normalization, the networks will tend to fail to learn. Subtracting the mean centers the data around zero and dividing by the standard deviation yields values of between -1 and 1. Furthermore, we matched the normalization when the models were trained. Hence, since we have applied pre-trained models on the ImageNet data set, in order to abide the normalization method applied for this data set, each color channel was normalized separately; the means were [0.485, 0.456, 0.406] and the standard deviations were [0.229, 0.224, 0.225]. Moreover, two methods of data augmentation techniques, including random rotation 45 degrees and vertical and horizontal flipping were applied. Besides that, Adam's method for optimization had been chosen while using backpropagation in our models. The dropout method had been practiced as a regularization method in our study. Moreover, in this part, we had employed six models with the best accuracy results on ImageNet as depicted in Fig.3. In order to have a comprehensive comparison, we have added and ordered these models to Table I ascendingly based on their acquired accuracy. These models are explained as follows.

During our investigation, a pre-trained SENet-154 network gained the highest accuracy score among all examined models. The size of the input images for this model was set to 244 ×2 44 × 3. All the layers of this model had been frozen except the last three fully connected layers with the size of 1024. The number of parameters of the last three fully connected layers was about 3 M. This model was retrained for 100 epoch and the highest accuracy score achieved by this model was 99.87% in epoch 27 which was quite a state of the art. Using data augmentation methods, the accuracy score improved by 0.63% from 99.24% to 99.87% and loss error decreased from 0.021 to 0.006 for the test dataset.

The second rank of accuracy results was gained by DualPathNet-131. In our examinations, we have used a pre-trained model and retrained all the layers of the network for 100 epochs by the size of input of 224 x 224 x 3 and the batch size of 32. With the aid of data augmentation techniques, we could improve the accuracy score by almost one percent from 98.73% to 99.74% for this model. The accuracy score with the data augmentation was received in epoch 50 and the error loss score in this epoch was 0.015. The number of parameters of DPN-131 was 79.25 M for 1000 classes, yet this number had decreased to 76.57 M due to the fact that we had replaced the 1000 classes with 2 classes in the last linear layer.

Inception-ResNet-V2 gained the highest accuracy results among all investigated previous reviews. In our investigation, this model gained 99.74% of the accuracy and 0.057 scores of the loss function in epoch 67. All the layers were frozen and we had added three fully connected layers whose number of parameters was about 2.6 M. Our examination ensured both accuracy and readability of this model, and the examination conducted by them [14].

The next cutting-edge model we explored was ResNeXt-101(32 × 4d). We had fine-tuned this model with three fully connected layers with the size of 1024. In this experiment, ResNeXt-101(32 × 4d) obtained 99.49% of the accuracy score and 0.006 scores of the test loss in epoch 66. The number of parameters of the last three fully connected layers was 3.149 M that were set to be trainable. This model gained 99.36% of the accuracy score and the loss score for the test dataset was 0.023 in epoch 83. By using a fine-tuning technique, 3.14 M of the parameters were dedicated to these trainable layers.

The highest test set accuracy score for the NASNet-A-Large model was gained in epoch 55, with 99.24% and the test loss score of 0.025. We had applied a fine-tuning technique with three fully connected layers of 1024 perceptrons. In this experiment, a pre-trained model on the ImageNet dataset had been applied, and all the layers were frozen except the last three layers. This model required a fixed input size of 331 × 331 × 3.

As discussed by Han *et al.* [50], accurate multi-classifications provided more valuable information for breast cancer diagnosis or prognosis than binary classifications. The staging of breast cancer is very important to help physicians in deciding the treatment modality [19]–[21]. However, the heterogeneity of color distribution of breast cancer histopathological images might lead to subtle differences or noisy labels in multiple classes that complicate the multi-classifications [50], [84]. In order to recognize the suitable models and techniques to deal with multi-classifications, several studies were compared and discussed in this review.

TABLE II
COMPARATIVE STUDY OF FOUR-CLASS CLASSIFICATION USING BACH DATABASE

Study	Pre-trained Model	Database	Data augmentation	Pre-processing	Transfer-learning	Optimization/ regularization	Accuracy % 200 ×
Our method	DPN-131	Even/ Total:400 patches	Rotation, flipping	Resizing, Normalization	Feature extraction	Adam/-	97.5
Our method	NASNet-A-Large	Even/ Total:400 patches	Rotation, flipping	Resizing, Normalization	Feature extraction	Adam/-	97.5
Our method	ResNeXt-101(32×4d)	Even/ Total:400 patches	Rotation, flipping	Resizing, Normalization	Feature extraction	Adam/-	97.5
Our method	Inception-ResNet-V2	Even/ Total:400 patches	-	Resizing, Normalization	Feature extraction	Adam/-	97.5
Our method	SENet-154	Even/ Total:400 patches	Rotation, flipping	Resizing, Normalization	Feature extraction	Adam/-	97.5
[47]	Inception-V3	Even/ Total:400 whole-slide, (for training: 33,600 patches with data augmentation)	Random vertical and horizontal flipping, rotation of 90, 180, 270◦	Stain normalization with color transfer between images	Fine-tuning two last FCs (1024 units and output units)	SGD + momentum 0.9/-	97.08
Our method	ResNeXt-101(64×4d)	Even/ Total:400 patches	Rotation, flipping	Resizing, Normalization	Feature extraction	Adam/-	95
[85]	Inception V3+ IDPN-26 +GBM, logistic regression, SVM	Even/ Total:400 whole-slide	Randomly flipped vertically and horizontally, randomly rotated by 90	Stain normalization, color perturbation scheme	Fine-tuning for Inception-V3	RMSprop+ momentum 0.9 / L1	87.50
[86]	Inception-Resnet-V2	Even/ Total:400 whole-slide, 5,600 patches	Random vertical, horizontal flipping, rotation 90, 180, 270◦	-	Fine-tuning FC with 2048 units and output units.	SGD+ momentum 0.9/ 50% dropout	87.00
[33]	ResNeXt-50	Even /Total:432 whole-slide	Random HSV color space augmentations Random, rotations, reflections, cropping	-	Fine tuning the last two FCs	SGD/ dropout	81.00

B. FOUR-CLASS CLASSIFICATION

1) PREVIOUSLY PUBLISHED LITERATURE REVIEWS

Table II compares the four classifications studies performed using the BACH database. As mentioned earlier, this database contains all four classes of breast cancer (normal, benign, in-situ carcinoma, and invasive carcinoma). The highest accuracy, 97.08%, was achieved by Inception-V3 [47]. In this study, the number of images was tripled to 33,600, employing data augmentation. Additionally, the pre-processing method

was applied; the transfer color between images was used to normalize the histopathological images. The comparison between the two studies showed that Inception-V3 [47] gained a 10% higher accuracy than Inception-Resnet-V2 [86].

Similarly, Vang *et al.* [85] applied Inception-V3 with the dual-path network (DPN) to separate the class of in-situ carcinoma from invasive carcinoma. Two other studies also employed Inception-V3 and shared the same data augmentation and transfer learning methods [47], [85]. Thus, a decent pre-processing method was more important than ever because these two studies were different in terms of the pre-processing methods. Furthermore, Vesal *et al.* [47] showed that Inception-V3 alone could gain better accuracy than the combination of Inception-V3 with DPN as done by another study [85]. Therefore, as more experiments conducted with DPN are needed to reach a better conclusion on the efficacy of this model for breast cancer histopathological images, we have examined DPN per se in our study on the same database to gain better insights. This model will be explained in the next part.

In another study, ResNeXt-50 was examined by using the combination of BACH and BISQUE databases [33]. As the BISQUE database consists of few but precious images, it was included in Table II. This model gained the lowest accuracy score (81%) among all the studies (Table II). All the pre-processing methods were eliminated, thus showing the significance of this technique in improving accuracy. Besides that, ResNet models with more layers were more accurate compared to the same model with fewer layers as shown in Table I. Thus, since further studies on ResNeXt models with more layers are needed to reach a comprehensive conclusion on the accuracy scores of these models, we have examined this model and compared the results in Table II.

In another study, the SDG optimizer was applied together with momentum but without the regularization method [47]. The SDG optimizer was also utilized with dropout regularization by researchers in another two studies [33], [86]. Among all the studies in Table II, only one study employed RMSprop and momentum for optimization, and L1 for regularization [85].

2) OUR EXAMINATIONS

Several experiments have been conducted in our study on the pre-trained models, including DPN-131, NASNet-A-Large, ResNeXt-101, Inception-ResNet-V2, SENet-154, to examine their performance on four classifications using the BACH database. We had applied feature extraction techniques in all of the experiments. The input images were normalized between -1 and 1. We had resized the images before fitting the models. Adam optimizer had been utilized in our study and as there was no fully connected layer, no regularization method was applied. Furthermore, through data augmentation techniques, including rotation (45°), and horizontal and vertical flipping, we could improve the results for almost all the models, except for Inception-ResNet-V2 and ResNeXt-101(32 × 4d). For instance, for DPN-131, we could improve the accuracy score from 92.5% to 97.5% by utilizing these methods. Moreover, by using data augmentation methods, the test accuracy results for NASNet-A-Large, SENet-154, and ResNeXt-101(64×4d) increased by 10%, 2.5%, and 2.5% respectively. However, this metric for Inception-ResNet-V2 decreased to 2.5% while using the data augmentation techniques, and for ResNeXt-101(32 × 4d), it remained unchanged. Additionally, the loss error scores for DPN-131, NASNet-A-Large, ResNeXt-101(32×4d), Inception-ResNet-V2, SENet-154, and ResNeXt-101(64×4d) were 0.05, 0.07, 0.1, 0.22, 0.28, and 0.37, respectively. We had trained all the models for 100 epochs and the best results for DPN-131, NASNet-A-Large, ResNeXt-101(32×4d), Inception-ResNet-V2, SENet-154, and ResNeXt-101(64×4d) were gained in epoch 42, 90, 54, 11, 62, and 86, respectively.

C. EIGHT-CLASS CLASSIFICATION

1) PREVIOUSLY PUBLISHED LITERATURE REVIEWS

According to Table III, the experiment done by Xie *et al.* [14] with Inception-ResNet-V2 achieved the highest accuracy of 97.63% for the eight classifications. In this experiment, the data augmentation techniques were applied to increase the number of images of each subtype besides balancing them, improving the accuracy from 92.07% to 97.63%. After increasing the number of images by the augmentation technique, the input images for benign tumors subtypes included adenosis (1,335), fibroadenoma (3,045), phyllodes-tumor (1,362), and tubular-adenoma (1,710). As for malignant tumor subtypes, the input images included ductal carcinoma (3,451), lobular carcinoma (1,881), mucinous-carcinoma (2,379), and papillary-carcinoma (1,683). Furthermore, pre-processing techniques such as normalization, cutting border, and saturation adjustment were used in this study. Although the researchers did not apply the fine-tuning method, the model of Inception-ResNet-V2 gained the best result (97.63%), followed by DenseNet (95.40%) [67]. As this model consisted of only a few trainable parameters, the researchers [67] retrained all of them using fine-tuning. The third rank of accuracy level of 95% was achieved by Nawaz *et al.* [43] whom employed ResNet-50 and fine-tuning techniques. The number of classes also became balanced through data augmentation methods. Furthermore, with the advantage of stain normalization as the pre-processing method, the extreme values in the breast cancer histopathological slides were normalized.

In general, based on a study by Han *et al.* [50], the CSDCNN model achieved a 93.20% accuracy score with data augmentation, pre-processing, and fine-tuning of the last layer. Among all the Inception models, Inception-V3 obtained the highest level of accuracy score (90.28%) for eight imbalanced classes [14].

TABLE III
COMPARATIVE STUDY OF EIGHT-CLASS CLASSIFICATION USING BREAKHIS DATABASE

Study	Pre-trained Model	Database	Data augmentation	Pre-processing	Transfer-learning	Optimization/ regularization	Results (Accuracy %)				
							40 ×	100 ×	200 ×	400 ×	Avg
[14]	Inception-ResNet-V2	Even /Total: 27,262	Turning, clockwise rotation	Normalization [-1,1]/, cutting border, adjust saturation	Feature extraction	Adam/ exponential decay method	97.63	97.00	96.89	97.49	97.25
Our method	SENet-154	Uneven/Total:7,909	Rotation, flip	Normalization	Fine-tuned the last three FC Layers	Adam/Dropout 0.5	-	-	-	-	96.33
Our method	ResNeXt-101(32×4d)	Uneven/Total:7,909	Rotation, flip	Normalization	Fine-tuned the last three FC Layers	Adam/Dropout 0.5	-	-	-	-	96.20
Our method	ResNeXt-101(64×4d)	Uneven/Total:7,909	Rotation, flip	Normalization	Fine-tuned the last three FC Layers	Adam/Dropout 0.5	-	-	-	-	95.81
Our method	Inception-ResNet-V2	Uneven/Total:7,909	Rotation, flip	Normalization	Fine-tuned the last three FC Layers	Adam/Dropout 0.5	-	-	-	-	95.44
[67]	DenseNet-121	Uneven/Total: 7,909	-	-	Fine-tuning all the layers	Adam/dropout	93.64	97.42	95.87	94.67	95.40
Our method	DualPathNet-131	Uneven/Total:7,909	Rotation, flip	Normalization	Fine-tuned the Layers	Adam/Dropout 0.5	-	-	-	-	95.32
[43]	ResNet-50	Even /Total: More than 7,909	Rotation, flipping to left and right, cropping	Stain normalization, normalizing to [0, 1]	Fine-tuning all the FC layers	RMSProp/ dropout	-	-	-	-	95.00
Our method	NASNet-A-Large	Uneven/Total:7,909	Rotation, flip	Normalization	Fine-tuned the last three FC Layers	Adam/Dropout 0.5	-	-	-	-	95.32
[65]	BHCNet-6	Uneven/ Total:7,909	Height and width shift, horizontal flip, constant fill mode	Down sampling to change image size, zero-mean normalization	-	SGD + momentum of 0.9/ weight decay of 1e-4.	94.43 ± 0.28	94.45 ± 0.15	92.27 ± 0.08	91.15 ± 0.43	93.07 ±0.23
[50]	CSDCNN	Even /Total: More than 7,909	Intensity variation, rotation, flip, translation, and random combination of all	Over-sampling, data augmentation to balance the classes	Fine-tuning the last layers	SGD/L2	92.8 ± 2.1	93.9 ± 1.9	93.7 ± 2.2	92.9 ± 1.8	93.32 ±2.0
[50]	CSDCNN	Even /Total: More than 7,909	-	Over-sampling, data augmentation to balance the classes	Fine-tuning the last layers	SGD/L2	89.4 ± 5.4	90.8 ± 2.5	88.6 ± 4.7	87.6 ± 4.1	89.1± 4.17
[14]	Inception-ResNet-V2	Uneven/Total:7,909	-	Normalization [-1,1]/, cutting border, adjust saturation	Feature extraction	Adam/ exponential decay method	92.07	88.06	87.62	84.50	88.06
[14]	Inception-V3	Uneven/Total:7,909	-	Normalization [-1,1]/, cutting border, adjust saturation	Feature extraction	Adam/ exponential decay method	90.28	85.35	83.99	82.08	85.42

