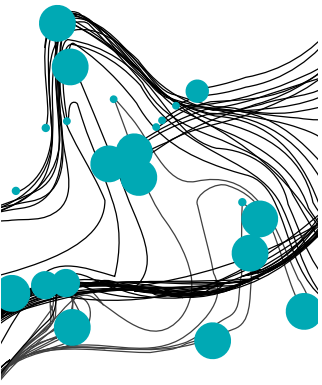
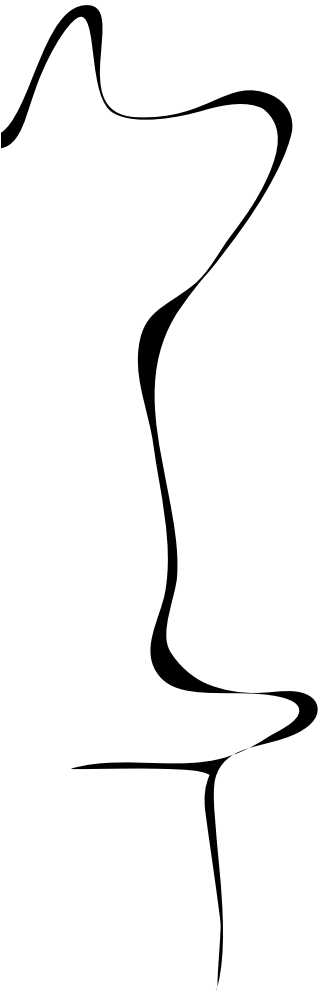


.92732

DMB

DATA MANAGEMENT AND BIOMETRICS



EVALUATING THE CORRUPTION ROBUSTNESS OF CONVOLUTIONAL NEURAL NETWORKS FOR THE TASK OF CLASSIFYING SKIN DISEASES

Tijmen Westeneng

BACHELOR THESIS ASSIGNMENT

Committee:
Estefanía Talavera Martínez
Luuk Spreeuwers
Faizan Ahmed

June, 2024

2024DMB0005
Data Management and Biometrics
EEMathCS
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

Abstract

Skin disease classification by machine learning models is an upcoming field of research that shows great potential to assist dermatologists. These techniques are, however, still vulnerable to corruptions that can be present in the dermoscopic images. In this paper, we tested two promising methods to improve the robustness against corruptions: augmenting the training datasets with corrupted images and preprocessing the images with Contrast Limited Adaptive Histogram Equalization (CLAHE). We found that the presence of corrupted images in the training datasets can greatly improve the corruption robustness while CLAHE harms the classification accuracy of the models when faced with corrupted images. Our benchmarks can be used as a starting point to further develop AI models that can be used as reliable diagnostic tools.

I. Introduction

Skin cancer is the most common type of cancer worldwide, for instance 1 in 5 Americans will develop skin cancer by the age of 70 [1]. That leads to a huge number of cases, which is mainly due to the fact that, although most people know melanoma, there are actually more types of skin cancer. The good news is that, when detected early, the 5-year survival rate for melanoma is 99 percent. This implies it is extremely important to diagnose skin cancer as early as possible, which means each technique that improves this diagnosis time could be very valuable.

Machine learning or deep learning is one of these techniques and has entered the field over the last decade or so. Deep learning uses Convolutional Neural Networks (CNN's) to classify certain types of skin cancer from images without the need for any medical expert to intervene. A lot of research has already been done in this field and it has been found that these CNN's can be as reliable or even superior to human experts. For example, in a study done in 2019, a deep learning algorithm outperformed 136 out of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task [2]. This shows the technique has great potential to assist doctors in their work.

However, where these CNN's lack reliability is when it comes to corruptions in the images, like blur or noise. A human expert could easily ignore these corruptions and still correctly diagnose the patient, but CNN's have a lot more trouble understanding the picture if there are small or larger differences present compared to the images they were trained on. This means the robustness of the CNN's is rather low, which can be a great issue if they are actually used to diagnose real patients [3].

That's where this research comes in. We have looked into the effects of corruptions in images on the accuracy of CNN models and tried to find ways to improve this accuracy. This research will not only focus on skin cancer but also on other types of (less harmful) skin diseases to broaden the scope.

We specifically looked at two ways of improving the corruption robustness of machine learning models classifying skin diseases. Firstly, adding corrupted data to the training datasets. This tries to improve accuracy by already exposing the machine learning models to different amounts of corrupted images. This has been proven in earlier research to be able to increase classifier robustness [4]. The second method is preprocessing the images with Contrast Limited Adaptive Histogram Equalization (CLAHE). This is a variant of adaptive histogram equalization that is meant to increase the contrast of an image and thus increase the visibility of details in an image (see Figure 1). It is widely used for medical imaging applications and therefore also has the possibility of improving the image classification robustness of machine learning models [5].

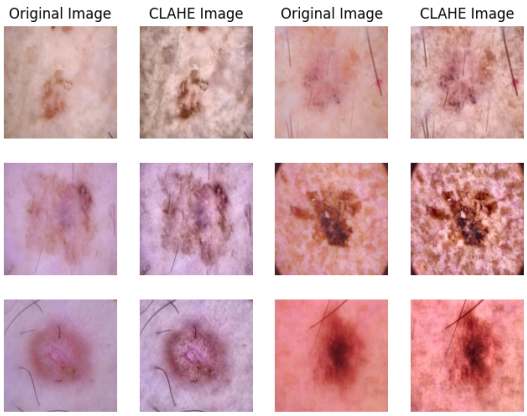


Fig. 1. Effects of CLAHE being applied to images from the HAM10000 dataset [6] used during this study.

Research Questions

This research focuses on two research questions. What is the influence of adding corrupted images to the training datasets of CNN's classifying skin diseases on the corruption robustness of these models? And what is the impact of preprocessing the images with Contrast Limited Adaptive Histogram Equalization on the corruption robustness?

II. Scientific Background

A. Benchmarking Neural Network Robustness To Common Corruptions And Perturbations [7]

This paper tests the robustness of multiple CNN architectures by feeding them with images with corruptions and perturbations. The datasets created by adding these corruptions and perturbations build on the already existing ImageNet dataset [8]. This dataset consists of more than a million images divided over 1000 object classes and is often used as a standard dataset for training new CNN architectures and models. The authors of the paper use the created ImageNet-C (Corruptions) and ImageNet-P (Perturbations) datasets to test the error rate of different CNN architectures that have been created over the years like AlexNet [9] and ResNet [10]. They found that the robustness for both corruptions and perturbations has improved over time, but that this is mainly due to general accuracy improving too. The actual improvement in corruption robustness seems to even have declined compared to older models. The authors also come up with multiple ways to increase the robustness of the models. Examples of these are Contrast Limited Adaptive Histogram Equalization (CLAHE), multiscale networks, which operate across multiple feature map scales, and using larger networks like DenseNet-121 [11]. In the end the authors point out that further testing and investigating new CNN architectures for robustness can be important, especially now the clean model accuracy reaches its limits. The datasets they built have been replicated using skin disease pictures for our research and their findings in using CLAHE to improve robustness laid the groundwork for testing it during this project.

B. A benchmark for neural network robustness in skin cancer classification [12]

This paper focuses especially on the model robustness of models classifying skin cancer, which means it's closer to our research questions. The paper was heavily inspired and adapted from the previously discussed ImageNet paper and also created a corruption dataset, called SAM-C, and a perturbations dataset, called SAM-P. The authors trained four different models (AlexNet, VGG16+BN [13], ResNet50, and DenseNet121) on multiple skin cancer datasets without any corruptions and perturbations. Then they tested the error rates when the models were fed with both clean images, images with corruptions, and images with perturbations. Just like in the previous paper, they found that although the newer networks like ResNet50 performed better on the clean images, the older AlexNet actually has a better perturbation robustness. The relative error rate for images with corruptions (the one that compensates for the overall better accuracy) was actually best in DenseNet121. Although our research also focuses on skin cancer, we broaden the scope by also including different skin diseases than skin cancer. We also expand on their research by specifically testing two ways to improve corruption robustness instead of just comparing the corruption robustness across different models. We do, however, use largely the same methods as they used during their research.

III. Methodology

A. The Dataset

The dataset used to test our research questions is the Human Against Machine with 10000 training images dataset (HAM10000 in short) [6]. It was part of a challenge hosted by the International Skin Imaging Collaboration (ISIC) in 2018 which aimed to assess the current state of skin cancer classification with the help of artificial intelligence [14]. The dataset consists of 10015 dermoscopic images divided into 7 classes of skin diseases. As can be seen in Table I the dataset is very unbalanced with the class nv having 6705 images while df has 115. In Figure 2 an example picture from each class of the dataset can be found. The original dimensions of the images are 600 pixels by 450 pixels but for this research they have been resized to 224 pixels by 224 pixels because these are the input dimensions of the machine learning models.

