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

Peter Bevan
School of Computing
Newcastle University
Newcastle, UK
p.bevan2@newcastle.ac.uk

Amir Atapour-Abarghouei
School of Computing
Newcastle University
Newcastle, UK
amir.atapour-abarghouei@newcastle.ac.uk

Abstract

Convolutional Neural Networks have demonstrated dermatologist level performance in the classification of melanoma and other skin lesions, but performance irregularities due to bias are an issue that should be addressed before widespread deployment. In this work, we robustly remove bias and spurious variation from an automated melanoma classification pipeline using two leading bias 'unlearning' techniques. We show that the biases introduced by surgical markings and rulers presented in previous studies can be partially mitigated using these methods. We also demonstrate the generalisation benefits of 'unlearning' spurious variation relating to the imaging instrument used to capture lesion images. Separately, we look to tackle skin tone bias in melanoma classification. We propose a simple algorithm for automatically labelling the skin tone of lesion images, and use this to annotate the benchmark ISIC dataset. Finally, we provide evidence that 'unlearning' skin tone improves generalisation and can reduce performance disparity between lighter and darker skin tones. The novel contributions of this work include a comparison of different debiasing techniques for artefact bias removal, the concept of instrument bias 'unlearning' for domain generalisation in melanoma detection, the application of bias 'unlearning' for skin tone bias removal in melanoma detection, and a novel variation of an existing method for automated skin tone labelling. Our experimental results provide evidence that each of the aforementioned biases are mitigated to some degree, with different debiasing techniques excelling at different tasks.

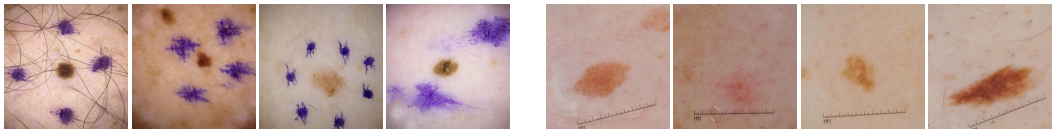
1 Introduction

In recent years, Convolutional Neural Networks (CNN) have demonstrated performance levels on par with experienced dermatologists in skin lesion diagnosis [26, 6, 7]. When diagnosed early, melanoma may be easily cured by surgical excision [57, 27], and so accessible and accurate diagnostic tools have the potential to democratise dermatology and save many lives worldwide.

While deploying such learning-based techniques far and wide could be massively beneficial, great care must be taken as any small pitfall could be replicated on a massive scale. For example, some dermatologists use visual aids such as skin markings to mark the location of a lesion, or rulers to indicate scale, as seen in Figure 1. In fact, Winkler et al. [57, 56] demonstrated how bias induced by the presence of these artefacts can result in diminished classification performance. They also suggest that dermatologists avoid using these aids in the future, which is a valid solution to the problem, but changing the habits of every dermatologist is an unrealistic task. Segmentation of the lesion from the surrounding skin has also previously been proposed, but is not a good option, as stated in [56]: "any kind of pre-processing or segmentation itself may erroneously introduce changes that impede a CNN's correct classification of a lesion". Cropping surgical markings out of the image was shown

Preprint. Under review.

to be effective at mitigating surgical marking bias in [57], but it was noted this must be done by an experienced dermatologist to prevent the loss of important information, which is costly and time consuming. Recent advances in debiasing architectures for CNNs [1, 33] present a good opportunity to robustly mitigate the aforementioned biases without any need to alter the behaviour of physicians or pre-process the training data.



(a) Surgical marking present.

(b) Ruler present.

Figure 1: Artefacts seen in the ISIC 2020 dataset [45].

Another issue that plagues many machine learning models is the domain shift between the training and real-world inference data, leading models to perform poorly upon deployment. One cause of this domain shift in skin lesion classification is likely to be spurious variation from minor differences in the imaging instruments used to capture lesions. Inspired by [19], we propose also using 'unlearning' techniques [33, 1] for domain generalisation by removing spurious variation associated with instrument type to create a more generalisable, instrument-invariant model.

Another issue that is commonly raised in the existing literature is skin tone bias in lesion classification tasks. Groh et al. [22] present a compiled dataset of clinical lesions with human annotated Fitzpatrick skin type [18] labels, and show that CNNs perform best at classifying skin types similar to the skin types in the training data used. We use the skin type labels in this dataset as the target for bias 'unlearning' heads to evaluate their effectiveness in helping melanoma classification models generalise to unseen skin types. We then use a debiasing variational autoencoder [2] (see Section 3.3) in an attempt to uncover hidden skin tone bias in ISIC data [45, 11]. We propose a novel variation on the skin tone labelling algorithm presented in [34] to annotate the ISIC data and subsequently use the generated labels as the target in a debiasing head towards improving the generalisation of models to images of individuals from differing ethnic origins.

In summary, this work aims to explore bias and domain 'unlearning' towards creating more robust, generalisable and fair models for melanoma classification. Our primary contributions (objectives set out at the beginning of the project) can be summarised as follows:

- *Artefact debiasing* - We mitigate the surgical marking and ruler bias shown in [57, 56] using 'Learning Not to Learn' [33] and 'Turning a Blind Eye' [1] (see Section 4.1).
- *Instrument debiasing* - We demonstrate the generalisation benefits of unlearning [33, 1] information relating to the instruments used to capture skin lesion images (see Section 4.2).
- *Skin tone detection* - We propose a novel and effective variation of Kinyanjui et al.'s [34] skin tone detection algorithm (see Section 4.3).
- *Skin tone debiasing* - We evaluate the effectiveness of unlearning [33, 1] skin tone bias for melanoma classification (see Section 4.3).

We introduce related work in Section 2, methods and data in Section 3, experimental results in Section 4, limitations of the project and potential future work in Section 5, and conclusions in Section 6.

2 Related work

2.1 Artefacts bias

The first problem addressed in this project was investigated in [9], which notes the algorithmic bias introduced by certain artefacts present in skin lesion images. Further precedent for investigating debiasing in skin lesion classification is found in [57], which compares the performance of a CNN classification model on 130 lesions *without* surgical markings present, versus the same 130 lesions *with* surgical markings present. Strong bias was demonstrated, with specificity hit hard, as well as Area Under the Curve (AUC) [57]. Another work [56] shows a similar level of bias caused by rulers in skin lesion images.

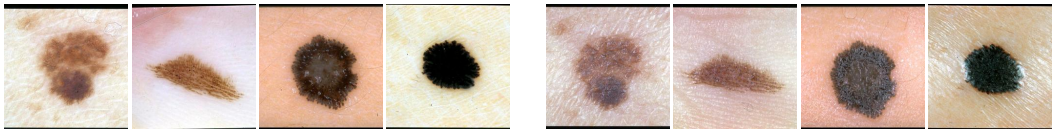
In an earlier work, Bissoto et al. [5] tackle the issue of artefact bias removal in a manner similar to the one proposed in this work by using a bias removal model [33] with seven debiasing heads in an attempt to remove the bias caused by seven artefacts. The authors conclude that the bias removal

method in [33] ('Learning Not to Learn') is not ready to tackle the issue, however ablation studies to isolate each head are lacking, and so the efficacy of each of the seven individual debiasing heads cannot be ascertained: it is possible that certain heads bring down the performance of the entire model, or interact with each other unfavourably. In addition to this, the paper does not experiment with other leading debiasing solutions such as the one proposed in [1] ('Turning a Blind Eye'), which may be more effective at the given task. In this work, we only focus on biases that are well documented as causing performance degradation, and compare individual debiasing heads across different methods before combining these heads. Bisotto et al. do note improvements in performance when testing their debiasing models on data with significant domain shift (Interactive Atlas of Dermoscopy data [38]), which indicates some improvement in generalisation. We build upon this notion in our domain generalisation experiments (Section 4.2).

2.2 Domain generalisation

A common assumption in machine learning is that the training and test data are drawn from the same distribution, however this assumption does not usually hold true in real-world applications [14]. Inconsistencies in prostate cancer classification performance between image samples originating from different clinics was shown in [3], and the authors hypothesised that this could be due to domain shift caused by variation in the equipment used. In skin lesion classification, there are two main imaging methods: dermoscopic (skin surface microscopy), and clinical (standard photograph) [55] (see Figure 2). This domain shift was shown to impact model performance in [23]. Within these two imaging methods, many different brands and models of instrument are used by different clinics, which may also introduce domain bias. Supporting this hypothesis, it was shown in [31] that CNNs can easily discriminate between camera models, which can lead models to overfit to this spurious variation during training.

Domain adaptation has been used to minimise the distance between source and target distributions [14, 24, 23], however these methods require knowledge of the target distribution. Domain generalisation is more robust than domain adaptation, and differs in that the target domain is unseen [36], aiming for improved performance on a wide range of possible test data. We explore applying bias unlearning techniques [33, 1] towards domain generalisation in melanoma classification, attempting to find an instrument invariant feature representation without compromising performance.



(a) Dermoscopic. (b) Clinical.

Figure 2: Comparison between Interactive Atlas of Dermoscopy [38] clinical and dermoscopic images of the same lesion. Dermoscopic images use skin surface microscopy for reduced surface reflection.

2.3 Skin tone bias

Groh et al. [22] showed that CNNs perform better at classifying images with similar skin tone to those the model was trained on. Performance is therefore likely to be poor for dark skinned patients when the training data is predominantly images of light skinned patients, which is the case with many of the current commonly-used dermoscopic training datasets such as the ISIC archive data [45, 11]. While melanoma incidence is much lower among the black population (1.0 per 100,000 compared to 23.5 per 100,000 for whites), 10-year melanoma-specific survival is lower for black patients (73%) than white patients (88%) or other races (85%) [12], and so it is of heightened importance to classify lesions in patients of colour correctly. One way to ensure a more even classification performance across skin tones is to re-balance the training data by collecting more high quality images of lesions on skin of colour, however the low incidence of melanoma in darker skin means this could be a slow process over many years. Rather than placing the deployment of automated skin lesion classification on hold until sufficient data is obtained, a robust automated method for removing skin tone bias from the model pipeline may allow models to operate with increased fairness across skin tones, even with unbalanced data. Skin tone/racial bias has been mitigated in other domains such as facial recognition [1, 2, 20, 54], but not in skin lesion classification to date.

3 Methods

In this work, two leading debiasing techniques within the literature are used, namely 'Learning Not To Learn' (LNTL) [33] and 'Turning a Blind Eye' (TABE) [1]. Both of these are often referred to as

'unlearning' techniques because of their ability to remove bias from the feature representation of a network by minimising the mutual information between the feature embedding and the unwanted bias. Further details of these unlearning methods are described in Sections 3.1 and 3.2. A debiasing variational autoencoder is used to uncover skin tone bias, and is described in Section 3.3.

3.1 Learning Not to Learn

'Learning Not to Learn' (LNTL) [33] proposes a novel regularisation loss combined with a gradient reversal layer [19] to remove bias from the feature representation of a CNN during backpropagation. Figure 13 in Appendix B shows a generic overview of the LNTL architecture. The input image, x, is passed into a feature extractor f: $x \rightarrow \mathbb{R}^K$, where K is the dimension of the embedded feature. The feature extractor is implemented as a pre-trained convolutional architecture such as ResNeXt [29] or EfficientNet [51] in this work. This extracted feature embedding is then passed in parallel into both g: $\mathbb{R}^K \rightarrow \mathcal{Y}$ and h: $\mathbb{R}^K \rightarrow \mathcal{B}$, the primary and auxiliary classification heads respectively, where, in the case of this project, $\mathcal{Y}$ represents the set of possible lesion classes and $\mathcal{B}$ represents the set of target bias classes.

The networks f and h play the minimax game, in which h is trained to classify the bias from the extracted feature embedding (minimising cross-entropy), whilst f is trained to maximise the cross-entropy to restrain h from predicting the bias, and also to minimise the negative conditional entropy to reduce the mutual information between the feature representation and the bias. The gradient reversal layer between h and f acts as an additional tool to remove information relating to the target bias from the feature representation. The gradient reversal layer works by multiplying the gradient of the auxiliary classification loss by a negative scalar during backpropagation, causing the feature extraction network f to 'learn not to learn' the targeted bias, $b(x)$, rather than learn it. By the end of training, f has learnt to extract a feature embedding independent of the bias, g has learnt to use this feature embedding to perform the primary classification task without relying on the bias, and h performs poorly at predicting the bias due to the lack of bias information in the feature embedding.

The minimax game along with the main classification loss are formulated as:

$$\min_{\theta_f,\theta_g} \max_{\theta_h} \mathbb{E}_{\tilde{x}\ P_X(\cdot)}[\underbrace{\mathcal{L}_g(\theta_f,\theta_g)}_{(a)} + \underbrace{\lambda \mathbb{E}_{\tilde{b}\sim Q(\cdot|f(\tilde{x}))}[\log Q(\tilde{b}|f(\tilde{x}))]]}_{(b)} - \underbrace{\mu \mathcal{L}_{\mathcal{B}}(\theta_f,\theta_h)}_{(c)}, \tag{1}$$

where (a) represents the cross-entropy loss of the main classification head, (b) represents the regularisation loss and (c) represents the cross-entropy loss of the auxiliary bias classification head. The hyperparameters λ and μ are used to balance the terms. The parameters of each network are denoted as θ_f, θ_g and θ_h. An auxiliary distribution, Q, is used to approximate the posterior distribution of the bias, P, which is paramaterised as the bias prediction network h.

3.2 Turning a Blind Eye

Figure 14 in Appendix B shows a generic overview of the 'Turning a Blind Eye' (TABE) [1] architecture. Similar to LNTL [33], this method also removes unwanted bias using an auxiliary classifier, θ_m, where m is the m-th unwanted bias. The TABE auxiliary classifier minimises an auxiliary classification loss, $\mathcal{L}_s$, used to identify bias in the feature representation, θ_{repr}, as well as an auxiliary confusion loss [52], $\mathcal{L}_{\text{conf}}$, used to make θ_{repr} invariant to the unwanted bias. Since these losses stand in opposition to one another, they are minimised in separate steps: first $\mathcal{L}_s$ alone, and then the primary classification loss, $\mathcal{L}_p$, together with $\mathcal{L}_{\text{conf}}$. The confusion loss is defined as follows:

$$\mathcal{L}_{\text{conf,m}}(x_m, y_m, \theta_m; \theta_{\text{repr}}) = -\sum_{n_m} \frac{1}{n_m} \log p_{n_m}, \tag{2}$$

where x_m is the input, y_m is the bias label, p_{n_m} is the softmax of the auxiliary classifier output and n_m is the number of auxiliary classes. This confusion loss works towards finding a representation in which the auxiliary classification head performs poorly by finding the cross entropy between the output predicted bias and a uniform distribution. The complete joint loss function being minimised is:

$$\mathcal{L}(x_p, y_p, x_s, y_s, \theta_p, \theta_s, \theta_{\text{repr}}) = \mathcal{L}_p(x_p, y_p; \theta_{\text{repr}}, \theta_p) + \mathcal{L}_s + \alpha \mathcal{L}_{\text{conf}}, \tag{3}$$

where α is a hyperparameter which determines how strongly the confusion loss impacts the overall loss. The feature extractor, f, is implemented as a pre-trained convolutional architecture such as ResNeXt [29] or EfficientNet [51] in this work.

As suggested in [33], a hybrid of LNTL and TABE can be created by utilising the confusion loss (CL) from TABE [1], and then also applying gradient reversal (GR) from LNTL [33] to the auxiliary classification loss as it is backpropegated to f. This configuration is denoted as CLGR in this report.

