JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
SICAIET
2024
Research Article
Collection: Artificial Intelligence and Emerging Technologies

Breast Cancer Diagnosis: A Transfer Learning-based System for Detection of Invasive Ductal Carcinoma (IDC)

Sukana Zulfqar[1], M. Azam Zia[1*], Faisal Mehmood[2], Athar Pervez[3], and Touqeer Abbas[2]

[1]Department of Computer Science, University of Agriculture Faisalabad, Faisalabad, 38000, Pakistan.
[2]Department of Computer Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China.
[3]Department of Pathology, Frontier Medical and Dental College, Abbottabad, 22020, Pakistan.
[*]Corresponding Author: M. Azam Zia. Email: mazamzia@uaf.edu.pk

Academic Editor: Salman Qadri **Published:** February 01, 2024

Abstract: Breast cancer (BC) is a form of cancer that originates in the breast. BC cells typically form a tumor readily detectable on an X-ray or can be inspected as a lump. Even though improvements in screening, treatment, and monitoring have made patient survival rates higher, BC is still the common type of cancer in women and the main reason they die from cancer. Invasive ductal carcinoma (IDC) makes up about 85% of all cases that have been studied. It is the most aggressive type of BC. The purpose of this reserach is to find IDC early on so that it can be treated quickly. In the present study, we attempted to apply deep learning principles to detect breast-invasive ductal carcinoma via transfer learning. This study implemented six transfer learning approaches on two widely used datasets: Agios Pavlos and Databiox. The approaches utilized were InceptionReNetV2, DenseNet-121, DenseNet201, VGG19, ResNet-50, and ResNet-101v2. The models' efficacy in identifying IDC demonstrates their potential utility in assisting pathologists in illness diagnosis. The experimental results that the authors have attained concerning the accuracy: DenseNet121 with 97.13%, DenseNet201 with 96.34%, VGG19 with 95.65%, ResNet-50 with 94.90%, ResNet-101v2 with 93.53%, and InceptionReNetV2 with 93.20%. It is evident from the experimental findings that DenseNet121 provides the highest level of accuracy for cancer detection, while InceptionReNetV2 provides the lowest level of accuracy.

Keywords: IDC; Deep learning; Breast cancer; Databiox; Agios Pavlos.

1. Introduction

Breast cancer (BC) is a form of cancer that begins in breast tissue. It develops when aberrant breast cells proliferate and grow out of control, resulting in a tumor. These tumors can infect nearby tissues and sometimes travel to other body areas via the lymphatic system or bloodstream. BC remains a significant cause of death worldwide, particularly among women. The number of BC deaths can vary from year to year due to various factors such as advancements in screening, early detection, and treatment options, as well as changes in population demographics and healthcare accessibility. It is estimated that 290,560 new cases were reported (287,850 women, 2,710 men), and 43,780 deaths occurred (43,250 women and 530 men) from BC in the United States in 2022 [1]. Before being diagnosed with BC, a biopsy or mammogram must be performed. The standard procedure involves staining the removed tissue sample with H&E and examining it beneath a microscope [2, 3]. To detect any abnormal cells in the tissue, a pathologist examines

it under a microscope. During the inspection, a critical stage before prognosis, pathologists identify the cancer phase based on the extent of the growth cells [4]. On the other hand, identifying and rating IDC are time-consuming and complicated procedures that frequently add to heterogeneity between and within observers during diagnosis [5-7].

As a result, CAD-based technologies can help pathologists find erroneous cells and streamline the procedure overall. Many DL models have been developed for years to identify malignant cells. These models have effectively identified and labeled diverse tumor cell patterns within histology pictures. This work uses ensembles of deep-learning models to identify and classify IDC BC. When analyzing the same dataset as other published approaches, our ensemble of DenseNets achieves the best F-score possible while maintaining a balanced level of accuracy. This enables us to utilize it for the identification of IDC. The Databiox dataset [5] is used for IDC grading, while an extra dataset called "Agios Pavlos" [6] is applied for validation purposes. It has been noted that the quantitative outcomes obtained with the identical methodology on the Agios dataset are significantly superior to those obtained with the Databiox dataset. This could imply that the complexity of the photos in the dataset is the cause of the model's difficulty in predicting the Databiox dataset. Real-world medical images are frequently varied and complex, making it challenging to produce high-quality results consistently using CAD-based technologies. Given this discovery, we can infer that our approach is capable of generating favorable results even when dealing with intricate sets of images. To our knowledge, this is the first time a Databiox dataset has been used to develop an automated IDC BC grading system.When we compare our DenseNets ensemble to other methods published on the same dataset, it gets the highest F-score and the most accurate results for identifying IDCs. The Databiox dataset was used to create an East cancer scoring system.

We have five contributions to this study as given below:

- On the IDC detection dataset, our ensemble model outperformed previously published approaches and obtained state-of-the-art results..
- • For IDC detection, a number of comparison analyses have been conducted.
- To grade IDC BC, a great deal of study has been done utilizing the Databiox dataset.. To our knowledge, no prior study has utilized this dataset for grading IDC.
- Databiox and Agios Pavlo's datasets were used to test the proposed system for IDC grading.
- Experimental results show that DenseNet121 provides the most accuracy (97.13%) while grading cancer, whereas InceptionReNetV2 provides the lowest accuracy (93.20%).

The following is the outline for this paper: Section 1 will offer a quick introduction to cancer, Section 2 will discuss previous work in this field, and Section 3 will explain the methodology, materials, and approaches employed in this study. The experimental research, with full results and discussion, is presented in Section 4, and the work is brought to a close in Section 5.

