Direct Research Journal of Public Health and Environmental Technology
Vol. 8(4), Pp. 22-39, May 2023
ISSN 2734-2182
DOI: https://doi.org/10.26765/DRJPHET11672419
Article Number: DRJPHET11672419

Original paper

Attention-Based End-to-End Hybrid Ensemble Model for Breast Cancer Multi-Classification

Chiagoziem C. Ukwuoma[1], Dongsheng Cai[1*], Elvis Selasi Gati[2], Victor K. Agbesi[3], GutemaMisgana Deribachew[4], Leta Yobsan Bayisa[5], and Turi Abu[5]

[1]The College of Nuclear Technology and Automation Engineering, Chengdu University of Technology, Sichuan P.R., 610059, China.
[2]C.K. Tedam University of Technology & Applied Sciences (Directorate of ICT Services), Navrongo, Upper East, Ghana.
[3]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Sichuan, PR China.
[4]School of Mechanical Engineering, University of Electronic Science and Technology of China, Sichuan P.R., China.
[5]School of Information and Software Engineering, University of Electronic Science and Technology of China, Sichuan P.R., China.
*Corresponding Author email: caidongsheng@cdut.edu.cn

Received 4 March 2023; Accepted 27 April 2023; Published 1 May 2023

ABSTRACT: The daily rise in female instances of breast cancer (BC) is largely due to misinformation and late-stage detection. Effective treatment for BC can only be administered by correctly diagnosing cancer in its very early stages of development. BC classification has been discovered to be accelerated and automated using deep learning models and medical image analysis techniques. However, these techniques, which are crucial for numerous other applications outside broad visual identification, usually use generic traits. Since the primary objective of deep learning models is to characterize complex boundaries of hundreds of classes in deep space, embracing higher-order qualities is essential for improving non-linear modeling abilities. This study employs the publicly available BreakHis to provide an end-to-end hybrid ensemble model for BC multi-classification utilizing an attention-based global second-order pooling network. Ensembling is accomplished by adding an attention-based second-order pooling network in the form of a convolutional layer to the separate models to increase their non-linear modeling skills before concatenating their output features. Finally, the output features are relied on a classification layer for the final forecast. The proposed model produced enhanced results for binary and multiclass (four classes and eight classes) classification with 97.6% (40x), 95.5% (100x), 96.6 (200x) and 95.9% (400x) accuracy for the eight classes experiment. The experimental results show that, when compared to state-of-the-art models, the proposed approach obtains the best BC multi-classification accuracy.

Keywords: Breast Cancer; Attention Mechanism; Second-Order Pooling; Ensemble model; Multi-classification

INTRODUCTION

The whole human race is afflicted with terrible diseases, such as Breast Cancer (BC) in women. BC is reportedly the second most lethal kind of disease that affects women, according to the World Health Organization (WHO) (2014) and the latest report of 36 different malignancies, with recently diagnosed cases nearing a rate of 7.3% and a death rate of 6.9% (Sung et al., 2020. According to estimates, 90% of BC might be discovered and attended to by specialists in the initial phases and be cured (Sabtu et al., 2019). Hence, making a precise and prompt detection of this fatal illness is very essential to reduce the mortality rate. Medical imaging serves as a useful technique for identifying the presence of different medical disorders and evaluating research results. Biomedical imaging plays a critical role in the cancer treatment process. BC can be analyzed and found using

imaging modalities (Zhang et al.,2019) such as diagnostic mammography (X-rays), ultrasound (sonography), magnetic resonance imaging (MRI), and thermography. The suggested study makes use of ultrasound images. Effectively identifying and pinpointing cancerous cells in BC images is extremely challenging since cancerous cells have distinct sizes, forms, and positions. Diagnoses of a patient might have several result interpretations from a different pathologist. Research found that the diagnostic discrepancy between several examiners utilizing tissue samples was around 25% (Ukwuoma et al., 2022a; Ukwuoma et al., 2022b; Elmore et al., 2015). Histopathology images are typically utilized to diagnose BC (Veta et al., 2014). In recent times, several studies have proposed a variety of methods for the automated categorization of cells in BC diagnosis. Machine learning (ML) and deep learning (DL) are now being widely used to simplify the manual assessment of several diseases making it easier for pathologists (Budd et al., 2021; Umer et al., 2021). The core elements of modern deep learning models that are required for accurate and effective classification are image pre-processing and attribute extraction. The method used in deep learning is capable of recognizing characteristics by employing layers of neural networks. Medical diagnosis using Deep Convolution Neural Networks (DCNN) showed encouraging results (Ukwuoma et al., 2022), hence incorporating DCNN models into BC image analysis using histopathological images will fasten the rate of BC as well as an accurate method. In order to accurately classify breast cancer histopathology images utilizing publicly accessible datasets like BreakHis (Man et al., 2020) and BACH (Pimkin et al., 2018) datasets, recent studies have deployed DCNN models (Ukwuoma et al., 2022c; Ukwuoma et al., 2022d; Ukwuoma et al., 2022e) which are pre-trained on the ImageNet. Corresponding to this, the majority of studies of the BACH competition (Pimkin et al., 2018) used a single pre-trained model or a collection of pre-trained models to classify breast histopathology images into many categories. Despite ongoing research on breast cancer technologies, the feature extraction method used by the suggested models makes it difficult to correctly identify or categorize breast disorders. Also, several studies have been identified to have been conducted in the binary categorization of histopathological images. However, only a few methods, primarily based on four classes, have been presented for the multi-classification of histopathological images. These methods' classification accuracy was subpar since they solely took into account texture-based extracted characteristics. Additionally, the multiclass classification of breast cancer into eight classes is seldom discussed because of the striking similarities across the eight classes of breast cancers. It is a very difficult assignment to properly categorize breast cancer into eight classes because of the considerable commonality in the input images of various classes of the disease.

Sequel to the identified research gap, this paper addresses the identified problems by proposing an accurate and enhanced deep learning feature extraction framework for breast cancer image multi-classification via two, four and eight classes. This study uses the publicly accessible BreakHis to offer an end-to-end hybrid ensemble model for BC multi-classification using an attention-based global second-order pooling network. The ensembling is done by introducing an attention-based second-order pooling network in form of a convolutional layer to the single models to improve their non-linear modeling capabilities before concatenating their output features. Finally, the output features are based on a Softmax layer for prediction. In addition, we reviewed several pre-trained models (deep learning models) capabilities in breast cancer diagnosis, after which three models (DenseNet201, ResNet110 and Xception) were selected for our model building. We highlight the contribution of this manuscript thus;

- We developed an Attention-based end-to-end hybrid ensemble model for BC Multi-classification.
- This research introduced a spatial attention-based second-order pooling network for higher-order quantitative feature extraction, considering that pre-trained models are based on first-order quantitative characteristics of the input images. In order to represent the spectral–structural information of breast cancer images, the first-order feature mechanism was carried out whereas a second-order pooling mechanism with an attention-based architecture is employed to simulate relevant and discriminatory characteristics
- This research went ahead to present a comprehensive study of pre-trained models including binary classification and Multi-class (Four classes and Eight classes) of the proposed model.

The rest of the manuscript is structured in the following manner: Section 2 examines previous research in the field, while Section 3 presents a brief summary of the proposed approach. Section 4 looks at the experimental outcomes, including performance assessment, discussion of results, and a comparison with the latest models. Lastly, Section 5 concludes the paper and explores possible avenues for future research.

Related works

Early detection and classification are crucial in reducing the death rate from breast cancer, which is the most commonly diagnosed cancer worldwide. Various deep-learning techniques have been suggested for the automatic diagnosis and classification of breast cancer. One of these techniques is a six-branched CNN model proposed by Umer et al., (2022) which employs a deep feature fusion and selection method for multiclass breast

cancer classification. Histopathology images are used to diagnose breast cancer in the majority of cases (Veta et al., 2014b). With advancements in computer technology, artificial Intelligence techniques are being extensively used to automate the manual diagnosis of various diseases. Wang et al. (2021) introduced a deep feature fusion and enhanced routing-based automatic classification approach for breast cancer histopathological images. They developed a network with two parallel channels that can simultaneously obtain convolution features and capsule features. Yan et al. (2019) also proposed a hybrid technique of convolutional and recurrent deep neural networks for histopathological image classification.

Identifying the computational expense as an obstacle to accurately categorizing cancer via convolutional neural networks, Albashishet al. (2021) proposed a transfer learning model that relied on the VGG 16-layer deep model architecture to obtain advanced features from the BreakHis benchmark histopathological image dataset, as a solution to this problem. Singh et al., (2021) used transfer learning techniques with VGG-19 as the base to classify imbalanced histopathological images. To help in the automatic identification and diagnosis of breast cancer, likewise, Arooj et al., (2022) also proposed a transfer learning technique based on a pre-trained AlexNet model. Wang et al., (2022) proposed a hybrid deep learning model for the automatic detection of cancer, in which different layers of CNN and GRU architecture were used. Aljuaid et al., (2022) proposed a new computer-aided diagnosis method for binary and multi-class breast cancer using a combination of ResNet 18, ShuffleNet, and Inception-V3Net. Obtaining essential image features from breast histopathology images is typically done from the entire image, regardless of magnification. To overcome the challenge of selecting generic features, Ukwuoma et al. (2022f) introduced DEEP_Pachi, a model that is capable of classifying breast histopathology images at different magnifications. Nawaz et al., (2018) presented a deep-learning method for multi-class classification via CNN. The proposed method classifies tumors as benign or malignant, but the prediction is done in subclasses. Mi et al., (2021) proposed a two-stage architecture via a patch-level CNN and XGBoost as a WSI-level classifier. This method classifies H&E-stained digital pathology images of the breast into four categories. Maleika et al., (2021), combined a pre-trained model based on ResNet50 with an enhanced CNN to segment and classify various types of breast abnormalities. Handcrafted feature extraction methods were used by Joseph et al. (2022) and this was then used to train the deep neural Network. Their proposed techniques achieved an accuracy score of 96.84% (400x). Zabitet al., (2022) used the pre-trained Xception model's six intermediate layers to obtain salient characteristics from input images. They optimized the model on an unnormalized dataset before being tested on a normalized dataset.

METHODOLOGY

This section explores the proposed model in detail. Figure 1 shows the flowchart of the work done in this paper starting from the data acquisition to the evaluation of the train models. The first step is data splitting, followed by data augmentation of the train set. The work done is in sequential order s seen in (Figure 1). We examined 7 deep learning pre-trained model performances on the preprocessed data (DenseNet121, ResNet110, GoogleNet, Xception, EfficientNNetB3, ResNet110, InceptionResNetV2), followed by the selection of the networks for the proposed hybrid ensemble network. In this research, the DenseNet201, ResNet110 and the Xception model were used for the ensembling task.

