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

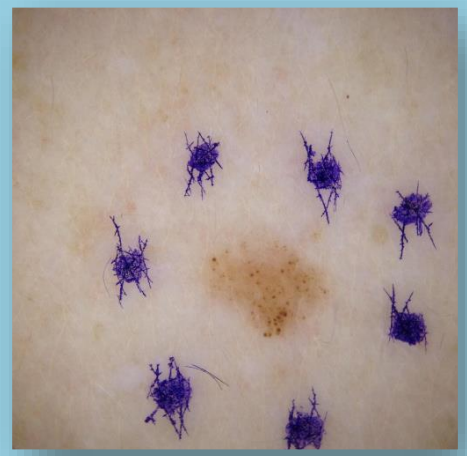
Peter Bevan and Amir Atapour-Abarghouei

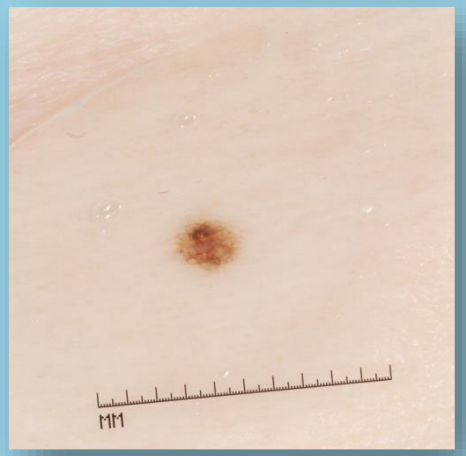
School of Computing, Newcastle University, Newcastle, UK

