Paired Augmentation for Improved Image Classification using Neural Network Models

Shikar Rajcomar
School of Mathematics, Statistics and Computer Science
University of KwaZulu-Natal
Durban 4041 South Africa
Centre for Artificial Intelligence Research, CAIR
shikar.rajcomar@gmail.com

Anban W. Pillay
School of Mathematics, Statistics and Computer Science
University of KwaZulu-Natal
Durban 4041 South Africa
Centre for Artificial Intelligence Research, CAIR
pillayw4@ukzn.ac.za

Edgar Jembere
School of Mathematics, Statistics and Computer Science
University of KwaZulu-Natal
Durban 4041 South Africa
Centre for Artificial Intelligence Research, CAIR
jemberee@ukzn.ac.za

2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) | 978-1-6654-1974-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/CSDE50874.2020.9411553

***Abstract*—The lack of sufficient data points and class imbalances in datasets is a serious problem that mitigates against the success of deep learning classification models. These problems result in model overfitting, poor accuracy, and poor generalization. Traditional augmentation techniques, and advanced augmentation, such as adversarial approaches, have individually been found to be effective. In this work, we present a method to determine the most effective augmentation techniques to combine into the machine learning pipeline. We propose using only two augmentations in the pipeline, an advanced technique applied in an offline manner followed by a simple technique applied in an online manner. This approach is validated by application to two medical image problems using datasets characterized by class imbalances and small sizes. The former is a binary classification problem using a brain tumor dataset and the latter is a multi-label classification problem using a white blood cell dataset. Exhaustive experimentation with approximately 170 different combinations of augmentations methods is reported. Experimental results indicate a 15% and 11% improvement in validation accuracy over using no augmentation and a 2% improvement over using a single augmentation. We conclude that the pairing of augmentation methods results in improvements in the classification tasks.**

Keywords—Augmentation, Neural Networks, Image Classification, Image Processing, Computer vision.

I. Introduction

Deep learning approaches have demonstrated superiority over traditional computer vision techniques in a large range of image processing tasks including classification, object detection, and object recognition. Image classification tasks, in particular, have seen remarkable progress in the recent past since the seminal work of Hinton and colleagues that demonstrated an order of magnitude improvements possible with deep neural networks [1].

A common challenge for these deep learning models is the requirement for large datasets to train and construct models, typically in the order of millions of data points, to ensure high accuracies. These models also have a high tendency to overfit i.e. the models learn the training data accurately but do not generalize well to unseen data. Accuracy and generalization problems are exacerbated when limited data is available. These issues are particularly acute in the medical domain where deep learning has been applied to problems such as cancer detection, skin lesion classification, brain MRI analysis, and others [2]. Much of the images captured in this domain are costly to collect and maintain. It is challenging to assemble these datasets due to the rarity of certain medical conditions, patient confidentiality, and access to medical experts to correctly label data points. Gathering raw data can be very time-consuming and expensive. It is common for these data sets to be poorly distributed among the classes.

To date, reducing model overfitting remains a challenge in image classification [3]. Research on reducing overfitting is either focused on the model architecture or data augmentation. Data augmentation approaches the overfitting problem from its root cause viz. the size of the data set. For image datasets, augmentation involves creating new data points by transforming existing images, for example by rotating or flipping existing images or synthesizing new images using adversarial approaches.

In this work, we distinguish between simple and advanced augmentation techniques. Simple techniques involve transformation of an input image such as rotation, flipping, shear, and translation. These techniques are computationally less expensive than advanced techniques. The latter approaches use adversarial and other techniques to synthesize new images from one or more input images.

Augmentation may be applied in an online or offline manner. Offline augmentation transforms all images by increasing the size of the dataset before input to the model whilst online augmentation transforms the images in a batch-wise manner during model training in a feedforward manner. Offline augmentation is more suitable for computationally expensive techniques.

