




Please cite the Published Version

Pewton, SW, Cassidy, B , Kendrick, C  and Yap, MH  (2024) Dermoscopic dark corner artifacts removal: Friend or foe? *Computer Methods and Programs in Biomedicine*, 244. 107986
ISSN 0169-2607

DOI: <https://doi.org/10.1016/j.cmpb.2023.107986>

Publisher: Elsevier

Version: Published Version

Downloaded from: <https://e-space.mmu.ac.uk/633786/>

Usage rights:  [Creative Commons: Attribution-Noncommercial-No Derivative Works 4.0](https://creativecommons.org/licenses/by-nc-nd/4.0/)

Additional Information: This is an open access article published in *Computer Methods and Programs in Biomedicine*, by Elsevier.

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)



Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Dermoscopic dark corner artifacts removal: Friend or foe?

Samuel William Pewton¹, Bill Cassidy^{*,1}, Connah Kendrick, Moi Hoon Yap

Department of Computing and Mathematics, Faculty of Science and Engineering, Manchester Metropolitan University, Chester Street, Manchester, M1 5GD, UK

ARTICLE INFO

Keywords:

Dark corner artifacts
ISIC
Melanoma
Skin cancer
Vignettes

ABSTRACT

Background and Objectives: One of the more significant obstacles in classification of skin cancer is the presence of artifacts. This paper investigates the effect of dark corner artifacts, which result from the use of dermoscopes, on the performance of a deep learning binary classification task. Previous research attempted to remove and inpaint dark corner artifacts, with the intention of creating an ideal condition for models. However, such research has been shown to be inconclusive due to a lack of available datasets with corresponding labels for dark corner artifact cases.

Methods: To address these issues, we label 10,250 skin lesion images from publicly available datasets and introduce a balanced dataset with an equal number of melanoma and non-melanoma cases. The training set comprises 6126 images without artifacts, and the testing set comprises 4124 images with dark corner artifacts. We conduct three experiments to provide new understanding on the effects of dark corner artifacts, including inpainted and synthetically generated examples, on a deep learning method.

Results: Our results suggest that introducing synthetic dark corner artifacts which have been superimposed onto the training set improved model performance, particularly in terms of the true negative rate. This indicates that deep learning learnt to ignore dark corner artifacts, rather than treating it as melanoma, when dark corner artifacts were introduced into the training set. Further, we propose a new approach to quantifying heatmaps indicating network focus using a root mean square measure of the brightness intensity in the different regions of the heatmaps.

Conclusions: The proposed artifact methods can be used in future experiments to help alleviate possible impacts on model performance. Additionally, the newly proposed heatmap quantification analysis will help to better understand the relationships between heatmap results and other model performance metrics.

1. Introduction

The first recorded example of using microscopy dates back to 1655 where Pierre Borel observed capillaries of the nailbed under a microscope. Ever since this moment there have been many studies resulting in improvements to the process, including the use of different immersion fluids to make the upper layers of the epidermis more translucent to improve examination. Portable devices were not available until 1990 where Kreuzsch and Rassner developed a portable stereomicroscope capable of magnification from 10-40x [9]. The downside to this device is that it was much more expensive than the devices used previously. These early advancements led to the development of the hand-held dermatoscope [9]. The use of a dermatoscope gives the dermatologist the ability to magnify and view features of the lesion that were obscure

or invisible to the naked eye allowing for a more accurate diagnosis [22].

Dark corner artifacts (DCA), also known as vignettes [35], in skin lesion images can be defined as regions around the edges of the image which are dark in appearance which can vary in size and intensity. This phenomenon is a result of the circular shape of the dermoscopic lens when pressed against the skin during a dermatological examination. DCA are not always present in dermatological images. Variations in dermatoscope calibration, camera zoom levels, and post-processing, as manually determined by the examining dermatologist, are the main contributing factors in the presence and degree of DCA [1].

Another type of DCA appears when a non-contact dermatoscope is used to take an image of a lesion located on a non-flat surface of skin,

* Corresponding author.

E-mail addresses: sam.pewton@hotmail.co.uk (S.W. Pewton), b.cassidy@mmu.ac.uk (B. Cassidy), connah.kendrick@mmu.ac.uk (C. Kendrick), m.yap@mmu.ac.uk (M.H. Yap).

¹ These authors contributed equally to this work.

<https://doi.org/10.1016/j.cmpb.2023.107986>

Received 11 July 2023; Received in revised form 9 December 2023; Accepted 16 December 2023

Available online 23 December 2023

0169-2607/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

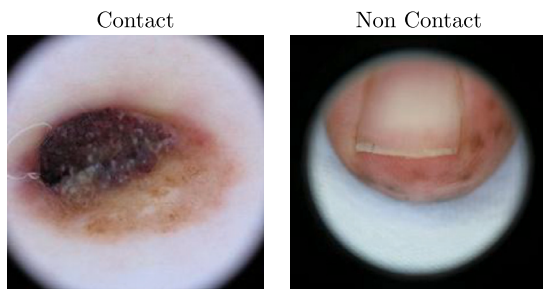


Fig. 1. Illustration of Contact vs Non Contact border artifact.

such as the ear. In this example, the non-ear region is black due to the focus, exposure and white balance settings of the camera. A card is often placed around the ear (and similar areas of the body such as nose and digits) if the exposure or focus is incorrect due to the camera's attempt to balance all parts of the image. Fig. 1 shows examples of both types of artifacts. Additionally, if the nevi of the lesion is large, the dermatologist may not be able to capture the entire lesion when zoomed-in to remove the DCA from the image - this happens regardless of the model of the dermatoscope.

The cost of dermatoscopes may be linked to the number of features available on the device, however, cost is not directly associated with the presence of DCA. The most expensive devices may exhibit DCA, while cheaper ones may not, and vice versa. An example of this is the DermLite HÜD[®], which is one of the cheapest dermatoscopes. This device provides a rectangular view of the lesion, and is therefore easy to remove DCA. Zoom settings used to remove DCA are a feature only of the camera used to acquire the skin lesion photograph, and not the dermatoscope itself. Cropping of DCA is common practice in dermatology, which may include cropping of small regions of the dermoscopic view. Some of the publicly available datasets [52] perform pre-processing to normalise colours and provide zoomed images to reduce the presence of the DCA when possible as a form of natural augmentation.

Dermoscopic images found in the ISIC datasets are sourced from a variety of dermatoscopes and cameras. They may include transparency slide scans, dedicated dermoscopy cameras, video stills, fixed focus devices, lenses of differing diameter (10 mm to 30 mm), shape and quality. Some are plastic while others are glass, may be scratched or dirty, and may have been acquired with or without immersion contact fluid. Most ISIC images were acquired using direct contact with skin and lesion, with rare cases showing non-contact.

The use of dermatoscope devices is known to cause numerous variations in contrast which may result from the use of unpolarised and cross-polarised light. Variations in illumination, and noise [17,3] may also be present, amongst other types of artifacts. Researchers have become increasingly focused on the effect of artifacts in recent years, however, such studies are limited, and tend to focus mainly on the effects of hair [26,5]. Studies which focus on detection of melanoma and other skin lesion pathologies have become a significant field of research, especially following the rise in the use of deep learning [15,8,7,16,39,25,24]. However, despite the acknowledgement of the presence of artifacts in the associated datasets, solutions are either not explored [31,56] or are limited [38].

Previous solutions in preventing overfitting in deep learning focused on the collection of more data, or utilising various data augmentation techniques. These methods include standard transformation of images, removal of artifacts (inpainting methods) and generation of synthetic data. The most popular methods used to overcome the effect of artifacts on skin lesion images on deep learning models has been by the proposition that artifacts should be removed, often with the use of inpainting methods. Although these studies [55,50] achieved some improvement in accuracy, it is unclear if this process is a friend (removed and inpainted artifacts) or a foe (removed and inpainted important features).

Additionally, this process involved localisation and inpainting of artifacts, which is usually computationally expensive and inaccurate as there is no ground truth to evaluate performance. Due to these reasons, focusing on the occurrence of DCA, we propose to superimpose DCA and train a deep learning network to learn these types of artifacts. The aim of this paper is to answer the following questions in binary classification of melanoma and non-melanoma:

1. What is the effect of DCA in skin lesion binary classification?
2. Will inpainting algorithms improve the accuracy of skin lesion binary classification?
3. Which data augmentation method provides the best results: inpainted DCA or superimposed synthetic DCA?
4. Will the deep learning algorithm learn to ignore DCA like the dermatologists?

The main contributions of this paper are as follows:

- Introduction of a new DCA split balanced dataset which we make available to the research community. To assist in answering research questions (1) and (2), we curate a new dataset to allow for fair comparison. To date, there are no publicly available DCA split balanced datasets.
- A proposed realistic DCA data augmentation method and compare its performance with a binary DCA data augmentation method [38]. We investigate different DCA augmentation techniques to study the effect of DCA inpainting versus the effect of superimposing DCA as in research question (3).
- A quantitative measure is proposed to evaluate the visualised activation maps that are commonly used in deep learning research. We measure the differences between deep learning methods when performing inference on images with DCA, and draw a new perspective in handling DCA which can be used in other artifact-related research.

2. Related work

The study of external ocular images or external photographs [2] in deep learning has gained popularity in explainability analysis. These experiments share a common goal to increase the trustworthiness of deep learning algorithms by indicating the area of interest used for prediction, rather than the background. There are several imaging domains with external ocular images, including dermoscopic [9], external eye photography [2], ophthalmic and endoscopy images [48], and colonoscopy images [29]. As each of these domains are device specific and application specific, this paper focuses on dermoscopic images for skin cancer analysis.

Tschandl et al. [53] conducted experiments using the ISIC datasets [19,12,11,52,13,45] to compare the diagnostic accuracy of deep learning algorithms with human readers for all clinically relevant types of benign and malignant pigmented skin lesions. Their findings showed that classifiers often had good performance when tested on data that is similar to the training data but performed worse or failed on out of distribution examples.

Han et al. [21] conducted multi-class classification experiments for 12 classes of skin diseases. They used a fine-tuned ResNet-152 model which was trained on a number of skin lesion datasets. This work highlighted the challenges inherent in the use of different datasets, and showed that deep learning predictions can be at least as accurate as dermatologists. These findings were further supported in subsequent studies [20,23]. However, more recent works, such as those by [33], found that a deep learning model which demonstrated superior performance in experimental studies performed poorly when compared to specialists in real-world settings. Such findings suggest that caution is required when extrapolating results of experimental studies to clinical practice.

Sies et al. [49] investigated the effect of small, medium, and large DCA on the performance of a market-approved CNN (Moleanalyzer-Pro[®], FotoFinder Systems) for skin lesion classification which provided malignancy scores in the range of 0 to 1. They observed that for small and medium DCA cases the system gave comparable diagnostic performance as control cases (those without DCA). However, prediction results for the large DCA cases showed a significant decrease in specificity performance, indicating that the CNN was less robust in its ability to correctly reject when a patient did not present a malignancy.

Nauta et al. [34] observed that artifacts can lead to shortcut learning. Their work focused on detecting and quantifying shortcut learning in trained classifiers for skin cancer diagnosis using the ISIC datasets. They trained a standard VGG16 skin cancer classifier with data split so that elliptical colour patches were present only in the benign images. They inserted colour patches onto images which did not already have them and used inpainting to automatically remove patches from images to assess the effect on predictions. They found that the classifier would partly base its benign predictions on the presence of the coloured patches, and that artificially inserting coloured patches into malignant images resulted in shortcut learning leading to a significant increase in misdiagnosis.

Zand et al. [54] conducted experiments to reduce the severity of several types of artifacts in the ISIC-2017 dataset by cropping lesions into rectangles, which reduced the amount of artifacts present in each image. However, although they report good accuracy results (0.8893), this work did not observe the effect of individual artifacts on classification. Cassidy et al. [10] trained a wide range of CNN architectures for binary and multi-class classification using the ISIC datasets and observed through the use of Grad-CAM heatmap visualisation that classifiers would frequently focus on artifacts such as air pockets, hair, immersion fluid, measurement overlays, and physical rulers. Such artifacts were shown to have negative effects on classification performance, most critically in cases where melanoma would be misclassified.

