Biomedical Signal Processing and Control 86 (2023) 105118

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Promoting accuracy in low-magnification histopathology grading: With augmentation and multi-dilation model

Zonghan Gan [a], Abdulhamit Subasi [b,c]

[a] Center for the Cellular Microenvironment, Department of Biomedical Engineering, University of Glasgow, Glasgow G11 6EW, United Kingdom
[b] Institute of Biomedicine, Faculty of Medicine, University of Turku, Turku 20520, Finland
[c] Department of Computer Science, EFFAT University, Jeddah 21478, Saudi Arabia

ARTICLE INFO

Keywords:
Whole slide images
Patch-based grading
Image augmentation
Deep learning
CovXNet

ABSTRACT

Advances in artificial intelligence have facilitated the automated grading of histopathology slides. Yet, the magnification of whole slide scanners (WSS) has restrained the accuracy of patch-based grading. In this work, we found that augmentation can significantly promote grading performance under this issue, even when the data volume is large (>140 K). With augmentation and a multi-dilation model, the CovXNet, we yielded a Balanced Accuracy of 92.13%, which is the current highest for the Breast Histopathology Dataset (40X magnification) also the first time both sensitivity and specificity >90%. However, in this focused grading task, augmentation only improves models with high invariance (the CovXNet and BCA-CNN). Pre-trained ResNet has lower invariance in this task, but fine-tuning can significantly improve both accuracy and invariance. For the CropNet attention model, adapting with max pooling but not augmentation offers promotions. Additionally, this work also found two types of common errors in high-starred codes, when using random.shuffle for data-label composited array, or the integrated shuffle function of ImageDataGenerator, which fake a higher accuracy by masking class 0 as class 1. Using Sklearn.shuffle instead is safer. All codes are available on our GitHub.

1. Introduction

Histopathology grading serves as a gold standard in tumor research and clinical diagnosis [2–4]. Established in the last century, the grading system translates histopathological data into semi-quantitative assessments [5] to standardize treatment, ideally offering reproducibility, definability, and interpretability [6]. However, manual grading is heavily reliant on the analysts' experience, resulting in challenges such as low user agreement, subjective assessments, inter-pathologist variability [7], and considerable time consumption. Over the past decade, rapid advancements in deep learning and Computer-Assisted Diagnosis (CAD), coupled with the widespread adoption of Whole Slide Scanners (WSS), have spurred efforts to automate histopathology grading, delivering benefits such as increased speed, accuracy, and normalization [7–10].

CAD-based grading is predominantly achieved through patch-based grading algorithms [9,11], where Whole Slide Images (WSI) are divided into small "patches" accompanied by metadata (Patient ID, Slide ID, location coordinates). Each patch is individually classified before being "reconstructed" to the original slides, based on the metadata. Subsequently, the area percentage of each class within the reconstructed slide is calculated to determine the grading score. The reliability of CAD grading results hinges on the classification accuracy of individual patches.

1.1. Motivation

The WSI's rate of magnification represents the WSS's resolution and directly decides information density in patches, thereby playing a critical role in determining the classification accuracy of individual patches (without specification, "accuracy" in this paper refers to the classification accuracy of single patches). At low magnification (40×), fewer details of individual cells are preserved, resulting in lower accuracy compared to higher magnifications [12–15]. Yet, not all histopathology laboratories worldwide have access to high-magnification WSSs. Moreover, discarding low-magnification (40×) WSIs would waste existing data, and higher magnification WSIs could increase computational demands [16,17], hindering real-time classification on local endpoints. Consequently, we propose a novel classification approach for patch-based grading of low-magnification WSIs, utilizing a combination of data augmentation and the multi-dilation CovXNet model. Our primary goal is to achieve higher and more balanced classification accuracy for patches derived from low-magnification WSIs. We focus on the widely

E-mail addresses: abdulhamit.subasi@utu.fi, absubasi@effatuniversity.edu.sa (A. Subasi).

https://doi.org/10.1016/j.bspc.2023.105118
Received 15 March 2023; Received in revised form 25 May 2023; Accepted 8 June 2023
Available online 20 June 2023

Nomenclature

Acronyms	*Definitions*
ANN	Artificial Neural Network
AUC	Area Under Receiver Operating Characteristic Curve
BHI	The Breast Histopathology Images dataset
CAD	Computer-Assisted Diagnosis
CNN	Convolutional Neural Network
COVID	Coronavirus Disease
GAN	Generative Adversarial Network
ID	Identity
IDC	Invasive Ductal Carcinoma
MS-DenseNet	Mobile Sequenze-Excitation DenseNet
NaN	Not a Number value
PPI	Pixels Per Inch
RAM	Random-Access Memory
ROC	Receiver Operating Characteristic Curve
SE	Squeeze-and-Excitation
WSI	Whole Slide Image
WSS	Whole Slide Scanner

researched 40X Breast Histopathology Images (BHI) dataset. Table 1 presents the accuracy of previous BHI studies and their applied methods.

As evident in Table 1, current methods have not achieved >90% classification accuracy in single patch recognition, with the exception of Celik's study [18]. Furthermore, these studies were unable to simultaneously attain high (>90%) sensitivity and specificity. The sensitivity and specificity largely diverged in most studies, with some demonstrating a delta of >10% [19–22]. Our aim is to advance patch classification to a level where accuracy, sensitivity, and specificity all exceed 90%, thereby enhancing the patch-based grading of whole slides.

Table 1
Comparison of previous attempted methods for identification among the BHI dataset.

Study	Year	Method	Sensitivity	Specificity	Testing Balanced Accuracy (BAC)
Cruz-Roa et al. [23]	2014	3-layer CNN	–	–	84.23%
Reza et al. [19]	2018	CNN + oversampling	80.85%	90.12%	85.48%
Romano et al. [24]	2019	InceptionNet	84%	86%	85.41%
Romero et al. [25]	2019	InceptionNet + multi-level batch normalization	90%	87%	89.00%
Alghodhaifi et al. [21]	2019	IDCNet	93.44%	71.14%	87.13%
Johnson [22]	2020	Conditional GANs	72.41%	94.66%	83.54%
Celik et al. [18]	2020	DenseNet161	89.59%	93.56%	**91.57%**
R. Singh et al. [20]	2021	VGG19 multiple input	93.31%	82.75%	88.03%
S. Singh et al. [14]	2022	ResNet + InceptionNet	82.01%	88.41%	85.21%
The proposed Method	**2023**	**CovXNet + augmentation**	**91.68%**	**92.58%**	**92.13%**

1.2. Proposed approach

Although not yet fully satisfactory, the classification of the low-magnification BHI has improved since the initial study in 2014 [23]. However, most previous research efforts have primarily focused on manipulating machine learning models, including multi-level batch normalization [25], multiple inputs [20], combined model [14], and new models [18,21]. Very few studies have concentrated on data preprocessing, such as oversampling strategies [19]. Therefore, this study aims first to enhance the learnability of data through preprocessing, and then to employ a specially-designed robust model, the CovXNet, to achieve optimal performance. The CovXNet [1] is a model designed for COVID-19 X-ray analysis, garnering increased attention for its multi-dilation design, which captures features at multiple observation levels. This aligns with the nature of histopathology, where information exists at various scales, ranging from tissue to sub-cellular structures.

Data augmentation is a cornerstone technique in machine learning. Despite its "age", new research continues to sprout [26–28], particularly in image classification [29,30]. While augmentation has traditionally been employed to address insufficient data volume [31–34], recent studies suggest that its potential extends beyond this job [35–38]. Investigations into the mechanisms of augmentation have revealed that it enhances data learnability through multiple aspects, including generalization capacity, regression optimization, and feature purification [33,39–43]. This work is the first to demonstrate that augmentation alone can significantly improve low-magnification histopathology grading. After identifying the best-performing model with augmentation, we also explored performance differences resulting from various model structures and training methods to elucidate the structural influence on augmentation's effects. The models evaluated covers a traditional CNN model (the Breast Cancer Analyzer CNN (BCA-CNN)) [44], location-wise and channel-wise Squeeze-and-Excitation attention models (the CropNet [45] and the MS-DenseNet [46]), and pre-trained transfer learning models (ResNet 50, ResNet 101, DenseNet 121, and DenseNet 169). Each model was selected for its unique structural features.

2. Contribution

Main contributions of this work are given as follows:

- This work demonstrates that augmentation can significantly improve low-magnification patch-based grading with certain models, including the CovXNet and BCA-CNN. The CovXNet model achieved a single patch classification accuracy of 92.13%, which is the highest among existing literature for the 40X BHI dataset. Additionally, this is the first time that both sensitivity and specificity have surpassed 90%. This study is also the first to employ the multi-dilation CovXNet for histopathology.
- This work elucidated that the improvement a model receives from augmentation depends on the model's structure and invariance. In this task, Complex models like ResNet (with ANN top, pre-trained top-tuning) demonstrated less invariance and saw a significant decline in performance with augmentation. For the tested attention model, adaptation with max pooling performed better than augmentation.
- Two types of common errors in shuffling were identified in this work. Using random.shuffle on a composite array of data and labels or incorporating shuffling in ImageDataGenerator can cause positive labels to be masked as negative, resulting in artificially inflated accuracy. Using sklearn.shuffle to co-shuffle data and labels in two separate arrays is proven to be a safer approach.

3. Material and methods

To overcome the constraint of classification accuracy in low-

magnification WSI grading, augmentation and other regulation methods were explored in several typical models, including the multi-dilation CovXNet, the Breast Cancer Analyzer CNN (BCA-CNN) [44], attention models (the CropNet [45] and MS-DenseNet [46]), and pre-trained transfer-learning models (ResNet 50, ResNet 101, DenseNet 121, and DenseNet 169). Each model was selected for its unique structural design.

3.1. Dataset

The Breast Histopathology Images (BHI) dataset was exploited in this study. Breast Cancer is one of the most fatal malignancies and the most common cancer in females [47,48]. In the BHI dataset, slides are labeled with invasive ductal carcinoma (IDC), which directly reflects metastasis. Due to these factors, the BHI dataset is widely used in CAD grading research [14,18–25,49].

