Dark Corner on Skin Lesion Image Dataset: Does it matter?

Samuel William Pewton and Moi Hoon Yap
Department of Computing and Mathematics, Manchester Metropolitan University
John Dalton Building, Chester Street, M1 5GD Manchester
M.Yap@mmu.ac.uk

Abstract

Skin lesion image datasets gained popularity in recent years with the successes of ISIC datasets and challenges. While the users of these datasets are growing, the Dark Corner Artifact (DCA) phenomenon is under explored. This paper provides a better understanding of how and why DCA occurs, the types of DCAs and investigates the DCA within a curated ISIC image dataset. We introduce new labels of image artifacts on a curated balanced dataset of 9,810 images and identified 2,631 images with different intensities of DCA. Then, we improve the quality of this dataset by introducing automated DCA detection and removal methods. We evaluate the performance of our methods with image quality metrics on an unseen dataset (Dermofit), and achieved better SSIM score in every DCA intensity level. Further, we study the effects of DCA removal on a binary classification task (melanoma vs non-melanoma). Although deep learning performances in this task show marginal differences, we demonstrate that with DCA removal, it can help to shift the network activations to the skin lesions. All the artifact labels and codes are available at: https://github.com/mmu-dermatology-research/dark_corner_artifact_removal.

1. Introduction

Deep learning has had a lot of movement in medicine in recent years. Binary classification on dermatoscopic images of skin lesions validate the effectiveness of deep learning algorithms in this sector [8]. Dermatoscopy generates highly detailed images of malignant and non-malignant skin lesions [4]. Dermatologists then use these images to diagnose a skin lesion as being cancerous or not. The accuracy of clinical diagnosis with the unaided eye is only about 60%, whilst with the use of a dermatoscope an expert dermatologist can correctly diagnose approximately 80% of the time [10]. Artificial intelligence and machine learning has been of high interest in efforts to improve this diagnosis accuracy to aid dermatologists in the diagnosis of skin cancer [1, 7].

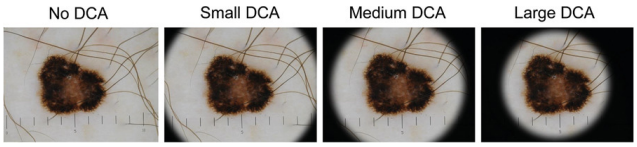

Figure 1. Illustration of different DCA Intensities. From left to right: No DCA, Small DCA (covers 7% of the image area), Medium DCA (covers 25% of the image area) and Large DCA (covers 50% of the image area) according to [17].

Due to the nature of dermatoscopic images, a variety of different artifacts can occur. For example, hair strands, air bubbles, ruler measurements and dark tubular periphery (more commonly known as dark corner artifact (DCA) or black frame). Figure 1 illustrates DCA of different intensities. These artifacts make segmentation and classification more complex [18]. Although DCA are commonly acknowledged in other research, they are less-recognised in efforts to understand its effect on skin lesions classification. DCA have a large potential to skew a CNN classifications, ultimately effecting the accuracy of the proposed deep learning models [17]. Since the role of DCA is less recognised in classification, there are limited efforts to remove this artifact from dermatoscopic images.

The most related study was conducted by [17], where the authors investigated the effect of DCA on the performance of skin lesion classification. The idea behind the study was to superimpose the DCA onto all images in the dataset with varying intensities (small, medium and large). [17] highlighted a handful of limitations encountered in this study. These limitations include the fact that the dataset was collected from patients at increased melanoma risk and that the majority of the images were acquired from fair-skinned patients living in Germany. It is not detailed in this study that the dataset is very small (233 images), and that having a much larger dataset would produce a much more accurate model. A large quantity of data would allow the CNN to more accurately define the features and relationships between the images. Another issue with this study is that only

one example of a CNN has been used to determine results. Using a cohort of different CNNs would give the opportunity to cross examine the results and ensure that the result was not anomalous. Before artificial intelligence can be incorporated into the dermatological process, more research is required to prove that any predictive model generated for this task is correctly identifying features of a skin lesion and not any artifact observed.

This paper aims to investigate the DCA phenomenon in skin lesion image dataset, gain a better understanding of how and why they occur and develop efficient DCA removal processes. With this understanding, the efficiency of modern deep learning methods applied to the binary classification of melanoma vs non-melanoma can be evaluated and the idea of a class bias correlating to images containing DCA can be determined. The main contributions of this paper are dataset-centric by: 1) Providing new insights and understanding of dark corner in skin lesions; 2) Introducing new labels and attributes (manually inspected under 600x magnification) of skin lesion image dataset that will be made available publicly; and 3) Improving the quality of skin lesion image dataset by proposing DCA removal method.

2. Related works

There are limited attempts in previous research in removing DCA. In a study conducted by [12], unsegmented images from PH2 dataset [13] were manually cropped so that any DCA was removed from the image. It was believed that these artifacts were causing a drop in model performance as a large proportion of the PH2 dataset [13] consisted of images with this artifact. There were multiple datasets used throughout this study, including ISIC 2018 [6], HAM [21] and PH2 [13], however the size of the datasets used were not specified. The manual cropping of DCA was only applied to the PH2 dataset [13]. Manual cropping of images is a task that is very time consuming and results in loss of potentially important data within images. Cutting a square out of a circle will lose the edges of the circle. These edges of the circle could contain the nevi and this would be lost in the cropping process.