2. Related Work

Deep learning algorithms have previously been applied to identify and categorize IDC tumor regions. Cruz-Roa et al. [8] using histopathological slides from 162 women with IDC, a three-layer CNN was constructed., gave an F-score of 71.80% and a balanced accuracy of 84.23%. Brancati et al. [9] proposed a FusionNet convolutional autoencoder-based IDC classification approach that achieved an F-score of 81.54% and a balanced accuracy of 87.76%. The model was built by Romano et al. [9] using a convolutional neural network (CNN). Parts of the design include completely linked layers, accept-reject pooling layers, convolutional layers, and dropout layers. With an F-score of 85.28% and an accuracy of 85.41%, the design was successful. Janowczyk et al. [10] successfully identified IDC tumors with an F-score of 76.48% and a

balanced accuracy of 84.68% utilizing an AlexNet-based technique. In their study, Romero et al. [11] reported the performance metrics of their multilevel batch normalization scheme, which was implemented using the Inception architecture. They attained an F-score of 89.70% and a balanced accuracy of 89.00%.

IDC tumors were categorized into several classes by Celik et al. [12] using pre-trained algorithms ResNet-50 and DenseNet-161. The F-score of the DenseNet-161 model was 92.38%, while the balanced accuracy of the ResNet-50 model was 90.96%, with an F-score of 94.11%. An awful lot of work has gone into making the process of scoring BC automatic. In their paper, Khan et al. [13] suggested a way to achieve nuclear atypia that uses both an image description and a GkNN-based classifier. A strategy for detecting BC and grading prostate and BC (high vs. low grade) was put out by Naik et al. [14]. In addition to template matching, nucleus segmentation, and architectural features, their method incorporates a Bayesian classifier. They were able to diagnose BC with an accuracy of 81.91 percent and grade it with an accuracy of 80.52 percent. To distinguish between low, middle, and high grades of BC.

Tao et al. [15] developed an ensemble strategy that led to a 69.00% overall accuracy. To achieve such accuracy, they used SVM classifiers to extract semantic, object, and pixel information from photos to differentiate between low- and high-grade BC. Doyle et al. [16] created a novel strategy involving spectral clustering and an SVM classifier to reduce the dimensionality of the feature set. The technique has been applied to differentiate between images with and without cancer and between images with and without high- and low-grade BC, with accuracy of 95.80% and 93.30%, respectively. Wetstein et al. [17] introduced a deep learning approach for grading ductal carcinoma in situ (DCIS) built upon DenseNet-121. They utilized the quadratic weighted Cohen's kappa to assess the inter-observer agreement between their developed system and three expert observers at the lesion and patient levels.

Despite the considerable amount of research dedicated to automating the grading process of BC, there is a scarcity of established works that specifically grade IDC BC. Dimitropoulos et al. [5] suggested a way to grade IDC. Grassmannian manifolds and VLAD encoding were implemented on histology images obtained from the "Agios Pavlos" General Hospital in Thessaloniki, Greece, affiliated with the Department of Pathology. Their approach was successful 95.80% of the time. They were utilizing the same dataset. Li et al. [18] suggested an approach that relies on a classification and grading method for IDC cancer predicated on the Xception model with an approximate 94.54% accuracy rate. Senousy et al. [19] employed the same "Agios Pavlos" dataset to create an entropy-based elastic ensemble method for IDC grading, which is based on deep learning models (3E-Net). They suggested two iterations of the procedure, yielding results with corresponding accuracy percentages of 96.15% and 99.50%. Combining the datasets BreakHis [20] and Agios Pavlos [21], Abdelli et al. [22] were able to detect cancer and grade photos. This was accomplished by supplementing the first three grades, IDC grades, with a fourth grade, grade 0, which signifies BC. Using ResNet-50 and MobileNet that had already been trained, they achieved a maximum combined accuracy of 97.03% and a total accuracy of 93.48% on the Agios Pavlos dataset, respectively. For their training, these networks relied on the ResNet algorithm. A combination of ResNet-50 and VGG-16 trained with different activation functions was suggested by Maguolo et al. [23]. They conducted experiments on over ten distinct datasets, with the Agios Pavlos dataset [21] being the most used for IDC grading, yielding an accuracy rate of 95.33%.

3. Materials and Methods

First, we will use data from [21, 23] and then six well-known Transfer Learning methods: DenseNet-121, DenseNet201, VGG19, ResNet-101v2, ResNet-50, and InceptionReNetV2. To ensure that the system is adequately trained, we divided the dataset into three equal parts and used 70% of it to train the models,

10% to validate them, and 20% to put them to the test. Evaluation criteria such as loss, accuracy, precision, area under the curve (AUC), and recall were then used to assess how well each TL model performed. The diagram in Figure 1 shows how the whole system works.

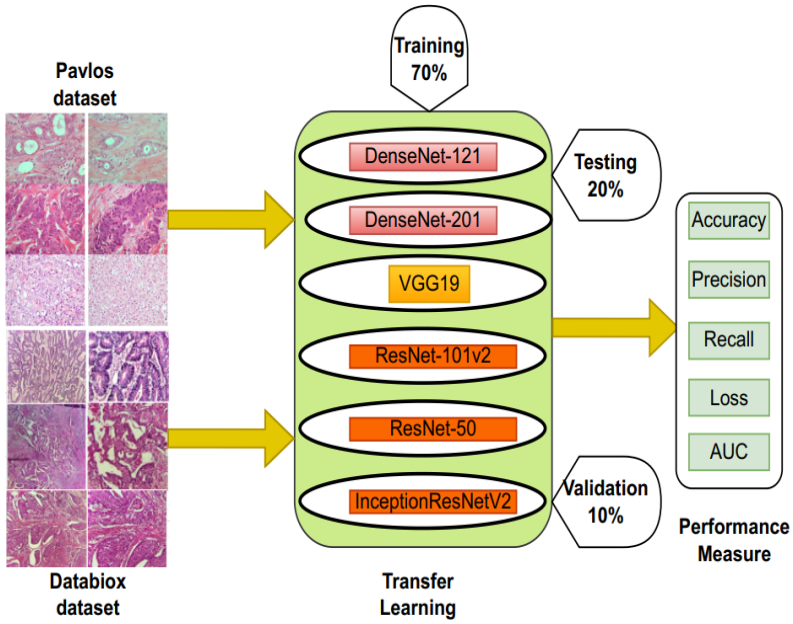

Figure 1. Proposed network architecture.

