Classification of Hematoxylin and Eosin-Stained Breast Cancer Histology Microscopy Images using Ensemble Transfer Learning with EfficientNets

Nikiel Ramawthar
School of Maths, Stats and Computer Science
University of KwaZulu-Natal
Westville,Durban,3629
218007620@stu.ukzn.ac.za
Correspondence should be addressed to Serestina Viriri
viriris@ukzn.ac.za

ABSTRACT

Breast cancer continues to be one of the most invasive cancers among women and is the second leading cause of cancer death in women after lung cancer. Cancer is best treated and fought once it is caught in the early stages. To pick up on this quickly, a proper analysis of histology images is vital. Computer-Aided Detection (CAD) systems have been developed to overcome restrictions such as human error, misdiagnosis by a radiologist, and image quality. While CAD systems have increased the reliability of a diagnosis, these systems are typically used as a second opinion as final decisions are made by the radiologist. Many of these systems use handcrafted approaches for feature extraction, each tailored for their own interpretation. This project aims at taking a different route on things by looking into deep learning and using ensemble transfer learning with EfficientNets to classify Hematoxylin and Eosin-Stained Breast Cancer Histology Microscopy Images. We investigate two stain normalization techniques, Macenko and Reinhard staining techniques, and the effects these techniques, along with the use of ensemble transfer learning, has on classifying whether the histology microscopy images are either normal, benign, in situ carcinoma, or invasive carcinoma breast tissue The experiments conducted have produced modes which have achieved accuracies of 80%, 92%, and 100%.

KEYWORDS

Ensemble Learning, EfficientNets, Histology Microscopy Images, Breast Cancer, Deep Learning

1. INTRODUCTION AND BACKGROUND

Cancer occurs when changes called mutations take place in genes that regulate cell growth. The mutations let the cells divide and multiply in an uncontrolled way [1]. Breast Cancer typically occurs in the breast cells. It affects both males and females, although it leans more towards the female spectrum.

In South Africa, 1 in 25 women will develop breast cancer in their lifetime [8]. Globally it has taken the lives of 650,000 registered people in the past year [7].

The effects of having breast cancer psychologically are that many begin to self-blame and, because of this, spiral into depression and develop anxiety. When it comes to treatment, many fear the removal of the breasts, causing more anxiety and depression within them. First, a physical exam is conducted, then a mammogram and/or an ultrasound is conducted. If your doctor suspects that it may be cancer, they will request a biopsy to be performed where a sample tissue is taken and stained with Hematoxylin and eosin and is observed via an optic microscope. Pathologists search for signs of cancer by analyzing its histological properties.

Hematoxylin is used to illustrate nuclear details in cells, while eosin is a counterstain to stain the cytoplasm and nuclei of cells. The two dyes used are typically not to get certain information mixed up. As known, human cells consist of a nucleus that contains nuclear DNA, cytoplasm, mitochondria, and cell membrane. To stop the risk of results that may negatively impact the pathologist's view of the cell. E.g., if the cytoplasm is stained with Hematoxylin, the

pathologist may get confused in differentiating the different structures of the cell, thereby possibly creating a misdiagnosis of the patient. [5]

Pathologists now look at the stained cells under a microscope and deduce by looking at various features whether the tissue is cancerous, noncancerous, or normal tissue. The information received can also tell pathologists the type of cancer, i.e., whether it is invasive or noninvasive, by looking at surrounding tissue.

This process, while being the most trusted and seemingly reliable way to determine cancer, is cost-inefficient, time-consuming, and prone to human error. Figure 1. Depicts the different breast tissue examples after undergoing the Hematoxylin and eosin-based staining procedure.

Carcinomas generally start in the milk duct or the milk-producing glands; thus, an in-situ carcinoma is a cancer that has not spread or invaded tissue outside these specific parts [2]. IF left untreated, the cancer cells may develop the ability to spread to surrounding tissue making the cancer invasive [4]. Carcinomas are said to be invasive if it has spread into the surrounding breast tissue. This is typically the worst-case scenario [2].

Humans have been prone to errors, which at times have cost the lives of many. A human is said to make around 3-6 errors per hour and averages to 50 errors per day. [6] Misdiagnosis is no stranger to the medical community. There is not a great deal of reliable data, but by some estimates, a misdiagnosis is anywhere between 5-28% [9]. It is not alarming, but in contrast, would you get on a plane with a 30% chance of death?

Misdiagnosis could range from a variety of factors – such as the similarities between carcinomas and non-carcinomas. Roughly everything is digitized, so why not add this task onto it? Why not use the power of image analysis to remove the error. A computer could see more than a human eye. An example is steganography and the art of hiding malicious code in images for attacks. A computer, if trained and prepared for this, can ultimately remove the stress doctors have to face about misdiagnosis and the time spent waiting for pathologists to reach a conclusion on a sample of tissue. Although there has been technology already introduced for this line of work (CAD), anything and everything can be improved with the arrival of new techniques. Deep learning being the newly found protagonist, has sparked surprising and promising results.

1.1 Research Problem Early and accurate detection of breast cancer is important in the fight, as well as treatment plans and strategies. Cancer is best fought when caught early. Technology has been invented to aid humans in this field, but many times is used as a second opinion, or the doctor gets a say once it is processed. Many techniques have been implemented to categorize breast tissue as normal, benign, in situ carcinoma, or invasive carcinoma, although none have reached the 100% mark in prediction. The problem at hand is how do we make a computer think like a human? How does a human brain categorize the different tissues into their correct category? For a start, ensemble transfer learning uses predictions from multiple models to come to a conclusion.

The proposed research is to explore the use of EfficientNets for the classification of Hematoxylin and Eosin-stained breast histology microscopy images. In more so, it will be training particular features across different EfficientNets architectures and combining their predictions at the end to classify an image as being normal tissue, benign lesion, in situ, or invasive carcinoma.

1.2 Research Contributions. In this research, the combination of the four best EfficientNets versions with transfer learning and different ensemble learning techniques for breast cancer histology image classification is investigated. The proposed architecture was able to effectively extract and learn global features. One of the combinations of EfficientNets models had produced a result of 100% accuracy.

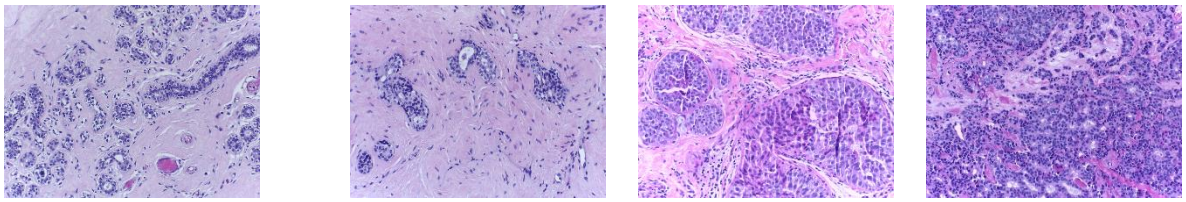

Figure 1: Example of breast tissue after hematoxylin and eosin-based staining procedure. Taken from the iciar 2018 dataset [3]

The remainder of the paper is structured as follows: Section 2, Literature Review, provides details on previous successful approaches. Section 3, the methods and techniques, provides insight into the framework followed in this study. Section 4, the results, provides details of the results which was obtained during the research. Finally, Section 5 concludes and elaborates on the insights of this work.