Tajeddin and Asl [19] made efforts to remove DCA alongside a variety of other image artifacts for the ISBI melanoma segmentation challenge. This method was tested on a dataset of 900 images. In order to remove the DCA, they used a simple method of masking the images containing the dark corners. Pre-defined masks were used and selected by evaluating 9 different pixel regions of the image and determining the pixel intensity of the region. This mask was then applied in a further pre-processing step prior to using Otsu's thresholding method [14]. The method does not remove the artifact, it only works around it by disregarding the area from the thresholding method. Moreover, the dataset this literature was conducted on only contains 900 images. The post-processing steps used to alter the mask may have been fine for this dataset, but it might not be so efficient for a larger dataset. A similar study conducted by [9] used Otsu's thresholding [14] to generate a mask of the dark corners in order to ignore them from segmentation. This paper did not attempt to classify the lesions, only to suggest an appropriate method to handle unwanted artifacts. This study also used a limited dataset of 200 images.

Another approach that has been used to remove DCA is cellular automata. [18] used this method in efforts to enhance the quality of the images ready for processing through a diagnostics tool. It was suggested that using the images without pre-processing steps might interfere with subsequent border detection steps. The approach was to inscribe both a circle and an ellipse centrally on the image with a radius of half the image width. The dataset used by [18] contained 45 dermatoscopic images. One thing that was noted during this study is that the darkness level of the DCA is not constant, so it is not as simple as applying a mask to identify the DCA region. When [18] evaluated the resulting images, it was found that this method was ineffective on images with a large nevi (the circle did not cover the entire area of interest). The radius was then changed to half the length of the image and yielded better results.

Although research into removal of this artifact is limited, there are multiple key points that are apparent from reviewing other literature surrounding this issue. The first key issue with other studies is that the dataset sizes used to remove the artifact in many of the methods are very small. Incorporating the methods used by other researchers onto a large dataset could result is a substantial difference in reliability and accuracy of the removal method. This is a common issue as access to a large dataset of dermatoscope images is very limited.

Another key issue raised from literature surrounds the occurence of the DCA itself. As the artifact is directly related to the device settings, the intensity of the artifact varies greatly. There is no standard size of artifact, it is directly influenced by the magnification of the device. This means that pre-defined masks would be inaccurate and inefficient due to the variety of artifact size. Not only do the artifacts vary greatly in size, but the darkness level of the artifact is not standardised either. As the artifact may not be solid black, it can be difficult to identify using the colour. This is why many of the studies focus on the shape rather than the colour. Finally, many of the studies do not compare the results generated post-removal of the artifact. Although the artifact may have been removed from the image, it would be useful to be able to compare and draw conclusion on the effect the removal has on the accuracy of the CNN.

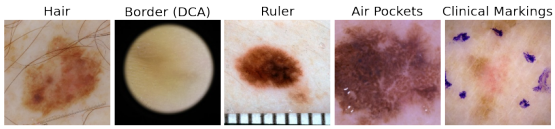

Figure 2. Examples of artifact on skin lesions.

3. Methods

3.1. Dataset

We use a curated balanced dataset proposed by [5], which consists of equal number of melanoma images and other category of skin lesion images. This dataset has a total of 9,810 images, where the authors split it to 7,848 training images and 1,962 validation images. This dataset has been evaluated for skin lesion classification tasks. The nature of our study is dataset-centric, to understand the properties and attributes of the vision dataset, and with no intention to improve the classification algorithms. There are a number of reasons as to why the curated balanced dataset has been used throughout this study. Firstly, this dataset contains a large number of images. The majority of research in this area uses small datasets and this is a key problem. Secondly, the number of melanoma images in this dataset is a significant improvement. Thirdly, this dataset contains a comprehensive amount of examples of varying artifacts, in particular there are a multitude of images containing DCA, which allows in depth testing of methods and results. Finally, this dataset contains balanced classes and has been thoroughly checked for duplicate images to help reduce bias in any deep learning models created.

3.2. New Labels on Image Artifacts

The extent of artifact types present in the curated balanced dataset is broad. In order to gain a more insightful understanding of the artifacts distribution, each image has been manually inspected under 600x magnification and annotated against the artifact categorical conditions. Figure 2 displays examples of all artifact types: Borders, Hair, Measurement Device, Air Pocket(s) and Clinical Markings. It is important to note that the annotations that have been made are subjective to what is believed to be exhibited in the images. New annotations made on the dataset may yield different results to the ones discovered in this process.

Table 1 shows the distribution of artifacts in the curated balanced dataset. It shows that the presence of hair is the most common artifact in dermatological images, followed by the presence of borders of any type. Another observation is the presence of borders is more common in melanoma. In the training set, the presence of borders is 2.6 times more frequent in the melanoma compared to the other. In the validation set, the presence of borders is 2.3 times more frequent in the melanoma compared to the other. More borders in the melanoma class may cause bias toward classification task in any trained predictive models. All other categories of artifact are closely matched across subsets.