3.1 Datasets

We have considered the Agios Pavlos and Databiox datasets for the current work. A histopathology microscope image from the Databiox and Agios Pavlos datasets. This dataset can be found at http://databiox.com and https://www.kaggle.com/pavlostsoukias. To address the imbalance in the data, data augmentation (DA) is used. After that, the AutoML model [24], created by Google and hosted on the cloud platform, is used. It uses 70% of the data for training, 10% for validation, 20% for testing, and some for the held-out test. By the current state of the art, AutoML demonstrates a greater average accuracy [24].

Databiox dataset: To validate our models' stability, we used the Databiox dataset to evaluate our proposed model, which consists of 922 images associated with 124 individuals with IDC. This dataset diverges from the comparable ones in that it includes an equal number of cases from each of the three IDC evaluations, each generating approximately 50 examples. There are four levels of amplification provided for each example: 4×, 10×, 20×, and 40×. Depending on the pathologist's evaluation, multiple images from a specific amplification level are occasionally introduced. To give an example, almost all of the instances have four 40× photos. We looked at the 40x enlarged images for this project. Seeing small details in an example test with a clear objective lens is easy. A dominant objective lens connected to a 10× eyepiece may magnify an object up to 400×, providing a prominent image of the specimen on the slide [25]. Three malignant cell spotlights are considered, and a score is given to each. After that, the scores are adjoined to produce a number between 3 and 9, which is then used to determine a grade of 1, 2, or 3, indicated on the pathology report.

Pavlos dataset: We used the Agios Pavlos dataset [26] to test our proposed model and ensure reliability. The collection, which is 1280 × 960 pixels, includes 300 photos of 21 IDC-infected individuals in grades 1, 2, and 3. Eight patient slides are used for inference, and thirteen are set aside for training. Before training and inference, every image is cropped to 250 × 250 pixels. Fig. 6 presents the sample photographs. During training, the following augmentations are applied: shear (0.16◦), width shift (interval - [− 0.6, +0.6)), rotation (40◦), horizontal and vertical flip, and zoom (range - [0.84,1.16]).

The augmentations utilized here are less extensive than those used in the Databiox dataset. The most effective use of these Agios Pavlos augmentation combinations has been observed in practice. The results might have been better because the additions reflected or covered the whole test set better. The ensemble grade model got 89.26% of the test sets right after the training. This proves that our model is capable of making precise grade predictions. The 40x enlarged pictures in Agios Pavlos were much more accurate than the 40x enlarged images in the Databiox datasetThe model may have perceived the Databiox photographs as more intricate and challenging to analyze compared to the Agios Pavlos ones.

3.2 Transfer Learning

TL is the process of getting better at learning something else by sharing information from a related task that you have already mastered. By gaining access to data from the source, TL aims to improvise learning in the goal function. There are primarily three fundamental estimates that move might use to improve learning. Before additional education, the entire show can be done in the objective function by merely employing the transferred data. This appears differently from the hidden show of a careless, trained professional. The second factor is how long it takes to become accustomed to the objective function, considering that the shifted data appeared differently in time to consume it unplanned. The third is the final exhibition level attained in the target task, as opposed to the last level without movement [27, 28]. In DL, TL has grown to be a sizable subfield. It has practical benefits since it can increase the productivity of DL and philosophical benefits since it is an essential component of human learning. Data transfers may become increasingly appealing with increasing registering power and experts applying deep learning to increasingly complex problems. The main benefits of transfer learning are resource conservation and increased efficacy when new models are trained. Since most models will be pre-trained, it can also help with model-training situations when only unlabeled datasets are available [29].

3.3 DenseNet121

The next step in deepening CNN's profundity is called DenseNets [30]. DenseNets leverages the network's potential by reusing features instead of deriving genuine power from extraordinarily expansive or profound designs. DenseNets require fewer parameters than an equivalent conventional CNN because learning repeated feature maps is not a compelling use case. Several limitations (such as 12 filters) and a few more feature maps were added by the DenseNets layer [31]. The 33 MB DenseNet121 has a 121-level depth, 7,188,035 parameters (150,531 trainable and 7,037,504 nontrainable), and Top 1 accuracy of 0.750 (out of a possible 1.00) and Top 5 accuracy of 0.923 (out of 1.00) [32]. Figure 2 (a) illustrates the model. The number of filters varies amongst DenseBlocks, which raises the channel's element count. The growth rate (G) is used to summarize the m-th layer. It determines the amount of data to be put into each layer, as mentioned in (1).

$$G_m = G_o + G \times \mathrm{G(m\text{-}1)} \tag{1}$$

3.4 Densenet201

Once more, the DenseNet group's [30] output of the model's representation to do photo classification is the Densenet201model. The size and precision of the densenet121 model are the main differences. The designers converted them from their original Torch preparation to a Caffe* design. Using the ImageNet image dataset, each DenseNet model has been pre-prepared [45]. Figure 9 presents the insight of the model. DenseNet201, with a size of 80 MB, has 18,604,227 parameters, of which 282,243 are trainable, and 18,321,984 are nontrainable. Its accuracy in the top five ranks is 0.936, and its Top 1 accuracy is 0.773 [32].