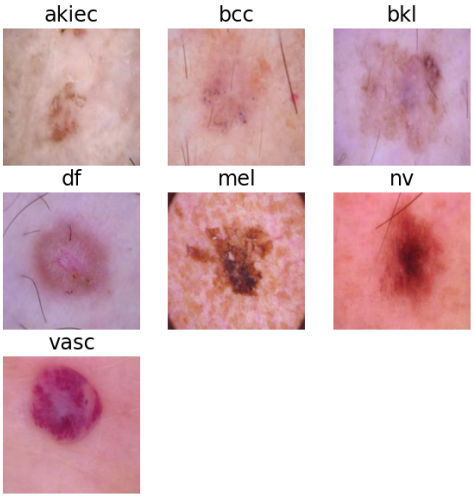


Fig. 2. Example image of each of the classes in the HAM10000 dataset [6].

TABLE I
AN OVERVIEW OF THE HAM10000 DATASET [6]

Class Name	Name Skin Disease	Number of Images
akiec	Actinic Keratoses	327
bcc	Basal Cell Carcinoma	514
bkl	Benign Keratosis	1099
df	Dermatofibroma	115
mel	Melanoma	1113
nv	Melanocytic Nevi	6705
vasc	Vascular Skin Lesions	142

B. Corruptions

To accurately evaluate the robustness of certain machine learning models, we had to develop a list of image corruptions to use on the images. Preferably, these corruptions resemble ones that could occur during real-life use cases. The corruptions used in this research were based on the corruptions used in both the earlier mentioned ImageNet robustness paper [7] as well as the skin cancer robustness paper [12]. The ImageNet paper already provided a library to recreate the corruptions they used, which made it easier to also implement those corruptions into our dataset. The other corruptions were implemented by us and those implementations can be found on this project's GitHub page via the url *https://github.com/TijmenWesteneng/BachelorThesis*. The corruptions can be divided into four categories: noise, blur, dermoscopy, and digital. Each corruption has a total of 5 severity levels, where 1 indicates a low corruption level and 5 is the highest corruption level. In Figure 3 the 14 different corruptions applied to an example image from the dataset can be found. Furthermore, in Section VII-B in the appendix a further overview and descriptions of the 14 different corruptions can be found.

C. Contrast Limited Adaptive Histogram Equalization

Adaptive histogram equalization works by stretching the contents of an image histogram across the whole possible range. It does this by first dividing the image into tiles and then equalizing the histogram per tile whereafter it merges the tiles together again. This increases contrast but also increases noise. To solve this issue Contrast Limited Adaptive Histogram Equalization (CLAHE) clips the histogram at a predefined value and redistributes the contents uniformly over the histogram (see figure 4). This improves contrast without greatly amplifying the noise [15].
CLAHE was implemented in our research using the OpenCV library [16]. In most cases, CLAHE is applied to a grayscale image but because the color contents of the images are important in our case we first converted the image to LAB color space and then applied CLAHE to the lightness channel. We used a clip limit of 2 and a tile grid size of 8. Figure 1 shows the result of a CLAHE transformation on a few images from our dataset.

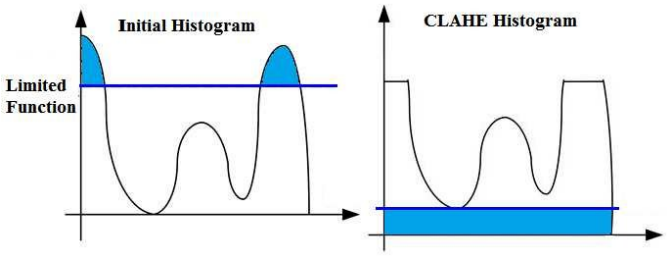


Fig. 4. Representation of the effects of CLAHE on an image histogram [17].

D. Creating The Training & Testing Datasets

The research done in this thesis revolves around changing the data used for a machine learning model to hopefully improve its corruption robustness. To achieve this, first the images in the dataset were all resized to 224 by 224 pixels to be usable for the machine learning model. The complete dataset was then split up into a test set (20%) and a training/validation set (80%). The test and training sets were then individually treated in the following ways:

1) Test Set: The test set was corrupted with all 14 different corruptions, each for all 5 different severities. This created a dataset that could be used to test the accuracy of the machine learning model for each different corruption and severity. To evaluate the influence of Contrast Limited Adaptive Histogram Equalization both the corrupted and the non-corrupted test sets were treated with CLAHE to create two new test sets.

2) Training Set: The training set was first corrupted in the same way as the test set to create a new corrupted set including all different corruptions and severities. To then create the different corrupted datasets, the non-corrupted dataset was taken and added to this were a number of random images from each corruption from the corrupted set. This was done in such a way to end up with a new dataset that had a certain ratio between the number of corrupted images in the set and the total number of images in the set. This ratio is called the corruption ratio. Section VII-C in the appendix shows a representation of two example training datasets. Five sets were created with a corruption ratio of 0.5 and a corruption severity ranging from 1 to 5 and another five sets were created with a corruption severity of 3 and a corruption ratio

ranging from 0.1 to 0.5.
The training sets were then augmented by rotating (in steps of 90 degrees) and mirroring / flipping the images. These augmentations created eight different instances of each image. In our case, this was done for all classes except the nv class, because this class was already largely overrepresented in the dataset. However, when working with a different dataset that is more balanced, augmenting can be done for all classes.
To test the influence of Contrast Limited Adaptive Histogram Equalization, all training sets were copied and then treated with CLAHE to end up with new training sets with the same characteristics and images as the non-CLAHE sets but with CLAHE applied.
After this all training sets were split up into a training set (80%) and a validation set (20%).

E. Implementation details

Nowadays there is an abundance of machine learning architectures available to use for testing. Based on literature research we chose to use the ResNet architecture [10] because of its good balance between accuracy and size [18] [19] [20]. It also was one of the better-performing models in the earlier discussed robustness papers [7] [12]. We chose to use ResNet-18 as the main model used for testing because it is the smallest out of all the ResNet architectures and will therefore take the shortest time to run. Because this research focuses on the difference between different configurations it is not necessary to use larger models to achieve a greater accuracy as long as the same model is used in each instance. ResNet-50 was used as a baseline benchmark, which use will be explained in Section III-F.
The ResNet models were implemented in Python using the PyTorch library [21] and transfer learning was used to make them able to classify the HAM10000 images. This means that we took the weights from a ResNet network that was pre-trained on the ImageNet [8] dataset and then replaced the last (fully-connected) layer with a new layer that could identify 7 different classes. We then froze all weights except the ones from the newly added layer so that the network only trained these weights. This transfer-learning approach greatly reduces training time by making use of the fact that a large part of an image-classification network can remain identical when changing its specific classification task [22].
Considering the hyperparameters used in training the machine learning models, we decided to keep them equal

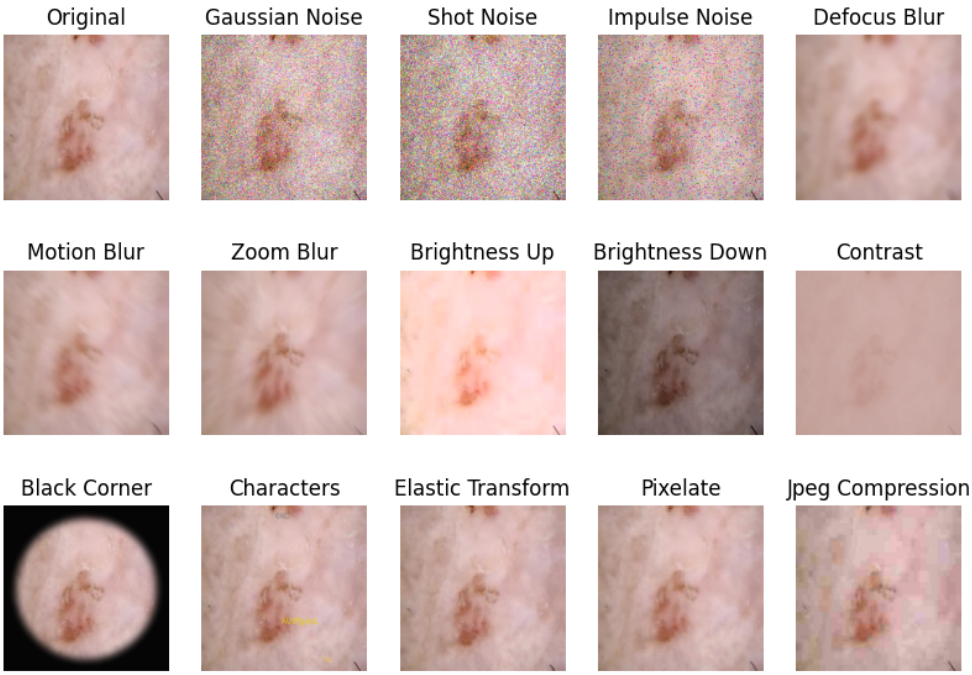


Fig. 3. Overview of all 14 corruptions used and the original image for reference (class: akiec & severity: 3).

for all instances. A batch size of 32 and a learning rate of 0.001 was used. The models were trained for 20 epochs or until the balanced accuracy of the network on the validation set didn't improve for 3 epochs. Then the model weights from the last best balanced accuracy were saved. The cross-entropy loss function was used as this is a common loss function used in image classification [23]. To compensate for the imbalance in the dataset we assigned weights to each class in the loss function that prioritizes the smaller classes.