In the computer vision domain, most image classification tasks are used with hand-picked augmentation strategies. There is little guidance on which augmentation techniques to use due to lack of research [4]. The general approach to deciding which method to use typically involves starting with an assumption of which method may work best and then validating it through testing. Literature has demonstrated that pairing augmentations in the machine learning pipeline produce better results. However, this has not been shown to be effective in the medical imaging domain. Most research on the effectiveness of augmentation uses mainstream datasets such as MNIST, CIFAR10, and Caltech [5].

Selecting suitable data augmentation strategies is even more important than choosing a network structure. There has been research into automating the process of selecting an augmentation method such as AutoML which have shown to produce state of the art results, however they are technically impractical due to high resource usage and processing time

(500 GPU hours for one augmentation search) [6]. In the absence of such guidance, practitioners have to rely on intuition and experience in choosing a data augmentation strategy.

This research presents a framework to determine the most effective augmentation techniques to combine into the machine learning pipeline. The proposed method determines which two augmentations (an advanced technique applied in an offline manner and a simple technique applied in an online manner) to apply for a given problem This approach is validated by exhaustive application to two medical image datasets: a brain tumor and white blood cell dataset. This research will also outline the importance of structuring the augmentations used rather than randomly selecting one method and trying to squeeze performance from model tuning. Besides being efficient and effective, the framework can be easily implemented and applied to different models and tasks. We do not test more than two augmentations as studies reported in the literature, and our own experimentation, demonstrates that this does not result in substantial performance gains or leads to performance decrease.

The results obtained validates the approach and shows statistically significant performance gains over approaches using no augmentation or a single augmentation.

The rest of the paper is organized as follows: Section 2 provides an overview of the relevant literature on augmentation techniques, section 3 describes our approach and section 4 details the experimental design. Results are given and discussed in section 5. Conclusions and avenues for future work conclude the paper.

II. RELATED WORK

Simple (or traditional augmentations) are based on image transformations where simple transforms are applied to subsets of images. These are widely used due to their speed, reproducibility, reliability, and ease of implementation. They are also used to balance datasets with skewed classes.

Several studies have introduced new transformation techniques. In random erasing [4], random pixels in a rectangular shape are erased. The author noted that datasets that were expanded by this method resulted in reduced overfitting and resulted in more robust models. Noise injection was developed as a technique to help convolutional neural networks defend against adversarial attacks. Applying different noise levels to images as an augmentation technique was studied in [7]. Experimental results on two sub-datasets of ImageNet suggested that Speckle noise leads to better CNN models.

Advanced augmentation methods use image synthesis techniques to augment image datasets. Sample pairing is an image synthesis technique developed by [3] that involves overlaying one image over another. This approach led to significantly improved classification accuracies for all the tested datasets. The images may seem convoluted to a human eye but demonstrated error rate improvements between 8.22% to 6.93% on the CIFAR 10 dataset [3]. A reduction in the error rate from 43.1 to 31.0% was achieved on the reduced CIFAR 10 dataset (1000 images).

Adversarial techniques, using generative models have also proved effective. The use of neural style transfer was examined in [8]. The authors compared it to traditional methods and as a secondary investigation, combined the traditional methods with neural style transfer. The Caltech 101 and Caltech 256 dataset was used. Using a pre-trained VGGNET architecture the author noted that without any augmentation a classification accuracy of 83.34% was reached. Using traditional augmentation strategies of rotation and flipping resulted in a marginal decrease in performance. A combination of flipping and rotation significantly reduced accuracy. By testing eight different artistic styles, neural style transfer outperformed traditional methods. Using traditional methods combined with style transfer demonstrated a marginal performance increase. Using the Caltech 101 dataset, produced roughly an accuracy of 84.5% with no augmentation applied. This is used as a baseline for comparison to other methods. There were no improvements when flipping or the wave style transfer method. However, there was much improvement when these methods were combined (85.81%).

The efficiency of using traditional data augmentation techniques (cropping, rotating) and advanced methods like GANS on a limited ImageNet data was studied in [9]. The authors noted that using traditional methods of image augmentation proved to be more effective on its own while consuming less time when compared to the advanced methods.