The BHI dataset contains 277,524 patches of 50x50 size, including 78,786 IDC positive (+, or class 1) and 198,738 IDC negative (-, or class 0). These patches were extracted from 162 WSIs using a 40X WSS. Each patch's filename contains metadata for reconstructing. For example, the filename "12749_idx5_x2051_y851_class1.png" refers to the patient/slide ID (12749_idx5), the coordinates on its original slide (x2051 and y851), and the class (class 1, i.e., IDC+). Notably, the BHI dataset was contributed by the Hospital of the University of Pennsylvania and the Cancer Institute of New Jersey [23,49].

To normalize the comparison among the models, 78,786 patches per class were employed for each model, with 70% allocated for training, 21% for testing, and 9% for validation.

3.2. Augmentation and max sharpening

The data augmentation was performed within the following ranges: width range 10%, height range 10%, zoom range 10%, and rotation range 20 degrees. Additionally, to normalize the patches, they were rescaled to a value of 1.0/255. It is important to note that no color augmentation was applied, as color is a crucial feature in histopathological classification. To evaluate the effect of augmentation, the augmented data was integrated back into the raw data (referred to as raw + aug) for the CNN models. Due to computational resource limitations, 40,000 images per class were used in the raw + aug session. For the attention models (CropNet and MS-DenseNet), replacing the average pooling with max pooling for extra sharpening was also employed.

3.3. The multi-dilation CovXNet model

The CovXNet is a new CNN-based model proposed by Mahmud in 2020 [1]. This model was designed for diagnosing COVID-19 using chest X-ray images and achieved 99% accuracy. Given the scarcity of COVID-19 X-ray training data at the time, the CovXNet could be pre-trained with a large number of similar images (chest X-Ray images of normal/viral/bacterial pneumonia). Fig. 1 illustrates the work scheme of CovXNet with augmentation.

The CovXNet characterizes a special feature extraction network, consisting of 2 basic units: the residual unit and the shifter unit (see Fig. 2). Both units are composed of depthwise convolution layers with ascending dilation rates in parallel. The most notable advantage of this design is the incorporation of features from multiple observation levels, ranging from very localized to the whole image [50–52], as the receptive field expands with an increasing dilation rate [53]. This advantage is particularly versatile in low-magnification grading, where features are dispersed. The units' output is squeezed through a strided depthwise convolution for 1/2 dimensions, which preserves more spatial information than using ordinary pooling [54].

Compared to the shifter unit, the residual unit includes a residual connection, which allows the unit's input feature map to access the output. The residual connection mitigates overfitting and gradient vanishing, and facilitates deeper spatial reduction [55].

3.4. The Breast Cancer Analyzer CNN (BCA-CNN) model

Marsh proposed a simple CNN model giving a surprisingly good

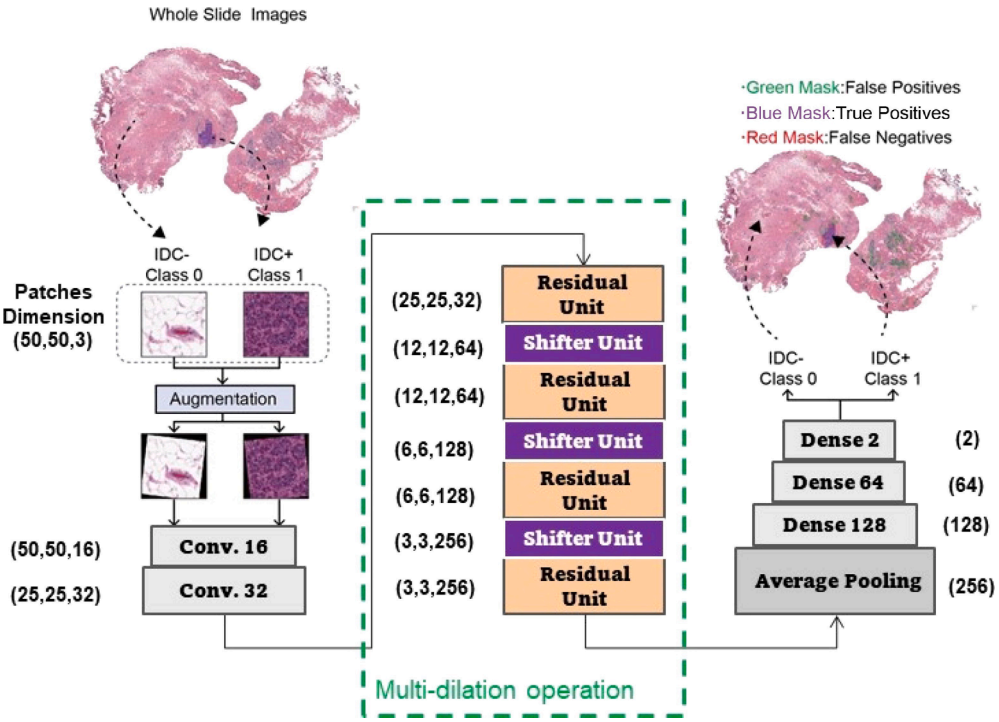

Fig. 1. Proposed scheme using augmentation with the CovXNet Model. The model's structure was adapted from [1] with permission. Models' schematics in this paper were drawn using the ML Visual tools [35]. The details of the dilation units are shown in Fig. 2. The Scheme is just showing the training. For testing and applying the model, no augmentation is needed.

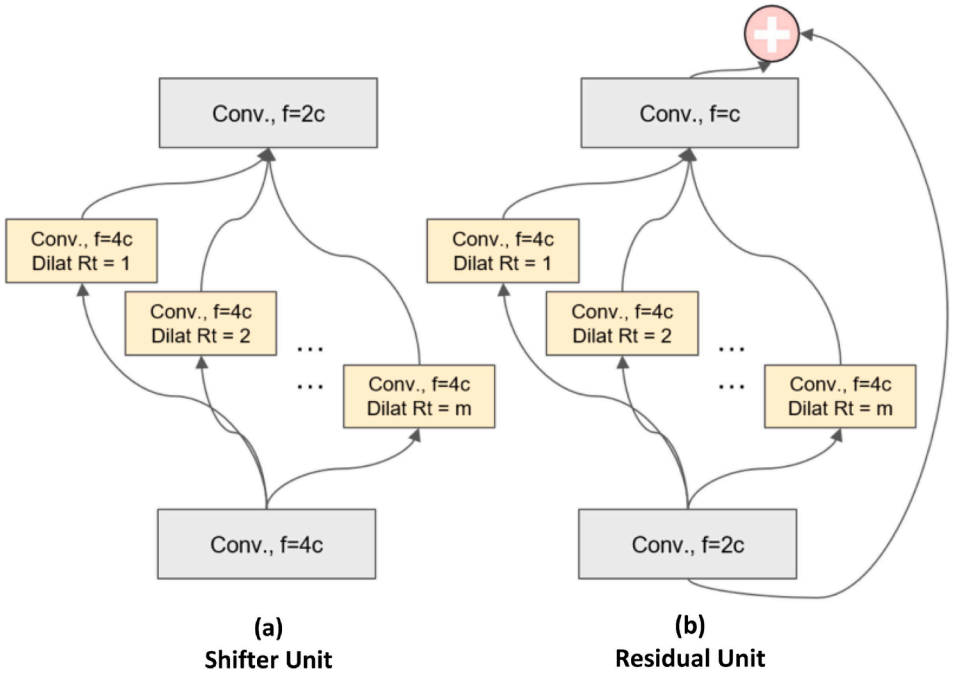

Fig. 2. Dilating Convolution Unit in CovXNet, f for the number of filters and c for the number of channels, m is set as 5. (a) shifter unit; (b) residual unit. Adapted from [1] with permission.

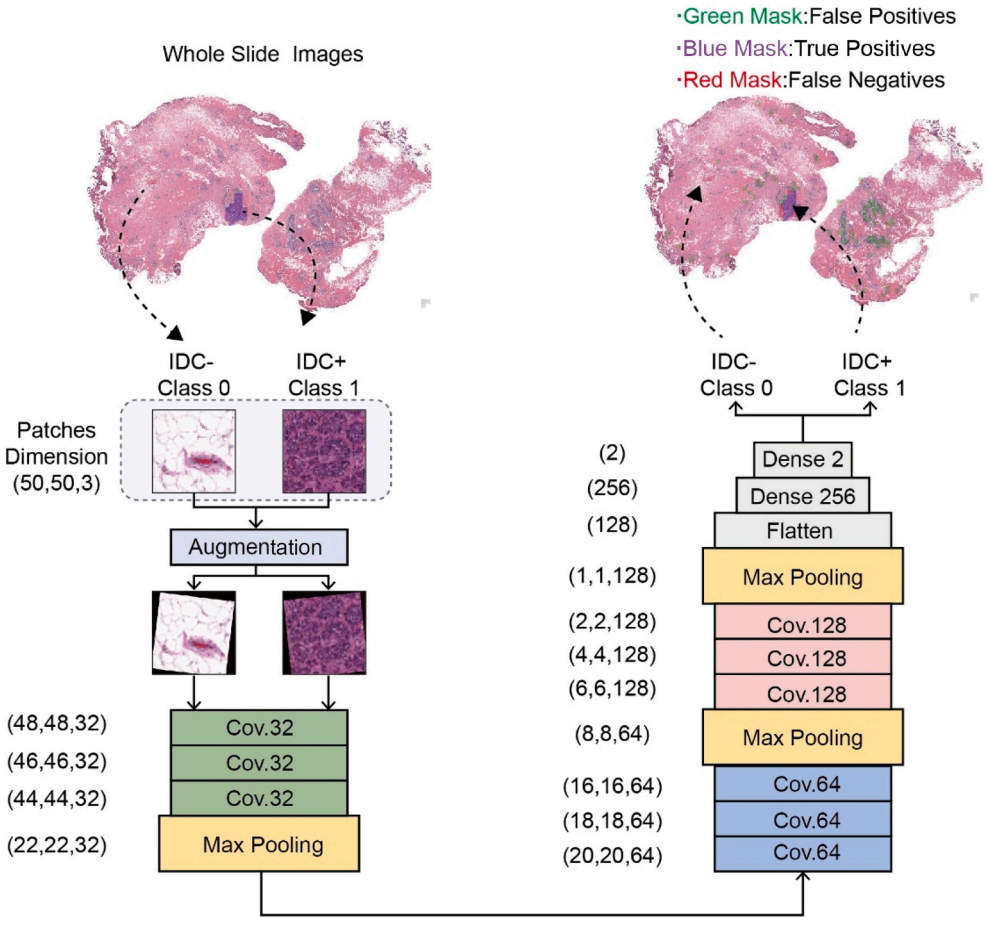

Fig. 3. Proposed scheme using augmentation with the BCA-CNN Model.

performance among the BHI dataset [44], with 87.99% accuracy, 91.03% sensitivity, and 84.96% specificity. The convolution network was constructed as 32 × 32 × 32 × 64 × 64 × 64 × 128 × 128 x 128. A max pooling layer tailed each triple convolution as dimension reduction, providing better sharpening than average pooling. Fig. 3 displays the scheme of the BCA-CNN.

