Algorithmic Fairness in Lesion Classification by Mitigating Class Imbalance and Skin Tone Bias

Faizanuddin Ansari[1], Tapabrata Chakraborti[2], and Swagatam Das[1]

[1] Indian Statistical Institute, Kolkata, India
faizanuddin_r@isical.ac.in, swagatamdas19@yahoo.co.in
[2] The Alan Turing Institute, University College London, London, UK
tchakraborty@turing.ac.uk, t.chakraborty@ucl.ac.uk

Abstract. Deep learning models have shown considerable promise in the classification of skin lesions. However, a notable challenge arises from their inherent bias towards dominant skin tones and the issue of imbalanced class representation. This study introduces a novel data augmentation technique designed to address these limitations. Our approach harnesses contextual information from the prevalent class to synthesize various samples representing minority classes. Using a mixup-based algorithm guided by an adaptive sampler, our method effectively tackles bias and class imbalance issues. The adaptive sampler dynamically adjusts sampling probabilities based on the network's meta-set performance, enhancing overall accuracy. Our research demonstrates the efficacy of this approach in mitigating skin tone bias and achieving robust lesion classification across a spectrum of diverse skin colors from two distinct benchmark datasets, offering promising implications for improving dermatological diagnostic systems.

Keywords: Class Imbalance · Skin-Tone Bias · Data Augmentation · Skin Cancer Classification · Melanoma Detection · Algorithmic Fairness

1 Introduction

The field of skin lesion classification using deep learning models has shown great promise in achieving high performance in recent years. However, a significant challenge these models face is the presence of bias towards dominant skin tones and imbalanced class representation within datasets. This is because skin lesions and cancer occurs more on skin with less melanin, and thus is more prevalent among patients with pale skin, compared to darker skin. This has given rise to benchmark skin cancer image datasets that are biased towards samples from white patients, which is statistically proportionate, but causes training imbalance for machine learning algorithms. For a machine learning algorithm to be equitable and generalisable, it needs to perform robustly across demographics in a fair manner, because skin cancer can happen for darker skin, though lesser in number, but no patient can be treated as an outlier. Hence this is an open persisting problem within the area of machine learning fairness and health equity, which we address in this work.

Researchers have proposed various methods to handle imbalanced datasets. The methods mainly used for handling imbalances in the skin lesion dataset are widely classified into sampling-based, reweighing, and augmentation-based methods. The reweighing-based methods include RW (Reweighing) [7], Focal-loss [9], CBRW (Class balanced reweighing) [2], and Balanced Softmax [13], are some of the state-of-the-art reweighing based methods used to handle imbalance. This genre of approach focuses on assigning different weights to each data point during training. While effective in some cases, reweighing techniques can be sensitive to the chosen weight function and might not always capture the true importance of each data point. Other methods include using modified sampling techniques, such as oversampling the under-represented categories. We can further divide this oversampling and undersampling technique into instance, class-balanced [14,10], reverse, and progressive-balanced sampling. Oversampling techniques, while seemingly intuitive, can introduce redundancy and lead to the model overfitting on the minority data. Undersampling, on the other hand, discards valuable information from the majority class, potentially affecting overall model performance. Another line of work focuses on augmentation-based techniques, such as Mixup [17], CutMix [16], and Balanced Mixup [3], aim to regularize training data by mixing instance and class-based sampling. Rebalanced Mixup (Remix) [1] gives greater weight to labels of minority classes, improving generalization. Another technique, CMO [11], incorporates minority class images into majority class backgrounds. Methods like CutMix and CMO improve performance on tasks like CIFAR and ImageNet by replacing image parts with patches from others. However, in medical datasets, such as skin cancer detection, they often reduce accuracy by losing crucial diagnostic information [12]. These methods also lack dynamic adaptation during mixing, limiting their effectiveness on unseen data.

Moreover, not many studies have been done on mitigating the racial skin-tone bias in lesion classification. Groh et al. [4] study highlights a bias in skin tone classification using convolutional neural networks (CNNs). CNNs trained on datasets with limited skin tone diversity perform better on images with skin tones similar to those in the training data. This leads to lower accuracy for individuals with darker skin tones, which are often underrepresented in current datasets, thus giving rise a serious problem in machine learning fairness and health equity.