3.3 Debiasing variational autoencoder (DB-VAE)

Figure 15 in Appendix B shows the generic debiasing variational autoencoder (DB-VAE) architecture [2], which utilises a modified VAE network to classify images, as well as identify and mitigate hidden biases in an unsupervised manner by using the underlying latent features, z, to dictate adaptive resampling of the training data. This resampling means the network is shown a more representative subset of the original dataset, and acts as an automatically-debiasing classifier. The custom CNN encoder however performs classification poorly in comparison to more complex architectures such as ResNeXt [58] and EfficientNet [51], and so is not a realistic option for deployment. We instead use the model as a tool to uncover evidence of whether skin tone bias exists in ISIC data [45, 11]: we use the resampling probabilities to visualise the least and most represented images, and also perturb individual latent variables to understand if skin tone is being used as a feature for classification.

In the DB-VAE model, the distribution of the latent variables, $q_\phi(z|x)$, is approximated by the encoder, however in contrast to classical variational autoencoders (VAEs), d additional supervised output variables are introduced where $\hat{y} \in \mathbb{R}^d$. This allows classification of the input images using the supervised loss, $\mathcal{L}_y(y, \hat{y})$, given by the cross-entropy loss where y is the vector of ground truth labels relating to the primary classification task. As is standard with VAEs, the decoder mirrors the encoder and reconstructs the input image from the latent vector by approximating $p_\theta(x|z)$. The reconstruction loss, $\mathcal{L}_x(x, \hat{x})$, is given by the L_p norm between the input and the reconstructed output. The Kullback-Liebler (KL) divergence is used as the latent loss, $\mathcal{L}_{KL}(\mu, \sigma)$. The total loss is defined as a weighted combination of these three losses:

$$\mathcal{L}_{TOTAL} = c_1 \underbrace{\left[\sum_{i \in \{0,1\}} y_i log\left(\frac{1}{\hat{y}_i}\right) \right]}_{\mathcal{L}_y(y,\hat{y})} + c_2 \underbrace{\left[\|x - \hat{x}\|_p \right]}_{\mathcal{L}_x(x,\hat{x})} + c_3 \underbrace{\left[\frac{1}{2} \sum_{j=0}^{k-1} (\sigma_j + \mu_j^2 - 1 - \log(\sigma_j)) \right]}_{\mathcal{L}_{KL}(\mu,\sigma)}, \tag{4}$$

where c_1, c_2 and c_3 are weighting coefficients which affect the relative importance of each individual loss function. It should be noted that independent histograms, $\hat{Q}_i(z_i(x)|X)$, are used for each latent variable, the product of which is used to approximate the joint distribution $\hat{Q}(z|X)$. The probability distribution of sampling a data point x is defined as:

$$\mathcal{W}(z(x)|X) \propto \prod_i \frac{1}{\hat{Q}_i(z_i(x)|X) + \alpha}, \tag{5}$$

where α is a hyperparameter used to tune the degree of debiasing during training. These probabilities are updated each epoch.

3.4 Data

This section briefly describes each dataset used in the project (see Appendix C for example images).

3.4.1 ISIC challenge training data

The International Skin Imaging Collaboration (ISIC) challenge [45] is a yearly automated melanoma classification challenge with several publicly available, high-quality dermoscopic skin lesion datasets (see ISIC archive[1]), complete with diagnosis labels and metadata. A combination of the 2017 and 2020 ISIC challenge data [45, 11] (35,574 images) is used as the training data in this project due to the higher representation of artefacts in these datasets than other competition years. Pre-processed (centre cropped and resized) images of size 256×256 are used for all training and testing [16]. The surgical markings are labelled using colour thresholding, with the labels double-checked manually, while the rulers are labelled entirely manually. A random subset (33%, 3326 images) of the 2018 [10] challenge data is used as the validation set for hyperparameter tuning.

3.4.2 Artificially skewed ISIC training data

The model and training data used in [57, 56] is proprietary, and so the bias in these studies could not be exactly reproduced. Alternatively, since the primary objective is to investigate the possibility of removing bias from the task, we skew the ISIC data [45, 11] to produce similar levels of bias in our baseline model to that shown in the aforementioned studies [57, 56]. Benign lesions in the training data that had surgical markings are removed and images that are both malignant and marked are duplicated and randomly augmented (treating each duplicate as a new data point) to skew the model towards producing false positives for lesions with surgical markings, as shown in [57]. The dataset is

[1] `https://www.isic-archive.com/`

processed similarly with rulers to demonstrate ruler bias. The number of duplications of melanoma images with surgical markings present, dm, and with rulers present, dr, are used as hyperparameters to control the level of skew in experiments. Note that this artificially skewed data is only used to demonstrate artefact debiasing, and the original data is used for all other experiments.

3.4.3 Fitzpatrick17k training data

A publicly available compilation of clinical skin condition images with human annotated Fitzpatrick skin types [18] called the 'Fitzpatrick17k' dataset is presented in [22], and we use this dataset to demonstrate the effectiveness of unlearning for skin tone debiasing, and to evaluate our automated skin tone labelling algorithm. Of the 16,577 clinical images of skin conditions, we focus on the 4,316 of these that are neoplastic (tumorous). The Fitzpatrick six-point skin type labelling system is the standard system in dermatology for categorising skin tones [8, 18], type 1 being the lightest and type 6 the darkest. These labels are not provided by dermatologists, so are likely to be imperfect.

3.4.4 Heidelberg University test data

The test set presented in [57] is used to evaluate the bias mitigation approach presented in this work. The dataset consists of 130 lesions: 23 malignant, 107 benign. There are two images of each lesion in the set: one with no surgical markings present, and one with surgical markings present. This allows a direct evaluation of the effect of surgical marking bias on the performance of a model, as shown in [57]. The test set from the ruler bias study [56] is not publicly available or shared, so the plain images from [57] are superimposed with rulers (see Appendix C, Figure 20) to be used as test images. The approach of superimposing rulers was validated as not statistically significantly different from in-vivo rulers in [56], which compared in-vivo rulers and superimposed rulers on the same lesions.

3.4.5 MClass benchmark test data

The MClass public human benchmark introduced in [7] was obtained as a test dataset for assessing domain generalisation, as well as providing a human benchmark. This dataset comprises a set of 100 dermoscopic images and 100 clinical images (*different* lesions), each with 20 malignant and 80 benign lesions. The dermoscopic and clinical image sets were classified by 157 and 145 experienced dermatologists respectively, with their average classification performances published in [7]. The dermoscopic MClass data is made up of images from the ISIC archive, some of which were also present in the ISIC training data, so these were removed from the training data to prevent data leakage.

3.4.6 Interactive Atlas of Dermoscopy and ASAN test data

Two additional test sets, the Interactive Atlas of Dermoscopy dataset [38], and the ASAN test dataset [28], were obtained to further test domain generalisation. The Atlas dataset comprises 1000 lesions across 7 classes, with one dermoscopic and one clinical image per lesion. The ASAN test dataset comprises 852 clinical images across 7 classes of lesions. Whilst the ISIC training data [45, 11] is mostly white Western patients, the Atlas seems to have representation from a broad variety of ethnic groups, and ASAN from predominantly South Korean patients, which should allow a good test of a models ability to deal with different domain shifts.

3.5 Implementation

All experiments are implemented in PyTorch [43] and carried out using two NVIDIA Titan RTX GPUs in parallel with a combined memory of 48 GB on an Arch Linux system with a 3.30GHz 10-core Intel CPU and 64 GB of memory. The baseline model is inspired by the winning entry from the 2020 ISIC challenge [25], which utilises the EfficientNet-B3 architecture [51], pre-trained on the ImageNet dataset [15]. ResNet-101 [29], ResNeXt-101 [58], DenseNet [30] and Inception-v3 [50] are each substituted for EfficientNet-B3 in order to select the optimal network for the task, whilst simultaneously demonstrating the effectiveness of the debiasing techniques across different architectures.

Early experimentation showed ResNeXt-101 to be the overall best performing architecture, as seen in Table 2, and it is therefore used as the feature extractor in the domain generalisation and skin tone debiasing experiments. EfficientNet-B3 is kept as the base architecture for surgical marking and ruler debiasing since the baseline performance is closest to the unknown proprietary model used in [57]. The primary and auxiliary classification heads are implemented as a single fully connected layer, as suggested for gender and age bias in [33]. Stochastic gradient descent is used across all models, ensuring comparability and compatibility between the baseline and debiasing networks.

Following a grid search, the learning rate (searched between 0.03 and 0.00001) and momentum (searched between 0 and 0.9) are selected as 0.0003 and 0.9 respectively. See Appendix E for full results. The learning rate of the TABE heads is boosted by a factor of 10 (to 0.003), as suggested in [1],

except when using multiple debiasing heads since this seems to cause instability. The best performing values of the hyperparameters α and λ in Equations 1 and 3 are also empirically chosen: α=0.03 and λ=0.01. For the DB-VAE [22], we use a latent dimension size of 512 after also experimenting with 128, 256 and 1024 units. A weighted loss function is implemented for all auxiliary heads to tackle class imbalance, with each weight $\mathcal{W}_n$ the inverse of the corresponding class frequency c. Since the proportion of benign and malignant lesions is highly imbalanced in the test sets, accuracy proved not to be a descriptive metric to use. Instead, AUC is used as the primary metric across all experiments, as is standard in melanoma classification [37, 41, 28, 25] given that it takes into account both sensitivity and specificity across all thresholds and is effective at communicating the performance when the target classes are imbalanced [39]. We use test-time augmentation to average predictions over 8 random flips along different axes, applied to all test images, to increase the reliability of our scores. The optimal number of epochs for training each architecture on each dataset is chosen through analysis of the 5-fold cross validation curves for the baseline models, selecting the epoch at which the AUC reached its maximum or plateaued (see Appendix E, Table 8).

4 Experimental results

The results of our artefact bias removal experiments are presented in Section 4.1. We present the instrument debiasing experimental results in Section 4.2. Finally, the results of our skin tone bias experiments are presented in Section 4.3. The three variations of debiasing heads that were implemented and combined in the experiments are: 'Learning Not to Learn' (**LNTL**) [33], 'Turning a Blind Eye' (**TABE**) [1], and a hybrid of the confusion loss (CL) from TABE with the gradient reversal layer (GRL) from LNTL (**CLGR**).

4.1 Artefacts bias removal

We attempt to remove the bias caused by two artefacts that have been shown to affect performance in melanoma classification: surgical markings [57] and rulers [56] (see Table 1 and Figure 3). Separate individually-skewed training sets are used with skew levels set at dm=20 (duplications of marked images) for examining the removal of surgical marking bias and dr=18 (duplications of ruler images) for ruler bias.

Each model is trained and evaluated 6 times using 6 different random seeds, allowing the mean and standard deviation to be reported. The surprisingly high scores are likely due to the inherent easiness of classification in the sample of images in the test set, and are consistent with the scores reported in [57, 56]. The creators of the test set ruled out any chance of a leak between their test set and the ISIC training data. Despite the ease of classification, bias and bias mitigation can still be demonstrated, and using the test set from the original paper provides direct evidence that the methods help mitigate the problem presented in these studies [57, 56]. While the baseline model performs very well for the unbiased images ('Heid Plain'), performance suffers when this model is tested on the same lesions with either artefact present, replicating the findings from [57, 56].

Table 1: Comparison of each unlearning technique against the baseline, trained on artificially skewed ISIC data. All scores are **AUC**. 'Heid Plain' test set is free of artefacts while 'Heid Marked' and 'Heid Rulers' are the same lesions with surgical markings and rulers present respectively.

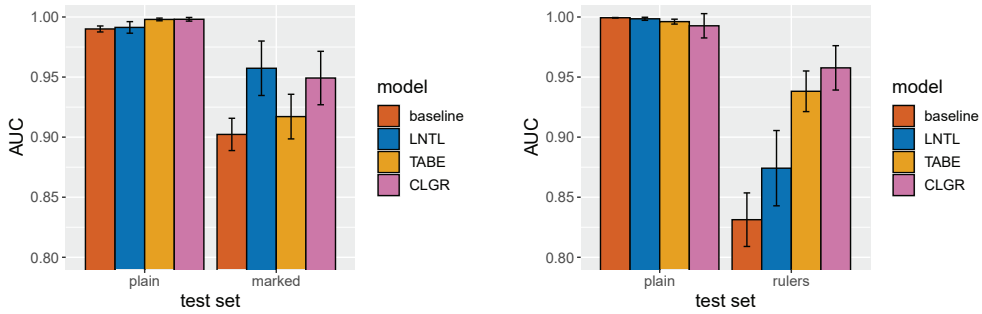
(a) Removal of surgical marking bias (dm=20).

Experiment	Heid Plain	Heid Marked
Baseline	0.990±0.002	0.902±0.013
LNTL†	0.991±0.005	**0.957**±0.023
TABE†	**0.998**±0.001	0.917±0.019
CLGR†	0.998±0.002	0.949±0.022

(b) Removal of ruler bias (dr=18).

Experiment	Heid Plain	Heid Ruler
Baseline	**0.999**±0.000	0.831±0.022
LNTL‡	0.997±0.001	0.874±0.031
TABE‡	0.992±0.002	0.938±0.017
CLGR‡	0.999±0.010	**0.958**±0.018

Figure 3a presents evidence that each debiasing method is successful at mitigating artefact bias to some extent. LNTL is the most effective at unlearning surgical marking bias, achieving comparable performance to the baseline on the plain images from Heidelberg University [57] ('Heid plain'), and improving on the baseline AUC by 0.055 (6.1% increase) on the equivalent marked images from the same set ('Heid marked'). All three techniques also mitigate ruler bias well, with CLGR being the most effective and showing a 0.127 increase in AUC compared to the baseline (15.3% increase). The results of our experiments suggest that unlearning techniques can be used to reduce the bias demonstrated in [57, 56], but are not a perfect solution, given that the artefacts still have a negative impact on performance.



(a) Removal of surgical marking bias (dm=20).

(b) Removal of ruler bias (dr=18).

Figure 3: Comparison of each artefact debiasing model against the baseline, trained on artificially skewed ISIC data.

4.2 Instrument bias removal

We attempt to remove instrument bias from the model pipeline using unlearning techniques [33, 1], with the aim of improving domain generalisation. According to ISIC, image dimensions in [45, 11] are a good proxy for the imaging instrument used to capture the image[2]. These dimensions were used as the auxiliary target for debiasing, attempting to remove spurious variation related to the imaging instrument from the feature representation. The vast majority (98%) of the ISIC training images [45, 11] made up the first 8 'instrument' categories, but there were many outlier categories with a very small number of observations, which were discarded to prevent class imbalance.

Table 2: Comparing generalisation ability of each debiasing method across different architectures, trained using ISIC 2017 and 2020 data [45, 11]. All scores are **AUC**. The 'dermatologists' row is the AUC scores from [7]. Dermoscopic and clinical sets identified by 'D' and 'C' respectively. The § symbol indicates the use of instrument labels for the auxiliary head. Bold numbers are the highest score for that architecture, underlined scores are the highest scores across all architectures.