3.5. The attention models (the CropNet & MS-DenseNet)

Chen J. & Zhang, D. [45] developed an attention model with high performance in colorful image classification of crop diseases, known as the CropNet. This location-wise soft-attention model achieved 99.71% accuracy on the PlantVillage dataset, a popular dataset of crop diseases. The model features a soft-attention module on top of the base layers of MobileNetV2, providing an ultra-compact size for using on mobile endpoints. The soft-attention module consists of a convolution trunk and a mask branch, which is composed of two up-sampling and two down-sampling layers to derive the attention map.

In a subsequent study, Chen, J. & Zhang, D. [46] developed a new lightweight attention model, namely Mobile Sequenze-Excitation DenseNet (MS-DenseNet). The model is adapted from DenseNet to copy its outstanding feature extraction capacity. Traditional convolutions were replaced with depthwise separable convolutions, reducing DenseNet's heavy computation to 1/9. A Squeeze-and-Excitation module (SE module) was added to each dense block for channel-wise recalibration.

3.6. Compared with typical transfer learning models

ResNet 50, ResNet 101, DenseNet 121, and DenseNet 169 were also employed for the BHI dataset. Proposed by Kaiming He in 2015 [55], ResNet was the first to introduce the residual connection, significantly accelerating learning with identity mapping and eliminating the gradient vanishing issue. ResNet consists of triple convolution blocks, and each block is enveloped with a residual connection. In 2015, ResNet won the ILSVRC championship. Additionally, Celik et al. reported achieving 90.96% accuracy on the BHI dataset using ResNet 50, and 91.57% using DenseNet 161 [18].

The DenseNet also features suppression of gradient vanishing. Instead of a residual connection of each block, DenseNet has inter-block connections. In DenseNet [56], each block's input is concatenated to all previous blocks, through a special bottle-neck layer. However, DenseNet reuses features for multiple times, massively increasing computation and video memory requirements.

In this work, to compare the pre-trained models, ResNet and DenseNet were imported without top layers and re-topped with a 4096 x 512 x 128 x 2 ANN classifier with sigmoid activation. To compare with Celik's study [18], ResNet 50 was also imported with the original top classifier and trained with top-tuning (only the top layers), as Celik et al. described.

3.7. Synthetic image augmentation generated by Generative Adversarial network (GAN)

To further study how augmentation affects the models' performance, we also generated synthetic patches from the raw using the Generative Adversarial Network (GAN) and trained the classification models. Invented by Goodfellow, L in 2014 [57], the GAN was soon widely accepted as a milestone and the technique of choice for image generation. The GAN consists of two contest neural networks: the generator, which attempts to synthesize images to deceive the discriminator; and the discriminator, which tries to distinguish synthetic images from the raw data. Both networks are trained simultaneously in a zero-sum game setting (the loss of the generator is defined by the gain of the discriminator) until reaching a Nash equilibrium [58–60]. The GAN has proved to have strong generative capacity and is widely employed for generative data augmentation [61,62].

The GAN network employed was adapted from the gold prize model in the 2019 Generative Dog Images competition [63]. The feature map of the generator was set as 10x10 and then upsampled to 50 x 50, while the channel evolved through 32-128-64-32-3, providing a 50 x 50 x 3 output as the synthetic patch. The discriminator was also reshaped accordingly, resulting in 225,007,502 parameters (225,007,500 trainable).

4. Experimental results

Different measurements (including the confusion matrix, F1 score, Cohen Kappa score (Kappa), and the Area under ROC Curve (AUC)) were employed for evaluations. F1 compares the models, Kappa examines the inter-rater reliability, and AUC assesses the resolution of models. In the context of histopathology grading, the significance of precision and sensitivity is not equal, so the confusion matrix can reflect the balance of identification more directly than F1 [64,65]. Kappa in two-class identification is specified as 2x (TPxTN-FNxFP)/[(TP + FP) × (FP + TN) + (TP + FN) × (FN + TN)], which equals the Heidke skill score (TP: true positives; FP: false positives; TN: true negatives; FN: false negatives) [66]. Noticeably, as testing data was balanced, the Test Accuracy is naturally equal to the Balanced Accuracy (BAC).

4.1. Augmented patches trained performance

In the first experimental session, all the employed models were respectively trained with raw and augmented patches. The epoch history of accuracy was plotted to monitor the overfitting (brief in Table 2 and Table 3′s Best Epoch). In raw data training, the train accuracies quickly rose to 90%, while the validation accuracy reached a ceiling of around 86% at an early Best Epoch. This overfitting trend was suppressed with augmentation (namely Aug 1st train). The validation accuracy climbed slower (later best epoch), and the difference between training and validation or testing accuracy was <1%.

The suppression of overfitting with augmentation implied an improvement in continued training. Thus, we trained both models three times more epochs (namely Aug 2nd, 3rd, and 4th train), until validation accuracy stopped increasing. The BAC kept increasing and reached 89.56% and 88.48% (CovXNet and BCA-CNN, see Table 2, Table 3 and Fig. 8), while the sensitivity and specificity first converged then diverged (see Fig. 5 and Fig. 6). The F1 and Kappa also kept increasing, while the AUC reached a ceiling at ~ 95%. Furthermore, we trained the CovXNet with 120,000 patches per class (oversampled), providing an accuracy of 92.13% (consistent testing), with the sensitivity and specificity both >90% (see Fig. 4 and Table 2). The training history is shown in Fig. 7. With 78,786 samples, the accuracies reached a ceiling at 90th epoch, together with a sign of overfitting (early best epoch in Table 2). With 120,000 samples, the accuracies climbed higher, and loss dived lower, but the validation loss also fluctuated a bit more. However, increased data volume demoted BCA-CNN's performance.

4.2. Max sharpening with attention model

Without augmentation, both attention models achieved a BAC of around 85% (see Table 4 and Table 5), which is a decent performance compared to the literature (see Table 1). In contrast to the CovXNet and BCA-CNN, the augmentation impaired the performances of the CropNet and MS-DenseNet. With augmentation, the MS-DenseNet's performance decline was quite small (<0.5% in BAC, F1, KAPPA, and AUC, see Table 5) while CropNet's performance dropped by >2% in multiple scores. This indicates that the channel-wise MS-DenseNet is more invariant than the location-wise CropNet. Still, augmentation led to a divergence in sensitivity and specificity with MS-DenseNet. It is not surprising that in attention models, augmentation can interrupt the concentrating of attention maps.

For extra sharpening in the CropNet, we used max pooling to

Table 2
Performance scores of CovXNet, with and without Augmentation (Aug).

	Best Epoch	Training Accuracy	Testing BAC	F1	KAPPA	Pre-cision	Specifi-city	Sensi-tivity	AUC
No Aug	15	90.94%	86.74%	86.73%	73.47%	86.75%	85.62%	87.85%	93.96%
Aug 1st train	25	89.41%	88.06%	88.06%	76.12%	88.12%	86.04%	90.08%	94.92%
Aug 2nd rain	17	90.63%	88.62%	88.62%	77.23%	88.62%	89.00%	88.23%	95.32%
Aug 3rd train	2	90.63%	89.24%	89.22%	78.47%	89.43%	85.75%	92.72%	95.83%
Aug 4th train	4	90.72%	89.56%	89.55%	79.12%	89.71%	86.49%	92.63%	95.92%
120,000 data Aug	18	92.36%	92.13%	92.13%	84.27%	92.14%	91.68%	92.58%	97.47%

Table 3
Performance scores of BCA-CNN, with and without Augmentation (Aug).

	Best Epoch	Training Accuracy	Testing BAC	F1	KAPPA	Pre-cision	Specifi-city	Sensi-tivity	AUC
No Aug	38	89.45%	86.67%	86.67%	73.34%	86.70%	85.19%	88.15%	93.81%
Aug 1st train	46	88.04%	87.45%	87.42%	74.90%	87.86%	82.23%	92.67%	94.66%
Aug 2nd rain	43	89.18%	88.20%	88.19%	76.40%	88.28%	85.83%	90.57%	95.04%
Aug 3rd train	38	89.90%	88.23%	88.23%	76.46%	88.26%	86.87%	89.59%	95.02%
Aug 4th train	36	90.48%	88.48%	88.47%	76.96%	88.57%	86.11%	90.85%	94.92%

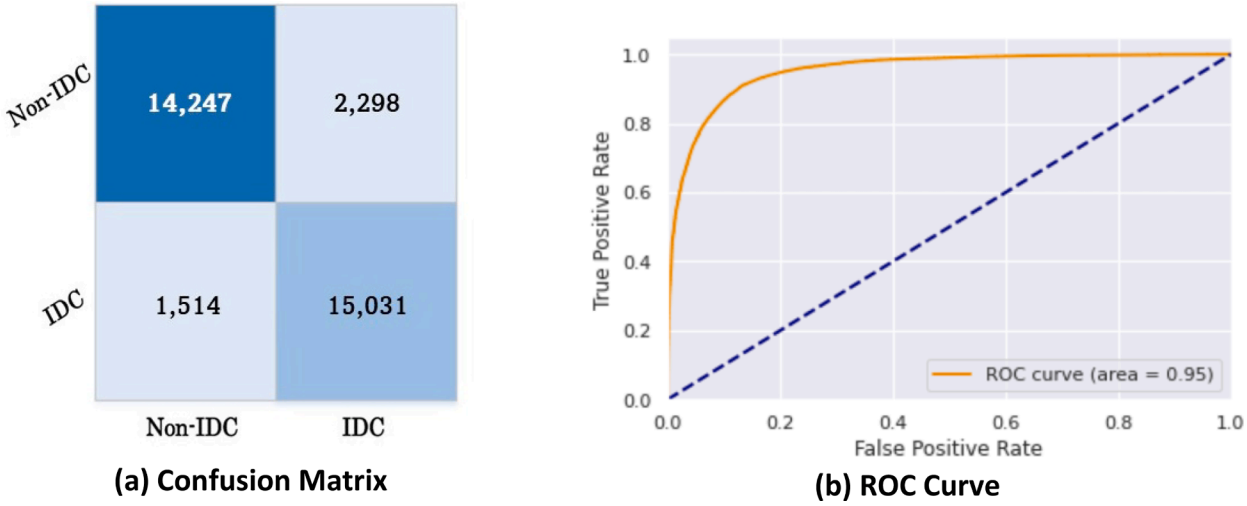

Fig. 8. Testing performance of the BCA-CNN trained with augmentation: a) Confusion Matrix, b) ROC curve.

