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

Peter J. Bevan
School of Computing
Newcastle University, UK
peterbevan@hotmail.co.uk

Amir Atapour-Abarghouei
Department of Computer Science
Durham University, UK
amir.atapour-abarghouei@durham.ac.uk

Abstract

Convolutional Neural Networks have demonstrated dermatologist-level performance in the classification of melanoma and other skin lesions, but prediction irregularities due to biases seen within the training data are an issue that should be addressed before widespread deployment is possible. 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 reasonably mitigated using these bias removal methods. We also demonstrate the generalisation benefits of 'unlearning' spurious variation relating to the imaging instrument used to capture lesion images. Contributions of this work include the application of different debiasing techniques for artefact bias removal and the concept of instrument bias 'unlearning' for domain generalisation in melanoma detection. Our experimental results provide evidence that the effects of each of the aforementioned biases are notably reduced, 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 [6, 7, 17]. This is particularly important since, when diagnosed early, melanoma may be easily cured by surgical excision [18, 43], and so accessible and accurate diagnostic tools have the potential to democratise dermatology and save numerous 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. [42, 43] 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, though changing the habits of every dermatologist is highly unrealistic and could potentially be detrimental to their performance. Segmentation of the lesion from the surrounding skin has also previously been proposed, but is not a good option, since "any kind of pre-processing or segmentation itself may erroneously introduce changes that impede a CNN's correct classification of a lesion" [42]. Cropping surgical markings out of the image has been shown to be effective at mitigating surgical marking bias in [43], but it is noted this must be done by an experienced dermatologist to prevent the loss of important information, which is costly and time-consuming.

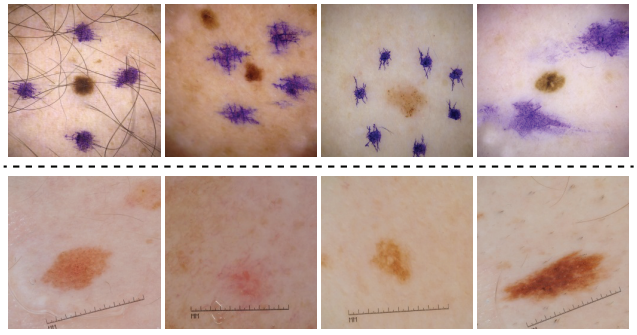


Figure 1: Examples of artefacts seen in ISIC 2020 data [33]. Top row shows images with surgical markings present, bottom row shows images with rulers present.

Consequently, the alternative path towards diminishing the effects of such artefacts would be not to remove the artefacts themselves from the image, but to reduce their influence on how the model functions, which translates to removing the 'bias' these artefacts introduce into the learning process. As such, recent advances in debiasing architectures for CNNs [1, 24] present an excellent opportunity

to robustly mitigate the aforementioned biases without any need to alter the behaviour of physicians or pre-process the image data.

Surgical artefacts left by physicians are not the only concern when it comes to skin lesion classification, however. 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 [13], we propose also using 'unlearning' techniques [1, 24] for domain generalisation by removing spurious variation associated with instrument type to create a more generalisable, instrument-invariant model.

In summary, this work aims to explore bias and domain 'unlearning' towards creating more robust, generalisable and fair models for melanoma classification. Code will be released post-review to preserve anonymity. Our primary contributions can be summarised as follows:

- *Melanoma classification* - The models presented in this paper demonstrate impressive melanoma classification performance, beating the average performance of experienced dermatologists on a benchmark dataset [7] (Section 4.2).
- *Artefact debiasing* - We mitigate the bias introduced by surgical markings and rulers, as shown in [42, 43] using 'Learning Not to Learn' [24] and 'Turning a Blind Eye' [1] (Section 4.1).
- *Domain generalisation* - We demonstrate the generalisation benefits of unlearning [1, 24] information relating to the instruments used to capture skin lesion images (Section 4.2).

2. Related work

We consider related work within two distinct areas, namely artefact bias in skin lesion images (Section 2.1) and domain generalisation (Section 2.2).

2.1. Artefacts bias

One of the problems addressed in this paper was investigated in [8], 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 [43], 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) [43]. Another work [42] shows a similar level of bias caused by rulers in skin lesion images.

Segmentation of skin lesions from the surrounding healthy skin has been suggested as a means of removing artefacts from the input of skin lesion classification models in [23, 30]. However, this is not commonly used at the time of writing, given that CNNs may utilise information in the surrounding skin regions [4, 42] and so removing this can impact classification performance. Artefacts themselves also often impede segmentation [29], and artefacts that are on the lesion itself are not separable from the lesion by image region.

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 model with seven debiasing heads [24] in an attempt to remove the bias caused by seven artefacts. The authors conclude that the bias removal method in [24] ('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, therefore, 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 such as the Interactive Atlas of Dermoscopy clinical data [27], 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, though this assumption does not usually hold true in real-world applications [10]. For instance, inconsistencies in prostate cancer classification performance between image samples originating from different clinics is shown in [2], and the authors hypothesise 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) [41] (see Figure 2). This domain shift has been shown to impact model performance in [14]. Additionally, 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 is shown in [22] that CNNs can easily discriminate between camera models, which can lead models

Figure 2: Illustration of the domain shift between clinical and dermoscopic images [27] of the same lesion. Top row shows dermoscopic images, bottom row clinical.

to overfit to this spurious variation during training.

Domain adaptation methods have been successfully used to minimise the distance between the underlying distributions of the training and test datasets, *i.e*. a model trained on a given dataset (source distribution) is enabled to perform well on a different dataset (target distribution) via domain adaptation [3, 10, 14, 15]. However, such methods require knowledge of the target distribution, which is not always readily available. Domain generalisation, on the other hand, is more robust than domain adaptation, and differs in that the target domain is unseen [25], aiming for improved performance on a wide range of possible test data. In this work, we explore applying bias unlearning techniques [1, 24] towards domain generalisation in melanoma classification, attempting to find an instrument-invariant feature representation without compromising performance.

3. Methods

In this work, two leading debiasing techniques within the literature are used, namely 'Learning Not To Learn' (LNTL) [24] 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.

3.1. Learning Not to Learn

'Learning Not to Learn' (LNTL) [24] proposes a novel regularisation loss combined with a gradient reversal layer [13] to remove bias from the feature representation of a CNN during backpropagation. Figure 3 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 [20] or EfficientNet [39] in this work. The 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 work, $\mathcal{Y}$ 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 step 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:

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

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 4 shows a generic overview of the 'Turning a Blind Eye' (TABE) [1] architecture. Similar to LNTL [24], 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 [40],

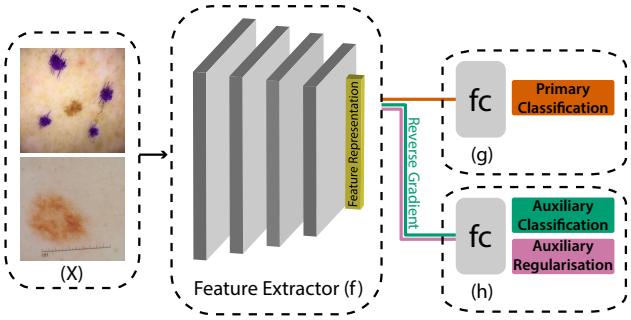


Figure 3: 'Learning Not to Learn' architecture. Feature extractor, f, is implemented as a convolutional architecture such as ResNeXt or EfficientNet in this work. 'fc' denotes a fully connected layer.

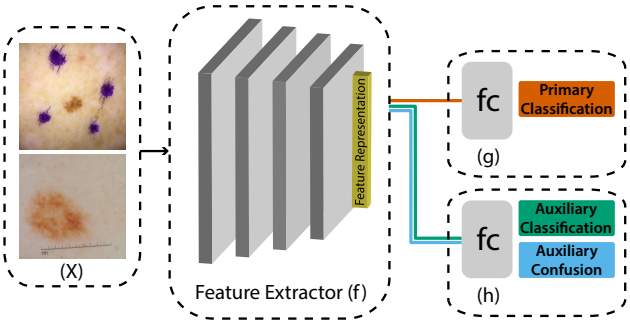


Figure 4: 'Turning a Blind Eye' generic architecture. Feature extractor, f, is implemented as a convolutional architecture such as ResNeXt or EfficientNet in this work. 'fc' denotes a fully connected layer.

$\mathcal{L}_{\text{conf}}$, 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:

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

where x_m is the input, y_m is the bias label, p_{nm} 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:

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

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 [20] or EfficientNet [39] in this work.

As suggested in [24], 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 [24] to the auxiliary classification loss as it is back-propegated to f. This configuration is denoted as CLGR in this work.

3.3. Datasets

This section briefly describes the datasets used in the experiments (see supplementary material Section B for example images).

3.3.1 ISIC challenge training data

The International Skin Imaging Collaboration (ISIC) challenge [33] is a yearly automated melanoma classification challenge with several publicly available dermoscopic skin lesion datasets (see ISIC archive[1]), complete with diagnosis labels and metadata. A combination of the 2017 and 2020 ISIC challenge data [9, 33] (35,574 images) is used as the training data in this work 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 [12]. 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 [9] challenge data is used as the validation set for hyperparameter tuning.

The model and training data used in [42, 43] are 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 [9, 33] to produce similar levels of bias in our baseline model to that shown in the aforementioned studies [42, 43]. 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, thus reproducing the level of performance shown in [43]. The dataset is 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 (Section 4.1), and the original data is used for all other experiments.

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

3.3.2 Heidelberg University test data

The test set presented in [43] is used to evaluate the artefact debiasing approach presented in this work (Section 4.1). The dataset consists of 130 lesions: 23 malignant, 107 benign. There are two images of each lesion in the set, one without 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 [43]. The test set from the ruler bias study [42] is not publicly available or shared, so the plain images from [43] are superimposed with rulers to be used as test images. The approach of superimposing rulers was validated as not statistically significantly different from in-vivo rulers in [42].

3.3.3 MClass benchmark test data

The MClass public human benchmark introduced in [7] is used as a test dataset for assessing domain generalisation (Section 4.2), also 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.

3.3.4 Interactive Atlas of Dermoscopy and Asan data

Two additional test sets, the Interactive Atlas of Dermoscopy dataset [27], and the Asan test dataset [19], are used to further test domain generalisation (Section 4.2). The Atlas dataset comprises 1,011 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 [9, 33] is mostly white Western patients, the Atlas and Asan datasets seem to have representation from a broad variety of ethnic groups, which provides a good test of a model's ability to deal with domain shift.

3.4. Implementation

All experiments are implemented in PyTorch [31] 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 [16], which utilises the EfficientNet-B3 architecture [39], pre-trained on the ImageNet dataset [11]. ResNet-101 [20], ResNeXt-101 [44], DenseNet [21] and Inception-v3 [38] are each substituted for EfficientNet-B3 to evaluate the optimal network for the task, simultaneously testing 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 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 [43]. The primary and auxiliary classification heads are implemented as a single fully connected layer, as suggested in [24]. 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 Section D of the supplementary material for full hyperparameter tuning 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 Equation (1) and Equation (3) are also empirically chosen to be α=0.03 and λ=0.01.