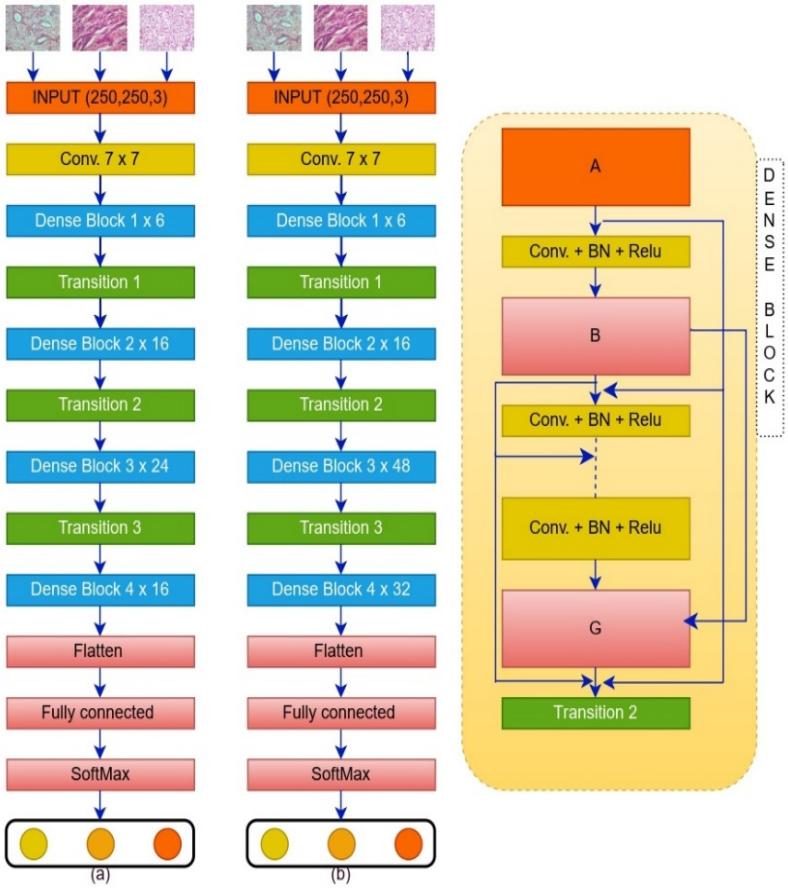

Figure 2. Model block diagrams used for IDC grading: (a) DenseNet-121; (b) DenseNet-201.

3.5 VGG19

The VGG19 [33] model contains of 19 layers, convolution layers is 16, 3 of which are entirely connected, 5 of which are MaxPool layers, and 1 of which is a SoftMax layer. For VGG19, the number of FLOPS is 19.6 billion [34]. At a depth of 23, VGG19's trainable parameters are 75,267 and non-trainable are 20,024,384 parameters provide an accuracy of 0.713 at the top level and 0.900 at the top 5 [32]. Figure 3 depicts the model. Equation (2) introduces the down-sampling layer.

$$Y_{q_i}^{(m)} = f(\gamma_i^m down(Y_i^{(m-1)}) + a_i^{(m)} \quad (2)$$

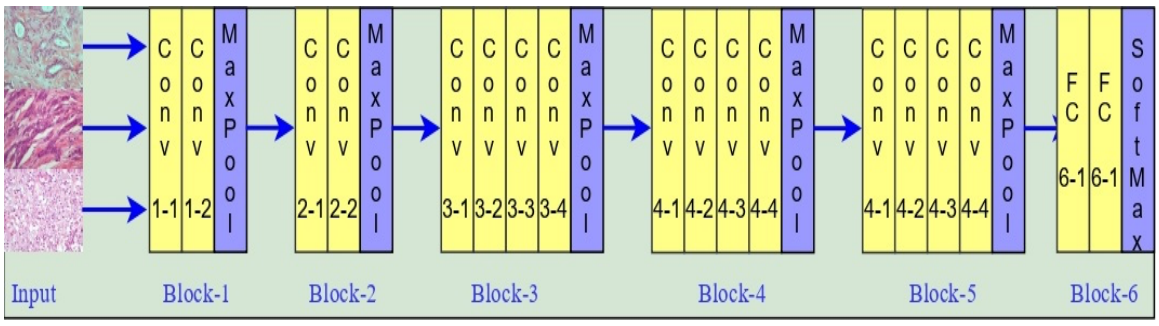

Figure 3. Block diagrams of model VGG19 used for the grading of IDC.

3.6 InceptionResNetV2

Recently, the utilization of deep convolutional neural networks (CNNs) has emerged as a crucial factor in enhancing the performance of image recognition tasks. The efficacy of the inception model was proved in achieving high performance while using relatively lower computer resources. The progressive version of inception and residual modules, known as Inception-v4, is the high-level version of Inception-

v3. Batch normalization is applied on top of traditional convolutional layers in inception-v4. The inception block size has increased due to these possessions [35]. The computational cost of Inception-ResNet v2 and Inception v4 is the same. However, their stems differ [36]. InceptionResNetV2 has a 215 MB size, Top 1 accuracy of 0.803, Top 5 accuracy of 0.953, 54,451,939 parameters (115,203 trainable and 54,336,736 not), and 572 depth [32]. ResNet uses a linear projection to copy the identity mapping and increase the number of shortcut channels that match the residual. Figure 4 shows that this agrees to have the inputs c and F(c) mixed as inputs to the next layer. We used F (c) and c when they had different sizes, like 30 x 30 and 35 x 35. The model gets more parameters when this Pi term is used with 1 × 1 convolutions.

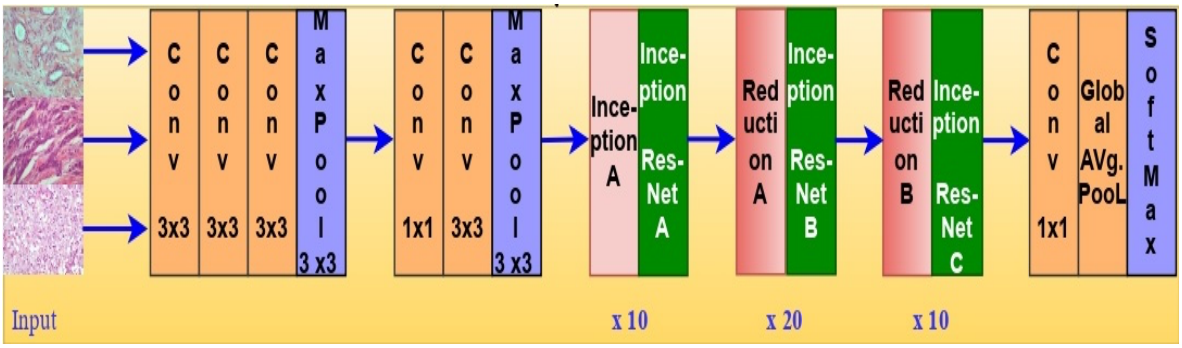

Figure 4. Block diagrams of the InceptionResNetV2 model used for the grading of IDC.