Experiment	Architechture	AtlasD	AtlasC	ASANC	MClassD	MClassC
Dermatologists	—	—	—	—	0.671	0.769
Baseline	EfficientNet-B3	0.757	0.565	0.477	0.786	0.775
LNTL§	EfficientNet-B3	0.709	0.562	0.570	0.830	0.630
TABE§	EfficientNet-B3	**0.811**	**0.629**	**0.685**	0.877	**0.889**
CLGR§	EfficientNet-B3	0.761	0.562	0.656	**0.882**	0.838
Baseline	ResNet-101	**0.802**	0.606	0.704	**0.877**	**0.819**
LNTL§	ResNet-101	0.776	0.540	**0.766**	0.817	0.748
TABE§	ResNet-101	0.746	0.541	0.617	0.809	0.808
CLGR§	ResNet-101	0.795	**0.615**	0.723	0.870	0.739
Baseline	ResNeXt-101	**0.819**	0.616	0.768	0.853	0.744
LNTL§	ResNeXt-101	0.776	0.597	0.746	0.821	0.778
TABE§	ResNeXt-101	0.817	**0.674**	**0.857**	**0.908**	0.768
CLGR§	ResNeXt-101	0.784	0.650	0.785	0.818	**0.807**
Baseline	DenseNet	0.775	0.559	0.655	0.851	0.695
LNTL§	DenseNet	0.760	0.548	0.750	0.859	0.689
TABE§	DenseNet	**0.809**	**0.622**	0.743	**0.863**	**0.788**
CLGR§	DenseNet	0.760	0.596	**0.872**	0.843	0.776
Baseline	Inception-v3	0.762	0.528	0.671	0.784	0.605
LNTL§	Inception-v3	**0.784**	0.556	0.729	0.809	0.583
TABE§	Inception-v3	0.751	**0.593**	0.735	0.818	**0.746**
CLGR§	Inception-v3	0.722	0.537	**0.775**	**0.847**	0.706

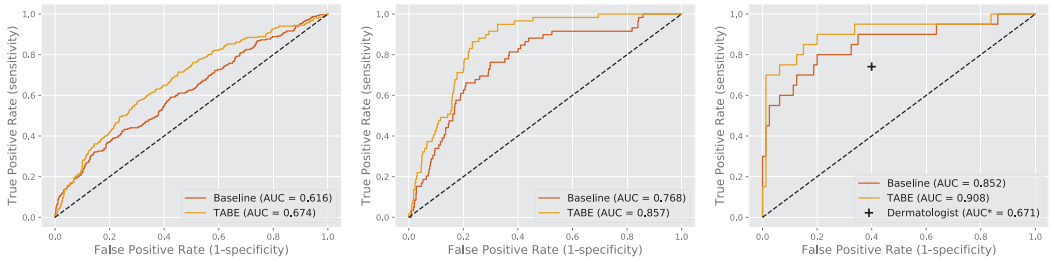
Table 2 compares the generalisation ability of each debiasing method when attempting to unlearn the imaging instrument used against the baseline. We test the models on a number of datasets of differing distributions to test generalisation. We apply the debiasing heads to several different model architectures and compare the results, allowing us to select a champion architecture for further experimentation. ResNeXt-101 is chosen for further experimentation since it achieves the highest

[2]The ISIC were contacted in search of labels for the origin clinics of their data and they pointed out the association between image dimensions and origin.

score on 3 out of the 5 test sets, as seen in Table 2. TABE and CLGR (TABE with gradient reversal) consistently outperform the baseline across all architectures. It is reasonable to attribute the increase in performance to the debiasing head given that this is the only change made from the baseline. On the MClass clinical test set, the CLGR head is the difference between the model performing below the dermatologist benchmark, and exceeding it (8.5% AUC increase), highlighting the potential impact of these domain generalisation methods.

In general, the greatest performance increases come on the clinical test sets, likely due to the fact these have the greatest domain shift compared to the dermoscopic training set. The models utilising a LNTL head were less successful, and even negatively impacted performance in some cases. This highlights that a single technique should not be applied in blanket fashion, as is done in [5], but rather experimentation should be carried out to select a technique suitable for the specific task.

Figure 4 illustrates the performance of using a TABE head for instrument removal compared to the baseline model (both ResNeXt-101). TABE can be differentiated from the baseline across each clinical test set, suggesting this to be a good tool for domain generalisation between dermoscopic and clinical data. Performance is generally poor on the Atlas clinical dataset, likely due to the noisiness of the images, however performance is still improved by the debiasing head. TABE also enhances performance on the MClass dermoscopic data [7] (see Figure 4c), indicating there may be benefits to unlearning instrument information even on data drawn from a similar distribution, through unlearning spurious variation relating to the specific type of dermoscopic instrument used.



(a) Atlas clinical [38] (b) ASAN clinical [28] (c) MClass dermoscopic [7]

Figure 4: ROC curves for TABE [1] instrument debiasing on three test sets [38, 28, 7], with **ResNeXt-101** as the base architecture. Model trained using the ISIC 2020 [45] and 2017 data [11].

We also conduct a set of experiments using two additional debiasing heads, each removing a different bias (either instrument, surgical marking or ruler). The best performing configurations are shown in Table 3. Using a single TABE head to remove instrument bias is still the most effective overall configuration, however scores were improved on the MClass clinical [7] test set using CLGR with ruler labels, and on the Atlas dermoscopic [38] test set using one TABE head and one LNTL head with instrument and ruler labels as their respective targets. For results across a more complete set of configurations, please refer to Table 10 in Appendix Section F.2.

Table 3: Generalisation of **ResNeXt-101** models trained using ISIC 2017 and 2020 data. The 'dermatologists' row is the AUC scores from [7]. Dermoscopic and clinical sets identified by 'D' and 'C' respectively. Instrument, surgical marking and ruler labels represented by §, † and ‡ respectively.

Experiment	AtlasD	AtlasC	ASANC	MClassD	MClassC
Dermatologists	—	—	—	0.671	0.769
Baseline	0.819	0.616	0.768	0.853	0.744
TABE§	0.817	**0.674**	**0.857**	**0.908**	0.768
CLGR‡	0.818	0.610	0.760	0.886	**0.882**
TABE§&LNTL‡	**0.828**	0.640	0.747	0.880	0.824

4.3 Skin tone bias

A CNN trained using Fitzpatrick [18] types 1 and 2 skin is shown to perform better at classifying skin conditions in types 3 and 4 than types 5 and 6 skin in [22]. We are able to reproduce these findings with our baseline ResNeXt-101 model, trained and tested on the neoplastic subset of the Fitzpatrick17k data. We attempt to close this gap with the addition of an auxiliary debiasing head which uses skin type labels as its target. The CLGR configuration proves to be most effective, and is shown in Figure 4. The disparity in AUC between the two groups is closed from 0.037 to 0.030, with types 3 and 4 boosted by 1.3% and types 5 and 6 boosed by 2.2%. While the improvement is modest,

the baseline performance is already strong on this dataset, meaning even very small improvements are highly valuable. This experiment serves as a proof of concept for the mitigation of skin tone bias with unlearning techniques, and gives us precedent to explore this for debiasing the ISIC data [45, 11].

Table 4: Attempting to improve model generalisation to skin tones different to the training data [22]. All scores are **AUC**. Trained using types 1&2 skin images from the Fitzpatrick17k dataset [22], tested on types 3&4 skin and types 5&6 skin from the same set.

Experiment	Types 3&4	Types 5&6
Baseline	0.872	0.835
CLGR	**0.883**	**0.853**

We use a DB-VAE model [2] to evidence the existence of skin tone bias in the ISIC data [45, 11] in an unsupervised manner, since skin tone bias has not yet been demonstrated in this dataset. We visualise the 9 images that the network assigns the greatest and least probabilities of resampling based on Eq. (5), uncovering the images which the model calculates to be under or over represented (see Figure 5). The images with the highest representation are light skinned with light coloured lesions, consistent with types 1 and 2 skin. The least represented images are more diverse, with most lesions generally darker. This gives us some evidence of the under-representation of darker skin tones in the ISIC data [45, 11].

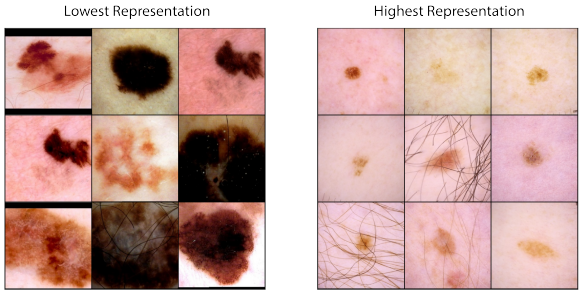


Figure 5: Top 9 images with highest and lowest probability of being sampled based on Eq. (5), indicating the images with the lowest and highest respective representation.

To visualise what individual features are being used to calculate under-representation, we find the 50 latent variables with the highest values of $(\hat{Q}_i(z_i(x)|X) + \alpha)^{-1}$ (most significant to calculating resampling probability) and then gradually perturb the value each of these individually, observing the effect on the decoded reconstruction. To ensure we are perturbing within a valid range, we interpolate between the individual variable and the value of the corresponding variable on a target image. Figure 6 shows that we have identified a disentangled variable within the latent space that encodes skin tone. Since the encoder is optimizing for melanoma classification, this is evidence that the model may be using skin tone as an important latent feature in this task. This provides further precedent for evaluating the use of debiasing techniques on the ISIC data to ensure this protected characteristic does not negatively impact the classification performance of melanoma detection models. The reconstructions are of comparatively low quality because the classification loss acts in opposition to the reconstruction loss. Better reconstruction quality can be achieved over more epochs, however the classification loss collapses after a certain point. We opt for a trade-off between the classification performance and reconstruction quality since we are attempting to show that the classifier is using skin tone as an important feature. We can still demonstrate the existence of the skin tone feature with these imperfect reconstructions.

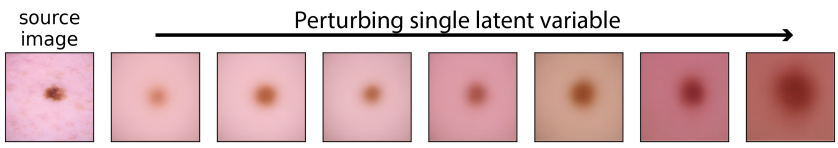


Figure 6: Gradually perturbing the value of a single latent variable in a DB-VAE [2] and decoding to view the impact on the reconstruction.

We use the DB-VAE solely as a bias identification tool and then use ResNeXt-101 with an auxiliary debiasing head for classification and bias removal. Since the ISIC data [45, 11] does not have the skin tone labels required for this method, we attempt to generate these in an automated way.

We calculate the individual typology angle (ITA) of the healthy skin in each image to approximate skin tone [34, 22], given by:

$$ITA = arctan\left(\frac{L-50}{b}\right) \times \frac{180}{\pi}, \tag{6}$$

where L and b are obtained by converting RGB pixel values to the CIELab colour space. We propose a simpler method for isolating healthy skin than the segmentation method used in [34, 22]. Across all skin tones, lesions and blemishes are usually darker than the surrounding skin, and so to select a non-diseased patch of skin, we take 8 samples of 20×20 pixels from around the edges of each image, and use the sample with the highest ITA value (lightest skin tone) as the estimated skin tone, converting ITA to Fitzpatrick skin type using the thresholds from [22], which are modified from the thresholds in [34] to map to the Fitzpatrick scale (see Eq. 8, Appendix A.5). Solely for this calculation, we pre-process each image using black hat morphology to remove hair [42], preventing dark pixels from hairs skewing the calculation. Figure 7 illustrates how the hair removal and healthy skin sampling works. It is clear that even with large lesions with hard to define borders, our method is likely to select a sample of healthy skin.

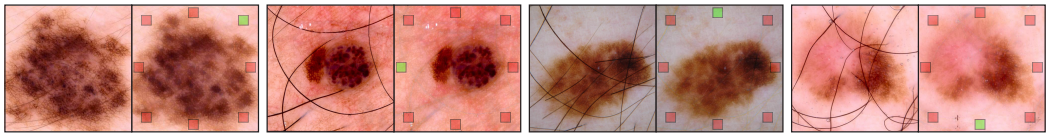


Figure 7: Left of each pair shows ISIC input images, right of each pair shows the placement of the 20×20 pixel samples on images with hair removed. Green square indicates chosen sample based on lightest calculated tone. This simple sampling method eliminates the need for segmentation.

To validate the effectiveness of our skin tone labelling algorithm, we re-label the Fitzpatrick17k data and compare these automated labels against the human annotated skin tones to calculate accuracy, with a correct prediction being within ±1 point on the Fitzpatrick scale [22]. Our method achieves 60.61% accuracy, in comparison to the 53.30% accuracy achieved by the algorithm presented in [22], which segments the healthy skin using a YCbCr masking algorithm. The authors of [22] improve their accuracy to 70.38% using empirically selected ITA thresholds, but we decide against using these to label the ISIC data, given that they are optimised to suit only the Fitzpatrick17k data. We expect our algorithm to perform better on the ISIC data [45, 11] than the Fitzpatrick17k data [22], since the images are less noisy, meaning the assumption that the lightest patch in the image is healthy skin is less likely to be undermined by artefacts or a lightly coloured background.

We use the automated ISIC skin tone labels as the target for debiasing heads. We see performance improvement across the board when debiasing with the TABE head, indicating that this model generalises to the test sets better than the baseline (see Table 5). Performance on the ASAN dataset is of particular interest since these images are known to be from Korean patients and so represent a definitive racial domain shift in comparison to the predominantly Western ISIC training data. Although the origins of the Atlas and MClass clinical data are unknown, these also look to be drawn from significantly different populations to the ISIC data (containing many more examples of darker skin tones), so improvement on these test sets is also evidence of the removal of skin tone bias.

Table 5: Comparison of bias removal techniques using automated **Fitzpatrick skin type** labels, with AUC used as the primary metric. Models are trained using ISIC 2020 & ISIC 2017 data [45, 11].

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

4.4 Ablation studies

Ablation was built into the experimentation process: individual bias removal heads were implemented in isolation before attempting combinations, and debiasing heads were tried with and without gradient

reversal. Using a single head to unlearn instrument bias was found to be more effective than combining this head with artefact bias removal heads.

TABE [1] with and without gradient reversal (named CLGR with gradient reversal) proved successful for different tasks, but ablation of the gradient reversal layer from LNTL [33] generally diminished performance, (see Table 6, full results in Appendix F.4). In our automated skin tone detection algorithm, removing sampling and calculating ITA across the entire image led to the accuracy on the Fitzpatrick17k data [22] decreasing from 60.61% to 56.09%, validating the utility of this addition. Deeper auxiliary heads were experimented with (additional fully connected layer), but didn't have a noticeable impact on performance (see Appendix F.2), which is consistent with the findings in [5].

Table 6: Ablation of gradient reversal from LNTL using **ResNeXt-101** for removal of instrument bias. All scores are **AUC**. Asterisk (*) represents model with gradient reversal ablated.

Experiment	AtlasD	AtlasC	ASANC	MClassD	MClassC
LNTL	**0.804**	**0.612**	**0.768**	0.819	**0.801**
LNTL*	0.783	0.605	0.710	**0.827**	0.747

5 Limitations and future work

One drawback of using these unlearning techniques [33, 1] for artefact debiasing is the need to manually label these artefacts in each training image, however these are quick and easy to identify by untrained individuals. Further research may look to uncover biases caused by other artefacts in a similar manner to [57, 56] and evaluate the effectiveness of unlearning techniques at mitigating these.

Image resolution cannot be universally assumed as a proxy for imaging instrument across all datasets, especially when images are only available in resized form, and so we recommend the actual instrument model be recorded as metadata when collecting training data for melanoma classification.

ITA is an imperfect method for estimating skin tone, given its sensitivity to lighting conditions, and so current labelling algorithms are limited in accuracy. Augmentation of existing data or creation of new samples using generative models to simulate more consistent lighting conditions may be an option for future work. Further work may also look at getting dermatologist annotated skin tone labels for dermoscopic datasets such as the ISIC data and evaluating the effectiveness of debiasing techniques using these more accurate labels. These labels would also allow a more robust evaluation of the existence of skin tone bias in the ISIC data than we were able to provide. A large enough dataset with ground truth skin tone labels may also enable supervised machine learning models to classify the skin tone of future data with high accuracy.