A weighted loss function is implemented for all auxiliary heads to tackle class imbalance, with each weighting coefficient, $\mathcal{W}_n$, being 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 [16, 19, 26, 30], 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 [28]. We use test-time augmentation [37, 38] to average predictions over 8 random flips along different axes, applied to all test images, to enable a fairer evaluation of our models. 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 Section 5 of the supplementary material).

4. Experimental results

The results of our artefact bias removal experiments are presented in Section 4.1. We present the domain generalisation experimental results in Section 4.2.

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, namely surgical markings [43] and rulers [42] (see Table 1 and Figure 5). Separate individually-skewed training sets are used with skew levels set at dm=20 (duplications of images with surgical markings) for examining the removal of surgical marking bias and dr=18 (duplications of images with rulers) for ruler bias. We use surgical marking and ruler labels as the target for the debiasing heads in each of these experiments respectively.

Each model is trained and evaluated 6 times using 6 different random seeds, allowing the mean and standard deviation to be reported. The high scores are due to the inherent clarity of the cues within the images and consequent simplicity of the classification of the test set [43], and are consistent with the scores reported in [42, 43]. Any chance of a leak between the test set and the ISIC training data has been ruled out [43]. Despite the ease of classification, both the existence of bias and the effectiveness of bias mitigation can still be demonstrated, and experiments using the original test set [43] provide direct evidence that we are able to mitigate the problem presented in these studies [42, 43]. While the baseline model performs very well for the unbiased images (Table 1 - 'Heid Plain'), performance suffers when this model is tested on the same lesions with either artefact present, replicating the findings from [42, 43].

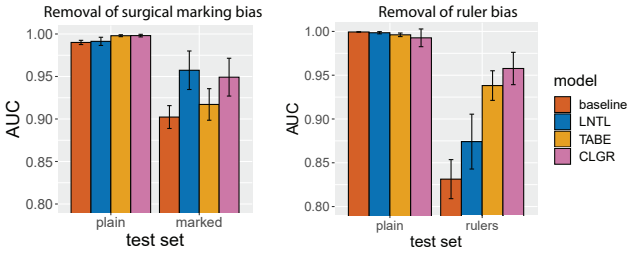


Figure 5: Comparison of artefact debiasing models against the baseline, trained on artificially skewed ISIC data.

Figure 5 presents evidence that each debiasing method is successful at mitigating artefact bias. LNTL is the most effective at unlearning surgical marking bias, achieving comparable performance to the baseline on the plain images from Heidelberg University [43] ('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 [42, 43], but are not a perfect solution, given that the artefacts still have a negative impact on performance.

4.2. Domain generalisation

Another significant issue within melanoma classification is instrument bias, hindering the application of a trained model to image data acquired via different imaging instruments. We attempt to address this issue by removing instrument bias from the model pipeline using unlearning techniques [1, 24], with the aim of improving domain generalisation. According to ISIC, image dimensions in [9, 33] 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 [9, 33] make up the first 8 'instrument' categories, but there are many outlier categories with a very small number of observations, which are discarded to prevent significant class imbalance.

Table 2 compares the generalisation ability of each instrument debiasing method 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 (EfficientNet-B3 [39], ResNet-101 [20], ResNeXt-101 [44], DenseNet [21], Inception-v3 [38]) 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 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. 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 are observed 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 certain techniques may only be suitable for specific tasks and datasets.

Figure 6 illustrates the benefits of using a TABE head for instrument bias removal compared to the baseline model (both ResNeXt-101), showing an 11.6% AUC improvement on the Asan clinical test set [19] and a 6.6% increase on the MClass dermoscopic test set [7]. TABE can be differenti-

[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.

Experiment	(a) Surgical Marking Removal (dm=20)	
	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

Experiment	(b) Ruler Bias Removal (dr=18)	
	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

Table 1: *Artefact debiasing*: Comparison of unlearning techniques against the baseline, trained on skewed ISIC data. Scores are **AUC**. 'Heid Plain' is free from artefacts while 'Heid Marked' and 'Heid Rulers' contain the same lesions with surgical markings and rulers present. The † symbol indicates the use of surgical marking labels as target for the auxiliary head and ‡ indicates ruler labels.

Experiment	Architecture	Atlas		Asan	MClass	
		Dermoscopic	Clinical	Clinical	Dermoscopic	Clinical
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: *Domain generalisation*: Comparing generalisation ability of each debiasing method across different architectures, trained using ISIC 2017 and 2020 data [9, 33]. All scores are **AUC**. The 'dermatologists' row is the AUC scores from [7]. 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.

ated from the baseline across each clinical test set, suggesting this to be a good tool for domain generalisation between dermoscopic and clinical data. Since both the training data and the MClass dermoscopic [7] data are drawn from the ISIC archive, the improved performance on this test set suggests the benefits of instrument invariance even on data drawn from a similar distribution. This is likely due to the mitigation of domain bias caused by variation in the specific type of dermoscopic instrument used.

We also experiment with using two debiasing heads, each removing a different bias (either instrument, surgical marking or ruler), with the aim of improving generalisation. 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, other configurations perform best on two of the five test sets (Table 3). For results across a more complete set of configurations, please refer to Section 8 in the supplementary material.

4.3. Ablation studies

Ablation was built into the experimentation process as individual bias removal heads were implemented in isolation before attempting combinations, and debiasing heads

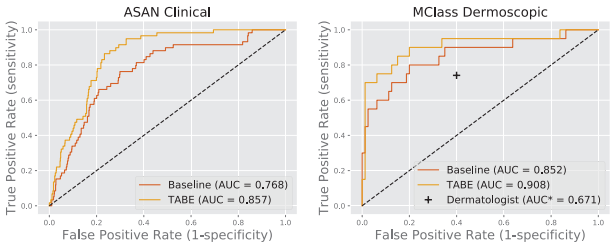


Figure 6: ROC curves for TABE [1] instrument debiasing on ASAN clinical [19] (left), and MClass dermoscopic [7] (right), with **ResNeXt-101** as the base architecture. Trained using ISIC 2020 [33] and 2017 data [9].