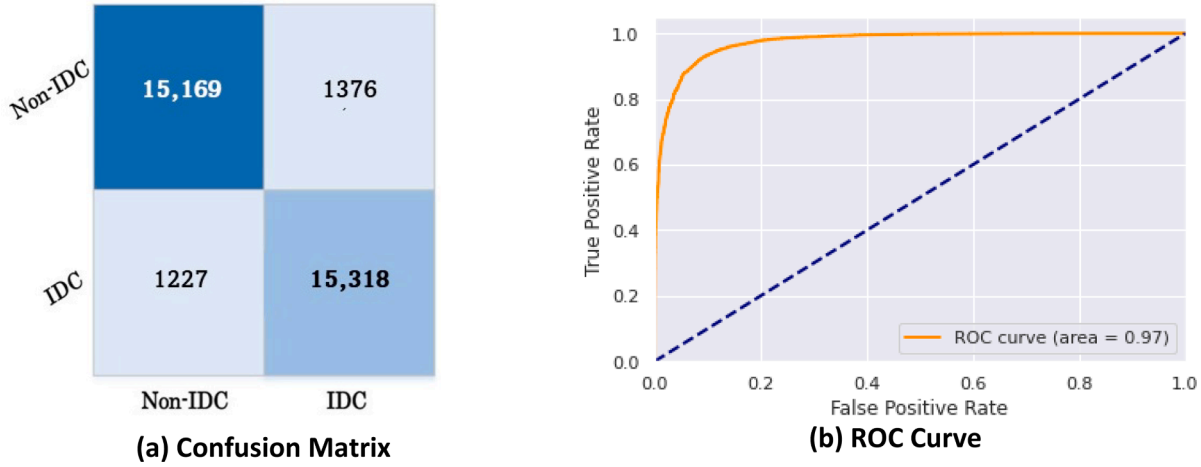

Fig. 4. Testing performance of the CovXNet trained with augmentation and oversampling 120,000 patches per class: a) Confusion Matrix, b) ROC curve.

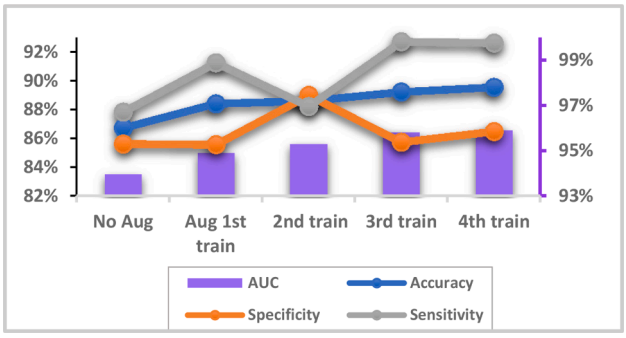

Fig. 5. Performance of CovXNet along training, 30 epochs each train.

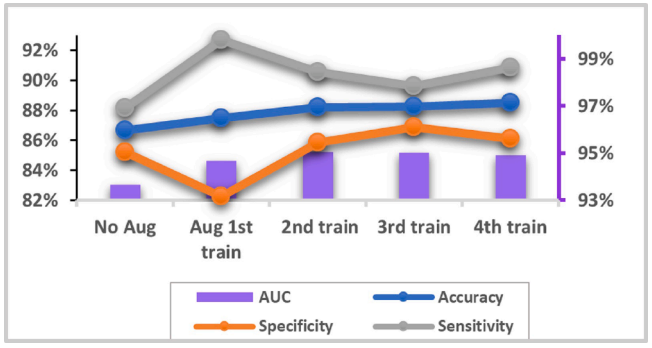

Fig. 6. Performance of BCA-CNN along training, 50 epochs each train.

substitute the average pooling, which improved performance with and without augmentation (see Table 4), in BAC, F1, KAPPA and AUC. The specificity and sensitivity were also more balanced with max pooling.

Interestingly, the improvement in performance with max pooling is larger when used with augmentation (0.6% in BAC) than without augmentation (0.2% in BAC). This indicates that max pooling can help attention models concentrate a bit more, especially under the variance induced by augmentation.

Additionally, the CropNet and MS-DenseNet were extremely sensitive to hyperparameters. With initial settings, the MS-DenseNet's accuracy remained at 50% and yielded NaN loss during training. The categorical output matrix also consisted of NaN, which is abnormal. The issue was resolved through the following adjustments: 1. Changing the learning rate of SGD from 0.01 to 0.00001; 2. Adding early stopping (patience = 20) to the callback. However, the accuracy-epoch plot revealed that the validation accuracy still exhibited considerable fluctuations (see Github). The CropNet performed better with initial hyperparameters, providing numerical loss but still maintaining an accuracy of ~50%. After changing the learning rate of Adam to 0.00001 and adding a clipnorm of 1, the BAC increased to ~85%. Following these adjustments, the CropNet's validation accuracy fluctuated less along epochs than that of the MS-DenseNet.

Another intriguing phenomenon was that, despite the CropNet and MS-DenseNet's accuracy elevating to ~85%, the AUC calculated by the standard AUC function was ~50%, which is contradictory. The reason behind this was later discovered to be that the attention models did not capture the mutual exclusivity between the two classes. Specifically, the CropNet and MS-DenseNet's categorical output vector contained many rows with sums $\neq 1$, e.g., [0.31, 0.34].

As is well-known, the categorical output vector represents the probabilities of each outcome class (each row corresponds to the prediction of a sample, and each column represents the prediction probability of each class). A row like [0.31, 0.34] is output as class 1 through numpy.argmax. However, in sklearn.metrics.roc_curve & auc's evaluation, the 0.34 would be considered as a bias towards class 0. We addressed this issue by normalizing each row's sum to 1 using the

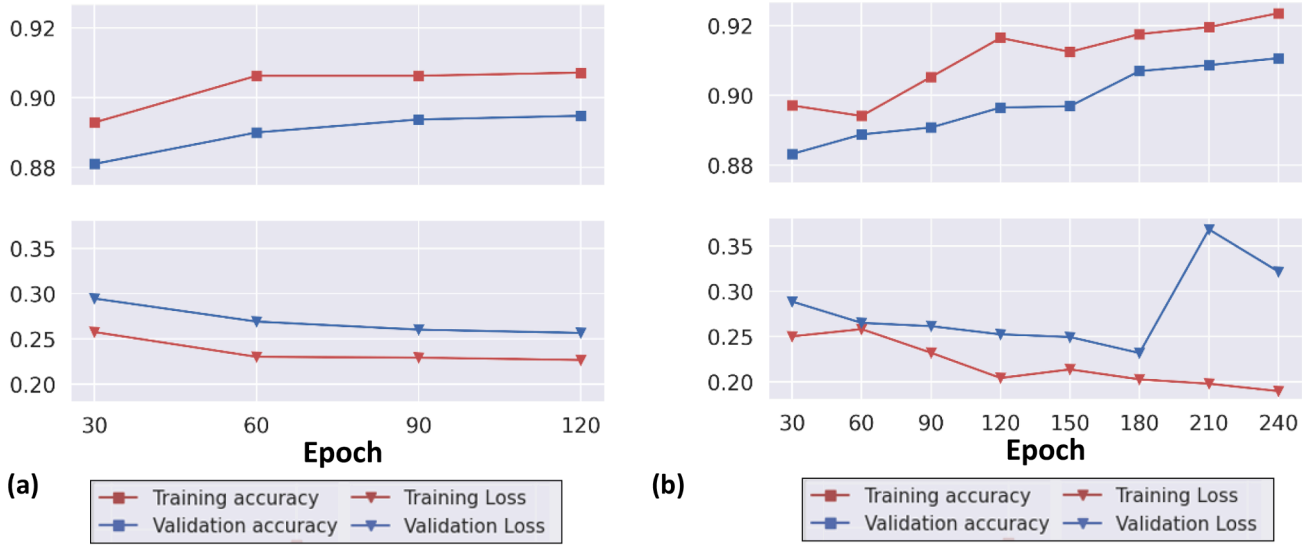

Fig. 7. Training history of CovXNet with augmentation, (a) 78,786 patches per class; (b) 120,000 patches per class.

Table 4
Performance of the Cropnet, regulated by augmentation and/or modified with max pooling.

Training	Best Epoch	Training Accuracy	Testing BAC	F1	KAPPA	Precision	Specificity	Sensitivity	AUC
Raw	20	87.45%	85.61%	85.60%	71.21%	85.64%	84.08%	87.13%	92.64%
Raw + Max	22	87.63%	85.88%	85.88%	71.77%	85.59%	85.12%	86.65%	92.81%
Aug	21	85.52%	83.22%	83.21%	66.45%	83.30%	80.79%	85.66%	90.87%
Aug + Max	30	84.90%	83.85%	83.84%	67.69%	83.88%	82.41%	85.28%	91.07%

Table 5
Performance scores of the MS-DenseNet, regulated by augmentation.

Training	Best Epoch	Training Accuracy	Testing BAC	F1	KAPPA	Precision	Specificity	Sensitivity	AUC
Raw	29	84.77%	84.67%	84.67%	69.60%	84.67%	84.54%	85.79%	92.46%
Aug	20	83.12%	84.62%	84.61%	69.18%	84.70%	82.19%	87.04%	92.11%

Partition Ratio Theorem. After normalization, [0.31, 0.34] would be interpreted as [47.7%, 52.3%]. In ROC evaluation, this row ("[47.7%, 52.3%]") would be regarded as a "weak" judgment of class 1 instead of a "wrong" bias towards class 0. Therefore, the ROC evaluation after normalization is more reliable.