Proposed hybrid ensemble model

First, the pre-trained architectures of choice were created to represent the spectral-spatial data of histopathology images of breast cancer. Second, discriminative and representative characteristics are modeled using an attention-based second-order pooling (ASP) operator. Since the selected pre-trained models used a first-order feature operator, we believed that their operators produce shallow future maps. Each model was expanded with an ASP operator before the concatenation layer to provide deeper features. The ASP is intended to derive from each model's first-order pooling (MaxPooling) by using the characteristics of the preceding convolutional layers as shown in (Figure 2). The modified models were DenseNet201, ResNet110 and Xception models as seen in (Figure 2). To reduce the scale dependency of the features from the prior layers of each selected Model, we included a l_2 normalization (Ukwuoma et al., 2022d). Second-order pooling is computed by using the real symmetric matrix $\mathrm{F_{SP}} \in \mathbb{R}^{k\times k}$;

$$\mathrm{F_{SP}} = \mathrm{F_{Fp}^{transpose}}\,\mathrm{F_{FP}}. \quad (1)$$

Eq. 1 calculates the internal product of each column-raw scalar pair in $\mathrm{F_{FP}}$. The relationship between the variables of the adjacent pixels is used to learn the attention mask thus;

$$\mathrm{AM} = \mathrm{F_{FP}}\mathrm{F_{Fp}^{transpose}}. \quad (2)$$

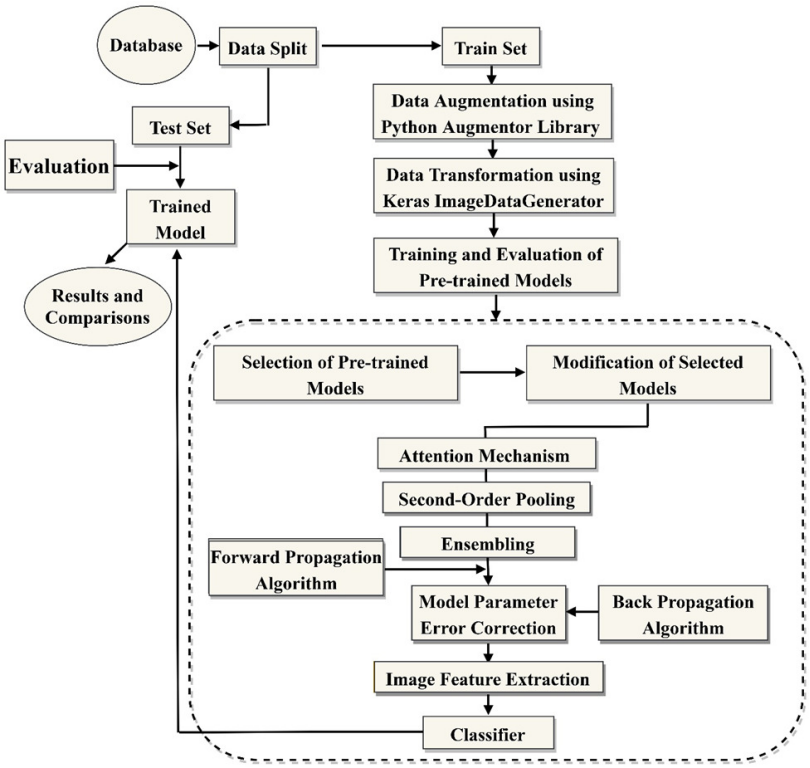

Figure 1: The Flowchart/Summary of the Proposed Architecture Starts from the Database/Data Acquisition to The Trained Model Evaluation and Comparison with The State-Of-The-Art Results.

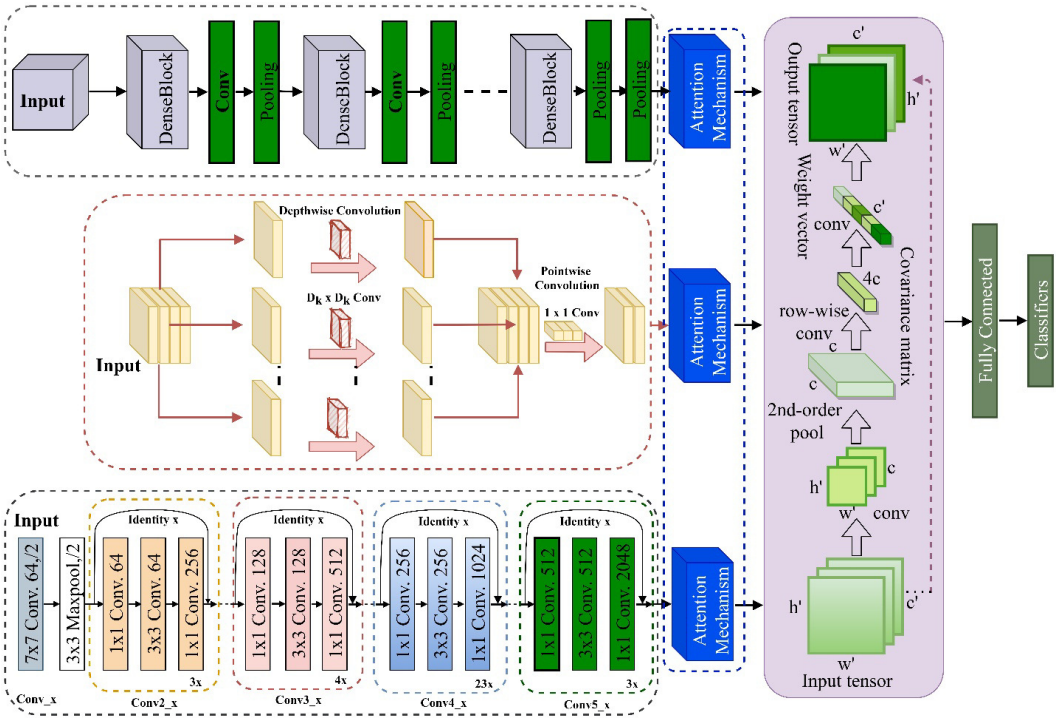

Figure 2: The Proposed Hybrid Ensemble Model. The Densenet201, Xception and ResNet110 Models were modified before concatenation of their output features after which a fully connected layer with SoftMax Classifier is employed for the Final Prediction.

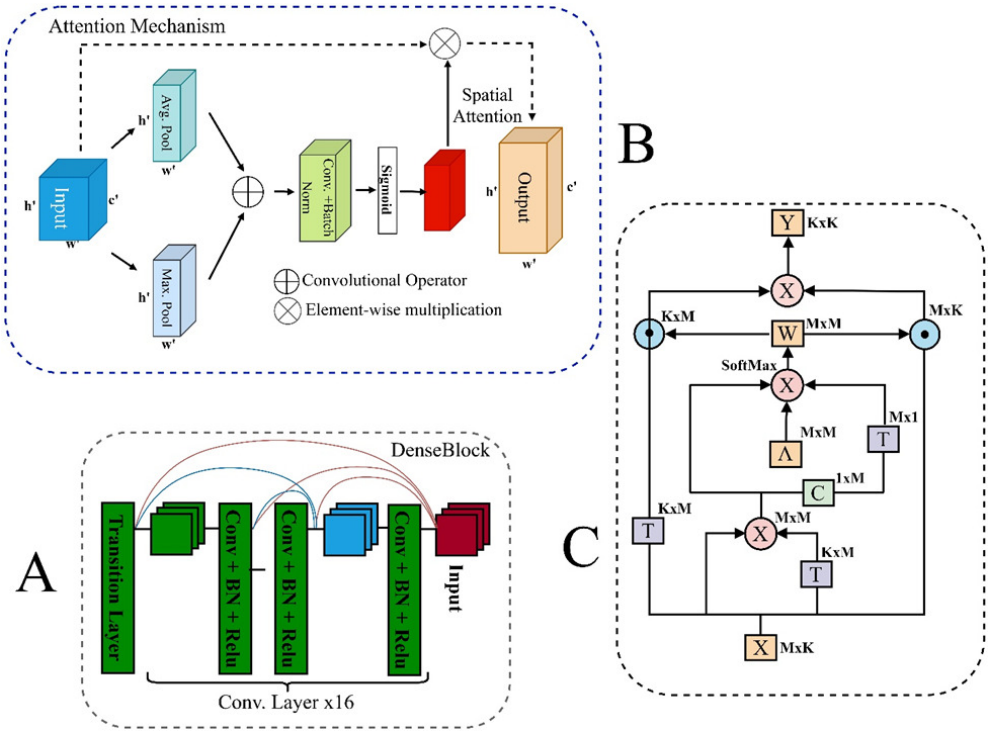

Figure 3: A Depicts the Dense Block of The Implemented Densenet201. B Depicts the Attention Mechanism Used for Second-Order Pooling Which Is Spatial Attention. C Depicts the Graphical Representation of the ASP.

Where $\mathbf{AM} \in \mathbb{R}^{M \times M}$ is the interchangeability of pixels that are next to one another in pairs. The trigonometric distance metric so determines how similar the surrounding pixels are to the core pixel;

$$\rho i = \frac{\mathbf{MA}_i \Lambda \mathbf{MA}_0^{transpose}}{\|\mathbf{MA}_i\| \left\|\mathbf{MA}_0^{transpose}\right\|}. \quad (3)$$

Where $i \in \{1, 2, \ldots, M\}$, $\mathbf{MA}_i$ signifies any adjacent pixel's vector, $\Lambda \in \mathbb{R}^{M \times M}$ denotes the diagonal matrix used to create ρ learning while $\mathbf{MA}_0$ designates the neighborhood's center pixel vector. According to Eq. 4, the attention scores are standardized to a uniform value using the SoftMax function.

$$\omega_i = \frac{e^{\rho_i + b_i}}{\sum_{j=1}^{M} e^{\rho_i + b_i}}, \quad (4)$$

b denotes the bias, $\sum \omega_i = 1$ for $i \in \{1, 2, \ldots, M\}$ and a cross-weight matrix $\mathbf{W} \in \mathbb{R}^{M \times M}$ can be gotten from eq. 4. The final implemented attention-based second-order pooling is seen in Eq. 5

$$\mathrm{F_{ASP}} = \mathrm{F_{Fp}^{transpose}} \mathrm{W}^2 \mathrm{F_{FP}}. \quad (5)$$

In order to depict the second-order pooling, the ASP takes the characteristics from the last convolutional layer of the pre-trained models and adds a variety of nearby pixels using an information and learning weighting method. This suggests that while the impact of the pixels with lesser weights is muted, the pixels with bigger weights have a significant impact.

The binary operations "⊗" and "⊙" stand for elementwise multiplication and matrix multiplication, respectively. The letters "**T**" and "**C**" which stand for transpose and picking out the indicated pixel respectively are examples of unary operations while features/parameters are represented by the yellow boxes. The resulting features are then vectorized using a Frobenius normalizing layer and a feature extraction layer. After that, classification is performed using the final FC layer with SoftMax loss. It should be noted that because each component of the proposed hybrid ensemble model is differentiable, the entire network is an end-to-end deep learning model as shown in (Figure 3).

Implementation details

In order to conduct this experiment, a desktop pc with a 64-bit OS, an Intel(R) Core (TM) i5-8300H CPU clocked at 2.30GHz, 15.9 GB of useable RAM and an NVIDIA GEFORCE GTX-1050Ti, 4.0GB GPU were used. This work uses the open-source frameworks Keras and TensorFlow for implementation. The model hyperparameters include Adam optimizer, images size of 224 × 224, batch size = 4, epoch = 300. The data transformation function includes; horizontal flip = True, rescale = 1./255, rotation range=40, width shift range=0.2, height shift range=0.2, and shear range=0.2. The experiment's loss function is the categorical cross-entropy.

Materials

This section explains the dataset, data preprocessing and the evaluation metrics used. The BreakHis (Man et al., 2020) dataset with Binary class (Benign vs Malignant), Multiclass-4 classes which have two sets; set 1 comprises Adenosis, Fibroadenoma, Phyllodes_tumor and Tubular_adenoma while set 2 comprises Ductal_carcinoma, Lobular_carcinoma, Mucinous_carcinoma, and Papillary_carcinoma and multiclass-8 classes which includes; Adenosis, Ductal_carcinoma, Fibroadenoma, Lobular_carcinoma, Mucinous_carcinoma, Phyllodes_tumor, Papillary_carcinoma and Tubular_adenoma is used for our experiment.