3.7 ResNet-50 and ResNet-101v2

The ResNet, a deep residual network, was initially presented in 2015 by academics affiliated with Microsoft Research. The idea behind this design was to address the problem of disappearing or exploding slopes. Our experiments were conducted using the ResNet-101v2 and ResNet-50 designs [37]. After grading IDC with DenseNet-121, DenseNet-201, ResNet-101v2, and ResNet-50, TTA is used. To train WSI pictures, which are intractable and demand a significant amount of processing effort, the images are first cropped by removing as much of the black background as possible. After the photos have been cropped, they are converted into 250-by-250-pixel patches that do not overlap. The training package contains photographs of various magnifications for instruction. The dataset is partitioned into three discrete sets to mitigate the risk of data leakage: a training set, two validation sets (val1 and val2), and a third set comprising unique patient transparencies. The training dataset includes 48,220 image patches that represent various magnification levels (4×, 10×, 20×, and 40×). Additionally, image patches representing 12 patients are present in val1 and val2. After the models have been trained using the training set, they are independently inferred using the two validation sets for each degree of magnification.

ImageNet-trained models that have already been used for training are used. DenseNet-121's first 88 layers, DenseNet-201's first 39 levels, ResNet-50's first 43 layers, and ResNet-101v2's first 40 layers are all frozen throughout training. A dense SoftMax-activated layer in the output layer simplifies image classification into many categories. DenseNet-121 is the only model with dropout and flattening layers. After the Global Average Pooling (GAP) layer, these strata precede the thick layer. These supplementary layers are not included in any of the other models. All models, with the exception of DenseNet-201, are equipped with the categorical cross-entropy loss function during the training process. In order to train the DenseNet-201 model, the Kullback-Leibler divergence loss function is utilized [39]. Illustrations of the various models are presented in the form of block diagrams in Figure 5.

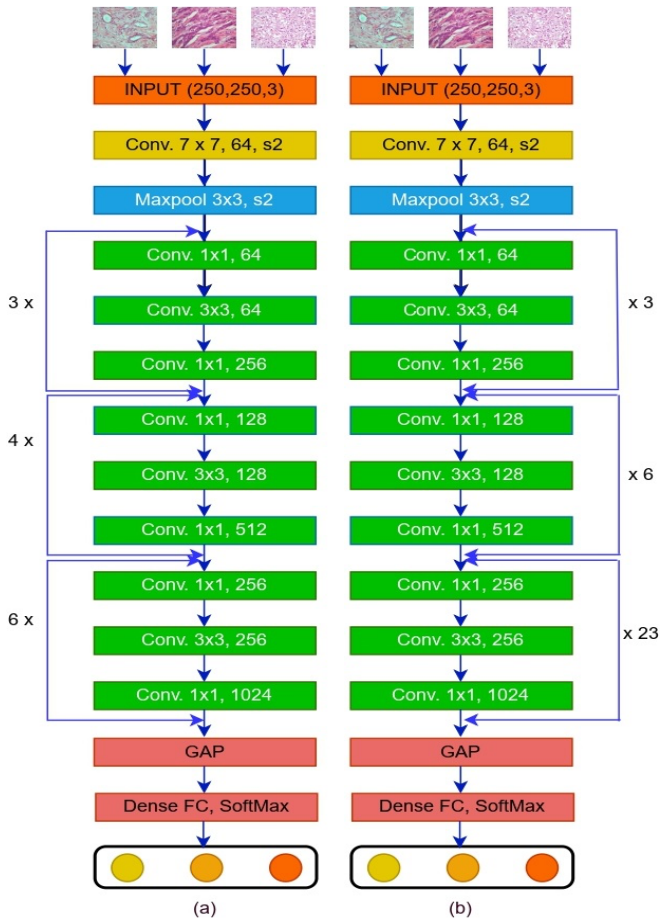

Figure 5. Model block diagrams used for IDC grading: (a) ResNet-50; (b) ResNet-101v2.

During the training process, the images in the training dataset are rescaled from 0 to 1, and the dataset is substantially improved. To improve the quality of the images, the following operations are performed on the remaining 50%: horizontal flipping of 50%, vertical flipping of 20%, sharpening, embossing, darkening, and lightening by multiplying each pixel by a random value between -5% and 50%, adding values between 10% and -10% of the pixels, and causing 1%–10% of the pixels to become black. In addition to these augmentations, the pictures undergo affine transformations, which include rotation (between - 10◦ and + 10◦) and translation (between - 25◦ and + 25◦). SGD, with a learning rate of 10− 6 and momentum of 0.9, is paired with CLR during training; the latter has a bottom bound of 10− 6 and an upper limit of 10− 4. Every batch consists of 32 elements, and each model is trained for 30 iterations. The probability that each picture patch belongs to each class are added to create DenseNet-121 and DenseNet-201, which are built together (ensembledensenets + DenseNet-201). TTA is then applied to these networks. With ResNet-101v2 and ResNet-50, the same outcome can be obtained by creating ensembles (ensembleresnets + TTA) and applying TTA. During TTA (Test Time Augmentation) for ensemble DenseNets, several enhancements are employed, such as vertical flipping and randomly blacking 2% to 20% of the pixels in the image [31]. Next, an ensemble model known as ensemble grade is created by combining the two ensemble models. The most excellent predicted probability of each picture patch for each class label is considered when determining the final output for the patch-level classification. Individual slides from various patients are viewed for computing the accuracy score when classifying patients at the patient level rather than using patches. The final class output is determined by selecting the class with the highest projected probability, calculated by counting the expected probabilities of each patch on every patient's slide. This class is the one with which the patient was ultimately classified.

4. Results

Implementation details: All testing are carried out on a Windows PC equipped with an Intel Core (TM) i7-7700 CPU @3.60 GHz, a one-terabyte hard disk (HDD), 32 gigabytes of random-access memory (RAM), and a CUDA-enabled Nvidia GTX 1050 4 gigabyte GPU. The programs are executed in Keras with the TensorFlow backend.

4.1. Evaluation metrics

Given below is a list of the evaluation metrics that were applied in this study:

$$Acuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Balanced\ \ Accuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{4}$$

$$F1 - score = \left(\frac{TP}{TP+\frac{1}{2}(FP+FN)}\right) \tag{5}$$

FN stands for False Negative, *FP* for False Positive, *TN* for True Negative, and *TP* for True Positive. The following formula is used to determine the accuracy of patient-level classification:

$$Acc_p = \frac{Slides\ predicted\ Correctly}{Total\ Number\ of\ Slides} \tag{6}$$

4.2. Results Discussion

The results of the Databiox test set for the individual are displayed in Table 2. The accuracy of the DenseNet-121 model was 97.13%, the balance accuracy was 96.25%, and the F-score was 96.38%. This was better than the DenseNet-201 model.

Table 1. Model accuracy utilizing Databiox dataset.

Model	4x	10x	20x	40x	Test set
DenseNet-121	71.31%	76.97%	63.05%	63.01%	val1
	63.94%	81.14%	77.62%	73.05%	val2
	67.09%	78.26%	69.99%	66.85%	Avg
DenseNet-201	64.18%	73.28%	63.66%	62.71%	val1
	60.37%	74.22%	73.47%	71.14%	val2
	62.28%	72.45%	67.57%	65.90%	Avg
VGG19	64.45%	71.41%	59.57%	63.44%	val1
	61.28%	76.35%	76.50%	62.82%	val2
	60.11%	75.12%	66.75%	63.53%	Avg
ResNet-50	62.42%	64.58%	62.55%	62.03%	val1
	63.37%	73.45%	63.97%	65.22%	val2
	64.28%	72.67%	64.97%	63.42%	Avg
ResNet-101v2	62.34%	56.68%	56.76%	56.38%	val1
	51.29%	72.61%	76.62%	67.40%	val2
	50.44%	73.49%	64.47%	65.39%	Avg
InceptionReNetV2	62.58%	74.38%	62.76%	58.71%	val1
	59.47%	76.52%	73.67%	75.54%	val2
	61.28%	75.65%	64.57%	66.30%	Avg

Table 2. IDC classification results.

Models	Acc. %	F1-score %	Balanced Acc.%
DenseNet-121	97.13	96.38	96.25
DenseNet-201	96.34	96.00	95.64
VGG19	95.65	96.62	94.56
ResNet-50	94.90	95.30	93.55
ResNet-101v2	93.53	95.23	93.00
InceptionReNetV2	93.20	94.90	92.65

Compared to the each models. This finding makes it clear that InceptionReNetV2 has 93.20 accuracy, 94.90 F1-score and 92.65 Balanced accuray while DenseNet had 97.13 accuracy, 96.38 F1-score and 96.25 Balanced accuray. Form above given Table 2 its clear that DenseNet-121 has thee more accuracy as compare to other pproposd models. So, we can say that DenseNet-121 haa good model for IDC classification for Databiox dataset.

Table 3. IDC results comparison of different methods on Databiox.

Method	F1-score (%)
Zavareh et al. [38]	72
Kumaraswamy et al. [39]	72
Sujatha et al. [29]	92.64
Talpur et al. [40]	92.81
Our (DenseNet-121)	96.38

4.3. IDC grading results

Table 1 presents the outcomes of the individual models applied to the test datasets. The DenseNet-121 model demonstrates the highest accuracy among the cohorts (val1 and val2). A third test set (Avg) is used to infer the models, and it comprises every single picture from the val1 and val2 test sets. The accuracy (Accp) of the DenseNet-121 for patient-level classification was 71.31% on val1 81.14% on val2, 77.62% on val3, and 73.05% on val4 for 4x, 10x, 20x, and 40x magnifications, respectively. Quantitative results show that the patient-level classification outperforms the patch-level classification. This might be the case because the model is more sure of its grade predictions for each patient's slides when it's doing patient-level classification instead of patch-level classification, which doesn't consider these things.

4.4. Evaluation of the proposed model on the "Agios Pavlos" dataset

We performed tests on the Agios Pavlos dataset [26] to validate the stability of our proposed model. There are 300 photos in the collection, representing 21 individuals with grades 1, 2, and 3 IDC infections. Thirteen patient slides are set aside for instruction, and eight are viewed for inference. Before training and inference, all photos are cropped to 250 250 px. Figure 6 displays several examples of the images.

Training enhancements include horizontal and vertical flips, rotation by 40 degrees, width shift by 0.6 units (interval: -0.6, +0.6), zoom by 0.84 to 1.16 times, and shear by 0.016 units. We employ less stringent augmentations compared to those discovered in the Databiox dataset. It has been empirically determined that the most favorable outcome was achieved when these combinations of augmentations were utilized for Agios Pavlos. One possible explanation for the improved performance is that the augmentations employed here better represent or generalize the test set. Our DenseNet-121 model could predict grades precisely after training, as it achieved an accuracy of 98.73% on the test set as shown in Table 4. Compared to the 40x photos in the Databiox dataset, the 40x photographs in Agios Pavlos obtained much greater

accuracy. One possible explanation is that the model considered analyzing photos from Databiox significantly more challenging than analyzing photographs from Agios Pavlos.