4.3. GAN-generated patches trained model's performance

The GAN-generated patches are shown in Fig. 9. In human eyes, the GAN-generated patches contained the same patterns as the raw. Both the raw and GAN-generated IDC + patches presented tubular structures formed by malignant cells (see the yellow circles in Fig. 9), which is the signature of IDC [67,68]. The IDC is usually referred to as IDC-non specific type [69], as the IDC tumor tissue has no unified characteristics, but 3 commonalities in morphology: the tubular malignant cluster, pleomorphic cell nucleus, and high mitosis count. However, the tubular malignant cluster is the most obvious at the 40X magnification. Also, both raw and GAN-generated IDC- patches often contained large adipose cells (see the green circles in Fig. 9), which are usually narrowed and lessened by the invasion of malignant cells and fibroblasts in the IDC + area [70].

With the GAN-generated patches, the training accuracy rose to ~99% within 10 epochs, while the validation accuracy remained at ~50% (see Fig. 11). Looking into the confusion matrix (see Fig. 10), the GAN-trained models classified most patches either to class 0, or to class 1. This shows that the GAN-trained models failed to draw the threshold between the 2 classes.

4.4. IDC prediction maps

To evaluate how promotions in single-patch identification affect the whole slide grading, the patches after identification were reconstructed to whole slides as Merii described in [71]. With higher single-patch identification accuracy, the CovXNet and BCA-CNN with augmentation presented an improvement in reconstructed maps, especially with fewer green marks (false positive) (see Fig. 12). However, as training

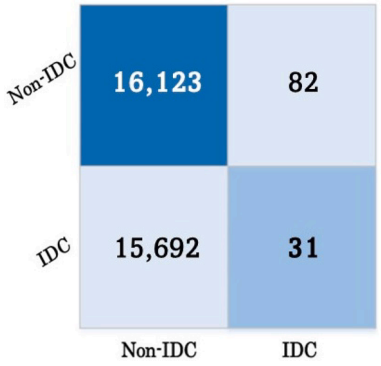

Fig. 10. Testing confusion matrix of CovXNet trained by GAN-generated patches.

continued, the performance was not 100% synchronous with increasing accuracy. The second train of the CovXNet and the third train of BCA-CNN had better reconstructed maps, even though the 4th train of both models gave the highest single patch accuracy.

Additionally, the improvement in reconstructed maps was not monotonic. As shown in the magnified area & arrow in Fig. 12, patches previously classified correctly may be misidentified while the overall performance was improved (Green: false-positive; Red: false-negative). Thirdly, false positives were more of a problem than false negatives, as the true positive area was much lower than the true negative area.

The reconstructed maps of CropNet, MS-DenseNet, ResNet, and DenseNet are in Supplementary Figs. 1 and 2. The reconstructed map of 92.13% CovXNet is in Supplementary Fig. 3.

4.5. Augmented + raw images trained performance

To investigate the mechanism of augmentation, augmented patches were mixed with raw ones, and were used to train CovXNet and BCA-CNN models (performances are shown in Table 6 and Table 7). Both

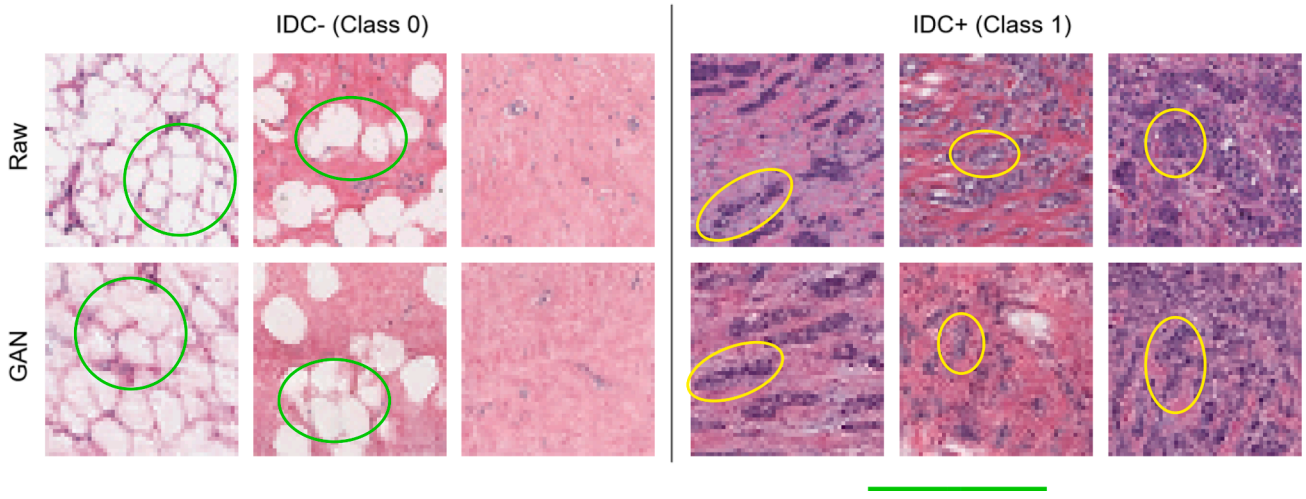

Fig. 9. Raw and GAN generated patches of BHI, the green circles highlight the adipose cells and the yellow circles highlight the tubular structure formed by malignant cells. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

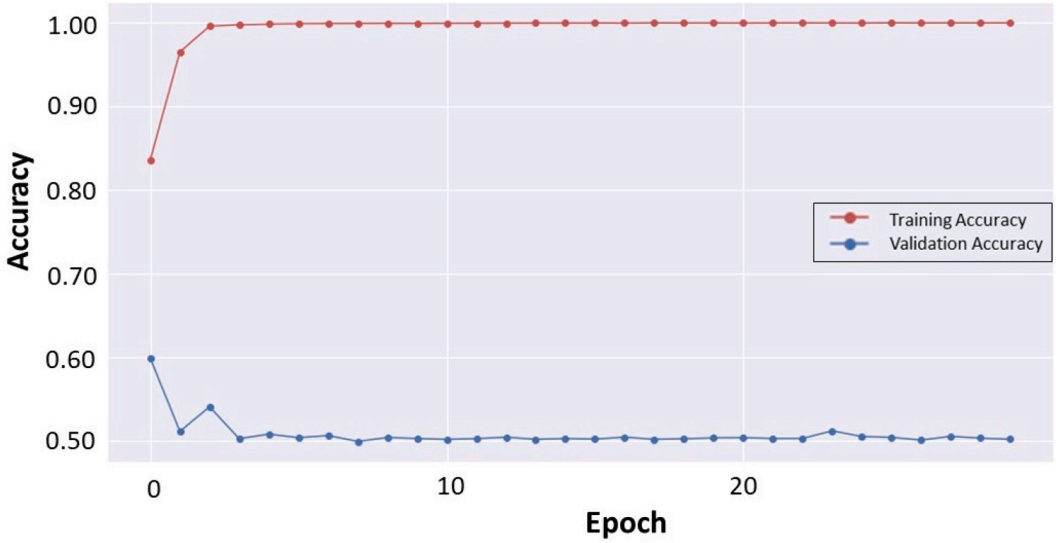

Fig. 11. Training history of GAN-generated patches with trained CovXNet, validated by the raw data.

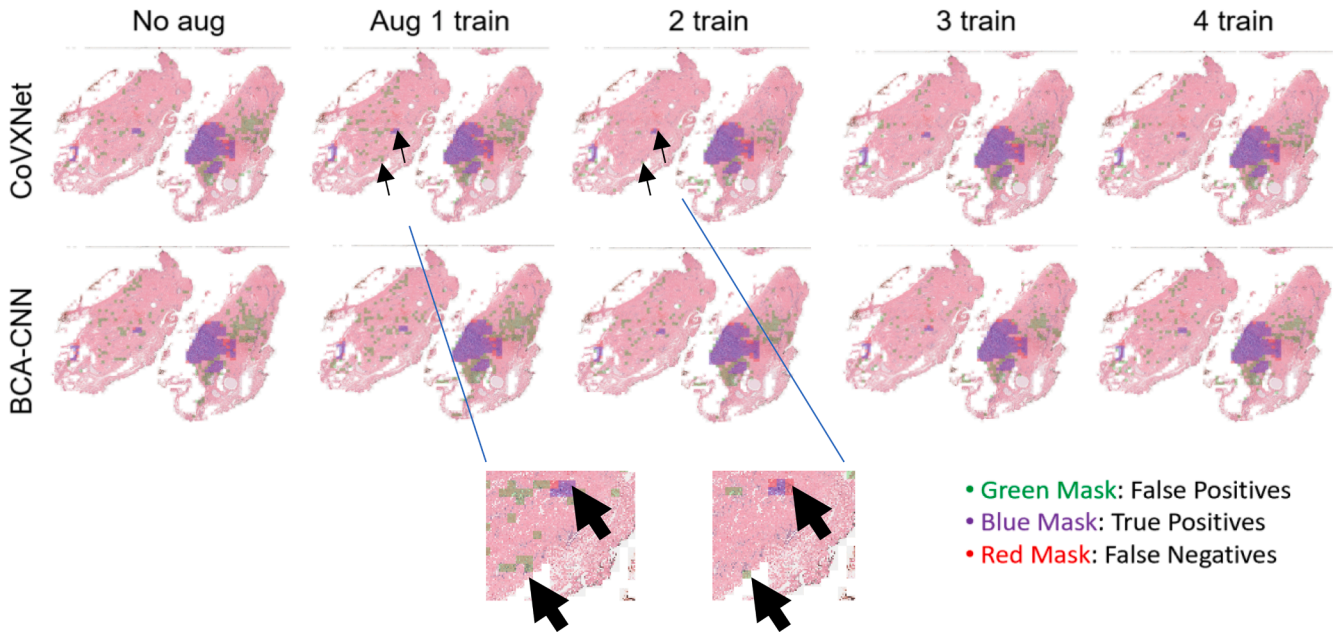

Fig. 12. Reconstructed maps of patches classified by ConVXNet and BCA-CNN. Each colored block is a single patch. The blue line links a magnification to the arrow area. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 6
CovXNet Performance with Combined Training.