BreakHis (Man et al., 2020) Dataset

The Pathological Anatomy and Cytopathology (P&D) Lab in Brazil provided the database. Eighty-two individuals were diagnosed, producing malignant and benign microscopic images at various magnifications. Table 1 shows the breakdown of the dataset which is visualized in (Figure 4).

Dataset preprocessing

The employed dataset was first split in the ratio of 70:20:10 for the training, validation and test set. Only the train set was subjected to data augmentation using the python augmentor library to 1002 samples. The train sets were upsampled with the following transformations: cropping, random rotation, random flipping, skew, shear, zoom, etc.

Evaluation metrics

These are the metrics we employed;

$$Accuracy = \frac{TP + TN}{(TP + TN) + (FP + FN)} * 100 \quad (6)$$

$$Precision = \frac{TP}{TP + FP} * 100 \quad (7)$$

$$Specificity = \frac{TN}{N} * 100 = \frac{TN}{TN+FP} * 100 \quad (8)$$

$$Sensitivity = \frac{TP}{P} * 100 = \frac{TP}{TP+FN} * 100 \quad (9)$$

$$F_1\ score = \left(\frac{SEN^{-1} + PRC^{-1}}{2}\right)^{-1} = \frac{2 * TP}{2 * TP + FP + FN} \quad (10)$$

$$AUC = \frac{\sum_{i=1}^{k}\left(\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}\right)}{k} \quad (11)$$

RESULTS

This section presents the results of all the experiments performed in this paper. Since our work is based on ensemble models, this section presents various experiments starting from the pre-trained model's experiment, to the modification of the selected models to the developed ensembled models. The developed ensemble model results are based on three scenarios; binary classification, four-class multi-classification (Benign classes and malignant classes) and lastly Eight class multiclass classification.

Backbone selection and fine-tuned backbone experiment

The backbone model selection experiments were done on different magnifications for both benign and malignant classes. Table 2 shows that the DenseNet architecture recorded superior performance among the other models in both benign and malignant classes, followed by the Xception Model, and the ResNet architectures. However, the ResNet architecture had equal metrics in some magnifications compared to the VGG16 and InceptionResNetV2 architecture. From (Table 2), we concluded that the DenseNet, ResNet and Xception architecture's performance led to their selection as the proposed hybrid ensemble model backbone.

Table 3 summarizes the results of the modification experiments using the multiclass (four-classes) set. The results show that the selected models recorded a 0.01 - 0.05 performance improvement in all magnifications and evaluation metrics. The DenseNet architecture, its best performance is at 40x magnification with 0.925 accuracy, 0.844 sensitivity, 0.946 specificity, 0.855 precision, 0.846F1_score and 0.895 AUC for the benign while the modified DenseNet architecture best performance is seen at the 100x magnification with an accuracy of 0.960, Sensitivity of 0.920, specificity of 0.973, precision of 0.920, F1_score of 0.922and AUC of 0.951 for the benign class. For the Malignant class, the conventional DenseNet model's best performance was seen at 400x magnification with 0.913 (accuracy), 0.825(sensitivity), 0.941 (specificity), 0.825 (precision), 0.825 (F1_score)

Table 1: Breakdown of the BreaKHis dataset which includes the original partition, the splitting and the number of samples per split.

Class	Sub_Class	Magnification				Total	Nos_Patients
		40x	100x	200x	400x		
Benign	Adenosis	114	113	111	106	444	24
	Fibroadenoma	253	260	264	237	1014	
	Phyllodes_tumor	109	121	108	115	453	
	Tubular_adenoma	149	150	140	130	569	
Malignant	Ductal_carcinoma	864	903	896	788	3451	58
	Lobular_carcinoma	156	170	163	137	626	
	Mucinous_carcinoma	205	222	196	169	792	
	Papillary_carcinoma	145	142	135	138	560	
Total		1995	2081	2013	1820	7090	82
Augmented-Train							
Benign	Adenosis	1002	1002	1002	1002	4008	-
	Fibroadenoma	1002	1002	1002	1002	4008	
	Phyllodes_tumor	1002	1002	1002	1002	4008	
	Tubular_adenoma	1002	1002	1002	1002	4008	
Malignant	Ductal_carcinoma	1002	1002	1002	1002	4008	-
	Lobular_carcinoma	1002	1002	1002	1002	4008	
	Mucinous_carcinoma	1002	1002	1002	1002	4008	
	Papillary_carcinoma	1002	1002	1002	1002	4008	
Test							
Benign	Adenosis	11	11	11	10	43	-
	Fibroadenoma	25	25	26	23	99	
	Phyllodes_tumor	10	12	10	11	43	
	Tubular_adenoma	14	14	13	12	53	
Malignant	Ductal_carcinoma	86	90	89	78	343	-
	Lobular_carcinoma	15	16	16	13	60	
	Mucinous_carcinoma	20	22	19	16	77	
	Papillary_carcinoma	14	14	13	13	54	
Validation							
Benign	Adenosis	98	98	98	98	392	-
	Fibroadenoma	98	98	98	98	392	
	Phyllodes_tumor	98	98	98	98	392	
	Tubular_adenoma	98	98	98	98	392	
Malignant	Ductal_carcinoma	98	98	98	98	392	-
	Lobular_carcinoma	98	98	98	98	392	
	Mucinous_carcinoma	98	98	98	98	392	
	Papillary_carcinoma	98	98	98	98	392	

and 0.854(AUC) while for the modified DenseNet, the best malignant performance is seen at the 40x magnifications with an accuracy of 0.952, sensitivity of 0.904, specificity of 0.968, precision of 0.904, F1_score of 0.904, and AUC of 0.909. For the ResNet architecture, both the conventional and modified recorded their best performance at the 100x magnification with 0.911(Accuracy), 0.863(sensitivity), 0.943(specificity), 0.823(precision), 0.828(F1_score), 0.903(AUC) against 0.95(Accuracy), 0.910(sensitivity), 0.968(specificity), 0.910(precision), 0.909(F1_score), 0.938(AUC) for the benign class while that of the malignant is seen at 40x magnification with 0.892(Accuracy), 0.785(sensitivity), 0.928(specificity), 0.785(precision), 0.785(F1_score), 0.806(AUC) against 0.922(Accuracy), 0.844(sensitivity), 0.948(specificity), 0.844(precision), 0.844(F1_score), 0.882(AUC). For the Xception model (Table 3), the conventional best performance for the benign class is seen at 400x magnification whereas the modified Xception is at 40x magnification with an accuracy of

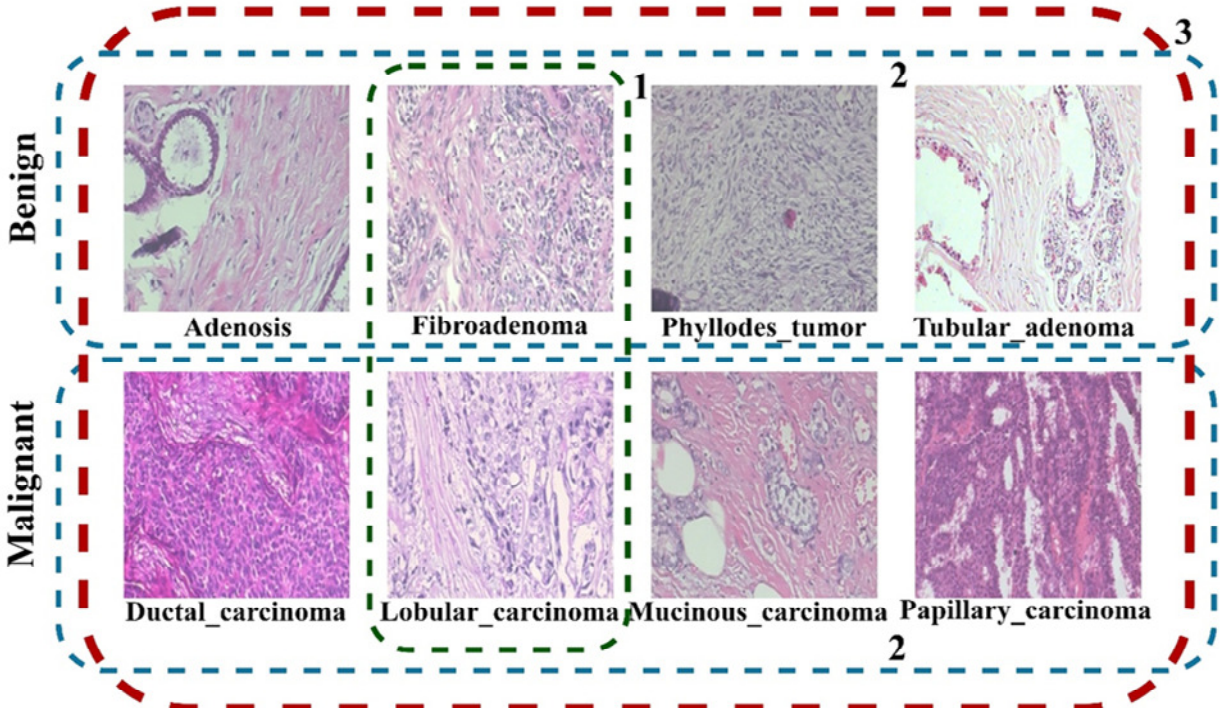

Figure 4: Sample Visualization of the BreaKHis (Man et al., 2020) Data. The Red Dash-Lines Indicate the Multiclass (8 Classes) Classification Labeled With 3, While the Blue-Dash Lines Indicate the Multi-Class (4 Class) Classification Labeled With 2 and the Binary Classification Is the Images with The Green Dash-Lines Labeled 1.

0.925, sensitivity of 0.850, specificity of 0.950, precision of 0.875, F1_score of 0.850, and AUC of 0.887 against 0.893(Accuracy), 0.800(sensitivity), 0.929(specificity), 0.786(precision), 0.788(F1_score), 0.863(AUC) for the conventional Xception. For the Malignant class, both the conventional and modified recorded their best performance at the 100x magnification with 0.900(Accuracy), 0.800(sensitivity), 0.933(specificity), 0.800(precision), 0.800(F1_score), 0.858(AUC) against 0.922(Accuracy), 0.844(sensitivity), 0.948(specificity), 0.844(precision), 0.844(F1_score), 0.868(AUC). The individual class performance for both the benign and malignant is also evaluated as shown in (Table 3).

Attention-based end-to-end hybrid ensemble model experiment

This section presents the result of the proposed model as recorded in (Table 4). For the binary classification, the 40x magnification yielded the highest classification result with an accuracy of 0.979, sensitivity (0.981), specificity (0.981), precision (0.979), F1_score (0.976) and AUC (0.981) whereas the 200x magnification yielded the lowest classification result with an accuracy of 0.939, sensitivity (0.942), specificity (0.942), precision (0.939), F1_score (0.939) and AUC (0.941). The highest area (ROC) and AP is seen in the 40x magnification, followed by the 100x, 400x and lastly the 200x magnification. The model performs very well in the malignant class compared to the benign class.