Experiment	Atlas		Asan	MClass	
	Dermoscopic	Clinical	Clinical	Dermoscopic	Clinical
Dermatologists	—	—	—	0.671	0.769
Baseline	0.757	0.565	0.477	0.786	0.775
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

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

were implemented both with and without gradient reversal. Using a single head to unlearn instrument bias is found to be more effective for generalisation than combining this head with artefact bias removal heads.

TABE [1] both with and without the gradient reversal layer (named CLGR with gradient reversal) has proven successful for different tasks (Table 1, Table 2), but ablation of the gradient reversal layer from LNTL [24] generally diminished performance, (see Table 4).

5. Limitations and future work

While we have demonstrated the impressive performance of unlearning techniques [1, 24] for artefact debiasing, one drawback of such approaches is the need to manually label these artefacts in each training image. These artefacts, however, are often quick and easy to identify by untrained individuals. Further research may look to uncover biases caused by other artefacts in a similar manner to [42, 43] and evaluate the effectiveness of unlearning techniques at mitigating these. Future work could also incorporate an algorithm which accurately labels artefacts and dynamically changes the model architecture to apply the required bias removal heads for the task.

As for potential improvements in domain generalisation, image resolution cannot be universally assumed as a proxy for imaging instrument across all datasets so we recommend the actual instrument model be recorded as metadata when collecting training data for melanoma classification. Further research could include incorporating an instrument identification system which could provide labels for the image acquisition instrument. Further work may also evaluate the effectiveness of these debiasing techniques [1, 24] in improving generalisation for diagnostic smartphone apps [32].

Experiment	Atlas		Asan	MClass	
	Dermoscopic	Clinical	Clinical	Dermoscopic	Clinical
LNTL	**0.804**	**0.612**	**0.768**	0.819	**0.801**
LNTL*	0.783	0.605	0.710	**0.827**	0.747

Table 4: Ablation of gradient reversal from LNTL using **ResNeXt-101** for removal of instrument bias. All scores are **AUC**. Asterisk (*) represents ablation of gradient reversal.

6. Conclusion

This work has compared and demonstrated the effectiveness of debiasing methods in the context of skin lesion classification. We have successfully mitigated the surgical marking and ruler bias presented in the work of Winkler et al. [42, 43] in an automated manner using unlearning techniques (Section 4.1). We have investigated the use of bias removal models against a baseline model on two test sets for each artefact bias, one comprising lesion images with no artefacts present and one comprising the same lesion images with artefacts present. We have shown that the debiasing models perform on par with the baseline on the images without artefacts, and better on the images with artefacts, with the 'Turning a Blind Eye' [1] (plus gradient reversal) model improving on the baseline AUC by 15.3% on the test set with rulers present (Table 1). This suggests the addition of these debiasing heads leads to a model more robust to each artefact bias, without compromising performance when no bias is present. Utilising these techniques could be an alternative to the behaviour change amongst dermatologists as suggested by Winkler et al. [42, 43].

We have also provided evidence of the generalisation benefits of using unlearning techniques to remove instrument-identifying information from the feature representation of CNNs trained for the classification of melanoma (Section 4.2). We have demonstrated this using the ISIC training data [9, 33], with image resolution as a proxy for the imaging instrument. To test the generalisation capabilities of our bias removal approaches, we have used five popular skin lesion test sets with varying degrees of domain shift. Utilising the 'Turning a Blind Eye' [1] debiasing head is most effective, achieving improved performance across the board, most notably inducing an 11.6% AUC increase compared to the baseline on the Asan dataset [19]. Our models perform better than experienced dermatologists, consist-

ently beating their average AUC score on the MClass test sets [7]. Generalisation methods such as this are powerful for ensuring consistent results across dermatology clinics, and may have utility in the emerging diagnostic app space [32], given that differences between smartphone cameras are likely to introduce spurious variation in a similar 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. 1, 2, 3, 4, 5, 6, 8

[2] 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, Apr. 2018. 2

[3] Amir Atapour-Abarghouei and Toby P. Breckon. Real-Time Monocular Depth Estimation Using Synthetic Data with Domain Adaptation via Image Style Transfer. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2800–2810, Salt Lake City, UT, June 2018. IEEE. 3

[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. 2, 12

[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. 2, 6, 15

[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, Apr. 2019. 1, 12

[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, Apr. 2019. 1, 2, 5, 6, 7, 8, 9, 13, 14, 17, 18

[8] 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. 2

[9] 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, Apr. 2018. 4, 5, 6, 7, 8, 12, 13, 14, 15, 16, 18

[10] 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. 2, 3

[11] 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. 5

[12] Chris Deotte. SIIM-ISIC Melanoma Classification - JPEG Melanoma 256x256. https://www.kaggle.com/cdeotte/jpeg-melanoma-256x256. 4

[13] 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. 2, 3

[14] 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. 2, 3

[15] Hao Guan and Mingxia Liu. Domain Adaptation for Medical Image Analysis: A Survey. *arXiv:2102.09508 [cs, eess]*, Feb. 2021. 3