Up until recently, there have not been many comparative studies performed between the performance of different augmentation methods. The use of traditional augmentation techniques against advanced methods like GAN`s, WGAN`s. was studied in [10]. This research was performed using the ImageNet and CIFAR-10 datasets. The work demonstrated that using a combination of augmentation methods demonstrated better performance than using a single augmentation. Overall, using a combination of cropping and flipping and WGAN and flipping presented the highest performance with classification improvements of roughly 3-3.5 % on the CIFAR-10 dataset and 2-2.5% on the ImageNet dataset. The author also experimented with triple combination augmentation but noted that the performance was lower than the paired combination and further some triple augmentations decreased performance. From examining this research, it is crucial to know which methods should be paired together as there are definite performance improvements that occur when using a combination of methods, however, the increase in performance will be dependent on the pair of methods chosen.

Using the Caltech 256 dataset, comparable results are indicated using flipping (which produced an accuracy of 66.66%) and the scream style transfer method (which produced an accuracy of 63.13%). Combining both methods increased accuracy to 67.28%, Whilst combining flipping with wave produced an accuracy of 66.32 which is higher than just using wave style transfer by itself. Combing augmentations methods can greatly increase the size of any image datasets. However, this is not guaranteed to be advantageous. In domains with very limited data, this could result in further overfitting. Therefore, it is important to apply domain knowledge for deriving an optimal subset of augmented data to train Deep Learning models [5].

III. APPROACH

The proposed approach is to apply augmentation in an offline manner and then apply online augmentation during model training.

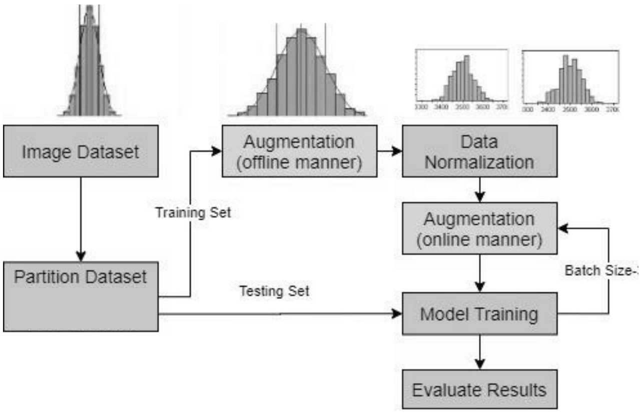

Fig. 1. Proposed paired augmentation methodology

The offline augmentation transforms all images by increasing the size of the dataset before any input is made to the model increasing the dataset twofold. Whilst the online augmentation transforms the images in a batch-wise manner during model training in a feedforward manner. It is noted that the entire point of using augmentation is to ensure that the network is exposed to new images at every epoch. By including the original image with the augmented image during each batch of training, the network will see the same image multiple times. To overcome this issue, the advanced methods were combined with traditional methods in an online manner to ensure that the network sees new variations of data at each and every epoch. The online method is used for traditional augmentations only because applying online methods to advanced techniques will increase the training time substantially and increase resources. To mitigate this, it was decided to combine them (see Fig.1).

IV. Experimental Design

In this paper, all experiments were carried out on the two datasets independently. Different datasets and model structures will bring different effects with the same data enhancement. However, using these independent datasets and architectures will verify that the suggested technique will always achieve improvements demonstrating its meta-heuristics ability.

The first set of experiments trains the model using the original dataset. This will identify the minimum classification accuracy. Thereafter the process will be repeated using single augmentation. The purpose of these initial tests is to provide a benchmark for comparison. Thereafter the images will be augmented using a combination of traditional and advanced augmentation methods. The results of the paired augmentation experiments will be compared to the baseline results to validate if there have been any improvements. To validate the generalization ability of the proposed framework the top-performing methods will be repeated using the VGG16 net architecture to demonstrated that the framework is independent of the dataset or model architecture used.

Neural networks are stochastic algorithms and as a result, a defined model fitted on the same data may produce different results each time. To overcome this issue K-fold cross-validation is performed. This cross-validation test randomly dividing the dataset into k groups of equal size. The first fold is kept for testing and the model is trained on k-1 folds. The process is repeated K times and each time a different group of data points is used for validation [11].

