2023 IEEE 6th International Conference on Multimedia Information Processing and Retrieval (MIPR) | 979-8-3503-0781-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/MIPR59079.2023.00030

Toward Fairness Across Skin Tones in Dermatological Image Processing

Abdulaziz A. Almuzaini*[1,3], Srujan K. Dendukuri*[1], and Vivek K. Singh[2]

[1]Department of Computer Science, Rutgers University
[2]School of Communication and Information, Rutgers University
[3]College of Computer Science and Information Systems, Islamic University of Madinah
{*a.almuzaini, sd1278, v.singh*}*@rutgers.edu*

Abstract

Skin cancer is a prevalent and concerning form of cancer, with an annual incidence rate estimated to be more than 3 million cases in the US. In recent years, the field of medical image processing has made a remarkable progress in the domain of skin cancer detection, surpassing the diagnostic capabilities of dermatologists in certain settings. However, it has been reported that the performance of these deep learning detection models varies significantly across different skin tones (e.g., light versus dark), motivating the need for fair and unbiased classification results. Here, we evaluate DeepDerm [10], a state-of-the-art skin cancer detection model, specifically focusing on its performance across skin types classified by the Fitzpatrick Skin Tones (FST). By analyzing the model's accuracy and fairness, we observe notable discrepancies in its performance across different FST categories. We propose a novel architecture that leverages fine-tuning, an ensemble architecture, and fairness-based resampling for supporting high accuracy and fairness in skin cancer detection. The proposed framework demonstrates promising outcomes, marking a significant stride toward achieving fairness and accuracy in dermatological image processing.

1. Introduction

Multimedia processing is rapidly impacting the healthcare industry, enabling a variety of health-related and medical diagnostic applications to enhance healthcare. These applications include chest X-ray diagnosis [9], brain MRI segmentation [15], and skin cancer detection [11, 18]. In particular, skin cancer detection is one of the crucial applications that can benefit from multimedia processing, as it can reduce the reliance on human experts, improve the accuracy performance, and facilitate the sharing of information. With an estimated annual incidence rate of more than 3 million cases in the US, skin cancer is a common and alarming form of cancer [5]. Early detection and diagnosis are crucial for improving the survival rate and reducing the treatment costs of this disease. However, the availability and accessibility of dermatologists are limited, especially for vulnerable populations and in rural areas. Therefore, there is a growing demand for automated and reliable methods for skin cancer detection using digital images.

However, building accurate and robust deep learning models for skin cancer detection requires large and diverse datasets, which are often scarce or private in the field of medical imaging. A common solution to this challenge is to use transfer learning, which allows adapting the structure and parameters of pre-trained models on different domains (e.g., GoogleNet-Inception-V3 [24]) to the specific task of skin cancer detection. Yet, transfer learning alone may not be sufficient to ensure high performance and generalization across different subgroups of patients, as it may introduce biases and disparities that affect the fairness and inclusiveness of the system. For instance, it has been reported that some deep learning models for skin cancer detection exhibit unfair behaviors towards minority or disadvantaged groups (e.g., female, black, etc.) if not carefully diagnosed and constrained [10, 1].

In this paper, we address the problem of building an accurate and fair model for skin cancer detection using multimedia processing. We follow the transfer learning approach to fine-tune a state-of-the-art model, DeepDerm [10], on a large and diverse dataset of skin images, combining two ex-

*Equal Contribution

2770-4319/23/$31.00 ©2023 IEEE
DOI 10.1109/MIPR59079.2023.00030

isting benchmarks from the field of skin dermatology: DDI and Fitzpatrick17k datasets [10], [17]. We evaluate the performance and fairness of DeepDerm on different subgroups of patients, classified by their Fitzpatrick Skin Tones (FST) [14]. The FST is a widely used scale that categorizes human skin tone into six types, ranging from very light (FST I) to very dark (FST VI). Following [10], we analyze the accuracy and fairness performance of DeepDerm on different FST groups (FST I-II and FST V-VI), and we observe notable discrepancies in its performance. We find that DeepDerm tends to favor lighter skin tones (FST I-II) over darker ones (FST V-VI), resulting in higher false positives and lower true positives for the latter group. This implies that DeepDerm may misdiagnose or miss malignant lesions in patients with darker skin tones, potentially leading to adverse health outcomes and reduced trust in the system.

