# Data Augmentation for Skin Lesion Analysis

Fábio Perez[1], Cristina Vasconcelos[2], Sandra Avila[3], and Eduardo Valle[1]

[1]RECOD Lab, DCA, FEEC, University of Campinas (Unicamp), Brazil
[2]Computer Science Department, IC, Federal Fluminense University (UFF), Brazil
[3]RECOD Lab, IC, University of Campinas (Unicamp), Brazil

**Abstract.** Deep learning models show remarkable results in automated skin lesion analysis. However, these models demand considerable amounts of data, while the availability of annotated skin lesion images is often limited. Data augmentation can expand the training dataset by transforming input images. In this work, we investigate the impact of 13 data augmentation scenarios for melanoma classification trained on three CNNs (Inception-v4, ResNet, and DenseNet). Scenarios include traditional color and geometric transforms, and more unusual augmentations such as elastic transforms, random erasing and a novel augmentation that mixes different lesions. We also explore the use of data augmentation at test-time and the impact of data augmentation on various dataset sizes. Our results confirm the importance of data augmentation in both training and testing and show that it can lead to more performance gains than obtaining new images. The best scenario results in an AUC of 0.882 for melanoma classification without using external data, outperforming the top-ranked submission (0.874) for the ISIC Challenge 2017, which was trained with additional data.

**Keywords:** Skin lesion analysis · Data augmentation · Deep learning

## 1 Introduction

Deep learning has achieved impressive results in computer vision tasks, including skin lesion analysis [4]. However, deep learning models are data-hungry, and collecting and annotating skin lesion images can be challenging.

In image classification tasks, knowledge transfer and data augmentation are regularly employed for small datasets. Knowledge transfer usually takes place by initially training a Convolutional Neural Network (CNN) in a large source dataset (e.g., ImageNet) and using its weights as a starting point for training in the smaller target dataset [10]. Data augmentation goal is to add new data points to the input space by modifying training images while preserving semantic information and target labels. Thus, it is used to reduce overfitting.

In this work, we: (i) investigate the impact of applying diverse data augmentation techniques to three different CNN architectures (namely Inception-v4 [13], ResNet [5], and DenseNet [6]); (ii) investigate the impact of data augmentation on different dataset sizes; and (iii) evaluate the use of different data augmentation methods during test-time, aiming to reduce generalization error. We con-

ducted the experiments on the ISIC Challenge 2017 dataset [3] for melanoma classification task.

## 2 Related Work

Data augmentation is broadly used in CNN architectures, such as AlexNet [8], Inception [7,13,14], ResNet [5], and DenseNet [6]. These architectures are trained on the ImageNet dataset, which contains millions of annotated images. Some examples of data augmentation techniques are color modifications and geometric transforms (rotation, scaling, random cropping).

Models can also benefit from data augmentation on test-time. Krizhevsky et al. [8] average the predictions on 10 patches (cropped from the center plus the four corners and then flipped) extracted from each test image. Szegedy et al. [14] report gains with a method that generates 144 patches by cropping images at different resolutions, when compared with the 10-crop method. These methods are commonly used in competitions to increase final performance but can be expensive for production.

Data augmentation is also extensively employed in skin lesion classification, a task that has much less available training data. Data augmentation is ubiquitous among top-ranked submissions in the ISIC Challenge 2017 [1, 9, 11].

Some works specifically explore data augmentation for skin lesion analysis [12, 15, 16]. Vasconcelos and Vasconcelos [16] report gains in performance by using data augmentation with geometric transforms (rotations by multiples of 90 degrees; flips; lesion-preserving crops), PCA-based color augmentation, and specialist warping that preserves lesions symmetries and anti-symmetries. Valle et al. [15] highlight the importance of using data augmentation for both training and testing. They averaged the predictions for 50 augmented test samples. Pham et al. [12] compare the effects of data augmentation on classifiers (SVM, neural networks, and random forest) trained with features extracted with a pre-trained Inception-v4. Their results indicate that using more samples in test data augmentation (100 vs. 50) increases the model's performance.