Datasets

The two datasets used are the Brain Tumor(dataset1) and the White Blood cell (dataset 2) respectively. Dataset 1 consists of MRI scans of two classes namely; no tumor images, encoded as 0, and tumor images, encoded as 1. Fig. 2 presents a sample of the images in the dataset. The datasets contain 91 images of Class 0 images and 100 images of Class 1. The dataset was previously classified with a genetic algorithm and support vector machine and classification accuracy of 83.22% [12] was reported.

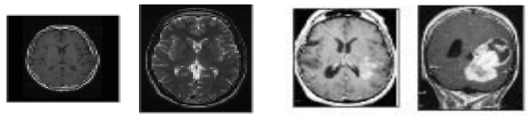

Fig. 2. Sample of images from the Brain Tumor

The dataset contains 299 120×120 images with a colour depth is 24 bits and contains four classes: neutrophil (176 images), lymphocyte (53 images), monocyte (48 images), and eosinophil (22 images) **[13]**. Fig.3 presents sample images from the dataset.

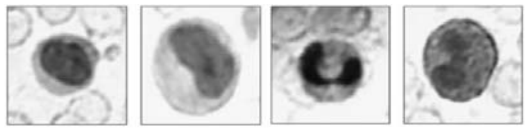

Fig. 3. Sample of Blood cell images

Neural Architectures

To evaluate the generalization of our approach over network architectures, the experiments were run VGG16 network. It was established that using the VGG16 model across both datasets did demonstrate that pairing of methods always resulted in the best performing networks. However, using one architecture for both datasets that have very different features does not always result in optimum performance[14]. It is for this reason that two unique models were used. Also using different models also indicate that the proposed method is independent of the architectures and when used, indicates that the pairing of methods results in better performing networks. The architectures for each dataset is presented below.

Architecture for Dataset 1: The input layer converts the image to a 3x3 tensor. The tensor size is 120x120x3, with the first two values being the image row and width respectively, and the last value representing the primary colours (Red, Green, and Blue). The first layer consists of a convolutional layer made up of 136 filters and with the second layer having 64 filters. The weights of the first layer are initialized from a Gaussian distribution with 0 mean. The initializer is also seeded at every neuron to produce the same random tensor on every run. Each layer uses a 3×3 kernel. With a max-pooling filter of size 2×2 and a dropout layer (25%). The pre-processing stage has a 128 filter layer. This layer is followed by a dropout layer (50%) which feeds directly to the final output layer. The final output has 128 filters and applies the softmax function to classify the images.

Architectures for Dataset 2: The architecture consisted of five trainable layers: four convolutional layers, a fully connected layer, and a softmax layer. The input to the CNN is a 3 channel (RGB) 224x224 pixel image as an input.

Convolved layers over the image produce a feature map of size 222×222×32. This layer also contains a max-pooling filter of size 2×2. The second layer contains a set of 64 filters producing a new feature map of size 109×109×64. The third and fourth layers contain a set of 160 filters followed with a dropout layer (20%) reducing the feature map size to 107×107×160 and 105×105×160 respectively. The fully connected layer is of size 160, connected to a soft-max layer of size 160.

TABLE 1.OPTIMIZED HYPER-PARAMETERS

Hyper-parameter	Type
Activation Function	Rectified Linear(RELU)
Weight Initialization	Glorot normal
Learning Rate	0.0001
Optimization Algorithm	ADAM[a], Adadelta[b]
Regularization	L2 vector norm(0.001)
Batch	32
Epochs	30

[a] Optimisation Algorithm Dataset 1
[b] Optimisation Algorithm Dataset 2

All weights for both architectures were updated using back-propagation with a learning rate of 0.0001. Tensorflow on GPU was used to train the models. Also, regularization was used in the form of L2 vector norm also called weight decay with a regularization parameter of 0. 001. The regularization helps reduce the overfitting of the model to the training dataset and improve the performance on the holdout set. The optimized hyperparameters are presented in TABLE 1.