Early attempts to address the impact of DCA were conducted by [50]. They performed simple rectangular cropping to remove DCA and inpainting of hairs. However, this may result in the removal of lesion details where the lesion is not centred within the image, or if the lesion details naturally extend beyond the rectangular crop.

Pewton and Yap [38] observed the effects of DCA on the ISIC image datasets. They found that DCA that occupied a large percentage of the image influenced the classification of melanoma vs non-melanoma with a bias towards the melanoma class. DCA were annotated and dynamic masking and removal methods were proposed. The methods proposed were evaluated with a variety of deep learning architectures, with results showing that the predictive accuracy was comparable, however the network activations showed a large improvement in focus towards lesion regions when images containing DCA were passed through DCA removal methods.

Ramella [41] created a method to detect DCA regions in skin lesion images as a pre-processing step which would be performed prior to detection and removal of hair from the images. This was achieved by using the saliency [40,42] and proximity of the DCA to the image frame. Their method was developed as part of a larger workflow to improve hair detection and subsequent removal.

While the focus of this paper is on the appearance of artifacts, we note that there are other issues which could also limit the research in this field. The main challenges include the use of machine and deep learning algorithms and the use of mobile applications. We refer readers with interest in these topics to recent review papers, such as [4] and [47] which focused on reviewing the use of machine learning and deep learning in skin lesion detection and classification. Additionally, we refer to [27] and [28] who focused on the use of mobile applications in skin cancer recognition.

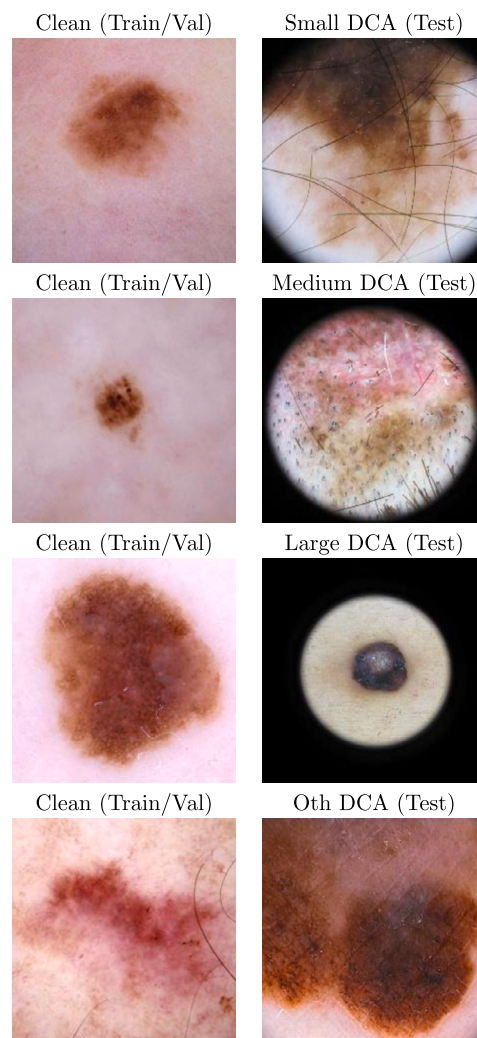


Fig. 2. Example images from the training set (clean) and the testing set (DCA).

3. Methods

This section describes the data curation processes used to introduce a DCA split balanced dataset, a proposed data augmentation method, i.e., introduction of realistic DCA, and the experimental settings used to evaluate the hypothesis.

3.1. DCA split balanced dataset

To further understand the capabilities of deep learning networks for DCA in binary classification (melanoma vs non-melanoma), and the effect of the removal methods, proposed by [38], on classification, a new balanced dataset is formed based on the publicly available dermoscopic datasets (comprising polarised, unpolarised, and a combination of both). Our proposed dataset consists of 10,250 skin lesions images, with a training set without DCA (clean images) and a testing set with DCA. Fig. 2 shows examples of images from both the training set (clean images) and the testing set (DCA images). This dataset will enable us to observe the performance of the deep learning algorithm on images with different distributions [53]. Additionally, the training set without DCA will enable us to superimpose synthetic DCA to study the differences between different DCA types. The testing set with DCA will be used to study the effect of DCA removal and inpainting algorithms (as proposed in [38]).

The baseline testing set has been created by including all DCA images from the ISIC balanced dataset proposed by [10] and separating

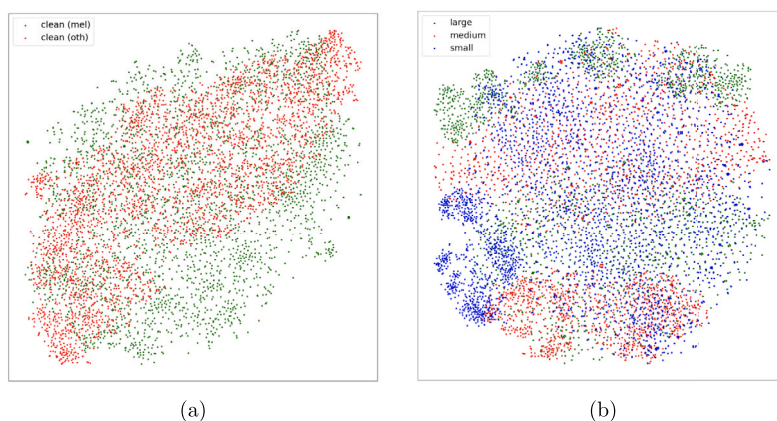


Fig. 3. Illustration of the t-SNE distribution for the curated balanced datasets: (a) Training set (train and val): clean datasets (without DCA) curated from ISIC and Fitzpatrick 17k datasets, a comparison between melanoma (mel) and others (oth); and (b) Testing set: datasets with DCA, curated from ISIC and Fitzpatrick 17k, a comparison of different sizes of DCA. In (a) green regions represent melanoma, and red regions represent ‘other’. In (b) green regions represent large DCA, red regions represent medium DCA, and blue regions represent small DCA.

Table 1

Summary of the DCA Split Balanced Dataset which contains a total of 10,250 images. Mel - melanoma; Non-Mel - Non-melanoma.

	Training set		Testing set (DCA sizes)			
	Train	Val	Small	Medium	Large	Other
Mel	2756	307	909	488	423	242
Non-Mel	2756	307	909	488	423	242

by DCA size categories (small, medium, large and others) into an individual testing set. As there are more melanoma images containing DCA than non-melanoma, the non-melanoma category is padded with 1493 images that had been excluded from the original ISIC balanced dataset [10]. The images used for padding are manually selected and measured using the DCA masking process proposed by [38]. The result of this process produces balanced testing sets where all images in each category contain DCA.

With all DCA images extracted into testing sets, the training and validation sets are unbalanced. These sets contain 1493 fewer melanoma images overall compared to non-melanoma images. In efforts to rebalance the dataset, melanoma images from the Fitzpatrick 17k dataset [18] have been inspected to determine if they are dermatoscopic images and if they contain DCA. Of the 16,529 images contained within the Fitzpatrick 17k dataset, 490 images are annotated as being melanoma by the dataset curator. Of these 490 melanoma images, we annotated 220 of these images to be free from DCA and usable in the DCA Split Balanced Dataset.

Following the incorporation of the Fitzpatrick 17k dataset into the training/validation sets, the sets required rebalancing. To rebalance the dataset, all of the melanoma images are shuffled to ensure a good distribution and then split using the holdout method. The training set contains 90% of the melanoma images, whilst the validation set contains the further 10%. As per the melanoma images, the non-melanoma images are shuffled. As there are many more non-melanoma images than melanoma - the extra non-melanoma images are removed to leave an equal number of images per class. The non-melanoma images are then split in the same way as the melanoma images.

Table 1 shows the final distribution of balanced training, validation and testing sets, with 5512, 614 and 4124 images, respectively. Although more images within the training sets would be desirable, there are a limited number of publicly available melanoma datasets containing DCA-free images.

Fig. 3 illustrates the distribution of curated datasets from ISIC and Fitzpatrick 17k, i.e., the balanced train/val datasets without DCA, and the distribution of test datasets with different sizes of DCA.

3.2. Recreating a realistic DCA

Our experiments utilise two types of synthetic DCA: (1) binary DCA as proposed by [38], and (2) a proposed more realistic DCA. Fig. 4 illustrates the processes to create a realistic DCA.

Building on the DCA masks extraction process by [38], the mask is applied to an image which is then processed using a Gaussian blur. Once a blurred image is generated, the original centre point and radius is extracted from the original DCA mask. The radius of the circle is reduced to determine the gradient of the transitional area between the image and DCA. Using this modified radius, a new mask is generated. The newly generated mask is used to determine the area of a new image which should contain the data from the blurred image, and the remaining area to contain data from the original unblurred image. The final stage involves the merging of the Gaussian blurred image and the reduced mask.

Fig. 5 illustrates the t-SNE plot of the datasets. It is noted that with superimposed of augmented binary DCA and realistic DCA, the distribution of the data changes, which demonstrate the importance to further study the effects of DCA in skin cancer analysis. When visually comparing Fig. 5(b) to Fig. 5(c), the two distributions are more separable when we augmented DCA on mel, this a result of mel tending to be darker in colour. However, this observation is inconclusive as it is limited by the datasets used in the experiment.

3.3. Classification

To evaluate the effect of DCA on binary classification, Inception-ResNetV2 [51] was selected for our experiments, as this network was the best overall performing network in our prior work on DCA [38]. Our intention is not to produce the best classification algorithm, but to investigate on the new strategy of our proposed dataset and training approaches. The three models are trained by using three types of training and validation sets: (1) clean (original images without DCA), (2) binary DCA (original images superimposed with binary DCA), and (3) realistic DCA (original images superimposed with realistic DCA).

The models were trained with no pre-trained weights to a maximum of 200 epochs, and early stopping after 10 epochs if no validation accuracy increase is achieved. The models were trained using a batch size of 64 with stochastic gradient descent as the optimiser. The model exhibiting the highest validation accuracy was saved. No model fine tuning was completed to ensure fairness and equality between the three models trained. The hardware configuration used to train all of the networks was an AMD Ryzen 7 3700X 8-core 16-thread 4.4 GHz CPU with 16 GB DDR4 3000 MHz Dual-Channel RAM and an NVIDIA Geforce RTX 3090 FE 24 GB GDDR6X GPU. The software configuration used was Python

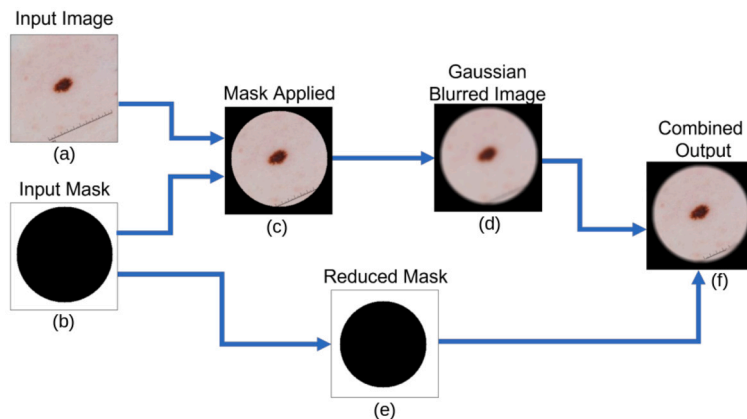


Fig. 4. Illustration of our proposed realistic DCA creation process. First, the input image (a) and input mask (b) are combined to generate a binary mask image (c). Then, the binary mask is passed through a Gaussian blur (d). Finally, the input image and the Gaussian blur image are combined using a reduced mask (e) as a filter. Any pixels in the filter that is contained within the circle are replaced with the original input image, and pixels outside are replaced with the Gaussian blurred image equivalent. The final output is the combined output (f).