[16] 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, Oct. 2020. 5, 12, 14

[17] 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, Aug. 2018. 1, 12

[18] 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, Jan. 2020. 1

[19] 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. 5, 6, 8, 13, 14, 18

[20] 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. 3, 4, 5, 6

[21] 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. 5, 6

[22] 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. 2

[23] 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, Dec. 2016. 2

[24] 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. 1, 2, 3, 4, 5, 6, 8

[25] 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]*, Oct. 2020. 3

[26] Yuexiang Li and Linlin Shen. Skin Lesion Analysis towards Melanoma Detection Using Deep Learning Network. *Sensors (Basel, Switzerland)*, 18(2), Feb. 2018. 5, 14

[27] 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. 2, 3, 5, 12, 13, 18

[28] Jayawant N. Mandrekar. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, Sept. 2010. 5, 14

[29] Nabin Kumar Mishra and M. Emre Celebi. An Overview of Melanoma Detection in Dermoscopy Images Using Image Processing and Machine Learning. *arXiv e-print*, arXiv:1601.07843, Jan. 2016. 2

[30] Erdem Okur and Mehmet Turkan. A survey on automated melanoma detection. *Engineering Applications of Artificial Intelligence*, 73:50–67, Aug. 2018. 2, 5, 12, 14

[31] 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. 5

[32] 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, Apr. 2018. 8, 9

[33] 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, Jan. 2021. 1, 4, 5, 6, 7, 8, 12, 13, 16, 18

[34] 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, Oct. 2017. 15

[35] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences. *arXiv:1704.02685 [cs]*, Oct. 2019. 15

[36] 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. 12, 15, 17

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 5

[38] 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. 5, 6

[39] 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. 3, 4, 5, 6

[40] 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, Dec. 2015. 3

[41] 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. 2

[42] 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. 1, 2, 4, 5, 6, 8, 15

[43] 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, Oct. 2019. 1, 2, 4, 5, 6, 8, 12, 13, 15

[44] 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. 5, 6

Supplementary material

This section presents supplementary material that can be referenced to enhance the readers' understanding of the details of the work. We could not fit the general literature review for skin lesion classification in the paper, so this is presented in Section A. Samples of each training and test dataset are illustrated in Section B, to give a feel for the images present in each. Justification for our choice of metrics is given in Section C. Hyperparameter tuning results are presented in Section D. Additional experimental results in the form of ROC curves and tables that were not included in the main paper can be found in Section E. Our attempt at interpreting artefact bias using vanilla gradient saliency maps [36] is presented in Section E.1.1.

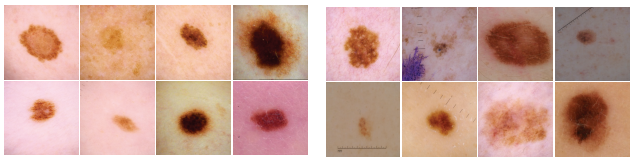
A. Skin lesion classification

The task of classifying skin lesions using machine learning has received attention within the research community since as early as 1988, initially using traditional machine learning methods such as decision trees in combination with segmentation [30]. Originally, lack of model sophistication, compute power and quality data meant that performance was not at the level of dermatologists. Like with many other areas of computer vision, the rise of convolutional neural networks and ever increasing compute power has seen the performance of skin lesion classification models rapidly increase to the point where there is evidence of machine learning techniques matching or even surpassing dermatologists at the task [6, 17]. 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. 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 [16]. We opt for the more common binary approach of classifying using benign/malignant as classes.

B. Examples of data

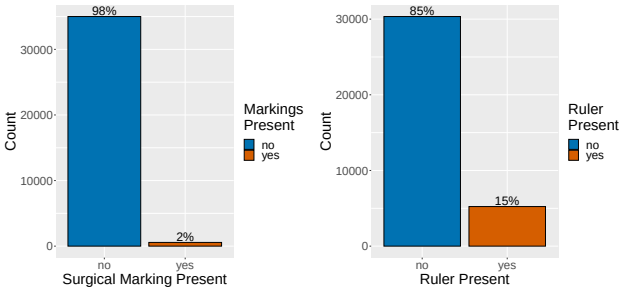
Figure 7 shows a sample of the images from the ISIC dermoscopic training data [9, 33], including some examples of surgical markings and some examples of rulers.



(a) Benign (b) Malignant

Figure 7: Example images from the ISIC dermoscopic training set [9, 33].

Figure 8 shows the class distribution for the surgical marking and ruler labels in the ISIC data. The distribution of artefacts is highly imbalanced, pointing to why weighted loss functions were needed to stabilise training.



(a) Marking distribution (b) Ruler distribution

Figure 8: Class distribution of artefacts in ISIC 2020 & 2017 training data [9, 33].

Figure 9 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 10 shows a sample of the 'Heid Plain' images from Heidelberg University [43]. These are dermoscopic images collected by the university of a variety of neoplastic lesions. Figure 11 shows a sample of the 'Heid Marked' images from Heidelberg university [43]. 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 12 shows a sample of the 'Heid Ruler' images, which was made by electronically superimposing rulers onto the 'Heid Plain' images.

Figure 13 shows a sample of the 'Interactive Atlas of Dermoscopy' [27] *dermoscopic* images, while Figure 14 shows the equivalent *clinical* images from the same set. The

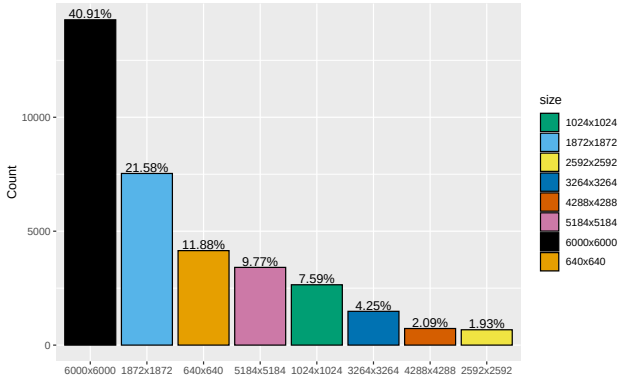


Figure 9: Class distribution of instruments in ISIC 2020/2017 combined data [9, 33]. Instruments inferred as separate through image resolution.