In this work, we further investigate the use of data augmentation for skin lesion analysis, by comparing: test techniques (testing on a single image; test data augmentation; and test cropping, commonly employed in CNN architectures for image classification); 13 different data augmentation scenarios, including a novel augmentation; and the effects of data augmentation on different dataset sizes.

## 3 Methodology

### 3.1 CNN Architectures

We evaluated every experiment on three very deep CNNs that are widely used in computer vision problems: Inception-v4 [13], ResNet-152 [5], and DenseNet-161 [6]. We chose these networks as they achieve increased depth with different design choices and represent the state of the art in image classification.

The Inception-v4 [13] architecture has modules that concatenate feature maps from parallel convolutional blocks, leading to increased width and depth. Residual Networks (ResNets) [5] use shortcut connections between layers, allowing even deeper networks. Densely Connected Networks (DenseNets) [6] concatenate the output of each layer to all subsequent layers inside a dense block, increasing the parameter efficiency and reducing overfitting.

Since we used the same optimization hyperparameters for the three networks, we do not intend to compare the numeric values alone, but rather compare big-picture results and trends.

### 3.2 Data Augmentation Techniques

We evaluated 13 data augmentation scenarios, comprising different image processing techniques, and some combinations of them. Table 1 describes the implementation details for each scenario. Fig. 1 shows examples of all scenarios.

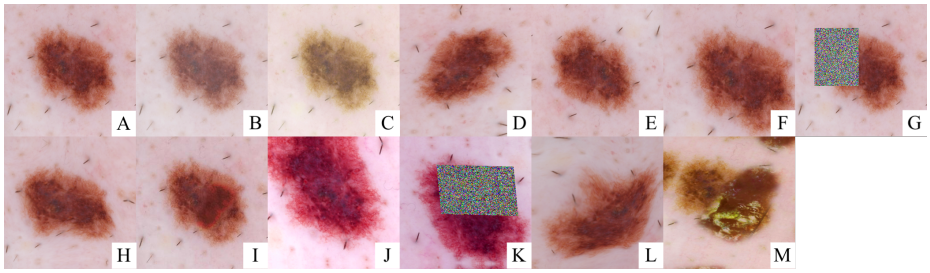


Fig. 1: Examples of augmentation scenarios, described in Table 1.

### 3.3 Training and Evaluation

We trained each network with Stochastic Gradient Descent (SGD) with a momentum factor 0.9, batch size of 32, starting learning rate 1e-3, reduced to 1e-4 after the $10^{th}$ epoch. The training data was shuffled before each epoch. The networks were initialized with weights trained on the ImageNet dataset, and fine-tuned with the ISIC Challenge 2017 train dataset (2000 images) [3]. The experiments were implemented with PyTorch (pytorch.org). Augmentations were implemented with torchvision and imgaug (github.com/aleju/imgaug).

All images were resized offline to a maximum width or height of 1024 pixels to avoid expensive resizing during training. On training, images were resized to the default input sizes for each network ($224 \times 224$ for DenseNet and ResNet; $299 \times 299$ for Inception-v4), although larger sizes were possible due to global average pooling. Images were normalized (subtract from the mean and divide by the standard deviation) based on the ImageNet dataset, in which the networks were pretrained. Augmentations were randomly applied online during training.

Table 1: Augmentation scenarios. Scenarios **J** to **M** represent augmentations compositions applied in the presented order.