For four-class multi-classification on the Benign class, the highest model performance is seen at the 100x magnification with an accuracy of 0.968, sensitivity (0.952), specificity (0.978), precision (0.936), F1_score (0.942) and AUC (0.965) followed by 200x, 40x and lastly the 400x magnification. The malignant model's best performance is seen at 200x magnification with an accuracy of 0.956, sensitivity (0.912), specificity (0.970), precision (0.912), F1_score (0.912) and AUC (0.907) followed by 40x, 400x and lastly the 100x magnification. Likewise, the ROC and Precision-recall are used to evaluate the model's class performance. The Benign class had an area (ROC) above 90% in all magnifications with an area of 97% as the highest which is seen in 100x magnification. Although the AP results are poor compared to the ROC area, the Phyllodes_tumor, Adenosis and Tubular_adenoma all had an AP of 1.0 for 40x, 100x and 200x magnification. Looking at the malignant Roc (area) and AP performance of each class, the benign experiment result is superior. The Ap result is indeed poor as some classes had an Ap below 0.5.

Table 2: Classification performance of the various models under different magnifications for the ensemble backbone model selection. AC= Accuracy, SE=Sensitivity, SP= Specificity, PR=Precision, F1_S=F1_score.

	40× Magnification						100× Magnification						200× Magnification						400× Magnification					
	AC	**SE**	**SP**	**PR**	**F1_S**	**AUC**	**AC**	**SE**	**SP**	**PR**	**F1_S**	**AUC**	**AC**	**SE**	**SP**	**PR**	**F1_S**	**AUC**	**AC**	**SE**	**SP**	**PR**	**F1_S**	**AUC**
DenseNet	Benign						Benign						Benign						Benign					
	0.925	0.844	0.946	0.855	0.846	0.895	0.919	0.839	0.946	0.857	0.839	0.885	0.883	0.795	0.922	0.794	0.777	0.858	0.902	0.804	0.935	0.815	0.803	0.863
	Malignant						Malignant						Malignant						Malignant					
	0.904	0.807	0.936	0.807	0.807	0.819	0.863	0.725	0.908	0.725	0.725	0.784	0.887	0.774	0.925	0.774	0.699	0.804	0.913	0.825	0.941	0.825	0.825	0.854
InceptionResNet	Benign						Benign						Benign						Benign					
	0.883	0.767	0.922	0.770	0.767	0.835	0.863	0.729	0.909	0.737	0.728	0.814	0.875	0.764	0.916	0.757	0.755	0.838	0.839	0.679	0.892	0.679	0.679	0.763
	Malignant						Malignant						Malignant						Malignant					
	0.870	0.741	0.914	0.741	0.741	0.801	0.768	0.570	0.845	0.535	0.535	0.700	0.847	0.701	0.898	0.693	0.693	0.787	0.858	0.717	0.906	0.716	0.716	0.789
InceptionV3	Benign						Benign						Benign						Benign					
	0.858	0.735	0.906	0.716	0.721	0.818	0.863	0.748	0.909	0.730	0.737	0.826	0.842	0.722	0.894	0.683	0.687	0.808	0.866	0.732	0.911	0.732	0.732	0.817
	Malignant						Malignant						Malignant						Malignant					
	0.815	0.640	0.877	0.630	0.630	0.755	0.760	0.624	0.848	0.521	0.521	0.736	0.799	0.598	0.866	0.599	0.598	0.700	0.808	0.61	0.872	0.616	0.617	0.665
ResNet	Benign						Benign						Benign						Benign					
	0.883	0.766	0.922	0.767	0.767	0.834	0.911	0.863	0.943	0.823	0.828	0.903	0.883	0.798	0.922	0.783	0.777	0.860	0.875	0.75	0.917	0.751	0.75	0.817
	Malignant						Malignant						Malignant						Malignant					
	0.892	0.785	0.928	0.785	0.785	0.806	0.838	0.676	0.892	0.676	0.676	0.785	0.858	0.723	0.905	0.715	0.715	0.806	0.85	0.900	0.900	0.700	0.700	0.715
VGG16	Benign						Benign						Benign						Benign					
	0.883	0.766	0.922	0.810	0.767	0.837	0.839	0.713	0.888	0.742	0.687	0.800	0.808	0.659	0.869	0.617	0.627	0.764	0.902	0.804	0.935	0.821	0.800	0.864
	Malignant						Malignant						Malignant						Malignant					
	0.826	0.652	0.884	0.652	0.652	0.742	0.824	0.665	0.882	0.648	0.648	0.767	0.821	0.642	0.880	0.642	0.642	0.642	0.821	0.642	0.881	0.641	0.642	0.684
Xception	Benign						Benign						Benign						Benign					
	0.892	0.783	0.928	0.813	0.796	0.850	0.887	0.786	0.925	0.774	0.774	0.855	0.891	0.806	0.928	0.787	0.783	0.866	0.893	0.800	0.929	0.786	0.788	0.863
	Malignant						Malignant						Malignant						Malignant					
	0.900	0.800	0.933	0.800	0.800	0.858	0.856	0.742	0.904	0.711	0.711	0.822	0.858	0.715	0.905	0.715	0.715	0.789	0.871	0.742	0.914	0.742	0.742	0.820

Table 3. Performance of the Finetuned Models (Densenet, Resnet and Xception). AC= Accuracy, SE=Sensitivity, SP= Specificity, PR=Precision, F1_S=F1_score, Ad=Adenosis, Fi=Fibroadenoma, Phy=Phyllodes_Tumor,Tu=Tubular_Adenoma, Dc=Ductal_Carcinoma,Lc=Lobular_Carcinoma, Mc=Mucinous_CarcinomaandPc=Papillary_Carcinoma

	40× Magnification						100× Magnification						200× Magnification						400× Magnification					
	AC	**SE**	**SP**	**PR**	**F1_S**	**AUC**	**AC**	**SE**	**SP**	**PR**	**F1_S**	**AUC**	**AC**	**SE**	**SP**	**PR**	**F1_S**	**AUC**	**AC**	**SE**	**SP**	**PR**	**F1_S**	**AUC**
DenseNet	Benign						Benign						Benign						Benign					
	0.925	0.850	0.941	0.886	0.850	0.882	0.960	0.920	0.973	0.920	0.922	0.951	0.933	0.867	0.960	0.873	0.870	0.910	0.938	0.875	0.955	0.887	0.875	0.910
	Malignant						Malignant						Malignant						Malignant					
	0.952	0.904	0.968	0.904	0.904	0.909	0.898	0.796	0.932	0.796	0.796	0.810	0.945	0.891	0.964	0.901	0.890	0.880	0.938	0.875	0.958	0.875	0.875	0.864
ResNet	Benign						Benign						Benign						Benign					
	0.925	0.850	0.950	0.910	0.859	0.886	0.95	0.910	0.968	0.910	0.909	0.938	0.933	0.894	0.956	0.867	0.872	0.925	0.920	0.839	0.946	0.847	0.840	0.884
	Malignant						Malignant						Malignant						Malignant					
	0.922	0.844	0.948	0.844	0.844	0.882	0.863	0.725	0.908	0.725	0.725	0.766	0.916	0.832	0.944	0.832	0.832	0.866	0.917	0.833	0.944	0.833	0.833	0.830

Table 3

Xception	Benign						Benign						Benign						Benign					
	0.925	0.850	0.950	0.875	0.850	0.887	0.911	0.835	0.941	0.827	0.825	0.886	0.867	0.768	0.911	0.733	0.742	0.837	0.911	0.821	0.940	0.837	0.821	0.865
	Malignant						Malignant						Malignant						Malignant					
	0.922	0.844	0.948	0.844	0.844	0.868	0.873	0.746	0.915	0.746	0.746	0.753	0.920	0.839	0.946	0.839	0.839	0.867	0.892	0.794	0.928	0.783	0.783	0.850
DenseNet	Benign – Class Performance																							
	Metrics		Ad	Fi	Phy	Tu	Ad	Fi	Phy	Tu	Ad	Fi	Phy	Tu	Ad	Fi	Phy	Tu						
	Accuracy		0.967	0.883	0.950	0.900	0.984	0.935	0.968	0.952	0.967	0.883	0.883	1.00	0.946	0.929	0.911	0.964						
	Precision		0.909	0.800	1.00	0.833	1.00	0.957	0.857	0.867	1.0	0.88	0.615	1.00	1.0	0.88	0.75	0.917						
	Sensitivity		0.909	0.960	0.700	0.714	0.909	0.880	1.00	0.929	0.818	0.846	0.800	1.00	0.700	0.957	0.818	0.917						
	Malignant - Class Performance																							
	Metrics		Dc	Lc	Mc	Pc	Dc	Lc	Mc	Pc	Dc	Lc	Mc	Pc	Dc	Lc	Mc	Pc						
	Accuracy		0.919	0.985	0.956	0.948	0.838	0.887	0.908	0.958	0.891	0.942	0.956	0.993	0.892	0.942	0.950	0.966						
	Precision		0.931	0.882	0.850	0.818	0.860	0.50	0.696	0.833	0.885	0.786	0.933	1.00	0.892	0.750	0.917	0.846						
	Sensitivity		0.942	1.00	0.850	0.643	0.889	0.438	0.727	0.714	0.955	0.688	0.737	0.923	0.949	0.692	0.688	0.846						

Table 3

ResNet	Benign – Class Performance															
	Metrics	Ad	Fi	Phy	Tu	Ad	Fi	Phy	Tu	Ad	Fi	Phy	Tu	Ad	Fi	Phy	Tu
	Accuracy	0.983	0.850	0.983	0.883	1.00	0.919	0.951	0.935	0.983	0.883	0.933	0.933	0.946	0.893	0.893	0.946
	Precision	1.00	0.750	1.00	0.889	1.00	0.917	0.909	0.812	1.00	0.952	0.75	0.765	0.769	0.840	0.778	1.00
	Sensitivity	0.909	0.960	0.900	0.571	1.0	0.880	0.833	0.929	0.909	0.769	0.900	1.00	1.00	0.913	0.636	0.750
	Malignant - Class Performance															
	Metrics	Dc	Lc	Mc	Pc	Dc	Lc	Mc	Pc	Dc	Lc	Mc	Pc	Dc	Lc	Mc	Pc
	Accuracy	0.881	0.933	0.941	0.933	0.754	0.859	0.915	0.923	0.847	0.920	0.956	0.942	0.850	0.933	0.950	0.933
	Precision	0.938	0.650	0.833	0.647	0.816	0.389	0.750	0.588	0.905	0.667	0.810	0.647	0.866	0.727	0.917	0.667
	Sensitivity	0.872	0.867	0.750	0.786	0.789	0.438	0.682	0.714	0.854	0.625	0.895	0.846	0.910	0.615	0.688	0.769
Xception	Benign – Class Performance															
	Metrics	Ad	Fi	Phy	Tu	Ad	Fi	Phy	Tu	Ad	Fi	Phy	Tu	Ad	Fi	Phy	Tu
	Accuracy	0.933	0.866	0.950	0.950	0.951	0.855	0.919	0.919	0.933	0.750	0.883	0.900	0.929	0.857	0.893	0.964
	Precision	0.888	0.793	0.818	1.00	0.833	0.833	0.733	0.909	0.769	0.762	0.636	0.733	0.875	0.778	0.778	0.917
	Sensitivity	0.727	0.920	0.900	0.786	0.909	0.800	0.917	0.714	0.909	0.615	0.700	0.846	0.700	0.913	0.636	0.917
	Malignant - Class Performance															
	Metrics	Dc	Lc	Mc	Pc	Dc	Lc	Mc	Pc	Dc	Lc	Mc	Pc	Dc	Lc	Mc	Pc
	Accuracy	0.859	0.926	0.948	0.956	0.775	0.866	0.923	0.930	0.847	0.927	0.920	0.985	0.783	0.933	0.883	0.967
	Precision	0.904	0.647	0.783	0.833	0.802	0.412	0.867	0.643	0.886	0.650	0.750	0.923	0.871	0.631	0.556	0.846
	Sensitivity	0.872	0.733	0.900	0.714	0.856	0.438	0.591	0.643	0.876	0.813	0.632	0.923	0.782	0.923	0.625	0.846