Table 1. Distribution of artifacts in the curated balanced dataset.

Artifact Category	**Subset**				**Artifact Totals**
	Train Mel	***Train Oth***	***Val Mel***	***Val Oth***	
Borders	1721	663	417	179	2980
Hair	2224	2595	560	617	5996
Measurement Device	962	749	202	183	2096
Air Pockets	1129	637	442	142	2350
Clinical Markings	124	90	29	20	263
Other	100	55	55	18	228
No Artifacts	377	616	57	172	1222

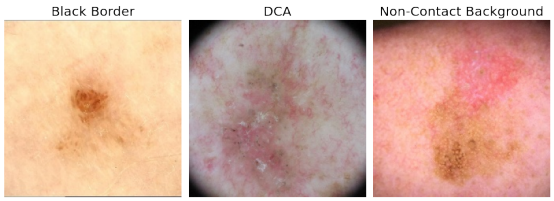

Figure 3. Border Category Examples (Black Border - left, DCA - middle, Non-Contact Background - right)

Table 2. Distribution of border artifact types across the curated balanced dataset.

Border Type	**Subset**				**Border Type Totals**
	Train Mel	***Train Oth***	***Val Mel***	***Val Oth***	
Black Bar(s)	56	212	10	58	336
DCA	1657	451	405	118	2631
Non-Contact BG	8	0	2	3	13
Dataset Totals	1721	663	417	179	2980

3.3. DCA Labels

The borders artifact category is the most important as this category encapsulates DCA. It is necessary to understand how many images across the entire dataset are affected by DCA of any size. Figure 3 shows examples of each border artifact category and Table 2 shows a breakdown of the borders artifact category.

When comparing the different border sub-categories, it is clear that DCA are by far the most common border type found in the dataset. Of this category, DCA occupy 88.3%, with the black bars and non-contact backgrounds occupying the further 11.7%. As the dataset size is 9,810 images, 26.8% of the entire dataset exhibits DCA.

Following the guidelines set out in the DCA Diagnostic Performance research conducted by [17], the images categorised as having a DCA have been further categorised into small, medium and large subsets. In order to get a true measure of the DCA size for each image, the masks generated from the masking process detailed in Section 3.4 will be used. Figure 4 shows examples of the DCA types and Table 3 shows a breakdown of the DCA size categories. The

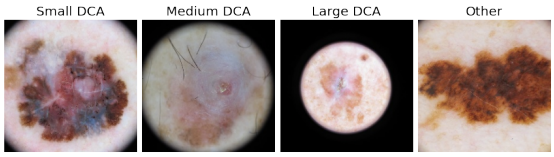

Figure 4. DCA Category Examples (Small DCA - left, Medium DCA - middle-left, Large DCA - middle-right, Other - right)

Table 3. Distribution of DCA sizes across the curated balanced dataset.

DCA Type	**Subset**				**DCA Type Totals**
	Train Mel	***Train Oth***	***Val Mel***	***Val Oth***	
Small DCA	742	237	167	58	1204
Medium DCA	393	79	95	16	583
Large DCA	343	78	80	25	526
Other	179	57	63	19	318
Dataset Totals	1657	451	405	118	2631

categories are determined by the following criteria:

- Small DCA: Any image with a DCA covering between 1% and 24% of the image (inclusive)
- Medium DCA: Any image with a DCA covering between 25% and 49% of the image (inclusive)
- Large DCA: Any image with a DCA covering over 50% of the image (inclusive)
- Other: This extra category has been added for any image containing a DCA covering less than 1% of the image.

From the results in Table 3, it can be seen that of all DCA artifacts found in the curated balanced dataset, there are more small DCA than any other type with a total coverage of 45.8% of the dataset. The second most common DCA artifact type is the medium DCA with a total coverage of 22.2% of the dataset, followed by the large DCA which cover 19% of the dataset. 12% of the DCA identified in the curated balanced dataset have less than 1% total image pixels.

3.4. DCA Detection and Dynamic Masking

We proposed an efficient process to masking images with DCA more effectively, ensuring a minimum loss of data. This section details a new, dynamic process to masking images with DCA and ensuring as much data is retained as possible.

Contour Extraction. When inspecting an image with a DCA, the most prominent shape on the image is the outer ring of the DCA. Finding all of the contours in the image displays an almost perfect outline of the DCA region, along with all other contours found within the image with the given threshold value. These contours can be found by retrieving a binary threshold image from the greyscale transform of the original image and passing the image through the computer vision inbuilt function to find all contours in the image. Figure 5 shows the binary threshold result from the original image, followed by the contours that are extracted from the threshold. The left-most example in Figure 5 contains a DCA with a small area. The central and right-most examples in Figure 5 exhibit lesions that protrude the artifact boundary and display results that do not resemble full circles. The images in this figure have been selected to display the effectiveness of the masking process.

