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

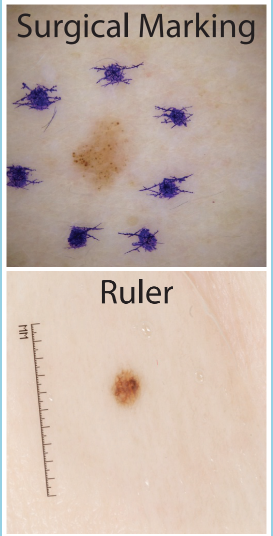
Peter Bevan and Amir Atapour-Abarghouei
School of Computing, Newcastle University, Newcastle, UK

https://github.com/pbevan1/Melanoma-Bias-Unlearning

Motivation

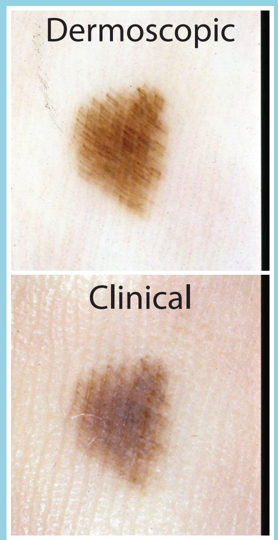
Artefact Bias

- Surgical markings and rulers introduce bias that causes performance irregularities in melanoma classification models [1,2].
- This could be solved if dermatologists stop using these visual aids, but this is not realistic.
- Cropping and segmentation are expensive and innefective.
- ***We investigate an automated solution to mitigating these biases using LNTL [4] and TABE [5] (right).***



Instrument Bias

- Domain bias is caused by differences in the instrument type (dermoscopic/clinical) or instrument model used to capture lesion images.
- This could lead to irregular performance accross different clinics.
- ***We investigate removing this domain bias towards domain generalisation, by using LNTL [4] and TABE [5] (right) with instrument labels as the target bias.***



Skin Tone Bias

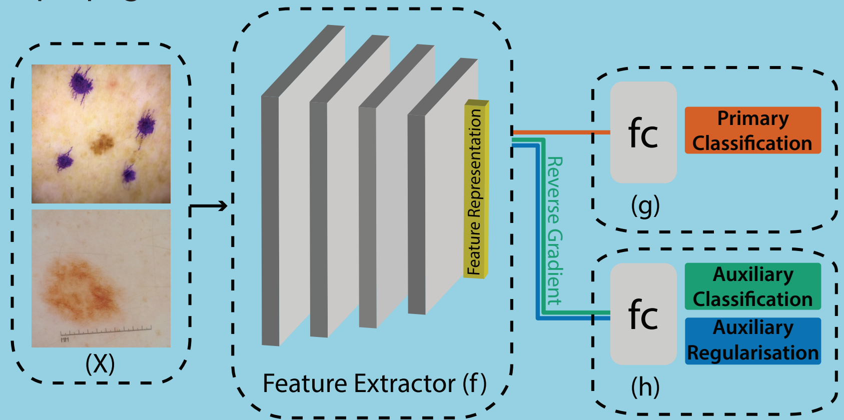
- CNNs perform best on skin tones similar to the model training data: a model trained on Fitzpatrick skin types 1&2 performed better on types 3&4 than types 5&6 [3].
- ***We investigate if using LNTL [4] and TABE [5] (right) to remove skin tone information from the feature representation can mitigate this performance disparity and improve generalisation to datasets with different distributions of skin types.***

Methods

Learning Not to Learn (LNTL) [4]

Auxiliary classifier head to identify and remove a labelled bias:

- Auxiliary regularisation loss minimises mutual information between feature embedding and bias.
- Gradient reversal applied to auxiliary classification loss during backpropegation as additional bias removal tool.

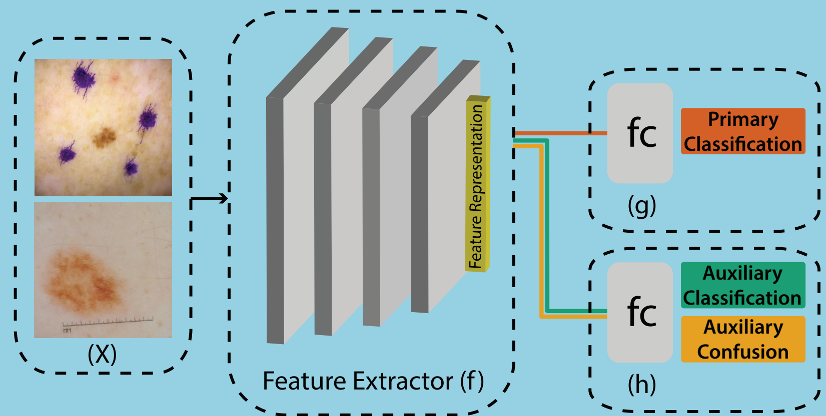


- Result of this is primary classification head learns to classify using a feature embedding that is independent of the target bias.

Turning a Blind Eye (TABE) [5]

Auxiliary classifier head to identify and remove a labelled bias:

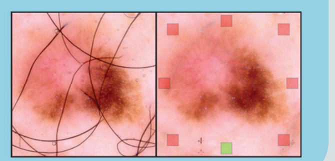
- Auxiliary confusion loss finds cross entropy between output predicted bias and uniform distribution towards finding a bias invariant feature representation.



- Gradient reversal can also be applied to the auxiliary classification loss for additional bias removal. We refer to this as **CLGR**.

Skin Tone Detection & Bias Identification

- We use a DB-VAE [6] as a tool to provide evidence of skin tone bias.
- We label skin type automatically using our healthy skin sampling method (right).

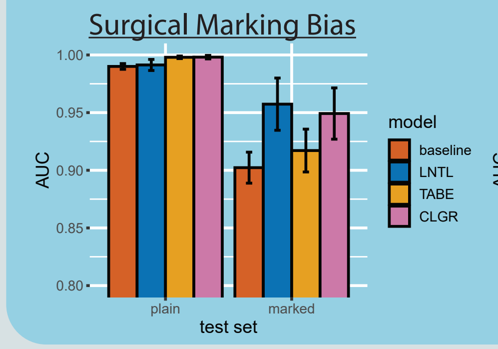


Experimental Results

Artefact Bias Removal

Models tested on the same lesions with and without artefacts present.

Debiasing heads mitigate both surgical marking and ruler bias.



Instrument Bias Removal

Using Turning a Blind Eye to unlearn instrument information leads to improved generalisation, with improved performance across several dermoscopic and clinical test sets.

Note: 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**

Skin Tone Bias Removal

Unlearning skin tone leads to improved generalisation to datasets with different distributions of skin tones.

Note: 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	0.803	0.608	0.765	0.858	0.787
TABE	**0.825**	**0.707**	**0.809**	0.865	**0.859**
CLGR	0.820	0.641	0.740	**0.918**	0.771

TABE head allows model trained using mostly White western data to generalise better to Korean data (ASANC).

[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] Groh et al., 'Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset'
[4] Kim et al., 'Learning Not to Learn: Training Deep Neural Networks With Biased Data'
[5] Alvi et al., 'Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings'
[6] Amini et al., 'Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure'