Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 6, No. 4, August, 2022

Brightness as an Augmentation Technique for Image Classification

Ibrahem Kandel [1*], Mauro Castelli [1], Luca Manzoni [2]

[1] *Nova Information Management School (NOVA IMS), Campus de Campolide, Universidade Nova de Lisboa, 1070-312 Lisboa, Portugal.*

[2] *Dipartimento di Matematica e Geoscienze, Università degli Studi di Trieste, Via Alfonso Valerio 12/1, 34127 Trieste, Italy.*

Abstract

Augmentation techniques are crucial for accurately training convolution neural networks (CNNs). Therefore, these techniques have become the preprocessing methods. However, not every augmentation technique can be beneficial, especially those that change the image's underlying structure, such as color augmentation techniques. In this study, the effect of eight brightness scales was investigated in the task of classifying a large histopathology dataset. Four state-of-the-art CNNs were used to assess each scale's performance. The use of brightness was not beneficial in all the experiments. Among the different brightness scales, the [0.75–1.00] scale, which closely resembles the original brightness of the images, resulted in the best performance. The use of geometric augmentation yielded better performance than any brightness scale. Moreover, the results indicate that training the CNN without applying any augmentation techniques led to better results than considering brightness augmentation. Therefore, experimental results support the hypothesis that brightness augmentation techniques are not beneficial for image classification using deep-learning models and do not yield any performance gain. Furthermore, brightness augmentation techniques can significantly degrade the model's performance when they are applied with extreme values.

Keywords:

Image Classification;
Deep Learning;
Medical Images;
Augmentation Techniques;
Supervised Learning.

Article History:

Received:	13	December	2021
Revised:	09	May	2022
Accepted:	24	May	2022
Available online:	31	May	2022

1- Introduction

With the global increase in cancer prevalence [1], the workload has increased extensively for pathologists, as timely and accurate classification of histopathological slides is necessary. However, the workforce has not increased in the last few years; rather, it has started to decrease rapidly [2–4]. An accurate and automatic classification method will decrease the time needed for and increase the accuracy of classification by pathologists [5–7]. Automatic classification has undergone significant progress since 2012, when deep-learning methods became de facto in the computer vision domain [8]. Many studies have already demonstrated the ability of deep-learning methods to classify histopathology slides [9-11]. One of the main strengths of deep-learning models is that they can be incorporated into the classification process either in addition to the electronic microscope itself or as stand-alone software. However, several challenges still limit the progress of deep-learning methods in the histopathology domain [12]. The models are usually trained on thousands of images, which are insufficient to train a deep-learning model accurately. In contrast, deep-learning models typically require millions of images, which are very scarce in the medical field and the histopathology domain.

One way to tackle the problem of the datasets' size is through augmentation techniques [13]. Image augmentation is a robust technique that has proved its effectiveness in computer vision models. However, not all augmentation techniques can produce a favorable result, and sometimes they can degrade the model's performance significantly. Augmentation techniques are not only used to increase a dataset's size but also as a starting point of the algorithm, i.e.,

* **CONTACT**: d20181143@novaims.unl.pt

DOI: http://dx.doi.org/10.28991/ESJ-2022-06-04-015

they can be used to train the deep-learning models differently than the standard method, or "the supervised method," such as self-supervised learning [14]. In self-supervised learning, the augmentation techniques will create several versions of the same image to construct a positive pair and then create other versions of the images to construct a negative pair. Then, the model will be trained to differentiate between these two classes. Self-supervised learning is essential in problems characterized by the scarcity of large datasets [14, 15]. Many augmentation techniques have emerged in recent years, which can be classified into two main categories: geometric augmentations and synthetic augmentations. Geometric augmentations are a class of techniques involving the cropping, zooming, and shifting of original images. Synthetic augmentations are techniques that introduce artificially made changes to the original images. One of the main methods in this category is generative adversarial networks (GANs).

Brightness is an augmentation method that cannot be easily assigned to the previous categories because it is neither a geometric nor a synthetic transformation. Brightness changes in the images will modify their underlying structure. Sometimes, when used with extreme values, they can change the images entirely, i.e., setting brightness to be very low will make the image black, and it will no longer represent the original content. Due to its ease of use and its logical explanation, many authors have used it when training deep-learning models, in which brightness has been used in a supervised manner and recently in a self-supervised manner. The use of brightness augmentation in the self-supervised algorithm can have a more severe effect on the classification because the entire algorithm will be based on the representations the brightness augmentation provides.

This paper, in which the aim is to understand the effect of brightness as an augmentation technique in the training process of deep-learning networks, complements existing research efforts dedicated to the analysis of color distortion techniques, geometric augmentation, the relevance of noise, and image quality in the context of deep-learning architectures. In more detail, many color distortion techniques have been studied in the literature as a means of augmentation. Chen [16] investigated the effect of five image enhancement algorithms on image classification performance. The enhancement techniques used were SMQT, CLAHE, Gamma, wavelet, and Laplace. The author used two datasets to conduct the experiments, a black and white X-ray images dataset, and a colored CatsVsDogs dataset. In addition, the author used a LeNet convolutional neural network (CNN). The results showed that these five image enhancement techniques had similar performance across the two datasets. Sometimes, the techniques would produce a poorer performance than the baseline model. It is worth noting that the performance with the colored dataset was similar to that with the black and white dataset. Rodríguez et al. [17] studied the effects of five noise distortions on images using two brightness ranges: the original brightness and 0.5 brightness (half the brightness of each image). The authors used the following noise sources: Poisson, Gaussian, salt and pepper, speckle, and uniform. The authors considered six CNNs to conduct their experiments: ResNet, DenseNet, InceptionV3, MobileNet, NASNet, and WideResNet. They selected 1000 images from the ILSRVC 2012 dataset and stated that the noise degraded all the CNNs' performance. Another important observation was that the performance with the 0.5 brightness level was always less than that with the original brightness level.

Taylor and Nitschke [18] compared the performance of six augmentation techniques, lipping, rotating, cropping, color jittering, edge enhancement, and fancy PCA. The authors used a custom-made CNN that was inspired by Dodge & Karam [19] and considered the Caltech101 dataset for their experiments. The best augmentation technique was the cropping technique, which increased the model's performance by 14% compared to the baseline. Additionally, the color jittering technique had a similar performance as the baseline without a noticeable difference. Dodge & Karam [19] and Nazaré et al. [20] studied noise's impact on the image classification process. Dodge & Karam [19] studied the effect of five noise types: blur, noise, JPEG, contrast, and JPEG2000. They used four CNNs, and they showed that CNNs are very prone to noise and that any noise presence can degrade the classification process's performance. Nazaré et al. [20] reached a similar conclusion, suggesting that noisy images can degrade CNNs' performance and that the images' quality is crucial. Haque et al. [21] trained an InceptionV3 model to classify maize crop leaves to detect healthy leaves. They noticed that the brightness in the dataset was not uniform because the dataset was taken with on-file and not in-lab-controlled settings. The authors trained the model using four brightness ranges [1.25, 1.5, 1.75, 2.0], and they reported that the model trained using brightness augmentation achieved slightly better performance than the model trained using rotation and color distortion, with a loss score of 0.1787 compared to 0.1861.