Training	Best Epoch	Training Accuracy	Testing BAC	F1	KAPPA	Precision	Specificity	Sensitivity	AUC
Raw	11	88.43%	87.36%	87.35%	74.72%	87.42%	85.28%	89.44%	94.33%
Aug	18	87.96%	87.92%	87.92%	75.85%	87.92%	87.87%	87.98%	94.74%
Aug + Raw	15	88.45%	86.88%	86.87%	73.75%	86.91%	85.30%	88.45%	94.10%

BAC and other scores (such as F1, KAPPA, and AUC, etc.) decreased in raw + aug training, compared to solely augmented training. This again verified that increasing data volume is not the mechanism of augmentation [39].

4.6. Transfer learning

To investigate the effectiveness of augmentation for common CNN-type pre-trained models, we employed DenseNet 169 & 121 and ResNet 50 & 101 models. Both ResNet and DenseNet were previously explored by Celik in the BHI dataset [18]. Interestingly, both DenseNets' and ResNets' performance dropped after augmentation, as shown in

Table 7
BCA-CNN Performance with Combined Training.

Training	Best Epoch	Training Accuracy	Testing BAC	F1	KAPPA	Precision	Specificity	Sensitivity	AUC
Raw	20	86.29%	85.55%	85.54%	71.10%	85.57%	84.18%	86.92%	92.91%
Aug	49	87.55%	86.10%	86.08%	72.19%	86.21%	83.27%	88.92%	93.51%
Aug + Raw	16	87.51%	85.86%	85.86%	71.73%	85.87%	85.29%	86.44%	93.09%

Table 8
Performance of Transferred Learning models (ResNet and DenseNet) with the same ANN Top and trained in top-tuning.

	Train -ing	Best Epoch	Training Accuracy	Testing BAC	F1	KAPPA	Pre-cision	Specifi-city	Sensi-tivity	AUC
DsNet 121	raw	6	84.45%	82.51%	82.50%	65.01%	82.51%	81.80%	83.22%	90.29%
	aug	21	82.10%	81.78%	81.78%	63.55%	81.78%	81.89%	81.67%	82.95%
DsNet 169	raw	5	84.00%	82.48%	82.49%	64.97%	82.52%	84.09%	80.89%	89.48%
	aug	15	81.89%	81.61%	81.61%	63.23%	81.64%	83.03%	80.19%	89.18%
RsNet 101	raw	14	73.63%	73.85%	73.71%	47.71%	74.36%	81.09%	66.61%	78.50%
	aug	13	72.58%	65.54%	63.95%	31.09%	68.87%	44.53%	86.55%	71.82%
RsNet 50	raw	14	78.46%	77.85%	77.85%	55.69%	77.85%	77.86%	77.83%	84.06%
	aug	13	73.72%	69.78%	69.19%	39.56%	71.45%	55.86%	83.71%	79.59%

Table 8. The drop in DenseNets was small (~1% in BAC, precision, F1, and KAPPA) and larger (7%~8%) in ResNets.

Secondly, with augmentation, the ResNets' specificity dropped to ~50% (KAPPA dropped to ~30%), indicating an ineffective learning of class 0′s pattern. Furthermore, the augmentation gave rise to larger divergence of training–testing accuracy in ResNets. The best epoch also occurred earlier, indicating signs of overfitting. Additionally, without augmentation, ResNet 50 outperformed ResNet 101 by 4–8% in terms of BAC, F1, KAPPA, and AUC. After augmentation, the drop in scores were smaller in ResNet 50 than in ResNet 101.

Contradictorily, in DenseNets the augmentation converged the training and testing accuracy, and delayed the best epoch, indicating overfitting suppression. However, the accuracy and other performance still slightly dropped. Both phenomena elucidate that augmentation is not always suppressing overfitting, and the suppression of overfitting is not 100% lead to performance promotion.

Using Celik's method [18], the ResNet 50 was imported with the original top layers and top-tuned (training only the top layers), resulting in an accuracy of 79.86% (see Table 9 for detailed results). However, with the original top, fine-tuning (training the whole model), and augmentation, ResNet 50 reached a BAC of 87.84%. The reason for not reaching the reported 90.96% accuracy [18] may be attributed to differences in unrevealed hypermeter settings. Interestingly, with the original top, ResNet 50′s performance was improved by augmentation, more in fine-tuning (3~5% in BAC, F1 and KAPPA) and less (0.3~0.5%) in top-tuning.

4.7. Two types of common bugs mis-using shuffle function

A noticeable issue is that some of the notebooks for the BHI dataset on Kaggle [72–74] reached an overall accuracy >90%, while presenting a highly imbalanced confusion matrix. The number of class 0 samples was much higher than class 1 (class zero: class one >3:1). Additionally, the sensitivity was just <75% while the specificity was >90%. Such a class 0-class 1 ratio is very unlikely given by random sampling. By adding a testing module to map the class distribution between each step, this issue was found attributed to using random.shuffle over data-label composited arrays. 5000 patches per class were inputted, and after shuffling the testing output turned to 8497 class 0 and 1503 class 1 (see the GitHub repository under mask label error folder). Further mining showed that the first 5000 labels in the array were all masked as class 0.

This shed light on the reliability of other shuffling functions. We investigated another more commonly used integrated shuffling in the ImageDataGenerator function. 700 patches per class were inputted into the ImageDataGenerator with integrated shuffling activated. After processing by the generator, the output consisted of 703 class 0 and 697 class 1 samples (see the GitHub repository under the mask label error folder). The distortion from ImageDataGenerator's shuffling was much smaller. Subsequently, this work verified that using sklearn.shuffle, which can co-shuffle the data and labels in two individual arrays, was the safest method.

5. Discussion

In the literature, it has been established that single patch classification and the corresponding histopathology grading can be hindered by the 40X magnification [14]. Previous studies have explored various deep learning strategies to enhance identification among the 40X BHI (see Table 1). In this study, with augmentation and the CovXNet model, the BAC was promoted to 92.13%, which is currently the highest. In previous publications, only Romero's [25] and Celik's [18] methods acquired BAC >89% [18,25]. Furthermore, this proposed method pushed the specificity and sensitivity beyond 90% simultaneously for the first time. The augmentation strategy also significantly improved other models for this task, such as the BCA-CNN and ResNet 50 (original top, fine-tuning).

The promotion of single-patch identification was consistent in whole-slide mapping, as demonstrated in Fig. 12. The statistics behind

Table 9
ResNet50 with original top layers' performance, either Fine-tuning (training all the layers), or top-tuning (training only the top layers). With (50, 50, 3) input (dimension of single patch), the original top classifier network is 2048x2.

	Training	Best Epoch	Training Accuracy	Testing BAC	F1	KAPPA	Preci-sion	Specifi-city	Sensiti-vity	AUC
Fine-Tune	raw	10	89.12%	84.82%	84.81%	69.64%	84.91%	82.25%	87.39%	91.98%
	aug	49	88.28%	87.84%	87.84%	75.69%	84.85%	87.73%	87.96%	94.75%
Top-tune	raw	43	79.63%	79.86%	79.86%	59.73%	79.90%	81.48%	78.25%	86.60%
	aug	50	79.55%	80.10%	80.08%	60.20%	80.19%	82.91%	77.29%	86.77%

reconstructing can be modeled as a binomial distribution, considering each single-patch identification as one Bernoulli test. Consequently, the improvement in single-patch identification is linearly synchronous with the enhancement in whole-slide mapping. However, as highlighted in section 3.4, specificity is more critical than sensitivity in tumor grading.

Among the models tested in this work, the CovXNet provided the highest performance, with a specially designed structure for image classification. As described in section 2.3, CovXNet has multi-dilation filter groups in the shifter and residual units to extract features from different observation levels. The residual unit is also enveloped with a residual connection to suppress gradient vanishing [55], which was first introduced in the ResNet. However, Researchers have established that residual connections also limit feature representing capacity [75–77] and thus classification accuracy, despite suppressing gradient vanishing. Yet CovXNet shows no sign of this issue with its elaborate structure. Reasoning in details, the residual connection is identical and linear, leading to a domain collapsing issue in feature presentation, thus reducing the learning capacity [78]. Yet, in the residual unit of CovXNet, multiple filter groups provide extra de-coupling, which is also the idea of ResXNet [79]. Compared to ResXNet, CovXNet's filter groups have different dilation rates, offering even more decoupling. Additionally, the shifter unit intermediate possesses no residual connection, breaking the continuity to directly transfer very bottom features to the top classifier.

Different models have different feature presentation and invariance provided by the model structure. Commonly elucidated, data augmentation increases data diversity, hence capable to encourage the model to learn more about the underlying patterns, instead of simply memorizing the data (i.e. increasing the generalization capacity) [80–82]. However, models with low invariance can be demoted by augmentation. Among the transfer learning models, the DenseNets performed much better than ResNets in multiple aspects, including classification capacity, invariance, and overfitting suppression. Since the ANN top classifier is the same and both models were pre-trained with ImageNet [55,56], the only difference is the feature presentation network. Even before augmentation, DenseNets significantly outperformed ResNets, indicating that DenseNet's feature representation is better suited for this task. As previously discussed, ResNet has a domain collapsing issue given by linear residual connections. ResXNet and CovXNet solve this problem with multiple filter groups. DenseNet, on the other hand, employs another approach. Instead of a direct residual connection, each block is jumped with a bottle-neck concatenation, and linked to all previous blocks, thus suppressing the gradient vanishing and domain collapsing simultaneously, leading to both better feature presentation and higher parameter efficiency [56]. After augmentation, ResNets' (ANN top) specificity dropped to ~50% (drop >20%) (see Table 8**)**, meaning for class 0 the model hardly learns any patterns applicable to unknown data. The variance induced by augmentation and after ResNets' feature presentation overwhelms the classifier, rather than prompting the model to overcome the variance and achieve higher generalization. DenseNet did not exhibit such specificity stall, and overfitting was suppressed after augmentation, demonstrating that DenseNet has higher invariance and increased generalization with augmentation. However, the ~1% drop in classification scores also reflects that augmentation also complicates underlying patterns and toughens the learning [83]. This is further confirmed by the drop in training accuracy.

Compared to standard image recognition (eg. "cat vs dog"), this grading task also shows a specialty in ResNet's feature presentation issue. Compared to top-tuning, fine-tuning significantly improved the performance as well as the invariance (in fine-tuning there was a large improvement after augmentation) (see Table 9). In fine-tuning, feature extraction weights are re-adjusted. This aligns with the specificity of histopathology, where information is held at multiple levels, from tissue to sub-cellular structures, which differs from normal images in ImageNet. The top classifier is also vital to this task. Both in top-tuning, the original top (2048×2) performed better than the refitted ANN top ($4096 \times 512 \times 128 \times 2$), and manifested a higher invariance under augmentation. This reflects that networks with higher complexity are prone more to overfitting, while simpler models reduce the chance to fit noise [84,85]. Moreover, BCA-CNN's performance was better than ResNet's. Both networks consist of triple convolution blocks, yet BCA-CNN only has 3 blocks (ResNet 50 has 4) and no residual connection.