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



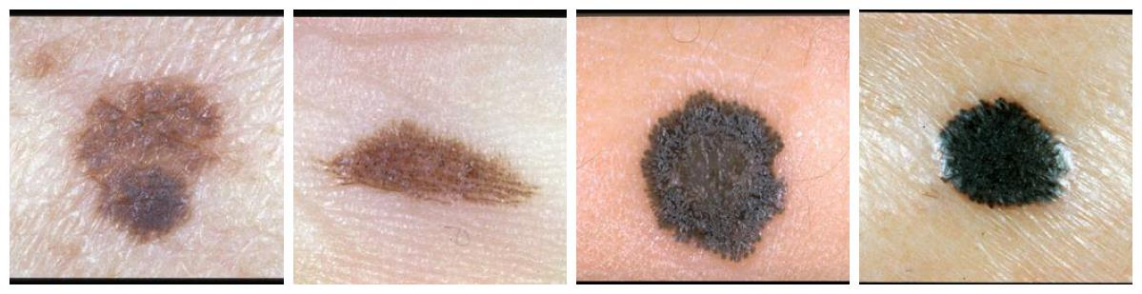


Motivation: Instrument Bias

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

Dermoscopic

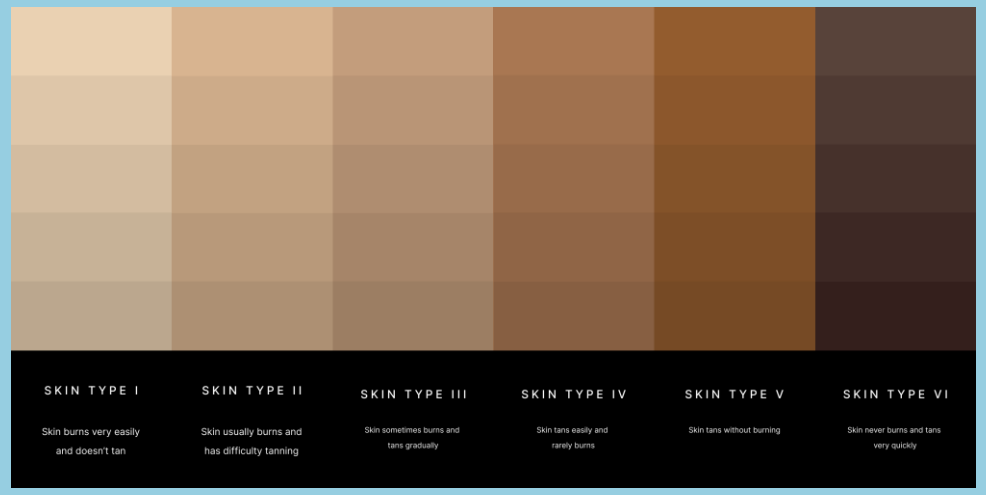
Clinical



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

Problems: Skin Tone Bias

CNNs perform best on skin tone similar to the model training data: a model trained on skin types 1&2 performed better on types 3&4 than types 5&6 [3].



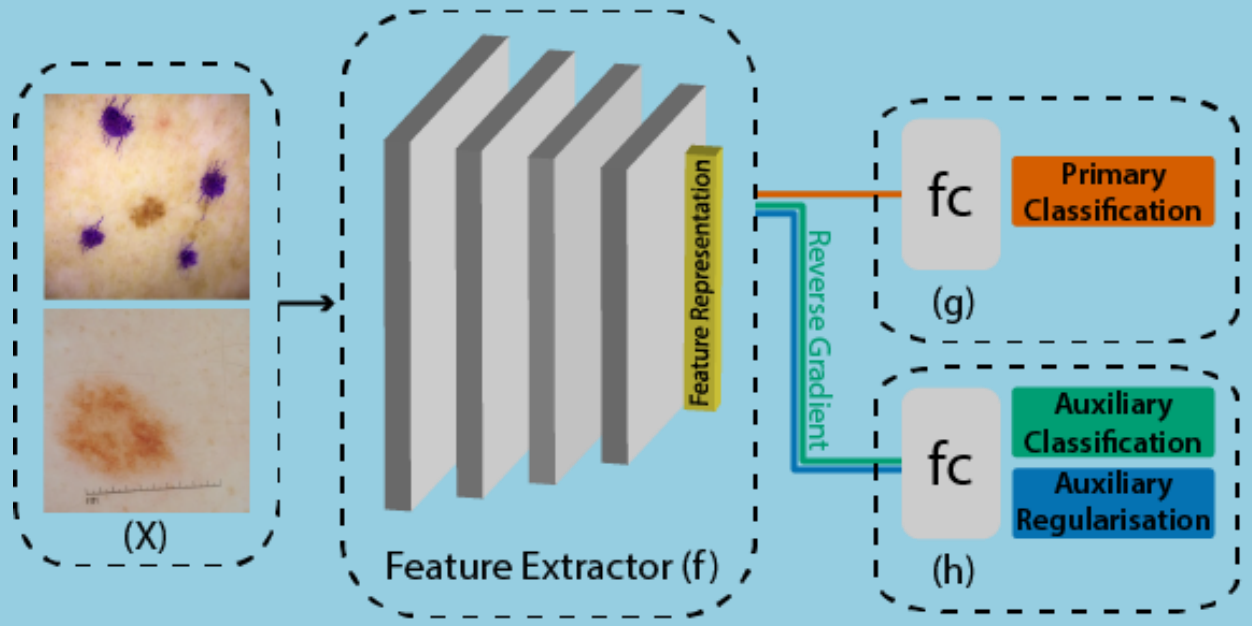
Type 1 Type 6

We investigate if removing skin type information from the model 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:

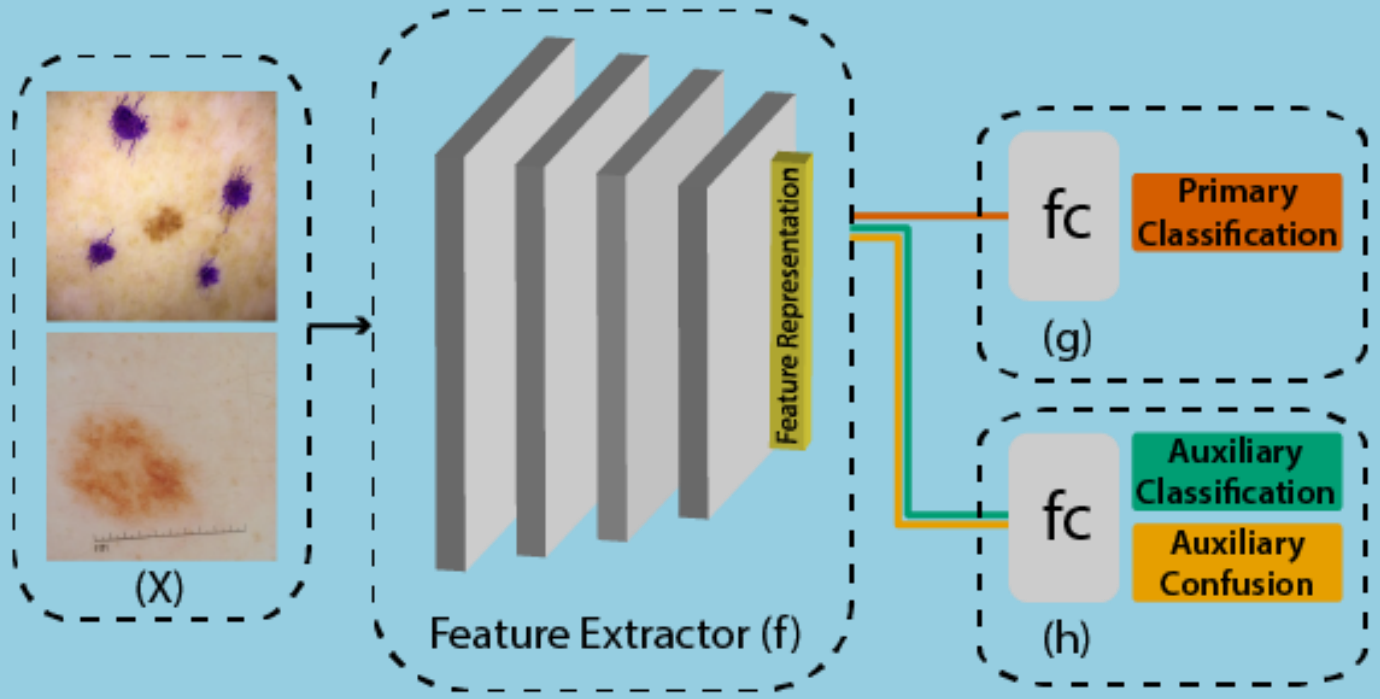
- ❑ Auxiliary regularisation loss minimises mutual information between the feature embedding and the targeted bias.
- ❑ Gradient reversal applied to auxiliary classification loss during backpropegation as additional bias removal tool.
- ❑ Goal is that the primary classification head learns to classify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



Methods: Turning A Blind Eye (TABE) [5]