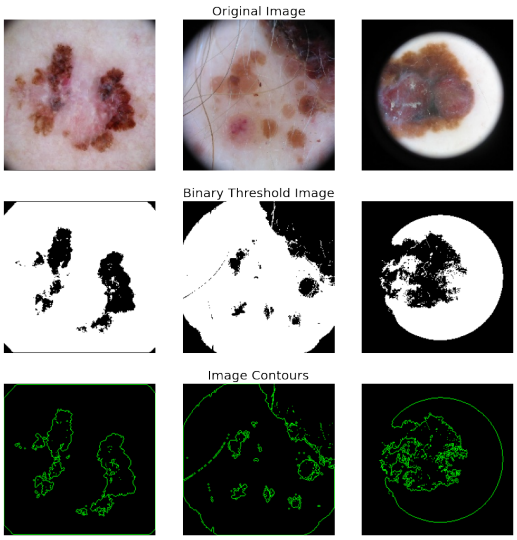

Figure 5. Contour extraction from images with DCA.

Largest Contour Identification. Once all of the contours have been identified within the image, it is necessary to recover the contour with the largest area. The contour with the largest area in the image is the contour which most closely matches the edge of the DCA. This contour is retrieved by calculating the area of all existing contours. Once the area of each contour is calculated, the contour with the maximum area is selected. The top row of Figure 6 shows the largest contour that is extracted from all contours located in the image and the result of using this contour as a mask directly. As can be seen in the central row, the masks generated at this stage appear to work well for DCA with a small surface area, however the lesions in the images with protrusions suffer from large amounts of data loss around the nevi.

Cellular Automata. In order to only capture the DCA present in the image and retain as much data in the image as possible, the idea of cellular automata used in the research conducted by [18] has been incorporated into the process. Instead of iteratively increasing the circle from the centre of the image, a minimum enclosing circle has been used to contain the largest contour within. This uses a bounding

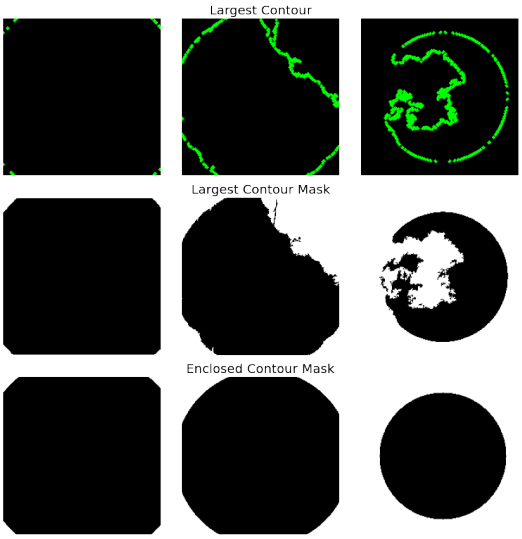

Figure 6. Generating a mask from the largest contour

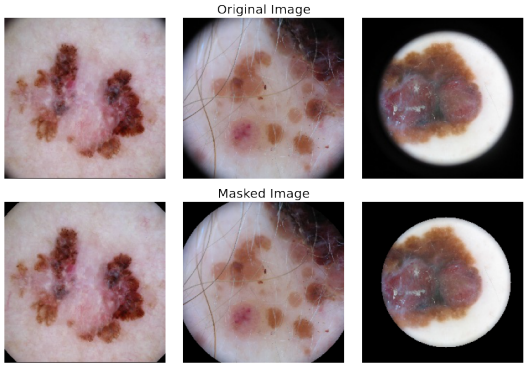

Figure 7. Comparison of original images against masked images

circle which iteratively reduces in size until it can no longer reduce. The centre point and radius of the minimum enclosing circle is calculated and a new circle is drawn. The new circle has a reduced radius of 2 pixels to account for the darkened area of the DCA. The bottom row in Figure 6 displays the new mask that is created from the minimum enclosing circle surrounding the largest contour.

Figure 7 shows the original image against a copy of the image with the DCA mask superimposed on top. Once the masks are superimposed on top of the original images, it can be seen that the DCA present in each of the images have been correctly identified, whilst retaining as much data within the image as possible. The results produced are accurate regardless of the size of the DCA and the positioning of the lesion. A small amount of data has been sacrificed in order to eliminate the shaded edges that are present in the transition into the DCA.

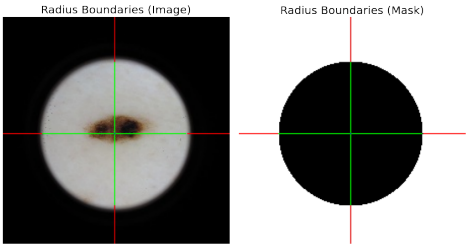

Figure 8. Radius boundaries surrounding the DCA outline (red = area to remove, green = area to keep)

3.5. Proposed DCA Removal Method

DCA intensity reduction. When dealing with medium and large DCA, the area of the DCA can be extreme. To reduce a lot of the processing required for predicting the DCA region, the edges of these images need to be removed as much as possible. As the sizes of the DCA are not static, this process also needs to be dynamic. From the masks generated of the DCA, the centre point and radius are retrieved. This information is crucial to accurately determine the DCA boundary on both the x and y axis. Figure 8 displays the area which can be removed outside of the DCA boundary using the radius of the circle.