Table 4: Performance of The Proposed End-End Hybrid Ensemble Model. Ac = Accuracy, Se=Sensitivity, SP= Specificity, PR=Precision, F1_S=F1_score, Ad=Adenosis,, Fi=Fibroadenoma, Phy=Phyllodes_Tumor,Tu=Tubular_Adenoma, Dc=Ductal_Carcinoma,Lc=Lobular_Carcinoma, Mc=Mucinous_CarcinomaandPc=Papillary_Carcinoma

	40× Magnification						100× Magnification						200× Magnification						400× Magnification					
	AC	**SE**	**SP**	**PR**	**F1_S**	**AUC**	**AC**	**SE**	**SP**	**PR**	**F1_S**	**AUC**	**AC**	**SE**	**SP**	**PR**	**F1_S**	**AUC**	**AC**	**SE**	**SP**	**PR**	**F1_S**	**AUC**
	Two Classes																							
	0.979	0.981	0.981	0.979	0.976	0.981	0.957	0.957	0.957	0.957	0.957	0.947	0.939	0.942	0.942	0.939	0.939	0.942	0.950	0.950	0.950	0.950	0.950	0.950
	Eight Classes																							
	0.976	0.903	0.986	0.903	0.903	0.938	0.955	0.819	0.974	0.819	0.819	0.878	0.966	0.863	0.980	0.863	0.863	0.907	0.959	0.835	0.976	0.850	0.835	0.886
Four Classes	Benign						Benign						Benign						Benign					
	0.933	0.867	0.956	0.889	0.874	0.906	0.968	0.952	0.978	0.936	0.942	0.965	0.942	0.891	0.961	0.883	0.883	0.926	0.929	0.857	0.952	0.915	0.858	0.886
	Malignant						Malignant						Malignant						Malignant					
	0.937	0.874	0.958	0.874	0.797	0.880	0.887	0.775	0.925	0.775	0.775	0.828	0.956	0.912	0.970	0.912	0.912	0.907	0.921	0.842	0.947	0.842	0.842	0.850
Two Classes	**Receiver Operating Characteristics (ROC)**																							
	Mi-A		Ma-A		Benign		Malignant		Mi-A		Ma-A		Benign		Malignant		Mi-A		Ma-A		Benign		Malignant	
	0.98		0.98		0.98		0.98		0.96		0.95		0.95		0.95		0.94		0.94		0.94		0.94	
	Precision-Recall Curve (AP)																							
	Mi-A		Benign		Malignant				Mi-A		Benign		Malignant				Mi-A		Benign		Malignant			
	0.97		0.94		0.99				0.94		0.89		0.96				0.91		0.84		0.96			
	0.95		0.95		0.95		0.95																	

Table 4.

	Benign - Receiver Operating Characteristics (ROC)																							
	Mi-A	Ma-A	Ad	Fi	Phy	Tu	Mi-A	Ma-A	Ad	Fi	Phy	Tu	Mi-A	Ma-A	Ad	Fi	Phy	Tu	Mi-A	Ma-A	Ad	Fi	Phy	Tu
	0.91	0.91	0.90	0.89	1.00	0.84	0.96	0.97	1.00	0.93	0.98	0.95	0.92	0.93	0.90	0.89	0.91	1.00	0.90	0.89	0.94	0.87	0.82	0.92
	Benign - Precision-Recall Curve (AP)																							
	Mi-A	Ad	Fi	Phy	Tu		Mi-A	Ad	Fi	Phy	Tu		Mi-A	Ad	Fi	Phy	Tu		Mi-A	Ad	Fi	Phy	Tu	
	0.78	0.77	0.79	1.00	0.66		0.89	1.00	0.89	0.86	0.88		0.81	0.77	0.84	0.64	1.00		0.77	0.83	0.74	0.71	0.87	
Four Classes	**Malignant - Receiver Operating Characteristics (ROC)**																							
	Mi-A	Ma-A	Dc	Lc	Mc	Pc	Mi-A	Ma-A	Dc	Lc	Mc	Pc	Mi-A	Ma-A	Dc	Lc	Mc	Pc	Mi-A	Ma-A	Dc	Lc	Mc	Pc
	0.92	0.86	0.90	0.92	0.88	0.74	0.85	0.83	0.80	0.78	0.91	0.83	0.94	0.90	0.91	0.80	0.95	0.95	0.89	0.85	0.83	0.88	0.82	0.88
	Malignant - Precision-Recall Curve (AP)																							
	Mi-A	Dc	Lc	Mc	Pc		Mi-A	Dc	Lc	Mc	Pc		Mi-A	Dc	Lc	Mc	Pc		Mi-A	Dc	Lc	Mc	Pc	
	0.80	0.90	0.77	0.64	0.44		0.66	0.83	0.37	0.70	0.42		0.85	0.92	0.56	0.91	0.75		0.75	0.85	0.67	0.51	0.67	

Table 4

	ROC	Mi-A	Ma-A	Ad	Dc	Fi	Lc	Mc	Pc	Phy	Ta	ROC	Mi-A	Ma-A	Ad	Dc	Fi	Lc	Mc	Pc	Phy	Ta
	40× Magnification	0.94	0.93	0.95	0.94	0.99	0.96	0.92	0.81	1.00	0.88	100× Magnification	0.90	0.88	0.95	0.88	0.89	0.83	0.85	0.78	0.95	0.88
	200× Magnification	0.92	0.90	0.91	0.94	0.86	0.77	0.81	0.99	0.94	0.99	400× Magnification	0.91	0.89	0.95	0.88	0.93	0.76	0.87	0.88	0.91	0.92
	PR/AP	Mi-A	Ad	Dc	Fi	Lc	Mc	Pc	Phy	Ta		PR/AP	Mi-A	Ad	Dc	Fi	Lc	Mc	Pc	Phy	Ta	
Eight Classes	40× Magnification	0.83	0.91	0.91	0.93	0.77	0.74	0.51	1.00	0.63		100× Magnification	0.69	0.91	0.81	0.75	0.47	0.61	0.44	0.73	0.55	
	200× Magnification	0.76	0.75	0.88	0.70	0.46	0.62	0.81	0.58	0.87		400× Magnification	0.72	0.82	0.80	0.67	0.38	0.67	0.66	0.83	0.84	

Table 5: Class Performance of The Proposed End-To-End Hybrid Ensemble Model Using Accuracy, Sensitivity, Specificity, Precision, F1_scorein All Magnifications. Ad=Adenosis, Fi=Fibroadenoma, Phy=Phyllodes_Tumor,Ta=Tubular_Adenoma, Dc=Ductal_Carcinoma,Lc=Lobular_Carcinoma, Mc=Mucinous_Carcinomaand Pc=Papillary_Carcinoma.

	Metrics	40× Magnification				100× Magnification				200× Magnification				400× Magnification			
Two Classes		Benign		Malignant		Benign		Malignant		Benign		Malignant		Benign		Malignant	
	Accuracy	0.979		0.979		0.957		0.957		0.939		0.939		0.950		0.950	
	AUC	0.981		0.981		0.947		0.947		0.942		0.942		0.950		0.950	
	F1_score	0.967		0.985		0.929		0.969		0.905		0.955		0.924		0.963	
	Precision	0.952		0.992		0.937		0.965		0.864		0.977		0.902		0.975	
	Specificity	0.978		0.983		0.972		0.922		0.934		0.95		0.951		0.948	
	Sensitivity	0.983		0.978		0.922		0.972		0.95		0.934		0.948		0.951	
Four Classes		Benign – Class Performance															
	Metrics	Ad	Fi	Phy	Tu	Ad	Fi	Phy	Tu	Ad	Fi	Phy	Tu	Ad	Fi	Phy	Tu
	Accuracy	0.950	0.883	1.00	0.900	1.00	0.935	0.968	0.968	0.950	0.900	0.917	1.00	0.964	0.857	0.929	0.964
	AUC	0.899	0.889	1.00	0.835	1.00	0.926	0.980	0.954	0.899	0.894	0.910	1.00	0.939	0.872	0.818	0.917
	F1_score	0.857	0.867	1.00	0.769	1.00	0.917	0.923	0.929	0.857	0.880	0.782	1.00	0.900	0.846	0.778	0.909
	Precision	0.900	0.821	1.00	0.833	1.00	0.957	0.857	0.929	0.900	0.917	0.692	1.00	0.900	0.757	1.00	1.00
	Specificity	0.980	0.857	1.00	0.957	1.00	0.973	0.960	0.979	0.980	0.941	0.920	1.00	0.978	0.788	1.00	1.00
	Sensitivity	0.818	0.920	1.00	0.714	1.00	0.880	1.00	0.929	0.818	0.846	0.900	1.00	0.900	0.957	0.636	0.833
		Malignant - Class Performance															
	Metrics	Dc	Lc	Mc	Pc	Dc	Lc	Mc	Pc	Dc	Lc	Mc	Pc	Dc	Lc	Mc	Pc
	Accuracy	0.911	0.970	0.933	0.933	0.796	0.894	0.944	0.915	0.927	0.942	0.985	0.971	0.850	0.958	0.917	0.958
	AUC	0.895	0.925	0.878	0.742	0.798	0.777	0.911	0.826	0.910	0.804	0.947	0.949	0.830	0.875	0.820	0.875
	F1_score	0.932	0.867	0.780	0.609	0.830	0.571	0.826	0.625	0.945	0.714	0.944	0.857	0.886	0.800	0.688	0.800
	Precision	0.911	0.867	0.762	0.778	0.877	0.526	0.792	0.555	0.925	0.833	1.00	0.800	0.875	0.833	0.688	0.833
	Specificity	0.837	0.983	0.957	0.983	0.808	0.929	0.958	0.938	0.854	0.983	1.00	0.976	0.762	0.981	0.952	0.981
	Sensitivity	0.953	0.866	0.800	0.500	0.789	0.625	0.864	0.714	0.966	0.625	0.895	0.923	0.897	0.769	0.688	0.769

Table 5

	Magnification	Metrics	Ad	Dc	Fi	Lc	Mc	Pc	Phy	Ta	Magnification	Ad	Dc	Fi	Lc	Mc	Pc	Phy	Ta
Eight Classes	40×	Accuracy	0.995	0.944	0.990	0.979	0.969	0.959	1.00	0.969	100×	0.995	0.882	0.966	0.946	0.951	0.956	0.980	0.961
		AUC	0.955	0.942	0.994	0.958	0.916	0.813	1.00	0.885		0.955	0.883	0.894	0.828	0.853	0.778	0.951	0.880
		F1_score	0.952	0.936	0.962	0.875	0.85	0.692	1.00	0.786		0.952	0.870	0.851	0.666	0.762	0.64	0.846	0.733
		Precision	1.00	0.941	0.926	0.824	0.85	0.75	1.00	0.786		1.00	0.851	0.909	0.647	0.8	0.727	0.786	0.688
		Specificity	1.00	0.954	0.988	0.983	0.983	0.983	1.00	0.983		1.00	0.877	0.989	0.968	0.978	0.984	0.984	0.974
		Sensitivity	0.909	0.930	1.00	0.933	0.850	0.643	1.00	0.786		0.909	0.889	0.800	0.688	0.727	0.571	0.917	0.786
	200×	Accuracy	0.985	0.934	0.954	0.949	0.959	0.985	0.970	0.990	400×	0.989	0.881	0.943	0.943	0.966	0.972	0.989	0.989
		AUC	0.906	0.937	0.860	0.773	0.813	0.992	0.937	0.995		0.947	0.881	0.930	0.757	0.869	0.878	0.909	0.916
		F1_score	0.857	0.930	0.809	0.643	0.750	0.897	0.750	0.929		0.900	0.868	0.808	0.583	0.800	0.800	0.900	0.909
		Precision	0.900	0.896	0.905	0.750	0.923	0.813	0.643	0.867		0.900	0.852	0.724	0.636	0.857	0.833	1.00	1.00
		Specificity	0.995	0.907	0.988	0.983	0.994	0.983	0.973	0.989		0.994	0.878	0.9478	0.975	0.988	0.988	1.00	1.00
		Sensitivity	0.818	0.966	0.731	0.563	0.632	1.00	0.900	1.00		0.900	0.885	0.913	0.538	0.750	0.769	0.818	0.833