As noted in the literature, brightness is very popular due to its ease of implementation and logical explanation. However, the use of brightness can change an image's underlying structure, thereby negatively affecting the CNN models' ability to classify images. This study investigated the brightness technique in detail and compared it to geometric techniques and training without any augmentation. Eight brightness scales were used and their effects were analyzed. The scales range from complete darkness [0-0.25] to double the initial brightness [1.75-2.0]. A large colored histopathology image dataset with more than 250,000 images was used to train, validate, and test the considered models and to investigate the effects of brightness augmentation fully. To quantify the effect of brightness scales better, four state-of-the-art CNNs were considered: two inception-based CNNs, InceptionV3 and Xception networks, and two residual connection-based CNNs, ResNet50 and DenseNet121 networks. Four evaluation metrics were used to evaluate

the obtained results: accuracy, kappa, AUC, and recall. Each experiment was repeated 30 times to calculate the confidence interval and examine each setting's stability and consistency.

The rest of the paper is organized as follows: Section 2 discusses the methodology used. Section 3 presents the experimental settings and the results achieved. Section 4 discusses the results and compares them to various state-of-the-art results. Finally, Section 5 concludes the paper and suggests future research directions.

2- Research Methodology

2-1- CNN Architectures

CNNs were introduced to address the problem of the spatial nature of images [22-24]. They successfully addressed various computer vision problems, such as segmentation, detection, and classification. CNNs have been used in various domains, such as agriculture, industry, and medicine. The main idea of using CNNs is to apply a convolution filter to each image and extract various features in a cascading manner that will be used to classify the images. The convolution operation can be formally defined as in Equation 1.

$$O[i,j] = F(u,v) * I(i,j) = \sum_u \sum_v \sum_{c\in\{R,G,B\}} F_c(u,v) \odot I_c(i+u, j+v) \quad (1)$$

where $I(.)$ is the input image, c is the color channels, $F(.)$ is the kernel, and $O(i,j)$ is the output pixel in the (i,j) position.

Multiple architectures were introduced to address various problems in computer vision. One of the first designs was the block design introduced by [23, 24], in which multiple convolution layers are stacked to create a convolution block. A pooling layer and a normalization layer separate the blocks. Szegedy et al. [25] designed a new Inception network with multiple convolution layers connected in parallel to address various aspect ratios in the same image. Chollet [26] introduced a novel architecture called Xception, inspired by the Inception network with some changes, such as the use of point-wise convolution. He et al. [27] introduced a novel architecture called ResNet, in which the author stated that after a certain depth, the CNN would experience the problem of vanishing gradients. To solve this problem, the authors introduced the residual connection, in which a connection will be made from later layers to proceeding layers to solve the problem of vanishing gradients. Finally, Huang et al. [28] introduced a novel architecture inspired by ResNet architecture, in which residual connections to all the layers are used. In this study, four CNNs were used: the InceptionV3, its update of the Xception network, the ResNet network, and its update of the DenseNet network. Using these four architectures, the objective is to study brightness's effect on various designs to generalize its effect. Below is a brief description of each network used in this study.

2-1-1- Inception Block

InceptionV3 architecture [25] was introduced to address the problem of sparse structure in CNNs. First, the author [25] introduced a novel connection between convolution layers, which is called the inception module. The convolution layers are connected in a parallel manner; then these layers' output is concatenated together to form a single convolution block. The following kernels were used in each inception block: two 1×1 kernels, one 3×3 kernel, and one 5×5 kernel. To reduce the computational power and increase the network's efficiency, a 1×1 kernel was used before the 3×3 kernel and the 5×5 kernel. Next, Chollet introduced the Xception architecture [26]. He modified the inception module, in which a point-wise convolution and separable convolutions follow a depth-wise convolution. Also, he noted that the intermediate activation function would degrade the network's performance, so he removed it. For more details, the reader is referred to the corresponding papers [25, 26].

2-1-2- Residual Connection

ResNet architecture [27] was introduced to address the problem of vanishing gradients faced when increasing a CNN's depth. The authors noted that adding a residual connection (skip connection) can prevent gradients from going to minimal values, or "vanishing." The idea of the residual connection, which was named the identity shortcut connection, is that by skipping some layers, the gradients will not follow the usual route during the backpropagation. To mitigate some drawbacks that occurred due to the ResNet network's identity shortcut connection, the DenseNet architecture [28] was introduced to update the ResNet network. One of the main differences between ResNet and DenseNet is that in the DenseNet network, the use of concatenation instead of the summation operation, as in the ResNet, can protect the features [29]. Another difference is the connection in DenseNet of each layer to its subsequent layers so that every layer will have an input of all the previous layers to maximize parameter reusability. In other words, any

important features learned by any layer are shared with all the networks through dense connections. For more details, the reader is referred to the corresponding studies [27, 28].

2-2- Brightness Range

Image augmentation techniques are usually used to increase the efficiency of the feature extraction operation. Image augmentation entails providing various iterations of the exact image to the classifier, C_i, as Equation 2 shows.

$$C\left(\sum_{j=1}^{n} x_{i_j}\right) = \tilde{x}_i \tag{2}$$

where n refers to the number of augmentation iterations used. Many forms of augmentation have been introduced in the literature, including geometric augmentation methods, such as transformation, zooming, and brightness. The most commonly used augmentation technique is geometric transformation. Many authors have stated that geometric augmentation can provide very accurate features from the image. As Figures 1 and 2 show, the histogram of the geometric augmentation techniques was approximately similar to the original histogram; however, with the use of the brightness augmentation technique, the histogram is very different from that of the original image, which may indicate that using brightness can confuse the classifier, C_i, and that features from the images will therefore be extracted incorrectly. To better quantify and measure the effectiveness of the brightness on the images, eight ranges were constructed, ranging from 0 (complete darkness) to 2 (twice the brightness of the original image), and including 1 (the original brightness). The variable γ is the brightness factor, which usually ranges from $[0, 2]$, where 0 indicates complete blackness, 1 indicates the original brightness, and 2 indicates double the original brightness. The variable γ was randomly selected from a range in Keras and TensorFlow, calculated using Equation 3. In our study, the augmented image, based on the brightness, was calculated using Equation 4. Figures 1 and 2 present brightness's effects.

$$\gamma_{Selected} = Rand(\gamma_{min}, \gamma_{max}) \tag{3}$$

$$\tilde{x}_i = \gamma * \sum_{R=1}^{n} \sum_{G=1}^{n} \sum_{B=1}^{n} (x_i) \tag{4}$$

where x_i is the original image, R is the image's red channel, G is the image's green channel, B is the image's blue channel, γ is the brightness factor, $\tilde{x}_i$ is the augmented image, and n is the number of pixels on the image x_i.