6 Conclusion

This work compares and demonstrates the effectiveness of debiasing methods in the context of skin lesion classification. We successfully mitigate the surgical marking and ruler bias presented in work by Winkler et al. [57, 56] in an automated manner using unlearning techniques. Utilising these techniques could be an alternative to the behaviour change amongst dermatologists suggested by Winkler et al.

We provide evidence of the generalisation benefits of using unlearning techniques, particularly 'Turning a Blind Eye' [1] to remove instrument-identifying information from the feature representation of CNNs trained for the classification of melanoma. We demonstrate this using the ISIC training data [45, 11], with image resolution as a proxy for the imaging instrument. Generalisation tools such as this are potentially powerful for ensuring consistent results across dermatology clinics, and may have utility in the emerging diagnostic app space [44], given that differences between smartphone cameras are likely to introduce spurious variation in a similar manner.

We provide evidence that the skin tone bias shown in [22] can be at least partially mitigated by using skin tone as the target for an auxiliary debiasing head. We subsequently present a novel and effective variation of Kinyanjui et al.'s skin tone detection algorithm [34], and use this to label ISIC data. We use these labels to unlearn skin tone when training on ISIC data and demonstrate some improvements in generalisation, especially when using a 'Turning a Blind Eye' [1] debiasing head. Given that current publicly available data in this field is mostly collected in Western countries for a number of reasons, generalisation and bias removal tools such as these may be important in ensuring these models can be deployed to less represented locations as soon as possible in a fair and safe manner.

References

[1] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, volume 11129, pages 556–572. Springer International Publishing, Cham, 2019.

[2] Alexander Amini, Ava P. Soleimany, Wilko Schwarting, Sangeeta N. Bhatia, and Daniela Rus. Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 289–295, Honolulu HI USA, January 2019. ACM.

[3] Ida Arvidsson, Niels Christian Overgaard, Felicia-Elena Marginean, Agnieszka Krzyzanowska, Anders Bjartell, Kalle Åström, and Anders Heyden. Generalization of prostate cancer classification for multiple sites using deep learning. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 191–194, April 2018.

[4] Alceu Bissoto, Michel Fornaciali, Eduardo Valle, and Sandra Avila. (De) Constructing Bias on Skin Lesion Datasets. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2766–2774, Long Beach, CA, USA, June 2019. IEEE.

[5] Alceu Bissoto, Eduardo Valle, and Sandra Avila. Debiasing Skin Lesion Datasets and Models? Not So Fast. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3192–3201, Seattle, WA, USA, June 2020. IEEE.

[6] Titus J. Brinker, Achim Hekler, Alexander H. Enk, Joachim Klode, Axel Hauschild, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Dirk Schadendorf, Stefan Fröhling, Jochen S. Utikal, Christof von Kalle, and Collaborators. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer (Oxford, England: 1990)*, 111:148–154, April 2019.

[7] Titus J. Brinker, Achim Hekler, Axel Hauschild, Carola Berking, Bastian Schilling, Alexander H. Enk, Sebastian Haferkamp, Ante Karoglan, Christof von Kalle, Michael Weichenthal, Elke Sattler, Dirk Schadendorf, Maria R. Gaiser, Joachim Klode, and Jochen S. Utikal. Comparing artificial intelligence algorithms to 157 German dermatologists: The melanoma classification benchmark. *European Journal of Cancer*, 111:30–37, April 2019.

[8] Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, January 2018.

[9] Stephanie Chan, Vidhatha Reddy, Bridget Myers, Quinn Thibodeaux, Nicholas Brownstone, and Wilson Liao. Machine Learning in Dermatology: Current Applications, Opportunities, and Limitations. *Dermatology and Therapy*, 10(3):365–386, June 2020.

[10] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). *arXiv:1902.03368 [cs]*, March 2019.

[11] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. 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)*, pages 168–172, April 2018.

[12] Karen Kadela Collins, Ryan C. Fields, Dadrie Baptiste, Ying Liu, Jeffrey Moley, and Donna B. Jeffe. Racial Differences in Survival after Surgical Treatment for Melanoma. *Annals of Surgical Oncology*, 18(10):2925–2936, October 2011.

[13] Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C. Halpern, Susana Puig, and Josep Malvehy. BCN20000: Dermoscopic Lesions in the Wild. *arXiv:1908.02288 [cs, eess]*, August 2019.

[14] Gabriela Csurka. A Comprehensive Survey on Domain Adaptation for Visual Applications. In Gabriela Csurka, editor, *Domain Adaptation in Computer Vision Applications*, Advances in Computer Vision and Pattern Recognition, pages 1–35. Springer International Publishing, Cham, 2017.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.

[16] Chris Deotte. SIIM-ISIC Melanoma Classification - JPEG Melanoma 256x256. https://www.kaggle.com/cdeotte/jpeg-melanoma-256x256.

[17] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, February 2017.

[18] Thomas B. Fitzpatrick. The Validity and Practicality of Sun-Reactive Skin Types I Through VI. *Archives of Dermatology*, 124(6):869–871, June 1988.

[19] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. In Gabriela Csurka, editor, *Domain Adaptation in Computer Vision Applications*, pages 189–209. Springer International Publishing, Cham, 2017.

[20] Sixue Gong, Xiaoming Liu, and Anil K. Jain. Jointly De-biasing Face Recognition and Demographic Attribute Estimation. *arXiv:1911.08080 [cs]*, July 2020.

[21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[22] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset. *arXiv:2104.09957 [cs]*, April 2021.

[23] Yanyang Gu, Zongyuan Ge, C. Paul Bonnington, and Jun Zhou. Progressive Transfer Learning and Adversarial Domain Adaptation for Cross-Domain Skin Disease Classification. *IEEE Journal of Biomedical and Health Informatics*, 24(5):1379–1393, May 2020.

[24] Hao Guan and Mingxia Liu. Domain Adaptation for Medical Image Analysis: A Survey. *arXiv:2102.09508 [cs, eess]*, February 2021.

[25] Qishen Ha, Bo Liu, and Fuxu Liu. Identifying Melanoma Images using EfficientNet Ensemble: Winning Solution to the SIIM-ISIC Melanoma Classification Challenge. *arXiv e-prints*, 2010:arXiv:2010.05351, October 2020.

[26] H.A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. Ben Hadj Hassen, L. Thomas, A. Enk, L. Uhlmann, Christina Alt, Monika Arenbergerova, Renato Bakos, Anne Baltzer, Ines Bertlich, Andreas Blum, Therezia Bokor-Billmann, Jonathan Bowling, Naira Braghiroli, Ralph Braun, Kristina Buder-Bakhaya, Timo Buhl, Horacio Cabo, Leo Cabrijan, Naciye Cevic, Anna Classen, David Deltgen, Christine Fink, Ivelina Georgieva, Lara-Elena Hakim-Meibodi, Susanne Hanner, Franziska Hartmann, Julia Hartmann, Georg Haus, Elti Hoxha, Raimonds Karls, Hiroshi Koga, Jürgen Kreusch, Aimilios Lallas, Pawel Majenka, Ash Marghoob, Cesare Massone, Lali Mekokishvili, Dominik Mestel, Volker Meyer, Anna Neuberger, Kari Nielsen, Margaret Oliviero, Riccardo Pampena, John Paoli, Erika Pawlik, Barbar Rao, Adriana Rendon, Teresa Russo, Ahmed Sadek, Kinga Samhaber, Roland Schneiderbauer, Anissa Schweizer, Ferdinand Toberer, Lukas Trennheuser, Lyobomira Vlahova, Alexander Wald, Julia Winkler, Priscila Wölbing, and Iris Zalaudek. Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8):1836–1842, August 2018.

[27] Holger Haenssle, Christine Fink, Ferdinand Toberer, J. Winkler, Wilhelm Stolz, Teresa Deinlein, Rainer Hofmann-Wellenhof, S. Emmert, Timo Buhl, M. Zutt, A. Blum, M.S. Abassi, Luc Thomas, Isabelle Tromme, Philipp Tschandl, A. Enk, Albert Rosenberger, Christina Alt, and Pascale Zukervar. Man against machine reloaded: Performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. *Annals of Oncology*, 31:137–143, January 2020.

[28] Seung Seog Han, Myoung Shin Kim, Woohyung Lim, Gyeong Hun Park, Ilwoo Park, and Sung Eun Chang. Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. *Journal of Investigative Dermatology*, 138(7):1529–1538, July 2018.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE.

[30] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, Honolulu, HI, July 2017. IEEE.

[31] Philip T. Jackson, Stephen Bonner, Ning Jia, Christopher Holder, Jon Stonehouse, and Boguslaw Obara. Camera Bias in a Fine Grained Classification Task. *arXiv:2007.08574 [cs]*, July 2020.

[32] M. H. Jafari, N. Karimi, E. Nasr-Esfahani, S. Samavi, S. M. R. Soroushmehr, K. Ward, and K. Najarian. Skin lesion segmentation in clinical images using deep learning. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 337–342, December 2016.

[33] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning Not to Learn: Training Deep Neural Networks With Biased Data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9004–9012, Long Beach, CA, USA, June 2019. IEEE.

[34] Newton M Kinyanjui, Timothy Odonga, Celia Cintas, Noel C F Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R Varshney. Estimating Skin Tone and Effects on Classification Performance in Dermatology Datasets. In *Fair ML for Health*, page 10, Vancouver, Canada, 2019. NeurIPS.

[35] Patel Krut. Convolutional Neural Networks — A Beginner's Guide | by Krut Patel | Towards Data Science. https://towardsdatascience.com/convolution-neural-networks-a-beginners-guide-implementing-a-mnist-hand-written-digit-8aa60330d022.

[36] Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex C. Kot. Domain Generalization for Medical Imaging Classification with Linear-Dependency Regularization. *arXiv:2009.12829 [cs, eess]*, October 2020.

[37] Yuexiang Li and Linlin Shen. Skin Lesion Analysis towards Melanoma Detection Using Deep Learning Network. *Sensors (Basel, Switzerland)*, 18(2), February 2018.

[38] Peter A. Lio and Paul Nghiem. Interactive Atlas of Dermoscopy: Giuseppe Argenziano, MD, H. Peter Soyer, MD, Vincenzo De Giorgio, MD, Domenico Piccolo, MD, Paolo Carli, MD, Mario Delfino, MD, Angela Ferrari, MD, Rainer Hofmann-Wellenhof, MD, Daniela Massi, MD, Giampiero Mazzocchetti, MD, Massimiliano Scalvenzi, MD, and Ingrid H. Wolf, MD, Milan, Italy, 2000, Edra Medical Publishing and New Media. 208 pages. $290.00. ISBN 88-86457-30-8.CD-ROM requirements (minimum): Pentium 133 MHz, 32-Mb RAM, 24X CD-ROM drive, 800 × 600 resolution, and 16-bit color graphics capability. Test system: Pentium III 700 MHz processor running Microsoft Windows 98. Macintosh compatible only if running Windows emulation software. *Journal of the American Academy of Dermatology*, 50(5):807–808, May 2004.

[39] Jayawant N. Mandrekar. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, September 2010.

[40] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science (New York, N.Y.)*, 366(6464):447–453, October 2019.

[41] Erdem Okur and Mehmet Turkan. A survey on automated melanoma detection. *Engineering Applications of Artificial Intelligence*, 73:50–67, August 2018.

[42] Vatsal Parsaniya. Melanoma Hair Remove. https://www.kaggle.com/vatsalparsaniya/melanoma-hair-remove, 2020.

[43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[44] Cédric Rat, Sandrine Hild, Julie Rault Sérandour, Aurélie Gaultier, Gaelle Quereux, Brigitte Dreno, and Jean-Michel Nguyen. Use of Smartphones for Early Detection of Melanoma: Systematic Review. *Journal of Medical Internet Research*, 20(4):e9392, April 2018.

[45] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, Allan Halpern, Brian Helba, Harald Kittler, Kivanc Kose, Steve Langer, Konstantinos Lioprys, Josep Malvehy, Shenara Musthaq, Jabpani Nanda, Ofer Reiter, George Shih, Alexander Stratigos, Philipp Tschandl, Jochen Weber, and H. Peter Soyer. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, 8(1):34, January 2021.

[46] Sumit Saha. A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53, December 2018.

[47] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, October 2017.

[48] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences. *arXiv:1704.02685 [cs]*, October 2019.

[49] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

[50] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826. IEEE Computer Society, June 2016.

[51] Mingxing Tan and Quoc V Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36th International Conference on Machine Learning*, 97:6105–6114, 2019.

[52] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous Deep Transfer Across Domains and Tasks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4068–4076, December 2015.

[53] Avinash Sharma V. Understanding Activation Functions in Neural Networks, March 2017.

[54] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial Faces in-the-Wild: Reducing Racial Bias by Information Maximization Adaptation Network. *arXiv:1812.00194 [cs]*, July 2019.

[55] K. Westerhoff, W. H. Mccarthy, and S. W. Menzies. Increase in the sensitivity for melanoma diagnosis by primary care physicians using skin surface microscopy. *British Journal of Dermatology*, 143(5):1016–1020, 2000.

[56] Julia K Winkler. Association between different scale bars in dermoscopic images and diagnostic performance of a market-approved deep learning convolutional neural network for melanoma recognition. *European Journal of Cancer*, page 9, 2021.

[57] Julia K. Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, and Holger A. Haenssle. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatology*, 155(10):1135, October 2019.

[58] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Appendix

This section presents material that can be referenced to enhance the readers' understanding of the details of the project. Supplementary background information relating to machine learning, convolutional neural networks and how these concepts intersect with dermatology is presented in Appendix A. Additional background information relating to the tasks and additional implementation information is also presented in Appendix A. Please refer to the cited sources for a more in-depth explanation of concepts. Schematic figures of the models used in the project are presented in Section B. Samples of each training and test dataset are illustrated in Appendix C, to give a feel for the images present in each. We plot the class distributions of the primary and auxiliary targets in Appendix D. Hyperparameter tuning results are presented in Appendix E. Additional experimental results in the form of ROC curves and tables that were not included in the main report can be found in Appendix F. Our attempt at interpreting artefact bias using vanilla gradient saliency maps [49] is presented in Section F.1.1. A reflective personal statement can be found in Section G.

A Background

A.1 Deep learning

Machine learning is a branch of computer science and artificial intelligence which aims to create algorithms that learn through exposure to data, from which patterns are identified in order to draw inferences. A particular subset of machine learning named 'deep learning' has become the state of the art solution for many machine learning tasks. As described by Goodfellow et al. [21], "Deep learning is a particular kind of machine learning that achieves great power and flexibility by learning to represent the world as a nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones".

A standard problem in traditional machine learning is selecting the correct set of features, or characteristics, to be provided to the algorithm, and engineers have hence spent much time doing 'feature engineering' to get the best out of their models. Deep learning leverages an approach known as 'representation learning', which essentially means the algorithm learns the optimal set of features (i.e., the feature representation), mostly eliminating the manual feature engineering required in more traditional machine learning approaches. These learned representations are often more informative to the model than manually engineered features, which contributes to the success of deep learning models. Deep learning algorithms express representations as other simpler representations: "Deep learning allows the computer to build complex concepts out of simpler concepts" [21]. The standard example of deep learning is a multilayer perceptron (MLP). "A multilayer perceptron is just a mathematical function mapping some set of input values to output values" [21]. An MLP is a network of 'neurons' (computational units) that take weighted inputs, as well as a weighted bias, and apply an activation function to produce an output. The activation function is usually a non-linear function (see ReLU, Sigmoid, Tanh [53]) that maps the weighted neuron inputs to the neuron output. The neuron output is calculated as:

$$f\left(b + \sum_{i=1}^{n} x_i w_i\right), \tag{7}$$

where f is the activation function, b is the bias value, x_i is an input from the previous layer and w_i is it's corresponding weight. These neurons are arranged into a network as shown in Figure 8, with an input layer (first layer), output layer (last layer), and so called hidden layers in between. The input layer is used to pass the input values to the first hidden layer, and these neurons do not apply an activation function. The output layer uses a suitable activation function to output values in the format required for the task, with each node representing a single class. The Sigmoid function can be used to output a single value between 0 and 1 for binary classification tasks; a threshold (often 0.5) can be applied to this value to obtain a 0 or 1 classification. The Softmax function can be applied to multi-class classification problems to output a probability for each class, with the probabilities across all classes summing to one. Increasing the number of neurons per layer (width) and number of layers (depth) of the hidden layers can often lead to better performance on more complex tasks, but as width and depth increase, computational cost also increases. To train a network, an input is passed forward through the network, activating neurons as it passes through, eventually resulting in an output. The output prediction is compared to the ground truth using a loss function such as cross entropy to produce a loss value. This loss value is then 'backpropegated' back through the network using the chain rule, to update the weights and biases accordingly.