For Eight-class multi-classification, the lowest result was seen at the 100x magnification with an accuracy of 0.955, sensitivity (0.819), specificity (0.974), precision (0.819), F1_score (0.819) and AUC (0.878) whereas the highest performance is seen at the 40x magnification with an accuracy of 0.976, sensitivity (0.903), specificity (0.986), precision (0.903), F1_score (0.903) and AUC (0.938). The AUC and AP are used to evaluate the 8-classes multi-classification of the proposed model. The 40x magnification result is the best for both area (AUC) and AP. Since the BreakHis dataset used has a huge class imbalance, we carried out a class evaluation of each classification scenario as seen in (Table 5). For the Binary techniques, the results of the malignant supersede that of the benign in all magnifications thus showing that the proposed model can detect images from classes with less data augmentation. The four-class benign multi-classification illustrates that our model performance on the Adenosis and Phyllodes_tumor have superior results in all magnifications whereas, the Mucinous_carcinoma and Papillary_carcinoma class recorded the best for malignant. The eight classes experiment shows that the proposed model is adequate for a multi-classification task with a huge data sample. We graphically illustrate the class performance of all the experiments in terms of accuracy as shown in (Figure 5a) denoting Eight classes experiment, 5(b) and 5(c) denoting four classes experiment and 5(d) denoting the binary experiment. Figures 6 - 9 summaries the performance of the proposed model using the confusion metrics evaluation.

DISCUSSION

Since the inception of deep learning techniques, it has been much easier to extract discriminative breast cancer features from high-resolution images. From our literature survey, we noticed that binary classification is prominent in research via breast cancer classification where models are developed to classify breast liaison as benign or malignant. Furthermore, researchers who engaged in multi-class classification mainly focused on the four class classifications which are Adenosis, Fibroadenoma, Phyllodes_tumor, and Tubular_adenoma for the benign set or Ductal_carcinoma, Lobular_carcinoma, Mucinous_carcinoma, Papillary_carcinoma for the malignant set based on the BreakHis data. Rarely is the eight-class multi-class classification of breast cancer discussed due to its low classification accuracy and rectified receptive field. However, this can besolved by using a variety of techniques, including ensemble models, CNN models with various receptive fields, extraction of features and fusion, attention mechanisms, and more which is the key objective of this study.

The proposed model performed very well via the binary classification with its optimal performance at 40x magnification with an accuracy of 0.979, sensitivity (0.981), specificity (0.981), precision (0.979), F1_score (0.976) and AUC (0.981). Comparing the class performance, the malignant was very much detected by the proposed model. This is linked to the fact that fewer images were generated for the malignant class as the class had bigger samples of the original samples compared to the benign class. This can also be traced to the multi-classification performance of the model using the four classes. The model performance is outstanding for the classes with large original samples than that of the augmented classes. For the Benign class, the100x magnification had the best performance with an accuracy of 0.968, sensitivity (0.952), specificity (0.978), precision (0.936), F1_score (0.942) and AUC (0.965) whereas the 200x magnification yielded the best for the malignant class with an accuracy of 0.956, sensitivity (0.912), specificity (0.970), precision (0.912), F1_score (0.912) and AUC (0.907). In support of the claims that the model performed very well on the original images than the augmented images, taking a look at (Table 1) and multi-classification (Eight class), we can see that the individual classes that had fewer augmented images all had an accuracy, sensitivity, specificity, precision of 1.0. This led to the future work of this study; augmenting the images using different generative models and running an analysis using our proposed model.

Comparison with state-of-the-art results

The result comparison is based on the networks that made use of the same BreakHis (Man et al., 2020) dataset for a fair comparison as shown in (Table 6). We focused more on recently published papers since technically their result and approach are state-of-the-art. For Binary classification, Authors like Chattopadhyay et al. (2022) proposed the DRDA-Net7 model (DRDA-Net, which stands for residual dual-shuffle attention network, is a dual-shuffle attention-guided deep learning model that improves the model's capacity to recognize intricate patterns in images. It was inspired by the bottleneck unit of the ShuffleNet design), Sharma and Kumar et al. (Sharma and Kumar, 2022 used the Xception model and SVM classifier and Liu et al. (2022) employed Autoencoder and Siamese Network for binary classification. Although the recorded results are promising, the result of Liu et al. (2022) exhibited more reliability among them

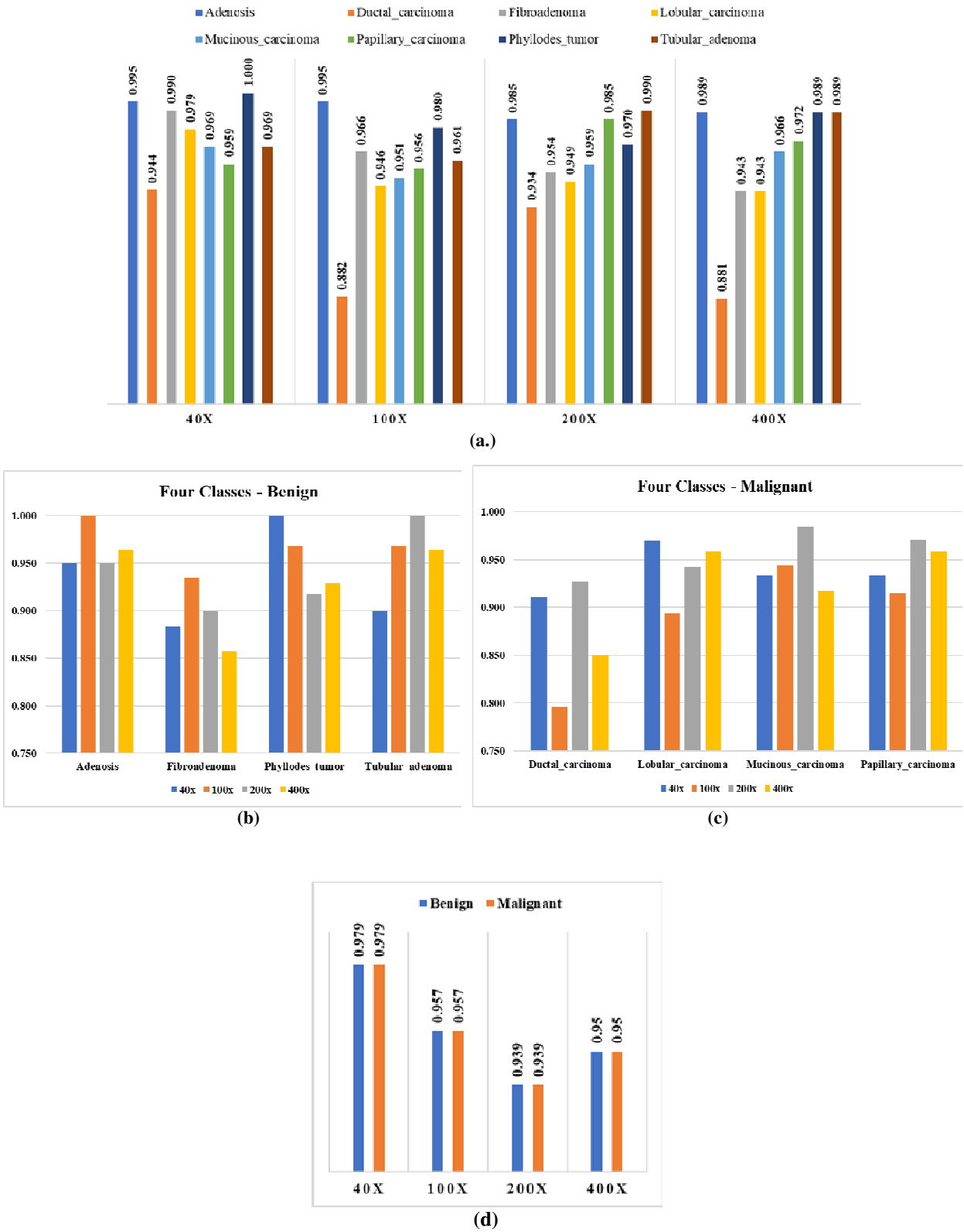

Figure 5: Graphically Illustration of The Class Performance (Accuracy) On The 3-Classification Scenario (Eight Classes, Four Classes-Benign and Malignant and The Binary Classes) Using the Proposed Model.

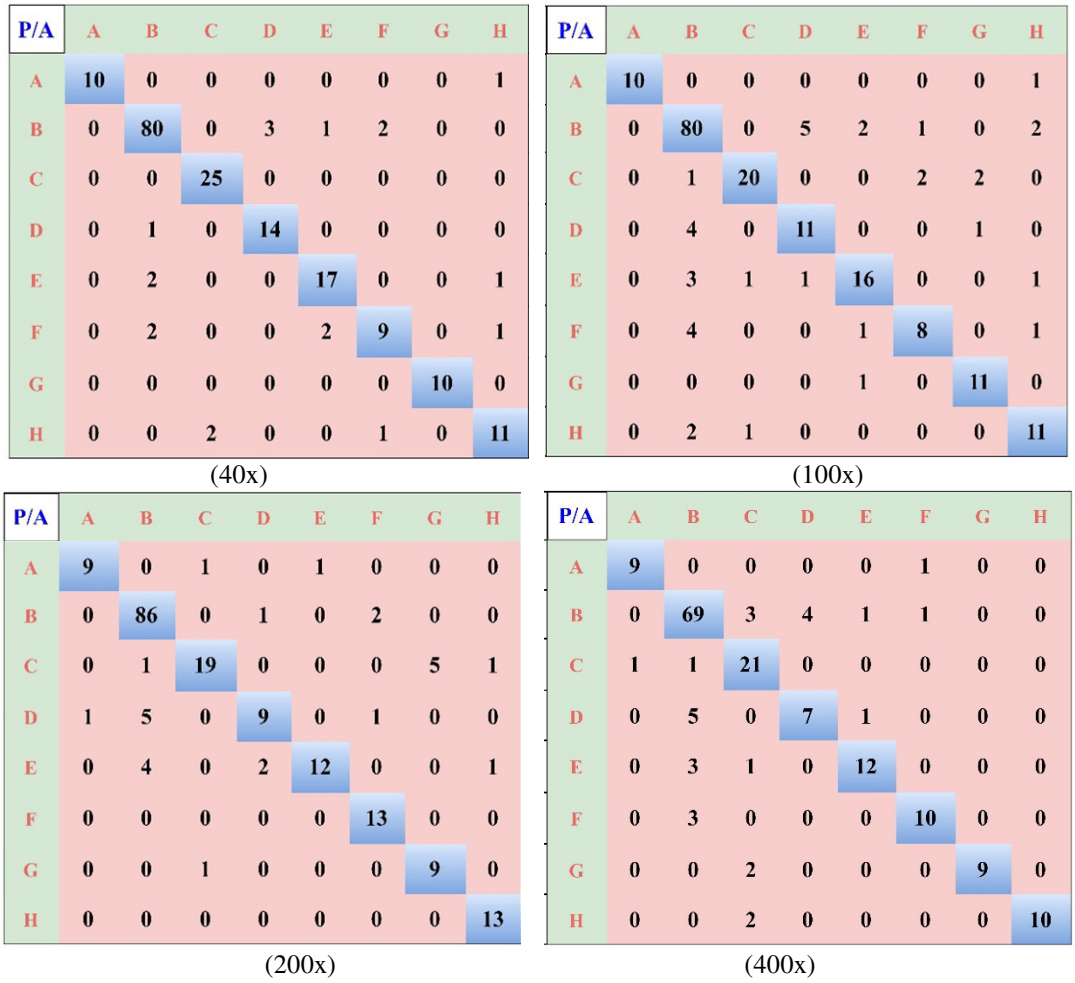

Figure 6. Eight Classes Experiment Confusion Matrix on All Magnifications. P/A Denotes Predicted and Actual. A = Adenosis, B = Fibroadenoma, C = Phyllodes_Tumor, D = Tubular_Adenoma, E = Ductal_Carcinoma, F = Lobular_Carcinoma, G = Mucinous_Carcinoma and H = Papillary_Carcinoma