By critically evaluating strengths and limitations discussed above, our work develops a more robust approach for handling imbalanced datasets and inherent color-tone bias in skin lesion classification simultaneously, ultimately leading to more accurate and fair models for skin lesion diagnosis.

We propose a novel approach utilizing a mixup-based algorithm guided by an adaptive sampler. This method tackles these issues on two fronts:

1. *Mixup-based Algorithm*: We augment and create new training data points that blend features from existing samples by leveraging the mixup technique. This process encourages the model to learn more robust representations that are less susceptible to biases based on skin tone.

2. *Meta Adaptive Sampler*: We introduce an adaptive sampling strategy that dynamically adjusts the selection probabilities of training data during each iteration. This strategy prioritizes samples that pose more significant challenges to the network based on their performance on a dedicated meta-validation set. This targeted approach fosters more balanced learning and improves overall classification accuracy.

Our research demonstrates significant progress in mitigating skin tone bias within the model through this combined strategy. This translates to achieving robust lesion classification performance across a diverse range of skin colors on multiple datasets. These findings hold promising implications for developing more accurate dermatological diagnostic systems, that work across patient demographics with equity and fairness.

2 Methodology

In this section, we first present some concepts relevant to this work, and then introduce the proposed method.

2.1 Preliminaries

Sampling Strategy Sampling strategies of training data refer to techniques used to select and represent data instances in deep learning models, particularly when dealing with imbalanced datasets. These strategies aim to mitigate the underfitting of minority classes and prevent the overfitting of majority classes. Modified sampling strategies can include oversampling under-represented categories, leading to counter-productive outcomes like repeatedly showing the same training examples to the model. Given the provided training set notation, one can describe data sampling strategies mathematically as: given a training set $D = \{(x_i, y_i), i = 1, ..., N\}$ for a multi-class problem with K classes where each class k contains n_k examples and $\sum_{k=1}^{K} n_k = N$, we can describe some common data sampling strategies mathematically with the probability p_k, associated with the particular class as $p_k = (\frac{(|n_k-\psi|)^\gamma}{\sum_{l=1}^{K}(|n_l-\psi|)^\gamma+\epsilon})$, where different values of γ and ψ, guides the different sampling strategies, where $\gamma \in [0, 1]$, and ψ, some statistic related with the examples in the dataset and ϵ added to avoid divide by zero error. $\gamma = 1$, and $\psi = 0$ forms the instance sampling strategy, followed by $\gamma = 1$, and $\psi = N$ forms reverse sampling strategy. However, such static strategies that fix the sampling probability at the start and use it throughout training are not feasible.

Mixup The mixup technique constitutes a data augmentation approach wherein synthetic training instances are created through the linear interpolation between pairs of authentic examples alongside their corresponding labels. The formula for mixup is:

$$\hat{x} = \lambda x_1 + (1 - \lambda)x_2 \tag{1}$$

$$\hat{y} = \lambda y_1 + (1 - \lambda)y_2, \tag{2}$$

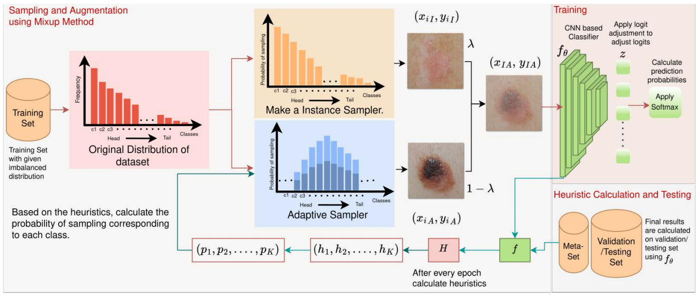
Sampling and Augmentation using Mixup Method
Training Set
Training Set with given imbalanced distribution
Frequency
c1 c2 c3
Head
Tail
Classes
Original Distribution of dataset
Probability of sampling
Make a Instance Sampler.
Adaptive Sampler
(x_{iI}, y_{iI})
λ
(x_{IA}, y_{IA})
$1-\lambda$
(x_{iA}, y_{iA})
Based on the heuristics, calculate the probability of sampling corresponding to each class.
$(p_1, p_2, \ldots, p_K)$
$(h_1, h_2, \ldots, h_K)$
H
f
After every epoch calculate heuristics
Training
CNN based Classifier
f_θ
Apply logit adjustment to adjust logits
z
Calculate prediction probabilities
Apply Softmax
Heuristic Calculation and Testing
Meta-Set
Validation /Testing Set
Final results are calculated on validation/ testing set using f_θ

Fig. 1. Our Proposed Framework

where λ is a random variable sampled from a Beta distribution $B(\alpha, \alpha)$. One underlying concept behind mixup is that by employing linear interpolation among data points, we encourage the network to smoothly and seamlessly transition between data points, minimizing abrupt changes. Mixup improves deep neural network performance by enhancing robustness to adversarial attacks and promoting better generalization through data-adaptive regularization, ultimately leading to more accurate and reliable model predictions.