In order to remove the edges from the DCA, the thickness of the horizontal and vertical borders need calculating. The thickness of border to remove from an image (r) is calculated with equation 1:

$$r = \min\left((L_{dist} + R_{dist}), (T_{dist} + B_{dist})\right) \qquad (1)$$

where: L_{dist} is equal to the spacing of the left-most edge of the image and the DCA; R_{dist} is equal to the spacing of the right-most edge of the image and the DCA; T_{dist} is equal to the spacing of the upper-most edge of the image and the DCA; and B_{dist} is equal to the spacing of the lower-most edge of the image and the DCA.

With the total removal amount calculated from the horizontal and vertical border thickness, the border can be removed from the image whilst retaining a square image. Figure 9 shows the result after intensity reduction processing has been completed against the original image. When this process encounters an image that does not have a medium or large DCA or the DCA protrudes more than 2 edges of the image, no intensity reduction steps are applied as no edges can be removed without removing data from the image.

Super Resolution and Rescaling Once the intensity reduction process has been completed, there is a new issue present with the images. As some of the image has been removed, the images are no longer all the same size. Simply resizing the images that have been cropped will cause a large reduction in quality. The cropped images which are enlarged in figure 9 show the level of distortion that

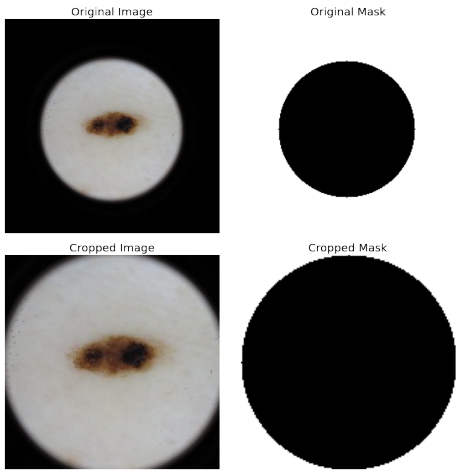

Figure 9. Comparison of before and after Intensity Reduction

results from resizing an image directly. In order to retain image quality, super resolution steps are applied to the images when upscaling the size.

Lim et al. [11] proposed new EDSR network to reconstruct high-resolution images at a higher scale. To train this network, [11] used a baseline model resembling the ResNet architecture is created with ReLU activation layers outside of the residual blocks removed. This baseline model uses 64 feature maps for each convolutional layer, so no residual scaling is used in this step. Once the baseline model is trained, it is expanded to contain 32 feature channels with a scaling factor of 0.1. This model uses a pre-trained baseline model to improve performance and results. [11] won the NTIRE2017 Super-Resolution Challenge with this proposed method. To restore image quality whilst the images are up-scaled, the DCA images which are cropped are passed through a pre-trained EDSR model. As the new images are scaled 4x the original size, the resulting images will be larger than required. The resulting images are then resized back to the original size of 224×224.

Corner Removal Once the DCA intensity has been reduced as much as possible and the image quality is restored, further image processing techniques can be applied in efforts to predict the remaining region within the DCA area. We use two popular image processing inpainting methods. The first method that has been used to remove the DCA region is inpainting with a Navier Stokes [3]. This method was created by [3] where ideas from fluid dynamics equations have been adapted for use in image inpainting. Each of the inputs used in the original equations are matched with a feature of the image. To summarise this method, the image smoothness is calculated using the Laplacian operator and is spread across isophote lines detected within the image. The second method that has been used to remove the DCA region is inpainting with an algorithm proposed by [20]. To approximate the value of a pixel, this method takes the known neighborhood of the pixel in question (the known pixels in the image within a given radius) and sums the weighted, normalised estimates that are calculated from the edges of the radius. Each point within the region to predict is iteratively estimated starting from the edges of the area until the entire region has been predicted. This method applies techniques that are used in manual inpainting processes at a much quicker rate.

3.6. Image Quality Comparison

In order to calculate and evaluate the effectiveness of the DCA removal processes detailed above, a second image dataset is introduced to limit any future bias. The dataset used for this task is the Dermofit Image Library [2]: https://licensing.edinburgh-innovations.ed.ac.uk/product/dermofit-image-library. This dataset contains 1300 high quality images of skin lesions from multiple classes with a minimal amount of artifacts present. Due to the licensing surrounding this dataset, no images are able to be shared in this report.

The images are loaded and shuffled to ensure a good class distribution before splitting into 4 subsets of 318 images. These subsets represent the different DCA intensity containing a balanced amount of images for each DCA size. This is determined by the lowest number of DCA masks generated from the ISIC curated balanced dataset. The masks that are generated from the ISIC curated balanced dataset are superimposed onto the Dermofit images to recreate the DCA effect, giving both a masked image and a ground truth image. To evaluate the performance of our proposed DCA removal method, we compare the inpainted area of the DCA image with the expected area in the ground truth values.