F. Evaluation

To accurately determine the corruption robustness of a machine learning model we had to define a way to measure this corruption robustness. This was done by first calculating a model's balanced accuracy score on the uncorrupted test set. The balanced accuracy is similar to the normal accuracy of a model but compensates for the possible class unbalance in the test set, which is a great factor in our case. This balanced accuracy is then subtracted from 1 to get the clean balanced error rate: BE^f_{clean}. After this the balanced error rate $BE^f_{s,c}$ for each corruption type c and severity level $s(1 \leq s \leq 5)$ is calculated. We then use ResNet-50 trained on an uncorrupted training dataset as a baseline to calculate an average of the balanced error rate over the five severities that compensates for the fact that not each corruption has the same difficulty. This then gives us the following formula for the balanced corruption error BCE^f_c for a single corruption type c:

$$BCE^f_c = (\sum_{s=1}^{5} BE^f_{s,c}) \ / \ (\sum_{s=1}^{5} BE^{ResNet-50}_{s,c}) \qquad (1)$$

This balanced corruption error is then averaged out over all the different corruption types to calculate the mean balanced corruption error $mBCE$.
Although this mean balanced corruption error already is a good indicator of the corruption robustness of a model, it can happen that a model performs relatively well on the uncorrupted dataset and therefore also performs well on the corrupted datasets. A good statistic that only focuses on robustness would disregard this edge that a model has due to its good general accuracy and would instead focus on the difference between the clean error rates and the corruption error rates. This is where the relative mean balanced corruption error $relative\ mBCE$ comes in. It is calculated by subtracting the clean balanced error rates from the balanced corruption errors:

$$relative\ BCE^f_c = (\sum_{s=1}^{5} BE^f_{s,c} - BE^f_{clean}) \ / \ (\sum_{s=1}^{5} BE^{ResNet-50}_{s,c} - BE^{ResNet-50}_{clean}) \qquad (2)$$

These balanced corruption errors are then averaged across all corruption types, just as with the normal $mBCE$, to get the $relative\ mBCE$.

IV. Results

After averaging the results over 5 runs of training and testing we end up with the following results.

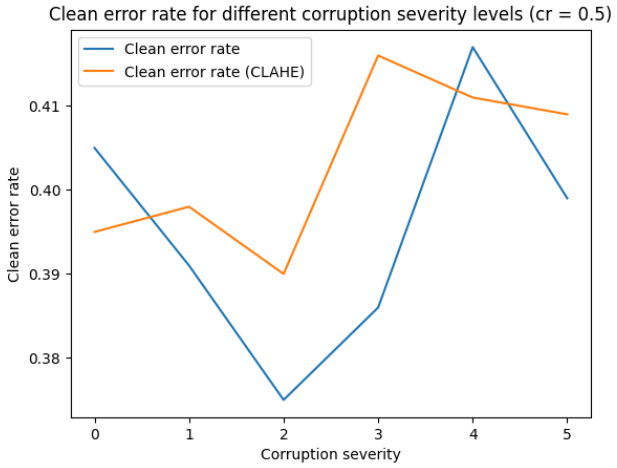


Fig. 5. Clean error rate of models trained with different corruption severities and possibly pre-processed with CLAHE.

Fig. 6. Clean error rate of models trained with different corruption ratios and possibly pre-processed with CLAHE.

Figure 5 and figure 6 show the clean error rates for different corruption severities and corruption ratios respectively. As can be seen, no trend is visible in the lines and the total range of values is relatively small (between 0.375 and 0.415). For different corruption severities, non-CLAHE is mostly outperforming CLAHE, but for different corruption ratios it is the other way around.

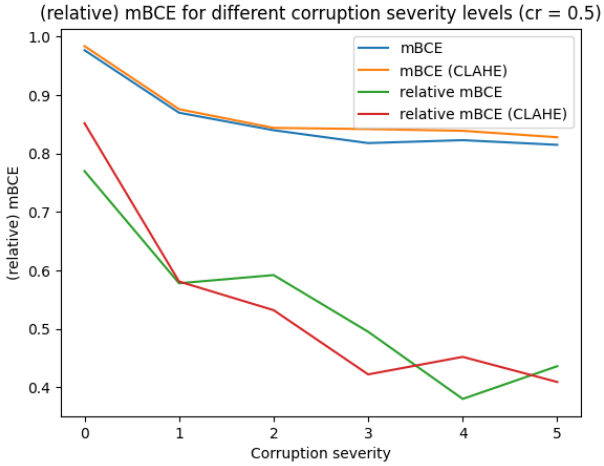


Fig. 7. (Relative) mBCE of models trained with different corruption severities and possibly pre-processed with CLAHE (corruption ratio (cr) = 0.5).

Figure 7 and Figure 8 show the (relative) mBCE values for different corruption severities and corruption ratios respectively. There is a clear downward trend visible in both plots, which indicates a better corruption robustness. The difference between CLAHE and non-CLAHE is relatively small with non-CLAHE mostly outperforming CLAHE when it comes to mBCE. To get a better idea of the reason behind CLAHE often underperforming we looked at the difference between images that are frequently wrongly classified by CLAHE models and rightly classified by non-CLAHE models and vice-versa. An overview of a few of these images can be found in Section VII-H of the appendix. We also looked into the classification improvement caused by training with corrupted data by listing a few example images that are correctly classified by the models trained on corrupted data but incorrectly classified by the models trained without corrupted data. This overview can be found in Section VII-I of the appendix.

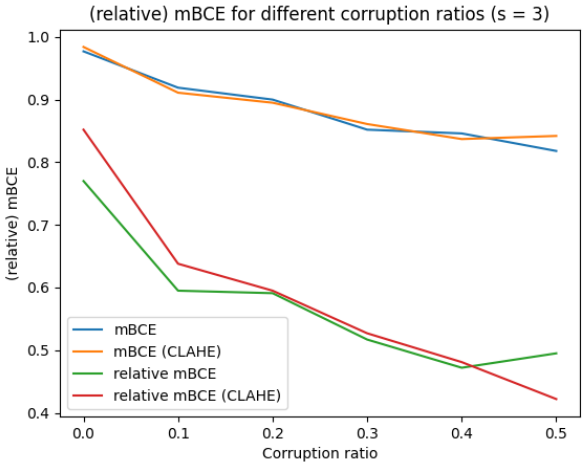


Fig. 8. (Relative) mBCE of models trained with different corruption ratios and possibly pre-processed with CLAHE (corruption severity (s) = 3).

To get a better overview of the corruption robustness to specific corruptions of the different models Table II shows the error rates averaged over all severities for each particular corruption. These values have not been compensated by a baseline, but are just the standard mean values. What is immediately noticeable is that the standard model is especially vulnerable to the different noise corruptions, proven by the fact that the top three error rates are all of the corruption-type noise. It's also noticeable that CLAHE is performing (slightly) worse on almost every corruption in both the models trained on the non-corrupted and the corrupted datasets. In Section VII-F of the appendix a table highlighting the difference in performance can be found.

Table III shows the top-1 accuracies per corruption of the best-performing models per severity. As can be seen, average performance degrades when severity increases and the best performing models are trained on severities close to the ones they were tested on. Also, none of the best-performing models were models whose images were pre-processed with CLAHE. Section VII-G of the appendix includes a full table summarizing the

TABLE II
AVERAGE ERROR RATES PER CORRUPTION SORTED BY THE NON-CORRUPTED TRAINING DATASET ERROR RATES
(MODEL: NUMBER = SEVERITY, C = CLAHE, RATIO = 0.5)

Model	Clean	Shot.	Impul.	Gauss.	Contrast	Bl.Cor.	Bri.Do.	JPEG	Defoc.	Motion.	Zoom.	Bri.Up	Pixel.	Elast.	Char.
0	0.41	0.83	0.83	0.81	0.73	0.73	0.68	0.65	0.60	0.58	0.55	0.55	0.48	0.46	0.44
0C	0.40	0.84	0.83	0.84	0.74	0.62	0.66	0.65	0.61	0.60	0.58	0.55	0.51	0.48	0.44
5	0.40	0.62	0.59	0.59	0.60	0.51	0.53	0.54	0.51	0.48	0.50	0.50	0.44	0.43	0.43
5C	0.41	0.65	0.61	0.64	0.64	0.50	0.52	0.55	0.51	0.49	0.52	0.48	0.46	0.44	0.43

accuracies of all 12 models with a corruption ratio of 0.5.

To further evaluate the performance of certain models the appendix also discusses a few confusion matrices. Section VII-D discusses four confusion matrices of non-corrupted, corrupted, and CLAHE models on a clean test set and Section VII-E discusses two confusion matrices of models tested on a corrupted test set.

V. Discussion

A. Corruption Robustness

Our results show that classifier reliability greatly declines when presented with corrupted images. As can be seen in Table II, models trained on purely non-corrupted data have a twice as high error rate on the worst corruption than on a clean test set and the average accuracy already decreases from 59.5% to 46.6% when confronted with severity 1 corruptions (see Section VII-G of the appendix). There is however a large difference visible in the impact of the different corruptions. For example, characters, the corruption with the smallest impact, only decreases performance by about 5%, while shot noise decreases performance by about 71%. This corresponds with what was found in earlier research [7] and is also to be expected. Noise can greatly change an image, while a few random characters in a photo don't change the contents a lot as long as they don't interfere with the important details of the image.
What is interesting to see is that the biggest improvement in classifier performance is the black corner corruption. Its error rate improved from 0.73 to 0.51 after it was trained on severity 5 and ratio 0.5 corruptions. This shows that training on corrupted data has taught the classifier not to focus on the corners of the image, but just on the middle of the image where the disease is the most visible.