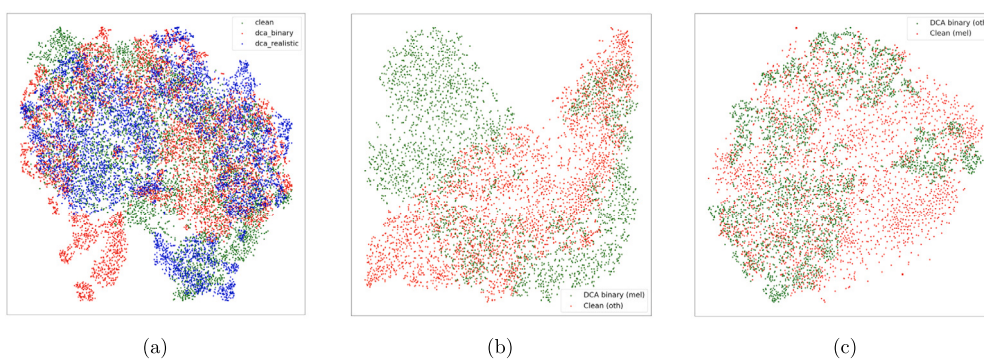


Fig. 5. The effects of DCA augmentation on the datasets used in our experiments using t-SNE plot: (a) An illustration of clean (train/val datasets) vs DCA augmented train/val datasets (binary DCA and realistic DCA); (b) An illustration of augmented DCA on melanoma (mel) and clean others (oth) distributions; and (c) An illustration of augmented DCA on oth and clean mel. In (a) green regions represent clean (non-DCA), red regions represent binary DCA, and blue regions represent realistic DCA. In (b) green regions represent binary DCA for melanoma, and red regions represent clean ‘other’ (non-DCA). In (c) green regions represent binary DCA ‘other’, and red represents melanoma clean (non-DCA).

Table 2

Trained model metrics for each training set. Acc - accuracy, AUC - Area Under the Curve.

Training dataset	Best epoch	Val Acc	Val AUC
Clean	29	0.82	0.88
Binary DCA	15	0.81	0.88
Realistic DCA	30	0.81	0.89

3.9.7, TensorFlow GPU 2.9.0-dev20220203, CUDA 11.2.1, and cuDNN 8.1 running on Windows 10.

Table 2 shows the overall model accuracy achieved with the validation set across each of the models trained.

3.4. Experiments

To provide an in-depth understanding of the effect of DCA, we conduct three experiments. In Experiment I, we test the clean model on the testing set with DCA, the testing set with DCA inpainted by Navier-Stokes (NS), and the testing set with DCA inpainted by Telea. Experiment II studies the effect of superimposed synthetic DCA (binary DCA and realistic DCA) training models on the testing set, i.e., the binary DCA model and realistic DCA model are tested on the testing set. Due to the varied size of DCA, we report the detailed results according to each category. Experiment III investigates the performance of Binary and Realistic DCA models on the testing set with DCA removal and inpainted by NS and Telea.

3.5. Performance metrics

For evaluation on the binary classification task, common performance metrics including Accuracy (Acc), True Positive Rate (TPR, also known as sensitivity), True Negative Rate (TNR, also known as specificity), Precision, F1-Score, and Area under the Receiver Operating Curve (AUC) are used. To further elaborate on the differences between the network performance of the different models, Grad-CAM [46] is used to extract the network activation gradients from the last convolutional layer of each network. The gradients extracted produce a heatmap to show which parts of an image the network focuses on to determine the classification result. For our experiments, a class implementation of Grad-CAM created by [44] has been used. Within the resulting heatmaps, the bright yellow regions are areas used heavily by the network for prediction.

We propose a new quantitative method by introducing intensity measures for prediction activation heatmaps. To quantify the area targeted by the heatmaps produced for the Grad-CAM experiments, the heatmaps for all test sets across all networks are extracted from the test predictions. The corresponding mask for each of the images is used to segregate the two areas in the image - the external section of the image is the DCA region, and the internal section of the image is the area in which the lesion resides.

As the areas of the image that are focused on by the network are brighter than those that are not - the brightness values make it possible to measure the difference that each method has made on the

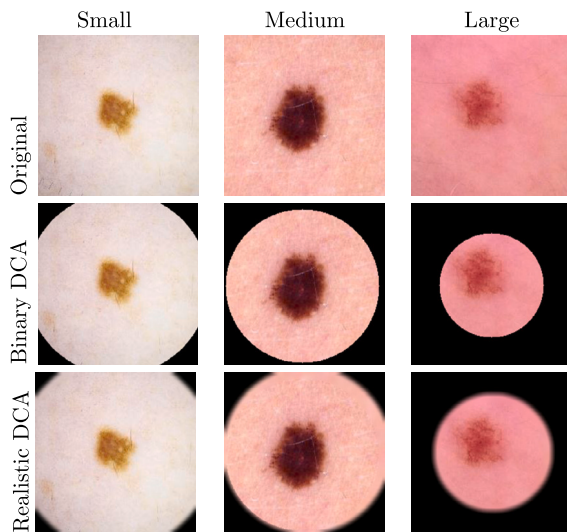


Fig. 6. Illustration of different superimposed DCA sizes on skin lesion images. The first row shows the original images from the training set. The second row shows the superimposed binary DCA, and the third row shows the superimposed realistic DCA.

corresponding heatmaps. Using the internal and external areas of the heatmap image, the root mean square (RMS) contrast and the average brightness value is calculated for each heatmap. This process was completed using the ImageStat method which is part of the Pillow Python library [30]. RMS contrast is not dependent on the spatial distribution or the angular frequency content of contrast in the image, and is defined as the standard deviation of pixel intensities [37]. The relevant mathematical expression for RMS contrast is as follows:

$$RMS = \sqrt{\frac{1}{MN} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} (I_{ij} - \bar{I})^2} \quad (1)$$

where intensities I_{ij} are the i -th j -th element of the two-dimensional image of size M by N . \bar{I} is the average intensity of all pixel values in the image. The image I is assumed to have pixel intensities normalized in the range of $[0, 1]$.

4. Results

This section presents the results of our proposed synthetic DCA and experiment I-III.

4.1. Synthetic DCA

Fig. 6 shows an example of superimposed the binary DCA and realistic DCA on a clean image from training set.

As can be seen in Fig. 6, the proposed realistic DCA method produces visually effective results for each of the DCA sizes. Fig. 7 presents a close up visual patch comparison (35x35 px) of a true DCA (real DCA from the testing set), a binary DCA superimposed on an image from training set, and a realistic DCA superimposed on a similar image from training set. It can be clearly seen that the realistic DCA has a smooth transition into the DCA region from the image, much like the true DCA image whereas the binary DCA forms a distinct solid boundary between lesion and DCA region.

4.2. Experiment I: the effect of DCA on skin lesions classification

Table 3 shows the evaluation metrics for the clean model (without DCA) on the testing set (with DCA). Due to the composition of the training set, it is expected that the results will be mostly predicted as

Table 3

The performance of the clean model on the test set (Original); test set with DCA inpainted by Navier-Stokes (NS); and test set with DCA inpainted by Telea (Telea).

Test set	Metrics					
	Acc	TPR	TNR	Precision	F1	AUC
Original	0.57	0.90	0.23	0.54	0.68	0.61
NS	0.59	0.86	0.31	0.56	0.68	0.64
Telea	0.59	0.84	0.34	0.56	0.67	0.65

Table 4

The effect of DCA according to DCA size on the performance of the clean model on the test set (Original); test set with DCA inpainted by Navier-Stokes (NS); and test set with DCA inpainted by Telea (Telea).

Test set - DCA size	Metrics					
	Acc	TPR	TNR	Precision	F1	AUC
Original - small	0.59	0.86	0.32	0.56	0.68	0.63
NS - small	0.58	0.87	0.30	0.55	0.67	0.62
Telea - small	0.58	0.86	0.30	0.55	0.67	0.62
Original - medium	0.57	0.91	0.24	0.54	0.68	0.64
NS - medium	0.58	0.90	0.26	0.55	0.68	0.65
Telea - medium	0.59	0.88	0.30	0.56	0.68	0.66
Original - large	0.51	0.99	0.01	0.50	0.67	0.58
NS - large	0.62	0.81	0.43	0.59	0.68	0.67
Telea - large	0.61	0.72	0.50	0.59	0.65	0.68
Original - other	0.58	0.90	0.26	0.55	0.67	0.65
NS - other	0.57	0.87	0.27	0.54	0.67	0.65
Telea - other	0.57	0.87	0.27	0.54	0.67	0.64

melanoma, with high sensitivity (TPR) but low specificity (TNR), as illustrated in the first row of Table 3. This result is aligned with previous research which indicates that superficial spreading melanoma are dark brown or black in appearance [43]. After removal of DCA by inpainted with Navier-Stokes and Telea, we observed marginal improvement to TNR, Precision, F1-Score and AUC.

To further support the efficiency of DCA removal methods on the testing set, the performance of all test sets with different DCA sizes are compared. Table 4 shows the performance of the clean model evaluated on the test set.

The results are comparable across the 'small' and 'other' DCA sizes, however the medium and large DCA sizes show a more notable increase in accuracy, TNR and precision. The largest accuracy increase is seen in the large DCA test sets where the NS inpainting method achieves 11% greater accuracy compared to the baseline test set containing original DCA images. Another observation is that the Original large DCA has 0.01 TNR, which means almost all the large DCA were classified as melanoma. With Telea inpainting, the TNR improved 49%, with 0.50 TNR.

4.3. Experiment II: the effect of superimposed DCA

When comparing the overall accuracy performance of the original test set across the 3 networks in Table 5, an accuracy increase of 3% is seen when the images are predicted using the model trained with binary DCA, and an increase of 4% when predicted with the model trained with realistic DCA. The performances across the original test set also shows a significant increase in TNR on the realistic DCA model - meaning that there is less of a blanket classification of melanoma across the network. Due to the variance of predictions made, this model also suffers from a reduction in TPR when compared to the predictions made with the cleanly trained model.

To further investigate the performance differences across the different models and test sets, metrics were then generated for all test sets across all DCA sizes. Table 6 shows the metrics that were generated

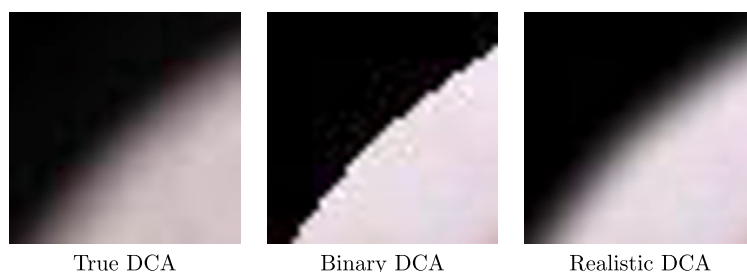


Fig. 7. A visual comparison (close-up 35 × 35 pixels) of true DCA and sythnetic DCAs. (Left) True DCA from the test set; (Middle) Binary DCA superimposed onto an image in the training set as proposed by [38]; (Right) Our proposed realistic DCA on an image in the training set. Note that visually, our proposed realistic DCA closely resembles the true DCA.

Table 5

The performance of all trained networks (clean model, superimposed binary DCA model and superimposed realistic DCA model) on the test set.

Model	Metrics					
	Acc	TPR	TNR	Precision	F1	AUC
Clean	0.57	0.90	0.23	0.54	0.68	0.61
Binary DCA	0.60	0.91	0.29	0.56	0.70	0.66
Realistic DCA	0.61	0.73	0.49	0.59	0.65	0.66

Table 6

The performance of all trained networks (clean model, superimposed binary DCA model and superimposed realistic DCA model) on the test set with different DCA sizes.