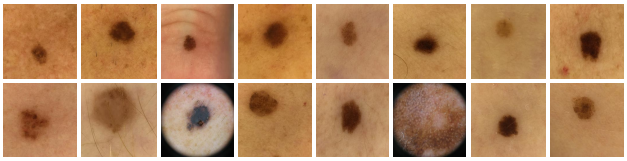


Figure 10: Example images from the Heidelberg University training set with no artefacts [43].

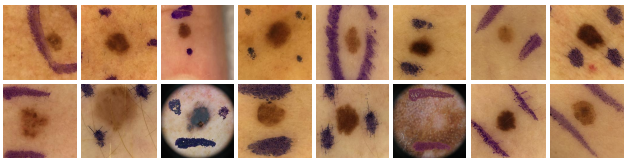


Figure 11: Example images from the Heidelberg University training set with surgical markings [43].

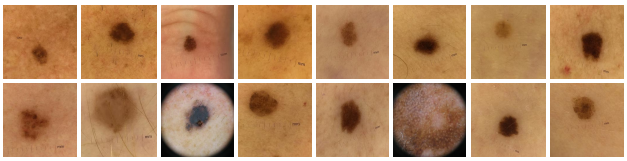


Figure 12: Example images from the Heidelberg University training set with superimposed rulers [43].

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 15 shows a sample of the Asan [19] clinical test set. This dataset is collected from the Asan medical centre, Seoul, South Korea and so features predominantly South Korean patients.

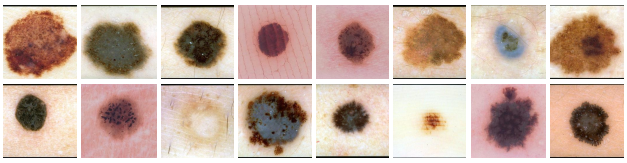


Figure 13: Example images from the Interactive Atlas of Dermoscopy dermoscopic test set [27].

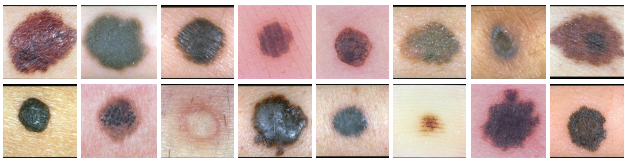


Figure 14: Example images from the Interactive Atlas of Dermoscopy clinical test set [27].

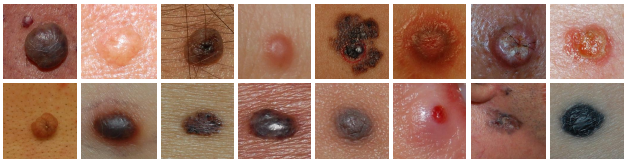


Figure 15: Example images from the Asan clinical test set [19].

Figure 16 shows a sample of the MClass [7] dermoscopic benchmark test set, and Figure 17 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 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.

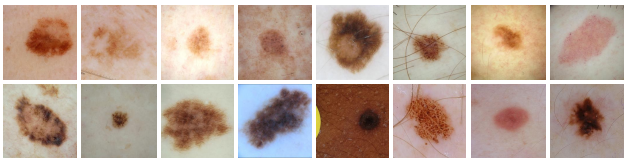


Figure 16: Example images from the MClass dermoscopic test set [7].

C. Metrics

The best evaluation metric for the task was carefully considered. We took into account the commonly used metrics in similar studies, as well as the specific requirements of the experiments.

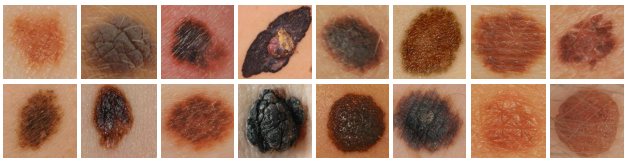


Figure 17: Example images from the MClass clinical test set [7].

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 is 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 21 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) [28]. 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. This is standard practice in melanoma classification [16, 19, 26, 30].

We anticipate that the binary classification 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 imbalance between benign/malignant lesions in the test sets meaning accuracy is not as descriptive of performance as AUC.

D. Hyperparameter tuning

Table 5 shows the chosen number of epochs for training each architecture for each dataset. These were chosen as the point at which the AUC reached its maximum or plateaued.

Table 6 shows 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 was not computationally feasible for the grid search, and so a random subset (33%, 3326 images) of the 2018 [9] challenge data is used as the validation set for hyperparameter tuning. 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 but it is important to note that the optimal performance is not the primary focus of this paper and as such, a detailed hyperparameter tuning procedure does not significantly contribute to the objectives of this paper.

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

Table 5: Optimal number of epochs for training, selected through analysis of cross validation curves.

E. Additional results

E.1. Artefact bias removal

To label the artefacts in the training 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 identifying 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 the remainders. This method does not work well at all for labelling rulers, likely due to the fact that hairs have similar pixel values to rulers. As a result, we manually label each image for rulers. The manual

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 6: Hyperparamter tuning of baseline ResNeXt-101 model, trained for 4 epochs and using a random subset (33%, 3326 images) of the 2018 [9] challenge data is used as the validation set for hyperparameter tuning.

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 18 shows the ROC plots from the surgical marking bias experiments. All models perform almost perfect classification of the easy test set with no artefacts (Figure 18a). The test set with surgical markings present causes performance to drop for all models. However, it is clear from Figure 18b 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 19b). The baseline is again affected most by this bias, with TABE clearly most robust to it.