B. Training with corrupted data

Let's take a look at our first research question: What is the influence of adding corrupted images to the training datasets of CNN's classifying skin diseases on the corruption robustness of these models? Our results prove that training with corrupted data improves model robustness, which was also what we expected to see. When we look at the difference between training without corrupted data and with corrupted data of severity 5 we see a decrease in mBCE from 0.98 to 0.82 (16.3%, see Figure 7). The relative mBCE even decreases from 0.77 to 0.44, a decrease of 42.9%. There doesn't seem to be much of an influence on the performance on the clean test set, although the performance does decrease slightly in some instances (see Figure 5 and 6).
When it comes to the influence of the corruption severity of the images in the training dataset on the corruption robustness of the model we can see that a higher corruption severity results in both a better mBCE and relative mBCE. For example, as shown in Figure 7 the average mBCE of non-CLAHE models for corruption severity 1 is 0.87, while the mBCE for severity 5 is 0.82. This is a decrease of 5.8%. The relative mBCE decreases even further from 0.58 to 0.44, a decrease of 24.1%. However, it is interesting to note that Table III shows that for lower severities, the best-performing models are not those trained on the highest severity, but rather models trained on images with a severity closer to the one they were tested on. This indicates that models perform best on corruptions that resemble the ones they were trained on, which is to be expected since those corruptions were part of their training data. However, on average over all severities, the severity 5 models are still the best-performing models. This is caused by the accuracy of models trained on lower severity images degrading greatly when faced with more severe corruptions, while models trained on higher severities are still quite good at working with lower severities (see Section VII-F of the appendix). Overall, models are good at classifying

TABLE III
Models with the best top-1 accuracies per severity (model: number = severity, ratio = 0.5)

Severity	Model	Corruption (top-1 acc. in %)														Avg.
		Gauss.	Shot.	Impul.	Defoc.	Motion.	Zoom.	Bl.Cor.	Char.	Bri.Up	Bri.Do.	Contr.	Elast.	Pixel.	JPEG.	
1	2	50.0	51.5	50.5	59.9	60.7	55.5	50.7	60.8	58.7	57.4	52.5	62.9	60.0	56.2	56.2
2	2	52.7	46.5	46.8	55.7	55.8	52.4	50.4	59.4	53.0	48.5	48.0	52.6	62.7	51.1	52.5
	3	49.5	47.8	45.5	56.4	54.8	52.5	49.7	58.9	54.0	52.2	48.2	50.8	62.4	51.9	52.5
3	3	44.7	42.4	44.3	51.2	53.1	50.6	49.4	57.1	52.6	41.9	41.0	63.3	58.2	50.7	50.0
4	5	41.2	38.9	38.6	44.9	47.3	49.4	48.4	56.2	48.8	40.4	35.0	59.2	49.9	40.0	45.6
5	5	33.0	31.5	35.6	45.1	47.1	48.2	47.6	54.5	46.7	34.5	34.1	55.9	50.7	34.8	42.8

equal or lower severities than the ones they were trained on. Therefore, it is advisable to estimate the expected severities of the corruptions a model will encounter during its use and adjust the training data accordingly. An alternative approach could be to train with corruptions with multiple different severities, but the impact of such a configuration should first be researched in future studies. As shown in Figure 8, a higher corruption ratio exhibits the same trend as a higher corruption severity. Between a corruption ratio of 0.1 and 0.5, the mBCE decreases from 0.92 to 0.82, a decrease of 10.9%. The relative mBCE even decreases from 0.60 to 0.50, a decrease of 16.7%. The results even seem to suggest that corruption ratios beyond 0.5 would result in even better corruption robustness, but these ratios weren't tested during this research so their effects remain uncertain. The results can be explained by the fact that when the model can correctly classify more corrupt images, it will also be able to better classify corrupt images when presented with them during testing. The relative mBCE does increase from the ratio 0.4 to 0.5 (from 0.48 to 0.50, see Figure 8), but this is only caused by a steep decline in the clean error rate (see Figure 6) and since there is no trend visible in these lines this result does not bear any significance.

C. Contrast Limited Adaptive Histogram Equalization

With our results, we can also answer the second research question: What is the impact of preprocessing the images with Contrast Limited Adaptive Histogram Equalization (CLAHE) on the corruption robustness? If we look at Figures 7 and 8 we see that the CLAHE models are often underperforming the non-CLAHE models. The mBCE of the CLAHE models in the corruption severity graph (Figure 7) is never better than that of the non-CLAHE models and also the relative mBCE is often worse. At the best performing, severity 5, models the non-CLAHE models achieve an mBCE of 0.82, and the CLAHE models an mBCE of 0.83, which is an increase (and thus decrease in performance) of 1.2%. When we look at the corruption ratio (Figure 8) the results are less definitive, with the mBCE lines crossing each other multiple times. However, because we concluded earlier that a corruption ratio of 0.5 is performing the best, the results of the other ratios don't have a lot of significance in our conclusion. CLAHE not improving the corruption robustness is not what we expected to find. The earlier mentioned ImageNet paper [7] found an improvement of the mBCE of almost 3% when using CLAHE. This difference could be explained by the fact that ImageNet is a vastly different dataset than HAM10000. Although it has many more classes (1000 against 7), the classes are much more distinctive than the HAM10000 classes. Distinguishing a bird from a dog relies on much less detail in the image than differentiating between skin diseases. We see in some cases that the CLAHE is amplifying details in an image that are not part of the disease itself and thus could hinder the recognition of the disease (see Section VII-H in the appendix).

D. Future research

Our findings leave quite some room for future research. One of the most interesting things that could be looked at is the effect that training with corruptions has on the focus regions of the neural network. We already saw that the model showed a large improvement in the black corner corruption and we hypothesized this could be because it learned to focus on the center regions of the image. This could be further investigated by looking at the class activation maps of the specific convolutional neural networks. This has been done in other image classification fields, but not yet for skin diseases [24] [25].

Next to this, further research into different methods of histogram equalization could lead to different insights than those found during this study. There exist other types of histogram equalization that try to only increase the contrast in the regions of interest of the image. This could be a solution for the amplification of non-important details that we found during our research. Examples of these approaches are Selective Energy-Based Histogram Equalization (SEBHE), which has been proven to improve contrast in breast mammograms [26], and Selective Apex Adaptive Histogram Equalization (SLAAHE), which has been used to improve contrast in ultrasound and MRI images [27]. Applying methods like these to dermoscopic images could lead to an improvement in corruption robustness that we didn't find in this study.

VI. Conclusion

In this paper, we examined the impact of training with corrupted images and pre-processing images with Contrast Limited Adaptive Histogram Equalization (CLAHE) on the corruption robustness of machine learning models classifying skin diseases. This was done to improve the diagnostic reliability of these models, so they are less impacted by imperfections in images and can therefore operate in a variety of circumstances. We found that the performance of models that are not trained with corrupted images or whose images are not pre-processed with CLAHE declines greatly

when presented with corrupted images. Even severity 1 corruptions already decrease the accuracy from 59.5% to 46.6%. These models can therefore not be trusted to classify skin diseases correctly when a corruption, in any severity, is present. Our findings indicate that augmenting the training datasets with corrupted images can greatly improve the corruption robustness. We found a decrease in mBCE and relative mBCE of 16.3% and 42.9% respectively when models are trained with severity 5 corruptions and a corruption ratio of 0.5. Adding more corrupted images and images with a greater corruption severity during training has a positive effect on the classification of corrupted images during testing. However, it is important to note that models perform best on severities that resemble the ones in their training set, so it is important to choose severities based on the expected use cases of the model. The improvement in the classification of corrupted images doesn't compromise the classification reliability of non-corrupted images.

Pre-processing the images with Contrast Limited Adaptive Histogram Equalization has a minor negative effect on the recognition of corrupted images. We found a decrease in performance of 1.2% on the most robust, severity 5 and ratio 0.5, models. CLAHE should therefore, according to our research, not be used as a method to improve corruption robustness.

Our work can be used as a starting point to further investigate methods to improve the corruption robustness of machine learning models classifying skin diseases. Where our research focused on alternating the training data, there are also other promising methods, like multiscale networks, that focus on alternating the model structure. Using our research and these methods, machine learning models can hopefully once be reliably used as a dermatological diagnostic tool.

References

[1] Skin Cancer Foundation. *Skin Cancer Facts & Statistics*. Feb. 2024. URL: https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/.

[2] Titus J Brinker et al. "Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task". In: *European Journal of Cancer* 113 (2019), pp. 47–54. ISSN: 0959-8049. DOI: https://doi.org/10.1016/j.ejca.2019.04.001. URL: https://www.sciencedirect.com/science/article/pii/S0959804919302217.

[3] Aharon Azulay and Yair Weiss. "Why do deep convolutional networks generalize so poorly to small image transformations?" In: *Journal of Machine Learning Research* 20.184 (2019), pp. 1–25. URL: http://jmlr.org/papers/v20/19-519.html.

[4] Roman C Maron et al. "Robustness of convolutional neural networks in recognition of pigmented skin lesions". In: *European Journal of Cancer* 145 (2021), pp. 81–91. ISSN: 0959-8049. DOI: https://doi.org/10.1016/j.ejca.2020.11.020. URL: https://www.sciencedirect.com/science/article/pii/S0959804920313575.

[5] Stephen M Pizer et al. "Adaptive histogram equalization and its variations". In: *Computer Vision, Graphics, and Image Processing* 39.3 (1987), pp. 355–368. ISSN: 0734-189X. DOI: https://doi.org/10.1016/S0734-189X(87)80186-X. URL: https://www.sciencedirect.com/science/article/pii/S0734189X8780186X.

[6] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions". In: *Scientific Data* 5.1 (2018), p. 180161. ISSN: 2052-4463. DOI: 10.1038/sdata.2018.161. URL: https://doi.org/10.1038/sdata.2018.161.

[7] Dan Hendrycks and Thomas G Dietterich. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations". In: *CoRR* abs/1903.12261 (2019). URL: http://arxiv.org/abs/1903.12261.

[8] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F Pereira et al. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[10] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). URL: http://arxiv.org/abs/1512.03385.

[11] Gao Huang, Zhuang Liu, and Kilian Q Weinberger. "Densely Connected Convolutional Networks". In: *CoRR* abs/1608.06993 (2016). URL: http://arxiv.org/abs/1608.06993.

[12] Roman C Maron et al. "A benchmark for neural network robustness in skin cancer classification".

In: *European Journal of Cancer* 155 (2021), pp. 191–199. ISSN: 0959-8049. DOI: https://doi.org/10.1016/j.ejca.2021.06.047. URL: https://www.sciencedirect.com/science/article/pii/S0959804921004421.

[13] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: (Sept. 2014). URL: http://arxiv.org/abs/1409.1556.

[14] Noel C F Codella et al. "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)". In: *CoRR* abs/1902.03368 (2019). URL: http://arxiv.org/abs/1902.03368.

[15] Parthiban Marimuthu. *Image Contrast Enhancement Using CLAHE*. July 2023. URL: https://www.analyticsvidhya.com/blog/2022/08/image-contrast-enhancement-using-clahe/.

[16] OpenCV team. *Open Computer Vision Library*. 2024. URL: https://opencv.org/.

[17] Omar Boudraa, Walid Khaled Hidouci, and Dominique Michelucci. "Degraded Historical Documents Images Binarization Using a Combination of Enhanced Techniques". In: (Jan. 2019). URL: http://arxiv.org/abs/1901.09425.

[18] Aqeel Anwar. *Difference between AlexNet, VGGNet, ResNet, and Inception*. June 2019. URL: https://towardsdatascience.com/the-w3h-of-alexnet-vggnet-resnet-and-inception-7baaaecccc96.

[19] Enrico Randellini. *Image classification: ResNet vs EfficientNet vs EfficientNet_v2 vs Compact Convolutional Transformers*. Jan. 2023. URL: https://medium.com/@enrico.randellini/image-classification-resnet-vs-efficientnet-vs-efficientnet-v2-vs-compact-convolutional-c205838bbf49.

[20] Vaibhav Kumar. *MobileNet vs ResNet50 – Two CNN Transfer Learning Light Frameworks*. June 2020.

[21] The PyTorch Foundation. *PyTorch*. 2024. URL: https://pytorch.org/.

[22] Asmaul Hosna et al. "Transfer learning: a friendly introduction". In: *Journal of Big Data* 9.1 (2022), p. 102. ISSN: 2196-1115. DOI: 10.1186/s40537-022-00652-w. URL: https://doi.org/10.1186/s40537-022-00652-w.

[23] Sivaramakrishnan Rajaraman, Ghada Zamzmi, and Sameer K. Antani. "Novel loss functions for ensemble-based medical image classification". In: *PLoS ONE* 16.12 December (Dec. 2021). ISSN: 19326203. DOI: 10.1371/journal.pone.0261307.

[24] Bolei Zhou et al. "Learning Deep Features for Discriminative Localization". In: *CoRR* abs/1512.04150 (2015). URL: http://arxiv.org/abs/1512.04150.

[25] Anh Pham Thi Minh. *Overview of Class Activation Maps for Visualization Explainability*. 2023.

[26] Nabila Elsawy, Mohammed Sayed, and Fathi Farag. "Selective energy-based histogram equalization for mammograms". In: June 2017, pp. 115–118. DOI: 10.1109/JEC-ECC.2017.8305791.

[27] Manas Sarkar and Ardhendu Mandal. "SLAAHE: Selective Apex Adaptive Histogram Equalization". In: *Franklin Open* 3 (2023), p. 100023. ISSN: 2773-1863. DOI: https://doi.org/10.1016/j.fraope.2023.100023. URL: https://www.sciencedirect.com/science/article/pii/S2773186323000178.

VII. Appendix

A. Artificial Intelligence statement

During the preparation of this work, the author used Grammarly and Overleaf to improve the spelling and grammar of sections of this paper. The author also used ChatGPT to help with debugging the Python code written by the author himself. No code used for the final product or sections of the paper submitted was fully written by ChatGPT or any other AI tool. After using the AI tools/services, the author reviewed and edited the content as needed and takes full responsibility for the content of the work.

B. Further overview of corruptions

TABLE IV
Overview and short description of the used corruptions

Category	**Corruption**	**Description**
Noise	Gaussian	Appears in low-lighting conditions
	Shot	Electronic noise that is caused by the discrete nature of light
	Impulse	Salt-and-pepper noise caused by bit errors
Blur	Defocus	The image is out of focus
	Motion	The camera moves quickly in some direction
	Zoom	The camera moves quickly towards the object
Dermoscopy	Black corner	Black image corners caused by the dermatoscope
	Characters	Letters, numbers and punctuation marks which might be overlaid by the camera
Digital	Brightness up/down	Varies with daylight intensity or the camera
	Contrast	Depends on lighting conditions and object colour
	Elastic Transform	Transformations stretching or contracting small image regions
	Pixelate	Artefacts occurring when upsampling low-resolution images
	Jpeg compression	Artefacts occurring due to lossy image compression format

Table IV shows an overview of all the fourteen different corruptions used during this study. As can be seen, they are divided into four categories: noise, blur, dermoscopy, and digital. The corruptions were chosen because they can occur during the real-life photographing and processing of dermoscopic images.

C. Representation of training dataset creation

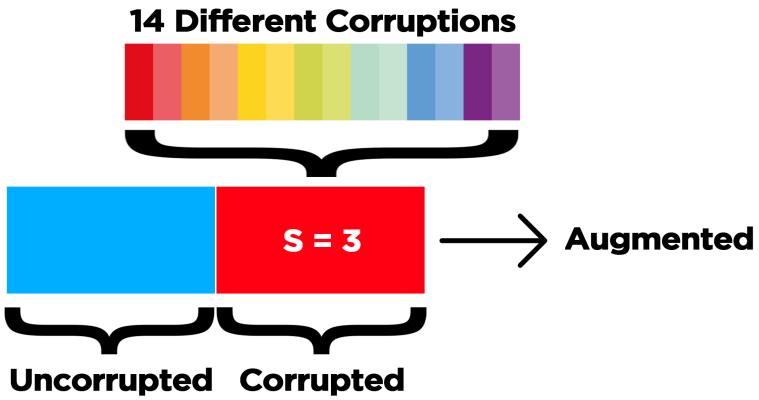


Fig. 9. Representation of a training dataset with severity 3 and corruption ratio 0.5.

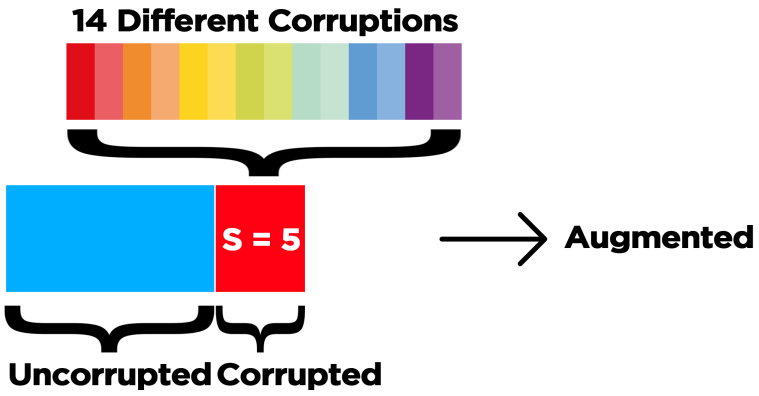


Fig. 10. Representation of a training dataset with severity 5 and corruption ratio 0.3.

Figure 9 and Figure 10 show representations of two datasets with different corruption severities and corruption ratios. All corrupted datasets consist of the full uncorrupted training dataset combined with an equal number of images from each corruption type depending on the corruption ratio. The class distribution is kept the same, so each class's corruption ratio is equal.