To overcome this limitation, we propose a novel **SAFE** (*S*kin cancer detection with *A*daptive *F*airness-aware *E*nsemble) framework that aims to improve both the accuracy and fairness of skin cancer detection across skin tones. Our framework consists of three main components: (1) fine-tuning, (2) ensemble architecture, and (3) fairness-based resampling. First, we fine-tune DeepDerm on a subset of images from different FST categories separately, creating specialized models that are more tailored to the specific characteristics of each skin tone. Second, we combine these specialized models into an ensemble architecture that leverages their complementary strengths and reduces their individual weaknesses. Third, we apply a fairness-based resampling technique that balances the distribution of skin tones in the training data, mitigating the effects of data imbalance and underrepresentation. We evaluate our framework by comparing the performance and fairness metrics on different FST based groups. We report that our framework achieves notable improvements over DeepDerm, and its full layers fine-tuned version in both aspects, demonstrating its effectiveness and robustness for fair and accurate skin cancer detection.

The main contributions of this paper are:

(1) To audit the performance and fairness of DeepDerm, a state-of-the-art skin cancer detection model, across different skin tones classified by the FST using a custom-aggregated large, diverse dataset.

(2) To propose a novel SAFE framework that leverages fine-tuning, ensemble architecture, and fairness-based resampling for improving both the accuracy and fairness of skin cancer detection across skin tones.

This work contributes to the growing literature on the need for fairness in multimedia processing algorithms [22, 2]. Such bias has been reported in tasks such as face matching, cyberbullying detection, and pedestrian detection [8, 2, 7]. Similarly, several bias reduction techniques have been proposed that include building fairness-centric adversarial networks [2], reweighing multimodal features [3], reweighing and combining multiple model outputs [4], or creating new mid-level representations that maintain high prediction accuracy but minimize the correlation with sensitive/demographic attributes [6]. In this work, we use a combination of fine-tuning, adaptive reweighing and ensembling.

Transfer learning is increasingly being used for dermatological image processing [13, 10]. Esteva et al., showed that such an approach, called DeepDerm (fine-tuned from Google-Inception-v3 CNN architecture pre-trained on the ImageNet dataset) can outperform human dermatologists in certain settings [13]. However, Daneshjou et al., showed that when tested over a Diverse Dermatological Images (DDI) dataset, DeepDerm yielded modest accuracy and significant differences in performance for different skin tones. They further showed that another round of fine-tuning over such diverse images helped with improving the fairness and accuracy. However, the DDI dataset was small (e.g., only 48 images for the malignant category for dark skin) and this could be a limiting factor in the validity, accuracy, and fairness of their approach. Hence, in this work, we complement DDI dataset with another similar dataset (Fitzpatrick17k), and propose a SAFE approach for fairness and robustness in dermatological image processing.

2. Methodology

2.1 Problem Formulation

We consider a dataset $\{(x_i, a_i, y_i)\}$ that has n samples. Each sample consists of an image (x), a sensitive attribute (a) and a target label (y). A sensitive attribute is a binary random variable where a can take values of $\{a^+, a^-\}$ which represents *advantaged* and *disadvantaged* demographic group, respectively. The definition of advantaged group depends on the societal context and can include aspects like race, gender, skin tone, age, etc. [19]. Here, we consider skin tone as the sensitive attribute. Similarly the target label is $y = \{y^+, y^-\}$, where *+, (-)* represents *positive class (negative class)*, respectively. In this work, we consider x as a photographed skin image, a is a binary value of whether an image belongs to a light (FST I-II)/dark (FST V-VI) skin tone [17]. Lastly, y represents the presence of cancer or not (i.e., benign versus malignant). The prediction algorithm f will map x to y, i.e., $f : x \rightarrow y$.

In such a setting our goal is to maximize two factors: (i) prediction correctness, and (ii) prediction fairness. We use four different metrics to quantify correctness: accuracy, area under the ROC curve (AUROC), true positive rate (TPR) and false positive rate (FPR). Consequently, following extant literature on group fairness metrics [20], we examine the disparity (i.e., Δ) of the aforementioned

metrics performance on the two groups (e.g., $\Delta_{TPR} = |TPR_{FST(I-II)} - TPR_{FST(V-VI)}|$) to quantify fairness. In effect, we quantify fairness as the lack of disparity in algorithm's performance based on the sensitive attribute [19].