DCA size	Model	Metrics					
		Acc	TPR	TNR	Precision	F1	AUC
Small	Clean	0.59	0.86	0.32	0.56	0.68	0.63
	Binary DCA	0.61	0.90	0.33	0.57	0.70	0.67
	Realistic DCA	0.60	0.85	0.35	0.57	0.68	0.65
Medium	Clean	0.57	0.91	0.24	0.54	0.68	0.64
	Binary DCA	0.63	0.94	0.31	0.58	0.72	0.68
	Realistic DCA	0.64	0.75	0.53	0.62	0.68	0.70
Large	Clean	0.51	0.99	0.01	0.50	0.67	0.58
	Binary DCA	0.55	0.96	0.13	0.53	0.68	0.62
	Realistic DCA	0.60	0.39	0.80	0.66	0.50	0.63
Other	Clean	0.58	0.90	0.26	0.55	0.67	0.65
	Binary DCA	0.60	0.83	0.36	0.57	0.67	0.67
	Realistic DCA	0.58	0.81	0.35	0.56	0.66	0.65

for the original DCA sets across each of the models and Fig. 8 shows a line graph comparing the model accuracy performance between the different DCA sizes. Full model performance metrics generated across the three networks are available on our GitHub repository: https://github.com/mmu-dermatology-research/dca_artifact_removal.

As can be seen in Table 6 and Fig. 8, the overall accuracy has increased for each of the test sets when the network is trained using both binary and realistic DCA. The largest accuracy increase can be seen for the large DCA test sets where the network trained with realistic DCA shows an increase in accuracy of 9%. Significant increases can also be seen for TNR on medium and large augmented DCA models indicating that more non-melanoma images were correctly classified.

Fig. 8 shows that network accuracy performance was equal to or improved from the baseline DCA set tested on the cleanly trained network. The binary DCA model shows the best overall accuracy for both the small and 'other' sized DCA test sets, the realistic DCA showed the best overall accuracy for the medium sized test set, and the Telea in-painted images on the clean model showed the best overall accuracy for the large testing set (Realistic DCA). Although both of the results for the large testing set from the models trained on synthetic DCA images improves from the baseline performance - the models still appear to be hindered due to the large DCA size.

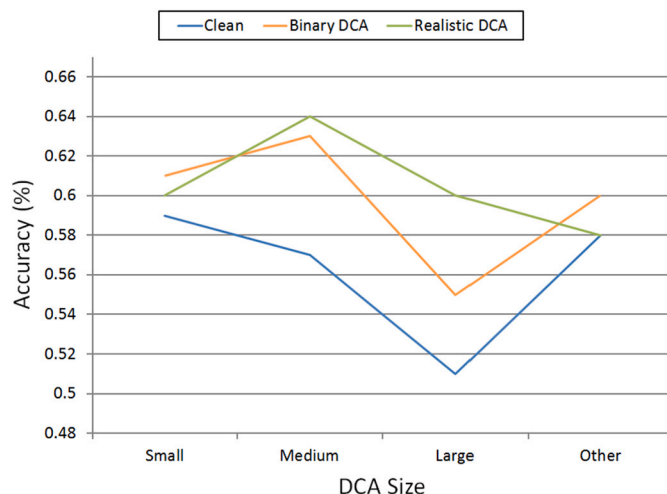


Fig. 8. Model accuracy for small, medium, large, and other DCA sizes.

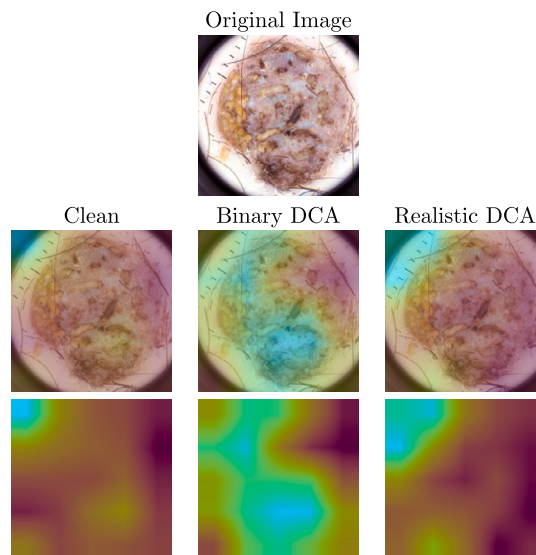


Fig. 9. Illustration of Grad-CAM results on a small DCA image for clean, binary, and realistic DCA models.

Fig. 9 shows the output results of a small DCA image across each of the three networks. The activations on the heatmap generated from the clean model focus largely on the area of the lesion, however, the top-left corner also exhibits significant focus. When comparing the results from the binary and realistic DCA trained models, the corner region becomes less of an area of interest when DCA is used for training. Although improvement can be seen in the activation area where the lesion is located, activations

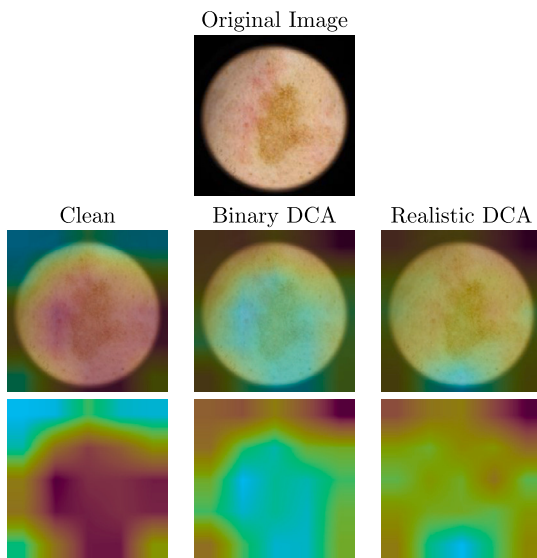


Fig. 10. Illustration of Grad-CAM results on a medium DCA image for clean, binary, and realistic DCA models.

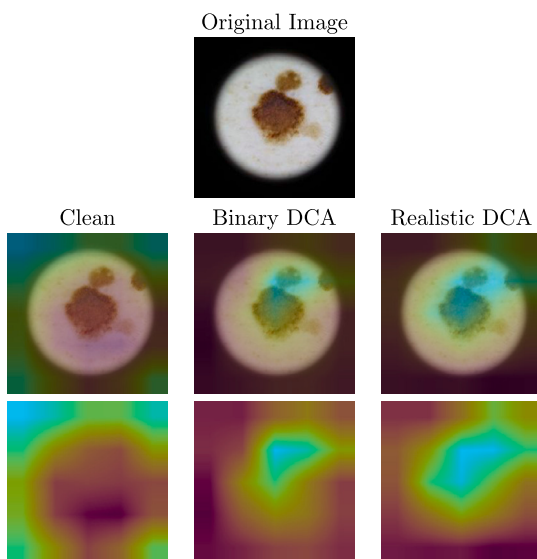


Fig. 11. Illustration of Grad-CAM results on a large DCA image for clean, binary, and realistic DCA models.

are still partially present on the dark regions. The network showing the best activations for small DCA images is the realistic DCA model.

Fig. 10 shows the output results of a medium DCA image across each of the three networks. The network activations on the cleanly trained model mostly focus on the upper corners of the image for medium sized DCA. When compared to the results from the binary and realistic DCA models, we observe that the network is able to effectively ignore the majority of the DCA region. The activations for the binary and realistic DCA models are similar, though the binary DCA activations appear to encapsulate the lesion more finely and ignore more of the DCA region than the model trained on realistic DCA. Both the binary and realistic DCA model results show a large improvement in the focus of activations for medium DCA images.

Fig. 11 shows similar results to Fig. 10 for large DCA. The model trained on a clean dataset almost entirely focuses on the DCA region, while the models trained on binary DCA and realistic DCA show a large improvement on the focus towards the activations. The results for both DCA trained models are both similar, with the binary DCA image show-

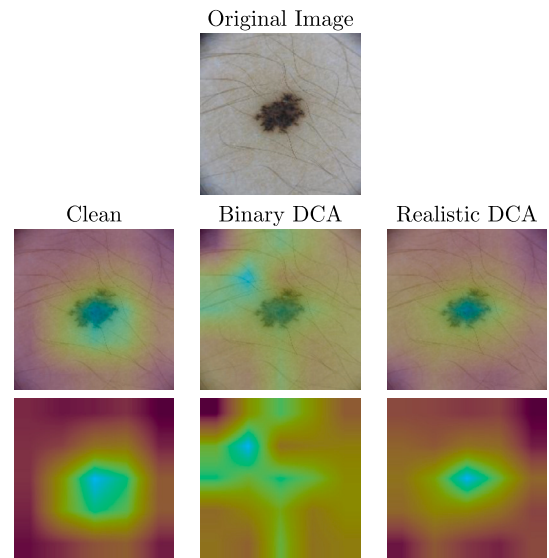


Fig. 12. Illustration of Grad-CAM results on a DCA image from the 'other' category for clean, binary, and realistic DCA models.

Table 7

The performance of the best TNR from the inpainted method and the superimposed method on the test set with different DCA sizes.

DCA size	Method	Metrics					
		Acc	TPR	TNR	Precision	F1	AUC
Small	Inpainted	0.58	0.87	0.30	0.55	0.67	0.62
	Superimposed	0.60	0.85	0.35	0.57	0.68	0.65
Medium	Inpainted	0.59	0.88	0.30	0.56	0.68	0.66
	Superimposed	0.64	0.75	0.53	0.62	0.68	0.70
Large	Inpainted	0.61	0.72	0.50	0.59	0.65	0.68
	Superimposed	0.60	0.39	0.80	0.66	0.50	0.63
Other	Inpainted	0.57	0.87	0.27	0.54	0.67	0.65
	Superimposed	0.60	0.83	0.36	0.57	0.67	0.67

ing a smaller surface area of activations as opposed to the realistic DCA model. Both DCA models show significant improvements on large DCA images when compared to the results from previous experiments using small and medium DCA.

Fig. 12 shows the output results for a 'other' DCA image across each of the three networks. Strong activations for both the model trained on the clean dataset and the model trained on realistic DCA are clearly visible. The heatmap generated from the model trained on binary DCA images completely loses focus on the lesion area. Between the results from the clean model and the realistic DCA model, the activations in the clean model are more strongly focused on the area of interest.

4.4. Experiment III: DCA removal testing set vs superimposed synthetic DCA training set

To understand the differences in model performance between the DCA removal processes and superimposed synthetic DCA in the training process, it is necessary to cross compare the results produced across each of the models. From Experiment I and II, we observed that both approaches resulted in improved accuracy, TNR and precision. Therefore, we compare their performance on different DCA sizes based on the best TNR from each approach. Table 7 compares the results. Overall, the superimposed synthetic DCA training model performed the best in accuracy, TNR, and precision. However, the method using inpainted achieved a superior result in TPR.

Fig. 13 and Fig. 14 show the different results generated for the inpainted DCA images using the Navier-Stokes and Telea inpainting

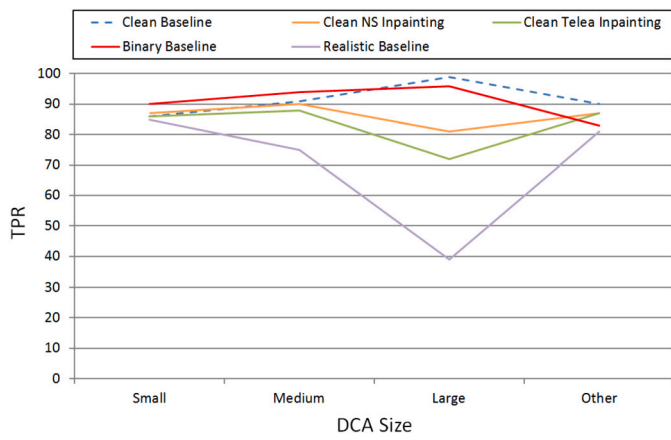


Fig. 13. TPR plots for result metrics generated across each trained network.