V. RESULTS AND DISCUSSION

This section presents the analysis of the paired augmentation. The results using no augmentation and singular augmentation are collected as the baseline. Thereafter experiments are performed using the pairing of samples. All results are reported to two decimal places but machine precision has been used in all calculations. The highest test accuracy is reported as the best score.

A. Baseline Results

Table 3 presents the results of using the original dataset with no augmentation. a model training accuracy of 67% and validation accuracy 64% for dataset 1 and training accuracy of 55% and validation accuracy 60% dataset 2 was achieved.

TABLE 2. EXPERIMENTAL RESULTS USING ORIGINAL DATASETS

	F1-Score	Precision	Recall
Dataset 1			
No Tumor (0)	41± 16%	97 ± 4%	29 ± 14%
Tumor (1)	74 ± 3%	60 ± 5%	99 ± 2%
Training Accuracy			67 ± 5%
Validation Accuracy			64 ± 2%
Dataset2			
Lymphocytes (0)	75 ± 0%	60 ± 0%	100 ± 0%
Monocytes (1)	0 ± 0%	0 ± 0%	0 ± 0%
Neutrophils (2)	0 ± 0%	0 ± 0%	0 ± 0%
Eosinophils (3)	0 ± 0%	0 ± 0%	0 ± 0%
Training Accuracy			55 ± 1%
Validation Accuracy			60 ± 0%

The accuracies of the models seem to be relatively good. However, when examining the loss curve, after epoch 7 and 15 for each dataset respectively indicates that the model suffers from overfitting. This is common with datasets that are small, as in the case of the datasets used. The inflection point at these epochs indicated that the variance between the training and testing curves increased rapidly. Furthermore, when analysing the confusion matrix(see Fig. 4.), it is evident by looking at the recall metric that the model is only able to identify one class (class 0). This is typical of scenarios where there is an imbalance of a class or limited data. Consequently, when the validation dataset does not have enough information the model`s ability to generalize is poor. This issue is evident in the models learning curve, where the training loss seems to be a good fit whilst the learning curve for validation loss shows erratic movements around the training loss. Since it was evident that the limited data in the validation dataset posed problems it was decided to augment the testing datasets. This involves creating multiple augmented copies of each image in the test set and having the model make a prediction for each. It was ensured that no augmentations from the training dataset were present in the testing set.

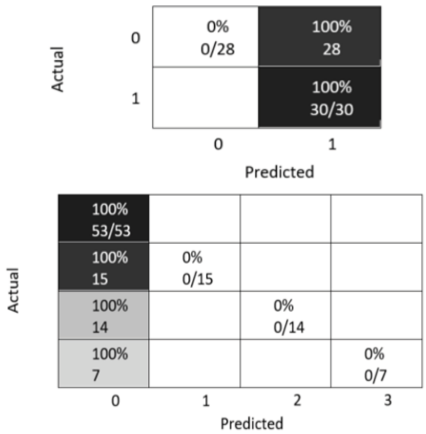

Fig. 4. Confusion matrix of original datasets (no augmentation)

The second aspect of the experimental procedure paired traditional and advanced methods together. The augmentations performed all look different from each other and each image is fed through the image generator to create the synthetic dataset. Twenty-six different augmentations were tested. The list of results is too extensive to be shown in this paper and only the augmentation with the highest validation accuracy is presented in TABLE 3. Shear in an online augmentation manner for dataset 1 and vertical flipping also in an online manner for dataset 2 resulted in the highest validation accuracy.

TABLE 3. EXPERIMENTAL RESULTS USING SINGLE AUGMENTATION

	F1-Score	Precision	Recall
Dataset 1			
No Tumor (0)	81± 3%	71± 1%	93 ± 8%
Tumor (1)	71 ± 1%	91 ± 11%	59 ± 5%
Training Accuracy			75 ± 5%
Validation Accuracy			77 ± 2%
Dataset2			
Lymphocytes (0)	77 ± 2%	62 ± 2%	100 ± 0%
Monocytes (1)	31 ± 26%	60 ± 49%	21 ± 18%
Neutrophils (2)	3 ± 5%	20 ± 40%	1 ± 3%
Eosinophils (3)	0 ± 0%	0 ± 0%	0 ± 0%
Training Accuracy			67 ± 11%
Validation Accuracy			65 ± 5%

As expected using augmentation helps significantly improve model performance and decrease overfitting. The artificially expanded training dataset leads to a more skilful model since the performance of the model scales relative to the size of the training set. In addition, the modified or augmented versions of the images using the traditional methods assisted the models in extracting and learning features in a way that is invariant to their position and lighting rather than the more advanced methods.