| ID | Name | Description |
|---|---|---|
| **A** | No Augmentation | No data augmentation. Only preprocess images, as described in Section 3.3. |
| **B** | Saturation, Contrast, and Brightness | Modify saturation, contrast, and brightness by random factors sampled from an uniform distribution of $[0.7, 1.3]$, simulating changes in color due to camera settings and lesion characteristics. |
| **C** | Saturation, Contrast, Brightness, and Hue | As described in B, but also shift the hue by a value sampled from an uniform distribution of $[-0.1, 0.1]$. |
| **D** | Affine | Rotate the image by up to 90°, shear by up to 20°, and scale the area by $[0.8, 1.2]$. New pixels are filled symmetrically at edges. This can reproduce camera distortions and create new lesion shapes. |
| **E** | Flips | Randomly flip the images horizontally and/or vertically. |
| **F** | Random Crops | Randomly crop the original image. The crop has $0.4 - 1.0$ of the original area, and $3/4 - 4/3$ of the original aspect ratio. |
| **G** | Random Erasing | Fill part of the image (area up to 30% of the original image) with random noise. The transformation is applied with a probability of 0.5. Implemented as described in [17]. The network may benefit from occlusion by learning to look for different lesion attributes. |
| **H** | Elastic | Warp images with Thin Plate Splines (TPS). The warp is generated by defining the origins as an evenly-spaced $4 \times 4$ grid of points, and destinations as random points around the origins (by up to 10% of the image width on each direction). This can produce new lesion shapes while maintaining medical attributes. |
| **I** | Lesion Mix | Mix two lesions, by inserting part of a foreground lesion (cut by its segmentation mask) into a background lesion. We apply Gaussian blur to the foreground lesion to avoid sharp edges, and equalize its color histogram with respect to the segmented background lesion. The resulting image is labeled as melanoma only if one of the two original lesions was labeled as melanoma. This can simulate clinical conditions with two lesions occur at the same location. We did not apply this transform at test-time. |
| **J** | Basic Set | F → D → E → C. |
| **K** | Basic Set + Erasing | F → G → D → E → C. |
| **L** | Basic Set + Elastic | F → D → H → E → C. |
| **M** | Basic Set + Mix | I → F → D → E → C. |

We applied early stopping to interrupt the training, monitoring the AUC value for the ISIC Challenge 2017 official validation dataset (150 images) for each epoch. The AUC value was calculated by averaging the predictions for 16 randomly augmented copies of each validation image, by applying the same

transforms used during training. The early stopping monitor interrupted the training when the validation AUC did not improve after 8 epochs. The final test AUC was calculated on the ISIC Challenge 2017 official test dataset (600 images) in three different ways: i) inputting the original test images to the network; ii) averaging the predictions for 64 randomly augmented copies of each test image; iii) averaging the predictions for 144 patches produced by cropping each test image as described in [14]. The weights used for testing were selected from the best AUC in the validation dataset. The validation-time and test-time augmentations followed the same transforms as the training.

For every setup, we run 6 separate trainings to reduce the effects of randomness. We used Sacred (github.com/IDSIA/sacred) to organize all experiments.

To guarantee reproducibility, we provide the documented source code used in the experiments (github.com/fabioperez/skin-data-augmentation).

## 4 Results and Discussion

### 4.1 Augmentation on Training and Testing

In this section, we discuss the results of train and test data augmentation for the proposed scenarios. Fig. 2 summarizes the results.

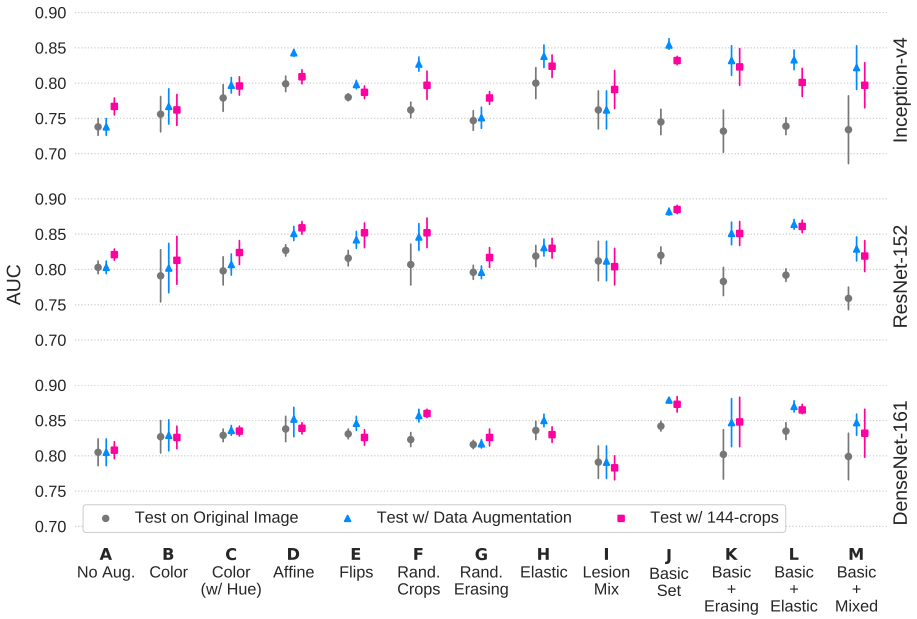


Fig. 2: Mean AUC values for augmentation scenarios. Each color and marker represent a prediction method: ● original image; ▲ test-time data augmentation (64 images); ■ 144 crops. Error bars represent the standard deviation for 6 runs. Values reported on ISIC Challenge 2017 test set.

