

Supplementary data

Clinically-relevant vulnerabilities of deep machine learning systems for skin cancer diagnosis

Authors:

Dr Xinyi Du-Harpur^{1,2,5} MA MBBChir MRCP*, Callum Arthurs BSc^{1*}, Dr Clarisse Ganier PhD¹, Dr Rick Woolf MBChB PhD MRCP(Derm)⁵, Dr Zainab Laftah MBChB MRCP(Derm)⁵, Dr Manpreet Lakhan MBChB MRCP⁵, Dr Amr Salam BSc MBChB MRCP⁵, Dr Bo Wan MD¹, Prof Fiona M. Watt PhD^{1,2}, Prof Nicholas M. Luscombe PhD^{2,3,4}, Dr Magnus D. Lynch MA DPhil MRCS MRCP(Derm)^{1,5}

*co-first authors

Affiliations:

1. Centre for Stem Cells and Regenerative Medicine, King's College London, Great Maze Pond, London SE1 9RT
2. The Francis Crick Institute, 1 Midland Road, NW1 1AT
3. Okinawa Institute of Science & Technology Graduate University, Okinawa, 904-0495, Japan
4. UCL Genetics Institute, University College London, Gower Street, London, WC1E 6BT, UK.
5. St John's Institute of Dermatology, Guys Hospital, Great Maze Pond, London SE1 9RT

Work performed in London, United Kingdom

Corresponding author: Dr Xinyi Du-Harpur

Address: Dr Xinyi Du-Harpur, 28th Floor Tower Wing, Guy's Hospital, London SE1 9RT

Email: xinyi.du@kcl.ac.uk









Tel: +447505125968









Keywords:

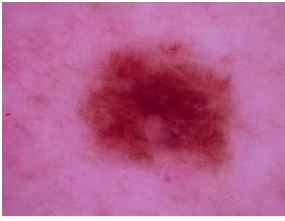



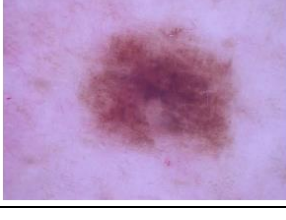
Bioinformatics, Methods/Tools/Techniques, Imaging, Melanoma

Abbreviations:

CNN	Convolutional neural network
FGSM	Fast gradient sign method
RGB	Red green blue

Image Perturbation	Correct classifications	False Negatives	False Positives	Accuracy	Sensitivity	Specificity	AUC	Example Image
Benchmark	4300	68	128	0.956	83.63	98.22	0.9876	
Rotate 180°	4276	76	144	0.951	81.33	97.95	0.9850	
Rotate 45°	4242	76	178	0.944	78.01	97.93	0.9844	
R channel (-5)	4276	84	136	0.951	81.46	97.85	0.9844	
G channel (-5)	4239	128	129	0.943	80.56	97.04	0.9790	
B channel (-5)	4247	79	140	0.951	81.07	97.23	0.9796	
R channel (+5)	4252	103	141	0.946	80.56	97.44	0.9817	
G channel (+5)	4261	88	147	0.948	80.43	97.71	0.9838	

B channel (+5)	4277	79	140	0.951	81.59	97.93	0.9855	
R channel (-10)	4270	91	135	0.95	80.82	97.68	0.9830	
G channel (-10)	4199	160	137	0.934	78.26	96.20	0.9733	
B channel (-10)	4206	179	111	0.935	80.95	96.07	0.9738	
R channel (+10)	4220	136	140	0.939	79.03	96.88	0.9761	
G channel (+10)	4255	92	149	0.946	79.41	97.63	0.9820	
B channel (+10)	4265	90	141	0.949	81.33	97.74	0.9851	
R channel (-50)	4158	116	222	0.925	71.23	97.04	0.9649	

G channel (-50)	4128	57	311	0.918	63.94	98.06	0.9677	
B channel (-50)	4086	248	162	0.909	72.12	95.34	0.9531	
R channel (+50)	3209	1150	137	0.714	0.000	100.0	0.8132	
G channel (+50)	3991	330	175	0.888	64.71	93.46	0.9269	
B channel (+50)	3983	234	279	0.886	58.44	94.78	0.9370	

Supplementary Figure 1: Results of systematic perturbation of image color balance and rotation on accuracy of Inception v3 classifier

The CNN was tested on a set of images that had not been used in network training (n=4496). The entire dataset was then perturbed, with either a 180° rotation, 45° rotation or intensity subtraction (-5,-10,-50) or addition (+5,+10,+50) for each channel of the RGB image. The number of correctly classified images, false negatives (melanoma images classified as naevus), false positives (naevus images classified as melanoma), and total accuracy of the classification (rounded to 3 decimal places), was recorded for each set of edited images.

a

Dear participant,

Many thanks for agreeing to participate in our study looking at dermatologists' diagnostic skills with regards to **dermoscopic** melanocytic lesions.