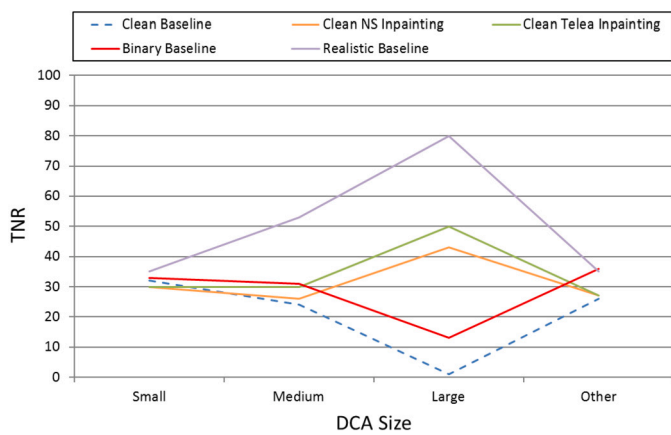


Fig. 14. TNR plots for result metrics generated across each trained network.

methods against the results generated for DCA images tested within the models trained on superimposed synthetic DCA images.

It can be seen that the results generated for the inpainted test sets are comparable to the results generated on the DCA test sets on models trained with synthetic DCA. Both methods produce similar accuracy improvements overall. In terms of TPR, the binary baseline model outperforms the other models for small and medium DCA sizes, while the clean baseline model performs best for large and other models. The realistic DCA model shows the largest decrease in TPR, however it also shows the largest increase in TNR. This suggests that the model is able to identify more features within non-melanoma images. The clean and binary models show the most notable drops in TNR for large DCA, both with < 0.20 TNR, a difference of > 0.60 compared to the best performing model (realistic DCA).

To further confirm and compare the differences between the methods, Fig. 15 shows the output of Grad-CAM heatmap activations across the different networks. Each column in this figure shows a different method used for evaluation, and each pair of rows in the figure represent the testing set and the activations shown with Grad-CAM.

Several observations are noted in Fig. 15. Without any bias, we selected the images randomly from each DCA size. Firstly, we can see that when comparing the performance on small DCA, every method improved the focus of the network and expanded the focus from the upper left corner to the rest of the image. It is difficult to determine the most efficient method from the results generated for this particular example of small DCA as none of the results show a perfect focus area. When analysing the performance on medium DCA, it can be seen that all evaluation methods showed a large improvement on the original network activations. For each of the methods the focus is taken away from the DCA and distributed across the image. The method showing the best fit-

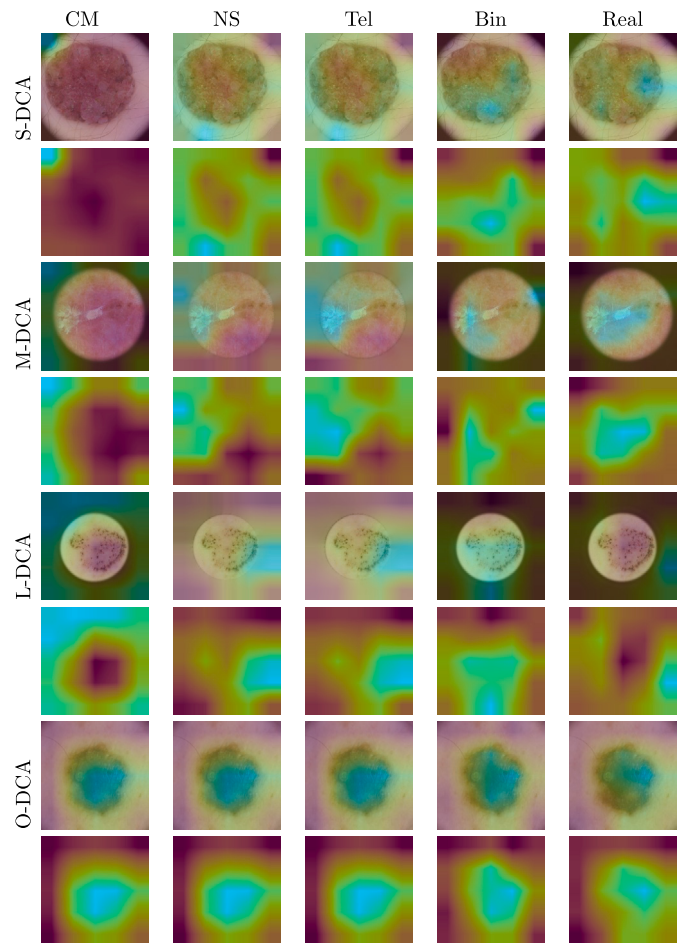


Fig. 15. Grad-CAM heatmap results to visually compare activations of different models on the test set with small (S-DCA), medium (M-DCA), large (L-DCA), and other (O-DCA) DCA sizes. Clean Model (CM) is the model (clean images) used for classification on the test set (with DCA); Navier-Stokes (NS) is the model used for classification on the test set inpainted by Navier-Stokes; Telea (Tel) is the model used for classification on the test set inpainted by Telea; Binary (Bin) is the model (clean images superimposed with synthetic binary DCAs) used for classification on the test set (with DCA); and Realistic (Real) is the model (clean images superimposed with synthetic realistic DCAs) used for classification on the test set (with DCA).

ting activations to what are expected can be seen in the results from the model trained on realistic DCA.

When comparing the activations for the large DCA test sets, a scenario similar to the medium DCA results is observed where the activations are mostly focused on the DCA region on the baseline results with the clean model, whereas most other methods manage to disrupt this focus and focus more on the areas of interest. The method displaying the best overall focus is the model trained on binary DCA, where the activations focus almost entirely on the central region of the image. The results for the DCA consuming less than 1% of the image do not display much variability compared to the baseline generated on the clean model with original data. This is due to the minimal area that the DCA occupies.

4.5. DCA removal combined with synthetic training

When evaluating the networks across each of the testing sets, the images inpainted with both Navier-Stokes and Telea methods were also used to evaluate both of the models trained on synthetic DCA images. Table 8 shows the results generated from evaluating the Binary DCA and Realistic DCA with images inpainted from both methods.

Table 8

Performance evaluation of the superimposed synthetic DCA training models on DCA inpainted testing sets. It is noted that the superimposed binary DCA model performed better in TPR, but the superimposed realistic DCA model performed better in TNR on the original test set than the inpainted images.

Model used	Test set	Metrics					
		Acc	TPR	TNR	Precision	F1	AUC
Binary DCA	Original - small	0.61	0.90	0.33	0.57	0.70	0.67
	NS - small	0.61	0.89	0.33	0.57	0.70	0.67
	Telea - small	0.61	0.89	0.34	0.57	0.70	0.67
	Original - medium	0.63	0.94	0.31	0.58	0.72	0.68
	NS - medium	0.63	0.91	0.36	0.59	0.71	0.70
	Telea - medium	0.64	0.89	0.40	0.60	0.71	0.71
	Original - large	0.55	0.96	0.13	0.53	0.68	0.62
	NS - large	0.65	0.73	0.56	0.62	0.67	0.69
	Telea - large	0.64	0.65	0.62	0.64	0.64	0.71
	Original - oth	0.60	0.83	0.36	0.57	0.67	0.67
	NS - oth	0.59	0.81	0.38	0.56	0.67	0.66
	Telea - oth	0.59	0.81	0.38	0.56	0.67	0.66
Realistic DCA	Original - small	0.60	0.85	0.35	0.57	0.68	0.65
	NS - small	0.60	0.85	0.34	0.56	0.68	0.65
	Telea - small	0.60	0.85	0.35	0.57	0.68	0.65
	Original - medium	0.64	0.75	0.53	0.62	0.68	0.70
	NS - medium	0.63	0.87	0.39	0.59	0.70	0.67
	Telea - medium	0.63	0.84	0.43	0.60	0.70	0.68
	Original - large	0.60	0.39	0.80	0.66	0.49	0.63
	NS - large	0.59	0.60	0.58	0.59	0.60	0.64
	Telea - large	0.60	0.49	0.70	0.62	0.55	0.64
	Original - oth	0.58	0.81	0.35	0.55	0.66	0.65
	NS - oth	0.57	0.79	0.36	0.55	0.65	0.64
	Telea - oth	0.57	0.79	0.36	0.55	0.65	0.64

As can be seen in Table 8, the accuracy for superimposed binary DCA increases but the accuracy for superimposed realistic DCA decreases when comparing original test sets with inpainted test sets. For medium DCA test sets, the realistic DCA model achieved the best result on original test set. In contrast, the binary model achieved better results when evaluated using inpainted images. The largest discrepancy is exhibited in the large DCA test set, where the binary DCA model achieved the best result in TPR of 0.96, but the poorest result in TNR of 0.13. When evaluating with inpainted images, a large increase in TNR for binary DCA model can be seen.

To determine the significance of these results, we calculate the p-value for the F1-scores using analysis of variances (ANOVA) single factor statistical analysis with an alpha value of 0.05. In this analysis, we compared the binary DCA with the realistic DCA F1-scores. The analysis shows that $p = 0.0406$, indicating that the differences between the F1-score values are significant.

Fig. 16 shows the different heatmaps generated for an image containing a medium DCA and Fig. 17 shows the different heatmaps generated for an image containing a large DCA. The same image is inpainted and examined again in each network to determine the differences in activations. The activations for the ‘small’ and ‘other’ DCA sizes do not show significant differences. To aid the comparison between the base test set and the inpainted test sets, the model trained on binary DCA using the baseline test set is used as it produces the best TPR across each of the DCA sizes present as seen in Table 8.

It can be seen that the activations for both inpainting methods produce similar results for each of the DCA images sizes. For the medium Binary Base, Clean, and Binary NS/Telea images, the activations are generalised around both lesion and DCA regions, while the medium Realistic activations show significantly more concentrated focus towards the lesion region. For large Binary Base, Clean, and Binary NS/Telea, the activations are generally more focused on the lesion regions, with the large Realistic images showing a much more generalised spread of activations across lesion and DCA. Whilst the overall performance of the networks appears to improve when inpainted images are evaluated

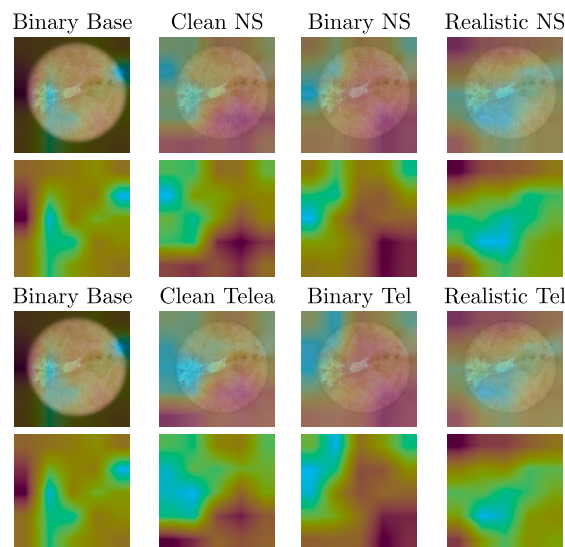


Fig. 16. Grad-CAM results for medium inpainted images on networks trained with synthetic DCA. Tel - Telea.

on a network trained by images containing synthetic DCA, the class activation maps show that the area in which the network is focused is not as targeted on the lesion regions as the intended test sets.

4.6. Heatmap contrast and brightness intensity measures

Table 9 shows a breakdown of heatmap contrast and brightness intensity measures for all images for each DCA size and all models. We observe from the heatmap intensity results that almost all models show a clear bias towards focussing on internal lesion regions, with the exception of the clean original small, medium, and large models which

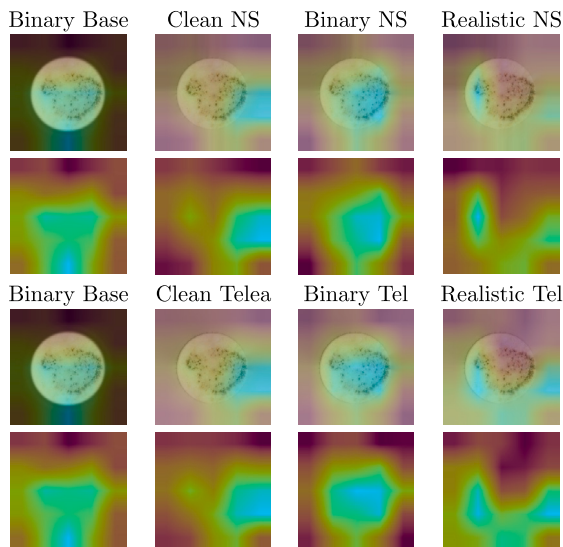


Fig. 17. Grad-CAM results for large inpainted images on networks trained with synthetic DCA. Tel - Telea.

show a clear bias of network focus towards the external DCA regions (shown in the negative mean difference values). The clean original small model shows a slight bias towards external DCA regions (RMS mean diff = -2.50, avg. brightness mean diff = -0.53), while the clean original medium model (RMS mean diff = -40.28, avg. brightness mean diff = -40.57) and the clean original large model (RMS mean diff = -44.87, avg. brightness mean diff = -48.44) show large biases towards external DCA regions (> -40).

The model with the highest RMS contrast and highest average brightness intensity bias towards internal (lesion) regions is the binary DCA original large model (RMS mean = 136.85, avg brightness mean = 135.22). These results correlate with the TPR for this model (Table 8) which has the highest TPR (0.96). However, the correlation is not present for accuracy, TNR, precision, and AUC, which were shown to be the lowest reported metrics for this model. This is likely to be a consequence of the binary DCA original large model comprising of mostly melanoma examples, which makes TPR occurrences more prevalent.

The model with the highest RMS mean and highest average brightness mean for external DCA regions is the clean original large model (RMS mean = 148.46, avg brightness mean = 146.95). These results correlate with the TPR (0.39) and F1-score (0.49) results for this model (Table 8), which shows the lowest results for these metrics, indicating that the model was focussed more on external DCA regions.

The model with the highest RMS mean difference and highest average brightness mean difference is the clean Telea 'other' model (RMS mean diff = 26.02, avg brightness mean diff = 27.23). The intensity metrics for this model indicate that it showed the highest shift in network focus towards internal (lesion) features. However, this shift alone was not sufficient enough to provide it with the highest overall scores in accuracy, TPR, TNR, precision, F1-score, and AUC.

4.7. Test results for external test sets

In this section we explore the ability of the clean, binary, and realistic models to detect other skin diseases and two external non-skin lesion datasets. These experiments can help to determine if these models are also affected by the presence of DCA. Fig. 18 demonstrates the output results of pyogenic granuloma and sebaceous hyperplasia using the three networks. For the clean model, we observe that the network activations focus mostly on the DCA regions. When compared to the results from the binary and realistic DCA models, we observe that the networks are able to effectively ignore the majority of the DCA regions.

For additional experiments, we used the test sets from the endoscopic SLAM dataset (EndoSLAM) [36], and the diabetic retinopathy dataset [14]. Both of these datasets exhibit naturally occurring DCA, which makes them ideal candidates for further investigation. Fig. 19 shows the output results of the other two external datasets on different domains using three networks. The network activations on the clean trained model indicate that most focus is directed to DCA regions. When compared to the results from the binary and realistic DCA models, we observe that the findings are aligned with the results for skin diseases, as presented in the previous experiment.

5. Discussion

Between Fig. 9, Fig. 10, Fig. 11, and Fig. 12 the Grad-CAM heatmap activations show relatively similar results across each of the DCA sizes fed into the network. For the small, medium and large DCA, improvements in the focus of the activations can be seen for both of the models trained on binary and realistic DCA. Both models exhibit similar results. Results for DCA covering less than 1% of the image show similar or worse activations on models trained with DCA images when comparing to the clean model.

We observe that the binary DCA model achieved better TPR than the realistic DCA model in our experiments, but it did not outperform the clean model. A possible reason for this could be that the gradient inherent in the smooth transitions between skin and DCA may represent an introduction of further complex features that creates an additional learning challenge to the network. However, with the use of superimposed realistic DCA, a notable improvement in TNR and precision indicate the network was able to learn to handle the DCA and was capable of reducing the biases of classifying DCA as melanoma.

Other studies, such as those by [32], demonstrated that removal of some surrounding features via cropping, when used in combination with other techniques such as ensembling, does not necessarily result in a deterioration in model performance. We therefore suggest that removal of some surrounding features, either by cropping, or by the introduction of DCA, may be beneficial to model performance. However, this is likely to be dependent on how much of the surrounding features are removed or obfuscated by DCA, and how much of the lesion is centred within the image.

During our analysis of the DCA images, we observed that there may be some edge cases where DCA occluded outer sections of the lesion. For future work, it may be useful to identify all such cases and analyse the possible effect they may have on classification tasks.

During our analysis of the heatmaps for the skin lesion images, we observed that many of the heatmaps showed that networks would focus on DCA regions. We hypothesised that the DCA regions may exhibit artifacts that were not visible to the human eye and were causing networks to focus on these areas. For example, we speculated that there may be JPG artifacts present within the DCA areas that introduced additional complex features into the DCA regions, or that the DCA region did not comprise of only black pixels. To test for this, we performed an additional analysis on images taken from the clean large DCA model. Images were selected from the test results of this model as it exhibited a heavy bias towards classifying almost all test images as melanoma. We adjusted the contrast of the original lesion images which naturally exhibit DCA. This process revealed that many lesion images with DCA exhibited complex pixel patterns that radiate outwards from the border regions between lesion area and DCA area that would not ordinarily be visible to the human eye. Due to the uniformity of these pixel patterns around the lesion / DCA borders, we speculate that these patterns are the result of light leakage from the dermoscope. Many dermoscope models are equipped with a built-in array of LED lights that surround the perimeter of the lens. Fig. 20 shows test images together with corresponding increased contrast images and Grad-CAM heatmap activation images from the clean large DCA model. The extent to which the light leakage is present can vary between examples. This may be due to the use

Table 9

Contrast and brightness intensity measures according to DCA size. RMS - root mean square; mean difference = *internal mean – external mean* (highest value shows a higher ratio of activations in the target area; positive for lesion area and negative for DCA area); Std - standard deviation.

Model Used	Test Set	RMS					Avg. Brightness				
		Internal		External		Diff (Mean)	Internal		External		Diff (Mean)
		Mean ↑	Std ↓	Mean ↓	Std ↓		Mean ↑	Std ↓	Mean ↓	Std ↓	
Clean	Original - small	111.10	22.85	113.60	23.10	-2.50	107.43	23.21	107.96	23.57	-0.53
	NS - small	125.07	14.22	107.11	27.06	17.96	121.25	15.60	102.14	27.51	19.11
	Telea - small	125.12	14.13	107.05	27.18	18.07	121.28	15.51	102.12	27.64	19.16
	Original - medium	89.82	14.24	130.10	7.89	-40.28	85.04	14.81	125.61	9.00	-40.57
	NS - medium	123.69	15.42	115.37	19.44	8.32	120.35	16.57	110.86	20.19	9.49
	Telea - medium	125.10	14.97	115.06	19.94	10.04	121.84	16.15	110.58	20.70	11.26
	Original - large	103.59	10.07	148.46	6.89	-44.87	98.51	10.65	146.95	7.82	-48.44
	NS - large	132.14	17.52	115.95	17.23	16.19	130.03	18.69	111.67	18.07	18.36
	Telea - large	132.26	17.18	116.78	17.69	15.48	130.14	18.30	112.53	18.62	17.61
	Original - oth	124.23	14.34	99.18	28.46	25.05	120.18	15.53	93.93	28.67	26.25
	NS - oth	124.84	14.31	98.88	28.18	25.96	120.82	15.57	93.64	28.36	27.18
	Telea - oth	124.81	14.30	98.79	28.19	26.02	120.78	15.56	93.55	28.39	27.23
Binary DCA	Original - small	132.79	11.00	111.27	23.86	21.52	129.88	11.93	106.24	24.92	23.64
	NS - small	132.40	10.84	112.95	24.74	19.45	129.43	11.71	108.13	25.93	21.30
	Telea - small	132.23	10.83	113.14	25.02	19.09	129.23	11.69	108.40	26.25	20.83
	Original - medium	135.98	10.67	111.41	13.70	24.57	133.84	11.52	106.79	14.07	27.05
	NS - medium	134.52	11.01	117.09	16.56	17.43	132.20	11.86	112.85	17.34	19.35
	Telea - medium	134.30	10.88	118.40	17.54	15.90	131.94	11.73	114.27	18.50	17.67
	Original - large	136.85	12.71	114.98	14.08	21.87	135.22	13.46	110.93	14.64	24.29
	NS - large	135.34	15.42	120.65	17.64	14.69	133.53	16.32	116.98	18.46	16.55
	Telea - large	133.78	16.08	123.18	18.69	10.60	131.85	17.06	119.70	19.70	12.15
	Original - oth	130.74	10.41	109.04	32.06	21.70	127.27	11.03	104.42	33.83	22.85
	NS - oth	130.82	10.89	109.06	32.16	21.76	127.37	11.52	104.46	33.95	22.91
	Telea - oth	130.81	10.97	109.06	32.23	21.75	127.34	11.61	104.47	34.02	22.87
Realistic DCA	Original - small	130.83	11.24	111.81	24.62	19.02	127.66	12.28	106.73	25.38	20.93
	NS - small	130.17	11.16	112.31	25.38	17.86	126.93	12.14	107.47	26.28	19.46
	Telea - small	130.23	11.11	112.20	25.63	18.03	126.98	12.10	107.35	26.55	19.63
	Original - medium	130.95	12.29	118.04	16.91	12.91	128.32	13.26	113.74	17.61	14.58
	NS - medium	131.61	12.57	118.56	16.27	13.05	129.19	13.46	114.12	16.90	15.07
	Telea - medium	132.58	11.92	118.89	16.91	13.69	130.17	12.81	114.48	17.64	15.69
	Original - large	130.29	17.74	123.03	14.72	7.26	128.26	18.83	119.59	15.54	8.67
	NS - large	136.25	15.52	118.74	13.16	17.51	134.54	16.54	114.66	15.78	19.88
	Telea - large	135.18	15.53	120.55	16.21	14.63	133.31	16.59	116.68	17.06	16.63
	Original - oth	129.19	10.81	108.41	27.92	20.78	125.67	11.51	103.62	29.27	22.05
	NS - oth	129.42	11.50	108.20	27.89	21.22	125.90	12.23	103.41	29.25	22.49
	Telea - oth	129.47	11.53	108.34	27.97	21.13	125.95	12.28	103.56	29.35	22.39

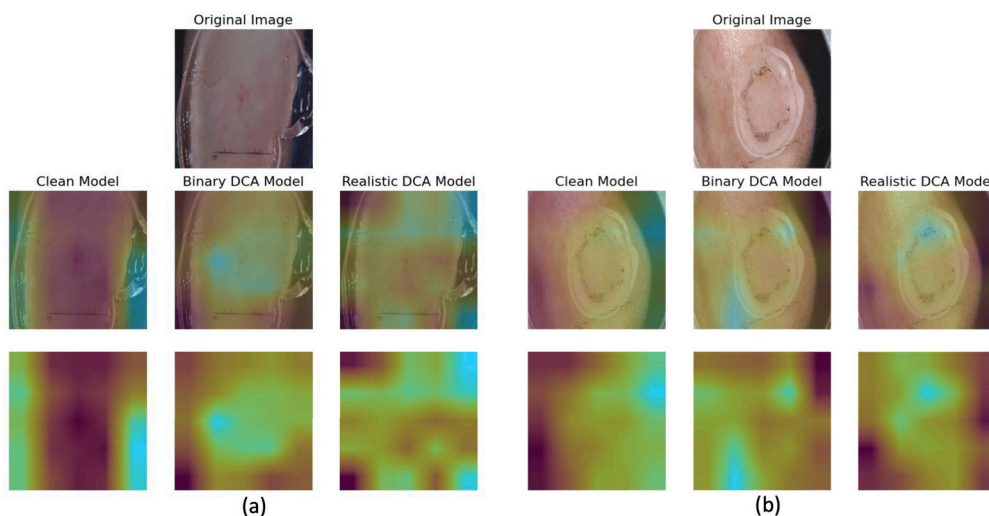


Fig. 18. Illustration of Grad-CAM results on other skin diseases for the clean, binary, and realistic DCA models: (a) pyogenic granuloma, and (b) sebaceous hyperplasia.