D. Confusion matrices clean test set

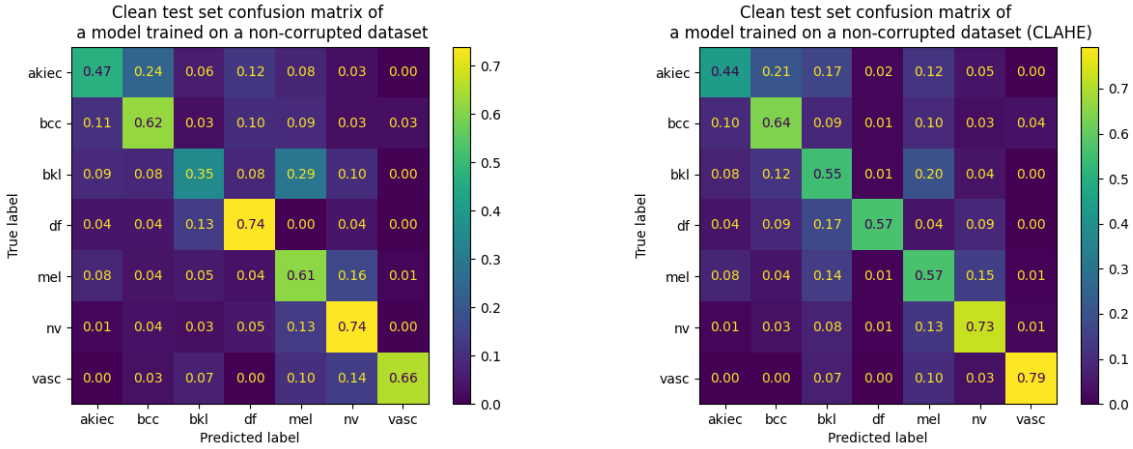


Fig. 11. Confusion matrices of model trained on a non-corrupted dataset with or without CLAHE.

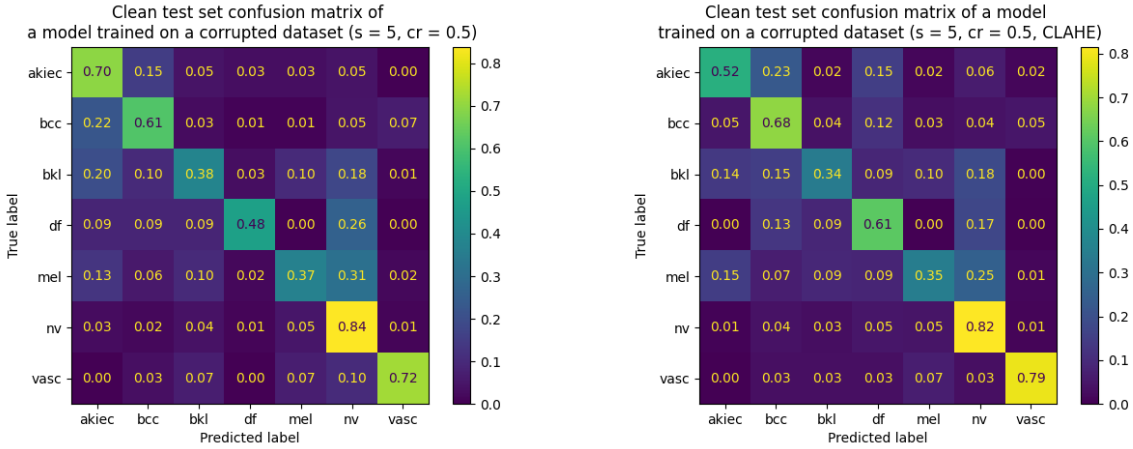


Fig. 12. Confusion matrices of model trained on a corrupted dataset (s = 5, cr = 0.5) with or without CLAHE.

Figure 11 shows the confusion matrices of models trained on non-corrupted data. The left matrix is of a model whose images have been pre-processed with CLAHE and the right one is without. Figure 12 shows the same but for models trained on corrupted data with a corruption severity of 5 and a corruption ratio of 0.5. It can be seen that all models perform very similarly with the non-corrupted CLAHE model having the smallest outliers when it comes to underperforming classes. There are no classes that are recognized much worse than the rest although the models are often underperforming on the class bkl. This once again proves that both CLAHE as well as training with corruptions don't have a big effect on the clean test set accuracy or the recognition of certain classes.

E. Confusion matrices corrupted test set

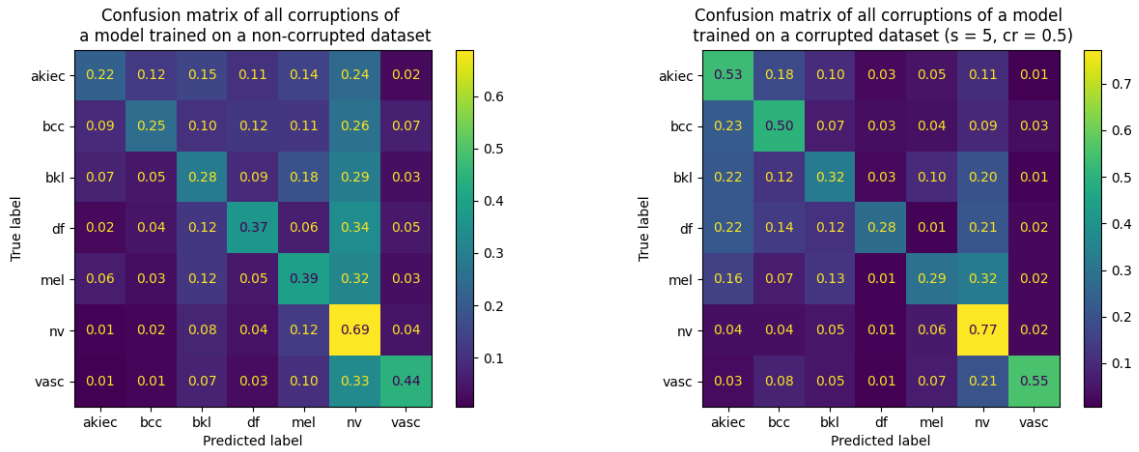


Fig. 13. Normalized (by rows) confusion matrices of a model trained on a non-corrupted dataset and a model trained on a corrupted dataset (s = 5, cr = 0.5) tested on all corruptions (s: 1-5).

To get a better idea on where the difference in corruption robustness is coming from, we tested a non-corrupted and a corrupted model on all corruption types and severities (total of 70) and then averaged out their confusion matrices to get one average confusion matrix per model. These matrices can be found in Figure 13. It can be seen that the model trained on non-corrupted data has a lot of false positives of the nv class and also is performing especially bad on the akiec, bcc and bkl classes. The model trained on corrupted data has an improved performance on almost all the classes and has a fewer false positives of the nv class. It is interesting to see that there are so many false positives of the nv class in the first place because this seems to suggest that the model has learned to guess the class with the most images if it isn't sure. This should have been prevented by implementing class weights in the loss function and augmenting all classes except the nv class. We also don't see the same amount of false positives in the confusion matrix of the clean test set (see Section VII-D), proving that these measures did have an effect.

F. Corruption error rates differences

TABLE V
PERCENTAGE DIFFERENCE BETWEEN NON-CORRUPTED AND OTHER ERROR RATES FOR DIFFERENT CORRUPTIONS
(MODEL: NUMBER = SEVERITY, C = CLAHE, RATIO = 0.5).

Model	Clean	Bl.Cor.	Impul.	Gauss.	Shot.	Bri.Do.	Contr.	Motion.	JPEG.	Defoc.	Bri.Up	Zoom.	Pixel.	Elast.	Char.
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0C	-2.44	-15.07	0.0	3.7	1.2	-2.94	1.37	3.45	0.0	1.67	0.0	5.45	6.25	4.35	0.0
5	-2.44	-30.14	-28.92	-27.16	-25.3	-22.06	-17.81	-17.24	-16.92	-15.0	-9.09	-9.09	-8.33	-6.52	-2.27
5C	0.0	-31.51	-26.51	-20.99	-21.69	-23.53	-12.33	-15.52	-15.38	-15.0	-12.73	-5.45	-4.17	-4.35	-2.27

Table V highlights the impact of training with corrupted data by showing the difference in error rates per corruption between the non-corrupted, non-CLAHE models and other models. You can immediately notice that the biggest error rate improvement is found in the black corner corruption. This is an outlier because the biggest improvements are generally found in the corruptions which started with the highest error rates: the noise corruptions. As discussed in the paper, this indicates that, by training on corrupted data, the model has learned to focus on the regions of the image that are not blacked out.

G. Accuracies of different models

Table VI shows a complete overview of top-1 accuracy percentages for all corruption types and severities for models trained on different severities with a corruption ratio of 0.5. It can be seen that, as expected, the accuracies get worse for increasing severities. It is also interesting to note that there is only a 4% difference between the average accuracy of the non-corrupted model on the severity 1 corruptions and the average accuracy of the severity 5 corrupted model on the severity 5 corruptions. This proves how much training with corruptions can make up for the difficulty of recognizing more corrupted images.