2. LITERATURE REVIEW

A machine learning approach had been taken, where the tissue was analyzed and based on inspection of the nuclei, classified as being benign or malignant [10]. This approach by Veta et al. [10] uses the feature analysis method. Where images are classified if they share similar features with respective classes. Here you are training the machine to look only for specific traits/ similarities. Various clustering algorithms were used for nuclei segmentation on fine needle biopsy microscopy images; this approach was undertaken by Kowal et al. [11]. To train the classifier, they had used three different types of features, namely morphological, topological, and texture. This technique, upon running, achieved a remarkable 84%-93% accuracy. A patient-wise classification resulted in a 96%-100% accuracy rating. Building on this, Filipczuk et al. [12] and George et al. [13] decided to look at nuclei-based features from fine needle biopsies, more particularly the extraction of shape and texture features of the nuclei using the watershed method by George et al. [13]. The use of color-texture graphs in the exploration of tissue organization and using statistical texture features for training the classifier was an approach undertaken by Belsare et al. [14].

Brook et al. [15] and Zhang et al. [16] considered the 3-class classification with Brook et al. [15] decided to go with the Support vector machine approach by using multiple threshold values to binarize images and train the machine. A cascade classification strategy was employed by Zhang et al. [16]. A first set of concurrent SVM classifiers was randomly fed subsets of Curvelet Transform and local binary pattern (LBP) features. Images were rejected if a certain number of classifiers disagreed. Then compared to other random features using a second set of artificial neural networks (ANN). The traditional methods does have a downfall thou. They're only as good as the programmer. Meaning they see what they are programmed to see, which could cause downfall or incorrect results. These traditional machine learning techniques have a flaw – overfitting. Where the classifier looks for characteristics from certain features. This does not impose a threat on supervised learning where the outcomes are known, but real-world problems are unsupervised, and outcomes are not known. This could lead to misdiagnosis when deployed in the field. These approaches also need to be fed features to carry out classification. These features are typically chosen beforehand and can lead to overfitting.

2.1 Neural Networks. The application of neural networks has a major advantage. It doesn't require a large dataset to be trained on as compared to the traditional methods. The Multilayer perceptron Neural network is a feed-forward neural network that utilizes the backpropagation technique for learning. The input layer acts as receivers, and one or more hidden layers of neurons compute the data, undergoing iterations. The output layer predicts an output. Ayer et al. [17] used a 3-layer ANN for classifying 256 mammogram images. With the use of

4 ANN, the ML-NN was 84% sensitive and 75% specific.

Ting and Sim. [18] implemented a supervised ML-NN by considering 170 labeled mammograms and achieved a 91% accuracy with a sensitivity of 90.53%. ANNs can ease model structure, capacity in catching communications among indicators, and capacity to think about convoluted nonlinearities among indicators and results [17]. They also come with a downside as ANNs are prone to overfitting due to the complexity of the model structure. ANNs are seen as black boxes, meaning we cannot know how much each independent variable is influencing the dependent variables [17].

2.2 Convolutional Neural Network. Due to the advancement of computational power, the use of CNN has increased. Contrary to the traditional approaches of handcrafted feature extraction, CNNs learn useful features straight from the training set by the optimization of classification loss problems.

Spaniel et al. [19] utilized a CNN architecture with ImageNet network to classify H&E breast tissue biopsy samples into either benign or malignant tumors with the use of multiple magnifications. 32 x 32 and 64 x 64-pixel patches were obtained with the combination of the patch probabilities with sum, product, or maximum roles. Two of these extraction methods were studied, namely sliding window and random extraction. These patches allowed to reduce the complicity of the model as the size of input in the layers was decreased. They found that the accuracy decreased for higher magnifications, suggesting the CNN architecture used could not extract features for higher magnifications.

Ciresan et al. [20] adjusted the architecture of the CNN by using 101 x 101 patches to train a CNN for the detection of mitosis in H&E stain breast histology images. The architecture studied nuclei of different sizes and their neighborhoods. The architecture was a feed-forward DNN consisting of a series of convolution, and max-pooling layers are employed. Using the public MITOS dataset, which included 50 images, it was split into three different subsets. Their model obtained an accuracy of 88%. Nuclei, although can be in various positions or rotations, thus the use of augmented rotations and mirroring as used to provide a way to increase the training set without damaging the quality.

Ragab et al. [21] used Alexnet (a type of DCNN) in the classification of breast mammogram images into two classes) the network consisted of 5 convolutional layers, three pooling layers, and two connected layers. The CLAHE method was utilized for image preprocessing. Rotation was used for data augmentation to increase the size of the input. Converting the images to greyscale and using threshold values to convert the image to binary images was done. The number of pixels is counted, and thus the largest binary object is labeled as a tumor. Multiplication of the binary image along with the input image gives us the output image. The accuracy of this algorithm was given a 71%, but upon using SVM in place of the final layer, it was increased to 87%. From this increased inaccuracy, it can be noted that replacing the final layer with a different classifier can yield better results as certain classifiers are better at retrieving data while others are better at interpreting the data.

Many people went further on to classify breast histology images into four classes. Using the BioImaging 2015 breast histology classification challenge dataset [25], Araújo et al. [26] proposed a CNN with the ability to integrate information from several histological scales. Normalizing the images using the Macenko et al. [27] method, which considers the staining process used in the preparation of histological slides. Stain normalization is a criterion for enhancing detection performance. The patches of 512 x512 were extracted from images of size 2040 x1536. Using data augmentation techniques, the patch-wise was trained on a CNN +SVM and CNN classifiers to produce the patch class probability. The image-wise classification was then used through the use of majority voting, maximum probability, or sum probabilities as a patch probability fusion method. The accuracy of the patch-wise classification achieved was 66% with the use of CNNs and 65% with the use of CNN+SVM. The accuracy of image-wise into the four classes was 77.8%, with the majority voting being the best.

Vo et al. [28] used the same dataset but decided to use an Ensemble of CNN and gradient boosting tree classifiers. This means employing a combination of deep CNNs on multiscale images to extract visual features, and then a boosting framework is utilized to obtain an improved classification performance. Macenko et al. [27] were once again used in preprocessing, and they decided to apply a technique consisting of geometric augmentations for data augmentation, namely reflection, rotation, etc. The training set was then used to train Inception-ResNet-v2 networks with multiscale images used as input. Discriminate features were extracted from the network and used as inputs for the training of the gradient boosting tree classifier. The majority voting strategy was used to merge the GTBC into a prosperous classifier. It is noted that the recognition rate was 96% for this classifier. The breaKHis dataset was also used for this method, and the results obtained were a staggering 96% for 200x.

What is intriguing is that replacing one component with another, in this case, the fully connected layers with global average-pooling layers once again leads to some optimization in the program, i.e., reduction of network parameters.

H.Kassani et al. [29] decided to ensemble 3 pre-trained CNNs- VGG19 [30], MobileNetV2 [31] and DenseNet [32] in binary classification. Using BioImaging 2015, BreaKHis, -patch-Cameleon, and ICIAR2018 datasets, they went about preprocessing and augmenting the data. Augmentation this time was enhancement, zoom, shear, and fill mode. Image normalization obtained a similar range of values for each image. The images were resized to 224x 224 and then fed into different pre-trained models. Each producing a feature vector flattens to form a much better feature vector. The classification was done using a multilayer perceptron, and the model achiever satisfactory results and performed substantially well in comparison to state-of-the-art machine learning algorithms. The accuracy achieved by the model was surprisingly poor when tested on the BioImaging dataset. The authors stipulated that a possible reason for this was due to using pre-trained models. Input images were required to be resized, and due to this factor, the model could have lost discriminative information.