These accuracy and fairness metrics are not collapsible into a single score and may even be impossible to optimize at the same time [21] and we will look for trends in their values when comparing models. Specifically, we prefer models that score high (↑) on accuracy, AUROC, and TPR, while scoring low (↓) on FPR, $\Delta_{accuracy}$, Δ_{AUROC}, Δ_{TPR}, Δ_{FPR}.

Dataset	Label	Sensitive		# Label
		FST I-II	FST V-VI	
DDI	Benign	159	145	304
	Malignant	49	48	97
	# Sensitive	208	193	
Fitzpatrik17K	Benign	359	77	436
	Malignant	775	181	956
	# Sensitive	1134	258	
Combined	Benign	518	222	740
	Malignant	824	229	1053
	# Sensitive	1342	451	

Table 1. Datasets Summary

2.2 Datasets

In this paper, we use two datasets, **Diverse Dermatology Images (DDI)** [10] and **Fitzpatrick17k** [17]. Both datasets are free for scientific, non-commercial use. **DDI** is a clinically curated and pathologically confirmed image dataset with diverse skin tones. The dataset is a compilation of benign and malignant lesions diagnosed in Stanford clinics between 2010 to 2020, which were reviewed histopathologically. The dataset consists of 656 images representing 570 unique patients [10]. For each image, the Fitzpatrick Skin Type (FST) was identified. The Fitzpatrick scale is a numerical classification schema for human skin tone. It ranges from I to VI and is a way to estimate the response of different types of skin to ultraviolet light (UV) [16, 17]. FST I-II represents light skin tones and FST V-VI represents dark skin tones while FST III-IV represents a class which lies in-between [17]. For this study, we focus on FST I-II and FST V-VI images. The unique positive aspect of the DDI dataset is the balance between number of samples of different skin tones. A major challenge is the small sample size, especially for the malignant class. The **Fitzpatrick17k** dataset consists of 16,577 images which have been collected from two dermatological atlases - DermaAmin and Atlas Dermatological - and the labels were assigned via open-source annotation platforms [17]. Similar to DDI, we focus on FST I-II and FST V-VI images. To maintain consistency, we pick diseases from Fitzpatrick17k which are common with DDI. As the dataset summary shows in Table 1, we can see the DDI dataset is small but relatively balanced for the sensitive attribute and the class label whereas the Fitzpatrick17k is larger but shows an imbalance behavior. Hence, we have decided to combine these two sources into a 'combined dataset' for the purpose of this study.

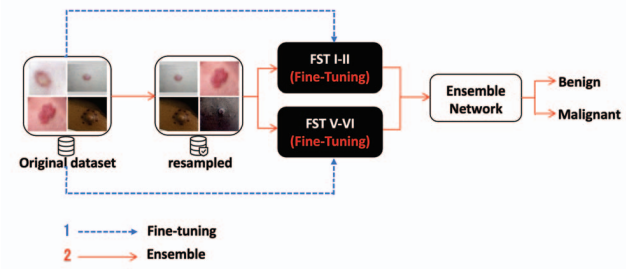


Figure 1. The Proposed SAFE Framework

2.3 Pre-Processing

Most of the images were originally cropped as part of de-identification. However, some images still contained background which needed to be removed, therefore all of the images were center-cropped. We have followed a 60-20-20 split for training, validation and testing. The training set was further transformed with random rotations, vertical flips, Gaussian blurring and color jitters. Furthermore, all splits have been normalized for model generalizability. During training we make use of mixup data augmentation with an alpha value of 0.2 [25]. Mixup data augmentation is used to augment training data by blending pairs of inputs and their corresponding labels to create virtual samples. This encourages a model to learn more robust and generalizable representations by exposing it to various combinations of features and labels.

2.4 Proposed Approach

In the proposed approach, we use DeepDerm as the starting point and undertake fine-tuning on it before creating an ensemble where reweighing is adopted to support fairness and accuracy. Thus, DeepDerm and full-layers fine-tuning work as baselines as well as conceptual building blocks for the proposed approach.