2.2 Proposed Method

Let us consider the K-class supervised classification problem with a training dataset $D = \{(x_i, y_i)\}_{i=1}^N$, where N is the total number of datapoints, x_i is the i^{th} image datapoint and y_i is the corresponding category label, such that $y_i \in L = \{1, 2, 3, \ldots, K\}$. Let f denote the classifier with parameter θ, which we need to train. The training set D is imbalanced with $n_1 > n_2 > \cdots > n_K$. Along with this we also have a meta set, $\mathcal{M} = \{(x_i, y_i)\}_{i=1}^M$, with total M datapoints. And also have heuristic function $\mathcal{H} : \mathcal{M} \rightarrow \{h_1, h_2, \cdots, h_K\}$, which gives heuristics corresponding to each class. The heuristic value increases in proportion to the degree of representation refinement achieved by the network or the level of accuracy attained by the class. This research proposes a novel imbalanced class learning approach that leverages a synergistic combination of instance sampling, adaptive sampling, and Mixup augmentation. Instance sampling ensures that all classes have a base representation during training. Adaptive sampling dynamically adjusts sampling probabilities based on a heuristic function utilizing a meta-set. This allows the model to focus on informative minority class examples. At the same time, the heuristic function, informed by the network's state and the meta-set, guides the sampling process to prevent overfitting the majority classes.

First, let us define the samplers to sample data points. The first sampler we will be using is the instance sampler where the probability of sampling from the Dataset D depends on the actual number of samples in the different classes, that is, the probability of choosing samples from the particular class k, $p_k = \frac{n_k}{N}$, where the Instance sampler denoted as $I(D, 1)$. This sampling approach will help to leverage the contexts present in the majority samples to enhance the limited context of the minority samples during the augmentation process.

The other sampler we will be using is the adaptive sampler $A(D, \mathcal{H})$, where the probability of sampling from the particular classes, depends on the heuristic calculate corresponding to each class using the meta-set $\mathcal{M}$, using the heuristic function as $\mathcal{H}(\mathcal{M}, f_\theta) = \{h_1, h_2, \cdots, h_K\}$, the calculate heuristic depend on the state of the network f and the meta set $\mathcal{M}$. The probability corresponding to each class for sampling calculated as $p_k = 1 - \frac{h_k}{\sum_{i=1}^{K} h_i}$. This technique dynamically adjusts the sampling probability based on a 'heuristic' function. This function considers the current state of the learning model and a separate 'meta-set' of data. The adaptive sampler refines the sampling process based on the model's learning progress. It can potentially identify classes still challenging for the model to learn, even within the minority class, and prioritize those for further training. This can lead to more focused learning and faster improvement in difficult classes.

For our augmentation method, the data is first sampled from instance sampler, $(x_{i_I}, y_{i_I}) \sim \mathcal{I}(\mathcal{D}, 1)$ and adaptive sampler, $(x_{i_A}, y_{i_A}) \sim \mathcal{A}(\mathcal{D}, f_\theta)$, and combined using mixup as follows:

$$\tilde{x}_{IA} = \lambda x_{i_I} + (1 - \lambda) x_{i_A}, \tag{3}$$

$$\tilde{y}_{IA} = \lambda y_{i_I} + (1 - \lambda) y_{i_A}. \tag{4}$$

This heuristic augmentation provides a better representation of the minority class by not neglecting classes that are not well-learned but instead striving to make the accuracy of such classes comparable to that of the majority classes using samples from the adaptive sampler. Furthermore, the data augmentation technique cleverly leverages the rich context found in abundant examples (majority samples) using an instance sampler to enhance the limited context surrounding the rarer examples (minority samples). By incorporating these additional details in the creation of new training data, we can significantly enhance our model's understanding of minority samples.