Transfer learning using the ResNet models by Mahbod et al. [33] on the BioImaging 2015 and ICIAR 2018 datasets. RGB histogram normalization was used to generate the best accuracy. ResNet models need a fixed size as input; thus, images were sized to 224x224. The model was pre-trained on natural images, and transfer learning fine-tuned the weights to accommodate breast histology slides. Accomplished by replacing the last fully connected layers of ResNet-50 and ResNet-101 models with new fully connected layers and randomly choosing weights assigned to nodes within layers. Data augmentation was used to produce eight different images with the application of rotation and flipping. The average output probabilities were chosen for classification. And the two network prediction vectors were taken as fusion results achieving 97% accuracy on BioImaging and 89% on ICIAR.

The use of boosting algorithms is said to theoretically design a strong learning model with the combination of multiple weak learners. The boosting algorithms collect and optimize decision trees, i.e., the weak learners. Assaad et al. [22] and Buabin et al. [23] demonstrated this with the use of AdaBoost and applied it to text or time series. Kariankis et al. [24] tried using AdaBoost decision trees to combine CNNs for the classification of images. Once again, it can be noted that histogram equalization, the use of the Reinhard and Macenko method for preprocessing, and the use of Data augmentation methods provided a greater dataset that did not lose information from the images.

2.3 EfficientNets. In 2019 Mingxing Tan and Quoc V. Le set out on a task of scaling up ConvNets using a simple yet effective compound calling method. It has been seen previously that ResNet can be scaled up from ResNet-18 to ResNet-200 by using more layers [34]. Alas, the scaling-up method has never truly been understood. So far, the methods of scaling up to include:

- By depth [34]
- Width [34]
- Image Resolution [34]

Generally, 1 of these dimensions is scaled up, while scaling two dimensions is possible, and it requires tedious manual tuning. Despite this, accuracy and efficiency are frequently sub-optimal.

Their study showed the criticality of balancing all dimensions of width/depth/resolution, and such balance can be achieved by scaling each of them using a constant ratio.

The conventional method was arbitrarily scaling the factors. Tan et al. [34] method include uniformly scaling depth, width, and resolution using a set of fixed scaling coefficients.

This model deemed EfficientNets outperformed other ConvNets. The EfficientNets B7 was computationally quicker and ran with fewer parameters as compared to the great pipe. The performance of the family of EfficientNets compared to other ConvNets architectures on the ImageNet database is shown in Figure 2.

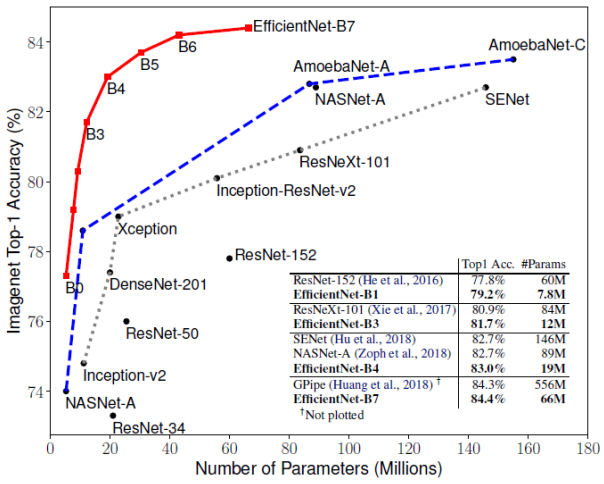

Figure 2. Represents the Family of EfficientNets against other ConvNet architectures on the ImageNet database. [34]

Munien et al. [35] used transfer learning along with EfficientNets to train a model on the classification of breast histological images using the ICIAR 2018 dataset. The authors had first preprocessed the images using the Reinhard and Macenko method, followed by an RBG color histogram equalization. She then went about augmenting the images, which provides various images for the model to be trained. Transfer learning using fine-tuning was then implemented, and then EfficientNets models were trained. The accuracy of this technique is absurdly magical, with EfficientNets-B2 achieving an accuracy of 98.33% using the Reinhard approach.

3. METHODS AND TECHNIQUES

Figure 3 depicts the process followed in this research. There were three crucial phases during this research. Phase 1 consisted of finding the perfect image for each class to be the target Image when carrying out the stain normalization aspect. Phase 2 dealt with extending the EfficientNets architecture to perform classification. Phase 3 dealt with picking which combination produced the best results. Sufficient regularization was necessary since the dataset used was relatively small compared to those generally seen in deep learning. The possibility of overfitting is quite high considering this is a small dataset, but by using regularization techniques, it is possible to counteract this. Munien et al. [35] proposed the model that was used for the classification. This proposed model had significantly high accuracies and was chosen to be the proposed model as well for this research. Figure 4. Depicts the model [35] had constructed.

