Skin Deep Unlearning: Artefact and Instrument Debiasing in the Context of Melanoma Classification

Peter Bevan[1] and Amir Atapour-Abarghouei[2]

[1]School of Computing, Newcastle University, Newcastle, UK

[2]Department of Computer Science, Durham University, Durham, UK

Newcastle University

arXiv

Durham University

Overview: Contributions

- **Artefact debiasing**: We mitigate the bias introduced by surgical markings and rulers when classifying skin lesion images.

- **Instrument debiasing for generalisation**: We demonstrate the generalisation benefits of unlearning information relating to the instruments used to capture skin lesion images.

Motivation: Surgical Marking Bias

JAMA Dermatology | Original Investigation

Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition

Julia K. Winkler, MD; Christine Fink, MD; Ferdinand Toberer, MD; Alexander Enk, MD; Teresa Deinlein, MD; Rainer Hofmann-Wellenhof, MD; Luc Thomas, MD; Aimilios Lallas, MD; Andreas Blum, MD; Wilhelm Stolz, MD; Holger A. Haenssle, MD

CONCLUSIONS AND RELEVANCE This study's findings suggest that skin markings significantly interfered with the CNN's correct diagnosis of nevi by increasing the melanoma probability scores and consequently the false-positive rate. A predominance of skin markings in melanoma training images may have induced the CNN's association of markings with a melanoma diagnosis. Accordingly, these findings suggest that skin markings should be avoided in dermoscopic images intended for analysis by a CNN.

- In this study, the CNN scored an AUC of **0.969** on images without surgical markings.
- When tested on the same lesions with surgical markings present, the CNN scored an AUC of **0.922**

- We also recreated this performance drop in our experiments, see table below:

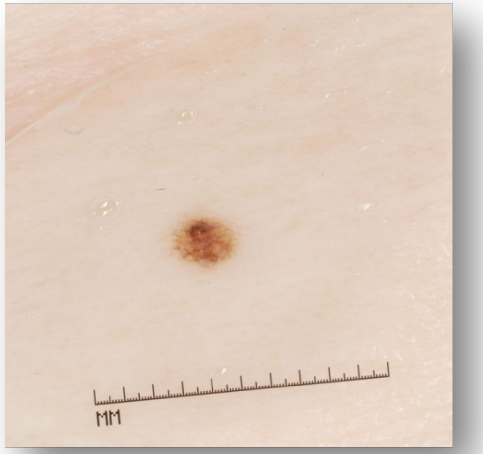
Plain images	Images w/ markings
0.990	**0.902**

Motivation: Ruler Bias

Association between different scale bars in dermoscopic images and diagnostic performance of a market-approved deep learning convolutional neural network for melanoma recognition

Julia K. Winkler [a], Katharina Sies [a], Christine Fink [a], Ferdinand Toberer [a], Alexander Enk [a], Mohamed S. Abassi [b], Tobias Fuchs [b], Holger A. Haenssle [a,*]

Conclusions: Superimposed scale bars in dermoscopic images may impair the CNN's diagnostic accuracy, mostly by increasing the rate of the false-positive diagnoses. We recommend avoiding scale bars in images intended for CNN analysis unless specific measures counteracting effects are implemented.



- In this study, the CNN scored an AUC of **0.953** on images without rulers.
- When tested on the same lesions with rulers present (3Gen-Dermlite I), the CNN scored an AUC of **0.774**.

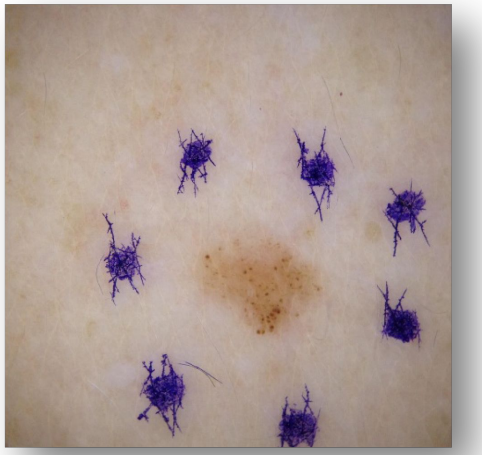
We also recreated this performance drop in our experiments, see table below:

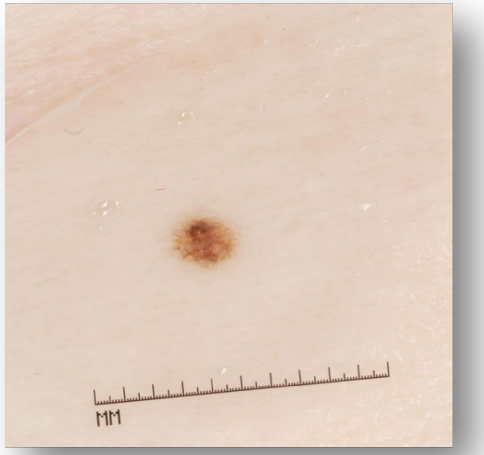
Plain images	Images w/ rulers
0.999	**0.831**

Motivation: Artefact Bias

- Surgical markings and rulers introduce bias that causes performance irregularities in melanoma classification models [1,2].
- Current suggestion is that dermatologists stop using these visual aids, but this is not realistic.
- Cropping and segmentation are expensive and ineffective.

We investigate an automated solution to mitigating these biases using leading debiasing techniques

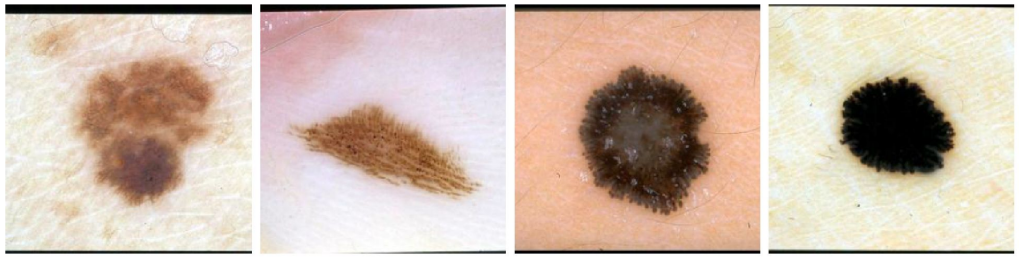




Motivation: Instrument Bias

Domain bias is caused by differences in the instrument type (dermoscopic/clinical) or instrument model used to capture lesion images.

Dermoscopic



Clinical

Atlas	
Dermoscopic	Clinical
0.819	**0.616**

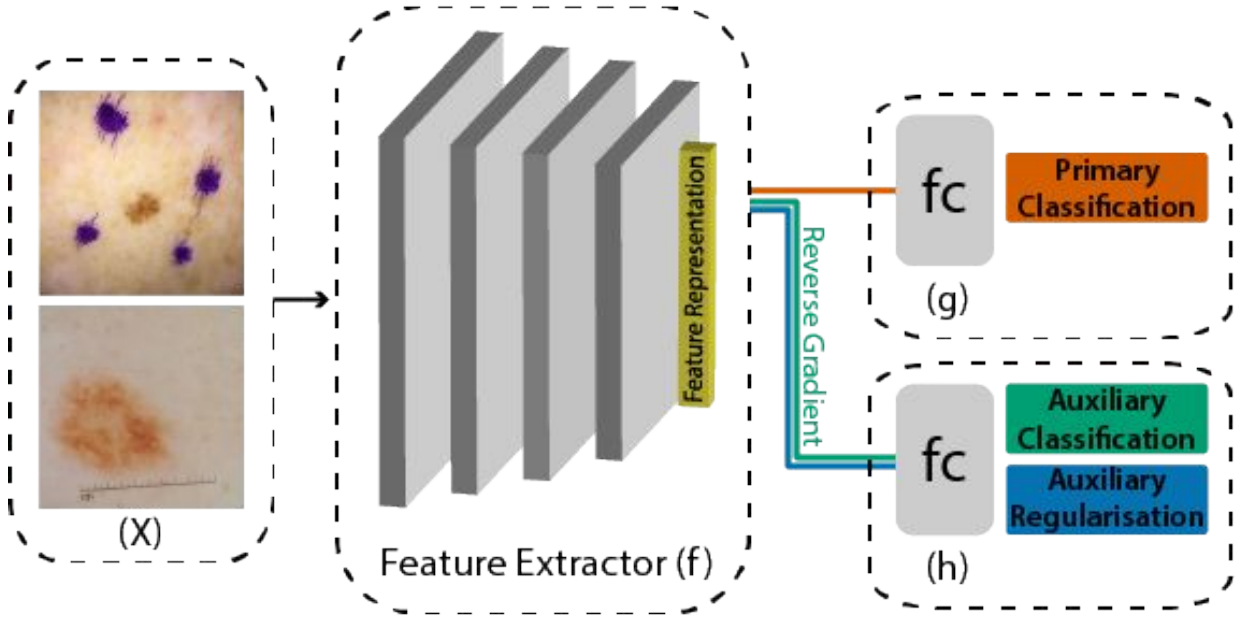
The Atlas dataset has 1000 pairs of clinical and dermoscopic images of the same lesion. Since the model is trained on dermoscopic data it doesn't generalise well to the clinical images. Similar results have been shown in previous studies [5].