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3029881, IEEE Access

Furthermore, by using six small SE-ResNet modules in the BHCNet-6 architecture, Jiang *et al.* [65] gained a better accuracy score of 92.24% in comparison with Inception-V3 models in another study [14]. However, for all eight classifications, the BHCNet-6 showed a lower accuracy score compared to ResNet-50 [43]. Both of these studies differed in terms of pre-processing and optimization methods [43], [65]. Furthermore, one of the studies utilized the BreakHis database with imbalanced classes [65].

As shown in Table I, BHCNet-3 [65] gained 1.8% more accuracy score than ResNet-50 [16] for the binary classifications. However, for the eight classifications (Table III), BHCNet-6 [65] scored 2.76% less in terms of accuracy than ResNet-50 [43]. Experiments by the ResNet-50 model in both of these studies differed in terms of pre-processing techniques and the number of classifications. Thus, it can be concluded that pre-processing methods such as stain normalization utilized by Nawaz *et al.* [43] can improve accuracy.

2) OUR EXAMINATIONS

Like the two classifications, in this section, we had employed six pre-trained models; DualPathNet-131, SENet-154, ResNeXt-101(32×4d), ResNeXt-101(64×4d), Inception-ResNet-V2, and NASNet-A-Large to classify the data images into eight classes. Each pre-trained model had been trained for about 100 epochs, and the data augmentation techniques like rotation and flipping were used. In all of the experiments, we had normalized the input images between -1 and 1. To have a comprehensive comparison, we added all of the examinations to Table III and arranged them ascendingly based on the accuracy scores. The results of each model are explained as follows.

Like the two classifications, SENet-154 reached the highest accuracy score for the eight classifications among all the examined models in our study. The accuracy score for this model was 96.33% for the test dataset with the loss error score of 0.12 in epoch 100. Additionally, all the layers of the model were frozen except the three fully connected layers with the size of 1024. In the binary classifications, this model obtained higher accuracy results compared to previous researchers [14], but for the eight classifications, this model was capable of obtaining the second rank of accuracy (Table III).

The next experiment on ResNeXt-101(32 × 4d) showed 96.20% of accuracy in epoch 63 for the eight classifications using the BreaKHis database (Table III). The loss error received by this model in epoch 63 was 0.21. Similarly, the experiments on ResNeXt-101 (64 × 4d) showed 95.81% of the accuracy score. Furthermore, the loss error function of this model was 0.11 in epoch 29.

We obtained a 95.44% of accuracy score by using a fine-tuned Inception-ResNet-V2 model with three fully connected layers with the size of 1024 in epoch 57. This model had a 0.14 score of loss error. Although previous studies showed a 97.25% accuracy score by using the feature extraction method with a large number of data images, we gained 93.44% of the accuracy score by these techniques. Previous studies had shown the impact of using a large amount of data images to improve the accuracy score; hence, a study [14] gained 97.25% of the accuracy score by using 27, 262 number of images which was almost 3.5 times more than ours. As examined by Xie *et al.* [14], for the two classifications, the author improved the accuracy score by 2% by increasing the number of input images. Thus, in our experiment, this model could have reached a higher level of score accuracy if a higher amount of input data was used. Thus, our examination assured the reliability and accuracy of this model to gain the best results for the eight classifications.

We had retrained all the layers of the pre-trained DualPathNet-131 model. This model gained 95.32% of the accuracy score and 0.13 of the loss error in epoch 75.

Similarly, in our investigation, NASNet-A-Large gained 93.67% of the accuracy score in epoch 63 and the loss error function was 0.18. This model gained the least accuracy score during our study.

Table III shows that the two highest accuracy scores are gained through experiments using Adam optimizer [14], [67]. Thus, in all of our experiments, we had applied this method to gain better results during training and evaluation. Only one study [43] used RMSProp while the other studies employed SGD [50], [65]. Likewise, only one study applied L2 regularization methods [50] while the rest preferred other methods, such as dropout or weight decay. Thus, the drop out method had been chosen and applied in all of our experiments.

Similarly, Inception-ResNet-V2 achieved the best accuracy for all the eight classes. However, all the models compared in Table III used only images from the BreakHis pathology database. Therefore, further research on other databases is needed to achieve a comprehensive conclusion for the eight classifications.

As shown in Table I, Habibzadeh Motlagh *et al.* [16] performed a binary classification on the BreakHis database. After separating the benign and malignant types of breast cancer, the author conducted other experiments for the four subclasses of benign and malignant cancer, respectively. The input images for the benign subtype of breast cancer included adenosis (1,335), fibroadenoma (3,045), phyllodes-tumor (1,362), and tubular-adenoma (1,710). The input images for the malignant subtypes included ductal-carcinoma (3,451), lobular-carcinoma (1,881), mucinous-carcinoma (2,379), and papillary-carcinoma (1,683).

In another study [16], two histopathological databases (BreaKHis and TMA) were used to conduct experiments on the four classifications. Data augmentation methods, such as resizing, rotating, cropping, and flipping were also applied in their study. In addition to that, they also applied pre-processing techniques of normalization and color distortion, along with the RMSProp optimizer, dropout, and batch normalization. Based on this study, the highest accuracy in classifying the benign subtypes was achieved by ResNet with 50 layers (94.80%) and ResNet with 152 layers (94.50%). Similarly, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3029881, IEEE Access

highest accuracy score to classify the malignant subtypes was 96.40% as accomplished by ResNet with 152 layers [16]. Since the four classifications conducted on the subtypes of benign and malignant cancer in this study were different from the classes illustrated in Table III, this study was not added to the table.

Based on the previous studies that examined breast cancer histopathological images, there were a few types of transfer learning techniques. Among the studies, the Inception-ResNet-V2 model gained far better accuracy scores with the feature extraction method [14] compared with only fine-tuning the last fully connected layers [86].