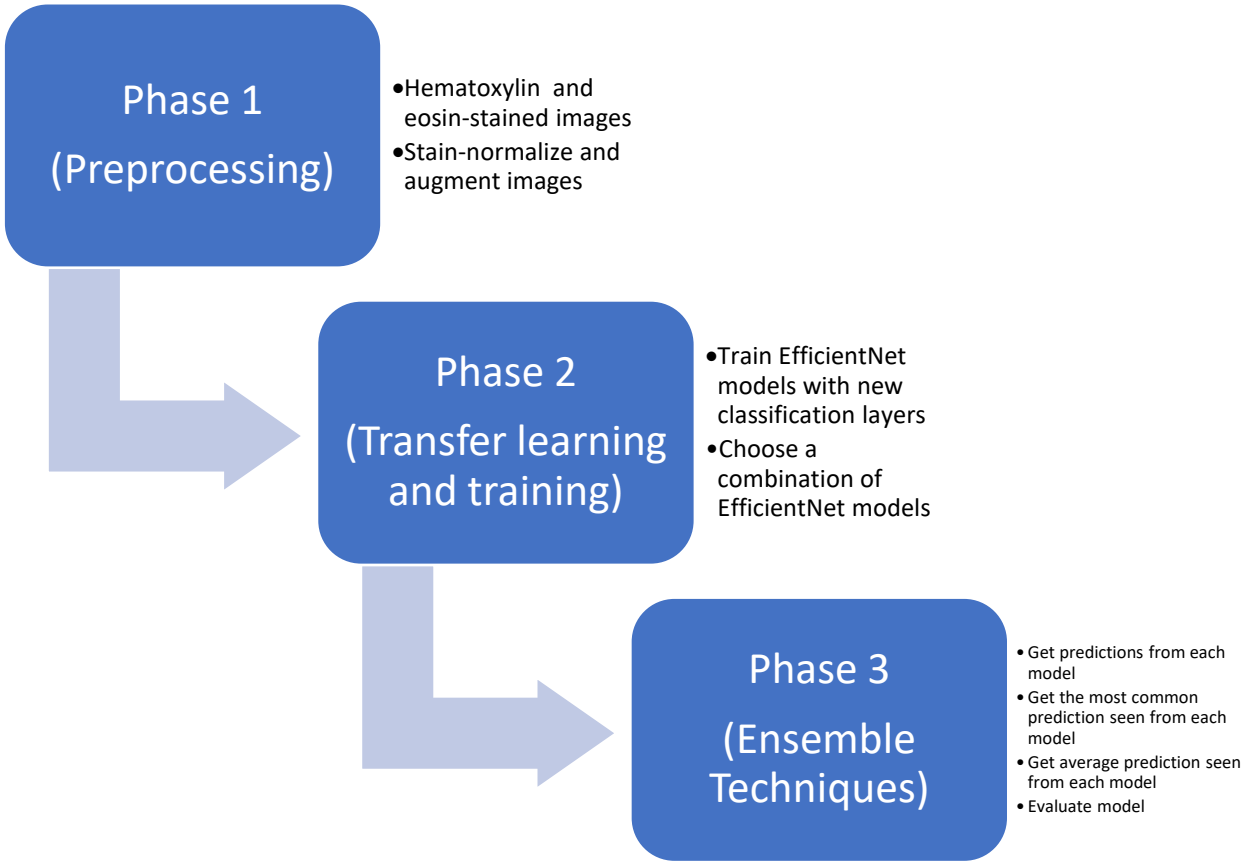

Figure 3. Training process followed and the different phases

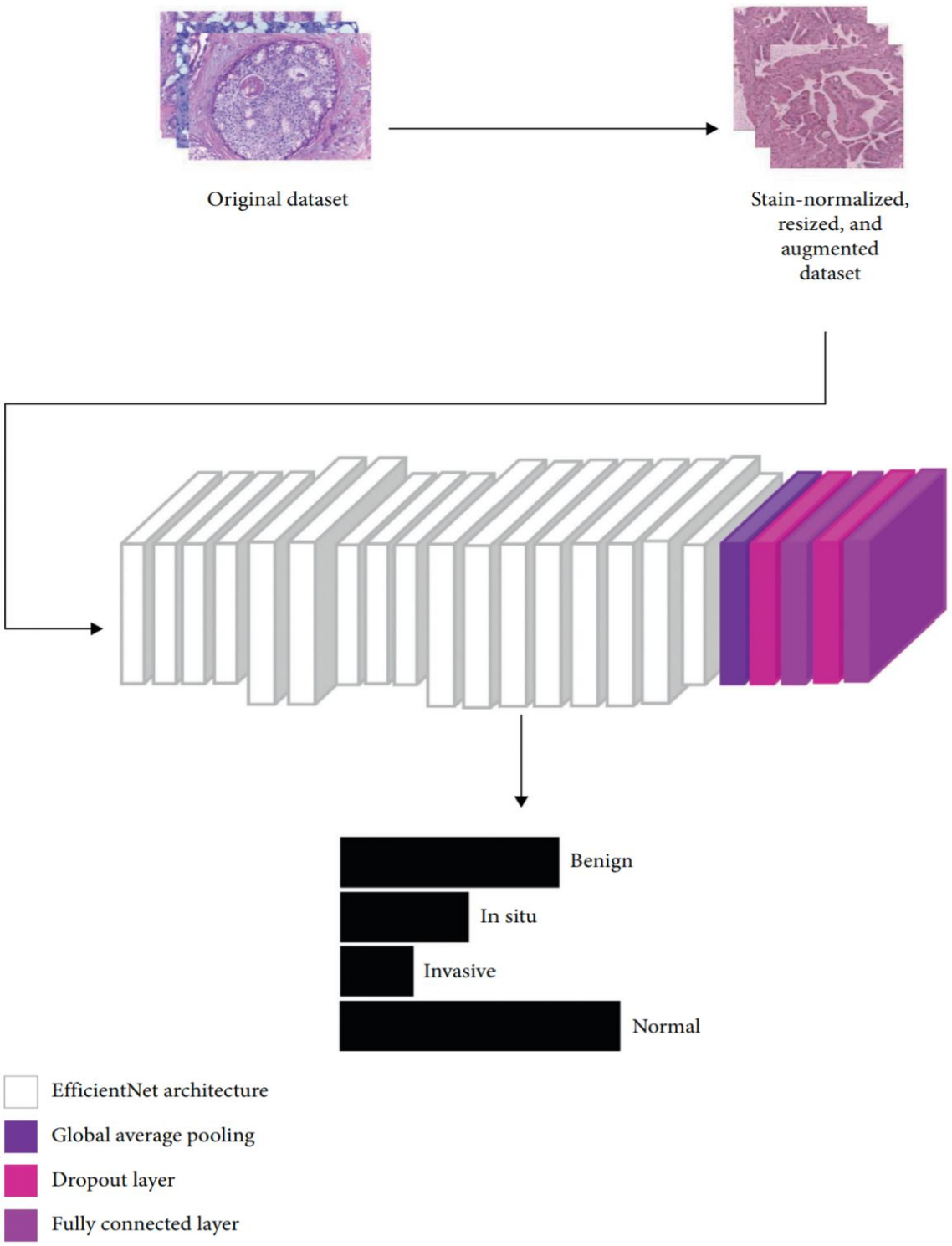

Figure 4. The model used by [35] classification of histology microscopy images using deep learning

3.1 Dataset. The dataset used was the ICIAR 2018 Breast Cancer histology images dataset [3]. It contains 400 microscopy images, which are evenly distributed across the four classes, Benign, Normal, InSitu Carcinoma, and Invasive Carcinoma. Two medical professionals annotated each image, and if the professionals disagreed on a particular image's annotation, the image was either discarded or confirmed through immunohistochemical analysis. The dataset is available in RGB.tiff format, and each image is 2048 × 1536 pixels in size, with a pixel scale of 0.42µm × 0.42µm (which refers to the area of tissue covered by a pixel) with a magnification of 200×. Augmented images are manually classified into three sets. The Training, Validation, and Testing set. Training is done on 85% of the dataset, validation is done with 11.25%, and testing is done on 3.75% of the dataset.

3.2 Preprocessing. The images from the dataset are extremely larger than those typically used when training neural networks. For the network to maintain important features, the rescaling of images needs to be done. The dataset once again is smaller than those typically used, and one way to get around this is by augmenting the images. Augmentation increases the amount of unique data and also contributes to overfitting.

3.2.1 Stain Normalization. Due to tissue preparation and the histology staining procedure, there may be color inconsistencies in certain slides. There are many factors that come into play when dealing with images firstly since there could be issues such as different light sources contributing to the image becoming a shade or so lighter or darker, the surface of which the slide was taken on as well as the angle which the camera was pointing in. These factors are amplified further when these slides have to undergo some preparation. While these factors contribute minimal to the digitized slide, small changes can make a big impact when it comes to the training process in CNNS. The technique of stain normalization had not been used by any of the top-performing methods reported in the ICIAR 2018 Grand Challenge paper [36], but a significant increase from nonnormalized images to stain normalized images in [35] can be seen, and thus the two methods of normalization where chosen. Stain normalization from Reinhard et al. [37] and Macenko et al. [27] These methods aided in improving the efficiency and accuracy for [35].

Images must be converted from the BGR colour space to the RGB colour space in order for the stain normalization techniques to function as expected.

3.2.1.1 Macenko Stain Normalization. This technique [19] accounts for the staining protocol used during the preparation of the tissue slide. Firstly, the colors are converted to optical density (OD) via the simple logarithmic transformation. A value, β, is specified and used as a threshold to remove data with higher OD intensity. Singular value decomposition (SVD) is applied to the optical density tuples from the first step in order to determine a plane. This plane corresponds to the two largest singular values found. The optical density transformed pixels are then projected onto this plane so that the angle at every point concerning the first SVD direction can be determined. Then, the color space transform resulting from the previous steps is applied to the original breast cancer histology image, and the histogram of the image is stretched such that the range covers the lower (100–α)% of the data. Minimum and maximum vectors are calculated and projected back into the optical density space. The hematoxylin stain corresponds to the former vector, and the eosin stain corresponds to the latter vector. The concentrations of the stains are appropriately determined, and the resulting matrix represents the RGB channels and OD intensity. The values α and β are recommended to be set to 1 and 0.15, respectively, and are kept the same for these experiments.

3.2.1.2 Reinhard Stain Normalization. This technique [22] focuses on mapping the color distribution of an over-or under-stained image to a well-stained image. The use of linear transformation from RGB to lαβ color space by matching mean and standard deviation values of the color channels achieves this. Essentially, the mean color within the selected target image is transferred onto the source image. This method preserves the intensity variation of the original image. This, in turn, preserves its structure while its contrast is adjusted to that of the target. In the lαβ color space, the stains are not precisely

separated. The lαβ color space must be converted back into RGB to attain the normalized image. Figure 7 shows examples of the stain normalization techniques applied in this study. Essentially, the techniques aim to normalize the colors in the original images to those of the target. Target images were found by normalizing the dataset to the default target image for the Macenko normalizing technique and choosing images from each class to be the target image of that class. The criterion in question was decided based on if the image had a mixture of light and dark undertones.