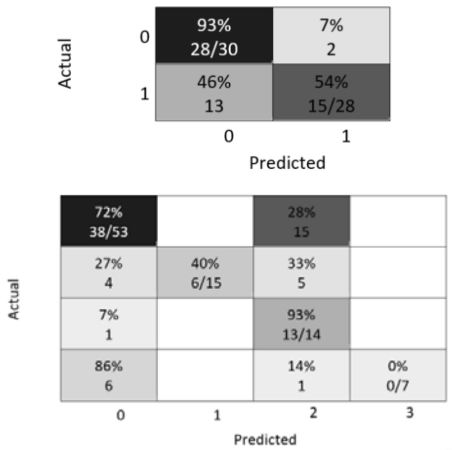

Fig. 1. Confusion matrix of with singular augmentation

B. *Paired Augmentation Results*

To evaluate the use of paired augmentation, the 26 types of augmentations were paired together. In total 170 different pairing combinations were tested for each dataset. The top-performing results are presented in TABLE 4.

TABLE 4. EXPERIMENTAL RESULTS USING PAIRED AUGMENTATION

	F1-Score	Precision	Recall
Dataset 1			
No Tumor (0)	74 ± 2%	88 ± 7%	65 ± 4 %
Tumor (1)	82 ± 2%	74 ± 1%	91 ± 6 %
Training Accuracy			88 ± 6 %
Validation Accuracy			79 ± 2 %
Dataset2			
0	72 ± 6 %	80 ± 17 %	71 ± 12 %
1	70 ± 4 %	84 ± 17 %	64 ± 11 %
2	61 ± 21 %	61 ± 7 %	76 ± 33 %
3	87 ± 14 %	96 ± 3 %	83 ± 21 %
Training Accuracy			89 ± 6 %
Validation Accuracy			70 ± 13 %

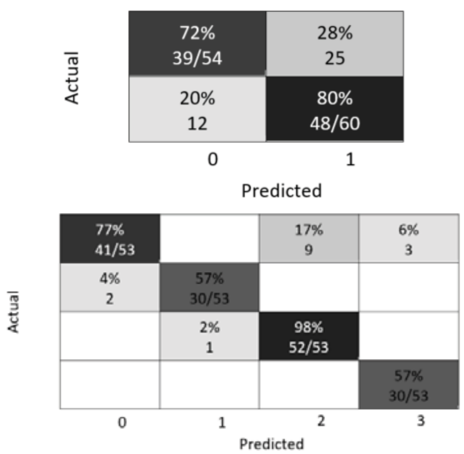

Fig. 2. Confusion matric with pairing augmentation

For dataset 1, sample pairing was paired with the shear method which leads to better performing networks. For dataset 2, shear with vertical flipping proved to be the best method. The flipping methods were also the best singular augmentation method used for dataset 2 as well. By applying the first augmentation initially, the overall semantic content of the original image is preserved, whilst the high-level discriminative features of the images are maintained, adding diversity to the dataset that is limited in size. However, if this original data along with the augmented data is used for training, then there is a high probability that the network would be exposed to the original data multiple times, effectively defeating the purpose. Secondly, the pairing aims to help with generalization problems faced by models with limited data. By applying the second typical augmentation in an online manner ensures that at every epoch the model is exposed to a new diverse batch of images allowing it to extract as many features as possible.