2.4.1 DeepDerm (Baseline 1)

DeepDerm uses Inception-V3 model architecture [24]. Inception-V3 consists of 42 layers and is known to be computationally efficient compared to its predecessors. While most models struggle to estimate the correct kernel size, Inception-V3 tries multiple kernels of different dimensions

Method	Metrics	All	I-II	V-VI	$\Delta \downarrow$
DeepDerm (Baseline)	Accuracy (%) ↑	78.60	79.93	75.00	4.93
	AUROC (%) ↑	85.90	87.21	82.33	4.88
	TPR (%) ↑	83.80	84.24	82.61	**1.63**
	FPR (%) ↓	28.60	26.92	32.61	5.69
Full-Layers (Fine-Tuning)	Accuracy (%) ↑	77.84 ±1.00	77.94 ±0.86	77.54 ±5.14	4.06 ±2.93
	AUROC (%) ↑	85.93 ±1.80	86.26 ±1.88	84.58 ±2.49	**2.10** ±1.48
	TPR (%) ↑	84.68 ±4.30	87.27 ±5.78	75.36 ±9.05	11.91 ±12.23
	FPR (%) ↓	31.78 ±6.16	36.86 ±6.96	20.29 ±4.53	16.57 ±2.99
SAFE (Proposed)	Accuracy (%) ↑	**85.13** ±0.16	**85.87**±0.37	**82.97**±0.63	**2.90**±0.97
	AUROC (%) ↑	**91.80**±1.00	**91.98**±1.49	**90.94**±1.13	2.21±0.96
	TPR (%) ↑	**87.36**±0.27	**88.48**±0.00	**83.33**±1.26	5.15±1.26
	FPR (%) ↓	**18.00**±0.67	**18.27**±0.96	**17.39**±0.00	**0.93**±0.88

Table 2. Performance of different models across different evaluation metrics. All indicates the full test set performance, I-II indicates the test set performance only for lighter skin subgroup and V-IV indicates the test set performance for the darker skin subgroup. Δ represents the absolute value of performance differences of (I-II and V-VI) subgroups for that metric. Values shown are the mean and the standard deviations of 3 different runs.

and concatenates the output. We use a pre-trained model provided by [13, 10]. We fine-tune our models using the SGD optimizer with the learning rate of 0.005 and a weight decay of 0.0001, and train for 200 epochs. We use the validation data to pick the best model during training.

2.4.2 Full-Layers Fine-Tuning (Baseline 2)

Since DeepDerm has been pre-trained on a different benchmarking dataset than ours, we follow [10] to apply the transfer learning approach by fine-tuning all-layers of DeepDerm architecture based on our combined dataset. Specifically, we keep the same architecture but update the model parameters based on the training of our dataset. This is similar to the approach adopted by [10] to reduce skin tone based disparities in DeepDerm.

2.4.3 SAFE (*S*kin cancer detection with *A*daptive *F*airness-aware *E*nsemble) Model

In this work, we adopt an ensemble approach with the aim of achieving both fairness and accuracy in cancer image detection. In a well-constructed ensemble model, each individual model should make predictions based on different aspects of the data or different algorithms, so that the ensemble can produce a more comprehensive and accurate prediction.

Here, we build upon the "decoupled classifiers for fairness" approach [12], to train (i.e., fine-tune all layers) two different DeepDerm models - one model on the FST I-II data and one model on the FST V-VI data - and then combine the models within an ensemble framework. The ensemble framework is summarized in Figure 1. It consists of two stages: (1) Individual Fine-Tuning and (2) A Fair Ensemble. To train a fair ensemble network, having separate fine-tuned models might not suffice, especially when the training dataset suffers from a significant imbalance with respect to both the sensitive attribute and the label as shown in Table 1. Therefore, to further support fairness and accuracy, we train the ensemble to ensure that the ratio of sensitive attribute (light/dark skin) and disease labels (benign/malignant) is equalized during the training process. Specifically, it will weight each sample proportionally to the inverse of its class/sensitive attributes frequency. Therefore, minority samples (i.e., benign or dark skin) will be *resampled* more frequently compared to the majority. Note that the ratio of the samples and labels for the test data is not altered. The intuition behind this idea is to first *specialize* disease classification for skin tone and then *generalize* the results across skin tones by training the ensemble weights. In such a scenario, the ensemble framework will likely be exposed to each subgroup equally during re-training; thus getting equal opportunity to learn about classifying skin diseases for the considered subgroups.