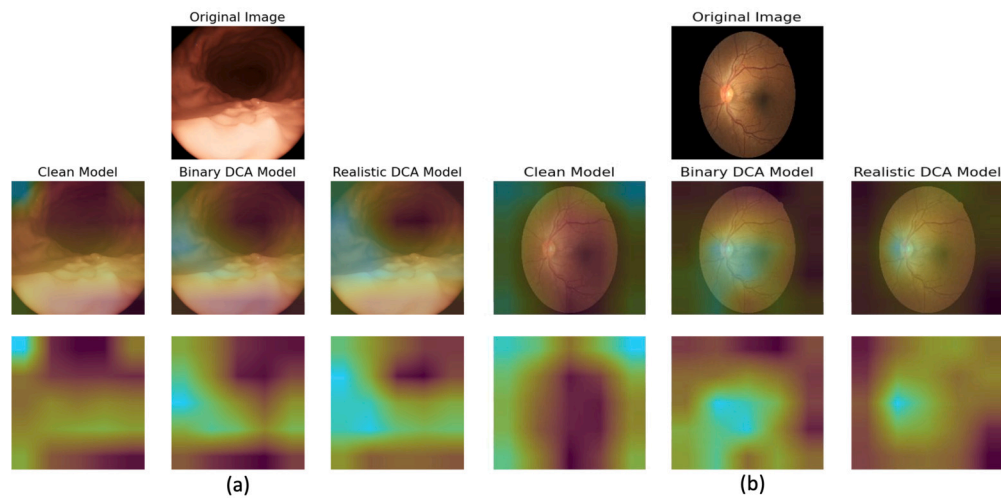


Fig. 19. Illustration of Grad-CAM results on external datasets for clean, binary, and realistic DCA models: (a) endoscopic SLAM dataset (EndoSLAM) [36] and (b) diabetic retinopathy dataset [14].

of different dermoscope models (size and quality), dermoscope settings, or the amount of pressure applied to the skin during an examination. If a dermoscope is powered by a battery, the amount of light leakage may also be affected by battery power levels which would affect the intensity of light being emitted by the device.

We note that at present, the ability to quantify heatmaps in classification tasks may be limited in that the ground truth labelling does not specify exactly where in the image the lesion is present. Moreover, the method may be of more use in scenarios where ground truth delineation is available, such as those found in segmentation tasks. However, the task covered by this present paper utilises masks that have been generated to produce DCA which give an approximate indicator as to where the skin lesion is present, i.e., within the masked lesion region. This allows for an approximate calculation of heatmap intensity in relation to lesion regions.

We observe from the images in Fig. 20 that most examples show that the network focused mainly on the surrounding DCA regions, including areas of light leakage, regardless of the prediction result. We draw two conclusions from this: (1) the network would use outer regions of the image to form both correct and incorrect predictions that may be due to the presence of black pixels, the complex features introduced by the light leakage, or a combination of both, and (2) the network may be learning spurious correlations between the features in the DCA and light leakage areas within the DCA, hence the apparent randomness of the results. As shown in Fig. 20, we indicate the brightest (blue circles) and darkest (red circles) regions of the heatmap activation images, with bright regions representing the highest levels of network focus, and dark regions indicating the lowest levels of network focus. In these examples, lesion details are present in the outer regions of the lesion area which are positioned close to the DCA regions. It may therefore be possible that although the network appears to be focusing mostly on DCA regions, for correct predictions the network is able to determine class using those lesion features that are close to the DCA perimeter. Fig. 21 shows the heatmaps from Fig. 20 with masks applied to show that the network still directs some of its focus towards the actual lesion regions.

The DCA and light leakage results shown in Fig. 20 clearly illustrate that the clean large DCA model is still prone to focusing on pure black DCA regions and light leakage regions to make correct and incorrect classification predictions. These results also indicate that the shift in activations causes the model to focus significantly less on the actual lesion regions. In the examples shown, the areas with the most focus all have light leakage artifacts present to varying degrees and varying levels of visual complexity.

While existing research attempted to remove artifacts and focused on creating improved deep learning models for melanoma classifica-

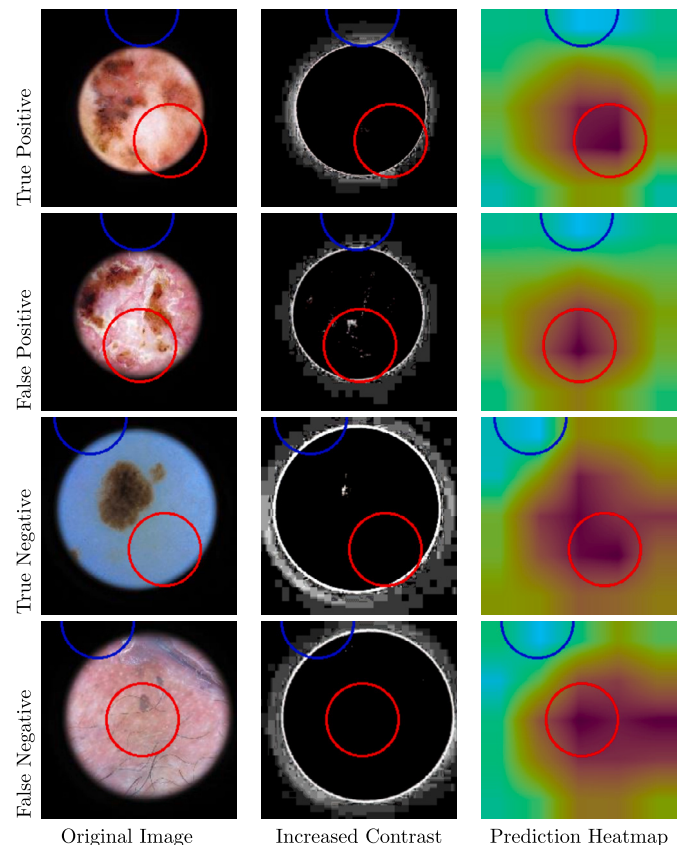


Fig. 20. Illustration of predictions on test images from the clean large DCA model. Images shown are the original unaltered test images (1st column), original images with increased contrast to expose light leakage (2nd column), and the corresponding prediction heatmaps (3rd column). The first row shows true positives, the second row shows true negatives, and the third column shows false negatives. Blue circles indicate the brightest region of the heatmap, red circles indicate the darkest region of the heatmap. Brightest and darkest heatmap regions were obtained using the minMaxLoc function in the OpenCV library [6].

tion, we emphasise on better understanding of the data and behaviour of the learning process, which are the keys to provide new insights into skin lesion analysis. Existing research shows that the limited work focusing on DCA is mostly inconclusive, mainly due to a lack of publicly available datasets with DCA cases with corresponding DCA labels

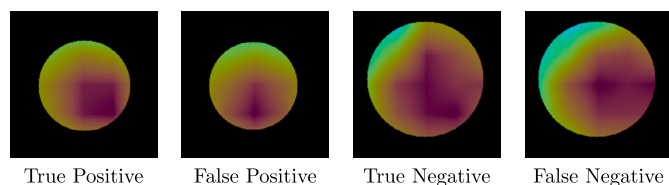


Fig. 21. Illustration of Grad-CAM heatmaps with the original generated masks overlaid to show network activation levels within lesion regions. Heatmaps were taken from inference results for the clean large DCA model.

to support the task. Therefore, we introduce a new curated balanced dataset with an equal number of melanoma and non-melanoma cases, drawn from publicly available skin lesion image datasets, which consists of 6126 training images without DCA and 4124 test images with DCA.

We investigated the effect of DCA in dermoscopic images in melanoma classification by producing a baseline result using the proposed training and test sets. As expected, we achieved high TPR and poor TNR, this is due to the tendency of the model to classify DCA as melanoma. We compared two data augmentation techniques, i.e., inpainted DCA and generated synthetic DCA. We demonstrated that DCA removal and inpainting methods improved the results marginally and proposed a new strategy to address the negative effect of DCAs, i.e., superimposed synthetic DCAs in the training set to train the deep learning model. In addition to existing Binary DCAs, we developed a new synthetic DCA method (namely, Realistic DCA) to improve the realism of the DCA appearance when compared to naturally occurring DCA. We present results from experiments performed on these two artificially generated DCA types and demonstrate their effect in comparison to inpainting of real DCA. Our results indicate that binary DCA provided the highest TPR but realistic DCA provided the highest TNR. Our experiments showed that the removal and inpainting of DCAs is not the sole solution to improve the performance of deep learning models. Instead, our experiments using superimposed synthetic DCAs improved the TNR and precision of melanoma classification. We recommend further investigation to focus on superimposed DCA rather than DCA removal and inpainting as the latter is computationally expensive, achieved marginal improvement, and it is not clear of what new element was introduced in the inpainting process. Moreover, a notable improvement in TNR and precision when using superimposed synthetic DCAs provide some early indication of the capability of such a proposal to reduce the biases of classifying DCA as melanoma.

We interpreted the performance of the deep learning model on different settings by using Grad-CAM heatmap visualisation and its association with the dermoscopy light leakage. We observed that the DCA regions may exhibit artifacts that were not visible to the human eye where the deep learning model might tend to use those features for decision making. Another interesting observation is that although the focus is on the region of interest (skin lesions), there is an apparent randomness in the predictions due to the challenging nature of melanoma classification.

We developed a new quantitative method based on heatmap contrast and brightness intensity measures to increase the understanding of the differences between internal and external DCA regions. Our method for quantifying Grad-CAM heatmap activation images shows a good correlation between heatmap contrast and brightness intensity and recorded metrics such as accuracy and F1-score. This measure can be used in other research domains where heatmap visualisation is used. As external ocular images exist in different imaging for other applications, such as eye imaging and colon imaging, this study potentially can be expanded in other domains. All relevant source code and guidelines to obtain the dataset will be made available upon acceptance of the paper.