3.2.2 Data Augmentation. The Keras library provides a combination of various methods for augmentation. These methods help overcome the issue of overfitting as well as contribute to improving the classification accuracy. Rotation should not negatively impact the training of the architecture. For this purpose, the already normalized images where rotated and saved (90k) clockwise k = [0,1,2,3]. Additionally, the following augmentations were made randomly on the training set. Table 1 depicts the augmentations along with their respective values. By using the Keras library, all augmentations are done dynamically and randomly. Hence no extra storage space is needed. Images are also resized to fit the recommended size for each EfficientNets architecture. Table 2 contains the recommended image size for each EfficientNets Architecture. The normalized rotated images are split manually into three sets, all equally balanced, the training, validation, and testing set. No augmentation is done on the testing set and validation set. 15% of the dataset is used for testing and validation. Equally balanced, three of the four rotated images are selected. The fourth rotated image, which is chosen randomly, is then sent into the testing set. Figure 9 depicts the additional augmentation carried out.

3.3 Transfer Learning. Transfer learning (TL) can be described as follows: "given a source domain DS and learning task TS, a target domain DT and learning task TT, transfer learning aims to help improve the learning of the target predictive function fT (·) in DT using the knowledge in DS and TS, where DS ≠ DT, or TS ≠ TT" [38]. Earlier and middle layers detect edges and generic shapes,

Table 1. Depicts the values used for Data Augmentation for Stain Normalized images.

Augmentation Type	*Value*
Rescale	1/255
Rotation	5°
Width Shift	0.1
Heigh Shift	0.1
Horizontal Flip	True
Vertical Flip	True
Fill Mode	Reflect
Additional Rotation	0°, 90°, 270°, 360°

Table 2. Recommended Image size for each EfficientNets architecture

EfficientNet Model	*Recommended Size*
B0	224x224
B1	240x240
B2	260x260
B3	300x300
B4	380x380
B5	456x456
B6	528x528
B7	600x600

layers towards the end detect problem-specific features in a CNN. The main benefits of TL are saving training time, improving the performance of the neural network, and circumventing the limitations caused by lack of data [39]. The challenges associated with transfer learning, especially with the application of transfer learning to medical image classification. These challenges include overparameterization, expensive computations, and insufficient labeled data [45], which leads to increased training time and longer epochs.

Dawud et al. [46] combined Alexnet and support vector machine for brain hemorrhage classification tasks, and while it did produce better results than the baseline Alexnet model, it took a greater number of epochs and longer training time to achieve higher accuracy. Exploring the use of lightweight architectures is one way to get around this issue. Large amounts of data in medical image classification are not properly annotated, and there is scarcity in

the availability of this resource due to the fact that it is expensive and complex to annotate images.

The concept of transfer learning is based on utilizing the general features learned in the earlier layers from the source dataset, and a specified number of layers at the end of the model are retrained on the target dataset.
This technique has been effective in overcoming the issue of small datasets [40].

There are various forms of transfer learning which have been proposed. These include weight initialization, feature extraction, and fine-tuning. The work of [35] concluded that feature extraction and fine-tuning did not have any impact on the accuracy, but it was found that feature extraction was ineffective as the extracted features could not transfer features from the source dataset to classify breast cancer. The reasoning behind this was deduced that the source dataset consists of natural images, those that have no resemblance to histology images. Thus fine-tuning the architecture and utilizing the low-level features yield was chosen as the transfer learning technique.

The most common workflow in transfer learning is as follows. Take layers from a previously trained model. Freeze them to avoid destroying information they contain during future training rounds. Add new trainable layers on top of the frozen layers; these learn to turn an old feature into predictions on a new dataset and train the new layers on your dataset.

Fine-tuning consists of unfreezing certain layers in the model you obtain and retraining it on new data with a very low learning rate. The generic features in EfficientNets come from the ImageNet dataset, which contains approximately 14 million natural images and 22 thousand visual categories. The most suitable layer to freeze was found to be the third block [35]. In addition to fine-tuning, the use of nosy-student weights was also appropriate for this application [35]. Figure 10 depicts the process of fine-tuning.

3.4 Ensemble Learning. The work of Opitz et al. [41] investigated various techniques used in ensemble learning. Ensemble learning can be defined as training several weak learners to solve the same problem and combined to get better results. The combination of multiple independent and diverse models leads to correct decisions being reinforced since random errors cancel themselves out. No single model is better than the rest, but a combination of models can assure the final decision taken. It is generally used for improving classification and prediction for the most part. The ensemble technique used in this study focuses on voting, of which two different voting schemes are investigated, hard voting and soft voting.

Hard voting is otherwise known as majority voting. This focuses on the most recurring class found when each individual classifier votes. If classifier 1 predicts the image belongs to class 1, classifier 2 predicts the image belongs to class 1, and classifier 3 predicts the image belongs to class 3, then the most occurring class, in this case, class 1, is chosen as the final class for the image. The algorithm is depicted in Figure 5.

Soft voting, which gives the average of the probabilities. Once a combination is chosen, the most recurrent prediction is taken as final. Using the example given above, soft voting would sum up the probabilities, 1 + 1 + 3, and then divide it by 3 to give us the final class, which will be class 2. All final predictions are rounded to the nearest integer. The algorithm is depicted in Figure 6.

These methods are the basis of Ensemble learning and were chosen because, as seen from [35], the proposed models generated significantly high accuracies for the Reinhard stain normalization.

3.5 Experimental Settings. Seeds were set for all methods, which allowed it so that results could be reproducible. Specifically, the images chosen for each epoch were the same throughout each model training for both stain normalization methods. Stain Normalization was done using the package provided by [42] and [43]. First, all images underwent the Macenko stain normalization procedure on an Image that was not included in the dataset but the default values in [42]. Once all were normalized, the best image from each class (as seen in figure 6) was manually picked, and all images had to undergo the Macenko and Reinhard staining procedure with

these images set as target images. Figure 5 shows the process an image had to undergo for the stain normalization process. Figure 8 shows the target images used for each class in the second stain normalization process. The ImageDataGenerator from the Keras library was used to create augmentation generators for training data.

Publicly available pretrained EfficientNet models were used in this research [44]. Each EfficientNet model followed the same design that [35] followed with the exception of flattening the layers after the global average pooling layer and before the softmax layer. The reasoning behind this is clearly and thoroughly explained in [35] and was chosen because of this reasoning.

Once again, the Adam optimizer was used with a learning rate of 0.0001. high learning rate leads to the neural network becoming reckless and failing to retain information. Lower learning rates lead to needing more epochs since training takes longer; this becomes computationally expensive and increases training time. Due to Colab having a time cap of 12 hours, having a very low learning rate would lead to models not training in time.

Having one constant learning rate could also lead to results such as overfitting thou. One way to overcome this was using the Reduce Learning Rate Plateau Callback available from the Keras Callbacks library. To overcome overfitting, the learning rate is reduced by a factor of 0.2 after every eight epochs if the validation loss does not change. The minimum learning rate is set to 0.0001. This means the learning rate will never go below 0.00001.

Early Stopping is another Callback used to overcome overfitting. Once again, this callback monitors the validation loss attribute, and after every ten epochs, if the validation loss attribute does not change, training is halted.