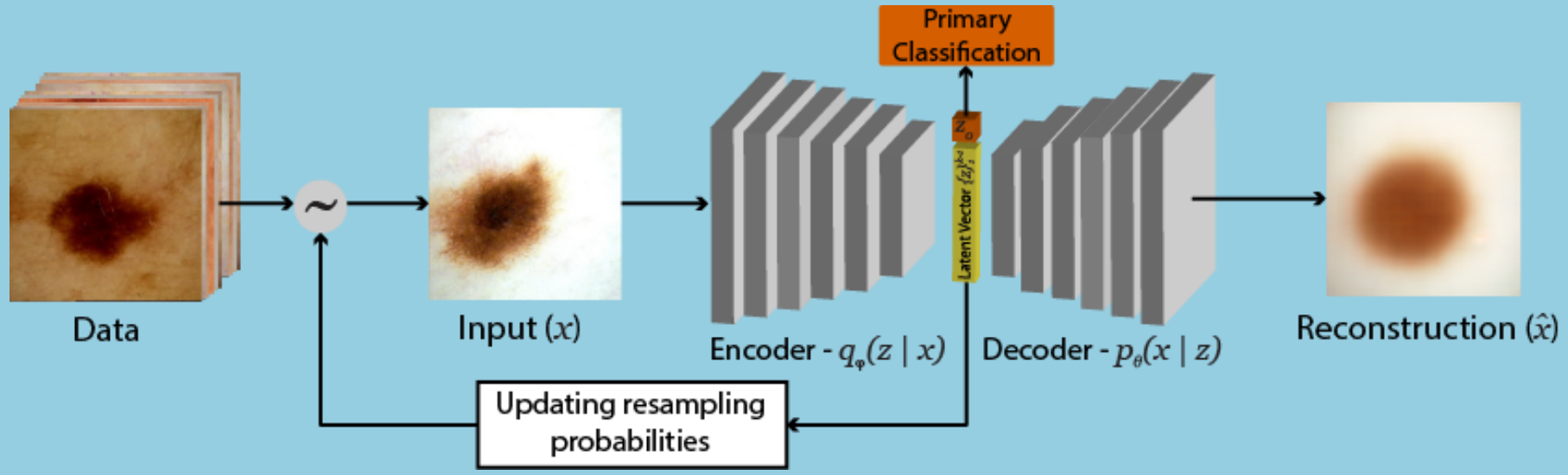
Auxiliary classifier head to identify and remove a labelled bias:

- ❑ 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 in TABE for additional bias removal. We refer to this configuration as **CLGR**.



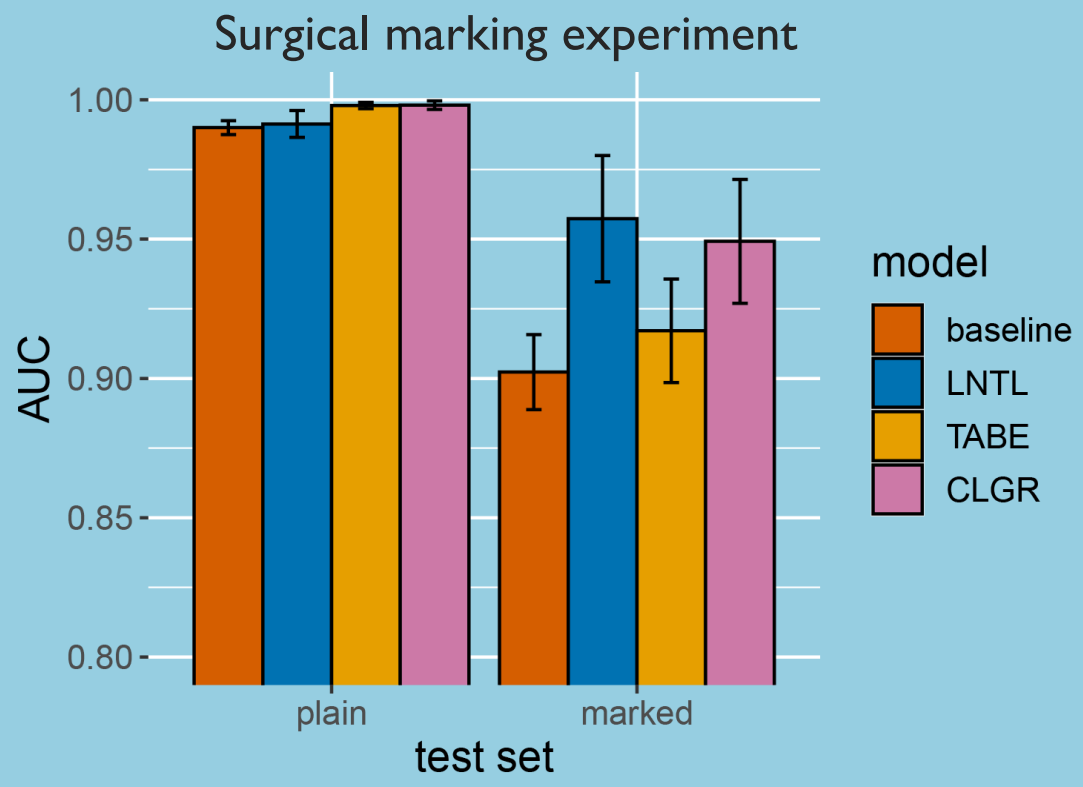
Methods: Debiasing Variational Autoencoder (DB-VAE)