We investigate removing this domain bias towards improved generalisation, using leading debiasing techniques.

Methods: Learning Not To Learn (LNTL) [3]

Auxiliary classifier head to identify and remove a labelled bias:

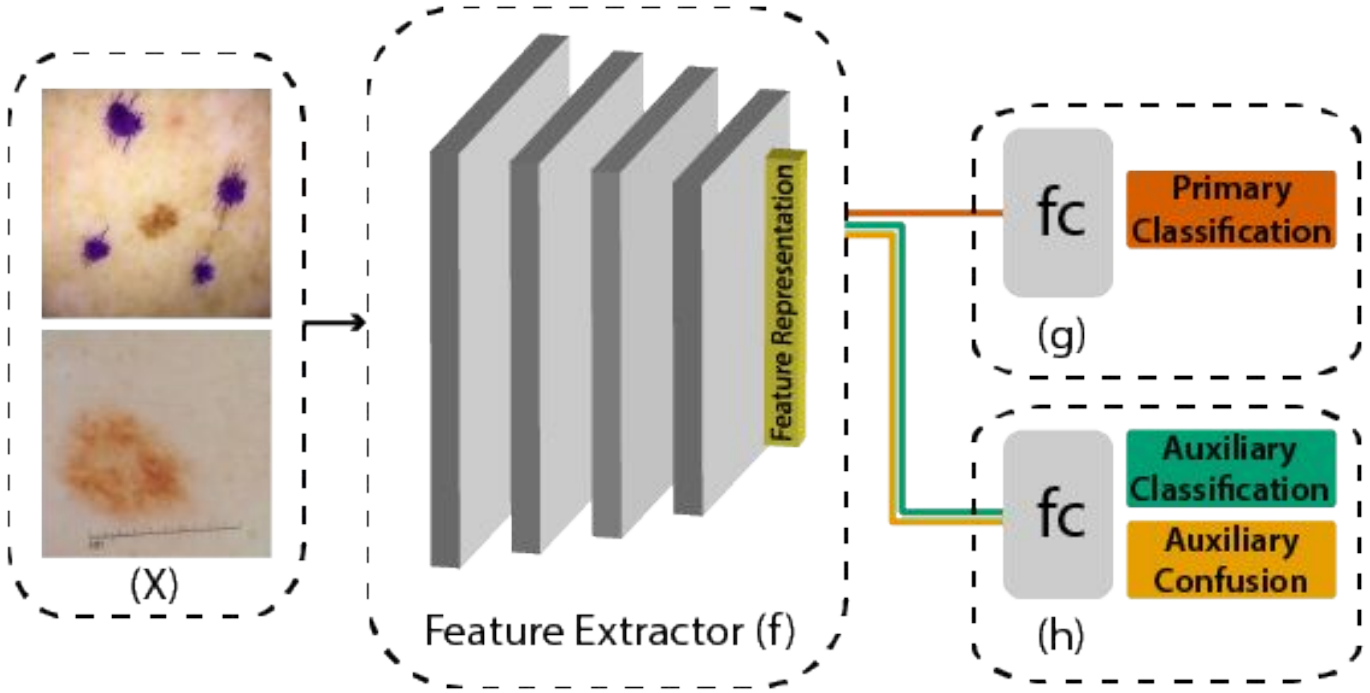
- ❑ Auxiliary regularisation loss minimises mutual information between the feature embedding and the targeted bias.
- ❑ Gradient reversal applied to auxiliary classification loss during backpropagation as additional bias removal tool.
- ❑ Goal is that the primary classification head learns to classify using a feature embedding that is independent of the target bias.