CRediT authorship contribution statement

Samuel William Pewton: Conceptualization, Data curation, Formal analysis, Methodology, Software, Writing – original draft, Writing – review & editing. **Bill Cassidy:** Conceptualization, Data curation, Formal analysis, Methodology, Software, Writing – original draft, Writing – review & editing. **Connah Kendrick:** Conceptualization, Data curation, Formal analysis, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Moi Hoon Yap:** Conceptualization, Data curation, Formal analysis, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors have no relevant financial or non-financial interests to disclose. The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

- [1] Q.U. Ain, Genetic programming based feature manipulation for skin cancer image classification, 2020.
- [2] B. Babenko, A. Mitani, I. Traynis, et al., Detection of signs of disease in external photographs of the eyes via deep learning, *Nat. Biomed. Eng.* 6 (12) (2022) 1370–1383.
- [3] C. Barata, M. Ruela, M. Francisco, et al., Two systems for the detection of melanomas in dermoscopy images using texture and color features, *IEEE Syst. J.* 8 (2013), <https://doi.org/10.1109/JSYST.2013.2271540>.
- [4] H. Bhatt, V. Shah, K. Shah, et al., State-of-the-art machine learning techniques for melanoma skin cancer detection and classification: a comprehensive review, *Intell. Med.* 3 (03) (2023) 180–190.
- [5] P. Bibiloni, M. González Hidalgo, S. Massanet, Skin hair removal in dermoscopic images using soft color morphology, in: A. ten Teije, et al. (Eds.), *AIME 2017*, in: *LNAI*, vol. 10259, 2017, pp. 322–326.
- [6] G. Bradski, *The OpenCV library*, Dr. Dobb's J. Softw. Tools (2000).
- [7] T.J. Brinker, A. Hekler, A.H. Enk, et al., A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task, *Eur. J. Cancer* 111 (2019) 148–154, <https://doi.org/10.1016/j.ejca.2019.02.005>.
- [8] T.J. Brinker, A. Hekler, A.H. Enk, et al., Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task, *Eur. J. Cancer* 113 (2019) 47–54, <https://doi.org/10.1016/j.ejca.2019.04.001>.
- [9] J. Buch, S. Criton, et al., Dermoscopy saga—a tale of 5 centuries, *Indian J. Dermatol.* 66 (2) (2021) 174, https://doi.org/10.4103/ijd.IJD_691_18.
- [10] B. Cassidy, C. Kendrick, A. Brodzicki, et al., Analysis of the isic image datasets: usage, benchmarks and recommendations, *Med. Image Anal.* 75 (2022) 102,305, <https://doi.org/10.1016/j.media.2021.102305>.
- [11] N. Codella, V. Rotemberg, P. Tschandl, et al., Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (isic), <https://doi.org/10.48550/arXiv.1902.03368>, 2018.
- [12] N.C. Codella, D. Gutman, M.E. Celebi, et al., Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic), in: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, IEEE, 2018, pp. 168–172.
- [13] M. Combalia, N.C.F. Codella, V. Rotemberg, et al., Bcn20000: dermoscopic lesions in the wild, <https://doi.org/10.48550/arXiv.1908.02288>, 2019.
- [14] Emma Dugas, J.W.C. Jared, Diabetic retinopathy detection, <https://kaggle.com/competitions/diabetic-retinopathy-detection>, 2015.
- [15] A. Esteva, B. Kuprel, R. Novoa, et al., Dermatologist-level classification of skin cancer with deep neural networks, *Nature* 542 (2017), <https://doi.org/10.1038/nature21056>.
- [16] Y. Fujisawa, Y. Otomo, Y. Ogata, et al., Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis, *Br. J. Dermatol.* 180 (2) (2019) 373–381, <https://doi.org/10.1111/bjd.16924>.
- [17] H. Ganster, A. Pinz, R. Röhner, et al., Automated melanoma recognition, *IEEE Trans. Med. Imag.* 20 (2001) 233–239, <https://doi.org/10.1109/42.918473>.
- [18] M. Groh, C. Harris, L. Soenksen, et al., Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset, <https://doi.org/10.48550/arXiv.2104.09957>, 2021.
- [19] D. Gutman, N.C.F. Codella, E. Celebi, et al., Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic), <https://doi.org/10.48550/arXiv.1605.01397>, 2016.

- [20] H. Haenssle, C. Fink, R. Schneiderbauer, et al., Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists, *Ann. Oncol.: Off. J. Eur. Soc. Med. Oncol.* 29 (2018), <https://doi.org/10.1093/annonc/mdy166>.
- [21] S. Han, M. Kim, W. Lim, et al., Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm, *J. Invest. Dermatol.* 138 (2018), <https://doi.org/10.1016/j.jid.2018.01.028>.
- [22] S. Hayes, Dermoscopy: an update and personal view, <https://www.thepmfjournal.com/features/post/dermoscopy-an-update-and-personal-view>. (Accessed 16 August 2022), 2018.
- [23] A. Hekler, J.S. Utikal, A.H. Enk, et al., Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images, *Eur. J. Cancer* 118 (2019) 91–96, <https://doi.org/10.1016/j.ejca.2019.06.012>.
- [24] J. Jaworek-Korjakowska, A. Brodzicki, B. Cassidy, et al., Interpretability of a deep learning based approach for the classification of skin lesions into main anatomic body sites, *Cancers* 13 (23) (2021), <https://doi.org/10.3390/cancers13236048>, <https://www.mdpi.com/2072-6694/13/23/6048>.
- [25] S. Jinnai, N. Yamazaki, Y. Hirano, et al., The development of a skin cancer classification system for pigmented skin lesions using deep learning, *Biomolecules* 10 (8) (2020), <https://doi.org/10.3390/biom10081123>.
- [26] J. Koehoorn, A. Sobiecki, D. Boda, et al., Automated digital hair removal by threshold decomposition and morphological analysis, in: *Lecture Notes in Computer Science*, in: LNIP, vol. 9082, 2015, pp. 15–26.
- [27] F.W. Kong, C. Horsham, A. Ngoo, et al., Review of smartphone mobile applications for skin cancer detection: what are the changes in availability, functionality, and costs to users over time?, *Int. J. Dermatol.* 60 (3) (2021) 289–308.
- [28] I. Kousis, I. Perikos, I. Hatzilygeroudis, et al., Deep learning methods for accurate skin cancer recognition and mobile application, *Electronics* 11 (9) (2022) 1294.
- [29] Y. Kudo Se Mori, M. Misawa, et al., Artificial intelligence and colonoscopy: current status and future perspectives, *Dig. Endosc.* 31 (4) (2019) 363–371.
- [30] F. Lund, A. Clark, Pillow, <https://github.com/python-pillow/Pillow>, 2013.
- [31] A. Mahbod, G. Schaefer, C. Wang, et al., Skin lesion classification using hybrid deep neural networks, in: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 1229–1233.
- [32] A. Mahbod, G. Schaefer, C. Wang, et al., Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification, *Comput. Methods Programs Biomed.* 193 (2020) 105,475, <https://doi.org/10.1016/j.cmpb.2020.105475>.
- [33] S. Menzies, C. Sinz, M. Menzies, et al., Comparison of humans versus mobile phone-powered artificial intelligence for the diagnosis and management of pigmented skin cancer in secondary care: a multicentre, prospective, diagnostic, clinical trial, *Lancet Digit. Health* 5 (2023) e679–e691, [https://doi.org/10.1016/S2589-7500\(23\)00130-9](https://doi.org/10.1016/S2589-7500(23)00130-9).
- [34] M. Nauta, R. Walsh, A. Dubowski, et al., Uncovering and correcting short-cut learning in machine learning models for skin cancer diagnosis, *Diagnostics* 12 (1) (2022), <https://doi.org/10.3390/diagnostics12010040>, <https://www.mdpi.com/2075-4418/12/1/40>.
- [35] D. Okuboyejo, O. Olugbara, Classification of skin lesions using weighted majority voting ensemble deep learning, *Algorithms* 15 (2022) 443, <https://doi.org/10.3390/a15120443>.
- [36] K.B. Ozyoruk, G.I. Gokceler, G. Coskun, et al., Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos: Endo-sfmlearner, *arXiv:2006.16670*, 2020.
- [37] E. Peli, Contrast in complex images, *J. Opt. Soc. Amer. A* 7 (10) (1990) 2032–2040, <https://doi.org/10.1364/JOSAA.7.002032>, <http://opg.optica.org/josaa/abstract.cfm?URI=josaa-7-10-2032>.
- [38] S.W. Pewton, M.H. Yap, Dark corner on skin lesion image dataset: does it matter?, in: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 4830–4838.
- [39] T.C. Pham, V.D. Hoang, C.T. Tran, et al., Improving binary skin cancer classification based on best model selection method combined with optimizing full connected layers of deep cnn, in: *2020 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, 2020, pp. 1–6.
- [40] G. Ramella, Automatic skin lesion segmentation based on saliency and color, in: *15th International Conference on Computer Vision Theory and Applications*, 2020, pp. 452–459.
- [41] G. Ramella, Hair removal combining saliency, shape and color, *Appl. Sci.* 11 (1) (2021), <https://doi.org/10.3390/app11010447>, <https://www.mdpi.com/2076-3417/11/1/447>.
- [42] G. Ramella, Saliency-based segmentation of dermoscopic images using colour information, *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* 10 (2) (2022) 172–186, <https://doi.org/10.1080/21681163.2021.2003248>.
- [43] L.C. Rose, Recognizing neoplastic skin lesions: a photo guide, *Amer. Fam. Phys.* 58 (4) (1998) 873–884, <https://www.proquest.com/scholarly-journals/recognizing-neoplastic-skin-lesions-photo-guide/docview/234312593/se-2>.
- [44] A. Rosebrock, Grad-cam: visualize class activation maps with keras, tensorflow, and deep learning, <https://pyimagesearch.com/2020/03/09/grad-cam-visualize-class-activation-maps-with-keras-tensorflow-and-deep-learning/>. (Accessed 3 October 2022), 2020.
- [45] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, et al., A patient-centric dataset of images and metadata for identifying melanomas using clinical context, *Sci. Data* 8 (2021) 34, <https://doi.org/10.1038/s41597-021-00815-z>.
- [46] R.R. Selvaraju, M. Cogswell, A. Das, et al., Grad-CAM: visual explanations from deep networks via gradient-based localization, *Int. J. Comput. Vis.* 128 (2) (2019) 336–359, <https://doi.org/10.1007/s11263-019-01228-7>.
- [47] A. Shah, M. Shah, A. Pandya, et al., A comprehensive study on skin cancer detection using artificial neural network (ANN) and convolutional neural network (CNN), *Clin. EHealth* (2023).
- [48] J.A. Sheindlin, T. Hirose, M.E. Hartnett, Ophthalmic endoscopy: applications in intraocular surgery, *Int. Ophthalmol. Clin.* 39 (1) (1999) 237–247.
- [49] K. Sies, J.K. Winkler, C. Fink, et al., Dark corner artefact and diagnostic performance of a market-approved neural network for skin cancer classification, *J. Dtsch. Dermatol. Ges.* 19 (6) (2021) 842–850, <https://doi.org/10.1111/ddg.14384>, <https://onlinelibrary.wiley.com/doi/abs/10.1111/ddg.14384>.
- [50] A. Sultana, I. Dumitrache, M. Vocurek, et al., Removal of artifacts from dermoscopic images, in: *2014 10th International Conference on Communications (COMM)*, IEEE, 2014, pp. 1–4.
- [51] C. Szegedy, S. Ioffe, V. Vanhoucke, et al., Inception-v4, inception-resnet and the impact of residual connections on learning, in: *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [52] P. Tschandl, The HAM10000 dataset, a large collection of multi-source dermoscopic images of common pigmented skin lesions, <https://doi.org/10.7910/DVN/DBW86T>, 2018.
- [53] P. Tschandl, N. Codella, B.N. Akay, et al., Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study, *Lancet Oncol.* 20 (7) (2019) 938–947, [https://doi.org/10.1016/S1470-2045\(19\)30333-X](https://doi.org/10.1016/S1470-2045(19)30333-X).
- [54] H. Zand, N. Nguyen, B. Zeinali, et al., A new preprocessing approach to improve the performance of cnn-based skin lesion classification, *Med. Biol. Eng. Comput.* 59 (2021) 1–9, <https://doi.org/10.1007/s11517-021-02355-5>.
- [55] H. Zhou, M. Chen, R. Gass, et al., Feature-preserving artifact removal from dermoscopy images, *Proc. SPIE Int. Soc. Opt. Eng.* 6914 (2008), <https://doi.org/10.1117/12.770824>.
- [56] H.M. Ünver, E. Ayan, Skin lesion segmentation in dermoscopic images with combination of yolo and grabcut algorithm, *Diagnostics* 9 (3) (2019), <https://doi.org/10.3390/diagnostics9030072>, <https://www.mdpi.com/2075-4418/9/3/72>.