Table 4. Performance metrics of IDC classification on Agios Pavlos dataset.

Method	Acc.%	F1-score %	Balanced Acc.%
DenseNet-121	98.73	98.28	98.45
DenseNet-201	97.74	97.00	96.24
VGG19	93.25	93.42	93.15
ResNet-50	93.98	93.40	93.65
ResNet-101v2	95.63	95.63	95.40
InceptionReNetV2	91.30	91.53	91.35

The Agios Pavlos dataset was utilized in this investigation; it has been used in other studies to investigate the grading of IDC BC. Convolutional neural networks (CNNs) were used to grade IDC in all of the other publications [8, 10, 41]. We investigated the efficacy of our algorithm in classifying IDC BC using the Agios Pavlos dataset. We divided the dataset slide-wise for 5-fold cross-validation instead of subject-wise, as we did for the holdout test set, which allowed for a more equitable comparison. Every fold is trained for 400 iterations while maintaining the same augmentations and hyperparameters as the holdout test set technique. Our model performed well, with a standard deviation of 0.04 and a mean accuracy of 98.28%.

Table 5. IDC results comparison of different methods on Agios Pavlos dataset.

Method	F1-score (%)	Balanced Accuracy (%)
Cruz-Roa et al. [8]	71.80	84.23
Janowczyk et al. [10]	76.48	84.68
Reza et al. [41]	84.78	85.48
Romero et al. [11]	89.70	89.00
Celik et al. [12]	92.38	91.57
Celik et al. [12]	94.11	90.96
Our (DenseNet-121)	**98.28**	**98.45**

Table 5 compares our technique's performance with the previous state-of-the-art. As can be observed, our method exceeds the state-of-the-art, achieving a 8.46% better-balanced accuracy and a 4.17% higher F-score.

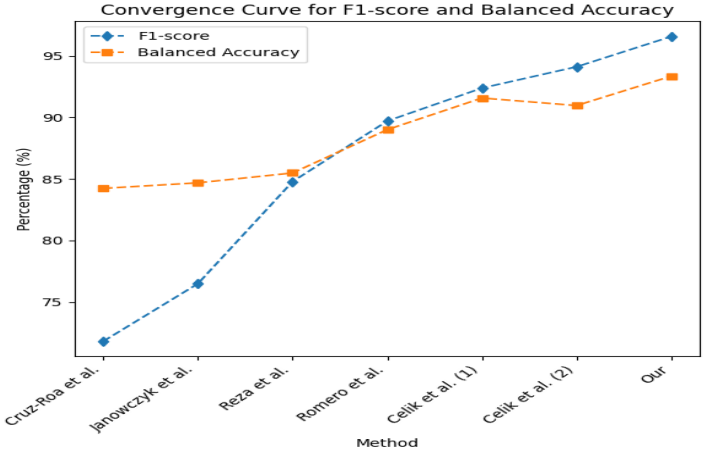

Figure 6. IDC results comparison convergence curve.

5. Conclusion

We have conducted a comprehensive comparative analysis of IDC classifications using different pre-trained deep-learning models. Our findings show that DenseNets have outperformed other models. Additionally, we have tested multiple preprocessing techniques and found that brightening the images resulted in the best performance. Out of the previous state-of-the-art results, the best performance was achieved by an ensemble of DenseNet-121 and DenseNet-201. An ensemble model for IDC grading has also been constructed and trained using two distinct datasets of enlarged images. Additionally, we have explored many approaches to picture patch extraction and worked with numerous pre-trained classifiers from deep learning. The ensemble model is further tested on the "Agios Pavlos" and Databiox datasets with a holdout test set and a 5-fold cross-validation set. A comparison with related literature has also been carried out for the 5-fold cross-validation set. Pathologists can utilize the models to aid in illness diagnosis and grading with more efficiency and accuracy after additional testing on various or more extensive datasets. In the future, we plan to make the data we develop publicly available for other researchers to evaluate in repositories. Experts suggest combining radiologists' findings with imaging data for BC classification is the future scope of this study. To enhance the performance of model, we recommend utilizing optimization-based algorithms.