Methods: Turning A Blind Eye (TABE) [4]

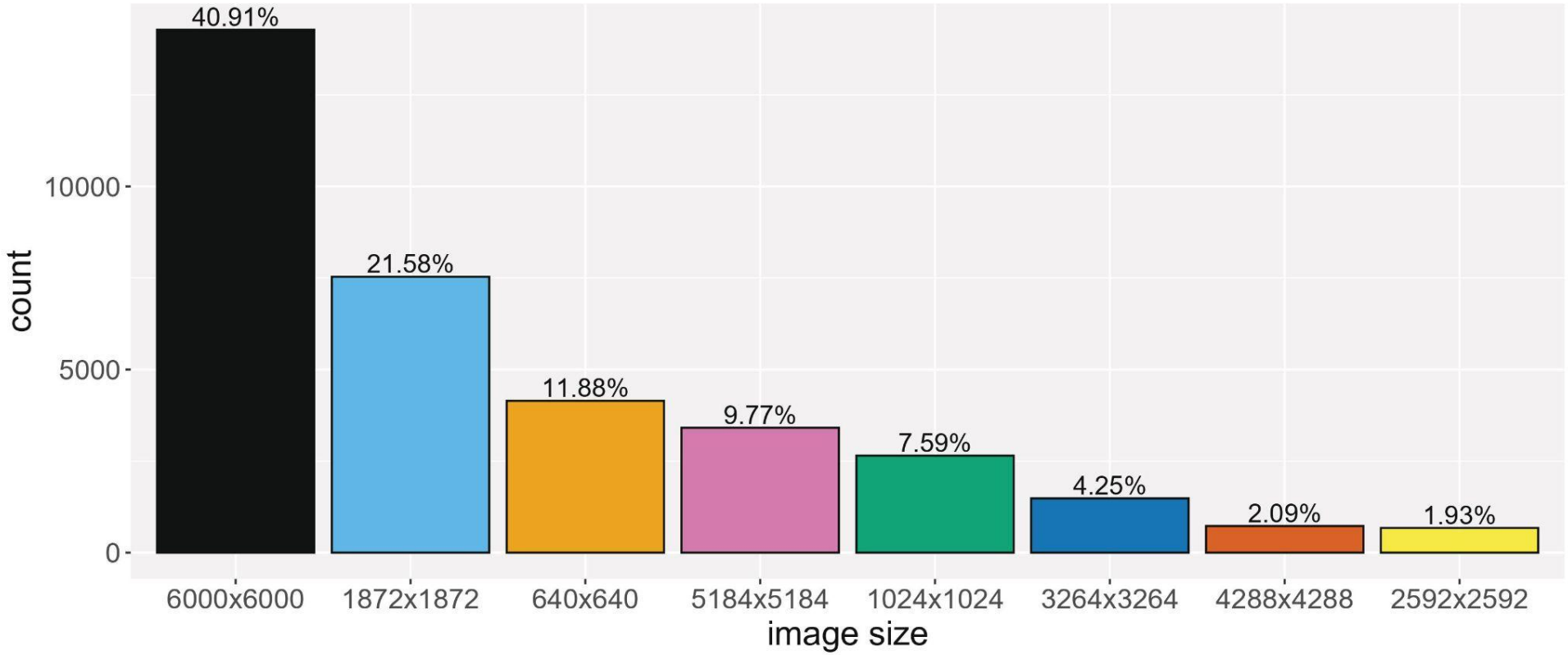
Auxiliary classifier head to identify and remove a labelled bias:

- ❑ Auxiliary confusion loss finds cross entropy between output predicted bias and uniform distribution towards finding a bias invariant feature representation.
- ❑ Gradient reversal can also be applied to the auxiliary classification loss in TABE for additional bias removal. We refer to this configuration as **CLGR**.



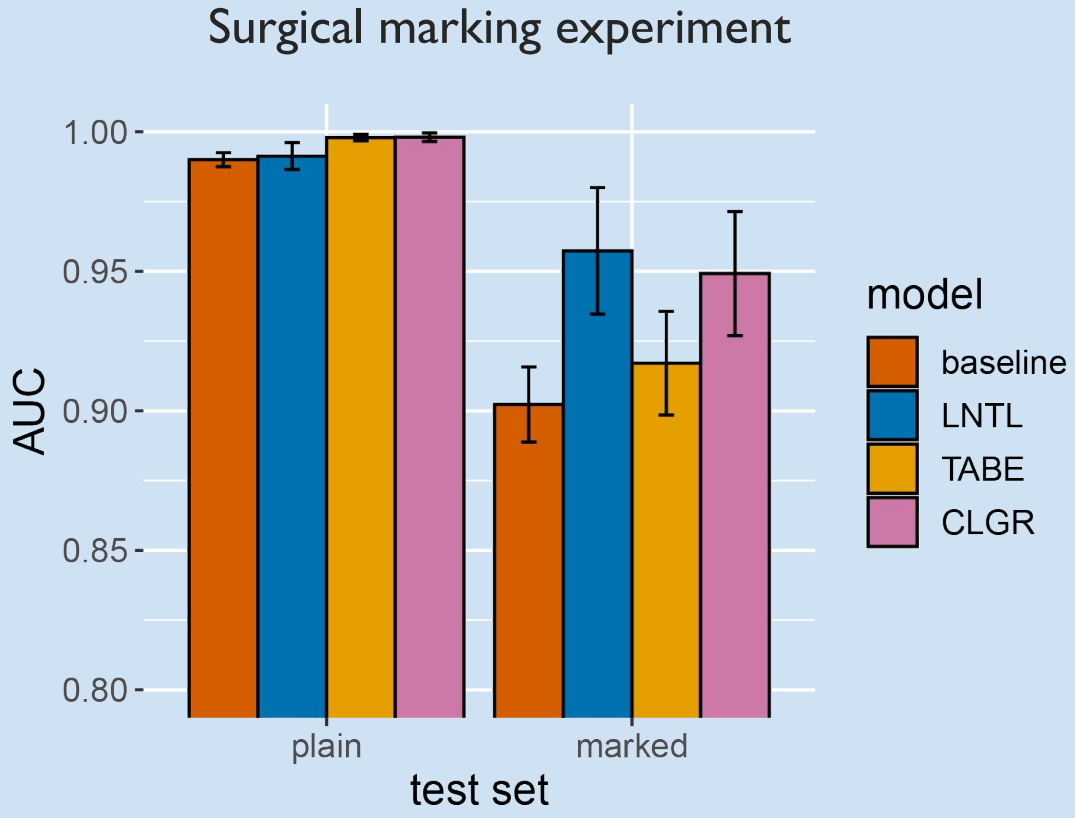
Methods: Instrument Bias Labels

Since we don't have labels for the instrument used to capture the images in the training set (ISIC competition data), we use the image size as a proxy.



Experimental Results: Artefact Bias Removal

- Models are tested on the same lesions with and without artefacts present.
- ***Debiasing methods seem to help mitigate both surgical marking and ruler bias.***



Experimental Results: Instrument Bias Removal

Using Turning a Blind Eye [4] to unlearn instrument information leads to improved generalisation, with improved performance (compared to the baseline) across several dermoscopic and clinical test sets. Scores are AUC.

Experiment	AtlasD	AtlasC	ASANC	MClassD	MClassC
Dermatologists	---	---	---	0.671	0.769
Baseline	**0.819**	0.616	0.768	0.853	0.744
LNTL	0.776	0.597	0.746	0.821	0.778
TABE	0.817	**0.674**	**0.857**	**0.908**	0.768
CLGR	0.784	0.650	0.785	0.818	**0.807**

References

[1] Winkler et al., ‘Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition’
[2] Winkler et al., ‘Association between different scale bars in dermoscopic images and diagnostic performance of a market-approved deep learning convolutional neural network for melanoma recognition’
[3] Kim et al., ‘Learning Not to Learn: Training Deep Neural Networks With Biased Data’
[4] Alvi et al., ‘Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings’
[5] Gu et al., ‘Progressive Transfer Learning and Adversarial Domain Adaptation for Cross-Domain Skin Disease Classification’