The Meta-set we are using does not necessarily need to belong to the same dataset as the training data. It might belong to a different dataset with varying skin tones, or it could be a combination of data similar to the training samples or samples with different skin tones. Thus, this data is not directly involved in the training samples used by the network for learning, but it will help to refine the heuristic. It will affect how the samples are chosen during augmentation, thus indirectly modifying the decision boundaries without exposing those samples to the network during training. The proposed framework is illustrated in Figure 1.

3 Experimental Analysis

3.1 Datasets, training protocol, comparison and evaluation metrics.

Dataset Used We have used the Asan Dataset [5], mainly containing patients from the Asia with darker skin tone, and the ISIC-2018 dataset [15] containing mainly caucasian patients with pale skin tone. We divide the Asan Dataset training set into two parts: 10% of the images are used as a Meta-set, while the remaining 90% are used for training. The provided test-set images are used for testing. ISIC-2018 skin lesion classification challenge adopted the HAM10000 dataset (HAM) as a training dataset. The HAM dataset is one of the largest and most used skin image datasets publicly available in the ISIC archive. It consists of 10,015 skin lesion images in seven skin lesion types. The test set comprises 1512 skin lesion images without published labels. The only method for performance evaluation is to upload the predicted results to the ISIC website. So, we divided the training set into three splits in the ratio 70:10:20 (train:meta:test). For both datasets, we convert the images to a size of 100x100px. The Asan dataset has 12 classes, and the ISIC-2018 dataset has 7. The classes both data have in common are five, namely melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, and dermatofibroma.

Training Details We utilize the CNN ResNeXt-50 (as used in [8]) for image classification. We trained a network using stochastic gradient descent (SGD) with a batch size of 128 for 100 epochs, starting with a learning rate (LR) of 0.01 that decayed if metrics didn't improve (via Reduce LR on plateau). We used early stopping if the LR reached 0, and set α to 0.2 for mixup ($B(\alpha, \alpha)$). We evaluated the performance of each model using task-specific metrics on a separate validation set, and we retained the best-performing model for further analysis. The code is available at: `https://github.com/fa-submit/Submission_M`.

Comparison Methods Not much work has addressed both the imbalanced problem and skin-tone bias together. To compare, we selected cost-sensitive methods: Reweighing-based methods like simple reweighing (RW) (weighting by the inverse of class frequencies), focal loss (FL), and class balanced reweighing (CBRW). We also considered resampling-based methods such as class balanced sampling (CBS), reverse sampling (RS), and progressive sampling (PS) - a combination of CBS and RS. Additionally, we examined Class balanced retraining, which adjusts sampling probabilities using the inverse frequency of each class raised to the power of $(1/8)^{\text{th}}$. We also evaluated balanced mixup (BalMixup) and mixup-based methods.

Evaluation Metrics We measure balanced accuracy (Bacc) (i.e. average recall of all classes). The macro-F1 score reflects improvements in smaller categories. We also include the geometric mean (GM) score. The heuristic used to determine sampling probabilities is calculated using per-class accuracy (i.e., for example, h_1 represents the accuracy of class 1). We assess model fairness using multi-class equalized odds (Eodd) and equalized opportunity (Eopp) [6]. Eopp0 evaluates the disparity in the True Negative Rate, Eopp1 evaluates the disparity in the True Positive Rate, and Eodd sums the disparities in the True Positive and False Positive Rates.

3.2 Results and Discussions

To demonstrate the effectiveness of our proposed model, we conducted several experiments. In the initial experiment, we aimed to address the existing imbalance in the dataset while enhancing performance. We accomplished this by training the ResNeXt-50 model using our proposed framework with the Asan dataset, comprising 12 classes. Subsequently, we compared its performance with other methods (as shown in Table 1) using the test set of the Asan dataset. Our findings indicate that our model exhibits superior performance compared to other methods. Additionally, when tested on the ISIC-2018 dataset, it demonstrates better performance compared to alternative methods. However, it is important to note that the ISIC-2018 dataset only shares 5 classes with the Asan dataset. To address this limitation, we utilized another subset of the Asan dataset containing the 5 common classes with the ISIC-2018 dataset. Furthermore, to showcase the generalizability of our trained model across various skin tones, we evaluated its performance on the entire ISIC dataset. Our method outperforms other methods, underscoring its superior generalizability. The lower value of fairness metrics in Table 2 highlight our method's predictive fairness compared to others, demonstrating its effectiveness. We demonstrate the performance of our main proposition regarding the use of the meta-set to regulate the sampling of samples in Table 3. We present various compositions of the meta-set and the corresponding results on the Asan test set and the ISIC-2018 dataset. Since the

Table 1. The ResNeXt-50 network was trained on the Asan dataset and evaluated on two sets: one with all 12 classes and another with 5 classes common to the ISIC-2018 dataset. The evaluation metrics include F-1 score (in %), GM (in %), and Bacc (in %).