References

1. Cancer Facts & Figures," Amercian Cancer Society, pp. 80, 2022.
2. F. Mehmood, E. Chen, M. A. Akbar, and A. A. Alsanad, "Human action recognition of spatiotemporal parameters for skeleton sequences using MTLN feature learning framework," Electronics, vol. 10, no. 21, pp. 2708, 2021.
3. M. J. Lakshmi, and S. Nagaraja Rao, "Brain tumor magnetic resonance image classification: a deep learning approach," Soft Computing, vol. 26, no. 13, pp. 6245-6253, 2022.
4. B. H. Segal, T. Giridharan, S. Suzuki, A. N. H. Khan, E. Zsiros, T. R. Emmons, M. B. Yaffe, A. A. Gankema, M. Hoogeboom, and I. Goetschalckx, "Neutrophil interactions with T cells, platelets, endothelial cells, and of course tumor cells," Immunological Reviews, vol. 314, no. 1, pp. 13-35, 2023.
5. K. Dimitropoulos, P. Barmpoutis, C. Zioga, A. Kamas, K. Patsiaoura, and N. Grammalidis, "Grading of invasive breast carcinoma through Grassmannian VLAD encoding," PloS one, vol. 12, no. 9, pp. e0185110, 2017.
6. P. Robbins, S. Pinder, N. De Klerk, H. Dawkins, J. Harvey, G. Sterrett, I. Ellis, and C. Elston, "Histological grading of breast carcinomas: a study of interobserver agreement," Human pathology, vol. 26, no. 8, pp. 873-879, 1995.
7. F. Mehmood, E. Chen, T. Abbas, M. A. Akbar, and A. A. Khan, "Automatically human action recognition (HAR) with view variation from skeleton means of adaptive transformer network," Soft Computing, pp. 1-20, 2023.
8. A. Cruz-Roa, A. Basavanhally, F. González, H. Gilmore, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, and A. Madabhushi, "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks." p. 904103.
9. N. Brancati, G. De Pietro, M. Frucci, and D. Riccio, "A deep learning approach for breast invasive ductal carcinoma detection and lymphoma multi-classification in histological images," IEEE Access, vol. 7, pp. 44709-44720, 2019.
10. A. Janowczyk, and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," Journal of pathology informatics, vol. 7, no. 1, pp. 29, 2016.
11. F. P. Romero, A. Tang, and S. Kadoury, "Multi-level batch normalization in deep networks for invasive ductal carcinoma cell discrimination in histopathology images." pp. 1092-1095.
12. Y. Celik, M. Talo, O. Yildirim, M. Karabatak, and U. R. Acharya, "Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images," Pattern Recognition Letters, vol. 133, pp. 232-239, 2020.
13. A. M. Khan, K. Sirinukunwattana, and N. Rajpoot, "A global covariance descriptor for nuclear atypia scoring in breast histopathology images," IEEE journal of biomedical and health informatics, vol. 19, no. 5, pp. 1637-1647, 2015.
14. S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology." pp. 284-287.
15. T. Wan, J. Cao, J. Chen, and Z. Qin, "Automated grading of breast cancer histopathology using cascaded ensemble with combination of multi-level image features," Neurocomputing, vol. 229, pp. 34-44, 2017.
16. S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features." pp. 496-499.
17. S. C. Wetstein, N. Stathonikos, J. P. Pluim, Y. J. Heng, N. D. Ter Hoeve, C. P. Vreuls, P. J. van Diest, and M. Veta, "Deep learning-based grading of ductal carcinoma in situ in breast histopathology images," Laboratory Investigation, vol. 101, no. 4, pp. 525-533, 2021.
18. L. Li, X. Pan, H. Yang, Z. Liu, Y. He, Z. Li, Y. Fan, Z. Cao, and L. Zhang, "Multi-task deep learning for fine-grained classification and grading in breast cancer histopathological images," Multimedia Tools and Applications, vol. 79, pp. 14509-14528, 2020.
19. Z. Senousy, M. M. Abdelsamea, M. M. Mohamed, and M. M. Gaber, "3E-Net: Entropy-based elastic ensemble of deep convolutional neural networks for grading of invasive breast carcinoma histopathological microscopic images," Entropy, vol. 23, no. 5, pp. 620, 2021.
20. F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," Ieee transactions on biomedical engineering, vol. 63, no. 7, pp. 1455-1462, 2015.
21. "Breast carcinoma histological images from the department of pathology, "agiospavlos" general hospital of thessaloniki, Greece," 2021.
22. A. Abdelli, R. Saouli, K. Djemal, and I. Youkana, "Combined datasets for breast cancer grading based on multi-cnn architectures." pp. 1-7.
23. H. Bolhasani, E. Amjadi, M. Tabatabaeian, and S. J. Jassbi, "A histopathological image dataset for grading breast invasive ductal carcinomas," Informatics in Medicine Unlocked, vol. 19, pp. 100341, 2020.
24. N. Noshiri, M. Khorramfar, and T. Halabi, "Machine Learning-as-a-Service Performance Evaluation on Multi-class Datasets." pp. 332-336.
25. M. Tan, and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks." pp. 6105-6114.
26. C. Zioga, A. Kamas, K. Patsiaoura, K. Dimitropoulos, P. Barmpoutis, and N. Grammalidis, "Breast carcinoma histological images from the department of pathology,"agios pavlos" general hospital of thessaloniki," Greece, July, 2017.
27. L. Torrey, and J. Shavlik, "Transfer learning," Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, pp. 242-264: IGI global, 2010.
28. N. Sideris, P. Dama, S. Bayraktar, T. Stiff, and L. Castellano, "LncRNAs in breast cancer: a link to future approaches," Cancer Gene Therapy, vol. 29, no. 12, pp. 1866-1877, 2022.

29. R. Sujatha, J. M. Chatterjee, A. Angelopoulou, E. Kapetanios, P. N. Srinivasu, and D. J. Hemanth, "A transfer learning-based system for grading breast invasive ductal carcinoma," IET Image Processing, vol. 17, no. 7, pp. 1979-1990, 2023.
30. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." pp. 4700-4708.
31. P. Ruiz, ""Understanding and visualizing DenseNets-towards data science," Medium, 2018.
32. K. Team, "Keras documentation: Keras applications," Inglés URL https://www. kerasio/api/applications, 2020.
33. K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
34. A. Kaushik, "Understanding resnet50 architecture," OpenGenus Foundation. Retrieved from: https://iq. opengenus. org/resnet50-architecture, 2020.
35. R. A. Aral, Ş. R. Keskin, M. Kaya, and M. Hacıömeroğlu, "Classification of trashnet dataset based on deep learning models." pp. 2058-2062.
36. B. Raj, "A simple guide to the versions of the inception network. Medium," Medium, 2020.
37. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition." pp. 770-778.
38. P. H. Zavareh, A. Safayari, and H. Bolhasani, "BCNet: A deep convolutional neural network for breast cancer grading," arXiv preprint arXiv:2107.05037, 2021.
39. E. Kumaraswamy, S. Sharma, and S. Kumar, "Invasive Ductal Carcinoma Grade Classification in Histopathological Images using Transfer Learning Approach." pp. 1-6.
40. S. Talpur, "Automatic Detection System to Identify Invasive Ductal Carcinoma by Predicting Bloom Richardson Grading from Histopathological Images," Journal of Independent Studies and Research Computing, vol. 20, no. 1, 2022.
41. M. S. Reza, and J. Ma, "Imbalanced histopathological breast cancer image classification with convolutional neural network." pp. 619-624.