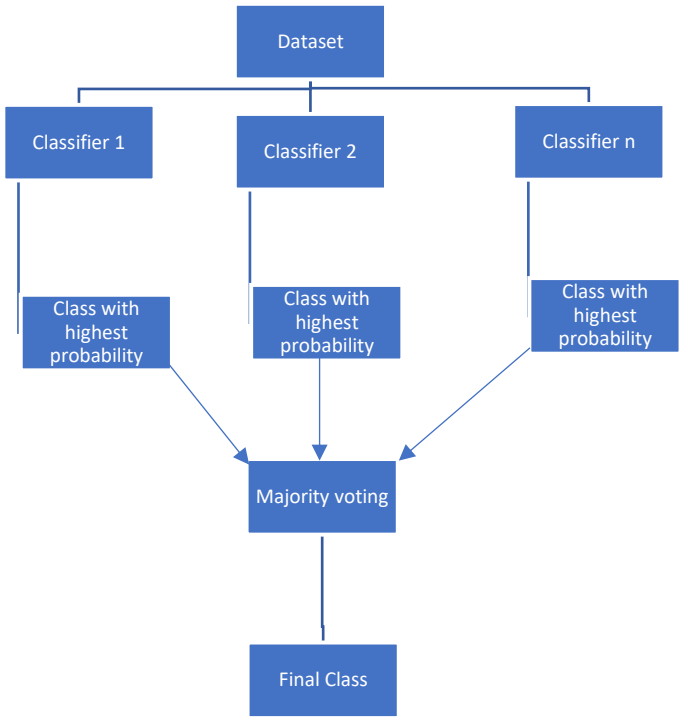

Figure 5. Majority voting algorithm

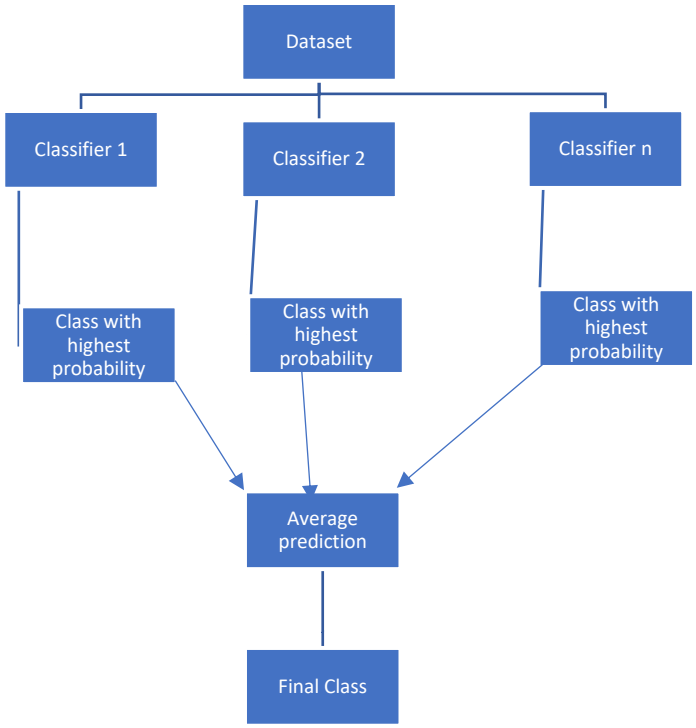

Figure 6. Average voting algorithm

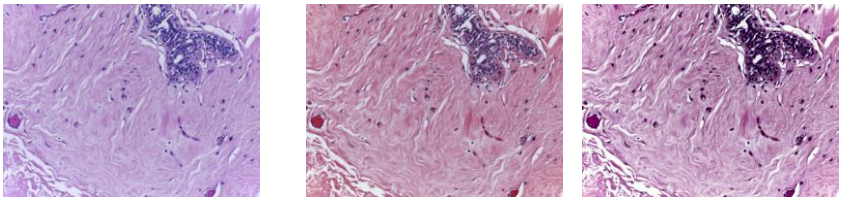

Figure 7. the process each image had to undergo. (a) represents the original image, (b) represents the image after undergoing Macenko normalization with a target image set, (c) represents the stain normalized image which had undergone a Reinhard normalization with a target image set

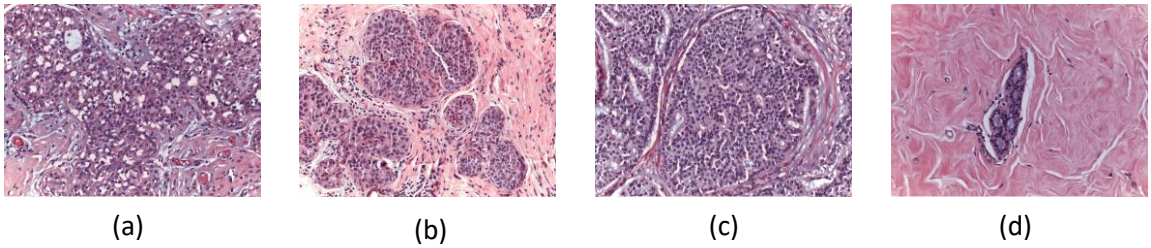

Figure 8. The target images used in the staining process, (a) target image for Benign, (b) target image for InSitu Carcinoma, (c) target image for invasive carcinoma, (d) target image for normal tissue.

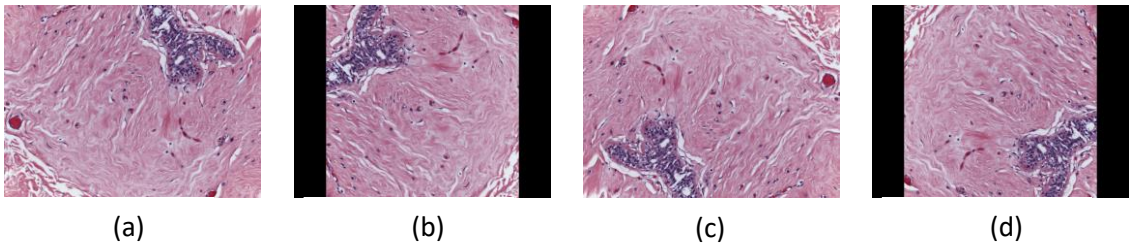

Figure 9. depicts the additional rotation applied to each image. A) depicts the original image, b) depicts the image after undergoing a 90° rotation to the left. C) depicts the image after undergoing a 180° rotation to the left d) depicts the image after undergoing a 270° Rotation to the left

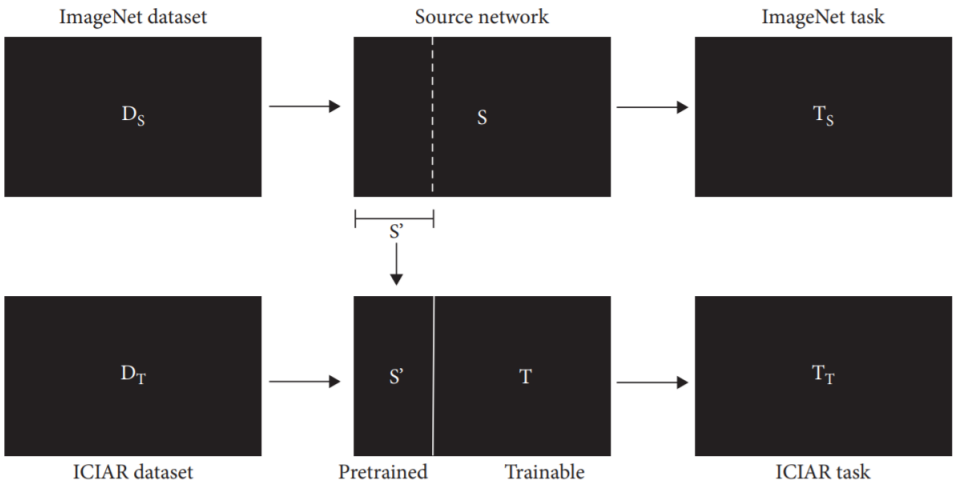

Figure 10. Visualization of fine tuning

The final callback used is the model checkpoint callback which monitors the validation accuracy and saves the best performing model, which has the highest accuracy. The batch size was set to 16 for models B0-B3 and 8 for models B4-B7 due to memory constraints. Categorical cross-entropy was chosen as the loss function since it is best suited for multiclass classification. The function is defined as:

$$CCE = -\frac{1}{N}\left(\sum_{i=1}^{N} \log P_{model}\left[y_i \in C_{y_i}\right]\right) \quad (1)$$

where N represents the number of instances and $\log P_{model}\left[y_i \in C_{y_i}\right]$ represents the probability predicted by the model for the i^{th} instance.