- ❑ Most popular melanoma classification dataset has not been investigated for skin type bias
- ❑ We attempt to uncover evidence using a debiasing variational autoencoder (DB-VAE)
- ❑ Calculates resampling probabilities from learned latent structure
- ❑ We use these resampling probabilities to visualise least represented images and view their skin tone.
- ❑ Attempt to find latent variable encoding skin type
- ❑ We use a debiasing variational autoencoder to provide evidence of skin type bias in one of the most popular melanoma classification datasets



Experimental Results: Artefact Bias Removal

- ❑ Models are tested on the same lesions with and without artefacts present.
- ❑ ***Debiasing heads mitigate both surgical marking and ruler bias.***



Experimental Results: Instrument Bias Removal

Using Turning a Blind Eye [5] to unlearn instrument information leads to improved generalisation, with improved performance (compared to the baseline) across several dermoscopic and clinical test sets.

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**

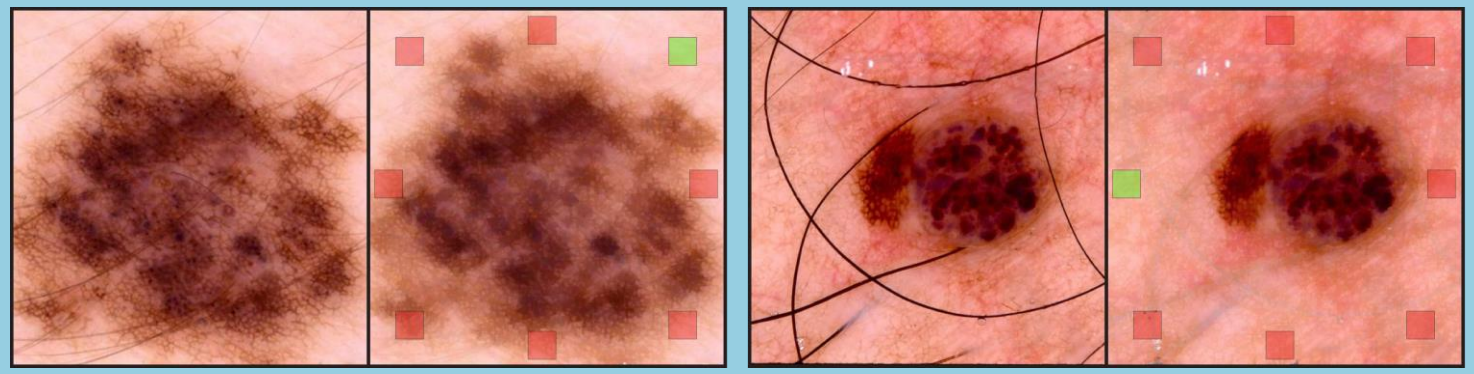
Methods: Skin Tone Detection Algorithm

We use a modified VAE [6] as a tool to uncover skin tone bias in a popular melanoma dataset (ISIC) by visualising the latent representation and finding underrepresented images.

We label the skin types of this dataset using our automated skin tone detection algorithm.

We use these labels as the target for debiasing heads to remove skin tone bias when training using this dataset.

Skin tone labelling algorithm sampling method



Experimental Results: Skin Tone Bias Removal

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

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).

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] 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’