Does the Fairness of Your Pre-Training Hold Up? Examining the Influence of Pre-Training Techniques on Skin Tone Bias in Skin Lesion Classification

Pratinav Seth, Abhilash K. Pai
Department of Data Science & Computer Applications
Manipal Institute of Technology,
Manipal Academy of Higher Education,
Manipal, India
seth.pratinav@gmail.com, abhilash.pai@manipal.edu

Abstract

Deep Neural Networks (DNNs) have found widespread application in various domains, but the challenge of addressing Algorithmic bias and ensuring fairness in their decision-making processes has emerged as a critical concern, particularly in mission-critical contexts. One of the main reasons for this concern is the inadequate representation of certain groups in the available datasets used for training. Pre-Training is a powerful technique for training DNNs, but it can be affected by pre-existing biases in the dataset. These biases can be transferred to the DNN during Pre-Training, leading to the DNNs making biased decisions, even when trained on unbiased datasets. This study investigates the impact on the fairness of popular Pre-Training methods, such as Masked Image Modeling (MAE, SimMIM) and Self-Supervised Learning (BYOL, MoCo, SimCLR, VICRegL), when used on skin lesion classification datasets with underrepresented demographic groups. The study compares the performance of pre-trained models to supervised learning backbones on two skin lesion datasets (ISIC-2019 and Fitzpatrick17k) with different skin tone distributions. The findings of this study reveal that Pre-Training improves performance but has a trade-off with fairness, which can be a potential danger associated with the model when applied in the real world. This study is one of the first to investigate how Self-Supervised Learning and Masked Image Modeling Pre-Training methods affect fairness in both in-distribution and out-of-distribution scenarios. Code is available at https://github.com/ptnv-s/PretrainingImpactOnSkinBias.

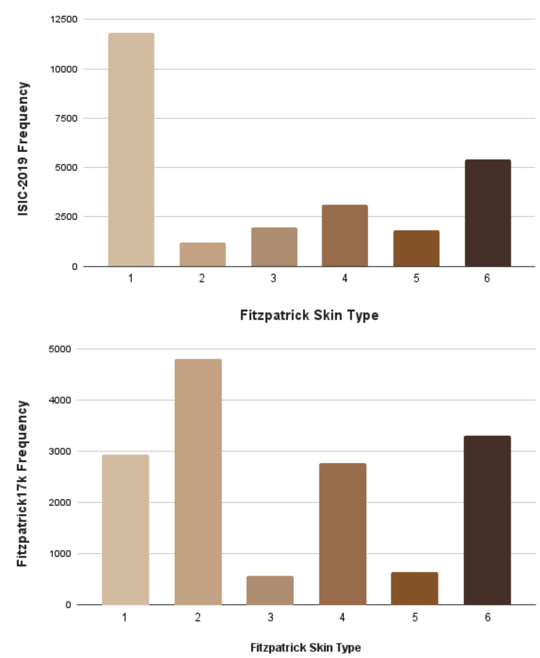

Figure 1. Fitzpatrick Labelled Skin-Types in ISIC-2019 [11] & Fitzpatrick17k Datasets

1. Introduction

In recent years, AI has profoundly transformed various aspects of our lives, including decision-making and daily activities, revolutionizing numerous domains. One critical sub-field within AI is medical data analysis, which focuses on processing and analyzing diverse medical data to extract crucial information for accurate diagnoses [20]

Deep neural networks are increasingly used in computer-aided health monitoring and diagnosis, but the need for large amounts of data is a challenge. Acquiring and annotating medical data is time-consuming and expensive, and it is especially difficult for rare or novel diseases. Cancer is the leading global cause of death, with around 10 million fatalities in 2020, representing approximately one

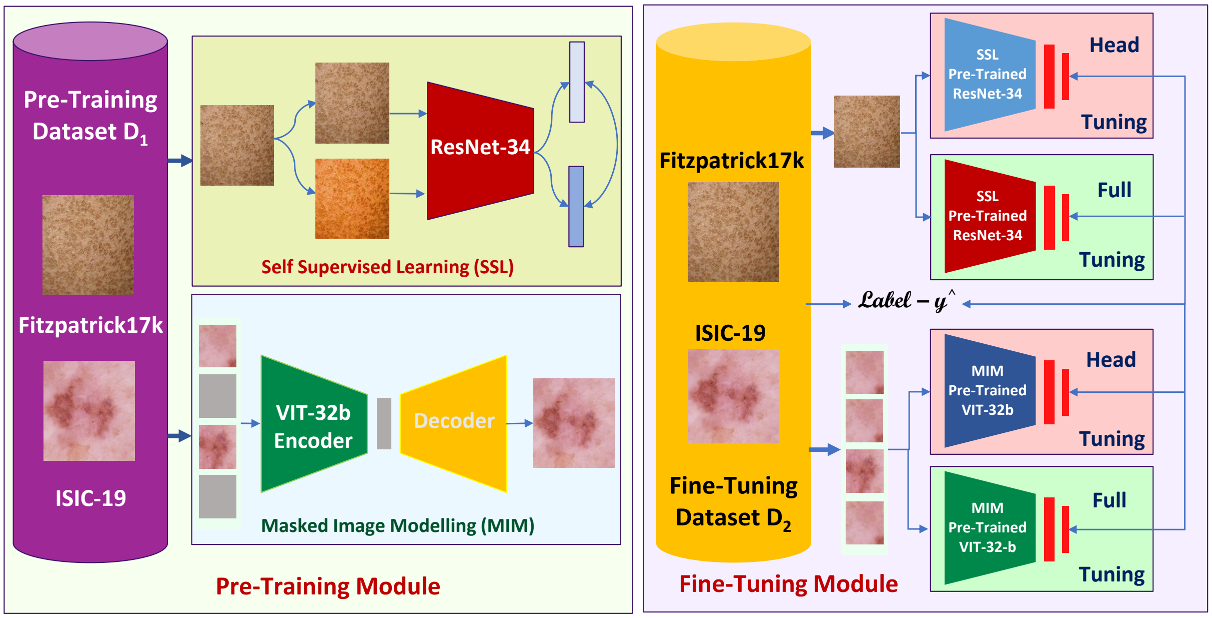

Figure 2. Illustration of Pipelines involved in experimentation of Self Supervised Learning (SSL) and Masked Image Modelling (MIM) Pre-Training followed by either Head Tuning and Full Tuning methodologies of Fine Tuning Module.

in six deaths [7]. Skin cancer is among the six most common cancers. Early detection and treatment of skin lesions can significantly improve patient survival rates. DNNs can extract robust features and make predictions directly from input images of skin lesions. However, their susceptibility to biases can lead to unfair decisions. Studies indicate that patients with darker skin tones experience lower diagnosis accuracy than those with lighter skin [12], limiting their reliable clinical use [15].

Fairness in skin lesion classification presents challenges due to limited annotated data with diverse skin types. Most publicly available datasets primarily represent fair-skinned demographics, leading to data imbalance and potential racial biases in a model's predictions. Despite these concerns, research evaluating Deep Learning (DL) based models on diverse skin types remains scarce, leaving their reliability as disease screening tools untested. Previous studies [21, 24, 25] have shown inherent racial disparities in machine learning algorithms across various domains, including healthcare applications like cardiac MR imaging, chest X-rays, and skin disease imaging.

Pre-Training has become a standard practice in training Deep Learning model pipelines due to its ability to mitigate the data scarcity challenge, especially in medical imagery. Inspired by Human Learning, Pre-Training techniques such as Self Supervised Learning (SSL) are used to learn relevant knowledge from unlabeled data to acquire universal feature representations through a two-step process: Pre-Training to learn these representations and Fine Tuning to adapt to specific tasks. SSL has gained attention in medical imaging as it overcomes the lack of annotated data by learning general-purpose representations without requiring human-annotated labels. A promising technique for training powerful DNN models like the transformer models is Masked Image Modeling (MIM), which involves masking parts of input images randomly and reconstructing them. Recent advancements in MIM-based techniques have surpassed the performance of SSL and supervised models.