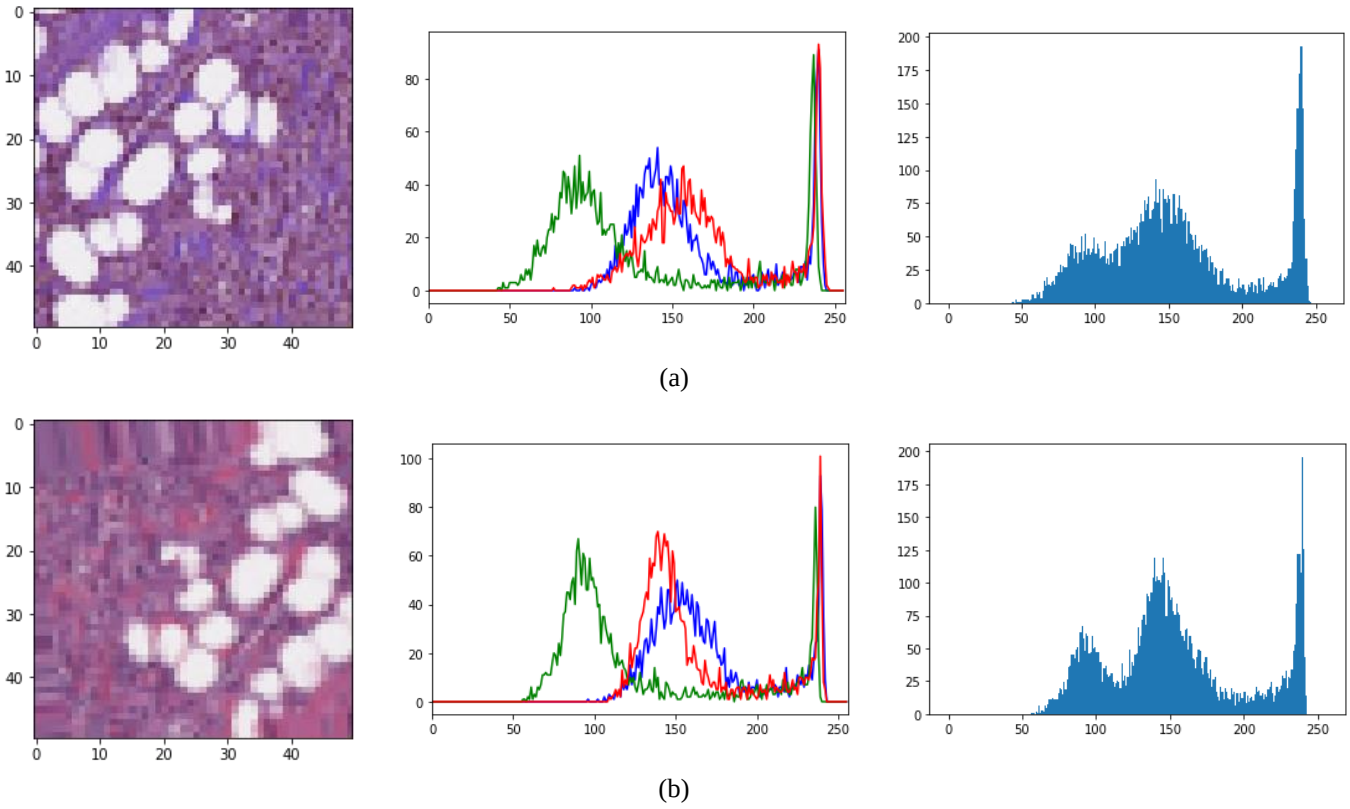

Figure 1. Geometric augmentation's effect on the image representation: (a) the original image without any changes and the RGB histogram and the overall histogram of the image, (b) the geometric augmentation's effect on the same image