If we take a look at the best-performing models per severity (shaded in gray) we find that the models trained on severity 5 corruptions are not always the ones with the best performance. For severity 1, the model trained on severity 2 corruptions is performing best. For severity 2, it is a tie for the models trained on severities 2 and 3. For severity 3, it is the model trained on severity 3 corruptions and for both severities 4 and 5, it is the model trained on severity 5 corruptions that is beating the other models in accuracy. This indicates that the models perform best on corruptions that resemble the ones they've been trained on. However, we do see that the models trained on corruptions less severe than the ones they're tested on degrade quickly in performance, while the models trained on more severe corruptions are still quite able to work with corruptions less severe than the ones they were trained on. For example, the severity 1 model is getting an average accuracy of 32.6% on severity 5 corruptions compared to the 42.8% accuracy of the severity 5 model. This is a big difference that indicates that a model trained on severity 1 corruptions cannot be trusted when confronted with severity 5 corruptions. On the other side, we see the severity 5 model getting an average accuracy of 52.7% on the severity 1 corruptions while the severity 1 model is getting an accuracy of 56.2%. This is still a performance gap, but less significant than the difference we found for severity 5 corruptions. This means that when it's uncertain what severity corruptions a model will encounter it's best to train the model with images of the highest level of corruption. However, when the model should be robust against only a certain range of corruption severities, it's best to train the model with severities in that range.

What is also interesting to see is that, if we look at the best-performing models per severity, the average accuracy of the CLAHE models is always worse than that of the non-CLAHE models. This shows that, according to our research, CLAHE doesn't improve the corruption robustness of models classifying skin diseases.

TABLE VI
TOP-1 ACCURACY OF A VARIETY OF MODELS FOR DIFFERENT CORRUPTIONS AND SEVERITIES
(MODEL: NUMBER = SEVERITY, C = CLAHE, RATIO = 0.5)

Severity	Model	Corruption (top-1 acc. in %)														Avg.
		Gauss.	Shot.	Impul.	Defoc.	Motion.	Zoom.	Bl.Cor.	Char.	Bri.Up	Bri.Do.	Contr.	Elast.	Pixel.	JPEG.	
	0	33.8	26.1	24.7	50.3	55.0	49.9	32.0	59.3	57.7	54.1	43.9	56.0	61.0	49.3	46.6
	0C	24.2	22.6	26.9	50.4	50.7	49.8	39.1	58.4	57.7	55.4	41.7	55.7	56.8	48.2	45.5
	1	52.1	47.9	47.7	58.7	60.0	55.9	52.1	60.2	59.7	56.2	52.4	59.7	62.1	54.8	55.7
	1C	52.6	47.9	51.4	55.7	57.2	54.5	51.8	57.8	56.5	56.1	48.5	59.1	60.4	53.1	54.5
	2	50.0	51.5	50.5	59.9	60.7	55.5	50.7	60.8	58.7	57.4	52.5	62.9	60.0	56.2	56.2
1	2C	48.1	47.5	52.8	57.7	55.9	54.2	50.1	60.4	58.3	59.0	50.1	60.7	60.5	52.9	54.9
	3	49.2	49.5	49.2	59.1	58.6	54.7	50.1	58.9	58.5	58.1	52.6	61.0	61.9	55.4	55.5
	3C	42.9	42.5	51.0	56.9	54.2	51.7	51.6	58.4	57.0	57.9	47.2	58.5	59.6	51.3	52.9
	4	44.1	43.9	45.5	56.6	57.6	53.3	49.2	55.9	54.8	55.6	49.0	59.2	59.0	53.8	52.7
	4C	41.7	40.7	47.9	55.9	52.3	51.5	46.6	57.6	55.8	58.2	45.3	58.7	59.8	51.4	51.7
	5	43.4	42.6	45.6	56.5	56.5	51.0	48.5	57.6	56.3	59.1	47.9	57.7	59.9	55.5	52.7
	5C	37.3	36.0	46.7	56.0	54.9	50.0	50.4	58.2	58.0	58.7	43.8	57.5	60.3	50.7	51.3
	0	19.5	15.3	17.5	45.0	48.0	48.4	28.5	56.9	48.0	40.1	36.4	47.7	61.8	43.4	39.8
	0C	15.6	14.0	17.1	41.8	45.8	44.8	40.3	56.0	48.5	43.3	33.8	44.7	60.3	41.6	39.1
	1	45.7	37.9	34.1	54.4	57.7	51.0	48.3	58.9	52.1	44.2	43.9	50.4	63.0	49.9	49.4
	1C	44.4	38.3	43.0	48.0	55.6	48.9	52.7	58.4	52.7	47.8	42.3	50.5	60.6	51.5	49.6
	2	52.7	46.5	46.8	55.7	55.8	52.4	50.4	59.4	53.0	48.5	48.0	52.6	62.7	51.1	52.5
2	2C	49.8	44.6	48.2	52.5	56.9	49.1	52.6	58.2	55.6	51.5	43.4	53.4	61.1	49.0	51.8
	3	49.5	47.8	45.5	56.4	54.8	52.5	49.7	58.9	54.0	52.2	48.2	50.8	62.4	51.9	52.5
	3C	43.9	42.7	45.8	53.9	53.1	49.8	52.1	57.0	52.5	52.8	42.4	49.5	58.5	50.0	50.3
	4	45.0	42.3	43.3	54.1	55.5	52.0	48.5	58.1	48.0	50.1	46.1	49.9	60.4	50.9	50.3
	4C	40.9	43.6	42.3	51.1	51.0	48.0	47.1	57.0	50.7	53.7	40.6	47.5	57.0	47.5	48.4
	5	43.3	40.0	44.2	52.5	54.7	48.8	49.7	57.4	48.3	54.5	44.5	47.0	62.1	50.6	49.8
	5C	37.0	37.5	40.0	51.2	53.0	47.6	51.0	56.5	52.5	53.4	39.2	47.2	60.2	49.0	48.2
	0	14.9	14.9	15.0	39.7	41.5	43.6	26.7	56.8	41.2	29.0	23.1	59.8	51.9	37.7	35.4
	0C	14.1	13.6	14.1	37.9	38.0	40.4	40.2	54.8	42.7	31.2	24.3	57.3	51.8	36.9	35.5
	1	32.3	25.0	31.0	47.5	49.4	46.1	47.5	58.4	49.3	31.9	35.5	63.2	56.5	45.1	44.2
	1C	33.5	26.2	37.8	40.1	48.5	45.6	51.2	56.2	51.6	35.1	35.2	62.5	50.5	46.4	44.3
	2	41.9	38.1	41.5	48.6	49.8	47.2	50.1	57.8	49.6	38.2	39.4	63.8	53.8	46.3	47.6
3	2C	40.3	37.6	46.6	45.7	49.8	47.4	51.8	58.5	51.4	41.9	37.0	60.9	53.4	45.8	47.7
	3	44.7	42.4	44.3	51.2	53.1	50.6	49.4	57.1	52.6	41.9	41.0	63.3	58.2	50.7	50.0
	3C	41.7	39.1	42.5	48.5	49.7	48.5	51.3	56.1	52.1	43.4	38.0	59.8	54.6	48.1	48.1
	4	42.4	40.5	42.1	49.1	51.1	51.0	48.0	55.8	49.1	42.0	40.6	61.9	55.6	46.6	48.3
	4C	42.4	41.8	41.6	45.7	49.1	48.6	48.5	55.5	49.7	47.3	38.0	59.0	51.6	45.3	47.4
	5	42.2	38.3	40.9	45.9	52.3	51.1	50.1	58.8	50.7	45.1	37.7	63.0	55.1	48.0	48.5
	5C	38.1	37.7	37.7	46.1	51.4	47.9	51.3	58.0	51.9	49.3	33.8	62.5	52.3	46.3	47.4
	0	14.7	14.3	14.7	33.7	34.1	42.7	25.3	55.3	40.4	19.9	14.9	57.4	43.2	28.6	31.4
	0C	14.3	14.1	14.2	32.9	34.1	39.8	37.2	55.0	39.9	23.1	14.3	55.1	38.4	28.0	31.5
	1	19.6	17.7	21.3	42.3	41.1	45.3	44.6	56.6	45.4	22.8	19.0	61.5	47.3	32.6	36.9
	1C	20.8	18.2	25.1	42.2	40.6	44.9	48.3	56.1	47.2	25.2	18.0	57.7	40.0	36.8	37.2
	2	27.7	22.1	27.7	42.6	40.8	47.0	47.9	57.7	45.1	26.9	27.5	59.5	47.6	35.3	39.7
4	2C	28.1	21.7	32.7	43.3	43.7	45.3	51.7	57.7	46.1	30.9	20.3	57.4	43.1	36.6	39.9
	3	38.2	24.0	35.6	48.0	45.3	51.0	47.2	54.0	48.5	32.1	30.8	61.7	49.8	39.2	43.2
	3C	33.0	25.9	37.8	46.4	46.1	49.5	49.7	54.7	47.3	35.1	24.5	57.2	46.0	39.4	42.3
	4	40.5	34.5	38.6	49.2	47.2	50.1	48.3	53.5	48.2	38.4	36.8	57.0	49.2	39.7	45.1
	4C	39.2	34.6	39.4	47.0	46.4	49.3	48.0	53.9	48.3	39.8	29.9	56.7	49.4	39.2	44.4
	5	41.2	38.9	38.6	44.9	47.3	49.4	48.4	56.2	48.8	40.4	35.0	59.2	49.9	40.0	45.6
	5C	37.8	36.4	37.1	46.5	47.9	48.5	49.1	56.1	50.1	40.7	32.7	58.3	49.6	40.7	45.1
	0	13.8	13.4	13.0	29.6	30.8	39.1	24.8	53.2	37.1	17.3	14.6	50.6	40.5	18.1	28.3
	0C	14.3	14.3	14.1	31.5	30.6	36.6	34.2	53.7	38.2	16.9	14.0	48.3	36.6	19.1	28.7
	1	15.7	15.7	17.3	37.5	39.7	40.5	41.1	55.6	43.8	17.2	15.2	51.7	45.4	19.5	32.6
	1C	16.4	17.4	17.9	36.4	39.1	40.1	46.9	53.3	44.5	19.2	14.8	50.0	38.3	26.5	32.9
	2	20.7	17.3	20.7	34.2	40.6	41.8	44.6	53.8	42.2	19.6	19.2	53.9	47.4	23.0	34.2
5	2C	17.2	18.3	20.4	38.4	42.2	40.6	48.7	57.6	44.3	21.1	15.6	51.5	42.3	28.9	34.8
	3	21.0	17.8	23.4	42.3	44.4	47.8	43.3	53.8	47.1	21.5	19.9	54.8	46.6	28.7	36.6
	3C	20.7	20.9	26.4	41.2	42.2	47.0	48.2	55.3	46.0	24.8	15.6	52.1	42.9	32.7	36.9
	4	29.5	28.3	32.0	48.3	47.0	47.9	46.7	53.2	46.4	30.5	30.1	51.1	49.0	32.6	40.9
	4C	26.6	27.8	31.6	46.6	44.8	47.3	48.8	53.7	45.4	33.9	23.0	51.6	46.0	34.0	40.1
	5	33.0	31.5	35.6	45.1	47.1	48.2	47.6	54.5	46.7	34.5	34.1	55.9	50.7	34.8	42.8
	5C	30.2	29.0	33.6	47.1	46.2	48.4	50.6	56.9	48.3	36.8	30.0	56.4	46.7	36.3	42.6