Method	12 Classes						5 Classes					
	Tested on the Asan Datasetset			Tested on the ISIC-2018 Dataset			Tested on the Asan Dataset Test Set			Tested on the ISIC-2018 Dataset		
	F1-score	GM	Bacc	F1-score	GM	Bacc	F1-score	GM	Bacc	F1-score	GM	Bacc
RW	59.18	74.78	57.89	8.92	25.06	26.71	79.75	86.3	78.48	28.71	58.86	40.12
FL	57.04	73.4	55.88	7.47	22.22	21.19	77.83	85.43	77.15	27.8	59.82	41.67
CBRW	58.62	74.72	57.88	8.04	23.63	23.75	79.84	86.45	78.72	33.37	59.35	40.95
CBS	58.13	74.11	56.94	7.96	23.95	24.52	**82.38**	**88.12**	**81.46**	22.67	55.93	36.79
RS	59.56	74.95	58.14	7.54	22.28	21.2	81.26	87.07	79.59	31.95	62.4	44.32
Cbrt	58.51	74.64	57.68	8.15	23.44	23.49	79.67	86.68	79.01	29.96	60.68	42.21
PS	56.91	73.7	56.33	7.81	25.33	27.36	81.1	87.61	80.54	26.67	56.81	37.62
BalMixup	60.04	75.91	59.62	9.77	26.64	30.18	81.10	87.25	79.58	34.25	60.91	42.63
Mixup	56.09	72.39	54.36	8.49	25.03	26.67	82.34	88.04	80.92	25.61	55.98	36.24
Ours	**61.05**	**77.54**	**62.16**	**10.36**	**28.97**	**35.5**	79.84	86.95	79.53	**40.57**	**65.33**	**47.81**

Table 2. Fairness results of different methods trained on the ASAN dataset (5 classes), tested on a combined ASAN test set and ISIC-2018 dataset.

Method(→)
Metric(↓)
	RW	FL	CBRW	CBS	RS	Cbrt	PS	BalMixup	Mixup	**Ours**
										(Meta-Set
										10% Asan)
											Ours
											(Meta-Set
											10% Asan
											+ 10% ISIC)

EOpp0	0.115	0.087	0.108	0.079	0.115	0.083	0.111	0.100	0.105	**0.075**	**0.055**
EOpp1	0.489	0.355	0.378	0.353	0.489	0.386	0.429	0.37	0.447	**0.317**	**0.336**
EOdds	0.298	0.185	0.19	0.186	0.298	0.200	0.241	0.164	0.185	**0.153**	**0.146**

Table 3. The table presents the ResNeXt-50 Network trained on the Asan Dataset, showcasing how results (in %) change with varying compositions of the Meta-Set.

Composition of Meta-Set	Tested on the Asan Data Test Set			Tested on the ISIC-2018 Dataset		
	F1-score	GM	Bacc	F1-score	GM	Bacc
10% of Asan Dataset	79.84	86.95	79.53	40.57	65.33	47.81
10% of ISIC Dataset	74.54	83.01	73.62	34.25	59.55	39.98
10% of ISIC Dataset+10% of Asan Dataset	**82.34**	**88.34**	**81.62**	**40.60**	**66.00**	**49.11**
20% of ISIC Dataset	73.31	82.69	72.81	33.37	59.91	41.37
30% of ISIC Dataset	74.81	83.87	75.17	34.95	61.98	43.58