The intuition is simple: flipping the images make sense due to the symmetric nature of the images. Other augmentations like neural style transfer add too much light and color to images making them almost unrecognizable. These images are anatomically incorrect, and will never occur in reality and create noise rather than valuable information. Consequently, data formed by these types of augmentations forces the network to overgeneralize to situations it will never see. Over generalization hurts performance because the network wastes its predictive capacity learning irrelevant features. A good learning curve is one which exhibits the loss of the model being lower on the training dataset than the validation dataset. As compared to the other learning curve. The learning curve of the model trained using the paired method displays the closest characteristics to a good learning curve the training loss is lower than the validation set, the validation loss does not fluctuate as much as the other curves. Furthermore, the validation and training curves are relatively close to each other indicating that the model is not experiencing overfitting problems.

The distribution of validation and training accuracy for baseline, single and paired augmentation is presented in Fig. . It is evident that the using paired augmentation provides the overall highest validation and training accuracy as compared to the other methods. Generally evaluating data augmentation methods are typically evaluated based on the performance improvements realized during the neural networks training. This typical approach may be a simple estimate, to determine which augmentation method may be best, but with this approach there is no confidence in the results being real or if they occur due to chance.

The paired augmentation method (research focus) was tested to verify if it improves model accuracy as compared to other augmentation techniques (singular augmentation, baseline) under the same operational conditions and inputs. Without this verification there can be no confidence that the paired augmentation can improve model accuracy. The test that will be applied here is the 2 sample t-test. In this test, the burden of proof lies in the alternative hypothesis (H1). The hypothesis tests are described as follows:

Null hypothesis H0: Paired augmentation accuracy equivalent to singular augmentation accuracy (the primary assumption)

Alternative hypothesis H1: Paired augmentation accuracy greater that singular augmentation accuracy (what needs to be proved)

In this report, a difference of greater than 1% between the two means is considered important at a confidence level of 80%. If the null hypotheses are rejected, then the difference falls within the equivalence interval and then it can be claimed that the means for the products are different – that is the paired augmentation accuracy is greater than the singular argumentation accuracy. This information is conveyed graphically using distribution plots presented in Fig. . The confidence interval for the distributions is different, indicating that the average validation accuracy of all paired augmentations tested is in fact greater than that of the single augmentation methods. Consequently, the alternative hypothesis is accepted, validating our claim that using paired augmentation is better than just using a single augmentation.

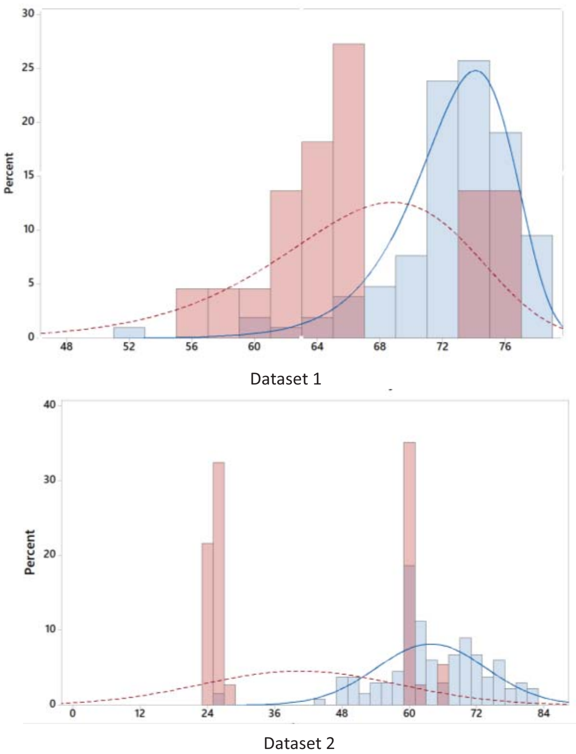

Fig. 7. Distribution of Validation accuracy (Legend- Red: Single Augmentation, Blue: Paired Augmentation)