You will be provided with an application, which will take you through 200 **dermoscopic** images and you will be asked to provide your decision regarding whether you consider the image to **more likely** represent a **naevus** or a melanoma (**melanoma** in situ, or invasive melanoma).

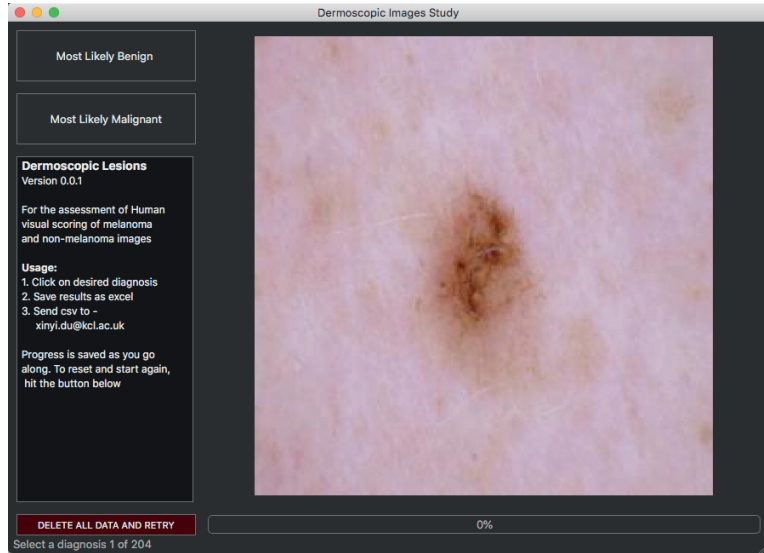
Once you have completed your assessment, an excel document will be automatically generated. Please email this, along with your post-assessment questionnaire (attached), to me (xinyi.du@kcl.ac.uk) for analysis.

Your results will be **anonymous** for analysis and publication purposes.