H. Influence of Contrast Limited Adaptive Histogram Equalization

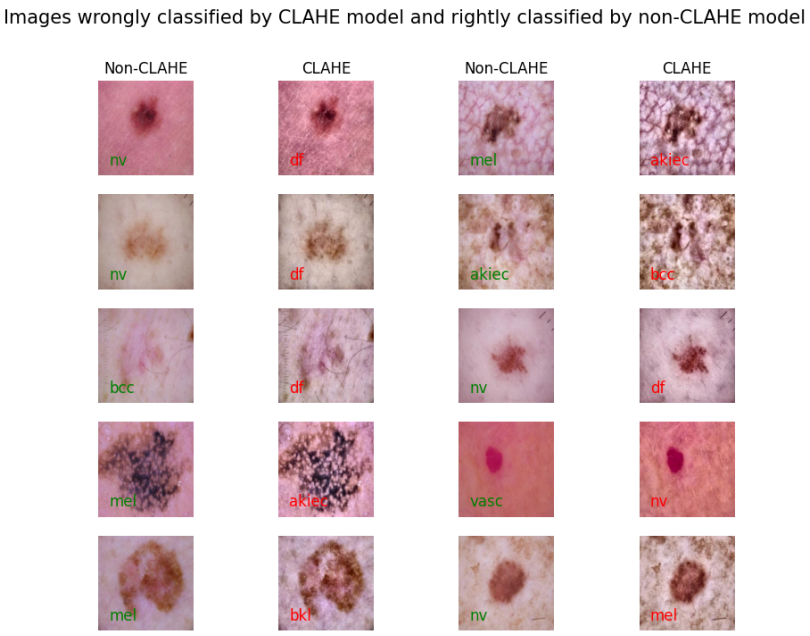


Fig. 14. 10 examples of images wrongly classified by CLAHE models and rightly classified by non-CLAHE models.

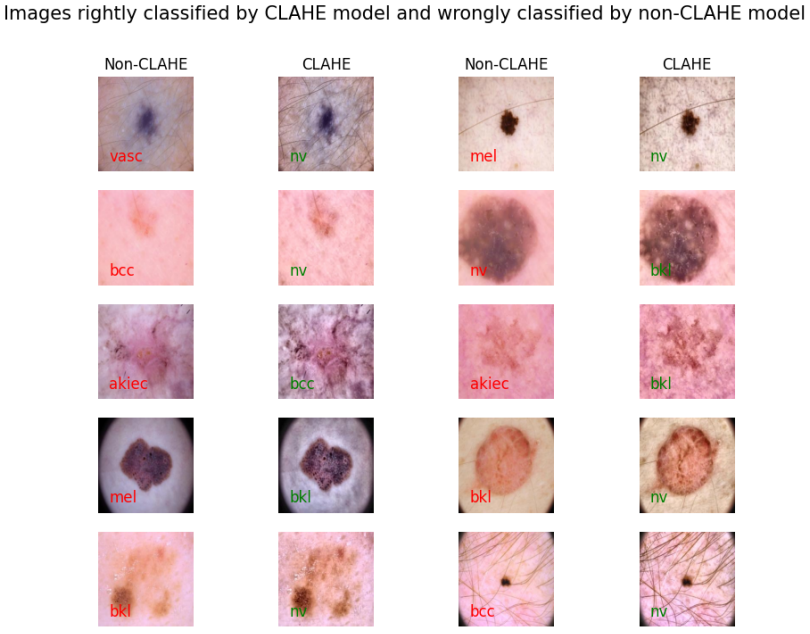


Fig. 15. 10 examples of images rightly classified by CLAHE models and wrongly classified by non-CLAHE models.

To better understand why Contrast Limited Adaptive Histogram Equalization is sometimes decreasing the performance of our models, we created an overview of images that were misclassified by multiple models trained on CLAHE pre-processed images but rightly classified by models that were not. Figure 14 shows an overview of

10 of these images where the right class is in green and the class predicted by the CLAHE models, the wrong class, is in red. Figure 15 shows 10 images where the exact opposite happened. They were misclassified by non-CLAHE models and rightly classified by the CLAHE ones.
Although the differences between these figures are not directly apparent, it can be seen that in multiple cases the wrongly classified images by CLAHE have higher contrast backgrounds after being processed with CLAHE. The background is often the normal skin and is therefore less important than the disease. When CLAHE amplifies especially these background regions it could achieve the exact opposite of what it's designed to do, which is amplifying the disease regions. This could worsen the performance of the CLAHE models on these images.

I. Examples of classification improvement through training with corruptions

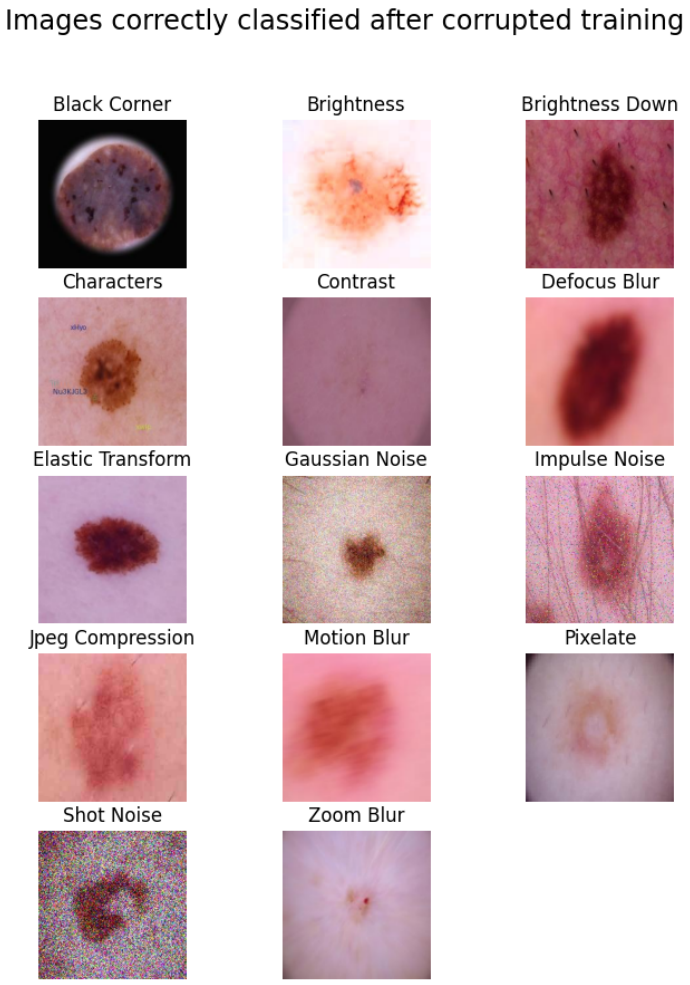


Fig. 16. Overview of images that have been rightly classified by all "s = 5, cr = 0.5"-models and wrongly classified by all models trained without corrupted data.

To better visualize the improved classifier robustness through training with corrupted data we provide the images in Figure 16. These images were correctly classified by all 5 models trained on corrupted data (severity = 5 and ratio = 0.5), but incorrectly classified by all 5 models trained on only non-corrupted data. This indicates that the models have learned to see through the corruptions in these images by training on different images with the same corruptions (the corrupted training dataset). Even some of the more corrupted images are still recognized correctly and by multiple models, which lowers the chances of them being recognized correctly by coincidence.