Moreover, the studies showed that Inception-V3 gained more immeasurable accuracy when the fine-tuning of the last fully connected layers was done [36], [47] compared to feature extracting [14] and fine-tuning all the layers [16]. Meanwhile, ResNet models gained satisfactory results by both fine-tuning the last fully connected layers [43] and all the layers [16]. Similarly, the DenseNet model that fine-tuned all the layers achieved more satisfactory results [67] compared to the other models such as CSDCNN [50], ResNeXt-50 [33], AlexNet [81], and VGG-16 [36] that fine-tuned only the last fully connected layers.

Although the pre-trained deep models achieved a high level of accuracy, the BHCNet-N model could also achieve competent results by using small SENet without using transfer learning [65]. Thus, it can be concluded that simple models without transfer learning are able to achieve acceptable results whereas pre-trained deep models are prone to overparameterization [80]. Besides that, a pre-trained SENet-154 model had been examined for the eight classifications in our study, and this model gained the second rank of the accuracy score among all the investigations.

D. EXPERIMENTAL SETTINGS

1) PREVIOUSLY PUBLISHED LITERATURE REVIEWS

Among all of the discussed studies, the ones that explained the experimental settings are outlined below.

First and foremost, in experimental settings, most of the studies applied GPU in their experiments. However, instead of using GPU, the study by Deniz *et al*. [81] used core i7 CPU with 32 GB memory RAM for the AlexNet model with 61 M parameters. Consequently, this study showed that it was possible to conduct experiments on the models with the same size of AlexNet, and that a lack of GPU would not hinder the process.

As for the software, the TensorFlow framework [87] was popular in the majority of the studies discussed earlier. However, only Du *et al.* [79] applied Caffe [88] as the framework. Another two studies applied MATLAB [89] to perform the pre-processing and data augmentation jobs [79], [50]. Besides that, libraries such as Keras [90] and Pytorch [91] were utilized in two of the studies [36], [33]. Among the different types of programming languages, Python [92] was the most popular language. Apart from that, some researchers [16], [43] conducted their experiments using Linux operating systems such as Centos [93] and Ubuntu [94].

With regard to the training and test split rates, the rates of 70/30 and 80/20 were mostly adopted. Only two studies used a validation set of 20% [85] and 25% [50]. Other methods like exponential decay, weight decay, or Gaussian error scheduler were also applied to decrease the learning rate in studies that showed satisfactory results. The majority of the researchers used 32 batch sizes.

2) OUR EXPERIMENTAL SETTINGS

In our experiments, Pytorch, an open-source machine learning library based on the Torch library [95], had been utilized. We also benefitted from the pre-trained models on the ImageNet database. We had applied the forward method to return the log-SoftMax for the output. Since SoftMax is a probability distribution over the classes, the log-SoftMax is a log probability. By using the log probability, computations are often faster and more accurate [96]. To get the class probabilities later, we employed exponential to inverse the log function. Since the model's forward method returned the log-SoftMax, we used the negative log loss as our criterion to calculate the loss function [97]. We also chose to use Adam optimizer, a variant of stochastic gradient descent which included momentum along with the learning rate of 0.0002. We also used ReLU activations for fine-tuning the layers, followed by drop out 0.5 to return the logits from the forward pass. We used dropout in the network to measure validation loss and accuracy, but in the inference phase, we did not use dropout. Otherwise, the network would appear to perform poorly because many of the connections were turned off. Hence, in the training mode, dropout and autograd were turned on while in the evaluation mode, they were turned off. Thus, before feeding our data into the model, we normalized, and resized the input images for each model.

Then, we trained the model through the batches in our dataset for 100 epochs, sent the data through the network to calculate the losses, obtained the gradients, and then run the optimizer. We also examined DualPathNet-131 through two rates of divisions. We examined this model by using a 65/35 rate of training and test dataset. Even though the data augmentation methods like rotation and flipping were applied, we obtained 97.72% of the accuracy in epoch 75. Yet, by the division of the database to 90/10 for the training and the test dataset, we reached to 99.74% of the accuracy score.

The results showed the superiority of the bigger number of training datasets over the smaller ones. Thus, we applied a 90/10 rate of division for all the examinations. We trained the models on 7,118 number of images and evaluated the models on 790 number of images in each epoch.

Our examinations were performed on Google Colaboratory (also known as Colab) which provided a runtime fully configured for deep learning based on Jupyter Notebooks [98]. We had applied the Colab service pro version with GPUs, like T4 or P100, and 27.4 Gigabytes of the available RAM. Using this service, models like Inception-ResNet-V2, ResNeXt-101

(32x8d), and SENet-154 were managed to be trained in less than 12 hours. Nonetheless, other models, such as NASNet-A-Large, DualPathNet-131, and ResNeXt-101(64x4d) required more time to be trained for 100 epochs, and we could train them within 24 hours.

Furthermore, the settings of the Colab Pro service enabled us to train the models, like Inception-ResNet-V2, SENet154, DualPathNet-131, and ResNeXt-101 (32 x 8d) with the batch size of 32. However, we had the difficulty to retrain ResNeXt-101(64 x 4d) with the 32-batch size because of the number of the parameters and the size of input images. Hence, we decreased the batch size to 16 while training this model to avoid GPU memory error. Moreover, NASNet-A-Large had been trained by using a batch size of 8 due to this error. Finally, during the evaluation, the batch size of one was applied.

VI. CONCLUSION

In the medical field, CNN models are preferred over traditional learning models due to the advantages in terms of speed and reliability. In this paper, the most current and relevant studies were scrutinized for the binary, four, and eight classifications of breast cancer histopathological image databases. For the binary and eight classifications, the studies that examined the deep learning models on the BreaKHis database were compared and outlined in Table I and Table III, respectively. Additionally, Table II shows the studies that compared the models using the BACH database for the four classifications. We also have examined the most current models with high accuracy results on the ImageNet database. These examinations were added to these tables and arranged based on the accuracy results. Among all the examinations, SENet-154 showed the highest accuracy results. Moreover, we have re-examined Inception-ResNet-V2 in our study and gained almost the same accuracy results for this model even though we had applied almost 3.5 times less amount of data while training this model. This examination ensured the accuracy of Inception-ResNet-V2 as this model had the best accuracy for both binary and eight classifications on the BreakHis database.

Although this model could not gain a satisfactory result for the four classes on the BACH database by the other researchers, during our examination, Inception-ResNet-V2 gained the highest accuracy results for the four classifications on the BACH database. Moreover, DPN-131, SENet-154, NASNet-A-Large, and ResNeXt-101 with 32 cardinalates gained the same accuracy results as Inception-ResNet-V2 in our examinations. Moreover, the highest rank of accuracy score for the four classifications on the BACH database was obtained by Inception-V3 by previous studies. Other studies that examined other pathology databases such as TMA and Camelyon were also discussed in this study.

Although SENet-154 could outperform the Inception-ResNet-V2 model in terms of two classifications even with less amount of data, this model achieved almost 1% of the accuracy results than that of Inception-ResNet-V2 trained on a larger amount of data with even numbers of images for the eight classifications. Therefore, the size of the database and balanced classes are important to improve the accuracy results for the eight classifications.