Scenario C (saturation, contrast, brightness, and hue) resulted in better AUC than scenario B (saturation, contrast, brightness) for all three networks. However, both color transforms performed worse than scenario A (no augmentation) with 144 crops on ResNet. Geometric transforms — affine (B), random crops (F), and elastic transformations (H) — had more consistent improvements among all three networks.

Random erasing (G) shows little improvements for Inception and DenseNet, but produce worse results than scenario A (no augmentation) with ResNet. Using 144 crops was better than test data augmentation, probably due to the destructive behavior of the method. When combined with other transformations (scenario K), random erasing reduced the test AUC in comparison with scenario J (basic set combining traditional augmentations).

Scenario H (elastic) shows promising results, but when applied with other common augmentation techniques (L) also performed worse than scenario J. This may occur due to deformations produced by the combined augmentation.

Lesion mix (I and M) had worse performances when compared to other augmentations, indicating that the generated images were not useful. We presume that the produced images were not able to preserve relevant features from both source lesions.

Scenario J (basic set) yields the best AUC values for all three networks: 0.854 for Inception-v4, 0.882 for ResNet, and 0.879 for DenseNet. The top-ranked submissions for melanoma classification scored 0.874 [11], 0.870 [1], 0.868 [9]. They used, respectively, 9640, 9600, and 3444 images for training. Our method achieved a higher AUC with ResNet and DenseNet without additional data. Scenario J also has the highest AUC for the validation set in all three networks.

For every scenario, averaging augmented samples or 144 crops resulted in better performance than predicting on the original image alone. Even when no data augmentation was employed during training, 144 crops significantly increased the AUC, indicating that the model can benefit from different representations of the input image.

For ResNet and DenseNet, 144 crops has similar results to using data augmentation on test-time. Considering that we used 64 augmented samples vs 144 crops, test data augmentation can lead to faster inference.

Particularly, Inception-v4 has a worse performance with 144 crops than with test data augmentation in most scenarios. This may indicate that Inception-v4 suffers from overfitting, considering that data augmentation produced similar patterns on both training and testing.

### 4.2 Impact of Data Augmentation on Different Dataset Sizes

We trained each network on random subsets of 1500, 1000, 500, 250, and 125 images of the original data to analyze the effects of having limited training data. We generated a random subset for each one of the 6 runs. Fig. 3 summarizes the results.

Applying data augmentation (scenario J) during both training and testing noticeably improved performance for datasets with 500 or more images. Data

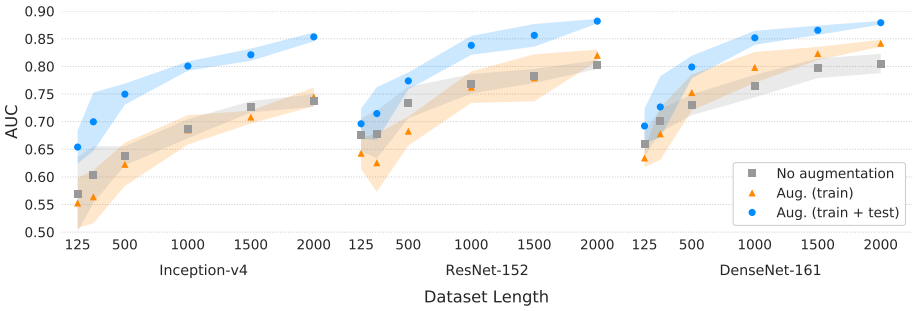


Fig. 3: Mean AUC values for different training dataset sizes, randomly sampled from the ISIC Challenge 2017 training dataset. Colors and markers represent the use of data augmentation: ■ no data augmentation; ▲ train data augmentation (scenario J); ● train and test data augmentation (scenario J, averaging each test image on 64 augmented samples). Bands represent the standard deviation for 6 runs. Values reported on ISIC Challenge 2017 test set.

augmentation for training only worsened the results for very small data sizes ($< 500$ images) and led to little or no improvement for other sizes, showing the importance of applying data augmentation during test-time.

The impact of data augmentation on Inception-v4 was more perceptible than on other networks, which may be caused by the regularizing properties of ResNet and DenseNet architectures. Training Inception-v4 with 500 images and data augmentation resulted in better performance than training with 1000, 1500 or 2000 images without augmentation. ResNet and DenseNet achieved a higher AUC with 1000 images and data augmentation than with 1500 and 2000 images without augmentation. This indicates that, in some cases, using data augmentation can be more effective than adding new training data. Nevertheless, employing data augmentation does not reduce the importance of adding new data, giving that the network can benefit from both.

## 5 Conclusion

The results highlight the positive impact of using data augmentation for training melanoma classification models. Moreover, models can also benefit from test data augmentation.

The best augmentation scenario (J), which combines geometric and color transformations, surpasses the top-ranked AUC values for the ISIC Challenge 2017 without any additional data. Fine-tuning hyperparameters and model ensembling may result in additional performance gains.

Lesion mix augmentation (scenarios I and M) have inferior results when compared with other scenarios. We implemented this augmentation through handcrafted image processing techniques, which may not be appropriate for producing reliable images. More advanced approaches, such as Generative Adversarial Networks or other generative architectures [2], might lead to better results.

## Acknowledgments

We gratefully acknowledge NVIDIA Corporation for the donation of GPUs and Microsoft Azure for the GPU-powered cloud platform used in this work. C. Vasconcelos and E. Valle are partially funded by Google Research LATAM 2017. E. Valle is also partially funded by CNPq PQ-2 grant (311905/2017-0) and Universal grant (424958/2016-3). RECOD Lab. is partially supported by diverse projects and grants from FAPESP, CNPq, and CAPES.

## References

1. Bi, L., Kim, J., Ahn, E., Feng, D.: Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. arXiv: 1703.04197 (2017)
2. Bissoto, A., Perez, F., Valle, E., Avila, S.: Skin lesion synthesis with generative adversarial networks. In: ISIC Skin Image Analysis Workshop (2018)
3. Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., et al.: Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). arXiv: 1710.05006 (2017)
4. Fornaciali, M., Carvalho, M., Bittencourt, F.V., Avila, S., Valle, E.: Towards automated melanoma screening: Proper computer vision & reliable results. arXiv:1604.04024 (2016)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE CVPR. pp. 770–778 (2016)
6. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. In: IEEE CVPR (2017)
7. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. pp. 448–456 (2015)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. pp. 1106–1114 (2012)
9. Matsunaga, K., Hamada, A., Minagawa, A., Koga, H.: Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. arXiv: 1703.03108 (2017)
10. Menegola, A., Fornaciali, M., Pires, R., Bittencourt, F.V., Avila, S., Valle, E.: Knowledge transfer for melanoma screening with deep learning. In: ISBI (2017)
11. Menegola, A., Tavares, J., Fornaciali, M., Li, L.T., Avila, S., Valle, E.: RECOD titans at ISIC challenge 2017. arXiv: 1703.04819 (2017)
12. Pham, T.C., Luong, C.M., Visani, M., Hoang, V.D.: Deep CNN and data augmentation for skin lesion classification. In: ACIIDS. pp. 573–582 (2018)
13. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI. vol. 4, p. 12 (2017)
14. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., et al.: Going deeper with convolutions. In: IEEE CVPR. pp. 1–9 (2015)
15. Valle, E., Fornaciali, M., Menegola, A., Tavares, J., Bittencourt, F.V., Li, L.T., Avila, S.: Data, depth, and design: Learning reliable models for melanoma screening. arXiv: 1711.00441 (2017)
16. Vasconcelos, C.N., Vasconcelos, B.N.: Experiments using deep learning for dermoscopy image analysis. Pattern Recognition Letters (2017)
17. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. arXiv: 1708.04896 (2017)