The GAN-generated data training resulted in unusable outcomes. The GAN-trained model failed to delineate the correct boundary between the two classes, suggesting that the GAN-generated data only exposed the most "learnable" features from the raw data and oversimplified the classification. Literature has elucidated that through GANs, raw features are lost and preserved in a black-boxed manner [86], requiring a large volume of data. However, when information concentration is low (as in low-magnification grading), even with a large data volume, critical features are still at risk of being lost as they are more dispersed. In contrast, shift-based augmentation does not discard the information.

Both location-wise and Squeeze-and-Excitation attention models gave ~85% BAC, which is moderate compared to the literature. The non-exclusive output categorical matrix indicates that the attention model is harder to "learn" the relation between classes. With augmentation, the SE-type MS-DenseNet enjoys higher invariance than the location-wise CropNet. With CropNet, despite the demotion of performance under augmentation, adapting with max pooling offers an improvement in scores (BAC, F1, Kappa, and AUC) and also a more balanced sensitivity and specificity. Furthermore, max pooling can rescue some of the performance demotion under augmentation. Considering augmentation as a perturbance, max pooling can assist the concentration of the attention model.

6. Conclusion

This work elucidated that augmentation could enhance classification and patch-based grading of low-magnification histopathology when using models with high invariance, such as the CovXNet, BCA-CNN, and fine-tuned ResNet 50. By incorporating augmentation, the multi-dilation CovXNet achieved a classification accuracy of 92.13%, which is the highest for the BHI dataset in the current literature. This study also identifies two types of common errors in integrated shuffling, which emphasizes the importance of data verification. Sklearn.shuffle can guarantee a more reliable data shuffling, ultimately leading to better performance and generalization.

CRediT authorship contribution statement

Zonghan Gan: Conceptualization, Methodology, Software, Writing – original draft. **Abdulhamit Subasi:** Supervision, Validation, Writing – review & editing, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to my code in my manuscript

Acknowledgments

During primary testing on Kaggle, three friends (Nikhil Menghani, Christopher Lush and Bo Wang) kindly lent their Kaggle's GPU duration. Also five friends (Ziyun Wang, Yizhu Wang, Lipeng Mao, Zhiying Ke, Yutao Tang) kindly answered some questions about attention model through emails.

Data Availability:

Our code is available at https://GitHub.com/Zonghan-Barry-Ga

n/2022-promoting-grading-and-classification-in-BHI-dataset

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bspc.2023.105118.

References

[1] T. Mahmud, M.A. Rahman, S.A. Fattah, CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization, Comput. Biol. Med. 122 (2020), 103869, https://doi.org/10.1016/j.compbiomed.2020.103869.

[2] P. Kleihues, F. Soylemezoglu, B. Schäuble, B.W. Scheithauer, P.C. Burger, Histopathology, classification, and grading of gliomas, Glia 15 (1995) 211–221, https://doi.org/10.1002/glia.440150303.

[3] X. Dai, L. Xiang, T. Li, Z. Bai, Cancer hallmarks, biomarkers and breast cancer molecular subtypes, J. Cancer 7 (2016) 1281–1294, https://doi.org/10.7150/jca.13141.

[4] M.J. Engstrøm, S. Opdahl, A.I. Hagen, P.R. Romundstad, L.A. Akslen, O.A. Haugen, L.J. Vatten, A.M. Bofin, Molecular subtypes, histopathological grade and survival in a historic cohort of breast cancer patients, Breast Cancer Res. Treat. 140 (2013) 463–473, https://doi.org/10.1007/s10549-013-2647-2.

[5] R.B. Greenough, Varying degrees of malignancy in cancer of the breast, J. Cancer Res. 9 (1925) 453–463, https://doi.org/10.1158/jcr.1925.453.

[6] K.N. Gibson-Corley, A.K. Olivier, D.K. Meyerholz, Principles for valid histopathologic scoring in research, Vet. Pathol. 50 (6) (2013) 1007–1015.

[7] L. Power, L. Acevedo, R. Yamashita, D. Rubin, I. Martin, A. Barbero, Deep learning enables the automation of grading histological tissue engineered cartilage images for quality control standardization, Osteoarthr. Cartil. 29 (2021) 433–443, https://doi.org/10.1016/j.joca.2020.12.018.

[8] R. Jaroensri, E. Wulczyn, N. Hegde, T. Brown, I. Flament-Auvigne, F. Tan, Y. Cai, K. Nagpal, E.A. Rakha, D.J. Dabbs, N. Olson, J.H. Wren, E.E. Thompson, E. Seetao, C. Robinson, M. Miao, F. Beckers, G.S. Corrado, L.H. Peng, C.H. Mermel, Y. Liu, D. F. Steiner, P.-H.-C. Chen, Deep learning models for histologic grading of breast cancer and association with disease prognosis, npj Breast Cancer 8 (2022) 1–12, https://doi.org/10.1038/s41523-022-00478-y.

[9] D. Komura, S. Ishikawa, Machine learning methods for histopathological image analysis, computational and structural, Biotechnol. J. 16 (2018) 34–42, https://doi.org/10.1016/j.csbj.2018.01.001.

[10] Y. Wang, B. Acs, S. Robertson, B. Liu, L. Solorzano, C. Wählby, J. Hartman, M. Rantalainen, Improved breast cancer histological grading using deep learning, Ann. Oncol. 33 (2022) 89–98, https://doi.org/10.1016/j.annonc.2021.09.007.

[11] L. Hou, D. Samaras, T.M. Kurc, Y. Gao, J.E. Davis, J.H. Saltz, Patch-Based Convolutional Neural Network for Whole Slide Tissue Image Classification, IEEE Computer Society, in, 2016, pp. 2424–2433.

[12] A. Ashtaiwi, Optimal histopathological magnification factors for deep learning-based breast cancer prediction, Appl. Syst. Innov. 5 (2022) 87.

[13] V. Gupta, A. Bhavsar, Breast cancer histopathological image classification: is magnification important? in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 17–24.

[14] S. Singh, R. Kumar, Breast cancer detection from histopathology images with deep inception and residual blocks, Multimed. Tools Appl. 81 (2022) 5849–5865, https://doi.org/10.1007/s11042-021-11775-2.

[15] M. Gour, S. Jain, T. Sunil Kumar, Residual learning based CNN for breast cancer histopathological image classification, Int. J. Imaging Syst. Technol. 30 (3) (2020) 621–635.

[16] C.F. Sabottke, B.M. Spieler, The effect of image resolution on deep learning in radiography, Radiol. Artif. Intell. 2 (2020) e190015.

[17] V. Thambawita, I. Strümke, S.A. Hicks, P. Halvorsen, S. Parasa, M.A. Riegler, Impact of image resolution on deep learning performance in endoscopy image classification: an experimental study using a large dataset of endoscopic images, Diagnostics (Basel) 11 (2021) 2183, https://doi.org/10.3390/diagnostics11122183.

[18] Y. Celik, M. Talo, O. Yildirim, M. Karabatak, U.R. Acharya, Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images, Pattern Recogn. Lett. 133 (2020) 232–239, https://doi.org/10.1016/j.patrec.2020.03.011.

[19] M.S. Reza, J. Ma, Imbalanced histopathological breast cancer image classification with convolutional neural network, in: 2018 14th IEEE International Conference on Signal Processing (ICSP), 2018, pp. 619–624, 10.1109/ICSP.2018.8652304.

[20] R. Singh, T. Ahmed, A. Kumar, A.K. Singh, A.K. Pandey, S.K. Singh, Imbalanced breast cancer classification using transfer learning, IEEE/ACM Trans. Comput. Biol. Bioinf. 18 (2021) 83–93, https://doi.org/10.1109/TCBB.2020.2980831.

[21] H. Alghodhaifi, A. Alghodhaifi, M. Alghodhaifi, Predicting invasive ductal carcinoma in breast histology images using convolutional neural network, in: 2019, pp. 374–378. 10.1109/NAECON46414.2019.9057822.

[22] J.W. Johnson, Detecting invasive ductal carcinoma with semi-supervised conditional gans, in: Proceedings of the Future Technologies Conference (FTC) 2020, vol. 3, Springer, 2021, pp. 113–120.

[23] A. Cruz-Roa, A. Basavanhally, F. González, H. Gilmore, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, A. Madabhushi, Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks, in: M.N. Gurcan, A. Madabhushi (Eds.), San Diego, California, USA, 2014, p. 904103. 10.1117/12.2043872.

[24] A.M. Romano, A.A. Hernandez, Enhanced deep learning approach for predicting invasive ductal carcinoma from histopathology images, in: 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), 2019, pp. 142–148, 10.1109/ICAIBD.2019.8837044.

[25] F.P. Romero, A. Tang, S. Kadoury, Multi-level batch normalization in deep networks for invasive ductal carcinoma cell discrimination in histopathology images, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), 2019, pp. 1092–1095, 10.1109/ISBI.2019.8759410.

[26] E.D. Cubuk, B. Zoph, J. Shlens, Q.V. Le, RandAugment: Practical automated data augmentation with a reduced search space, in: arXiv, 2019. 10.48550/arXiv.1909.13719.

[27] P. Chen, S. Liu, H. Zhao, J. Jia, GridMask Data Augmentation, 2020, 10.48550/arXiv.2001.04086.

[28] E.D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, Q.V. Le, AutoAugment: Learning Augmentation Policies from Data, 2019, 10.48550/arXiv.1805.09501.

[29] O. El Ogri, H. Karmouni, M. Sayyouri, H. Qjidaa, 3D image recognition using new set of fractional-order Legendre moments and deep neural networks, Signal Process. Image Commun. 98 (2021), 116410.

[30] H. Karmouni, M. Sayyouri, H. Qjidaa, A novel image encryption method based on fractional discrete Meixner moments, Opt. Lasers Eng. 137 (2021), 106346.

[31] C.M. Bishop, Training with noise is equivalent to Tikhonov regularization, Neural Comput. 7 (1995) 108–116, https://doi.org/10.1162/neco.1995.7.1.108.