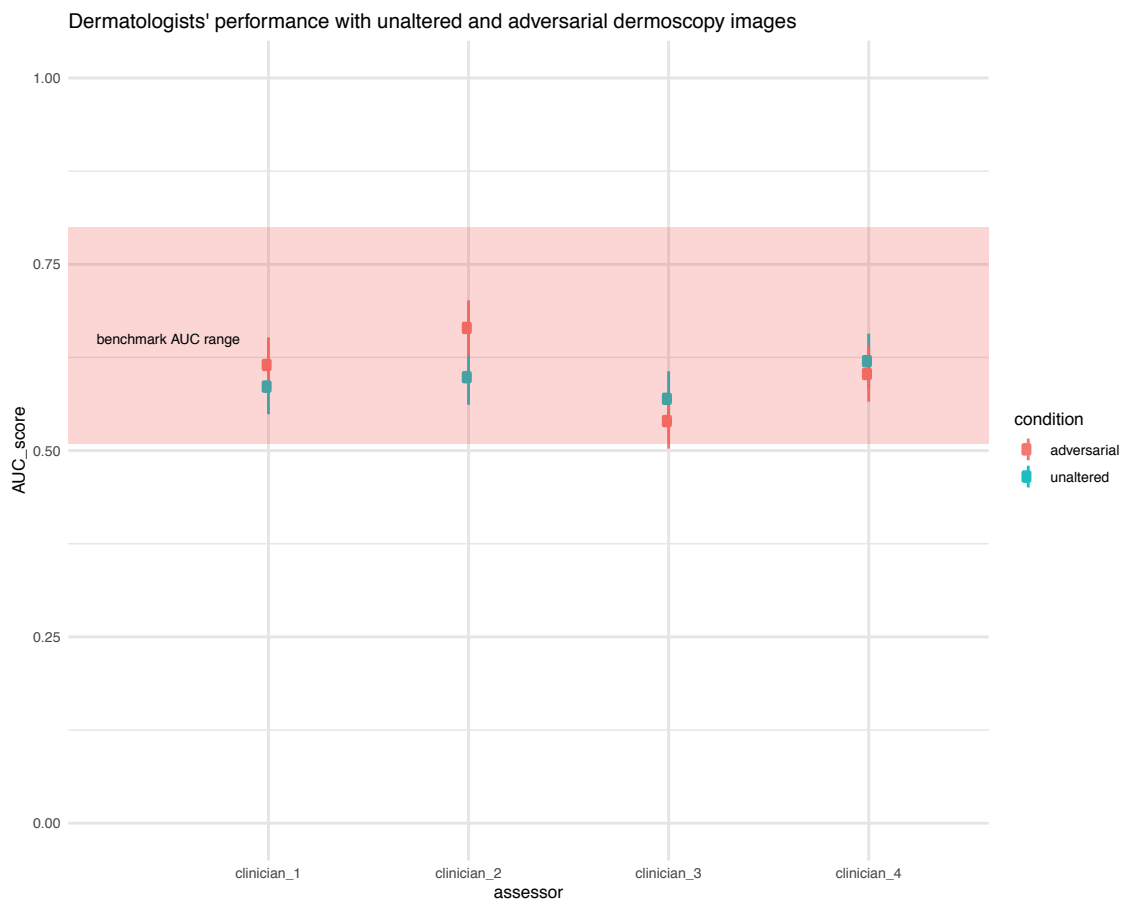
Best wishes,

Magnus Lynch
Principal Investigator

b



c



Supplementary figure 2:

a) Participant invitation letter for 'dermoscopy study'. Participants were informed only that this was a dermoscopy study and were unaware that a comparison was to be made between adversarial and unmodified images. All participants had greater than 4 years of dermatology experience, and used dermoscopy regularly in their clinical practice. b) Custom-designed application used to show dermoscopy images from ISIC 2018 dataset to study participants. On completion of the study, a .csv file of their decisions was automatically generated. c) AUC scores of participants, with 95% confidence intervals of the score added. The red shaded area represents the AUC range with 95% confidence interval from Brinker's 'dermatologist benchmark' study for comparison. There was no significant difference in the performance of Dermatologists with adversarial versus unmodified images ($p=0.337$; calculated by logistic regression with (i) identity of clinician (ii) type of image i.e. melanoma versus naevus and (iii) adversarial or unmodified as predictor variables)

Supplementary Table 1: Performance metrics of deep learning models

Model	AUC	Sensitivity	Specificity
Lynch Lab replicate 1	0.988	0.836	0.982
Lynch Lab replicate 2	0.957	0.703	0.965
DaisyLab	0.928	0.710	0.962
DysionAI	0.808	0.675	0.941
Almage Lab	0.933	0.684	0.956

Performance metrics of our deep learning model with two different splits of train/test data in comparison with other models in correctly evaluating melanoma, consisting of AUC (Area Under the Receiver Operating Curve), sensitivity (true positive rate) and specificity (true negative rate) at a decision boundary of 0.5, as compared to performance in diagnosing melanoma for the top three models in 2019 ISIC Challenge with equivalent threshold.

Supplementary Table 2: Results of color balance and rotation/translation adversarial attacks

Model	Attack Name	Number of test images	Number of successful attacks	Percentage successful attacks
Lynch lab (replicate 1)	Colour	782	80	10.23%
Lynch lab (replicate 1)	Rotation	782	357	45.65%
Lynch lab (replicate 2)	Colour	782	73	9.34%
Lynch lab (replicate 2)	Rotation	782	349	44.63%
Han <i>et al</i>	Colour	59	10	16.95%
Han <i>et al</i>	Rotation	59	22	37.29%

Efficacy of color balance and rotation/translation attacks. Results for retrained inception v3 (Lynch Lab replicate 1 and Lynch Lab replicate 2) are presented after training the network with random rotation and random variation in color balance ‘color jitter’. The percentage of correct classifications for the Model of Han *et al* on the malignant melanoma test images released along with their publication was 89.8% (n=59). Images in this image set were perturbed using the colour and rotation adversarial attacks. An attack was deemed successful if the model initially classified the test image as melanoma (the correct diagnosis) on the unperturbed image but classified the adversarial image incorrectly (one of the other 11 classes). The number of successful attacks for each attack is recorded in the table.