In this work, we analyze the impact of SSL and MIM Pre-Training methods on Fairness for Skin Lesion Analysis. The contributions of this work are:

- Our work is one of the first to explore the impact of Pre-Training methods on the fairness of models when trained on datasets that include underrepresented demographic groups.
- We analyze the impact of Pre-Training on the fairness and overall performance of models in Skin Image Analysis by considering different datasets having different representations of demographic groups (skin tone).
- We systematically compare eight Pre-Training based pipelines on six Pre-Training Methods against baseline methods involving two publicly accessible test datasets, comprising in-distribution and out-distribution scenarios of Pre-Training and Fine Tuning on downstream tasks against their Supervised Learn-

ing pipelines with a focus on performance and Fairness metrics.

2. Related Works

Self-Supervised Learning (SSL) involves Pre-Training models on auxiliary pretext tasks before Fine Tuning them on the downstream task. The base model, the encoder, transforms input images into latent representations. ResNet-50 [18] is commonly used as a backbone due to its simplicity and accuracy. Contrastive losses have been a critical advancement in SSL, organizing the feature space by bringing related samples together and pushing unrelated samples apart.

InstDisc [27] significantly contributed to SSL by treating class-level classification as instance-level discrimination. It involved using augmented views of training samples, a modified softmax loss function, and a temperature hyperparameter to handle multiple labels. It utilized a memory bank to store instance parameters and Noise-Contrastive Estimation to approximate the softmax, resulting in more concise representations. SimCLR [10] took an end-to-end approach and incorporated a projection network after the encoder network to reduce dimensions dynamically, similar to InstDisc with respect to the pretext task and loss function but calculates the loss based on the batch samples alone, eliminating the need for a memory bank.

BYOL [14] matches data-augmented views between positive pairs without using negative pairs. It compares the outputs of a fast and a slow network and utilizes the cosine distance between them as a loss. MOCO [17] introduces a representation dictionary whose size is determined by a hyperparameter, similar to InstDisc. Negative examples are sampled from the dictionary, and parameters are updated using a momentum update. VICRegL [4] expands on the VICReg objective to improve performance in image-level and dense prediction tasks. It introduced derived local features by considering feature and spatial location distances.

Recent advancements have demonstrated the effectiveness of **Masked Image Modeling (MIM)** as a Pre-Training strategy for Vision Transformers [13]. MIM involves masking a set of image patches at the input and reconstructing the masked patches at the output, encouraging the network to infer the masked target by leveraging contextual information. Masked Autoencoders (MAE) [16] is a simple approach with an asymmetric encoder-decoder architecture. The encoder receives only visible tokens and a lightweight decoder that reconstructs the masked patches from the encoder's patch-wise output and trainable mask tokens trained with L2 loss. SimMIM [29], another famous MIM architecture, employs a linear layer as a decoder and uses L1 loss instead of L2 loss.

There are two main approaches to **Self-Supervised Learning in Medical Applications**. The first adapts general pretext tasks [9], and the second combines medical knowledge with computer vision expertise [19]. In Skin Image Analysis, [26] used a clustering pretext task similar to SwAV [8]. SimCLR was used for skin-lesion by [3] and MAE for X-ray tasks by [28].

Bias and Fairness in machine learning is a growing concern [22], and approaches to address unfairness in deep learning can be categorized into pre-processing [5], in-processing [2], and post-processing [23]. Pre-processing methods transform data to remove discrimination, achieving a balanced trade-off between accuracy and non-discrimination. In-processing techniques modify model architecture or add Fairness-related penalties to train fairer models. Post-processing methods calibrate predictions using model outputs and sensitive attributes. However, skin type Fairness receives less attention than age, sex, and race Fairness.

3. Experimental Setup

3.1. Dataset

The experiments are performed using Fitzpatrick17k [15] and ISIC 2019 [11] that help us simulate an in and out-distribution scenario in terms of skin types. During the Pre-Training phase, we resize the images to 224x224 and apply the corresponding augmentation techniques specific to each method. In the Fine Tuning stage, we resize, perform random horizontal flipping, and normalize the image.

3.1.1 Fitzpatrick17k Dataset

Fitzpatrick17k Dataset (Fitz-17k) [15] comprises 16577 clinical images labeled with skin conditions and Fitzpatrick skin-type labels. It includes 114 unique skin conditions with corresponding Fitzpatrick skin-type labels. These are further categorized into 9 categories used in this study.

3.1.2 ISIC-2019

ISIC-2019 [11] Dataset contains 25331 samples with eight skin conditions and an unknown class. We use the Fitzpatrick labeling system for the six-point skin tone labeling.

3.1.3 Fitzpatrick labeling system

Fitzpatrick labeling system [15] is a six-point scale initially developed for classifying sun reactiViTy of skin treatment according to skin phenotype. In this, the skin types are categorized into six levels. Although commonly used for categorizing skin types, it has been used recently to evaluate algorithmic Fairness. For the Datasets used, the skin type distribution is shown in Figure 1.

Dataset	Backbone Model	Mean Performance Metrics			Fairness Metrics	
		Accuracy(↑)	ROC-AUC(↑)	F1 Score(↑)	DPR(↑)	DPD(↓)
ISIC-2019 [11]	ViT-32b	0.92214	0.89290	0.42526	0.70585	0.17906
	ViT-16b	**0.92609**	**0.90257**	**0.48205**	0.72679	0.17770
	ResNet-34	0.91586	0.87331	0.45153	**0.77735**	**0.13600**
Fitz-17k [15]	ViT-32b	**0.92039**	0.74292	0.24446	0.08761	**0.06676**
	ViT-16b	0.91292	**0.75895**	**0.27437**	**0.26766**	0.07063
	ResNet-34	0.91601	0.74326	0.24660	0.11710	0.07171

Table 1. Performance & Fairness Metrics of Backbone Models over ISIC-2019 & Fitzpatrick Dataset.

3.2. Pre-Training & Fine Tuning Methodology

We benchmark the tasks on three commonly used Backbones - ResNet-34, ViT-16b, and ViT-32b for supervised learning pipelines. For Pre-Training Methods, we used popular methods MAE and SimMIM with ViT-32b as the backbone for Masked Image Modelling (MIM). For Self-Supervised Learning (SSL), we have used - BYOL, MOCO, SimCLR & VICRegL with ResNet-34 as the backbone.

In order to train our supervised backbones, we utilize the BCE (Binary Cross-Entropy) loss function and the Adam optimizer with a 10^{-4} learning rate.

For the Pre-Training pipelines, we employ the conventional training approach that involves utilizing the loss function and image augmentation techniques associated with each specific Pre-Training method, followed by Fine Tuning similar to supervised backbones.

3.3. Experimental Design

We evaluate ten pipelines, which vary in the model's Pre-Training (PT) and Fine Tuning (FT) alongside the Datasets involved in the pipelines.

The **first two pipelines** involve supervised learning on ResNet34, ViT-16b and ViT-32b on Datasets $D_1 \in$ (ISIC-2019 [11], Fitzpatrick17k [15]), which acts as a baseline for our experiments.

The **other eight pipelines** as shown in Figure 2, involves Pre-Training on Dataset-1 (D_1) where $D_1 \in$ (ISIC-2019 [11], Fitzpatrick17k [15]), followed by Fine Tuning the pre-trained encoder weights, and evaluation on the Dataset-2 (D_2) where $D_2 \in$ (ISIC-2019 [11], Fitzpatrick17k [15]). Fine Tuning can be of two types: Head Tuning and Full Tuning. Head Tuning is where encoder weights remain frozen, and only the final linear layer is trainable. Full Tuning involves Fine Tuning the encoder alongside linear layers.