Furthermore, the application of the same deep learning models with different techniques can result in different accuracy scores, indicating the possibility of other significant factors that impact performance. The findings from this study revealed that data augmentation and balance class techniques could be used to improve the accuracy of the models. Moreover, to solve the unbalanced classes and lack of sufficient data, the generative adversarial networks [99] are suggested to generate more data and even out the classes.

Besides that, almost all the studies with high accuracy results applied pre-processing methods such as normalization. Other methods such as optimization and regularization methods were also discussed in this study. Although the pre-trained models showed more effective performances, the BHCNET-N model achieved an impressive result without the use of transfer learning. Thus, further study is needed to assess SENet blocks as they have the potential to be easily embedded in other cutting-edge models to improve accuracy. SENet-154 was also examined in our study, showing cutting edge results that ensured the efficacy of these blocks.

In short, this study shows that different results were obtained when models were examined using different resolutions. This differentiation indicates that deep learning models are weak against the low resolution and high noise. Thus, it is vital to work with the breast cancer histopathological images of suitable resolution and quality. However, the high cost of equipment, such as cutting-edge scanners and data storage, represents the challenges in acquiring high-resolution images [100]. To solve this challenge, super-resolution methods like super-resolution generative networks (SRGAN) [101] have been examined and found to successfully improve the resolution of the breast cancer histopathological images [26], [102]. Thus, future research should focus on investigating the performance of the deep learning models after employing SRGAN models for the pathology images.

ACKNOWLEDGMENT

The authors would like to express their appreciation to the Advanced Informatics Department, Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia for realizing and supporting this research work.

REFERENCES

[1] "WHO | Breast cancer," WHO, 2018. [Online]. Available: https://www.who.int/cancer/prevention/diagnosis-screening/breastcancer/en/, Accessed on: Feb. 15, 2019.

[2] M. Zeeshan, B. Salam, Q. S. B. Khalid, S. Alam, and R. Sayani, "Diagnostic Accuracy of Digital Mammography in the Detection of Breast Cancer," *J. Cureus*, vol. 10, no. 4, Apr. 2018, DOI: 10.7759/cureus.2448, [Online].

[3] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological Image Analysis: A Review," *IEEE Rev. Biomed. Eng.*, vol. 2, pp. 147–171, 2009, DOI: 10.1109/RBME.2009.2034865, [Online].

[4] J. D. Hipp, A. Fernandez, C. C. Compton, and U. J. Balis, "Why a pathology image should not be considered as a radiology image," *J. Pathol. Inform.*, vol. 2, p. 26, Jun. 2011, DOI: 10.4103/2153-3539.82051, [Online].

[5] M. D. Pickles, P. Gibbs, A. Hubbard, A. Rahman, J. Wieczorek, and L. W. Turnbull, "Comparison of 3.0 T magnetic resonance imaging and X-ray mammography in the measurement of ductal carcinoma in situ: A comparison with histopathology," *Eur. J. Radiol.*, vol. 84, no. 4, pp. 603–610, 2015, DOI: 10.1016/j.ejrad.2014.12.016, [Online].

[6] B. E. Bejnordi *et al*., "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA - J. Am. Med. Assoc.*, vol. 318, no. 22, pp. 2199–2210, Dec. 2017, DOI:10.1001/jama.2017.14585, [Online].

[7] L. C. Junqueira, and A. L. Mescher, "Histology & Its Methods of Study," in *Junqueira's basic histology: text & atlas/Anthony L. Mescher*, New York: Editora McGraw-Hill Medical, 2013. pp. 1-17.

[8] 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, July 2016. DOI: 10.1109/TBME.2015.2496264, [Online].

[9] P. W. S. A. Multitouch, Y. Wang, K. E. Williamson, P. J. Kelly, J. A. James, and P. W. Hamilton, "SurfaceSlide : A Multitouch Digital Pathology Platform SurfaceSlide : A Multitouch Digital Pathology Platform," *PloS one*, vol. 7, no. 1, 2012, DOI: 10.1371/journal.pone.0030783. [Online].

[10] S. U. Akram, T. Qaiser, S. Graham, J. Kannala, J. Heikkilä, and N. Rajpoot, "Leveraging unlabeled whole-slide-images for mitosis detection," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11039 LNCS, pp. 69–77, Jul. 2018, DOI: 10.1007/978-3-030-00949-6_9, [Online].

[11] J. Z. Cheng *et al.*, "Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans," *Sci. Rep.*, vol. 6, no. 1, pp. 1–13, 2016, DOI: 10.1038/srep24454, [Online].

[12] S. M. Ismail *et al*., "Observer variation in histopathological diagnosis and grading of cervical intraepithelial neoplasia," *British Medical Journal*, vol. 298, no. 6675, pp. 707–710, Mar. 1989, DOI: 10.1136/bmj.298.6675.707, [Online].

[13] A. Andrion *et al*., "Malignant mesothelioma of the pleura: Interobserver variability," *J. Clin. Pathol*., vol. 48, no. 9, pp. 856–860, Sep. 1995, DOI: 10.1136/jcp.48.9.856, [Online].

[14] J. Xie, R. Liu, J. Luttrell, and C. Zhang, "Deep learning based analysis of histopathological images of breast cancer," *Frontiers in Genetics*, vol. 10:80, Feb. 2019, DOI: 10.3389/fgene.2019.00080, [Online].

[15] H. Chen, Q. Dou, X. Wang, J. Qin, and P. Heng, "Mitosis detection in breast cancer histology images via deep cascaded networks," in *Proc. 13th AAAI Conf. Artif. Intell.*, Phoenix, Arizona, USA, 2016, pp. 1160–1166.

[16] N. H. Motlagh *et al*., "Breast Cancer Histopathological Image Classification: A Deep Learning Approach," *bioRxiv*, pp. 1–8, 2018, DOI: 10.1101/242818, [Online].

[17] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, "Methods for nuclei detection, segmentation, and classification in digital histopathology: A review-current status and future potential," *IEEE Rev. Biomed. Eng.*, vol. 7, pp. 97–114, 2013, DOI: 10.1109/RBME.2013.2295804, [Online].

[18] R. Barroso-Sousa and O. Metzger-Filho, "Differences between invasive lobular and invasive ductal carcinoma of the breast: results and therapeutic implications," *Ther. Adv. Med. Oncol.*, vol. 8, no. 4, pp. 261–266, Jul. 2016, DOI: 10.1177/1758834016644156, [Online].

[19] J. Sariego, "Breast cancer in the young patient," *Am. Surg.*, vol. 76, no. 12, pp. 1397–1400, 2010, DOI: 10.1016/S0002-9610(05)80007-8, [Online].

[20] F. A. Vicini, L. Kestin, P. Chen, P. Benitez, N. S. Goldstein, and A. Martinez, "Limited-Field Radiation Therapy in the Management of Early-Stage Breast Cancer," *J. Natl. Cancer Inst. (JNCI)*, vol. 95, no. 16, pp. 1205–1210, Aug. 2003, DOI: 10.1093/jnci/djg023, [Online].