Once each model is trained, the best four from each Stain Normalization are chosen and appended to a list. Each model predicts the result from the test set, and each prediction is appended to a list. Once each model in the list has made some predictions, using the mean function provided by the NumPy library takes the mean of all values along the axis 0. This will provide the mean value from each column.

The collections library provides a package, Collections, which contains a method that calculates the most common value across all models along each column. All tests were run on the Google Colaboratory platform, which provides 25 GB RAM and a 12 GB NVIDIA Tesla K80 graphics processing unit.

4. RESULTS AND DISCUSSION

4.1 Evaluation Criteria. Each model's performance was evaluated by calculating the precision, recall, F1-score and accuracy. The following equations below represent the manner in which these metrics are determined. Where TP, TN, FP, FN, represents true positives, true negatives, false positives, and false negatives, respectively. The precision and recall scores are recorded and averaged over all four classes.

Accuracy

$$Accuracy = \frac{TP \times TN}{TP + TN + FP + FN} \times 100 \quad (2)$$

The accuracy of a model is the fraction of prediction that the model got right.

Precision

$$Precision = \frac{TP}{TP + FP} \times 100 \quad (3)$$

The precision of a model is the number of items correctly labelled belonging to the positive class. The total precision is calculated by taking the sum of precision of every class and diving them by the total number of classes

Recall

$$Recall = \frac{TP}{TP + FN} \times 100 \quad (4)$$

Recall is also known as sensitivity which evaluates a model's ability to predict true classes of each available category. The total recall is calculated by taking the sum of each recall per class and dividing by the total number of classes.

F1 – score

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

The F score tells how precise and robust the model is

4.2 Experimental Results. Table 4 provides the accuracy for each model trained with the stain normalization methods. Table 5 provides the accuracy, precision and recall for the ensemble methods used for each stain normalization method

Table 3. Accuracies for each model for each Normalization Method

EfficientNet Model	*Macenko Normalized*	*Reinhard Normalized*
B0	90%	100%
B1	78%	100%
B2	92%	100%
B3	88%	97%
B4	Did not finish training	97%
B5	Did not finish training	68%
B6	77%	100%

The chosen models from the Macenko and Reinhard-stained procedure were B0, B2, B3, B6 as these provided the highest and most consistent accuracies. The choice of models was kept the same for both normalization methods so the comparisons between approaches would be fair and justified. Because of this The B1 model was not chosen despite it having a stronger accuracy than B6. This is counteracted by choosing the B3 model which has a lower accuracy than the B1 model for the Reinhard normalization procedure.

Table 4. Accuracy, Precision and Recall score for each ensemble method for each normalization procedure used.

	Voting used	**Accuracy**	**Precision**	**Recall**	**F1-score**
Macenko Ensembled model	Soft	80%	83.5%	80%	81.71%
Reinhard Ensembled Model	Soft	100%	100%	100%	100%
Macenko Ensembled Model	Hard	92%	92.25%	91.5%	91.87%
Reinhard Ensembled Model	Hard	100%	100%	100%	100%

4.3 Discussion

Compared to the best model found by [35] the mixture of using the Macenko and Reinhard approach is superior achieving a 100% accuracy score, while the Macenko-Macenko approach actually lost some of its accuracy compared to the best model found in [35] and this research.

A reason behind the astonishingly high accuracy could be because the test set used was far smaller than the training set, although when saving and training the model more emphasis is put on the validation scores than the training scores. While this is a limitation of the research it is justifiable since the validation set is unseen to the model during training as well as the testing set contains rotated images from the validation set. By combining two different stain normalization methods a higher accuracy is achieved. This has not been done in any study so far and can already be something to investigate.

Each image was normalized to a picture from its own class. This could have played a factor into why the model is so good at capturing main features while training because if all images are the same in one class, then it becomes easier to remove the needle from the haystack.

5. CONCLUSION.

From the ensemble methods it can be seen that by taking the prediction that recurs the most equals to the best possible result while taking the average prediction does decrease the accuracy slightly. This could be caused due to one of the indexes being 0 which negatively impacts the outcome of average predictions. For the Macenko process even though certain models resulted in low accuracies the end result from taking into account every model's output resulted in a much better classification model. Reinhard surprises us all by returning an accuracy of 100% which shocks mainly because it was not only 1 model that returned a high accuracy but 4 out of the 7 models trained had returned an accuracy of 100%.

For future work what can be interesting to investigate will be a universal target image, for which all images from all classes are normalized to. This image would need to have the correct undertones which does not favor one class over the other.

REFERENCES

[1] Herndon, J., 2021. *Everything You Need to Know About Breast Cancer*. [online] Healthline. Available at: <https://www.healthline.com/health/breast-cancer> [Accessed 28 May 2021].

[2] grand-challenge.org. 2021. *ICIAR 2018 - Grand Challenge*. [online] Available at: <https://iciar2018-challenge.grand-challenge.org/> [Accessed 28 May 2021].

[3] Cancer.org. 2021. *Types of Breast Cancer | Different Breast Cancer Types*. [online] Available at: <https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/types-of-breast-cancer.html> [Accessed 5 June 2021].

[4] Breast Cancer Now. 2021. *DCIS (Ductal Carcinoma In Situ) Explained - Breast Cancer Care*. [online] Available at: <https://breastcancernow.org/information-support/facing-breast-cancer/diagnosed-breast-cancer/primary-breast-cancer/ductal-carcinoma-in-situ-dcis> [Accessed 5 June 2021].

[5] Rolls, G. and Sampias, C., 2021. *H&E Staining Overview: A Guide to Best Practices*. [online] Leica Biosystems. Available at: <https://www.leicabiosystems.com/knowledge-pathway/he-staining-overview-a-guide-to-best-practices/> [Accessed 6 July 2021].

[6] Thomson, M., 2021. *Why do we tolerate human over machine error?*. [online] Blog.rmresults.com. Available at: <https://blog.rmresults.com/why-do-we-tolerate-human-over-machine-error> [Accessed 28 May 2021].

[7] Who.int. 2021. *Breast cancer*. [online] Available at: <https://www.who.int/news-room/fact-sheets/detail/breast-cancer> [Accessed 28 May 2021].

[8] CANSA - The Cancer Association of South Africa. 2016. *Women & Cancer | CANSA - The Cancer Association of South Africa*. [online] Available at: <https://cansa.org.za/womens-health/> [Accessed 6 July 2021].

[9] *Hale & Monica Blog*, 2021. How common are breast cancer misdiagnosis. Available at: <https://www.halemonico.com/2021/01/13/how-common-are-breast-cancer-misdiagnoses/> [Accessed 28 May 2021].