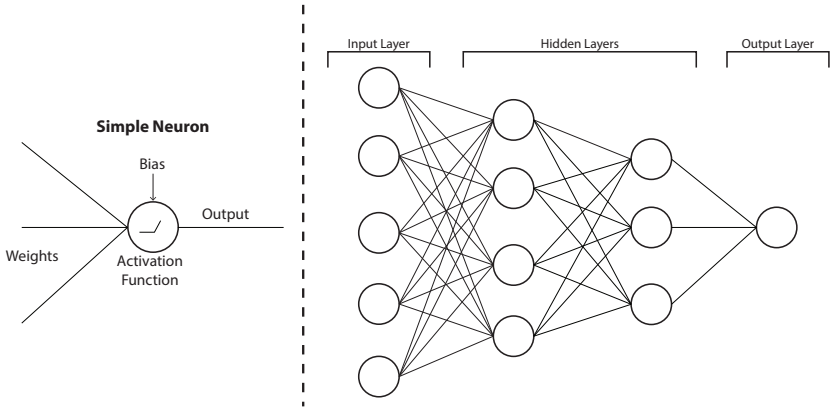


Figure 8: Multi-layer perception deep learning model simplified schematic.

Whilst images can be flattened into a one dimensional vector and passed into a multilayer perceptron, this means that spacial and temporal information is not utilised and so performance is often poor, especially for complex images and tasks. This is where convolutional neural networks (CNN) come in. CNNs are a subset of deep learning models, commonly used in computer vision and image processing, achieving state of the art performance in many domains. CNNs can be defined as: "simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers" [21]. These networks can take matrices/tensors as input, meaning images can be passed directly into the input layer. Convolutions work by sliding a kernel over the input image, performing matrix multiplication at each step, and mapping these values to a new feature map (see Figure 9). Kernels are then applied in the same fashion to these feature maps at the next layer and so on. The stride dictates how many pixels the kernel shifts each time it moves, and so increasing the stride increases down-sampling.

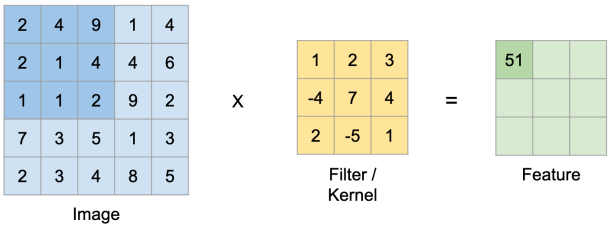


Figure 9: Convolution operation, kernel size 3 and stride of 1. Image courtesy of [35].

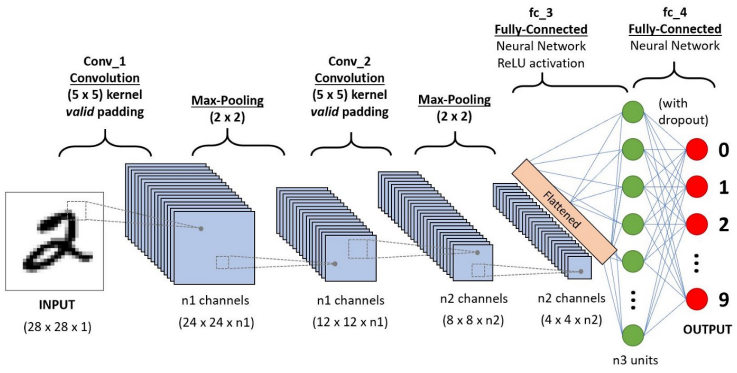


Figure 10: CNN model used for handwritten digit classification. Image courtesy of [46].

Several kernels are usually used at each layer, resulting in several feature maps, as seen in Figure 10. The elements in each kernel are the weights that are updated during backpropegation, and thus these

filters learn to extract useful feature maps. To get a classification output from the extracted feature maps, these are often flattened into a vector of values and passed into a fully connected layer (like the MLP above). As described above for the MLP, the output layer applies an activation function such as sigmoid or softmax to calculate the probabilities of each output class, allowing classification to be made based on a selected threshold.

Over time, the sophistication of convolutional neural networks has increased, with new architectures being proposed continually, and many state of the art solutions arising from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [15]. Some of the most successful CNN architectures we identified as potentially effective feature extractors for melanoma classification are: ResNet [29], ResNeXt [58], DenseNet [30] and Inception-v3 [50]. The performance of each architecture is data dependent and so we choose to compare the performance of each before deciding a champion. These architectures are briefly summarised below, but the corresponding papers should be referenced for full details:

- **EfficientNet, 2019**: Utilises compound scaling method to scale up CNNs efficiently with less parameters compared to previous SOTA solutions [51].
- **ResNet, 2016**: Solves the degradation problem using residual skip connections to allow effective training of deeper models [29].
- **Inception, 2016**: Combining multiple kernel sizes with 1×1 convolutions on each level to extract features at different scales [50].
- **ResNeXt, 2017**: Extension of ResNet which, inspired by the inception block, integrates branched paths within the residual blocks of the network [58].
- **DenseNet, 2017**: Allowing feature maps from all previous layers to be leveraged by every successive layer in the network using dense connections that forward concatenated feature maps through the network [30].

A.2 Deep learning for skin lesion classification

The task of classifying skin lesions using machine learning has been worked on since as early as 1988, initially using traditional machine learning methods such as decision trees in combination with segmentation [41]. Originally, lack of model sophistication, compute power and quality data meant that performance was not at the level of dermatologists. Like many areas of computer vision, the rise of convolutional neural networks and ever increasing compute power has seen model performance rapidly increase to the point where there is evidence of machine learning matching or surpassing dermatologists at the task [6, 26]. The power of deep learning to extract features has meant many modern models perform best without segmentation, and often use information in the surrounding skin in the classification task [4].

Skin diseases can be separated into many classes (see Figure 11). On the most granular scale, skin diseases can be separated into neoplastic and non-neoplastic conditions. A neoplastic condition is an abnormal growth of cells known as a tumour, while a non-neoplastic skin condition refers to any other type of skin condition. We focus on neoplastic lesions in this work. These neoplastic lesions can be separated into benign (non-cancerous) and malignant (cancerous), which is a very important classification to make, since cancerous tissue has the ability to invade the rest of the body and ultimately cause fatality. On a more fine-grained level, lesions may be classified by specific disease, such as cyst, basal cell carcinoma or melanoma. In terms of classification in machine learning, it is possible to use specific diseases as classes for prediction, allowing malignancy to be also inferred from this classification [25]. We opt for the more common binary approach of classifying using benign/malignant as classes.

A.3 Additional background on artefacts bias

While this project attempts to remove artefact bias using an unlearning approach, previous work has focused on other techniques such as segmentation, which was used to remove artefacts from the input of skin lesion classification models [32], however this is not commonly used at the time of writing, given that deep learning models may utilise information in the surrounding skin [56, 4] and so removing this can impact performance. Some biases are also not separable from the lesion by image region, for example surgical markings that are on the lesion itself. Finally, segmentation masks are expensive and so not widely available with current training datasets; they must be created manually by experienced dermatologists, and even these are imperfect. Automated debiasing within the model, without pre-processing, is a much more desirable prospect.

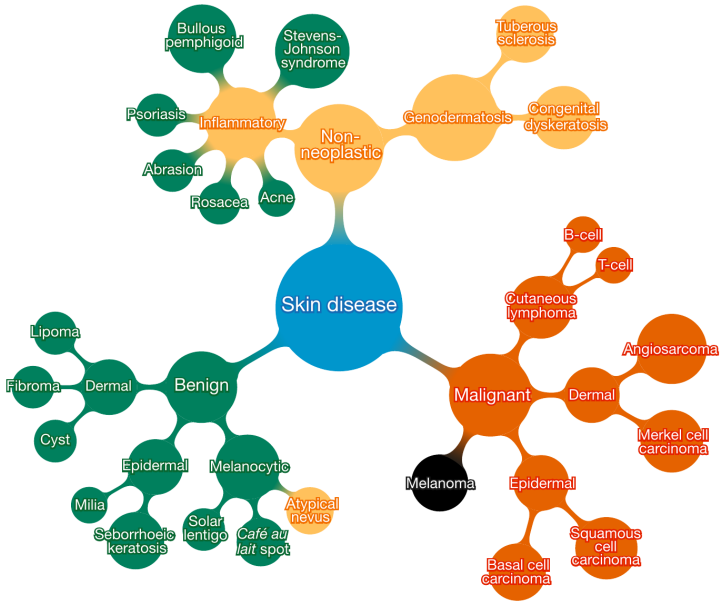


Figure 11: A schematic illustration of the taxonomy of skin diseases. Red indicates malignant, green indicates benign, orange indicates conditions that can be either and black indicates melanoma. Image courtesy of [17].

A.4 Additional background on domain generalisation

LNTL [33] and TABE [1] are adapted versions of [19, 52] respectively, which are both domain adaptation/generalisation methods. These papers were the inspiration behind attempting to use LNTL and TABE for domain generalisation.

A.5 Additional background on skin tone bias

Racial bias is a prominent problem in many areas of machine learning [8, 40, 22], often caused by datasets that are imbalanced or that reflect biases embedded within society. For example, current facial recognition technology was shown to perform poorly for classifying the gender of darker skinned individuals in comparison to lighter skinned individuals in [8]. Healthcare management algorithms in the USA were shown to exhibit bias against black patients in [40].

In order to begin quantifying racial/skin tone bias, it's often important to bin human skin tones into categories. The Fitzpatrick sun reactive skin type classification system [18] is a longstanding skin tone classification scale, which separates skin tones based on their reaction to exposure to sunlight. This system is widely accepted as the gold standard amongst dermatologists [8], and separates all human skin tones into six categories, with one being the lightest. Whilst this system is among the best devised to date, it's easy to see from Figure 12 how images may be misclassified given the small difference between each type, especially when there is varying lighting conditions for each image.

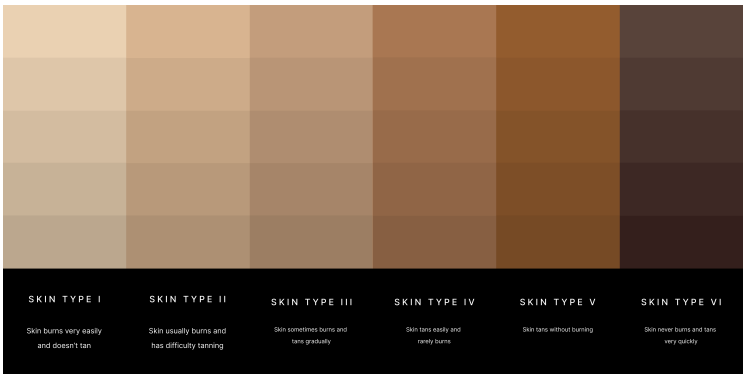


Figure 12: Visualisation of the Fitzpatrick 6 point scale [18]. Image courtesy of Groh et al. [22].

Equation 8 shows the thresholds set out in [22], which are taken from [34] and modified to fit the Fitzpatrick 6 point scale [18]. We use these thresholds in our skin tone labelling algorithm. Whilst these are also further adjusted to optimise performance on the Fitzpatrick17k data in [22], we choose not to utilise these since they are optimised specifically to that dataset and so are likely not generalisable.

$$Fitzpatrick(ITA) = \begin{cases} 1 & ITA > 55 \\ 2 & 55 \geq ITA > 41 \\ 3 & 41 \geq ITA > 28 \\ 4 & 28 \geq ITA > 19 \\ 5 & 19 \geq ITA > 10 \\ 6 & 10 \geq ITA \end{cases} \tag{8}$$

A.6 Metrics

Sensitivity (recall) is a measure of the proportion of the positive class that was correctly classified. Specificity is the proportion of the negative class that was correctly identified. These two metrics are defined as:

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

These are regularly used as metrics in the medical sciences, since it's important to both identify disease (leading to correct treatment) and rule disease out (preventing unnecessary treatment). In order to use these metrics, a threshold must be set at which the output of a model (between 0 and 1) is taken as a positive or negative classification. The default position for this threshold is 0.5, but this threshold may also be adjusted towards finding an acceptable trade-off between true positive, true negative, false positive and false negative predictions. Analysing the receiver operating characteristic (ROC) curve is a very useful way of finding this threshold, as it visualises how sensitivity and specificity vary over every possible threshold (see Figure 32 for example ROC curves).

The area under this curve (AUC) can hence be used as a single robust metric to evaluate the performance of a model where sensitivity and specificity are important, and where the threshold is open to adjustment (such as melanoma classification) [39]. We avoid relying on accuracy, sensitivity and specificity in this work, since these all rely on the assumption of a selected threshold, and instead use AUC as the primary metric, also plotting the ROC curves. We anticipate that the threshold would then be selected by a medical professional to suit their desired level of sensitivity and specificity. An AUC of 1 means the classifier is distinguishing positive and negative classes perfectly, and an AUC of 0.5 equates to random chance. Anything less than 0.5 and there may be an issue with the model or data labelling, since the model is actively predicting the wrong classes; in fact, inverting the data labels in this case would result in an AUC of over 0.5. We also avoid relying on accuracy due to the imbalaance between benign/malignant lesions in the test sets meaning accuracy is not as descripitive of performance as AUC.

A.7 Additional implementation details

For the main classification head we use PyTorch's 'BCEWithLogitsLoss' because our primary classification is always binary, while for the auxiliary heads we use PyTorch's 'CrossEntropyLoss', since the auxiliary target is sometimes multi-class. These functions combine the sigmoid layer with cross entropy in a single class, taking advantage of the log-sum-exp trick for numerical stability, and so directly take logits as input.

To help combat the undesirable effects of class imbalance, the auxiliary classification loss function is weighted as defined by the following equation:

$$\mathcal{W}_n, c = \frac{1}{\text{Number of samples in Class c}}. \tag{9}$$

The loss is not weighted for the primary classification head after experimentation showed it to have a detrimental impact on performance.

We took several measures to ensure reproducibility. Firstly, all random seeds (random, numpy, torch and cuda) are set to a default value, and cudnn set to deterministic. All code for the project is

annotated and made available, and the specific commands for each experiment in the project are also provided in the README. Weights are not provided due to their file size, but these can be provided on request.

B Model architecture schematics

In this section we present the generic architecture schematics for the models used in the project. Refer to Section 3 and the cited papers for full explanation of concepts. Figure 13 shows an overview of the 'Learning Not to Learn' architecture [33] (see Section 3.1). Figure 14 shows an overview of the 'Turning a Blind Eye' architecture [1] (see Section 3.2). The feature extractor , f, in both of these schematics is simplified and in the reality of our project is implemented as a more complex architecture such as ResNext-101 [58] or EfficientNet-B3 [51]. The blocks labelled 'fc' represent a fully connected layer(s). Figure 15 shows a schematic of the debiasing variational autoencoder model [2] (see Section 3.3), used as a tool to uncover skin tone bias in Section 4.3.