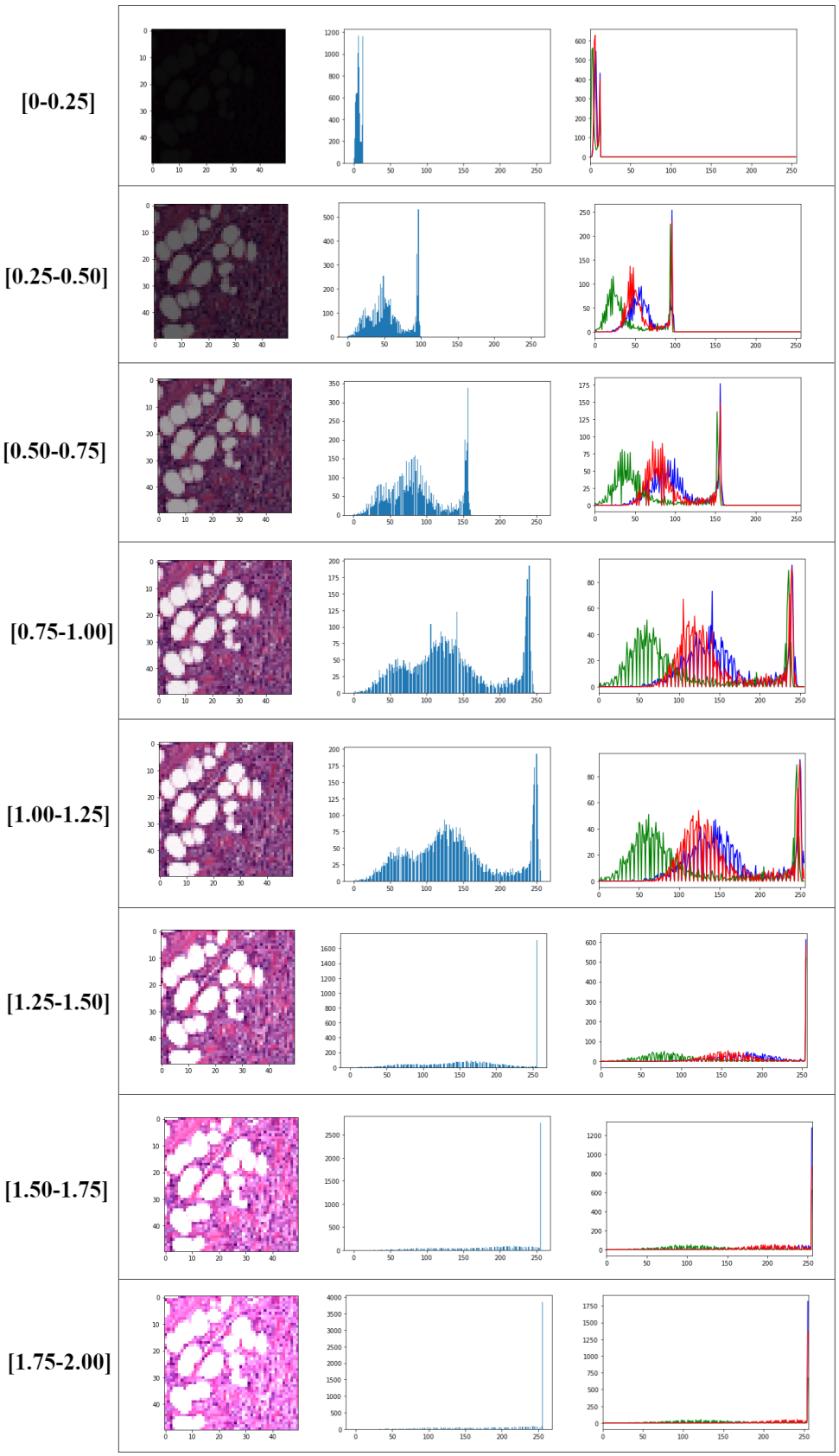

Figure 2. The effect of various brightness augmentations on the image representation

2-3- Dataset

This study considers an invasive ductal carcinoma dataset [30, 31]. The dataset contains 277,524 images with a size of 50×50 $pixels$. However, the original images were too small to be used for the CNN, so the images were rescaled to 75×75 $pixels$. The images were extracted from 162 whole-slide images scanned at $40 \times$ zoom. The dataset consists of 71% negative-class and 29% positive-class images. Figure 3 presents a sample of the dataset.

Figure 3. A sample of the ICD dataset

2-4- Evaluation Metrics

Evaluation of CNNs is crucial to estimate their performance with future and unseen datasets. Therefore, four metrics are considered for the purpose of comparison. Each one has its strength, and this process allows us to give a holistic overview of each network's performance. Below is a brief description of each metric used.

2-4-1- Accuracy

Accuracy is the classifier's overall performance. It can describe the classifier's ability to distinguish true labels. However, one main drawback of using accuracy occurs in cases of imbalance, in which the positive and negative classes are not equally represented. Equation 5 formally defines the accuracy.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{5}$$

where TP is the number of true positive labels being truly classified, TN is the number of true negative labels being truly classified, FP is the number of falsely positive labels, and FN is the number of falsely negative labels.

2-4-2- Kappa

Cohen's kappa [32] can be beneficial in evaluating imbalanced datasets. It measures the agreement/disagreement between the ground truth labels and the CNN's prediction. Kappa ranges in $[-1, +1]$, where -1 indicates random choice and $+1$ indicates a perfect classifier. *Kappa* is defined as reported in Equation 6.

$$Kappa = \frac{Aggrement_{Observed} - Aggrement_{Expected}}{1 - Aggrement_{Expected}} \tag{6}$$

$$Aggrement_{Observed} = \frac{TP+TN}{TP+FP+TN+FN}$$

$$Aggrement_{Expected} = \left(\frac{TP+FP}{N} * \frac{TP+FN}{N}\right) + \left(\frac{TN+FP}{N} * \frac{TN+FN}{N}\right)$$

2-4-3- AUC of the ROC curve

The ROC curve is very robust in avoiding false classification. The area under the ROC curve (AUC) is usually used instead of visually plotting the ROC curve. The AUC can be used to summarize each classifier's performance. AUC ranges from [0.5-1], where 0 indicates a random choice and 1 indicates a perfect classifier. The ROC is formally defined in Equation 7.

$$ROC = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{7}$$

where TP is the number of true positive labels being truly classified, TN is the number of true negative labels being truly classified, FP is the number of falsely positive labels, and FN is the number of falsely negative labels.

2-4-4- Recall

The recall metric can accurately describe the classifier's ability to classify positive classes. The recall metric is formally defined in Equation 8.

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

where, TP is the number of true positive labels being truly classified and FN is the number of falsely negative labels.

3- Results and Discussion

The performance of three techniques has been compared. The first technique consists of training the network without any augmentation techniques to form a baseline. The second technique consists of training the network using four geometric augmentation techniques: right rotation, left rotation, shifting, and zooming. Finally, the third technique consists of training the network eight times using eight brightness techniques ranging from 0 to 2. The value of 0 brightness indicates complete darkness, 1 indicates original brightness, and 2 indicates double the brightness. Figure 2 shows the eight brightness techniques used. Figure 4 shows a flowchart of the experiments.