Although we choose to use AUC as the primary metric rather than accuracy since accuracy depends on the threshold set (qualified in Appendix Appendix C), Table 7 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.

E.1.1 Saliency maps

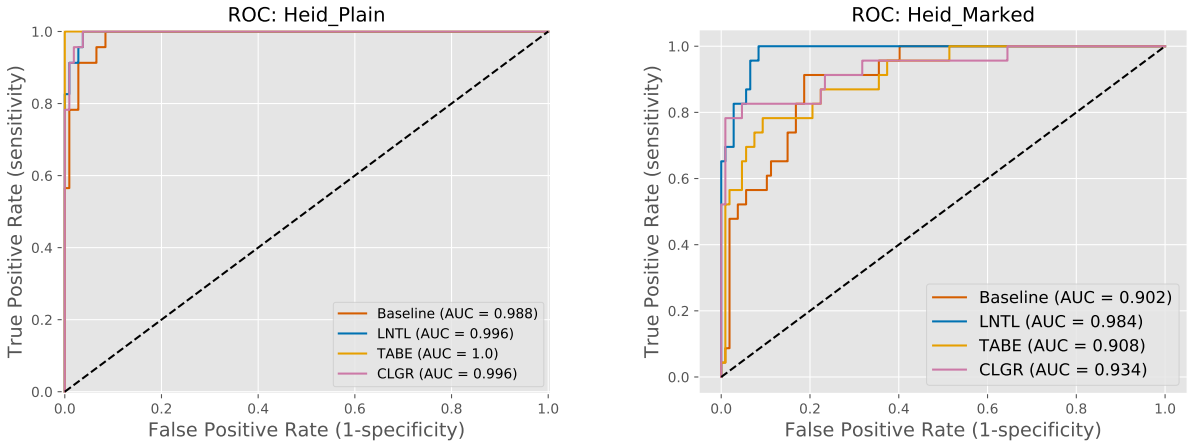
Since artefact bias can be located by image region, we attempt to identify whether the model is utilising the artefacts for classification by producing vanilla gradient saliency maps [36]. This is a pixel attribution method and is designed to highlight pixels that were most relevant for classification 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 [42, 43], to see if the focus of the model shifts from the lesion to the artefact (see Figure 20). It can be noticed in Figure 20 that for the baseline, there are less highlighted pixels in the lesion region when surgical markings are present compared to when there is not, 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 image region of the lesion, indicating the model has learned not to use the surgical markings for classification.

This method is a simple saliency map technique and suffers from certain issues, such as the ReLU activation function leading to a saturation problem [35]. For future work, a more sophisticated technique like the GRAD-CAM post-hoc attention method [34] may yield better quality visualisations.

E.2. Domain generalisation

Figure 21 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.

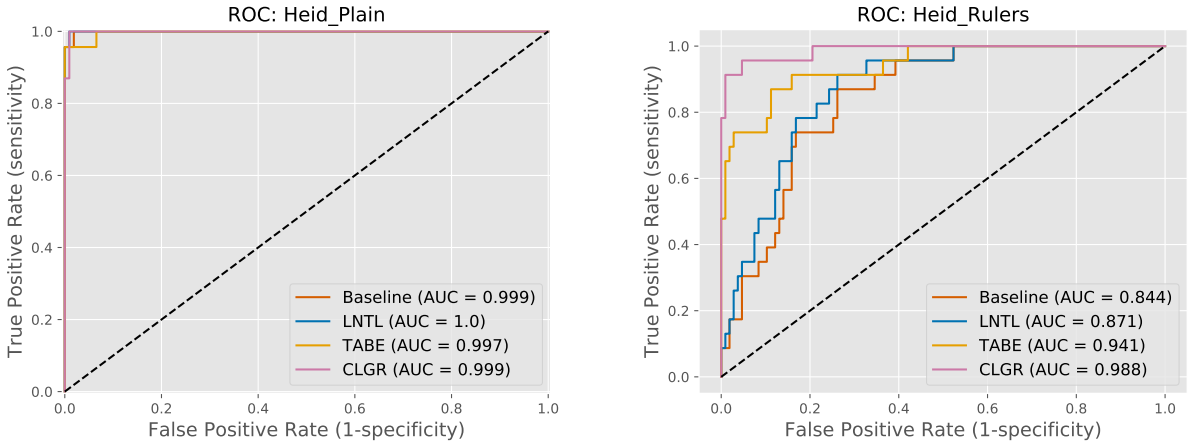
Table 8 is the full version of Table 3. A single debiasing head removing instrument bias is shown to be generally more effective than any combination of instrument, surgical marking or ruler bias removal. This is more evidence that combining debiasing heads can sometimes negatively impact performance, perhaps explaining the poor performance of the seven-head solution in [5].



(a) No surgical markings present.

(b) Surgical markings present.

Figure 18: Comparison between model performances with no surgical markings present (left) vs with surgical markings present (right). **EfficientNet-B3** trained on ISIC 2020 & 2017 data [9, 33], skewed to dm=20.



(a) No rulers present.

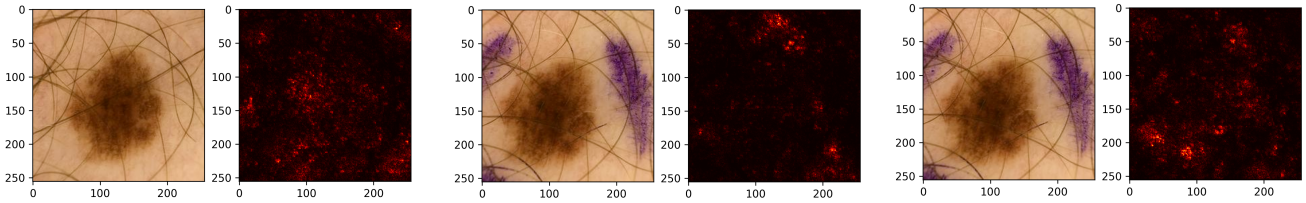
(b) Rulers present.