[21] *W. H. Organization: Diagnosis and treatment*. [Online]. Available: https://www.who.int/cancer/treatment/en/.

[22] R. G. Holzheimer and J. A. Mannick, "Multimodality treatment for hepatocellular carcinoma," in *Surgical treatment: evidence-based and problem-oriented*. Munich, Germany: Zuckschwerdt, 2001. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK6903/.

[23] R. Lin and P. Tripuraneni, "Radiation therapy in early-stage invasive breast cancer," *Indian J. Surg. Oncol.*, vol. 2, no. 2, pp. 101–111, Jun. 2011, DOI: 10.1007/s13193-011-0048-8, [Online] .

[24] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Trans. Neural Networks*, vol. 14, no. 6, pp. 1569–1572, Nov. 2003, DOI: 10.1109/TNN.2003.820440, [Online].

[25] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017, DOI: 10.1016/j.media.2017.07.005, [Online].

[26] G. Çelik and M. F. Talu, "Resizing and cleaning of histopathological images using generative adversarial networks," *Phys. A Stat. Mech. its Appl.*, p. 122652, Sep. 2019, DOI: 10.1016/j.physa.2019.122652, [Online].

[27] R. J. Marinelli *et al.*, "The Stanford Tissue Microarray Database," *Nucleic Acids Res.,* vol. 36, no. suppl 1, pp. D871-D877, 2008, DOI: 10.1093/nar/gkm861, [Online].

[28] C. Kampf, I. Olsson, U. Ryberg, E. Sjöstedt, and F. Pontén, "Production of tissue micro arrays, immunohisto chemistry staining and digitalization within the human protein atlas," *J. Vis. Exp. (JoVE)*, no. 63, p. e3620, May 2012, DOI: 10.3791/3620, [Online].

[29] (2016). *Camelyon16 Challenge on cancer metastases detection in lymph node*. [Online]. Available: https://camelyon16.grand-challenge.org.

[30] P.-H. C. Chen *et al.*, "An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis," *Nat. Med.*, vol. 25, pp. 1453–1457, Aug. 2019, DOI: 10.1038/s41591-019-0539-7, [Online].

[31] Z. Guo *et al.*, "A Fast and Refined Cancer Regions Segmentation Framework in Whole-slide Breast Pathological Images," *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, 2019, DOI: 10.1038/s41598-018-37492-9, [Online].

[32] A. Sarmiento and I. Fondón, "Automatic Breast Cancer Grading of Histological Images Based on Colour and Texture Descriptors," in *Proc. Image Analysis and Recognition*, Springer, Cham, Jun. 2018, pp. 887–894.

[33] I. Koné and L. Boulmane, "Hierarchical ResNeXt Models for Breast Cancer Histology Image Classification," in *Proc. Int. Conf. on Document Analysis and Recognition*, 2018, pp. 796–803, DOI: 10.1007/978-3- 319-93000-8_90, [Online].

[34] J. Lemley, S. Bazrafkan, and P. Corcoran, "Smart Augmentation Learning an Optimal Data Augmentation Strategy," *IEEE Access*, vol. 5, pp. 5858–5869, 2017. DOI: 10.1109/ACCESS.2017.2696121, [Online].

[35] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, pp. 216–231, 2019, DOI: 10.1016/j.patcog.2019.02.023, [Online].

[36] M. J. Lim, D. E. Kim, D. K. Chung, H. Lim, and Y. M. Kwon, "Deep convolution neural networks for medical image analysis," *Int. Eng. & Tech.,* vol. 7, no. 3, pp. 115–119, Aug. 2018, DOI: http://dx.doi.org/10.14419/ijet.v7i3.33.18588, [Online].

[37] M. A. Al-antari, M. A. Al-masni, M.-T. Choi, S.-M. Han, and T.-S. Kim, "A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification," *Int. J. Med. Inform.*, vol. 117, pp. 44–54, Sep. 2018. DOI: https://doi.org/10.1016/j.ijmedinf.2018.06.003, [Online].

[38] A. D. Belsare, and M.M. Mushrif, "Histopathological Image Analysis Using Image Processing Techniques: An Overview," *Signal Image Process. An Int. J.*, vol. 3, no. 4, pp. 23–36, Aug. 2012, DOI : 10.5121/sipij.2012.3403, [Online].

[39] J. Sun and A. Binder, "Comparison of deep learning architectures for H&E histopathology images," in *Proc. 2017 IEEE Conf. Big Data Anal. (ICBDA)*, Nov. 2017, pp. 43–48, DOI: 10.1109/ICBDAA.2017.8284105, [Online].

[40] E. Yuan and J. Suh, "Neural Stain Normalization and Unsupervised Classification of Cell Nuclei in Histopathological Breast Cancer Images," *CoRR*, vol. abs/1811.0, Nov. 2018. Available: https://arxiv.org/pdf/1811.03815.pdf, [Online].

[41] R. D. Fiete, "Image Enhancement Processing," in *Modeling the Imaging Chain of Digital Cameras*, 1000 20th Street, Bellingham, WA 98227-0010 USA: SPIE, 2010, pp. 127–161.

[42] M. Baatz and A. Schäpe, "Multiresolution segmentation: An optimization approach for high quality multi-scale image segmentation," *Strobl, J., Blaschke, T. and Griesbner, G., Eds., Angewandte Geographische Informations-Verarbeitung, XII, Wichmann Verlag*, pp. 12–23, 2000.

[43] M. A. Nawaz, A. A. Sewissy, and T. H. A. Soliman, "Automated classification of breast cancer histology images using deep learning based convolutional neural networks," *Int. J. Comput. Sci. Netw. Secur. (IJCSNS),* vol. 18, no. 4, pp. 152–160, Apr. 2018.

[44] M. Macenko *et al.*, "A method for normalizing histology slides for quantitative analysis,"in *Proc. - 2009 IEEE Int. Symp. Biomed. Imaging From Nano to Macro, ISBI 2009*, vol. 9, pp. 1107–1110, Jun. 2009, DOI: 10.1109/ISBI.2009.5193250, [Online].

[45] A. C. Ruifrok, R. L. Katz, and D. A. Johnston, "Comparison of quantification of histochemical staining by hue-saturation-intensity (HSI) transformation and color-deconvolution.," *Appl. Immunohistochem. Mol. Morphol. AIMM*, vol. 11, no. 1, pp. 85–91, Mar. 2003.

[46] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Comput. Graph. Appl.*, vol. 21, no. 4, pp. 34–41, 2001, DOI: 10.1109/38.946629, [Online].

[47] S. Vesal, N. Ravikumar, A. A. Davari, S. Ellmann, and A. Maier, "Classification of Breast Cancer Histology Images Using Transfer Learning," in *Proc. 15th Int. Conf. Image Anal. Recognit. (ICAIR 2018)*, Póvoa de Varzim, Portugal, Feb. 2018, pp. 812–819, DOI: 10.1007/978-3-319-93000-8_92, [Online].