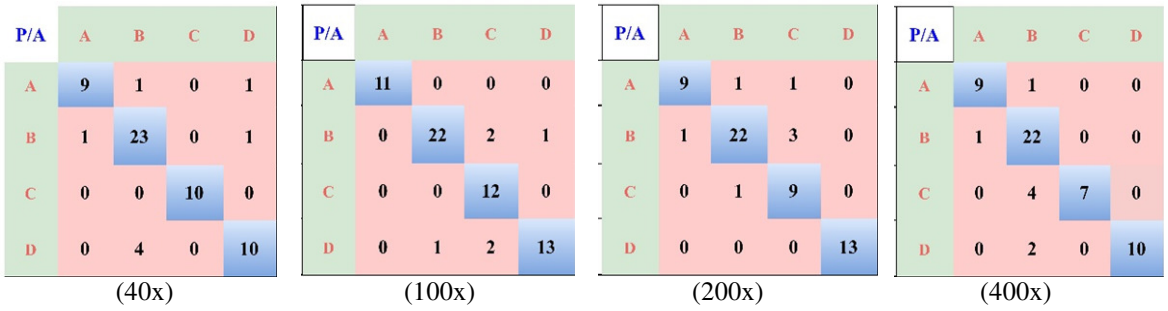

Figure 7: Four Classes Experiment (Benign) Confusion Matrix on All Magnifications. P/A Denotes Predicted and Actual. A = Adenosis, B = Fibroadenoma, C = Phyllodes_Tumor, D = Tubular_Adenoma.

with an accuracy of 97.3% for 40x, 96.1% for 100x, 97.8% for 200x, and 96.7% for 400x magnification respectively. Saini and Susan (2022) and Sharma et al., (2020) employed the VGG16 via multi-classification of Breast Cancer using the BreakHis Pimkin et al., (2018). Specifically, Saini and Susan (2022) proposed the VGGIN-Net in two scenarios; Finetuned and non-finetuned. The VGG16 pre-trained model's layers up to the block 4 pool layer were frozen and concatenated to create the VGGIN-Net, which also included a nonlinear

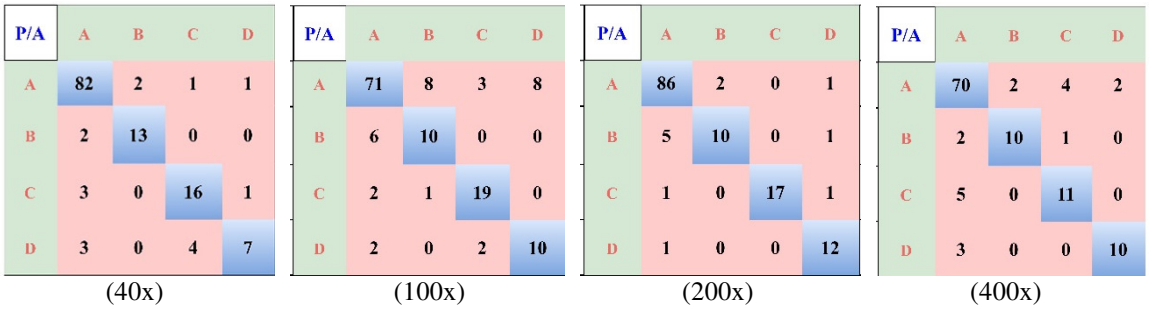

Figure 8: Four Classes Experiment (Malignant) Confusion Metrics on All Magnifications. P/A Denotes Predicted and Actual. A = Ductal_Carcinoma, B = Lobular_Carcinoma, C = Mucinous_Carcinoma, D = Papillary_Carcinoma

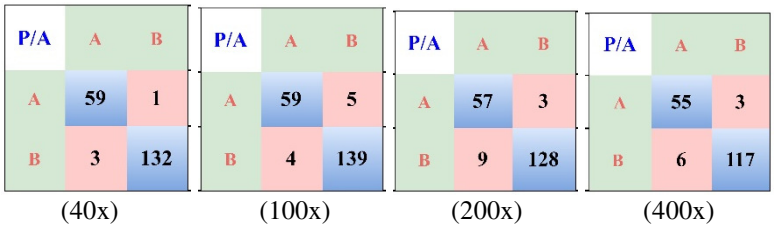

Figure 9: Binary Experiment Confusion Metrics on All Magnifications. P/A Denotes Predicted and Actual. A = Benign and B = Malignant.

Table 6: Summary of the state-of-the-art results against the proposed model.

Ref/Year	Approach	Class	Magnification				
			Merge	40x	100x	200x	400x
Chattopadhyay et al. [32] 2022	DRDA-Net7 model	Binary	-	96.10	96.03	96.08	96.02
Sharma and Kumar et al. [33] 2022	Xception model and SVM classifier	Binary	-	96.25	96.25	95.74	94.11
Liu et al. [34] 2022	AE + Siamese Network	Binary	-	97.3	96.1	97.8	96.7
Saini et al., [35] 2022	VGGIN-Net	Four	-	96.21	97.44	96.22	93.49
	Finetuned VGGIN-Net		-	97.56	96.89	97.49	94.21
Sharma et al., [36] 2020	VGG16 + Linear SVM (Patch Based)	Four	-	93.97	92.92	91.23	91.79
	VGG16 + Linear SVM (Patient Based)		-	93.25	91.87	91.5	92.31
Umer et al., [37] 2022	ResNet50 + ESKNN	Eight	80.10	-	-	-	-
	ResNet50 + ESD		69.40	-	-	-	-
	6B-Net + ESKNN		82.03	-	-	-	-
	6B-Net + ESD		70.43	-	-	-	-
	Fused + ESKNN		87.90	-	-	-	-
	Fused + ESD		74.30	-	-	-	-
	F-Selected + ESKNN		90.10	-	-	-	-
	F-Selected + ESD		75.60	-	-	-	-
Ameh et al., [38] 2022	HF + DNN (100) augmented	Eight	-	97.24	96.62	95.44	96.29
	HF + DNN (400) augmented		-	97.89	97.6	96.10	96.84
	HF + DNN (400)		-	90.87	89.57	91.58	88.67
	HF + DNN (1000)		-	90.01	88.35	92.95	89.54
Sharma et al., [39] 2020	CNN + Pooling Strategy	Eight	-	80.76	76.58	79.90	74.21
Asare [40] 2020	Self-Training and Self-Paced Learning	Eight	-	94.57	94.25	95.1	94.47
Boumaraf et al. [41] 2021	Transfer Learning (ResNet)	Eight	92.15	94.49	93.27	91.29	89.56
Ours	Proposed	Eight		97.6	95.5	96.6	95.9

mapping rudimentary Inception block unit (at the higher level), achieving a classification accuracy of 97.56% for 40x, 96.89% for 100x, 97.49% for 200x, and 94.21% for 400x using the finetuned approach and 96.21% for 40x, 97.44% for 100x, 96.22% for 200x, 93.49% for 400x magnification respectively without finetuning. On the other hand, Sharma et al., (2020) used a linear SVM as a classifier and VGG16 as a feature extractor via a patch-based approach. However, their results were poor compared to that of Saini and Susan (2022).

An observation is made via the eight classes' multi-classification. Only a few researchers tend to tackle this issue due to the low accuracy performance of models. However, notable authors have come up with improved models that were able to give higher accuracy. A unique 6BNet featuring six continuous nodes and various receptive fields is proposed by Umer et al. (2022). Among all the approaches recorded, the F-selected +ESKNN approach recorded the highest accuracy of 90.10%. Ameh et al., (2022) on the other hand, employed four different approaches which the HF (handcrafted features) + DNN (400) augmented approach recorded the highest accuracy with 97.89% for 40x, 97.6% for 100x, 96.10% for 200x, and 96.84% for 400x magnifications. Sharma et al., [39] proposed the use of CNN + Pooling Strategy, Asare (2020) proposed the use of Self-Training and Self-Paced Learning whereas Boumaraf et al. (2021) suggested the use of Transfer learning (ResNet) Which is based on a block-wise fine-tuning strategy. Asare (2020) contribution is based on a semi-supervised learning technique that generates and chooses pseudolabeled samples for categorizing breast cancer histopathology images while including self-training and identity learning. In summary, our proposed model result demonstrated superiority over all the discussed models in all experiments (Binary, Multi-classification).

Conclusion

This study proposes an end-to-end hybrid ensemble model for the multi-classification of BC based on the attention mechanism and global second-order pooling. The enhancement of deep learning pre-trained models is the foundation of the suggested model. Following the choice of the suggested ensemble model, six pre-trained models were thoroughly evaluated, and the DenseNet, ResNet, and Xception architectures were chosen. BreaKHis dataset is used for the experiment. The proposed model is utilized for binary classification as well as multi-classification (four classes and eight classes). The best accuracy for classifying BC was attained, with 97.9% for binary classification, 96.8% for benign multi-classification, 95.6% for malignant multi-classification, and 97.9% for eight-class classifications. According to the testing findings, the magnifications of 40x, 100x, 200x, and 40x yield the maximum accuracy. To expand the functionality of our model and increase the accuracy of BC diagnosis, we intend to investigate a variety of data augmentation strategies in the future, such as GAN for augmenting training samples.

Acknowledgments

The authors gratefully acknowledge the support of the National Natural Science Foundation of China (NFSC, Grant No. 52007025) and the Science and Technology Support Program of Sichuan Province (2022JDRC0025).

Compliance with ethical standards

Conflict of interest

The authors have disclosed that they do not have any conflicting interests.

Ethical Approval

The article did not involve any studies that included human or animal subjects conducted by the authors

Informed consent

All participants included in the study provided their informed consent.

REFERENCES

Albashish, D., Al-Sayyed, R., Abdullah, A., Ryalat, M. H., & Ahmad Almansour, N. (2021). “Deep CNN Model based on VGG16 for Breast Cancer Classification,” 2021 International Conference on Information Technology, ICIT 2021 - Proceedings, pp. 805–810, Jul. 2021, doi: 10.1109/ICIT52682.2021.9491631.