Online Methods

Development of the melanoma deep learning classifier

In order to explore the failure modes of convolutional neural networks (CNN) in the classification of skin cancer, we first trained a deep learning classifier on example images of melanoma and benign melanocytic lesions. As in previous work (Esteva et al. 2017) we train a CNN model, m , by fine-tuning a model previously trained on the ImageNet dataset. The network was trained against a dataset of 23,010 images of melanocytic lesions obtained from the publicly available ISIC challenge dataset. Training images were randomized and split with 80% used for training and 20% reserved for validation. We trained an Inception v3 network in PyTorch using the stochastic gradient descent method. To increase the diversity of training images, input images were subjected to random rotation, crop and horizontal and vertical flip. To assess whether random variation in color balance during the training process would protect against the color balance attacks, a second model was trained with these same modifications but with the addition of random variation, 'jitter', in the color balance of each image during each iteration of the training. A cohort of validation images were held back from the training set and not used to calculate gradients or update model weights but instead used to assess the accuracy of the classifier.

The final layer of the CNN comprising a vector \mathbf{z} containing K neurons was subject to the softmax function:

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \text{ for } i = 1, \dots, K$$

Model weights were updated by backpropagation using a learning rate of 0.001 and momentum of 0.9. The accuracy of the model is assessed on the output of the model for the validation set, following softmax transformation of the output of the network and using 0.5 as a decision boundary. Classification accuracy of the trained model on the validation dataset was 95.6%, sensitivity was 0.84, specificity was 0.98 and AUC was 0.99; this compares favourably with the performance of top-performing models from the 2019 ISIC challenge in distinguishing melanoma from benign melanocytic naevi at a decision boundary of 0.5 (Supplementary Table 1), however it is noted that these ISIC models were trained to differentiate multiple classes of skin lesions rather than simply to differentiate benign and malignant melanocytic lesions which may partially account for their lower performance. We also trained our model with a different random split of

train and test images with similar results (Supplementary Table 1). Confidence of the model in favour of a specific class is defined as the magnitude of that value following softmax transformation.

Fine-tuning of the parameters of the CNN uses established methods and is in keeping with other published papers (e.g. Brinker, 2019 and Esteva, 2017). The CNN is trained by gradient descent with cross entropy function as a loss function on the training set of images. Model parameters are updated by back propagation according to the gradient of the loss function.

Construction of adversarial attacks

For a pre-trained CNN model m and a sample x with class label y the concept of an adversarial attack is to modify x to create an adversarial image x_{adv} such that the classification error of m is maximized with the minimal changes to x as perceived by the human visual system. Adversarial attacks can be targeted such that the maximization function seeks to force the model to a specific class or untargeted such that misclassification to any other class is permitted. In the case of distinguishing benign from malignant melanocytic lesions there is no distinction between a targeted and untargeted attack since there are only two classes. Adversarial attacks can broadly be subdivided into white-box attacks which depend on knowledge of the internal gradients within the network and black-box attack which measure only the output of the classifier with no knowledge of the internal state. We ran adversarial attacks on all images in the validation set and recorded whether an adversarial image could be generated.

FGSM attack The fast gradient sign method (Goodfellow et al. 2014) perturbs each pixel in the input image for a single step in the direction that maximizes the probability of an incorrect class to generate an adversarial example, x_{adv} :

$$x_{adv} = \epsilon \text{sign}(\nabla_x L(\theta, x, y))$$

Where L represents the loss function, x the input image, θ parameters of the model and y the targets associated with x .

n-pixel attack An input image x is represented by a two dimensional array with each position in the array comprising a three dimensional vector corresponding to red, blue and green color channels. The optimization problem therefore is to identify firstly a small number, d of (x, y) pairs corresponding to pixels in the input image and secondly d vectors corresponding to (r, g, b) perturbations of the input pixel red, green and blue channels. Where m is the trained classifier function and e is an additive adversarial perturbation we seek to maximize the following function:

$$\begin{aligned} & \underset{e(x)^*}{\text{maximize}} && m_{\text{adv}}(x + e(x)) \\ & \text{subject to} && \|e(x)\|_0 \leq d \end{aligned}$$

Color balance attack For the color balance attack, rather than modifying a small number of pixels, for all pixels in the input image, we multiply each red, green or blue channel by a small fixed value. Where \mathbf{c} is the multiplicative color modification vector and δ is a small perturbation (we use 0.1), we maximize the following:

$$\begin{aligned} & \underset{\mathbf{c}^*}{\text{maximize}} && m_{\text{adv}}(\max(\mathbf{c}x, 255)) \\ & \text{subject to} && 1 - \delta \leq \mathbf{c} \leq 1 + \delta \end{aligned}$$