This results in a total of eight combinations of Pre-Training methods, with each combination representing a pipeline. SSL $\in$ (BYOL, MoCo, SimCLR, VICRegL) and MIM $\in$ (SimMIM, MAE) .

3.4. Evaluation Metrics

We aim for an accurate and fair skin condition classifier by assessing each pipeline's performance using metrics for both performance and Fairness.

3.4.1 Performance Metrics

The Model Performance is reported using well-known metrics such as mean ROC-AUC, Macro F1-score, and mean Accuracy.

3.4.2 Fairness Metrics

For quantification of **Fairness**, we use Disparity metrics [1, 6], namely Demographic Parity Difference (DPD) and Demographic Parity Ratio (DPR).

Demographic Parity Difference (DPD) reports the absolute difference between the highest and lowest group-level selection rates across different groups, with 0 indicating demographic parity as all groups have the same selection rate. Whereas, **Demographic Parity Ratio** (DPR) reports the ratio of the lowest and highest group-level selection rates across different groups, a result that all groups have the same selection rate. The mathematical notation for both of them is as follows :

$$DPD = ((\max_a \mathbb{E}[h(X)|A = a]) - (\min_a \mathbb{E}[h(X)|A = a])) \quad (1)$$

$$DPR = \frac{\max_a \mathbb{E}[h(X)|A = a]}{\min_a \mathbb{E}[h(X)|A = a]} \quad (2)$$

$\forall(a \mathfrak{X} A)$, where for classifier h, X is the denoted feature vector used for predictions, *h(X)* is the predicted value, A is a single sensitive feature and a denotes all distinct values of sensitive feature A.

4. Analysis & Findings

To analyze the change of performance with Self Supervised (SSL) and Masked Image Modelling (MIM) Pre-Training Methods, we experimented with supervised backbones - (ResNet34, ViT-16b, ViT-32b) over both datasets (ISIC-2019 [11], Fitzpatrick17k [15]) as shown in Table 1,

Pre-Training Method	Dataset		Average Change (Δ) from Supervised Backbone			
	Fine Tuning	Pre Training	DPD(↓)	DPR(↑)	Mean(↑) ROC-AUC	Macro(↑) F1 Score
MIM	Fitz-17k [15]	Fitz-17k [15]	0.003±0.034	0.15±0.114	0.046±0.026	0.123±0.046
		ISIC-19 [11]	**0.002±0.014**	**0.166±0.095**	**0.049±0.017**	**0.13±0.033**
	ISIC-19 [11]	Fitz-17k [15]	-0.017±0.015	0.032±0.022	**0.03±0.007**	**0.15±0.021**
		ISIC-19 [11]	**-0.035±0.004**	**0.066±0.016**	0.023±0.004	0.143±0.027
SSL	Fitz-17k [15]	Fitz-17k [15]	**0±0.037**	**0.168±0.083**	**0.089±0.019**	**0.176±0.081**
		ISIC-19 [11]	0.023±0.03	0.125±0.082	0.077±0.03	0.132±0.178
	ISIC-19 [11]	ISIC-19 [11]	0.01±0.027	-0.02±0.05	**0.053±0.016**	**0.132±0.083**
		Fitz-17k [15]	**0.008±0.025**	**-0.015±0.034**	0.041±0.014	0.118±0.074

Table 2. **Average Change** (Δ) in Fairness & Performance Metrics of Pre-Training Methods on Fine Tuning over same and out-of-distribution Dataset on Full & Head Fine Tuning of the model.