Figure 19: Comparison between model performances with no rulers present (left) vs with rulers present (right). **EfficientNet-B3** trained on ISIC 2020 & 2017 data [9, 33], skewed to dr=18.

Experiment	(a) Surgical Marking Removal (dm=20)	
	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

Experiment	(b) Ruler Bias Removal (dr=18)	
	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

Table 7: 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).

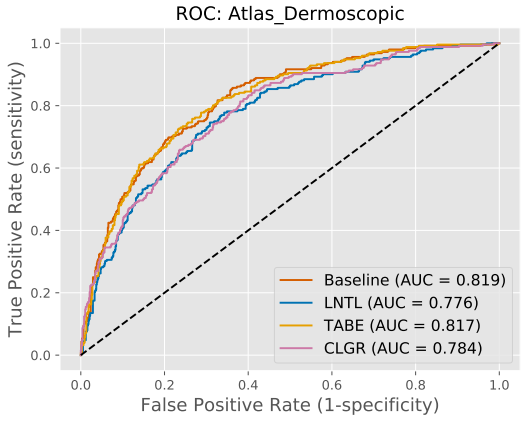


(a) No surgical markings, baseline. (b) Surgical markings, baseline. (c) Surgical markings, LNTL.

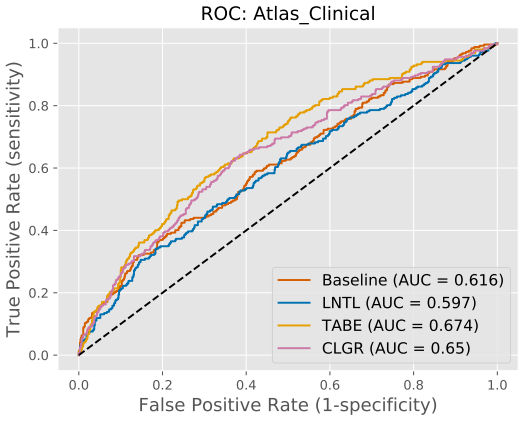
Figure 20: Vanilla gradient saliency maps [36] pointing to image regions most used by the model for classification. We compare the baseline on an unbiased and biased image of the same lesion, and also the LNTL model on the same biased image.

Experiment	Atlas		Asan	MClass	
	Dermoscopic	Clinical	Clinical	Dermoscopic	Clinical
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

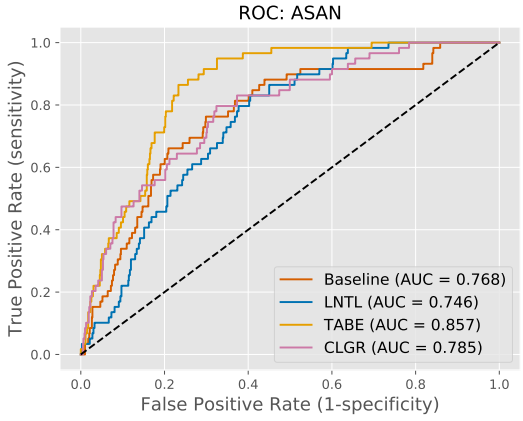
Table 8: *Domain generalisation*: Comparison of generalisation benefits of using different targets for the debiasing heads (**ResNeXt-101**), including some models with two debiasing heads. The 'dermatologists' row is the AUC scores from [7]. A capital 'D' indicates the images are dermoscopic, while a capital 'C' means the images are clinical. The § symbol indicates the use of instrument labels, † represents surgical marking labels and ‡ represents ruler labels.



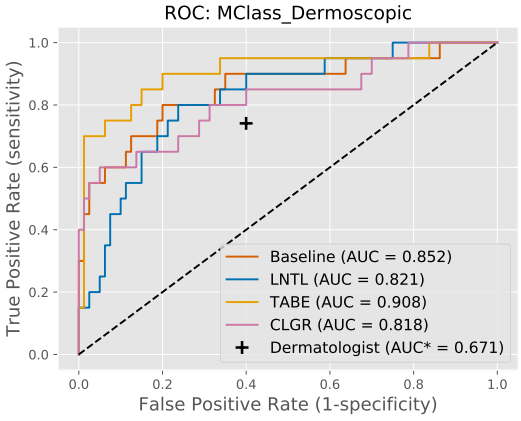
(a) Atlas dermoscopic.



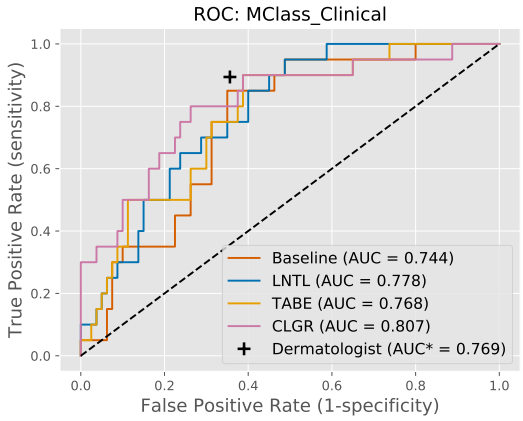
(b) Atlas clinical



(c) Asan (clinical)



(d) MClass Dermoscopic



(e) MClass Clinical

Figure 21: 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 [33] and 2017 data [9] and tested on five test sets [7, 19, 27].