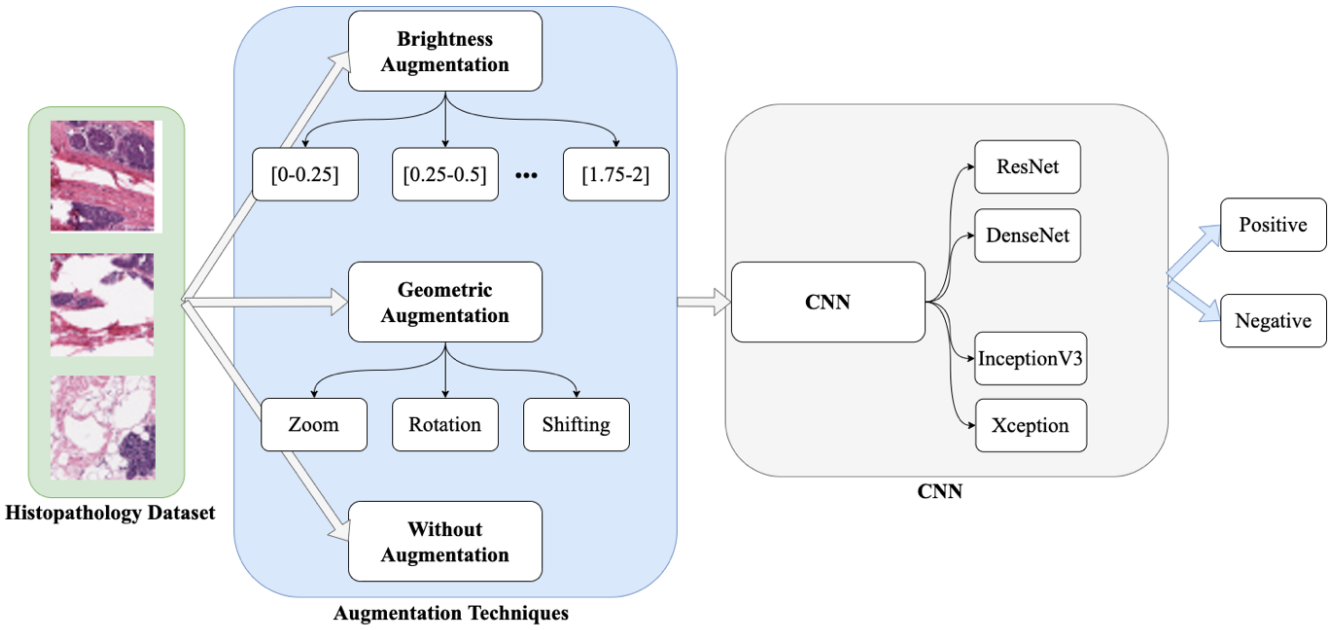

Figure 4. Flowchart of the experiments executed in this study

The dataset was divided into 80%/10%/10% for training/validation/testing. The hyperparameters used in this paper are as follows: the optimizer is the Adam optimizer [33], the batch size was equal to 32, and due to the dataset's size and the computational power available, an early stopping criterion of 10 epochs was considered. The image size was $75 \times 75\ pixels$. Instead of training the networks from scratch, the weights of the ImageNet dataset [34] were used to fine-tune the networks. The Keras package [35] with TensorFlow [36] as a backend was selected to train the models, and three Nvidia GPUs [37] were used for training: two NVIDIA TITAN and one Quadro GV100. Due to the CNN's stochastic nature, every experiment was repeated 30 times and then the average performance was calculated. Finally, the confidence interval with a 95% error rate was computed. A total of 1200 experiments have been performed (30 iterations × 10 techniques × 4 CNNs), and the total running time of all the experiments presented in this study was approximately 2400 hours.

In the first sets of experiments, the ResNet50 network was considered and four main evaluation metrics were used to measure each technique's performance. Table 2 shows the ResNet50 CNN's results. The results indicate that using brightness decreased the CNN's performance. The highest score any brightness technique achieved was the one with the brightness range of [0.75-1], where 1 indicates the original brightness. The lowest score was achieved using the brightness range of [0-0.25], where the brightness of 0 indicates black (complete darkness). Training the ResNet50 network without any augmentation technique led to better results than any brightness range. The highest score was achieved by training the network with geometric augmentation. This performance was better than that of any brightness range and higher than that of training the network without augmentation. Comparing brightness ranges, the performance of the three ranges between 0.5 and 1.25 was the best, which indicates that the ranges between the original brightness were the best, and the higher the distortion, the more confusing the network gets. The accuracy, kappa, and AUC metrics were consistent with each other, but for the recall metric, the score achieved with the brightness range of [0.5-0.75] was higher than that achieved with [0.75-1.00], by a close margin.

The confidence intervals (CI) can be used to measure the stability and consistency of each network's results, where high values indicate larger discrepancies between the results and low values indicate each time the network provided a close-by performance. The kappa metric and the recall metric reveal the largest discrepancies between the results. For the Kappa metric, the lowest brightness range [0-0.25] provided the highest CI, with a value of ±9.61% compared to only ±0.35% in the brightness range [0.75-1], which indicates that the lowest brightness range made the network very unstable, providing a different result each time. Comparing the recall metric to other metrics used, one can notice that the recall metric shows similar behavior. In particular, the lowest brightness range [0-0.25] is characterized by a CI of ±18.03%, and the brightness range [0.75-1] has a CI of only ±0.71%. Overall, for the ResNet50 network trained with geometric augmentation, the best results are characterized by the lowest CI, showing the training process's robustness. Table 1 shows the ResNet50 network's results.

Table 1. Results of ResNet50 network. (± 95% confidence interval over repeated trials)

Technique	Accuracy	Kappa	AUC	Recall
Without Augmentation	88.58% ±0.13%	0.7155 ±0.38%	85.24% ±0.32%	81.54% ±0.55%
With Geometric Augmentation	90.75% ±0.08%	**0.7690 ±0.28%**	87.81% ±0.32%	85.79% ±0.53%
Brightness Range [0.00-0.25]	75.91% ±1.89%	0.2336 ±9.61%	59.38% ±3.99%	70.55% ±18.03%
Brightness Range [0.25-0.50]	84.46% ±0.49%	0.5996 ±1.30%	78.56% ±0.77%	77.07% ±1.68%
Brightness Range [0.50-0.75]	87.02% ±0.14%	0.6752 ±0.40%	83.11% ±0.36%	79.12% ±0.73%
Brightness Range [0.75-1.00]	87.36% ±0.12%	**0.6881 ±0.35%**	84.23% ±0.36%	78.33% ±0.71%
Brightness Range [1.00-1.25]	86.48% ±0.14%	0.6629 ±0.46%	82.62% ±0.36%	77.73% ±0.33%
Brightness Range [1.25-1.50]	83.79% ±0.37%	0.5752 ±1.45%	76.95% ±1.02%	77.41% ±1.14%
Brightness Range [1.50-1.75]	78.82% ±0.97%	0.4122 ±3.37%	68.27% ±1.73%	70.76% ±2.08%
Brightness Range [1.75-2.00]	74.83% ±0.95%	0.2594 ±5.12%	60.98% ±2.48%	62.64% ±2.53%

In the second set of experiments, the DenseNet121 network was used. Comparing brightness ranges, the highest value was achieved with the brightness range of [0.50-0.75] and the lowest value with the brightness range of [1.75-2], keeping in mind that a brightness of 0 indicates complete darkness and 1 indicates the original brightness. Training the DenseNet121 network without augmentation produced better results than any brightness range. Also, in this set of experiments, the highest score was achieved by training the network with geometric augmentation. When comparing brightness ranges, the three ranges between 0.25 and 1 produced the best performance compared to other ranges. For the DenseNet121 network, the score of brightness range [0.50-0.75] led to better results than the ranges in which 1 is present, which indicates that slightly dimming the image led to better results; however, the use of brightness led to poorer results than training the network without it,

The accuracy and kappa metrics gave consistent results; however, the AUC and recall metrics were inconsistent. The highest AUC score was achieved by training the network with the brightness range [0.75-1.00], close to [0.50-0.75]. For the Kappa metric, the CI values of the highest brightness ranges were higher than those of the lowest brightness ranges, and the CI of the brightness range [0.50-0.75] was lower than the geometric augmentation, by a small margin. For the other metrics, the CI values of the highest brightness ranges were higher than those of the lowest brightness ranges. Overall, in the DenseNet121 network, training the network with geometric augmentation led to the best results and better consistency. Table 2 shows the results of the DenseNet121 network.