Conclusions and Future Work

This study's results demonstrate that an effective method of increasing NN classification task performance is to make use of combined augmentation. By evaluating various pairing of augmentation methods using a relatively simple NN architecture on a limited image dataset we were able to demonstrate that using a combination of augmentations is always more beneficial than using a single augmentation. The greatest performance improvement was yielded by using a combination of methods to produce validation accuracies that was 11% better than using no augmentation and 3 % better than using a single augmentation. As future work, we intend to study combined augmentations with ensemble modeling. Ensemble modeling is a machine learning method in which predictions of two or more models are combined. This modeling has proven to be an effective method in improving neural network performance but a combination of paired augmentations, where a combination of models trained on datasets exposed to paired augmentation has not yet been explored. Another direction for future work is to pair more advanced methods such as GANs, WGAN, noise filtering, and random erasing.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, 'ImageNet Classification with Deep Convolutional Neural Networks', in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[2] J. Gao, Y. Yang, P. Lin, and D. S. Park, 'Computer Vision in Healthcare Applications', J. Healthc. Eng., vol. 2018, Mar. 2018, doi: 10.1155/2018/5157020.

[3] H. Inoue, 'Data Augmentation by Pairing Samples for Images Classification', ArXiv180102929 Cs Stat, Apr. 2018, Accessed: Sep. 12, 2020. [Online]. Available: http://arxiv.org/abs/1801.02929.

[4] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, 'Random Erasing Data Augmentation', ArXiv170804896 Cs, Nov. 2017, Accessed: Sep. 21, 2020. [Online]. Available: http://arxiv.org/abs/1708.04896.

[5] Md. Z. Alom et al., 'A State-of-the-Art Survey on Deep Learning Theory and Architectures', Electronics, vol. 8, p. 292, 2019, doi: 10.3390/electronics8030292.

[6] X. He, K. Zhao, and X. Chu, 'AutoML: A Survey of the State-of-the-Art', ArXiv190800709 Cs Stat, Jul. 2020, Accessed: Nov. 15, 2020. [Online]. Available: http://arxiv.org/abs/1908.00709.

Dataset 2

[7] M. E. Akbiyik, 'Data Augmentation in Training CNNs: Injecting Noise to Images', Sep. 2019, Accessed: Sep. 21, 2020. [Online]. Available: https://openreview.net/forum?id=SkeKtyHYPS.

[8] X. Zheng, T. Chalasani, K. Ghosal, S. Lutz, and A. Smolic, 'STaDA: Style Transfer as Data Augmentation', Proc. 14th Int. Jt. Conf. Comput. Vis. Imaging Comput. Graph. Theory Appl., pp. 107–114, 2019, doi: 10.5220/0007353401070114.

[9] L. Perez and J. Wang, 'The Effectiveness of Data Augmentation in Image Classification using Deep Learning', ArXiv171204621 Cs, Dec. 2017, Accessed: Sep. 22, 2020. [Online]. Available: http://arxiv.org/abs/1712.04621.

[10] J. Shi-jie, W. Z. Ping, J. Peiyi, and H. Si-ping, 'Research on data augmentation for image classification based on convolution neural networks', 2017 Chin. Autom. Congr. CAC, 2017, doi: 10.1109/cac.2017.8243510.

[11] R. Khandelwal, 'K fold and other cross-validation techniques | by Renu Khandelwal | Data Driven Investor | Medium', Medium, Nov. 03, 2018. https://medium.com/datadriveninvestor/k-fold-and-other-cross-validation-techniques-6c03a2563f1e (accessed Sep. 14, 2020).

[12] M. Jafari and R. Shafaghi, '(PDF) A Hybrid Approach for Automatic Tumor Detection of Brain MRI Using Support Vector Machine and Genetic Algorithm', ResearchGate, Dec. 2012. https://www.researchgate.net/publication/234072915_A_Hybrid_Approach_for_Automatic_Tumor_Detection_of_Brain_MRI_Using_Support_Vector_Machine_and_Genetic_Algorithm (accessed Sep. 12, 2020).

[13] X. Zheng, Y. Wang, G. Wang, and J. Liu, 'Fast and robust segmentation of white blood cell images by self-supervised learning', Micron, vol. 107, pp. 55–71, Apr. 2018, doi: 10.1016/j.micron.2018.01.010.

[14] 'Generalization Performance - an overview | ScienceDirect Topics'. https://www.sciencedirect.com/topics/computer-science/generalization-performance (accessed Oct. 30, 2020).