3 Results and Discussion

3.1 Fairness Audit Across Skin Tones in Dermatological Image Processing

As a first step, we check if the existing DeepDerm algorithm exhibits algorithmic bias. In other words, does it yield statistically significant differences in its performance

Method	Metrics	All	I-II	V-VI	$\Delta \downarrow$
Ensemble (Vanilla)	Accuracy (%) ↑	85.23 ±0.16	85.87 ±0.37	83.33 ±0.63	2.54 ±0.97
	AUROC (%) ↑	86.16 ±10.15	87.65 ±8.43	80.69 ±16.25	6.96 ±8.04
	TPR (%) ↑	87.99 ±0.27	89.09 ±0.61	84.06 ±1.26	5.03 ±1.81
	FPR (%) ↓	18.67 ±0.00	19.23±0.00	17.39±0.00	1.84 ±0.00
Ensemble (Label)	Accuracy (%) ↑	**85.41 ±0.32**	86.00 ±0.43	83.70 ±0.00	2.30 ±0.43
	AUROC (%) ↑	89.96 ±3.48	89.75±4.87	90.50±0.11	3.15±3.91
	TPR (%) ↑	**88.47±0.27**	89.49±0.35	84.78±0.00	4.71±0.35
	FPR (%) ↓	18.89±0.38	19.55±0.56	17.39±0.00	2.16±0.56
Ensemble (Sensitive)	Accuracy (%) ↑	85.32±0.28	85.87±0.37	83.70±0.00	**2.18±0.37**
	AUROC (%) ↑	**92.44±0.04**	93.16±0.04	89.32±0.08	3.84±0.05
	TPR (%) ↑	88.15±0.47	89.09±0.61	84.78±0.00	**4.31±0.61**
	FPR (%) ↓	18.67±0.00	19.23±0.00	17.39±0.00	1.84±0.00
SAFE (Proposed)	Accuracy (%) ↑	85.13 ±0.16	85.87±0.37	82.97±0.63	2.90±0.97
	AUROC (%) ↑	91.80±1.00	91.98±1.49	90.94±1.13	**2.21±0.96**
	TPR (%) ↑	87.36±0.27	88.48±0.00	83.33±1.26	5.15±1.26
	FPR (%) ↓	**18.00±0.67**	18.27±0.96	17.39±0.00	**0.93±0.88**

Table 3. Ablation Study: Performance for different variants/components of the proposed approach.

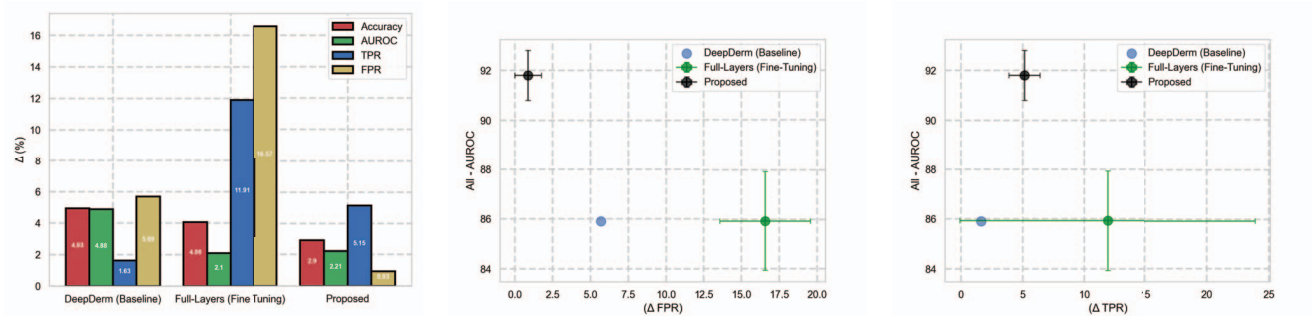
for groups that have lighter and darker skin tones. To test the hypothesis that the DeepDerm has a *significant* bias treating different skin tones differently, we use a t-test that compares the means for the two subgroups. To do so, for each subgroup we randomly and independently subsample 50 images and examine the DeepDerm average performance on each sample. Repeating the process 30 times, we use a significance level $\alpha = 0.05$ to test the null hypothesis. For comparisons made in terms of each of the four performance metrics, we found that the p-value for the t-test is less than 0.05 (Accuracy: $p < .001$, AUROC: $p < .001$, TPR: $p < .015$ and FPR: $p < .020$) and the performance was better for the lighter skinned group than the darker skinned group. Hence, the results indicate that there is a *significant difference* in the performance of the DeepDerm classifier for different skin tones. This is consistent with the results reported by Daneshjou et al. [10] on the performance of DeepDerm across skin tones. It is also consistent with other recent results which have reported difference in performance of image processing algorithms based on skin tone [8, 18], and motivate the need for approaches to increase fairness across skin tones in image processing and multimedia applications.