[48] M. I. Razzak, S. Naz, and A. Zaib, "Deep Learning for Medical Image Processing: Overview, Challenges and the Future BT - Classification in BioApps: Automation of Decision Making," in *Classification in BioApps*. Cham, Switzerland: Springer, 2018, pp. 323–350, DOI: 10.1007/978-3-319-65981-7_12, [Online].

[49] Y. Lecun, L. Bottou, Y. Bengio, and P. Ha, "Gradient-Based Learning Applied to Document Recognition," *IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998, DOI: 10.1109/5.726791, [Online].

[50] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, "Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model," *Sci. Rep.*, vol. 7, no. 1, 2017. [Online] Available: https://www.nature.com/articles/s41598-017-04075-z.

[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Commun. Acm*, vol. 60, no. 06, pp. 84–90, 2017, DOI: 10.1145/3065386, [Online]

[52] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. 2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016. pp. 770-778.

[54] M. Mahmoud, "Breast Cancer Classification in Histopathological Images using Convolutional Neural Network," *Int. J. Adv. Comput. Sci. Appl. (IJACSA)*, vol. 9, no. 3, pp. 64–68, 2018.

[55] M. Lin, Q. Chen, and S. Yan, "Network In Network," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–10.

[56] G. Zeng, Y. He, Z. Yu, X. Yang, R. Yang, and L. Zhang, "Going Deeper with Convolutions Christian," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, pp. 1-9, 2015.

[57] F. Sultana, A. Sufian, P. Dutta. "Advancements in Image Classification using Convolutional Neural Network", in *Proc 4th Int. Conf. on Research in Computational Intelligence and Communication Networks (ICRCICN)*, 2018, pp. 122-129, DOI: 10.1109/ICRCICN.2018.8718718, [Online].

[58] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proc. 2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2818–2826.

[59] S. Ioffe and C. Szegedy, "Batch Normalization : Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proc. the 32nd Int. Conf. on Machine Learning,* Lille, France, 2015. vol. 37.

[60] N. Kumaran and A. Vaidya, "Batch Normalization and Its Optimization Techniques : Review," *Int. J. Eng. R. Comput. Sci. Eng. (IJERCSE),* vol. 4, no. 8, pp. 211–215, Aug. 2017.

[61] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intel (AAAI-17)*, San Francisco, California, USA, 2017, pp. 4278- 4284.

[62] S. Zagoruyko and N. Komodakis, "Wide Residual Networks," in *Proc. Br. Mach. Vis. Conf. 2016, BMVC 2016*, York, UK, 2016, pp. 87.1-87.12.

[63] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),* Las Vegas, NV, USA, 2017, pp. 1492-1500.

[64] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. 2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2018, pp. 7132–7141.

[65] Y. Jiang, L. Chen, H. Zhang, and X. Xiao, "Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module," *PLoS One,* vol. 14, no. 3, pp. 1–21, Mar. 2019, DOI: https://doi.org/10.1371/journal.pone.0214587, [Online].

[66] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2017, pp. 4700-4708.

[67] M. Nawaz, A. A., and T. Hassan, "Multi-class breast cancer classification using deep learning convolutional neural network," *Int. J. Adv. Comput. Sci. Appl. (IJACSA)*, vol. 9, no. 6, pp. 316–322, 2018, DOI: 10.14569/IJACSA.2018.090645, [Online].

[68] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," *Adv. Neural Inf. Process. Syst.*, pp. 4468–4476, 2017.

[69] M. Wistuba, A. Rawat, and T. Pedapati, "A survey on neuralarchitecture search," *CoRR*, vol. abs/1905.01392, 2019. [Online]. Available: http://arxiv.org/abs/1905.01392.

[70] B. Zoph and Q. V. Le, ''Neural architecture search with reinforcement learning,'' 2016, arXiv:1611.01578. [Online]. Available: http://arxiv.org/abs/1611.01578

[71] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, ''Learning transferable architectures for scalable image recognition,'' in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 8697–8710.

[72] Y. Chen *et al.*, "RENAS: Reinforced evolutionary neural architecture search," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 4787–4796, 2019.

[73] K. Radhika, K. Devika, T. Aswathi, P. Sreevidya, V. Sowmya, and K. P. Soman, "Performance Analysis of NASNet on Unconstrained Ear Recognition," in *Nature Inspired Computing for Data Science*. Cham: Springer, 2020, pp. 57–82, DOI: 10.1007/978-3-030-33820-6_3, [Online].

[74] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64270–64277, 2018, DOI: 10.1109/ACCESS.2018.2877890, [Online].

[75] Z. Li and D. Hoiem, "Learning without Forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2018, DOI: 10.1109/TPAMI.2017.2773081, [Online].

[76] Y. You and J. Demmel, "Runtime Data Layout Scheduling for Machine Learning Dataset," in *Proc. 46th Int. Conf. Parl. Proc. (ICPP)*, 2017, pp. 452–461, DOI: 10.1109/ICPP.2017.54, [Online].

[77] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang, and Q. Sun, "Deep learning for image-based cancer detection and diagnosis − A survey," *Pattern Recognit.*, vol. 83, pp. 134–149, 2018, DOI: 10.1016/j.patcog.2018.05.014, [Online].

[78] R. Mehra, "Breast cancer histology images classification : Training from scratch or transfer learning ?," ICT Express, vol. 4, no. 4, pp. 247–254, 2018, DOI: 10.1016/j.icte.2018.10.007, [Online].

[79] Y. Du *et al.*, "Classification of Tumor Epithelium and Stroma by Exploiting Image Features Learned by Deep Convolutional Neural Networks," *Ann. Biomed. Eng.*, vol. 46, no. 12, pp. 1988–1999, Dec. 2018, DOI: 10.1007/s10439-018-2095-6, [Online].

[80] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding Transfer Learning for Medical Imaging," in *Advances in Neural Information Processing Systems(NeurIPS)*. 2019. pp. 3342-3352.

[81] E. Deniz, A. Sengür, Z. Kadiroglu, Y. Guo, V. Bajaj, and Ü. Budak, "Transfer learning based histopathologic image classification for breast cancer detection," *Heal. Inf. Sci. Syst.*, vol. 6, no. 1, pp. 1–7, 2018, DOI: 10.1007/s13755-018-0057-x, [Online].

[82] S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Classification of Histopathological Biopsy Images Using Ensemble of Deep Learning Networks," *arXiv preprint arXiv:1909.11870*, 2019.

[83] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*.

[84] S. Robertson, H. Azizpour, K. Smith, and J. Hartman, "Digital image analysis in breast pathology—from image processing techniques to artificial intelligence," *Translational Research*, vol. 194, pp. 19–35, 2018, DOI: 10.1016/j.trsl.2017.10.010, [Online].

[85] Y. S. Vang, Z. Chen, and X. Xie, "Deep Learning Framework for Multi-class Breast Cancer Histology Image Classification," in *Proc. Image Anal. Recognit. ICLR*, 2018, pp. 914–922, DOI: 10.1007/978-3-319-93000-8_104, [Online].

[86] S. Kwok, "Multiclass Classification of Breast Cancer in Whole-Slide Images," in *Proc. Image Anal. Recognit. (ICLR)*, 2018, pp. 931–940, DOI: 10.1007/978-3-319-93000-8_106, [Online].