4. Experiments and Results

This section present the results of the proposed DCA removal method and an additional experiment to evaluate its effect on a classification task. Figure 10 visually compare the results of our proposed DCA removal method, on a super resolution image. Both inpainted images show a closely matched skin colour in most areas of the image. The main visual difference between both resulting images is the upper left corner in the Telea method appears to more accurately reflect the skin colour expected - however there is a residual black line surrounding the perimeter of the DCA boundary. The Navier Stokes result also shows a clear variance in colour across the DCA perimeter. Both inpainting methods explained above appear to have made large visual improvements to the quality of images containing DCA and it is clear the colour gradient between the actual image and the DCA area is reduced drastically.

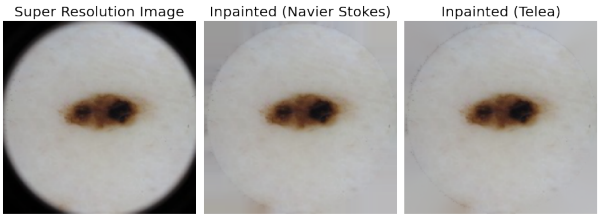

Figure 10. Illustration of inpainted results, from left to right: Super Resolution Image, inpainted Image with Navier Stokes method, and inpainted Image with Telea method.

Table 4. A comparison of image quality of baseline (augmented DCA) and DCA removal methods.

Method	**Set**	**MSE** ⇓	**MAE** ⇓	**SSIM** ⇑	**PSNR** ⇑
Baselines	Small	8.9923	7.6207	0.9133	16.2673
	Medium	40.8948	34.5211	0.6088	8.1716
	Large	67.1845	60.4687	0.3530	6.1891
	Oth	0.3729	0.3119	0.9966	30.4560
Inpainting (NS)	Small	3.0813	9.5022	0.9837	43.9194
	Medium	9.0683	30.0187	0.9553	36.4367
	Large	11.1234	35.6139	0.9584	34.6459
	Oth	0.0933	0.3520	0.9997	61.3411
Inpainting (Telea)	Small	3.2043	8.0621	0.9833	43.7565
	Medium	9.1789	26.8469	0.9544	36.4933
	Large	11.2175	32.7812	0.9573	34.7656
	Oth	0.1050	0.3204	0.9997	60.3124

4.1. Image Quality Assessment

Table 4 shows the vast majority of metrics have increased in performance from the baseline results for both removal processes. Overall both DCA removal methods gained performance on most metrics in comparison to the baseline results, but both methods show a decrease in MAE performance for images containing small DCA. The MAE for each DCA size has greater improvement rates for the Navier Stokes based method, the MAE shows less performance losses and more performance gains in the Telea based method, the SSIM shows more improvements in the Navier Stokes based method and the PSNR shows level increases between the methods. From these metrics, it can be seen that both methods produce comparable results, with the greatest affected category of DCA that both methods improve is the large DCA.

4.2. Evaluation on Skin Lesions Classification

Eighteen of the most widely used deep learning models in image classification tasks from the Keras Applications zoo have been trained using the dataset in its original state to form a series of benchmark results. These results enable comparisons with results generated using the datasets with new DCA labels. The deep learning models used are VGG16, VGG19, Xception, ResNet50, ResNet101, ResNet152, ResNet50V2, ResNet101V2, ResNet152V2, InceptionV3, InceptionResNetV2, DenseNet121, DenseNet169, DenseNet201, EfficientNetB0, EfficientNetB1, EfficientNetB3 and EfficientNetB4. Note that we only present the top-3 best models, the full results are available at: https://github.com/mmu-dermatology-research/dark_corner_artifact_removal.

Table 5. A Comparison of DCA removal methods on binary skin lesions classification.

Model	Method	Acc	TPR	TNR	F1	AUC
InceptionResNetV2	Baseline	**0.82**	0.80	**0.83**	**0.81**	**0.89**
	Telea	0.79	0.69	0.88	0.76	0.88
	NS	0.80	**0.81**	0.79	0.80	0.88
DenseNet121	Baseline	0.76	0.67	0.84	0.73	0.82
	Telea	0.80	0.80	0.80	0.80	0.88
	NS	0.80	0.77	0.83	0.79	0.88
EfficientNetB3	Baseline	0.75	0.63	0.88	0.72	0.82
	Telea	0.79	0.78	0.79	0.78	0.87
	NS	0.77	0.73	0.82	0.76	0.86

Each model is trained using stochastic gradient descent with a batch size of 64, no pre-trained weights and a maximum of 200 epochs with a patience of 10 epochs. It is important to note that no model fine tuning is used in this training process and all models are trained using the same hyperparameters to ensure fairness and validity when comparing model performance. The epoch for each model showing the maximum validation accuracy was saved and recorded. The hardware configuration used to train all of the networks was an AMD Ryzen 7 3700X 8-core 16-thread 4.4GHz CPU with 16GB DDR4 3000MHz Dual-Channel RAM and an NVIDIA Geforce RTX 3090 FE 24GB GDDR6X GPU. The software configuration used was Python 3.9.7, TensorFlow GPU 2.9.0-dev20220203, CUDA 11.2.1 and cuDNN 8.1 running on Windows 10.