3.2 Accuracy and Fairness Performance of the Proposed Approach

Next, we examine the performance of the DeepDerm model, the fine-tuning model, and the SAFE ensemble model on the whole test set and on (FST I-II, FST V-VI) subsets. For each model, we compare the performance based on the aforementioned four metrics and the difference in those metrics for samples belonging to the two different skin tones considered (FST I-II versus FST V-VI). The results (see Table 2) for each approach are based on the same test set. For Full-Layers and the proposed SAFE model, the results are an average across three runs, where each run had a different selection of training and validation sets. Since, no training/validation is done for DeepDerm (Baseline), it's results are reported on a single test set run.

In Table 2, we can see the baseline has a moderate overall accuracy (78.60%) and tends to perform well for lighter skin tone compared to the dark skin tone (79.93% versus 75.00%). The same trend is persistent with AUROC as well. Additionally, the Δ_{TPR} is the lowest for the DeepDerm (Baseline) but the Δ_{FPR} is relatively high in which the dark skin patients have the highest false positive rates. The transfer learning approach, namely, Full-Layers (Fine-Tuning) has helped in reducing the disparity between the two groups in terms of Δ_{AUROC}. However, it performs worse than the baseline in terms of overall accuracy, FPR, Δ_{TPR} and Δ_{FPR}. The proposed SAFE framework yields the best performance in terms of all four accuracy/correctness metrics for each of the demographic groups considered and the overall dataset. It also yields the best performance in terms of fairness metrics for $\Delta_{accuracy}$ and Δ_{FPR}. In all, for 14 of the 16 considered scores, the proposed approach outperformed the DeepDerm (Baseline) and the Full-Layers (Fine-Tuning) approach.

These results are also visualized in Figure 2. As a trend, we note that the proposed approach yields low bias (inverse of fairness) scores (Figure 2a). In Figure 2b, we see that if we were to consider a trade-off between overall AUROC and Δ_{FPR}, then the proposed approach will strictly dominate the other two baselines. If, instead, we consider a trade-off between overall AUROC and Δ_{TPR} then the proposed



(a) Proposed v.s. (baseline & fine-tuning)

(b) All-AUROC versus Δ_{FPR}

(c) All-AUROC versus Δ_{TPR}

Figure 2. Models performance for DeepDerm (Baseline), Full-Layers (Fine-Tuning) and the proposed method. (a) shows the disparity bar plot performance for each model in which the lower the bar the fairer the model is. (b) and (c) plot the trade-off performance of the overall AUROC in y-axis and the Δ_{FPR} and Δ_{TPR} in the x-axis where a circle represents the average and an errorbar is the standard deviation (See Table 2). The closer the model to the top-left corner the more accurate and fair.

approach will lie on a pareto curve [4] with baseline and dominate the Full-layers fine-tuning approach.

3.3 Ablation Study for the Proposed Approach

Our proposed approach has three important components: (a) ensembling two decoupled models, (b) reweighted training based on equal opportunities for the disease labels (benign/malignant), and (c) reweighted training based on the demographic groups.

To evaluate the relative impact of each of these steps on the overall output, we undertake an ablation study and report the results in Table 3. (The best results are highlighted in bold). The Vanilla Ensemble refers to the approach where the ensemble is retrained on the entire training dataset, Ensemble (Label) includes reweighing only based on labels, and Ensemble (Sensitive) includes reweighing only based on the sensitive attribute. We note that while the Vanilla Ensemble improves upon the results obtained with baseline and Full-Layers fine tuning, it falls short of the performance obtained via the reweighing based approaches. The label based reweighing approach yields the best correctness scores in terms of accuracy and TPR but does not seem to outperform others in terms of fairness. This is consistent with past literature that has suggested reweighing based on the classification label largely as a way to improve the accuracy without necessarily considering the fairness aspect [23]. The sensitive attribute based reweighing approach on the other hand focuses on the fairness aspect and outperforms the label based reweighing approach on most of the fairness metrics. The proposed approach reweighs based on both labels and the sensitive attribute and can be seen as trying to maximize a combination of fairness and accuracy based metrics. It's results are approaching the accuracy scores of the label and sensitive attribute reweighing, while also yielding the best fairness scores in terms of Δ_{AUROC} and Δ_{FPR}. It also yields the lowest overall bias in terms of an aggregated sum of the Δ scores for the four fairness metrics. This suggests that each of the components of the proposed approach has an important (albeit different) role to play in supporting its fairness and accuracy goals.