Pre-Training Method	Fine Tuning		Mean Performance & Fairness Metrics			
	Dataset	Method	DPR(↑)	DPD(↓)	Mean(↑) ROC-AUC	Macro(↑) F1 Score
MIM	Fitz-17k [15]	Full	**0.252±0.087**	0.087±0.022	**0.795±0.020**	**0.383±0.031**
		Head	0.240±0.121	**0.052±0.008**	0.786±0.023	0.359±0.0439
	ISIC-19 [11]	Full	**0.766±0.016**	**0.147±0.004**	**0.921±0.009**	**0.578±0.022**
		Head	0.744±0.031	0.159±0.018	0.918±0.003	0.565±0.024
SSL	Fitz-17k [15]	Full	**0.297±0.086**	**0.075±0.025**	0.823±0.027	**0.413±0.086**
		Head	0.229±0.067	0.091±0.042	**0.830±0.025**	0.388±0.178
	ISIC-19 [11]	Full	0.741±0.028	0.151±0.0283	0.919±0.016	**0.579±0.080**
		Head	**0.779±0.034**	**0.138±0.021**	**0.921±0.016**	0.574±0.078

Table 3. Performance & Fairness Metrics of Pre-Training Methods on Fine Tuning over same and out-of-distribution Dataset on Full & Head Fine Tuning of model.

which will help us to analyze the change of performance with Self-supervised and masked Image Modelling. Further, we performed the experiments by Pre-Training using various SSL and MIM algorithms followed by Fine Tuning for downstream tasks as described in Section 3.3.

4.1. Impact of Pre-Training & Fine Tuning Datasets on Model's Performance

We investigated the impact of using in-distribution and out-of-distribution datasets for Pre-Training and Fine Tuning on model fairness and performance metrics. In the SSL Pre-Training setting, fairness is maintained in the resulting models when Fine Tuning is performed on the same dataset as pre-trained. However, this setup leads to a more significant improvement in performance than Fine Tuning on different datasets. When examining the MIM approach, we observed that Fine Tuning on the same dataset results in an improvement in fairness metrics, as shown in Table 2. However, if the Pre-Training Dataset differs from the Fine Tuning Dataset, it leads to a more substantial performance boost. This can be attributed to specific Vision Transformer (ViT) architecture characteristics, which perform better when exposed to varied data sources during Pre-Training.

We found that including more underrepresented groups in the Pre-Training dataset improved both the performance and fairness of the model. This is consistent with the findings from models trained on the Fitzpatrick17k dataset, as shown in Table 2. The increase in performance is likely due to the more comprehensive representation of dark skin-toned samples, as shown in Figure 1.

We conclude that Pre-Training and Fine Tuning on the same distribution dataset have a limited impact on boost over fairness and performance metrics in MIM but the effect is much more evident in SSL-based CNN backbones. However, the distribution of demographic groups in Pre-Training plays the most important role in ensuring the fairness and robustness of the model.

4.2. Examining the effect of different Fine Tuning methods on the performance of a pre-trained encoder

We aim to examine how different Fine Tuning methods affect the performance metrics of a pre-trained encoder to gain insights into the Fine Tuning method's influence on the model's performance. While analyzing the results for

Pre-Training Method	Average Change (Δ) from Supervised Backbone				
	DPD(↓)	DPR(↑)	Accuracy(↑)	Mean ROC-AUC(↑)	F1-Score(↑)
SSL	0.01±0.03	0.065±0.105	**0.017±0.01**	**0.065±0.027**	**0.14±0.109**
MIM	**-0.012±0.024**	**0.104±0.089**	0.007±0.004	0.037±0.018	0.136±0.032

Table 4. **Average Change (Δ)** of Performance & Fairness Metrics of Self-Supervised Methods (SSL) & Masked Image Modelling (MIM) Pre-Training Methods over Supervised Backbones

Fine Tuning Dataset	Pre-Training Method	Average Change (Δ) over Supervised Backbone				
		DPD(↓)	DPR(↑)	Mean(↑) Accuracy	Mean(↑) ROC-AUC	Macro(↑) F1 Score
ISIC-2019 [11]	SSL	0.009±0.025	-0.017±0.041	**0.017±0.011**	**0.047±0.015**	0.125±0.076
	MIM	**-0.026±0.014**	**0.049±0.026**	0.009±0.003	0.027±0.006	**0.146±0.023**
Fitz-17k [15]	SSL	0.011±0.035	0.146±0.082	**0.016±0.01**	**0.083±0.025**	**0.154±0.135**
	MIM	**0.003±0.024**	**0.158±0.098**	0.005±0.003	0.047±0.02	0.127±0.037