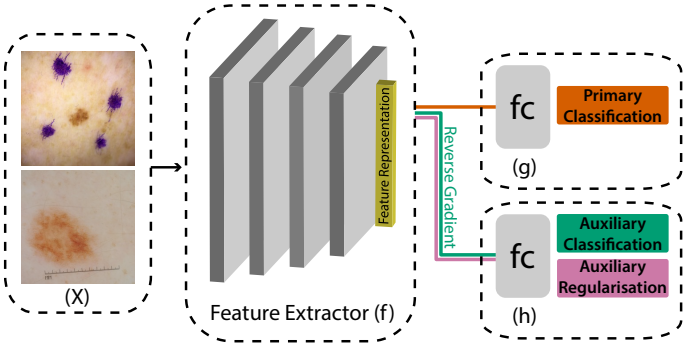


Figure 13: Overview of 'Learning Not to Learn' [33]. Feature extractor, f, is implemented as a convolutional architecture such as ResNeXt or EfficientNet in this work. 'fc' denotes a fully connected layer.

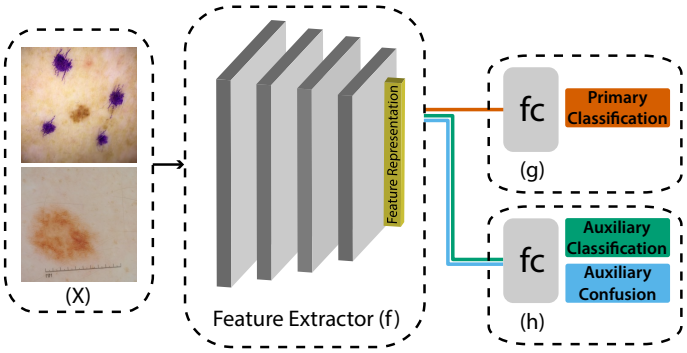


Figure 14: Overview of 'Turning a Blind Eye' [1]. Feature extractor, f, is implemented as a convolutional architecture such as ResNeXt or EfficientNet in this work. 'fc' denotes a fully connected layer.

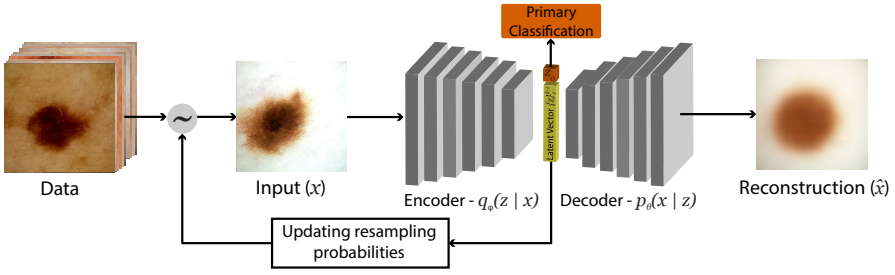
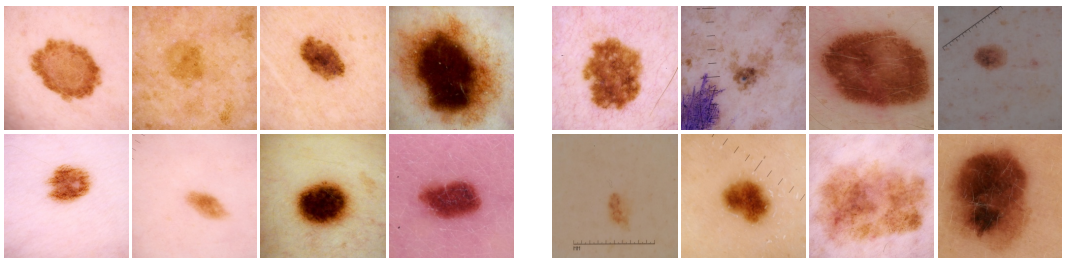


Figure 15: Debiasing Variational Autoencoder (DB-VAE) [2] generic architecture.

C Examples of data

Figure 16 shows a sample of the images from the ISIC dermoscopic training data [45, 11], including some examples of surgical markings and some examples of rulers. Figure 17 shows a sample of the Fitzpatrick17k clinical dataset [22]. It's clear that the Fitzpatrick dataset is much more noisy than the ISIC, with differing levels of zoom and some visible backgrounds.



(a) Benign. (b) Malignant.

Figure 16: Example images from the ISIC dermoscopic training set [45, 11].

(a) Benign. (b) Malignant.

Figure 17: Example images from the Fitzpatrick17k training set [22].

Figure 18 shows a sample of the 'Heid Plain' images from Heidelberg University [57]. These are dermoscopic images collected by the university of a variety of neoplastic lesions. Figure 19 shows a sample of the 'Heid Marked' images from Heidelberg university [57]. These are the same lesions from 'Heid Plain', but with surgical markings either applied in vivo (physically applied and images recaptured), or electronically superimposed. Figure 20 shows a sample of the 'Heid Ruler' images, which was made by electronically superimposing rulers onto the 'Heid Plain' images.

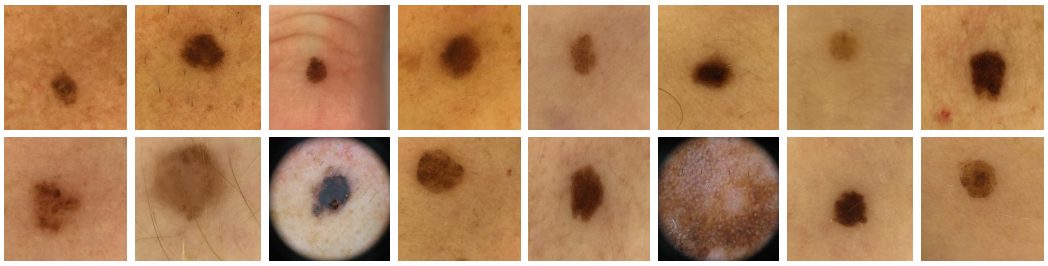


Figure 18: Example images from the Heidelberg University training set with no artefacts [57].

Figure 21 shows a sample of the 'Interactive Atlas of Dermoscopy' [38] *dermoscopic* images, while Figure 22 shows the equivalent *clinical* images from the same set. The domain shift between clinical and dermoscopic images is clearly illustrated: the skin/lesion can be seen in more detail in the dermoscopic images due to the reduction in surface shine.

Figure 23 shows a sample of the ASAN [28] clinical test set. This dataset is collected from the ASAN medical centre, Seoul, South Korea and so features predominantly South Korean patients. The dataset contains some images of the same lesion several times at different levels of zoom, but is generally less noisy than the Fitzpatrick17k data.

Figure 24 shows a sample of the MClass [7] dermoscopic benchmark test set, and Figure 25 shows a sample the MClass clinical benchmark test set. Both of these test sets were sent to a number of experienced dermatologists (157 for dermosocpic images, 145 for clinical images), who attempted

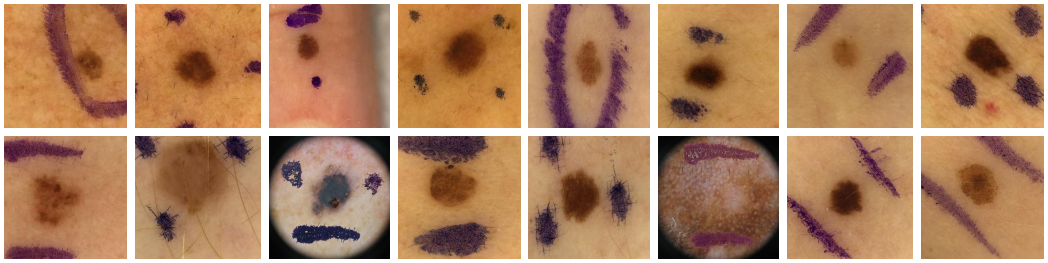


Figure 19: Example images from the Heidelberg University training set with surgical markings [57].

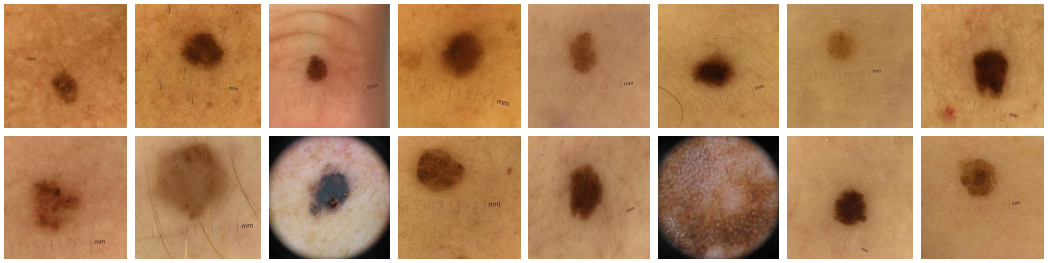


Figure 20: Example images from the Heidelberg University training set with superimposed rulers [57].

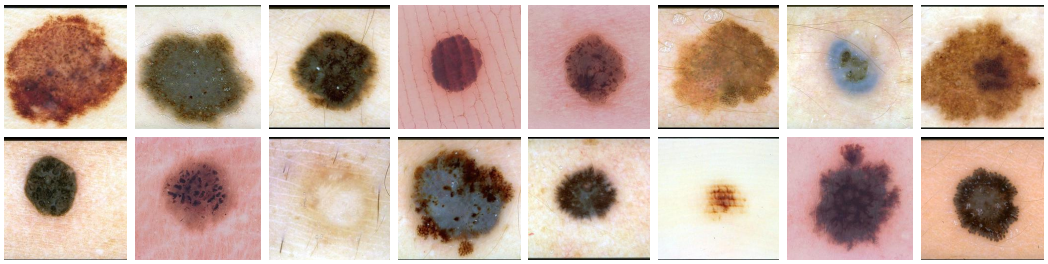


Figure 21: Example images from the Interactive Atlas of Dermoscopy dermoscopic test set [38].

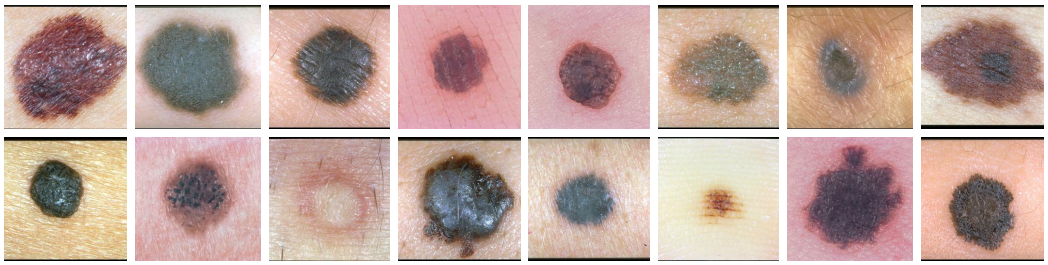


Figure 22: Example images from the Interactive Atlas of Dermoscopy clinical test set [38].

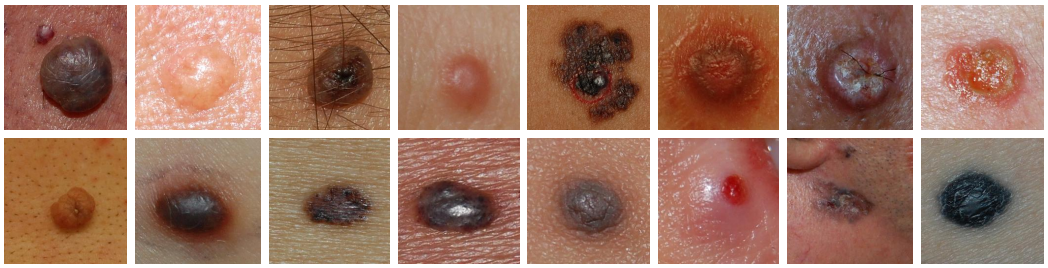


Figure 23: Example images from the ASAN clinical test set [28].

to classify the images, with AUC scores reported in [7]. Since true AUC cannot be calculated for dichotomous human predictions (we cannot adjust the threshold of human predictions), the authors use the average of sensitivity and specificity as a pseudo AUC score.

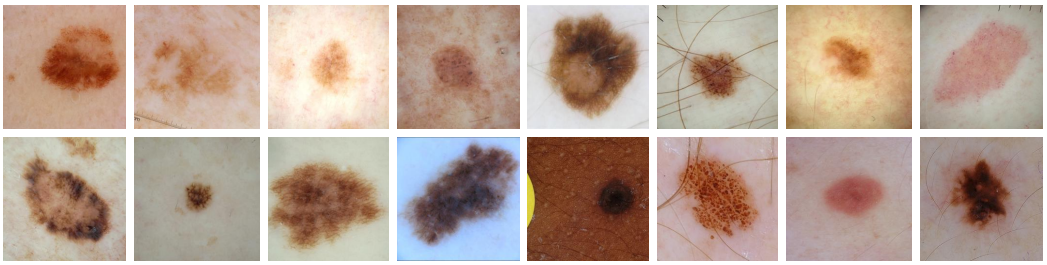
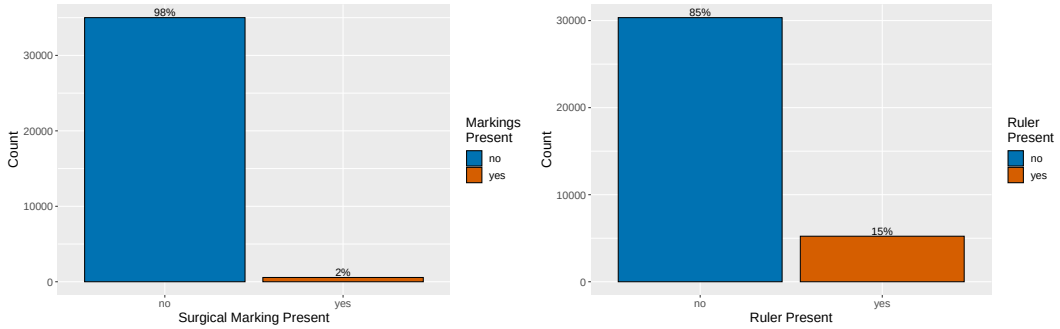


Figure 24: Example images from the MClass dermoscopic test set [7].

Figure 25: Example images from the MClass clinical test set [7].

D Dataset distributions

Figure 26 shows the class distribution for the surgical marking and ruler labels. The distribution of artefacts is highly imbalanced, pointing to why weighted loss functions were needed to stabilise training.

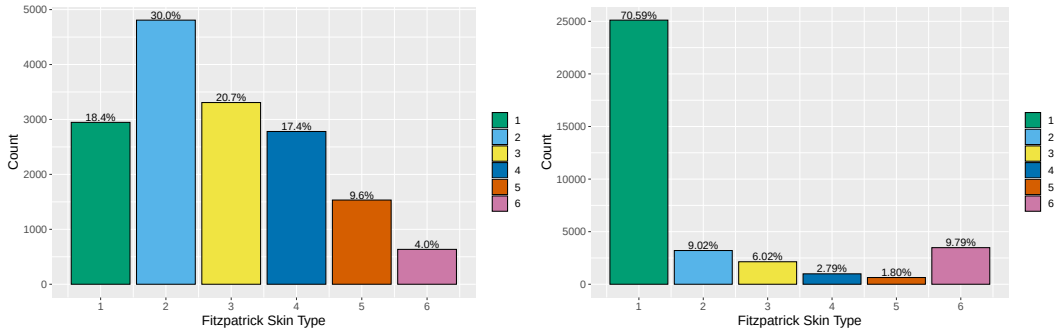


(a) Surgical marking distribution

(b) Ruler distribution

Figure 26: Class distribution of artefacts in ISIC 2020 & 2017 training data [45, 11].