[87] M. Abadi, A. Agarwal et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv:1603.04467*, 2016.

[88] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 2014 ACM Conf. Multimedia,* Orlando, Florida, USA, 2014, pp. 675–678, DOI: 10.1145/2647868.2654889, [Online].

[89] (1984). *MATLAB: A multi-paradigm numerical computing environment and proprietary programming language developed by MathWorks*. [Online]. Available: https://www.mathworks.com/.

[90] (2015). F. Chollet. *Keras*. [Online]. Available: http://keras.i /.

[91] (2016). PyTorch: An open source machine learning library based on the Torch library. [Online]. Available: https://pytorch.org/.

[92] (1991). *Python: An interpreted, high-level, general-purpose programming language*. [Online]. Available: https://www.python.org/.

[93] (2004). *CentOS: A Linux distribution that provides a free, communitysupported computing platform*. [Online]. Available: https://www.centos.org/.

[94] (2004). *Ubuntu: A free and open-source Linux distribution based on Debia*. [Online]. Available: https://ubuntu.com/.

[95] A. Paszke et al., ''PyTorch: An imperative style, high-performance deep learning library,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[96] A. De Brébisson and P. Vincent, "An exploration of softmax alternatives belonging to the spherical loss family,"in *Proc. 4th Int. Conf. Learn. Represent. (ICLR), 2015*. [Online]. Available: https://arxiv.org/abs/1511.05042

[97] H. Yao, D. lai Zhu, B. Jiang, and P. Yu, "Negative Log Likelihood Ratio Loss for Deep Neural Network Classification,"in *Proc. Future Tech. Conf. (FTC), 2019*, pp. 276–282.

[98] T. Carneiro, R. V. M. Da Nobrega, T. Nepomuceno, G. Bin Bian, V. H. C. De Albuquerque, and P. P. R. Filho, "Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications," *IEEE Access*, vol. 6, pp. 61677–61685, 2018.

[99] I. J. Goodfellow *et al.*, "Generative Adversarial Nets," in *Proc. Adv. in Neural Inf. Proc, Syst.*, pp. 2672-2680, 2014.

[100] L. Mukherjee, A. Keikhosravi, D. Bui, and K. W. Eliceiri, "Convolutional neural networks for whole-slide image superresolution," *Biomed. Opt. Express*, vol. 9, no. 11, pp. 5368-5386, 2018, DOI: 10.1364/BOE.9.005368, [Online].

[101] C. Ledig *et al.*, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition (CVPR)*, 2017, pp. 105–114.

[102] U. Upadhyay and S. P. Awate, "Robust Super-Resolution Gan, with Manifold-Based and Perception Loss," in *Proc. 2019 IEEE 16th Int. Symposium on Biom. Imaging (ISBI)*, 2019, pp. 1372-1376, DOI: 10.1109/ISBI.2019.8759375, [Online].

FAEZEHSADAT SHAHIDI received a B.Sc. degree from the Department of Computer Science and Engineering, University of Erfan Higher Education Institute (Sealed by Selected University of Kerman province), Iran. She has recently completed a master's degree in Master of Science Business Intelligence & Analytics, Universiti Teknologi Malaysia. Her research interests are computer vision and image processing, big data, cloud computing, deep learning, and generative adversarial networks.

SALWANI MOHD DAUD is a Professor of the Advanced Informatics Department in Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia (UTM). She obtained her B.Eng. (Hons) Electronics Engineering from the University of Liverpool in 1984. Then, she received her M.Eng. in Electrical Engineering in 1989 and Ph.D. in Electrical Engineering in 2006 from Universiti Teknologi Malaysia. She is a member of the Institute Electrical Electronic Engineer (IEEE), a registered Professional Technologist from Malaysia Board of Technologists (MBOT) and registered graduate engineer with the Board of Engineers Malaysia (BEM). She has been with UTM for more than 30 years and has vast experience in teaching and research. Her research area is focusing on artificial intelligence, blockchain, and IoT. Currently, she is teaching machine learning and system design for security for the postgraduate program. She is also leading few research grants in the related topics, securing more than RM2 million of R&D funds. She also has published more than 100 academic articles in journals, proceedings, and books. Currently, she is heading the Cyber-Physical Systems Research Group.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3029881, IEEE Access

HAFIZA ABAS serves Universiti Teknologi Malaysia (UTM) as an academic staff for 20 years. Today, she is a senior lecturer at the Razak Faculty of Technology and Informatics, UTM Kuala Lumpur. Hafiza Abas has a Ph.D. in Information Science from Universiti Kebangsaan Malaysia (UKM), MSc in Information Technology from Universiti Putra Malaysia (UPM), and BSc (Hons) in Information Technology from Universiti Utara Malaysia. Her perseverance permits her to grasp 28 research grants awarded by various agencies. She shines in emotional intelligence, social, and soft skills. Academically, she endures collaborating with academic scholars from various universities. At the same time, she has also published and present papers at local and international conferences. To sustain her role as an academic scholar, she is currently involved in a few professional associations. She has obtained research and other awards related to writing and publishing articles. She is also involved in corporate social responsibility (CSR) projects for dyslexic children and down syndrome to help them in learning.

NOOR AZURATI AHMAD serves as an Associate Professor at the Faculty of Technology and Informatics Razak, Universiti Teknologi Malaysia Kuala Lumpur. She obtained her B.Eng. in Computer Engineering in 2001 and Master of Electrical Engineering in 2006 from Universiti Teknologi Malaysia. She graduated with a Ph.D. in Embedded Systems from the University of Leicester in 2013. She is a Certified Tester Foundation Level (CTFL) under the Malaysian Software Testing Board (MSTB) and Certified Professional for Requirements Engineering (CPRE) under the International Requirements Engineering Board (IREB). She is a member of the Institute Electrical Electronic Engineer (IEEE), IEEE Computer Society, and Registered Graduate Engineer with the Board of Engineers Malaysia (BEM). She has been actively involved in research related to embedded real-time systems and mobile and pervasive computing. She also successfully delivered outputs in Big Data Cybersecurity using the Machine Learning project which costs RM2.2 million. She collaborated with many industrial projects include Sapura Secured Technology, Cybersecurity Malaysia, Unlock Design, and Drabot. She is now a Deputy Director for Innovation and Commercialization Center, UTM.

NURAZEAN MAAROP received her Ph.D. in Information System (Health Informatics) from the University of Wollongong, Australia, in July 2013. She is currently attached as a senior lecturer to Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, and has worked with the university for more than 16 years. She published more than 50 indexed papers and supervised more than 30 postgraduate students. Her research area of interest is related to any current Information Systems Issues, Technology Acceptance and Adoption, Information Systems Success Model, System Usability, Health Informatics, Business Informatics, Educational Informatics, Governance Informatics, Information Systems Security & Assurance Issues, Information Engineering & Data Science, Data Mining & Big Data Issues, Operation Research, and Enterprise Architecture. She is also interested in Research Methodology in Information Systems occupying Mixed Methods, Qualitative and Quantitative methods. Her main research grant works are related to the Information Systems research area and some application development.