When comparing the results as in Table 5, it shows that the model performances for each method have marginal differences. This is not surprise as the original training set with DCA will continue to provide prediction. However, there is a possibility that DCA has been used to predict melanoma (as there are more melanoma cases with DCA in the dataset).

To understand the effect of DCA and our proposed DCA removal methods on the model performance, it is necessary to further examine the network using Grad-CAM [16]. Grad-CAM extracts gradients from the final convolutional layer of a CNN and generates a heatmap. This heatmap shows the areas of the image which are most focused on by the convolutional layer. For this experiment, a class implementation of Grad-CAM created by [15] has been used. The colour scheme used in the following Grad-CAM images show the bright yellow areas as the targeted areas and the purple areas as the areas of least interest by the network.

Figure 11 displays the different Grad-CAM heatmaps generated for the same image exhibiting a DCA across each

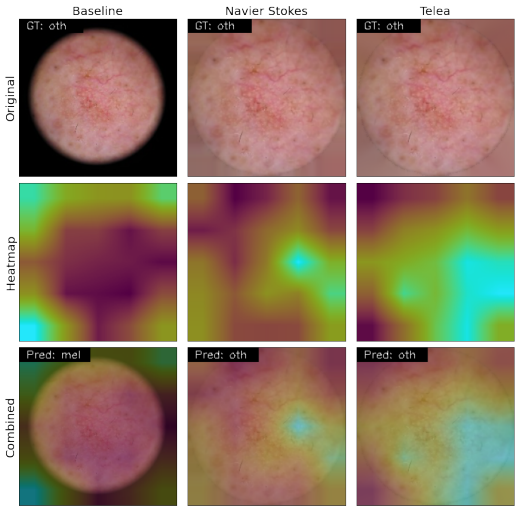

Figure 11. Comparison of Grad-CAM results for each removal-method using InceptionResNetV2.

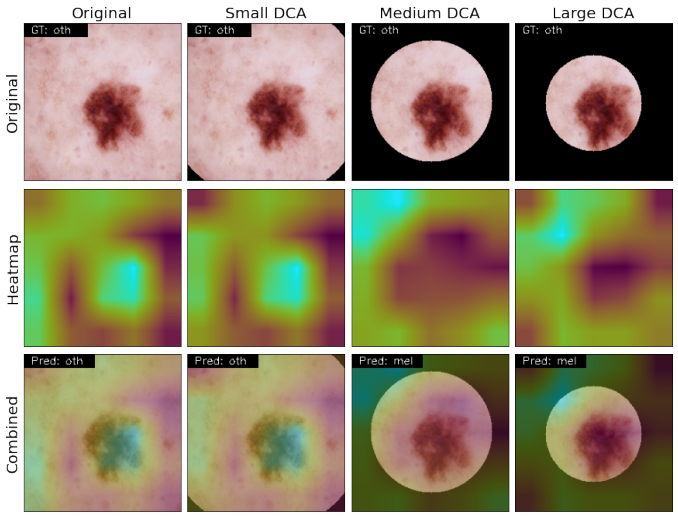

Figure 12. Grad-CAM results for the best performing baseline model for an image with varying DCA sizes superimposed. 'mel' for melanoma and 'oth' for other class.

of the InceptionResNetV2 models for the baseline model and each removal method used. InceptionResNetV2 has been used because it was the most common network with the best performance across each of the result sets.

As illustrated in Figure 11, the network activations in the baseline model are largely focused on the DCA region. In both of the networks using DCA removal methods, the activations shift to focus mostly on the skin lesion indicating that the network is no longer predicting based on the DCA and is focusing more on the features of the lesion itself. In this example, it can be seen that the prediction in the baseline results was incorrect, whereas the predictions made following the removal methods are correct. This example supports the concept that the DCAs give a natural bias toward the melanoma classification, which can be eliminated by using DCA removal techniques.

To further support this, Figure 12 takes an image belonging to the 'oth' classification which is correctly predicted by the baseline model and superimposes each a DCA of various sizes onto it. Grad-CAM images are generated for each image to show the difference in focus by the same baseline network (InceptionResNetV2).

As shown in Figure 12, as the DCA intensity increases - the area of focus in the centre of the image restricts. The network is able to correctly predict the class of the image when there is no DCA or if a small DCA is used. The activations in the original image and small DCA image also focus largely on the skin lesion. Once medium and large DCAs are superimposed onto the image, the activations no longer focus on the lesion are mostly predict based on the edges of the image. The medium and large DCA images are incorrectly predicted as belonging to the 'mel' classification.

5. Conclusion

This research has explored a widely acknowledged image artifact that has a large potential to skew the decision making of CNNs and generate a dynamic process to eliminate the artifact from images. It enables a deep understanding of the skin lesions image dataset, the artifacts and the DCA phenomenon. An effective masking and DCA removal method were created in order to limit the effect these artifacts have on the classification process. The original dataset and both removal methods were evaluated using popular deep learning methods for image classification tasks. Although comparable results have been generated, it is clear from the Grad-CAM heatmaps that the network activations were originally focused on the DCA region and with the removal processes, the activations focus more on the areas intended.