Table 5. **Average Change (Δ)** in Fairness Metrics (DPR, DPD) and Performance Metrics (Accuracy, ROC-AUC, F1 Score) of Self-Supervised & Masked Image Modelling Pre-Training Methods on Fine Tuning Dataset over Supervised Backbones

the MIM Pre-Training, we found that both Fine Tuning methods, Full Tuning and Head Tuning, lead to slight variations in performance metrics. We also observe that MIM Pre-Training performs better when followed by Full Tuning on the downstream task in Table 3. However, more pronounced differences are observed among the various Fine Tuning methods for the SSL Pre-Training. In the Fitzpatrick17k [15] Dataset, Full Tuning of the model appears to be a superior option with respect to both Fairness and performance. A reverse scenario is observed in the ISIC-2019 [11] Dataset, with Head-Tuning found to be fairer. Overall, we highlight the impact of the Fine Tuning method on the performance of a pre-trained encoder, with the results indicating variations in performance metrics across different Fine Tuning methods, suggesting that the choice of Fine Tuning methods significantly influences the model's performance. Full Tuning tends to provide a slight performance boost compared to Head-Tuning. To mitigate computational costs, employing Head-Tuning alone can be effective with a relatively minor trade-off between performance and Fairness.

4.3. Masked Image Modelling makes backbone fairer than Self-Supervised Learning

We evaluate the impact of Pre-Training approaches on the Fairness of the model backbone. We aim to compare the effects of these Pre-Training methods in ensuring Fairness. The results of our experiments reveal some interesting insights. We find that MIM leads to a more substantial improvement in Fairness metrics than SSL, suggesting that it is more effective in reducing bias in the model backbone as in Table 4 with a significant boost in DPR and a decrease in DPD. Furthermore, when considering performance metrics, we observe that both Pre-Training methods show improvements, with SSL having a better boost in performance. However, this can be due to fact-learning features related to skin-tone bias, resulting in a lower Fairness boost.

4.4. Self-Supervised Learning is much better in boosting the performance of backbone than Masked Image Modelling

We investigated the effects of Pre-Training methods on model performance. Our experiments indicate that both the SSL and MIM-based Pre-Training methods provide a boost over the supervised learning backbone in terms of performance. For Fairness, both Pre-Training methods show positive effects in DPR, indicating reduced bias. However, a small increase in DPD is observed, which isn't ideal. Based on assessing metrics, we infer that the SSL-based Pre-Training consistently demonstrates superior performance enhancement compared to MIM based on the evaluation done over downstream tasks on both Datasets, as shown in Table 5.

5. Conclusion

This study investigates the impact on the fairness of popular Pre-Training methods, such as Masked Image Modeling (MAE, SimMIM) and Self-Supervised Learning (BYOL, MoCo, SimCLR, VICRegL), when used on skin lesion classification datasets with underrepresented demographic groups. The study compares the performance of pre-trained models to supervised learning backbones on two skin lesion datasets (ISIC-2019 and Fitzpatrick17k) with different skin tone distributions.

We found that Pre-Training can improve the performance of these models' performance but also introduce fairness concerns. This is because pre-trained models are often trained on datasets that are not representative of the diversity of skin tones in the real world. As a result, these models can be more likely to make errors for patients with darker skin tones.

Our work is one of the first to examine the impact of Pre-Training on fairness in skin image analysis. We found that Pre-Training can lead to a trade-off between performance and fairness. This means that models that are pre-trained on large datasets may perform better, but they may also be more likely to make errors for patients with darker skin tones. Our results underscore the importance of considering fairness when using pre-trained models for skin image analysis.