Aljuaid, H., Alturki, N., Alsubaie, N., Cavallaro, L., & Liotta, A. (2022). “Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning,” Comput Methods Programs Biomed, vol. 223, Aug. 2022, doi: 10.1016/j.cmpb.2022.106951.

Ameh Joseph, A., Abdullahi, M., Junaidu, S. B., Hassan Ibrahim, H., & Chiroma, H. (2022). “Improved multi-classification of breast cancer histopathological images using handcrafted features and deep neural network (dense layer),” Intelligent Systems with Applications, vol. 14, p. 66, 2022, doi: 10.1016/j.iswa.2022.20.

Arooj, S., Atta-ur-Rahman, Zubair, M., Khan, M. F., Alissa, K., Khan, M. A., & Mosavi, A. (2022). “Breast Cancer Detection and Classification Empowered with Transfer Learning,” Front Public Health, vol. 10, Jul. 2022, doi: 10.3389/fpubh.2022.924432.

Asare, S. K., You, F., & Nartey, O. T. (2020). “A Semisupervised Learning Scheme with Self-Paced Learning for Classifying Breast Cancer Histopathological Images,” Computational Intelligence and Neuroscience, vol. 2020, pp. 1–16, Dec. 2020, doi: 10.1155/2020/8826568.

Budd, S., Robinson, E. C., & Kainz, B. (2021). A survey on active learning and human-in-the-loop deep learning for medical image analysis. Medical Image Analysis, 71, 102062. https://doi.org/10.1016/j.media.2021.102062.

Boumaraf, S., Liu, X., Zheng, Z., Ma, X., & Ferkous, C. (2021). “A new transfer learning-based approach to magnification dependent and independent classification of breast cancer in histopathological images,” Biomedical Signal Processing and Control, vol. 63, p. 102192, Jan. 2021, doi: 10.1016/j.bspc.2020.102192.

Chattopadhyay, S., Dey, A., Singh, P. K., & Sarkar, R. (2022). “DRDA-Net: Dense residual dual-shuffle attention network for breast cancer classification using histopathological images,” Computers in Biology and Medicine, vol. 145, p. 105437, Jun. 2022, doi: 10.1016/j.compbiomed.2022.105437.

Elmore, J. G., Longton, G. M., Carney, P. A., Geller, B. M., Onega, T., Tosteson, A. N. A., Nelson, H. D., Pepe, M. S., Allison, K. H., Schnitt,

S. J., O'Malley, F. P., & Weaver, D. L. (2015). Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens. JAMA, 313(11), 1122. https://doi.org/10.1001/jama.2015.1405.

Hameed, Z., Garcia-Zapirain, B., Aguirre, J. J., & Isaza-Ruget, M. A. (2022). Multiclass classification of breast cancer histopathology images using multilevel features of deep convolutional neural network. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-19278-2.

Heenaye-Mamode Khan, M., Boodoo-Jahangeer, N., Dullull, W., Nathire, S., Gao, X., Sinha, G. R., & Nagwanshi, K. K. (2021). Multi-class classification of breast cancer abnormalities using Deep Convolutional Neural Network (CNN). PLOS ONE, 16(8), e0256500. https://doi.org/10.1371/journal.pone.0256500.

Liu, M., He, Y., Wu, M., & Zeng, C. (2022). "Breast Histopathological Image Classification Method Based on Autoencoder and Siamese Framework," Information, vol. 13, no. 3, p. 107, Feb. 2022, doi: 10.3390/info13030107.

Man, R., Yang, P., & Xu, B. (2020). Classification of Breast Cancer Histopathological Images Using Discriminative Patches Screened by Generative Adversarial Networks. IEEE Access, 8, 155362–155377. https://doi.org/10.1109/access.2020.3019327.

[14] Mi, W., Li, J., Guo, Y., Ren, X., Liang, Z., Zhang, T., & Zou, H. (2021). "Deep learning-based multi-class classification of breast digital pathology images," Cancer Manag Res, vol. 13, pp. 4605–4617, 2021, doi: 10.2147/CMAR.S312608.

Nawaz, M., A., A., & Hassan, T. (2018). "Multi-Class Breast Cancer Classification using Deep Learning Convolutional Neural Network," International Journal of Advanced Computer Science and Applications, 9(6). https://doi.org/10.14569/ijacsa.2018.090645.

Perronnin, F., Sánchez, J., & Mensink, T. (2010). "Improving the Fisher kernel for large-scale image classification", Proc. Eur. Conf. Comput. Vis., pp. 143-156, 2010.

Pimkin, A., Makarchuk, G., Kondratenko, V., Pisov, M., Krivov, E., & Belyaev, M. (2018). "Ensembling Neural Networks for Digital Pathology Images Classification and Segmentation," Image Analysis and Recognition, pp. 877–886, 2018, doi: 10.1007/978-3-319-93000-8_100.

Sabtu, S. N., Abdul Sani, S. F., Bradley, D. A., Looi, L. M., & Osman, Z. (2019). A review of the applications of Raman spectroscopy for breast cancer tissue diagnostic and their histopathological classification of epithelial to mesenchymal transition. Journal of Raman Spectroscopy, 51(3), 380–389. Portico. https://doi.org/10.1002/jrs.5774

Saini, M., & Susan, S. (2023). "VGGIN-Net: Deep Transfer Network for Imbalanced Breast Cancer Dataset," IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 1–1, 2022, doi: 10.1109/tcbb.2022.3163277.

Sharma, S., & Kumar, S. (2022). "The Xception model: A potential feature extractor in breast cancer histology images classification," ICT Express, vol. 8, no. 1, pp. 101–108, Mar. 2022, doi: 10.1016/j.icte.2021.11.010.

Sharma, S., & Mehra, R. (2020). "Conventional Machine Learning and Deep Learning Approach for Multi-Classification of Breast Cancer Histopathology Images—a Comparative Insight," Journal of Digital Imaging, vol. 33, no. 3, pp. 632–654, Jan. 2020, doi: 10.1007/s10278-019-00307-y.

Sharma, S., Mehra, R., & Kumar, S. (2020). "Optimised CNN in conjunction with efficient pooling strategy for the multi-classification of breast cancer," IET Image Processing, vol. 15, no. 4, pp. 936–946, Dec. 2020, doi: 10.1049/ipr2.12074.

Singh, R., Ahmed, T., Kumar, A., Singh, A. K., Pandey, A. K., & Singh, S. K. (2020). "Imbalanced Breast Cancer Classification Using Transfer Learning," IEEE/ACM Trans Comput Biol Bioinform, vol. 18, no. 1, pp. 83–93, Jan. 2021, doi: 10.1109/TCBB.2020.2980831.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians, 71(3), 209–249. Portico. https://doi.org/10.3322/caac.21660.

Umer, M. J., Amin, J., Sharif, M., Anjum, M. A., Azam, F., & Shah, J. H. (2021). An integrated framework for COVID-19 classification based on classical and quantum transfer learning from a chest radiograph. Concurrency and Computation: Practice and Experience, vol. 34, no. 20, Jun. 2021, doi: 10.1002/cpe.6434.

Umer, M. J., Sharif, M., Kadry, S., & Alharbi, A. (2022). "Multi-Class Classification of Breast Cancer Using 6B-Net with Deep Feature Fusion and Selection Method," J Pers Med, vol. 12, no. 5, May 2022, doi: 10.3390/jpm12050683.

Ukwuoma, C. C., Zhiguang, Q., Bin Heyat, M. B., Mohammed Khan, H., Akhtar, F., Masadeh, M. S., Bamisile, O., AlShorman, O., & Nneji, Grace. U. (2022). Detection of Oral Cavity Squamous Cell Carcinoma from Normal Epithelium of the Oral Cavity using Microscopic Images. 2022 International Conference on Decision Aid Sciences and Applications (DASA). https://doi.org/10.1109/dasa54658.2022.9765023.

Ukwuoma, C. C., Urama, G. C., Qin, Z., Bin Heyat, M. B., Mohammed Khan, H., Akhtar, F., Masadeh, M. S., Ibegbulam, C. S., Delali, F. L., & AlShorman, O. (2022). Boosting Breast Cancer Classification from Microscopic Images Using Attention Mechanism. 2022 International Conference on Decision Aid Sciences and Applications (DASA). https://doi.org/10.1109/dasa54658.2022.9765013.

Ukwuoma, C. C., Qin, Z., Belal Bin Heyat, M., Akhtar, F., Bamisile, O., Muaad, A. Y., Addo, D., & Al-antari, M. A. (2022). A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images. Journal of Advanced Research. https://doi.org/10.1016/j.jare.2022.08.021.

Ukwuoma, C. C., Qin, Z., Agbesi, V. K., Cobbinah, B. M., Yussif, S. B., Abubakar, H. S., & Lemessa, B. D. (2022). "Dual_Pachi: Attention-based dual path framework with intermediate second order-pooling for Covid-19 detection from chest X-ray images," Computers in Biology and Medicine, vol. 151, p. 106324, Dec. 2022, doi: 10.1016/j.compbiomed.2022.106324.

Ukwuoma, C. C., Qin, Z., Agbesi, V. K., Ejiyi, C. J., Bamisile, O., Chikwendu, I. A., Tienin, B. W., & Hossin, M. A. (2022), "LCSB-inception: Reliable and effective light-chroma separated branches for Covid-19 detection from chest X-ray images," Computers in Biology and Medicine, vol. 150, p. 106195, Nov. 2022, doi: 10.1016/j.compbiomed.2022.106195.

Ukwuoma, C. C., Hossain, M. A., Jackson, J. K., Nneji, G. U., Monday, H. N., & Qin, Z. (2022). "Multi-Classification of Breast Cancer Lesions in Histopathological Images Using DEEP_Pachi: Multiple Self-Attention Head," Diagnostics, vol. 12, no. 5, May 2022, doi: 10.3390/diagnostics12051152.

Veta, M., Pluim, J. P. W., van Diest, P. J., & Viergever, M. A. (2014). "Breast Cancer Histopathology Image Analysis: A Review," IEEE Trans Biomed Eng, vol. 61, no. 5, 2014, doi: 10.1109/TBME.2014.2303852.

Wang, X., Ahmad, I., Javeed, D., Zaidi, S. A., Alotaibi, F. M., Ghoneim, M. E., Daradkeh, Y. I., Asghar, J., & Eldin, E. T. (2022). "Intelligent Hybrid Deep Learning Model for Breast Cancer Detection," Electronics (Switzerland), vol. 11, no. 17, Sep. 2022, doi: 10.3390/electronics11172767.

Wang, P., Wang, J., Li, Y., Li, P., Li, L., & Jiang, M. (2021). "Automatic classification of breast cancer histopathological images based on deep feature fusion and enhanced routing," Biomed Signal Process Control, vol. 65, p. 102341, Mar. 2021, doi: 10.1016/J.BSPC.2020.102341.

World Health Organization, WHO Position Paper on Mammography Screening, World Health Organization, 2014.

Yan, R., Ren, F., Wang, Z., Wang, L., Zhang, T., Liu, Y., Rao, X., Zheng, C., & Zhang, F. (2020). "Breast cancer histopathological image classification using a hybrid deep neural network," Methods, vol. 173, no. February 2019, pp. 52–60, 2020, doi: 10.1016/j.ymeth.2019.06.014.

Zhang, X., Zhang, Y., Zhang, Q., Ren, Y., Qiu, T., Ma, J., & Sun, Q. (2019). Extracting comprehensive clinical information for breast cancer using deep learning methods. International Journal of Medical Informatics, 132, 103985. https://doi.org/10.1016/j.ijmedinf.2019.103985.