4. Conclusion and Future Work

This paper adds to a small number of attempts at auditing image processing algorithms for skin tone based bias in cancer detection. It does so using multiple datasets and the state-of-the-art DeepDerm model [10]. Specifically, it finds that the fine-tuning based approach proposed by Daneshjou et al., [10] has limitations in terms of the fairness and accuracy levels achieved. This could in part be a function of the small sample size used. This work proposes a new SAFE (*S*kin cancer detection with *A*daptive *F*airness-aware *E*nsemble) approach that utilizes decoupled learning, ensembling, and fairness aware reweighing, to yield high performance in terms of both fairness and accuracy. The empirical results demonstrate the validity of the proposed ideas. Future work in this area could include more diverse datasets, more sophisticated multimedia processing, and devising more sophisticated bias mitigation strategies. However, the proposed approach marks an important step toward achieving fairness and accuracy in dermatological image processing.

Acknowledgements

This material is in part based upon work supported by the National Science Foundation under Grant No. SES-1915790.

References

[1] A. S. Adamson and A. Smith. Machine learning and health care disparities in dermatology. *JAMA dermatology*, 154(11):1247–1248, 2018.

[2] J. Alasadi, A. Al Hilli, and V. K. Singh. Toward fairness in face matching algorithms. In *Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia*, pages 19–25, 2019.

[3] J. Alasadi, R. Arunachalam, P. K. Atrey, and V. K. Singh. A fairness-aware fusion framework for multimodal cyberbullying detection. In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, pages 166–173. IEEE, 2020.

[4] A. A. Almuzaini and V. K. Singh. Balancing fairness and accuracy in sentiment detection using multiple black box models. In *Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia*, pages 13–19, 2020.

[5] American Cancer Society. Key statistics for basal and squamous cell skin cancers, 2023. Accessed: 2023-07-10.

[6] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

[7] M. Brandao. Age and gender bias in pedestrian detection algorithms. *arXiv preprint arXiv:1906.10490*, 2019.

[8] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[9] E. Çallı, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy. Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, 72:102125, 2021.

[10] R. Daneshjou, K. Vodrahalli, R. A. Novoa, M. Jenkins, W. Liang, V. Rotemberg, J. Ko, S. M. Swetter, E. E. Bailey, O. Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31):eabq6147, 2022.

[11] M. Dildar, S. Akram, M. Irfan, H. U. Khan, M. Ramzan, A. R. Mahmood, S. A. Alsaiari, A. H. M. Saeed, M. O. Alraddadi, and M. H. Mahnashi. Skin cancer detection: a review using deep learning techniques. *International journal of environmental research and public health*, 18(10):5479, 2021.

[12] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*, pages 119–133. PMLR, 2018.

[13] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[14] T. B. Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988.

[15] S. Gassenmaier, T. Küstner, D. Nickel, J. Herrmann, R. Hoffmann, H. Almansour, S. Afat, K. Nikolaou, and A. E. Othman. Deep learning applications in magnetic resonance imaging: has the future become present? *Diagnostics*, 11(12):2181, 2021.

[16] M. Groh, C. Harris, R. Daneshjou, O. Badri, and A. Koochek. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. *arXiv preprint arXiv:2207.02942*, 2022.

[17] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, and O. Badri. 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*, pages 1820–1828, 2021.

[18] N. M. Kinyanjui, T. Odonga, C. Cintas, N. C. Codella, R. Panda, P. Sattigeri, and K. R. Varshney. Fairness of classifiers across skin tones in dermatology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI*, pages 320–329. Springer, 2020.

[19] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck. Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, 31:611–627, 2018.

[20] K. Makhlouf, S. Zhioua, and C. Palamidessi. On the applicability of machine learning fairness notions. *ACM SIGKDD Explorations Newsletter*, 23(1):14–23, 2021.

[21] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

[22] V. K. Singh, E. André, S. Boll, M. Hildebrandt, D. A. Shamma, and T.-S. Chua. Legal and ethical challenges in multimedia research. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2514–2515, 2019.

[23] S. Subramanian, A. Rahimi, T. Baldwin, T. Cohn, and L. Frermann. Fairness-aware class imbalanced learning. *arXiv preprint arXiv:2109.10444*, 2021.

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[25] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.