Table 2. Results of DenseNet121 network. (± 95% confidence interval over repeated trials)

Technique	Accuracy	Kappa	AUC	Recall
Without Augmentation	89.06% ±0.06%	0.7280 ±0.14%	85.92% ±0.12%	82.23% ±0.36%
With Geometric Augmentation	91.06% ±0.13%	**0.7781 ±0.40%**	88.48% ±0.37%	85.61% ±0.52%
Brightness Range [0.00-0.25]	85.26% ±0.19%	0.6184 ±0.80%	79.32% ±0.64%	79.21% ±0.81%
Brightness Range [0.25-0.50]	87.02% ±0.16%	0.6710 ±0.41%	82.49% ±0.34%	80.47% ±0.84%
Brightness Range [0.50-0.75]	88.40% ±0.06%	**0.7115 ±0.26%**	85.10% ±0.35%	81.05% ±0.77%
Brightness Range [0.75-1.00]	88.11% ±0.12%	0.7104 ±0.23%	85.80% ±0.25%	78.43% ±0.72%
Brightness Range [1.00-1.25]	86.98% ±0.10%	0.6683 ±0.44%	82.20% ±0.47%	80.95% ±0.87%
Brightness Range [1.25-1.50]	83.66% ±0.55%	0.5640 ±2.11%	75.98% ±1.36%	79.08% ±0.94%
Brightness Range [1.50-1.75]	81.96% ±0.69%	0.5220 ±2.87%	74.28% ±1.91%	74.48% ±2.19%
Brightness Range [1.75-2.00]	79.95% ±0.72%	0.4671 ±2.60%	71.64% ±1.74%	70.37% ±3.12%

In the third and fourth sets of experiments, the InceptionV3 and Xception networks were used, which both had similar performance. When comparing the brightness ranges of these two networks, the highest value was achieved using the brightness range of [0.75-1], and the lowest value was obtained with the brightness range of [0-0.25]. Training the two networks, InceptionV3 and Xception, without any augmentation technique led to better results than any brightness range. Consistently, concerning the previously considered networks, the highest score was obtained by training the networks with geometric augmentation. Comparing brightness ranges to each other, the performance of the three ranges between 0.5 and 1.25 led to the best results among the brightness ranges.

The three evaluation metrics, accuracy, kappa, and AUC, behaved similarly. Recall was different in that the highest performance among the brightness ranges was achieved in the range of [0.25-0.50], which was slightly higher than the [0.75-1] range; however, the score of the geometric augmentation was similar to that of the other metrics in that it was the highest score. For the CI, the highest value was obtained with the range [0-0.25], which means that this range

produced the most inconsistent results, and the lowest CI value was obtained with the InceptionV3 and Xception networks, which were trained without any augmentation, followed by the geometric augmentation and the [0.75-1] range, which indicates the results' consistency with these settings. Overall, for the InceptionV3 and Xception networks, training the networks with geometric augmentation led to the best results with better consistency. Tables 3 and 4 show the InceptionV3 and Xception networks' results.

Table 3. Results of InceptionV3 network (± 95% confidence interval over repeated trials)

Technique	Accuracy	Kappa	AUC	Recall
Without Augmentation	88.65% ±0.09%	0.7167 ±0.21%	85.22% ±0.23%	81.93% ±0.67%
With Geometric Augmentation	90.40% ±0.09%	**0.7594 ±0.27%**	87.19% ±0.25%	85.61% ±0.48%
Brightness Range [0.00-0.25]	75.93% ±2.88%	0.2341 ±15.06%	60.11% ±6.67%	53.31% ±26.38%
Brightness Range [0.25-0.50]	81.73% ±1.55%	0.4875 ±6.07%	71.63% ±3.36%	79.82% ±0.93%
Brightness Range [0.50-0.75]	86.10% ±0.36%	0.6492 ±1.19%	81.59% ±0.89%	78.22% ±0.86%
Brightness Range [0.75-1.00]	87.54% ±0.09%	**0.6968 ±0.26%**	85.14% ±0.25%	77.36% ±0.43%
Brightness Range [1.00-1.25]	86.68% ±0.10%	0.6673 ±0.36%	82.78% ±0.34%	78.30% ±0.57%
Brightness Range [1.25-1.50]	82.78% ±0.40%	0.5764 ±1.42%	78.79% ±1.10%	69.94% ±0.93%
Brightness Range [1.50-1.75]	77.81% ±1.08%	0.4447 ±3.35%	71.89% ±2.02%	61.85% ±2.09%
Brightness Range [1.75-2.00]	74.13% ±1.54%	0.3024 ±4.47%	63.81% ±2.19%	56.69% ±3.76%

Table 4. Results of Xception network (± 95% confidence interval over repeated trials)

Technique	Accuracy	Kappa	AUC	Recall
Without Augmentation	89.07% ±0.06%	0.7290 ±0.11%	86.06% ±0.18%	82.00% ±0.56%
With Geometric Augmentation	91.06% ±0.06%	**0.7763 ±0.16%**	88.08% ±0.17%	86.62% ±0.45%
Brightness Range [0.00-0.25]	74.93% ±1.38%	0.2112 ±7.48%	58.49% ±3.13%	70.10% ±6.50%
Brightness Range [0.25-0.50]	84.04% ±0.81%	0.5770 ±2.77%	76.79% ±1.69%	79.11% ±1.48%
Brightness Range [0.50-0.75]	86.59% ±0.40%	0.6628 ±1.28%	82.36% ±0.91%	78.79% ±0.78%
Brightness Range [0.75-1.00]	88.05% ±0.08%	**0.7098 ±0.21%**	85.86% ±0.24%	78.07% ±0.50%
Brightness Range [1.00-1.25]	87.15% ±0.18%	0.6840 ±0.37%	84.13% ±0.44%	77.73% ±1.12%
Brightness Range [1.25-1.50]	82.57% ±0.83%	0.5418 ±2.85%	75.38% ±1.77%	74.83% ±1.66%
Brightness Range [1.50-1.75]	76.92% ±0.94%	0.3438 ±3.37%	64.88% ±1.63%	67.54% ±2.98%
Brightness Range [1.75-2.00]	75.00% ±1.13%	0.2814 ±5.91%	62.23% ±2.82%	61.05% ±2.93%