[32] S. Rajput, Z. Feng, Z. Charles, P.-L. Loh, D. Papailiopoulos, Does data augmentation lead to positive margin? arXiv (2019) https://doi.org/10.48550/arXiv.1905.03177.

[33] S. Wu, H. Zhang, G. Valiant, C. Ré, On the generalization effects of linear transformations in data augmentation, in: International Conference on Machine Learning, 2020, pp. 10410–10420.

[34] S. Yang, Y. Dong, R. Ward, I.S. Dhillon, S. Sanghavi, Q. Lei, Sample Efficiency of Data Augmentation Consistency Regularization, 2022. 10.48550/arXiv.2202.12230.

[35] Z. He, L. Xie, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Data augmentation revisited: Rethinking the distribution gap between clean and augmented data, ArXiv Preprint ArXiv:1909.09148 (2019).

[36] Q. Xie, Z. Dai, E. Hovy, T. Luong, Q. Le, Unsupervised data augmentation for consistency training, Adv. Neural Inf. Proces. Syst. 33 (2020) 6256–6268.

[37] L. Li, M. Spratling, Data augmentation alone can improve adversarial training, ArXiv Preprint ArXiv:2301.09879. (2023).

[38] S.-A. Rebuffi, S. Gowal, D.A. Calian, F. Stimberg, O. Wiles, T.A. Mann, Data augmentation can improve robustness, Adv. Neural Inf. Proces. Syst. 34 (2021) 29935–29948.

[39] R. Shen, S. Bubeck, S. Gunasekar, Data Augmentation as Feature Manipulation, in: Proceedings of the 39th International Conference on Machine Learning, PMLR, 2022, pp. 19773–19808. https://proceedings.mlr.press/v162/shen22a.html (accessed January 1, 2023).

[40] .C. Wong, A. Gatt, V. Stamatescu, M.D. McDonnell, Understanding Data Augmentation for Classification: When to Warp? in: 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2016, pp. 1–6, 10.1109/DICTA.2016.7797091.

[41] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (2019) 60, https://doi.org/10.1186/s40537-019-0197-0.

[42] B. Hanin, Y. Sun, How data augmentation affects optimization for linear regression, Adv. Neural Inf. Proces. Syst. 34 (2021) 8095–8105.

[43] Z. Allen-Zhu, Y. Li, Feature purification: How adversarial training performs robust deep learning, in: 2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS), IEEE, 2022, pp. 977–988.

[44] Marsh, Part 1 : Breast Cancer Analyzer + Web App, 2018, https://kaggle.com/vbookshelf/part-1-breast-cancer-analyzer-web-app (accessed June 16, 2022).

[45] J. Chen, D. Zhang, M. Suzauddola, A. Zeb, Identifying crop diseases using attention embedded MobileNet-V2 model, Appl. Soft Comput. 113 (2021), 107901, https://doi.org/10.1016/j.asoc.2021.107901.

[46] J. Chen, A. Zeb, S. Yang, D. Zhang, Y.A. Nanehkaran, Automatic identification of commodity label images using lightweight attention network, Neural Comput. Appl. 33 (2021) 14413–14428, https://doi.org/10.1007/s00521-021-06081-9.

[47] M. Ghoncheh, Z. Pournamdar, H. Salehiniya, Incidence and mortality and epidemiology of breast cancer in the world, Asian Pac. J. Cancer Prev. 17 (2016) 43–46, https://doi.org/10.7314/APJCP.2016.17.S3.43.

[48] J.L. Kelsey, L. Bernstein, Epidemiology and prevention of breast cancer, Annu. Rev. Public Health 17 (1996) 47–67, https://doi.org/10.1146/annurev.pu.17.050196.000403.

[49] A. Janowczyk, A. Madabhushi, Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases, J. Pathol. Inform. 7 (2016) 29, https://doi.org/10.4103/2153-3539.186902.

[50] H.X. Bai, B. Hsieh, Z. Xiong, K. Halsey, J.W. Choi, T.M.L. Tran, I. Pan, L.-B. Shi, D.-C. Wang, J.i. Mei, X.-L. Jiang, Q.-H. Zeng, T.K. Egglin, P.-F. Hu, S. Agarwal, F.-F. Xie, S. Li, T. Healey, M.K. Atalay, W.-H. Liao, Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT, Radiology 296 (2) (2020) E46–E54.

[51] T. Franquet, Imaging of pneumonia: trends and algorithms, Eur. Respir. J. 18 (1) (2001) 196–208.

[52] J. Vilar, M.L. Domingo, C. Soto, J. Cogollos, Radiology of bacterial pneumonia, Eur. J. Radiol. 51 (2) (2004) 102–113.

[53] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, 2015.

[54] X. Xie, Z. Zhong, W. Zhao, C. Zheng, F. Wang, J. Liu, Chest CT for typical coronavirus disease 2019 (COVID-19) pneumonia: relationship to negative RT-PCR testing, Radiolog. 296 (2020) E41–E45.
[55] K. He, X. Zhang, S. Ren, J. Sun, in: Deep Residual Learning for Image Recognition, IEEE, Las Vegas, NV, USA, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.
[56] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[57] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.
[58] A. Aggarwal, M. Mittal, G. Battineni, Generative adversarial network: An overview of theory and applications, International Journal of Information Management Data Insights. 1 (2021), 100004.
[59] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A.A. Bharath, Generative adversarial networks: an overview, IEEE Signal Process. Mag. 35 (1) (2018) 53–65.
[60] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, F.-Y. Wang, Generative adversarial networks: introduction and outlook, IEEE/CAA J. Autom. Sin. 4 (4) (2017) 588–598.
[61] J. Choi, T. Kim, C. Kim, Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6830–6840.
[62] S.-W. Huang, C.-T. Lin, S.-P. Chen, Y.-Y. Wu, P.-H. Hsu, S.-H. Lai, Auggan: Cross domain adaptation with gan-based data augmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 718–731.
[63] C. Deotte, Dog Memorizer GAN, 2019, https://kaggle.com/code/zonghangan/dog-memorizer-gan (accessed October 5, 2022).
[64] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genom. 21 (2020) 6, https://doi.org/10.1186/s12864-019-6413-7.
[65] D. Hand, P. Christen, A note on using the F-measure for evaluating record linkage algorithms, Stat. Comput. 28 (3) (2018) 539–547.
[66] P. Heidke, Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst, Geogr. Ann. 8 (1926) 301–349, https://doi.org/10.2307/519729.
[67] Z. Ping, Y. Xia, T. Shen, V. Parekh, G.P. Siegal, I.-E. Eltoum, J. He, D. Chen, M. Deng, R. Xi, A microscopic landscape of the invasive breast cancer genome, Sci. Rep. 6 (2016) 27545.
[68] M.A. Lopez-Garcia, F.C. Geyer, M. Lacroix-Triki, C. Marchió, J.S. Reis-Filho, Breast cancer precursors revisited: molecular features and progression pathways, Histopathology 57 (2010) 171–192.
[69] R.G. do Nascimento, K.M. Otoni, Histological and molecular classification of breast cancer: what do we know, Mastology 30 (2020) e20200024.
[70] A.J. Cozzo, A.M. Fuller, L. Makowski, Contribution of adipose tissue to development of cancer, Compr. Physiol. 8 (2017) 237.
[71] A. Merii, Breast Cancer Classification Guide PCA & SVMs | Kaggle, 2020a, https://www.kaggle.com/code/amerii/breast-cancer-classification-guide-pca-svms (accessed October 31, 2022).
[72] A. Merii, Breast Cancer Classification End to End, 2020b, https://kaggle.com/code/amerii/breast-cancer-classification-end-to-end (accessed December 4, 2022).
[73] S. Sammari, Breast Cancer images Classification, 2021, https://kaggle.com/code/midouazerty/breast-cancer-images-classification (accessed November 20, 2022).
[74] A. Verma, Breast Cancer Detection VGG16, 2022, https://kaggle.com/code/ayushv322/breast-cancer-detection-vgg16 (accessed December 4, 2022).
[75] S. Zagoruyko, N. Komodakis, Wide residual networks, ArXiv Preprint ArXiv: 1605.07146 (2016).
[76] C. Zhang, F. Rameau, S. Lee, J. Kim, P. Benz, D.M. Argaw, J.-C. Bazin, I.S. Kweon, Revisiting Residual Networks with Nonlinear Shortcuts., in: BMVC, 2019, p. 12.
[77] C. Zhang, P. Benz, D.M. Argaw, S. Lee, J. Kim, F. Rameau, J.-C. Bazin, I.S. Kweon, Resnet or densenet? introducing dense shortcuts to resnet, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3550–3559.
[78] G. Philipp, D. Song, J.G. Carbonell, Gradients explode-deep networks are shallow-resnet explained, 2018.
[79] S. Xie, R. Girshick, P. Dollar, Z. Tu, K. He, Aggregated Residual Transformations for Deep Neural Networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, 2017, pp. 5987–5995, https://doi.org/10.1109/CVPR.2017.634.
[80] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2012, https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (accessed February 19, 2023).
[81] S. Kornblith, J. Shlens, Q.V. Le, Do Better ImageNet Models Transfer Better? IEEE Computer Society, in, 2019, pp. 2656–2666.
[82] J. Yoo, N. Ahn, K.-A. Sohn, Rethinking Data Augmentation for Image Super-resolution: A Comprehensive Analysis and a New Strategy, in: arXiv, 2020. 10.48550/arXiv.2004.00448.
[83] S. Zheng, Y. Song, T. Leung, I. Goodfellow, Improving the robustness of deep neural networks via stability training, in: Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition, 2016, pp. 4480–4488.
[84] N. Romero, M. Gutoski, L. Hattori, H.S. Lopes, The Effect of Data Augmentation on the Performance of Convolutional Neural Networks, in: ABRICOM, 2017, pp. 1–12, 10.21528/CBIC2017-51.
[85] H. Noh, T. You, J. Mun, B. Han, Regularizing deep neural networks by noise: its interpretation and optimization, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 5115–5124.
[86] M. Luo, J. Cao, X. Ma, X. Zhang, R. He, FA-GAN: face augmentation GAN for deformation-invariant face recognition, IEEE Trans. Inf. Forensics Secur. 16 (2021) 2341–2355, https://doi.org/10.1109/TIFS.2021.3053460.