[10] Veta M, Pluim JPW, Van Diest PJ, Viergever MA. Breast cancer histopathology image analysis: A review. IEEE Transactions on Biomedical Engineering. 2014; 61(5):1400±1411. https://doi.org/10.1109/TBME.2014.2303852 PMID: 24759275

[11] Kowal M, Filipczuk P, Obuchowicz A, Korbicz J, Monczak R. Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images. Computers in Biology and Medicine. 2013; 43 (10):1563±1572. https://doi.org/10.1016/j.compbiomed.2013.08.003 PMID: 24034748

[12] Filipczuk P, Fevens T, Krzyzak A, Monczak R. Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies. IEEE Transactions on Medical Imaging. 2013; 32 (12):2169±2178. https://doi.org/10.1109/TMI.2013.2275151 PMID: 23912498

[13] George YM, Zayed HH, Roushdy MI, Elbagoury BM. Remote computer-aided breast cancer detection and diagnosis system based on cytological images. IEEE Systems Journal. 2014; 8(3):949±964. https://doi.org/10.1109/JSYST.2013.2279415

[14] Belsare AD, Mushrif MM, Pangarkar MA, Meshram N. Classification of breast cancer histopathology images using texture feature analysis. In: TENCON 2015Ð2015 IEEE Region 10 Conference. Macau: IEEE; 2015. p. 1±5.

[15] Brook A, El-Yaniv R, Issler E, Kimmel R, Meir R, Peleg D. Breast Cancer Diagnosis From Biopsy Images Using Generic Features and SVMs. 2007; p. 1±16.
[16] Zhang B. Breast cancer diagnosis from biopsy images by serial fusion of Random Subspace ensembles. In: 2011 4th International Conference on Biomedical Engineering and Informatics (BMEI). vol. 1. Shanghai: IEEE; 2011. p. 180±186
[17] Ayer T, Chen Q, Burnside ES. Artificial Neural Networks in Mammography Interpretation and Diagnostic Decision Making. Comput Math Methods Med. 2013;2013:1–10. https://doi.org/10.1155/2013/832509.
[18] Ting FF, Sim KS. Self-regulated multilayer perceptron neural network for breast cancer classification2017 International Conference on Robotics, Automation and Sciences (ICORAS). https://doi.org/10.1109/icoras.2017.8308074
[19] Spanhol FA, Oliveira LS, Petitjean C, Heutte L. Breast Cancer Histopathological Image Classification using Convolutional Neural Networks. In: International Joint Conference on Neural Networks (IJCNN 2016). Vancouver; 2016
[20] Ciresan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2013;8150 LNCS(PART 2):411±418.
[21] Ragab DA, Sharkas M, Marshall S, Ren J. Breast cancer detection using deep convolutional neural networks and support vector machines. PeerJ. 2019;7. https://doi.org/10.7717/peerj.6201 e6201.
[22] M. Assaad , R. Boné, H. Cardot , A new boosting algorithm for improved time-series forecasting with recurrent neural networks, Inf. Fusion 9 (1) (2008) 41–55 .
[23] E. Buabin , Boosted hybrid recurrent neural classifier for text document classification on the Reuters news text corpus, in: International Journal Machine Learning Computing, 2012, pp. 588–592 .
[24] N. Karianakis, T.J. Fuchs, S. Soatto, Boosting convolutional features for robust object proposals, Tech. Rep., 2015 arXiv: 1503.06350 , University of Califor- nia Los Angeles
[25] Teresa Araújo, Guilherme Aresta, Eduardo Castro, José Rouco, Paulo Aguiar, Catarina Eloy, António Polónia, Aurélio Campilho. "bioimaging challenge 2015 breast histology dataset". https://rdm.inesctec.pt/dataset/nis-2017-003, 2017. [Accessed 6th June 2021].
[26] Teresa Araújo, Guilherme Aresta, Eduardo Castro, José Rouco, Paulo Aguiar, Catarina Eloy, António Polónia, Aurélio Campilho. Classification of breast cancer histology images using convolutional neural networks.PLOS ONE, 12(6):1{14, 06 2017.
[27] M. Macenko, M Niethammer, J.S. Marron, D. Borland, J.T. Woosley, X. Guan, C. Schmitt, and N.E. Thomas. A method for normalizing histology slides for quantitative analysis. In Proceedings of the International Symposium on Biomedical Imaging (ISBI), page 1107{1110, 2009.
[28] Duc Vo and Quang Nguyen. Classification of breast cancer histology images using incremental boosting convolution networks. 11 2018.
[29] Sara H. Kassani, Peyman Hosseinzadeh Kassani, Mike Wesolowski, Kevin Schneider, and Ralph Deters. Classification of histopathological biopsy images using ensemble of deep learning networks. 09 2019.
[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2015.
[31] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: E_cient convolutional neural networks for mobile vision applications. CoRR, abs/1704.04861, 2017.
[32] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. CoRR, abs/1608.06993, 2016.
[33] Ecker R. Smedby • O. Wang C. Mahbod A., Ellinger I. Breast Cancer Histological Image Classification Using Fine-Tuned Deep Network Fusion, pages 754{762. Springer International Publishing, 06 2018.
[34] Mingxing Tan and Quoc V. Le. Effcientnet: Rethinking model scaling for convolutional neural networks.
CoRR, abs/1905.11946, 2019.
[35] C. Munien, S. Viriri. “Classification of Hematoxylin and Eosin-Stained Breast Histology Microscopy Images using Transfer Learning with EfficientNets.”, *Biomedical Applications of Computer Vision using Artificial Intelligence* , 04 2021, Accessed on June 8, 2021 [Online]. Available:

https://www.hindawi.com/journals/cin/2021/5580914/#references
[36] G. Aresta, T. Ara´ujo, S. Kwok et al., “Bach: Grand challenge on breast cancer histology images,” Medical Image Analysis, vol. 56, pp. 122–139, 2019
[37] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, “Color transfer between images,” IEEE Computer Graphics and Applications, vol. 21, no. 4, pp. 34–41, 2001.
[38] S. Pan and Q. Yang, “A survey on transfer learning. Knowledge and Data Engineering,” IEEE Transactions on, vol. 22, pp. 1345–1359, 2010.
[39]N. Donges, What Is Transfer Learning? Exploring the Popular Deep Learning Approach, 2019, https://www.builtin.com/ data-science/transfer-learning Accessed 23-June-2020.
[40] L. Alzubaidi, O. Al-Shamma, M. Fadhel, and L. Farhan, “Optimizing the performance of breast cancer classification by employing the same domain transfer learning from hybrid deep convolutional neural network model,” Electronics, vol. 9, p. 445, 2020
[41] D. Opitz, R. Maclin, “Popular Ensemble Methods: An Empirical Study”, Journal of Artificial Intelligence Research, vol. 11 p 169 – 198, 1999
[42] G.Schau, schaugf/HEnorm_python , 2020, https://github.com/schaugf/HEnorm_python
[43] WangHao, wanghao14/Stain_Normalization, 2018, https://github.com/wanghao14/Stain_Normalization
[44] P. Yakubovskiy, Qubvel/efficientnet, 2020, https://github.com/ qubvel/efficientnet.
[45] Godasu, Rajesh; Zeng, David; and Sutrave, Kruttika, "Transfer Learning in Medical Image Classification: Challenges and Opportunities" (2020). MWAIS 2020 Proceedings. 18. https://aisel.aisnet.org/mwais2020/18
[46] Dawud, A. M., Yurtkan, K., and Oztoprak, H. 2019. “Application of Deep Learning in Neuroradiology: Brain Haemorrhage Classification Using Transfer Learning,” Computational Intelligence and Neuroscience (2019), pp. 1–12. (https://doi.org/10.1155/2019/4629859)