Conducting 30 repeated trials was computationally expensive. However, it provided a clear indication of the performance in each experiment. A result consistent with the literature [13] was obtained by using geometric augmentation techniques without brightness. This provided better results than training the network without data augmentation. However, it is not easy to compare the results obtained in this paper to the ones published in the existing literature because the authors usually use brightness, among other techniques, without analyzing its effect. For example, in Choi et al. [38], the authors used the brightness range ±10% without stating its effect. Therefore, it is unclear whether modifying the brightness is beneficial for the task considered. Similarly, Hermsen et al. [39], Kitamura et al. [40], and Berral-Soler [41] used brightness, among other color noise augmentation techniques. However, the authors did not compare the results without the use of these techniques, which could have been higher than the stated results. Moreover, they did not analyze each augmentation technique's effects, thus making it impossible to determine whether brightness provides a solution to the problem in the images they studied. Perez et al. [42] compared augmentation techniques, including brightness; however, they did not separate the brightness but combined it with saturation and contrast or saturation, contrast, and hue. The authors stated that these two groups performed severely worse than other geometric techniques, which coincided with our findings. Therefore, it is possible to state that brightness augmentation techniques are not beneficial for deep-learning models and will not produce any performance gain. Based on experimental evidence, brightness augmentation can significantly degrade a model's performance. Therefore, researchers should be very careful when using brightness augmentation techniques and should test the model with and without them to ensure they will not degrade the model's performance. Additionally, researchers are encouraged to publish the results achieved by using only the brightness to determine its effect. Haque et al. [21] compared a model that was trained on rotation, distortion, and flipping to a model that was trained on brightness. The authors reported that the brightness-augmented model achieved slightly better results. However, when comparing these two models, they did not present brightness's effects, as the original model was trained on color distortion. Therefore, in this case as well, it is not possible to conclude that brightness is beneficial for deep-learning models because the authors did not discuss the baseline model's performance.

4- Conclusion

The use of augmentation transcends the need to enhance CNNs' performance. Today, augmentation techniques are used for self-supervised learning as the primary method to create data sources. Therefore, the study of augmentation techniques is now crucial. In fact, if the data source is biased, the models trained on such data will also be biased. Although image brightness has been frequently mentioned in the literature, it has not been studied thoroughly to assess its effectiveness and understand its effects on the performance of models trained using images produced with such an augmentation method. In this study, brightness's effect on CNNs' performance in classifying histopathologic images was investigated. In more detail, a colored histopathology image dataset with more than 250,000 images was considered to train, validate, and test our models. Four state-of-the-art CNNs were used (ResNet50, DenseNet, InceptionV3, and Xception), and three main experiments were performed. In the first experiment, the four CNNs were trained without any image augmentation techniques. In the second experiment, the CNNs were trained by only using geometric augmentation techniques, including horizontal shifting, vertical shifting, and zooming. Finally, in the third experiment, the CNNs were trained by considering eight brightness methods. Experimental results demonstrated that the ResNet network's classification performance was sensitive to small changes in brightness up to the point of non-convergence, as happened in the range of [0-0.25]. The DenseNet network produced superior performance compared to the ResNet network, and the Xception network was superior to the InceptionV3 network. However, the best performance was achieved by all the considered architectures without relying on brightness augmentation techniques. Additionally, experimental results suggest that across the considered brightness scales, the best results were obtained when the level of brightness was close to that of the original images. Therefore, there is clear empirical evidence suggesting that considering brightness modification among the augmentation methods is detrimental to deep-learning architectures' performance. These findings are relevant, and they highlight the need to analyze the effect of brightness augmentation separately before considering its use. This is an important suggestion, especially considering the existing literature in which brightness augmentation is used and analyzed in conjunction with other augmentation methods. Our results show that brightness augmentation techniques are not beneficial for image classification using deep-learning models and will not produce any performance gain. Furthermore, they can significantly degrade a model's performance when set to extreme values.

5- Declarations

5-1- Author Contributions

Conceptualization, I.K., and M.C.; methodology, I.K. and M.C.; software, I.K.; formal analysis, I.K., M.C., and L.M.; writing—original draft preparation, I.K., M.C., and L.M.; writing—review and editing, I.K., M.C., and L.M.; supervision, M.C., and L.M. All authors have read and agreed to the published version of the manuscript.

5-2- Data Availability Statement

Data used in this work are publicly available and can be downloaded from: *https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images*.

5-3- Funding

This work was supported by national funds through the FCT (Fundação para a Ciência e a Tecnologia) by the projects GADgET (DSAIPA/DS/0022/2018).

5-4- Ethical Approval

Ethical approval was not requested as no experimental procedure was applied.

5-5- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

6- References

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

[2] Metter, D. M., Colgan, T. J., Leung, S. T., Timmons, C. F., & Park, J. Y. (2019). Trends in the us and canadian pathologistworkforces from 2007 to 2017. JAMA Network Open, 2(5), e194337. doi:10.1001/jamanetworkopen.2019.4337.

[3] Robboy, S. J., Gross, D., Park, J. Y., Kittrie, E., Crawford, J. M., Johnson, R. L., Cohen, M. B., Karcher, D. S., Hoffman, R. D., Smith, A. T., & Black-Schaffer, W. S. (2020). Reevaluation of the US Pathologist Workforce Size. JAMA Network Open, 3(7), e2010648. doi:10.1001/jamanetworkopen.2020.10648.

[4] Bonert, M., Zafar, U., Maung, R., El-Shinnawy, I., Kak, I., Cutz, J. C., Naqvi, A., Juergens, R. A., Finley, C., Salama, S., Major, P., & Kapoor, A. (2021). Evolution of anatomic pathology workload from 2011 to 2019 assessed in a regional hospital laboratory via 574,093 pathology reports. PLoS ONE, 16(6), e253876. doi:10.1371/journal.pone.0253876.

[5] Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-Van De Kaa, C., Bult, P., Van Ginneken, B., & Van Der Laak, J. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific Reports, 6(1), 26286. doi:10.1038/srep26286.

[6] Kiani, A., Uyumazturk, B., Rajpurkar, P., Wang, A., Gao, R., Jones, E., Yu, Y., Langlotz, C. P., Ball, R. L., Montine, T. J., Martin, B. A., Berry, G. J., Ozawa, M. G., Hazard, F. K., Brown, R. A., Chen, S. B., Wood, M., Allard, L. S., Ylagan, L., Ng, A. Y., Shen, J. (2020). Impact of a deep learning assistant on the histopathologic classification of liver cancer. Npj Digital Medicine, 3(1), 23. doi:10.1038/s41746-020-0232-8.

[7] Steiner, D. F., Macdonald, R., Liu, Y., Truszkowski, P., Hipp, J. D., Gammage, C., Thng, F., Peng, L., & Stumpe, M. C. (2018). Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. American Journal of Surgical Pathology, 42(12), 1636–1646. doi:10.1097/PAS.0000000000001151.