Rotation / translation attack For the rotation / translation attack we subject the input image x to a combined rotation and translation. Rotation can occur in 360 degrees, however translation is limited to a fraction δ (we use 50 pixels) of the input image size (299,299 pixels). Where r is the rotation function and t is the translation function. t and r are parameterized by a vector \mathbf{a} that specifies the size of the rotation \mathbf{a}_r and the size of the translocation in both horizontal and vertical orientations, \mathbf{a}_t . We maximize the following:

$$\begin{aligned} & \underset{\mathbf{a}^*}{\text{maximize}} && m_{\text{adv}}(t(r(x, \mathbf{a}_r), \mathbf{a}_t)) \\ & \text{subject to} && 0 \leq \mathbf{a}_r \leq 360 \\ & && -\delta \leq \mathbf{a}_t \leq \delta \end{aligned}$$

Differential evolution Differential evolution is a population based optimization strategy that belongs to the class of evolutionary algorithms (Storn and Price 1997). It does not require that the objective function be differentiable permitting application to a wider range of optimization problems. It can be distinguished from other evolutionary algorithms since each member of the

population is represented by a vector of real numbers. It does not require knowledge of the internal state allowing it to be applied to black-box adversarial attacks. In each iteration it generates a population of similar input images with subtle random variations in the parameters that are to be optimized. By chance a subset of the images in this population may be classified less accurately than others. During each iteration, a set of candidate solutions is generated according to the parent population. Children are compared with their parents and retained if they are fitter according to the objective function. The differential evolution algorithm runs iteratively and, in order to allow the algorithm to complete in a reasonable time frame, terminates when either (i) the confidence of the network in favour of the incorrect diagnosis (i.e. benign naevus) exceeds 90%; or, where no example can be identified, the algorithm terminates after an arbitrary number (20) iterations - whichever happens first. Increasing the number of iterations beyond this has little impact on the conclusions since if an adversarial attack is to be found, it is usually in the early iterations. The differential evolution algorithm as implemented in the Python `scipy` (Jones and Oliphant 2014) package was employed for the pixel, color balance and rotation / translation attacks.

Testing of the model of Han *et al* with our adversarial attacks

Han et al (Han et al. 2018) have previously reported a CNN based on the ResNet-152 architecture that is able to distinguish 12 benign and malignant skin lesions (including melanoma and benign melanocytic naevi). Code, trained models and test images are available online: https://figshare.com/articles/Caffe_model_files_and_Python_Examples/5406223. In contrast to our model, which is implemented in PyTorch, the model of Han et al is implemented in the Caffe deep learning framework. However since our differential evolution-based approach does not require access to the internal state (gradients) of the network we were able to execute our adversarial attacks on this pre-trained model. We used the validation images of melanoma released by Han et al and a successful adversarial attack was achieved when the confidence of the model in favour of a benign melanocytic naevus exceeded that in favour of melanoma.

Systematic perturbation of images for testing of CNN

To understand whether naive (undirected) perturbation of the color balance and rotation of images had an impact upon the accuracy of the model. We perturbed the validation image set (not used

for training the network; n=4496) with either a 180° rotation, 45° rotation or intensity subtraction (-5,-10,-50) and addition (+5,+10,+50) for each channel of the RGB image, individually and evaluated the accuracy of the trained model (Supplementary Figure 1).

Testing of Dermatologists with adversarial images

To assess whether the accuracy of Dermatologists was influenced by adversarial images we developed a stand alone downloadable application that we used to compare the performance of Dermatologists on unmodified and adversarial images (modified color balance, 3 pixel, or rotation/translation, n=34 of each). 4 Dermatologists (2 consultants and 2 experienced residents) were presented with 204 images of which half were unmodified and half were adversarial images that defeated the CNN. The performance of these dermatologists with modified and adversarial images was evaluated by logistic regression (performed in R) with individual, type of perturbation and original/perturbed image as the predictor variables and accuracy of classification as the dependent variable.

Training of models and code availability

Deep learning models and adversarial attacks were implemented in Pytorch and were trained on an NVidia 2080ti GPU. Training images were obtained from publically available ISIC challenge dataset. Code for training of models, generation of adversarial challenges and testing of Dermatologists is available at the following URL: <https://github.com/thelynchlab/adversarial>.