The distribution of Fitzpatrick skin types in the neoplastic subset of the Fitzpatrick17k training images [22] is shown in Figure 27a. The data is slightly skewed towards lighter skin tones, with a very low representation of the darkest skin tone (type 6). Figure 27b shows the distribution of Fitzpatrick skin types in the ISIC training data [45, 11], as labelled by our automated labelling algorithm. The labelling suggests this data is much more skewed towards light skin, with the overwhelming majority (71%) of images labelled as type 1 skin. The relatively high number of type 6 classifications is likely due to failures of the labelling algorithm, perhaps picking up on dark lighting conditions, since upon visual inspection of the dataset it's clear there is not likely to be this many type 6 skin images in the dataset.



(a) Fitzpatrick17k training data [22].

(b) ISIC training data [45, 11].

Figure 27: Distribution of Fitzpatrick skin types in ISIC [45, 11] and Fitzpatrick17k [22] training data .

Figure 28 shows the distribution of image resolutions in the ISIC dataset, following omission of outlier classes. As suggested by the ISIC, these image resolutions can be used as a proxy for the imaging instrument used to capture the image.

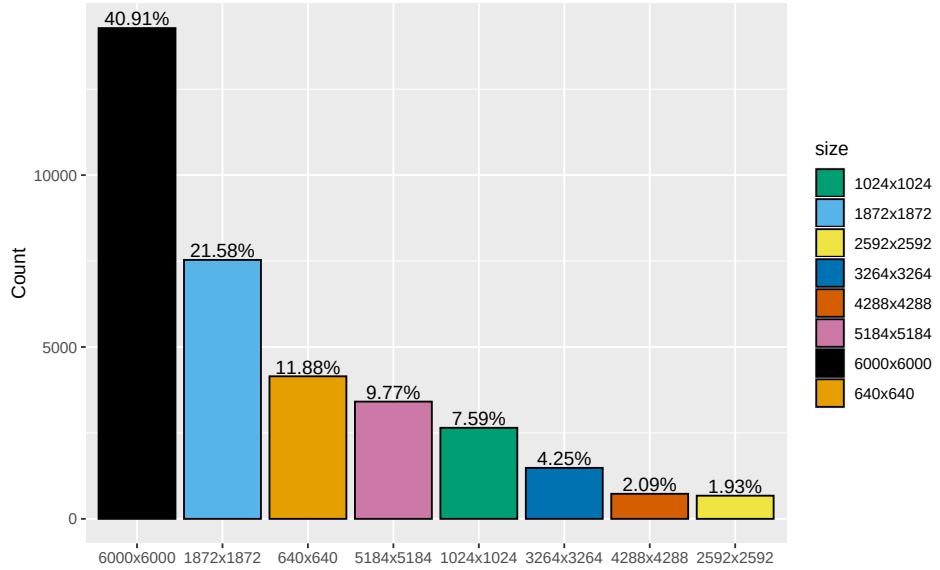


Figure 28: Class distribution of instruments in ISIC 2020/2017 combined data [45, 11]. Instruments inferred as separate through image resolution.

E Hyperparameter tuning

Below are the results of the grid search used to select learning rate and momentum, searching between 0.03 and 0.00001 for learning rate and 0 to 0.9 for momentum. We tune the baseline ResNeXt-101 model and also use these hyperparameters for the debiasing models to maximise cross-comparability. Whilst we used 5-fold cross validation for choosing the number of epochs, this wasn't computationally feasible for the grid search, and so the ISIC 2019 competition data [13] was used as a validation set. With more time and computational resources we could have optimised the number of epochs at the same time as these hyperparameters. In hindsight, perhaps a random search rather than a brute force grid search would have allowed more exhaustive tuning within the computational limitations.

Table 7: Hyperparamter tuning of baseline ResNeXt-101 model, trained for 4 epochs and using a random subset of ISIC 2019 data [13] as the validation data.

LR	Mom	AUC	LR	Mom	AUC	LR	Mom	AUC	LR	Mom	AUC
0.03	0	0.807	0.03	0.3	0.824	0.03	0.6	0.800	0.03	0.9	0.789
0.01	0	0.825	0.01	0.3	0.826	0.01	0.6	0.837	0.01	0.9	0.834
0.003	0	0.837	0.003	0.3	0.852	0.003	0.6	0.854	0.003	0.9	0.848
0.001	0	0.815	0.001	0.3	0.820	0.001	0.6	0.826	0.001	0.9	0.843
0.0003	0	0.783	0.0003	0.3	0.798	0.0003	0.6	0.809	**0.0003**	**0.9**	**0.866**
0.0001	0	0.681	0.0001	0.3	0.727	0.0001	0.6	0.770	0.0001	0.9	0.814
0.00003	0	0.469	0.00003	0.3	0.524	0.00003	0.6	0.627	0.00003	0.9	0.783
0.00001	0	0.398	0.00001	0.3	0.409	0.0001	0.6	0.445	0.00001	0.9	0.681

Table 8 shows the chosen number of epochs for training each architecture for each dataset. These were chosen as the point at which the AUC reached it's maximum or plateaued.

Table 8: Optimal number of epochs for training, selected through analysis of cross validation curves.

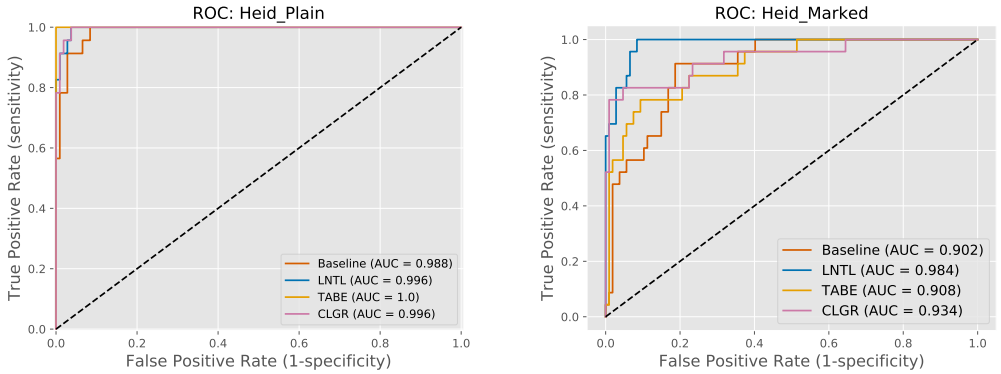
Training dataset	Architecture	Epochs
ISIC	EfficientNet-B3	15
ISIC	ResNet-101	6
ISIC	ResNeXt-101	4
ISIC	Inception-v3	5
ISIC	DenseNet	6
Fitzpatrick17k	ResNeXt-101	30

F Additional results

F.1 Artefact bias removal

To label the artefacts in the taining data, we attempt to use colour thresholding to automatically label both surgical markings and rulers. We set the script to separate the images into different directories for inspection. This method is somewhat successful for surgical markings, however by looking at the images labelled unmarked, we see that some are not picked up, and so we also go through and manually pick out surgical markings. This method does not work well at all for labelling rulers, perhaps since hairs have similar pixel values to rulers, and so we manually label each image for rulers. The manual labelling process is not difficult to the human eye since these artefacts are quite obvious and so this can be done quickly and accurately.

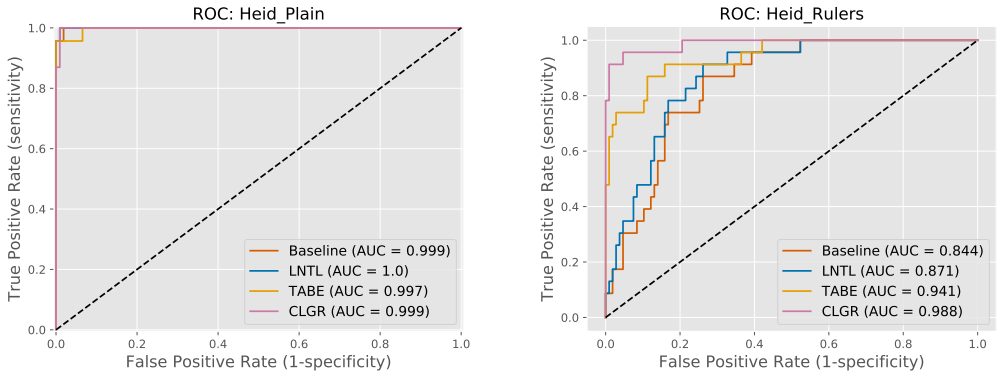
Figure 29 shows the ROC plots from the surgical marking bias experiments. All models perform almost perfect classification of the easy unbiased test set (Figure 29a). The biased set with surgical markings present causes performance to drop for all models, however it's clear from Figure 29b that the debiasing models are more robust than the baseline, especially LNTL, which retains close to the same AUC score across both test sets. Similarly, the introduction of rulers into the lesion images also causes a drop in the performance of all models (see Figure 30b). The baseline is again the most affected by this bias, with TABE clearly most robust to it.



(a) No surgical markings present.

(b) Surgical markings present.

Figure 29: Comparison between model performances with no surgical markings present (left) vs with surgical markings present (right). **EfficientNet-B3** trained on ISIC 2020 & 2017 data [45, 11], skewed to dm=20.



(a) No rulers present.

(b) Rulers present.

Figure 30: Comparison between model performances with no rulers present (left) vs with rulers present (right). **EfficientNet-B3** trained on ISIC 2020 & 2017 data [45, 11], skewed to dr=18.

Although we choose to use AUC as the primary metric rather than accuracy since accuracy depends on the threshold set (qualified in Appendix A.6), Table 9 shows the accuracy scores that correspond to the AUC scores in Table 1. These accuracy scores are calculated with a threshold of 0.5. These

accuracy scores also corroborate that the debiasing techniques improve the models robustness to artefact bias.

Table 9: Comparison of each unlearning technique against the baseline, trained on artificially skewed ISIC data. 'Heid plain' test set is free of artefacts while 'Heid Marked' and 'Heid Rulers' are the same lesions with surgical markings and rulers present respectively. All scores are accuracy (0.5 threshold).

(a) Removal of surgical marking bias (dm=20).

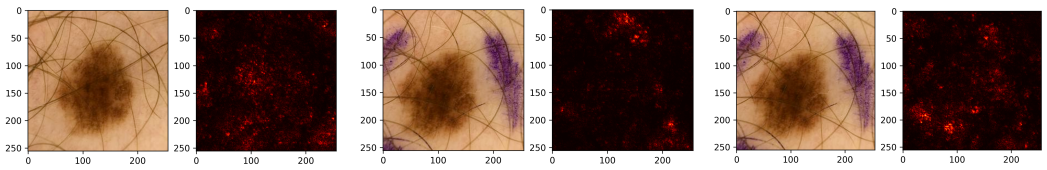
Experiment	Heid Plain	Heid Marked
Baseline	0.903±0.006	0.853±0.018
LNTL†	0.918±0.008	0.906±0.017
TABE†	**0.928**±0.023	0.836±0.058
CLGR†	0.927±0.021	**0.915**±0.015

(b) Removal of ruler bias (dr=18).

Experiment	Heid Plain	Heid Ruler
Baseline	0.961±0.009	0.682±0.057
LNTL‡	0.954±0.013	0.778±0.046
TABE‡	0.955±0.009	0.835±0.030
CLGR‡	**0.963**±0.005	**0.899**±0.002

F.1.1 Saliency maps

Since artefact bias can be located by image region, we attempt to identify whether the model is utilising the artefacts for classification by producing vanilla gradient saliency maps [49]. This is a pixel attribution method and is designed to highlight pixels that were most relevant for classification using a heatmap of the same resolution as the input image. This method leverages backpropagation to calculate the gradient of the loss function with respect to the input pixels. These pixel-wise derivative values can then be used to create a heatmap of the input image which highlights the location of pixels with high values. We output saliency maps for both the plain and biased images from [57, 56], to see if the focus of the model shifts from the lesion to the artefact (see Figure 31). The results are not as clear as hoped, however it can be noticed that for the baseline, there are less highlighted pixels in the lesion region when surgical markings are present (Figure 31b) compared to when there is not (Figure 31a), and potentially more in regions that correspond to surgically marked regions. When using the LNTL model, the most salient pixels look to be located back in the general region of the image lesion, indicating the model has learned not to use the surgical markings for classification.



(a) No surgical markings, baseline. (b) Surgical markings, baseline. (c) Surgical markings, LNTL.

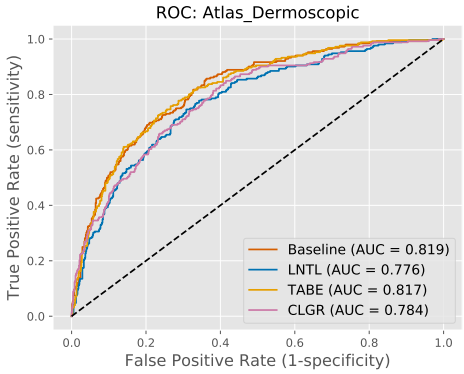
Figure 31: Vanilla gradient saliency maps, attempting to show the image regions most used by the model for classification. We compare the baseline on an unbiased and biased image of the same lesion, and also the LNTL model on the same biased image.

This method is the original and simplest of the saliency map techniques and does have some problems, one being that the ReLU activation function causes a saturation problem [48]. For future work a more sophisticated technique like the GRAD-CAM post-hoc attention method [47] may yield less noisy visualisations.

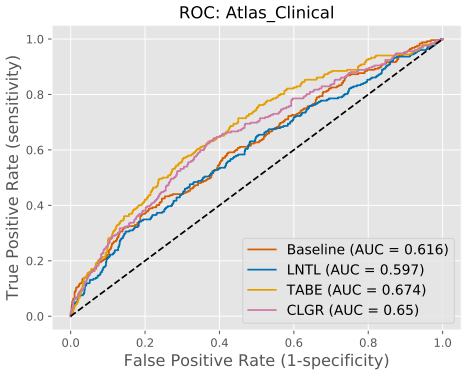
F.2 Domain generalisation

Figure 32 shows the ROC curves corresponding to Table 2. TABE and CLGR are able to be differentiated from the baseline across most test sets, providing evidence that these models generalise better than the baseline when removing instrument bias.

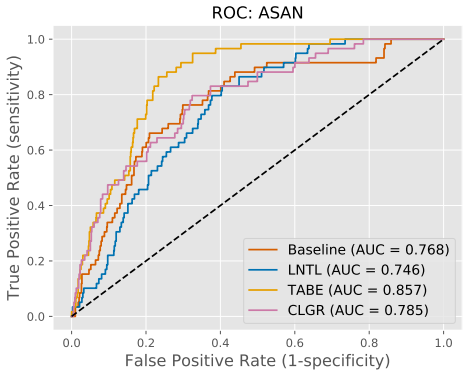
Table 10 is the full version of Table 3. A single debiasing head removing instrument bias is shown to be generally more effective than any combination of instrument, surgical marking or ruler bias removal. This is more evidence that combining debiasing heads can sometimes negatively impact performance, perhaps explaining the poor performance of the 7 head solution in [5].



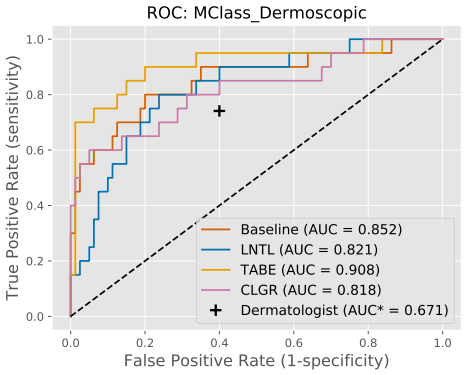
(a) Atlas dermoscopic.