6. Acknowledgments

The completion of this project was made possible through the assistance provided by the Student Seed Grant program (ID: 00000736) from Manipal Academy of Higher Education, Manipal, India. We express our gratitude to the Department of Data Science & Computer Applications at Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India, for their support and provision of essential computational resources, enabling the successful completion of numerous valuable experiments. Additionally, our appreciation extends to Mr. Aleti Vardhan, Mr. Rakshith Sathish, and Mr. Aditya Kasliwal for their valuable assistance in reviewing the manuscript.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 10–15 Jul 2018. 4

[2] Mohsan S. Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. *ArXiv*, abs/1809.02169, 2018. 3

[3] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zach Beaver, Jana von Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. Big self-supervised models advance medical image classification. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3458–3468, 2021. 3

[4] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features. *ArXiv*, abs/2210.01571, 2022. 3

[5] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Kr. Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *ArXiv*, abs/1810.01943, 2018. 3

[6] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, May 2020. 4

[7] Wei Cao, Hongda Chen, Yiwen Yu, Ni Li, and Wanqing Chen. Changing profiles of cancer burden worldwide and in china: a secondary analysis of the global cancer statistics 2020. *Chinese Medical Journal*, 134:783 – 791, 2021. 2

[8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *ArXiv*, abs/2006.09882, 2020. 3

[9] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, 58:101539, 2019. 3

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020. 3

[11] Marc Combalia, Noel C. F. Codella, Veronica M Rotemberg, Brian Helba, Verónica Vilaplana, Ofer Reiter, Allan C. Halpern, Susana Puig, and Josep Malvehy. Bcn20000: Dermoscopic lesions in the wild. *ArXiv*, abs/1908.02288, 2019. 1, 3, 4, 5, 6

[12] Roxana Daneshjou, Kailas Vodrahalli, Weixin Liang, Roberto A. Novoa, Melissa Jenkins, Veronica M Rotemberg, Justin M. Ko, Susan M. Swetter, Elizabeth E. Bailey, Olivier Gevaert, Pritam Mukherjee, Michelle Phung, Kiana Yekrang, Bradley Fong, Rachna Sahasrabudhe, James Zou, and Albert S. Chiou. Disparities in dermatology ai: Assessments using diverse clinical images. *ArXiv*, abs/2111.08006, 2021. 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 3

[14] Jean-Bastien Grill, Florian Strub, Florent Altch'e, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *ArXiv*, abs/2006.07733, 2020. 3

[15] Matthew Groh, Caleb Harris, Luis R. Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1820–1828, 2021. 2, 3, 4, 5, 6

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2021. 3

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2019. 3

[18] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 3

[19] Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Self-supervised learning for spinal mris. In *DLMIA/ML-CDS@MICCAI*, 2017. 3

[20] Mohd. Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, and Shanay Rab. Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, 2022. 1

[21] María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante. Addressing fairness in artificial intelligence for medical imaging. *Nature Communications*, 13, 2022. 2

[22] Ninareh Mehrabi, Fred Morstatter, Nripsuta Ani Saxena, Kristina Lerman, and A. G. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54:1 – 35, 2019. 3

[23] Felix Petersen, Debarghya Mukherjee, Yuekai Sun, and Mikhail Yurochkin. Post-processing for individual fairness. In *Neural Information Processing Systems*, 2021. 3

[24] Esther Puyol-Antón, Bram Ruijsink, Stefan K. Piechnik, Stefan Neubauer, Steffen Erhard Petersen, Reza Razavi, and Andrew P. King. Fairness in cardiac mr image analysis: An investigation of bias due to data imbalance in deep learning based segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021. 2

[25] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27:2176 – 2182, 2021. 2

[26] Dan Wang, Na Pang, Yanying Wang, and Hongwei Zhao. Unlabeled skin lesion classification by self-supervised topology clustering network. *Biomed. Signal Process. Control.*, 66:102428, 2021. 3

[27] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. *ArXiv*, abs/1805.01978, 2018. 3

[28] Junfei Xiao, Yutong Bai, Alan Loddon Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3577–3589, 2022. 3

[29] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: a simple framework for masked image modeling. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9643–9653, 2021. 3