There are many different avenues that could be explored in efforts to enhance this project further. The main avenue is to determine a better approach to removing the remaining DCA from the image once the intensity reduction steps have been applied. The majority of images are inpainted effectively however there are examples where the process has not worked as well. Another popular and actively researched approach in computer vision is outpainting with GANs. [22] proposed two outpainting methods to predict what is beyond the edges of a square image whilst retaining the style of the original image and reducing blur. If this method were to be modified to predict beyond a circular region as opposed to square, the DCA removal process may generate more accurate and less blurry predictions beyond the DCA border. This could possibly further divert the focus of the activations from the DCA region.

References

[1] Redha Ali, Russell C Hardie, Barath Narayanan Narayanan, and Supun De Silva. Deep learning ensemble methods for skin lesion analysis towards melanoma detection. In *2019 IEEE National Aerospace and Electronics Conference (NAECON)*, pages 311–316. IEEE, 2019. 1

[2] Lucia Ballerini, Robert B Fisher, Ben Aldridge, and Jonathan Rees. A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions. In *Color medical image analysis*, pages 63–86. Springer, 2013. 6

[3] Marcelo Bertalmio, Andrea L Bertozzi, and Guillermo Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001. 6

[4] Michael Binder, Margot Schwarz, Alexander Winkler, Andreas Steiner, Alexandra Kaider, Klaus Wolff, and Hubert Pehamberger. Epiluminescence microscopy: a useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists. *Archives of dermatology*, 131(3):286–291, 1995. 1

[5] Bill Cassidy, Connah Kendrick, Andrzej Brodzicki, Joanna Jaworek-Korjakowska, and Moi Hoon Yap. Analysis of the isic image datasets: Usage, benchmarks and recommendations. *Medical Image Analysis*, 75:102305, 2022. 3

[6] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 2

[7] Manu Goyal, Amanda Oakley, Priyanka Bansal, Darren Dancey, and Moi Hoon Yap. Skin lesion segmentation in dermoscopic images with ensemble deep learning methods. *IEEE Access*, 8:4171–4181, 2019. 1

[8] Imran Iqbal, Muhammad Younus, Khuram Walayat, Mohib Ullah Kakar, and Jinwen Ma. Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images. *Computerized Medical Imaging and Graphics*, 88:101843, 2021. 1

[9] Uzma Jamil, Shehzad Khalid, and M Usman Akram. Dermoscopic image enhancement and hair artifact removal using gabor wavelet. In *2016 Sixth International Conference on Innovative Computing Technology (INTECH)*, pages 279–284. IEEE, 2016. 2

[10] Harold Kittler, H Pehamberger, K Wolff, and MJTIO Binder. Diagnostic accuracy of dermoscopy. *The lancet oncology*, 3(3):159–165, 2002. 1

[11] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 6

[12] Roman C Maron, Achim Hekler, Eva Krieghoff-Henning, Max Schmitt, Justin G Schlager, Jochen S Utikal, and Titus J Brinker. Reducing the impact of confounding factors on skin cancer classification via image segmentation: Technical model study. *Journal of Medical Internet Research*, 23(3):e21695, 2021. 2

[13] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. Ph 2-a dermoscopic image database for research and benchmarking. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5437–5440. IEEE, 2013. 2

[14] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979. 2

[15] Adrian Rosebrock. Grad-cam: Visualize class activation maps with keras, tensorflow, and deep learning, 3 2020. [Accessed: 10-03-2022]. 7

[16] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 10 2017. 7

[17] Katharina Sies, Julia K Winkler, Christine Fink, Felicitas Bardehle, Ferdinand Toberer, Felix KF Kommoss, Timo Buhl, Alexander Enk, Albert Rosenberger, and Holger A Haenssle. Dark corner artefact and diagnostic performance of a market-approved neural network for skin cancer classification. *JDDG: Journal der Deutschen Dermatologischen Gesellschaft*, 2021. 1, 3

[18] Alina Sultana, Ioana Dumitrache, Mihai Vocurek, and Mihai Ciuc. Removal of artifacts from dermatoscopic images. In *2014 10th International Conference on Communications (COMM)*, pages 1–4. IEEE, 2014. 1, 2, 4

[19] Neda Zamani Tajeddin and Babak Mohammadzadeh Asl. A general algorithm for automatic lesion segmentation in dermoscopy images. In *2016 23rd Iranian Conference on Biomedical Engineering and 2016 1st International Iranian Conference on Biomedical Engineering (ICBME)*, pages 134–139. IEEE, 2016. 2

[20] Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1):23–34, 2004. 6

[21] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. scientific data. 2018; 5: 180161. *Search in*, 2018. 2

[22] Basile Van Hoorick. Image outpainting and harmonization using generative adversarial networks. *arXiv preprint arXiv:1912.10960*, 2019. 8