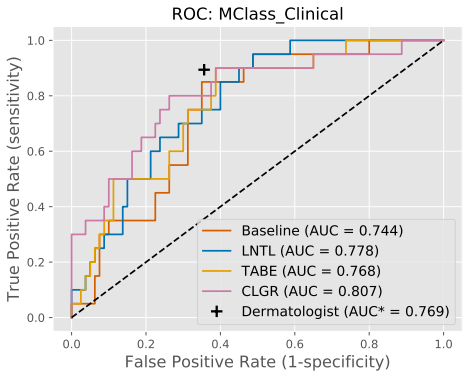
(b) Atlas clinical



(c) ASAN (clinical)



(d) MClass Dermoscopic



(e) MClass Clinical

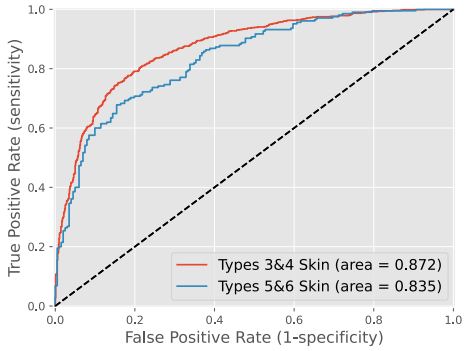
Figure 32: ROC curves for each debiasing method, with **ResNeXt-101** as the base architecture, aiming to remove spurious variation caused by the imaging instrument used. Model trained using the ISIC 2020 [45] and 2017 data [11] and tested on five test sets [38, 28, 7, 7].

Table 10: Comparison of generalisation ability of **ResNeXt-101** models trained using ISIC 2017 and 2020 data (not artificially skewed). The 'dermatologists' row is the AUC scores from [7]. A capital 'D' indicates the images are dermoscopic, while a capital 'C' means the images are clinical. The § symbol indicates the use of instrument labels, † represents surgical marking labels and ‡ represents ruler labels.

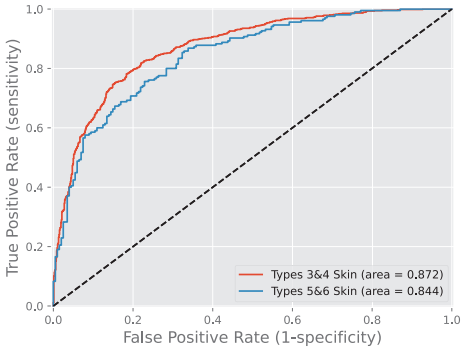
Experiment	AtlasD	AtlasC	ASANC	MClassD	MClassC
Dermatologists	—	—	—	0.671	0.769
Baseline	0.819	0.616	0.768	0.853	0.744
LNTL§	0.776	0.597	0.746	0.821	0.778
TABE§	0.817	**0.674**	**0.857**	**0.908**	0.768
CLGR§	0.784	0.650	0.785	0.818	0.807
LNTL†	0.737	0.589	0.631	0.731	0.799
TABE†	0.788	0.658	0.768	0.889	0.851
CLGR†	0.758	0.583	0.679	0.819	0.774
LNTL‡	0.818	0.616	0.705	0.849	0.759
TABE‡	0.813	0.667	0.679	0.865	0.846
CLGR‡	0.818	0.610	0.760	0.886	**0.882**
LNTL§&LNTL†	0.789	0.588	0.704	0.849	0.796
TABE§&TABE†	0.807	0.629	0.779	0.859	0.810
LNTL§&TABE†	0.802	0.591	0.766	0.864	0.705
LNTL§&CLGR†	0.573	0.574	0.645	0.717	0.617
CLGR§&CLGR†	0.801	0.656	0.840	0.811	0.820
CLGR§&LNTL†	0.763	0.615	0.767	0.833	0.790
TABE§&LNTL†	0.823	0.629	0.787	0.881	0.781
LNTL§&LNTL‡	0.786	0.604	0.686	0.837	0.779
TABE§&TABE‡	0.806	0.612	0.783	0.827	0.794
LNTL§&TABE‡	0.806	0.606	0.728	0.881	0.747
LNTL§&CLGR‡	0.816	0.618	0.740	0.872	0.792
CLGR§&CLGR‡	0.798	0.613	0.723	0.898	0.795
CLGR§&LNTL‡	0.793	0.586	0.704	0.876	0.776
TABE§&LNTL‡	**0.828**	0.640	0.747	0.880	0.824

F.3 Skin tone bias removal

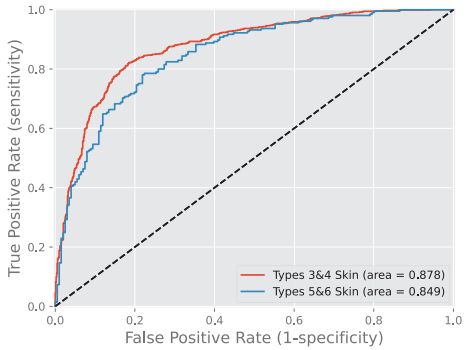
Figure 33 shows the ROC plots corresponding to Table 4, which are the results of attempting to reduce the performance difference between skin tones different to the training data using skin tone unlearning. The Fitzpatrick17k data is split into three subsets: types 1&2, types 3&4, types 5&6. The models are trained using types 1&2 skin, and tested on the other two subsets to monitor performance difference between them. Our baseline model reproduces the findings in [22] that models perform best on skin tones closest to the training data, with AUC score on types 3&4 4% better than types 5&6.



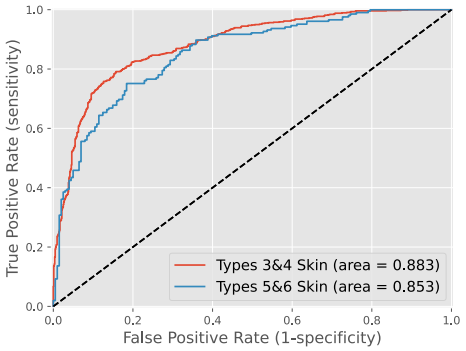
(a) ResNeXt-101 baseline



(b) ResNeXt-101 with LNTL [33] debiasing head.



(c) ResNeXt-101 with TABE [1] debiasing head.



(d) ResNeXt-101 with CLGR [33] debiasing head.

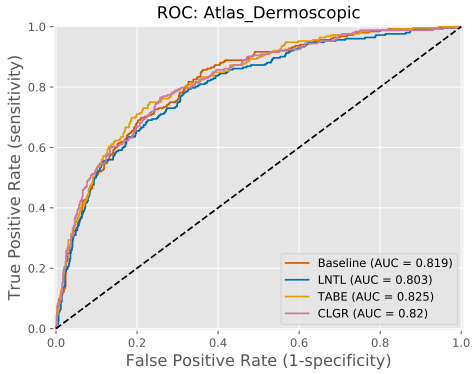
Figure 33: Attempting to tackle the issue of poor model generalisation to skin tones most different to the training data [22]. CLGR shows some reduction in disparity between skin types and small overall improvement.

The ROC curves that correspond to Table 5 are presented in Figure 34. The MClass dermoscopic data, like the training data, consists of ISIC archive data and so the improvement demonstrates the benefit of skin tone unlearning even when the target data is from a similar distribution.

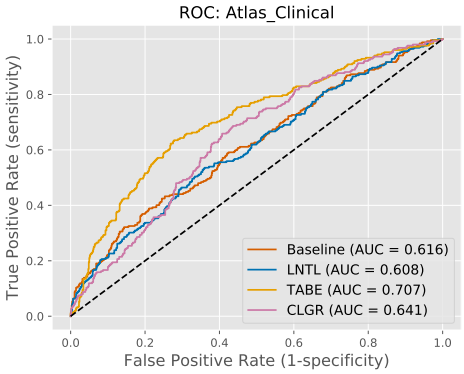
Table 11: Attempting to improve model generalisation to skin tones different to the training data. Trained using types 1 and 2 skin images from the ISIC 2017 & 2020 datasets [11, 45], tested on types 3&4 and 5&6 from the same set.

Experiment	Types 3&4	Types 5&6
Baseline	0.857	0.866
LNTL	0.864	0.867
TABE	0.882	0.811
CLGR	0.857	0.806

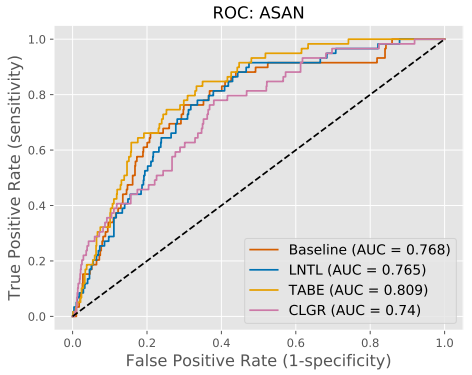
On top of attempting to show improved generalisation by removing skin tone bias when training using the full ISIC dataset, we also attempt to conduct the same experiment as was done on the Fitzpatrick17k data and split the dataset into three skin type groups, training on type 1&2. Table 11 presents the ROC curves from this experiment. We don't observe any improvement in disparity between the skin type groups. This is perhaps due to the imperfect automated labelling of the skin types meaning that the skin types cannot be separated into their actual groups well, in combination with unlearning not being optimal using these noisy labels.



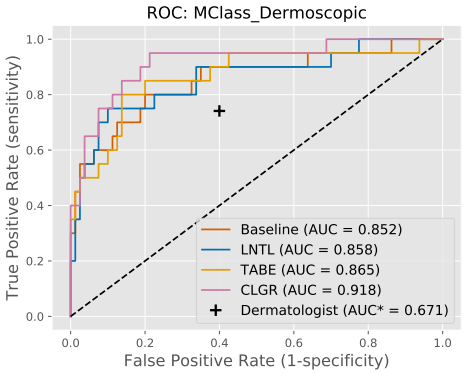
(a) Atlas dermscopic



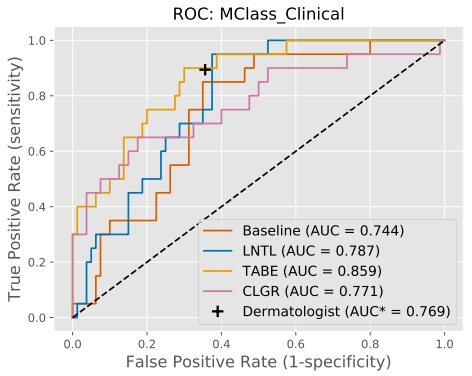
(b) Atlas clinical



(c) ASAN clinical



(d) MClass dermoscopic



(e) MClass clinical

Figure 34: ROC curves for each debiasing method, with **ResNeXt-101** as the base architecture, aiming to remove skin tone bias. Model trained using the ISIC 2020 [45] and 2017 data [11]. Tested on the test sets from the domain generalisation experiments.

F.4 Ablation studies

In Section 4.3 we attempted skin tone unlearning when training using a subset of the Fitzpatrick17k dataset (types 1&2 skin only), and testing on types 3&4 and 5&6 from the same set. Since the performance increase was only marginal, we experimented with using deeper auxiliary heads (additional fully connected layer), to see if this allowed the bias to be classified and removed more effectively. The results of these experiments can be seen in Table 12, with the deeper headed models denoted by an asterisk (*). This addition didn't have a noticeable impact on performance and so was not carried over to other experiments.

Table 12: Attempting to improve model generalisation to skin tones different to the training data [22]. Trained using types 1 and 2 skin images from the Fitzpatrick17k dataset [22], tested on types 3&4 and 5&6 from the same set. Asterisk (*) indicates use of **deeper head** (additional fully connected layer).

Experiment	Types 3&4	Types 5&6
Baseline	0.872	0.835
LNTL	**0.873**	0.834
LNTL*	0.872	**0.844**
TABE	0.878	**0.849**
TABE*	**0.884**	0.843
CLGR	0.883	**0.853**
CLGR*	**0.888**	0.849

G Personal reflections

In this project I attempted to tackle a real problem in computational dermatology, using state of the art deep learning techniques. This made for a challenging but very interesting project. Prompted by the suggested project topics, I began by reading into bias unlearning methods [33, 1], and went on to read around interesting domains where applying these methods may be useful. Having previously come across the ISIC challenge on Kaggle, I read a review paper [9] suggesting bias caused by artefacts causes performance irregularities in melanoma classification models, which I identified as an issue that may be helped using debiasing techniques. This sparked deeper research into the area, and the project eventually expanded to other biases in melanoma classification.

Once I had a good idea of the problem and potential solutions, I gathered training and test data and began to code up the debiasing models for experimentation, using a mixture of my own code and snippets found in github repositories corresponding to papers I had read. I then devised a set of experiments to assess the effectiveness of the proposed solution. Initially, I was experimenting using Kaggles notebooks and free GPU time, but later on in the project Amir (my supervisor for the project) kindly provided access to a machine in the school of computing which sped up the experimentation process significantly. The focus of the project morphed along the way as my understanding of the area increased and as new ideas came to mind. I learned many new machine learning concepts through reading papers and searching for existing solutions, as well as becoming proficient in several tools that I'd never used previously.

Since the problem of classifying skin lesions using machine learning is a very well researched one, many of the initial barriers to entry were lowered. For instance, I found that there were many good datasets already publicly available, alleviating the need to spend time creating datasets. The wealth of papers on the subject allowed me to get up to speed on the field quite quickly, and the Kaggle competition allowed me to get an idea of a good approach for a baseline model. I also found the community to be very helpful, receiving pointers from several researchers in the field.

A special thanks goes to Amir for his excellent support throughout the project. Knowing that expert help and advice was readily available when necessary really helped me to stay motivated during difficult parts of the work. Weekly meetings proved to be about the right interval, giving me time to amass questions and problems between each meeting, but not leaving so much time that work ever stalled. On top of this Amir was always very responsive via email in-between meetings for the times when pressing issues arose.

The technical aspect of the project I found most challenging was probably the DB-VAE, as it proved quite unstable and frustrating to train, and the implementation I found was quite difficult to understand at first in order to make modifications. I also found the report writing very difficult, struggling to stay engaged with the task. Learning to use LaTeX helped greatly as it took away the pains of formatting, as well as allowing me to place notes within the document. I think the technical programming feel of LaTeX may have also helped to keep my interest in the report writing process. I also found the switch from Keras to PyTorch and the switch from notebooks to Linux a little difficult initially, but once I overcame the learning curve I definitely found this worthwhile. When reading papers and the corresponding code, I sometimes found it difficult to understand how the mathematical details translated into code, especially because many of these were relatively new concepts and so I couldn't fall back on nicely presented tutorials or videos. I learnt to bridge the mental gap between how a formula looked on paper and how it presented as code (sometimes).

One of the biggest lessons I've learned is to properly plan out the necessary experiments in advance. Much time was wasted during the project running experiments that were, on further thought, not needed or incorrect in some way. Another lesson learnt was to commit to git much more regularly than I had been previously, since when I needed to revert to a previous version, this often rolled back too many other changes that I had made in between commits. Finally, I learnt that when it comes to implementing difficult machine learning/computer science concepts, perseverance is key, since the point at which I was almost ready to give up often turned out to be directly before the point at which it all came together and began to work.

The project has given me confidence that I am able to read emerging literature and make use of this for any task. I'm excited to continue reading machine learning research as it emerges and implement this for future projects. Going forward I would like to improve my skills at implementing papers without using other peoples implementations as a guide; often when copying a snippet from another implementation I was left thinking that there's no way I would have been able to figure this out myself from the paper alone.