[8] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS 2012), New Yourk, United States, 1097-1105.

[9] Kandel, I., & Castelli, M. (2020). A novel architecture to classify histopathology images using convolutional neural networks. Applied Sciences (Switzerland), 10(8). doi:10.3390/APP10082929.

[10] Van der Laak, J., Litjens, G., & Ciompi, F. (2021). Deep learning in histopathology: the path to the clinic. Nature Medicine, 27(5), 775–784. doi:10.1038/s41591-021-01343-4.

[11] Kassani, S. H., Kassani, P. H., Wesolowski, M. J., Schneider, K. A., & Deters, R. (2019). Classification of histopathological biopsy images using ensemble of deep learning networks. arXiv preprint arXiv:1909.11870. doi:10.48550/arXiv.1909.11870.

[12] Cheng, J. Y., Abel, J. T., Balis, U. G. J., McClintock, D. S., & Pantanowitz, L. (2021). Challenges in the Development, Deployment, and Regulation of Artificial Intelligence in Anatomic Pathology. American Journal of Pathology, 191(10), 1684–1692. doi:10.1016/j.ajpath.2020.10.018.

[13] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. Journal of Big Data, 6(1), 60. doi:10.1186/s40537-019-0197-0.

[14] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. 37th International Conference on Machine Learning (ICML 2020), July 12-18 2020, Vienna, Austria, 1575–1585.

[15] Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. E. (2020). Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems, 33, 22243-22255.

[16] Chen, X. (2019). Image enhancement effect on the performance of convolutional neural networks. Department of Computer Science, Faculty of Computing, Blekinge Institute of Technology, Blekinge, Sweden.

[17] Rodríguez-Rodríguez, J. A., Molina-Cabello, M. A., Benítez-Rochel, R., & López-Rubio, E. (2021). The Effect of Noise and Brightness on Convolutional Deep Neural Networks. Lecture Notes in Computer Science, 639–654. doi:10.1007/978-3-030-68780-9_49.

[18] Taylor, L., & Nitschke, G. (2018). Improving Deep Learning with Generic Data Augmentation. 2018 IEEE Symposium Series on Computational Intelligence (SSCI). doi:10.1109/ssci.2018.8628742.

[19] Dodge, S., & Karam, L. (2016). Understanding how image quality affects deep neural networks. 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX). doi:10.1109/qomex.2016.7498955.

[20] Nazaré, T. S., da Costa, G. B. P., Contato, W. A., & Ponti, M. (2018). Deep Convolutional Neural Networks and Noisy Images. Lecture Notes in Computer Science, 416–424. doi:10.1007/978-3-319-75193-1_50.

[21] Haque, M. A., Marwaha, S., Deb, C. K., Nigam, S., Arora, A., Hooda, K. S., Soujanya, P. L., Aggarwal, S. K., Lall, B., Kumar, M., Islam, S., Panwar, M., Kumar, P., & Agrawal, R. C. (2022). Deep learning-based approach for identification of diseases of maize crop. Scientific Reports, 12(1), 6334. doi:10.1038/s41598-022-10140-z.

[22] Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36(4), 193–202. doi:10.1007/BF00344251.

[23] LeCun, Y., Haffner, P., Bottou, L., & Bengio, Y. (1999). Object recognition with gradient-based learning. In Shape, contour and grouping in computer vision, 319-345. Springer, Berlin, Heidelberg. doi:10.1007/3-540-46805-6_19.

[24] Widiputra, H. (2021). GA-Optimized Multivariate CNN-LSTM Model for Predicting Multi-Channel Mobility in the COVID-19 Pandemic. Emerging Science Journal, 5(5), 619-635. doi: 10.28991/esj-2021-01300.

[25] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi:10.1109/cvpr.2015.7298594.

[26] Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi:10.1109/cvpr.2017.195.

[27] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 770–778. doi:10.1109/CVPR.2016.90.

[28] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi:10.1109/cvpr.2017.243.

[29] Zhang, C., Benz, P., Argaw, D. M., Lee, S., Kim, J., Rameau, F., Bazin, J. C., & Kweon, I. S. (2021). ResNet or DenseNet? Introducing Dense Shortcuts to ResNet. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). doi:10.1109/wacv48630.2021.00359.

[30] Janowczyk, A., & Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. Journal of Pathology Informatics, 7(1), 29. doi:10.4103/2153-3539.186902.

[31] Cruz-Roa, A., Basavanhally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J., & Madabhushi, A. (2014). Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. Medical Imaging 2014: Digital Pathology. doi:10.1117/12.2043872.

[32] Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20(1), 37–46. doi:10.1177/001316446002000104.

[33] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. doi:10.48550/arXiv.1412.6980.

[34] Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2010). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255. doi:10.1109/cvpr.2009.5206848.

[35] GitHub (2015). Keras-team/keras: GitHub Inc. 2015. Available online: https://github.com/fchollet/keras (accessed on January 2022).

[36] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467. doi:10.48550/arXiv.1603.04467.

[37] Vingelmann, P. and Fitzek, F. H. P. (2020). NVIDIA: CUDA, Release: 10.2.89. 2020. Available online: https://developer.nvidia.com/cuda-toolkit (accessed on January 2022).

[38] Choi, J. Y., Yoo, T. K., Seo, J. G., Kwak, J., Um, T. T., & Rim, T. H. (2017). Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database. PLoS ONE, 12(11), e187336. doi:10.1371/journal.pone.0187336.

[39] Hermsen, M., de Bel, T., den Boer, M., Steenbergen, E. J., Kers, J., Florquin, S., … van der Laak, J. A. W. M. (2019). Deep Learning–Based Histopathologic Assessment of Kidney Tissue. Journal of the American Society of Nephrology, 30(10), 1968–1979. doi:10.1681/asn.2019020144.

[40] Kitamura, G., Chung, C. Y., & Moore, B. E. (2019). Ankle Fracture Detection Utilizing a Convolutional Neural Network Ensemble Implemented with a Small Sample, De Novo Training, and Multiview Incorporation. Journal of Digital Imaging, 32(4), 672–677. doi:10.1007/s10278-018-0167-7.

[41] Berral-Soler, R., Madrid-Cuevas, F. J., Muñoz-Salinas, R., & Marín-Jiménez, M. J. (2021). RealHePoNet: a robust single-stage ConvNet for head pose estimation in the wild. Neural Computing and Applications, 33(13), 7673–7689. doi:10.1007/s00521-020-05511-4.

[42] Perez, F., Vasconcelos, C., Avila, S., & Valle, E. (2018). Data Augmentation for Skin Lesion Analysis. OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, 303–311, Springer, Cham, Switzerland. doi:10.1007/978-3-030-01201-4_33.