images in the meta-set are not directly used for training, the entire ISIC dataset can be used for testing. From the table, it is evident that the optimal composition involves utilizing both the Asan and ISIC-2018 datasets in the meta-set, resulting in an improvement in performance of nearly 2% in terms of Bacc for both test cases. Additionally, there is an enhancement of approximately 3% in the F-1 score for the Asan dataset when compared with the meta-set containing only Asan dataset images, indicating an improvement in the performance of the minority class. These results show that performance can be enhanced by using a small number of images in the meta-set, even without directly including images of different skin tones in the training set. Regulating the sampling process aids in better augmentation, resulting in improved and more generalizable decision boundaries. We also computed results on the ISIC-2018 dataset using a subset of 5 classes common with the Asan dataset. We used the trained network to evaluate results on the Asan dataset and the ISIC-2018 test set (Table 4). Our method performs well in both test cases, with nearly a 3% improvement in both Bacc and GM, albeit with a slight dip in the F-1 score compared to RW and Mixup. Overall, our method aims for balanced performance across all classes, enhancing generalization to unseen data, thus showing better performance in unseen skin-tone images than other methods. Fairness metrics are excluded in

Table 4. The results of the ResNeXt-50 network trained on the ISIC-2018 dataset are evaluated using F-1 score (in %), GM (in %), and Bacc (in %).

Method	Tested on the ISIC-2018 Dataset Test Set			Tested on the Asan Dataset		
	F1-score	GM	Bacc	F1-score	GM	Bacc
RW	**69.08**	81.61	71.53	**38.9**	55.97	35.98
FL	67.43	78.94	66.96	21.24	48.97	29.07
CBRW	69.27	81.94	71.6	31.17	52.18	32.53
CBS	62.48	74.13	59.79	22.15	50.48	30.67
RS	62.48	74.13	59.79	30.61	53.19	33.84
Cbrt	60.73	73.28	58.6	17.85	44.32	24.23
PS	62.61	74.51	60.1	25.88	51.17	31.63
BalMixup	65.89	75.90	62.77	14.95	41.97	21.85
Mixup	70.32	78.92	66.45	22.71	47.32	27.10
Ours	65.86	**83.6**	**74.79**	38.77	**57.4**	**38.77**

this case because the ISIC test set's minority class has far fewer samples than the ASAN dataset's minority, making estimates for the underrepresented group potentially unreliable.

4 Conclusion and future work

In conclusion, this study presents an innovative data augmentation technique that addresses the challenges of skin tone bias and imbalanced class distribution in deep learning models for skin lesion classification simultaneously through a novel adaptive mixup sampling strategy that uses cross sampling between the diverse skin tones in a judicious manner. To demonstrate the efficacy of our method across different skin tones and class imbalance (skin cancer is more common among caucasian patients and hence usually data available is biased towards that patient demographic), we choose two benchmark datasets: ISIC-2018 with mostly caucasian patients and Asan dataset with mostly Asian patients. Our results showcase the accuracy of our method compared to several recent competing approaches, and thus presents a classifier that can generalise across patient demographics with fairness and equity, and hence has the potential of practical translation across borders in clinical decision support systems (CDSS).

Disclosure of Interests. The authors have no competing interests in the paper as required by the publisher.

References

1. Chou, H.P., Chang, S.C., Pan, J.Y., Wei, W., Juan, D.C.: Remix: rebalanced mixup. In: European Conference on Computer Vision. pp. 95–110. Springer (2020)
2. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)
3. Galdran, A., Carneiro, G., González Ballester, M.A.: Balanced-mixup for highly imbalanced medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 323–333. Springer (2021)
4. Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., Badri, O.: Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1820–1828 (2021)
5. Han, S.S., Kim, M.S., Lim, W., Park, G.H., Park, I., Chang, S.E.: Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. Journal of Investigative Dermatology **138**(7), 1529–1538 (2018)
6. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
7. Huang, C., Li, Y., Loy, C.C., Tang, X.: Learning deep representation for imbalanced classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5375–5384 (2016)

8. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big transfer (bit): General visual representation learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. pp. 491–507. Springer (2020)
9. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
10. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., Van Der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: Proceedings of the European conference on computer vision (ECCV). pp. 181–196 (2018)
11. Park, S., Hong, Y., Heo, B., Yun, S., Choi, J.Y.: The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6887–6896 (2022)
12. Rao, A., Lee, J.Y., Aalami, O.: Studying the impact of augmentations on medical confidence calibration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2462–2472 (2023)
13. Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., et al.: Balanced meta-softmax for long-tailed visual recognition. Advances in neural information processing systems **33**, 4175–4186 (2020)
14. Shen, L., Lin, Z., Huang, Q.: Relay backpropagation for effective learning of deep convolutional neural networks. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14. pp. 467–482. Springer (2016)
15. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
16. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019)
17. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)