6.S952 Results Document

Anthony Khaiat and Jad Makki

1 Task

1.1 Problem: Skin conditions, such as acne, impact a vast number of individuals both through creating insecurities regarding appearance and through their actual physical dangers. As such, this project aims to develop a diagnostic tool for detecting skin conditions which is fair across different skin complexions.

1.2 Data: For this task, we aggregated two publicly available datasets from Kaggle, skin disease classification, and Face Skin Diseases. Through compiling these datasets, we end up with 715 images and cover the following skin conditions: Acne, Rosacea, Redness, Eye Bags, Actinic Keratosis, Basal Cell Carcinoma, and Eczema. In addition, we manually labelled individuals' skin complexion as either "lighter" or "darker" to obtain values for our sensitive attribute. It is also important to acknowledge the severe class imbalance in skin complexions, as over 90% of individuals had a lighter skin complexion. Furthermore, for the Redness and Eye Bag conditions, we had roughly 20 examples of each, making them the minority classes in our classification task.

1.3 Inputs/outputs: We learn two computer vision models from these images to classify skin condition of the depicted individuals, with the ultimate goal being both an accurate classifier and one that is fair across lighter and darker skin complexions.

1.4 Methods: We train a baseline Convolutional Neural Network with 3 million parameters and fine-tune a pre-trained Vision Transformer (ViT) model for this task. We perform weight-sampling to achieve fairness.

2 Evaluation Strategy

We evaluate the performance of each of our models in the following ways:

2.1 Overall Accuracy: We will evaluate the model's ability to accurately classify the skin conditions of individuals on a holdout test set - we utilize the same training and test set for training both of our models based on a stratified random sample of the sensitive attribute and overall class labels.

2.2 Fairness We will analyze the performance of the models in depth by observing the accuracy for individuals with lighter skin complexions and darker skin complexions separately. We also introduce a weight sampling adjustment motivated by our class imbalance which penalizes the models more for incorrect classification of the underrepresented group - darker skin - and analyze how this affects model performance.

2.3 Saliency Maps: To aid in our understanding of the performance of our models, we visually inspect saliency maps to observe the decision-making process.

3 Results

To begin, we will focus the discussion our CNN model. This model achieved an accuracy of 34.1% on the held-out test set. Evaluating the model strictly on lighter and darker complexions of skin yielded accuracies of 34% and 36.4% respectively. These results contend initially that the model is fair across the sensitive attribute. On the other hand, after fine-tuning the ViT model to our dataset, the model achieved an overall accuracy of 43.1%, with accuracies on lighter and darker complexions of skin at 46.2% and 9% respectively, revealing a large discrepancy in model performance based on the sensitive attribute. However, introducing the weight adjustment enhances performance both overall and within each subgroup for both models, these results, alongside the baseline results are summarized in table 1.

Metric	CNN	ViT	Adjusted CNN	Adjusted ViT
Training Accuracy	47%	45.61%	87%	88%
Test Accuracy	34.1%	43.1%	47.55%	46%
Lighter Skin	34%	46.2%	47.8%	46.2%
Darker Skin	36%	9%	45.5%	27.3%

Table 1: Model Performance

Furthermore, through plotting saliency maps, we discovered that the non-adjusted fine-tuned ViT model tended to discover more relevant features to the disease for lighter skin complexions than for darker ones. Examples of these saliency maps side-by-side for the two complexions can be found in the **Appendix** section. Interestingly, adding the weight adjustment for the ViT model also improved the model's ability to learn the characteristics of the skin conditions more prominently, these maps will be discussed more in depth in the **Analysis** section.

4 Analysis

4.1 Standard CNN: Since the CNN model was built from scratch, it does not have any pre-trained inductive biases. As such, we hypothesized that either the model will become biased towards diagnosing conditions for lighter skin conditions due to the data imbalance - which would optimize total accuracy - or the model will perform equally poor regardless of the sensitive attribute and satisfy error parity. The results indicate the latter hypothesis to be correct, as the model struggles regardless of the sensitive attribute, as indicated by the low accuracies. In addition, the model struggles to predict the redness and eye bags conditions, likely due to the lack of data for these diseases. Introducing weight sampling to this model singificantly improved the overall performance, though it did lead to increased overfitting. The weight sampling likely sparked this improvement because it forced the model to really learn the characteristics of the skin diseases particularly for the darker skin group, which also improves performance on the lighter group because the underlying characteristics of the diseases are the same regardless of the skin complexion.

4.2 Transfer Learning with ViT: Looking at the results of the Non-Adjusted ViT model, it is clear that the inductive biases learned from the model's pre-training lead to improved overall performance at the cost of fairness. Since this is a seven-class classification problem, a baseline model accuracy would be roughly 14.3% or $\frac{1}{7}$, and this model only exhibits 9% accuracy on those with darker skin complexions. While the weight adjustment did help bridge the gap between the performance across subgroups, the model is still not truly "fair". In contrast to the CNN, the weight adjustment purely improved the model's performance across the darker skin group, which indicates that the model was beginning to learn the underlying characteristics of the diseases, but perhaps was still separating the individuals by skin complexion. In addition the ViT model did not beat the CNN model in terms of performance on the underrepresented group, which speaks to the difficulty of overcoming the inductive biases of pre-trained models.

4.3 Saliency Maps: We also obtain useful information through analyzing the saliency maps in the **Appendix**. Looking at Figure 1, which details two individuals with different skin complexions both suffering from Basil Cell Carcinoma, we observe the non-adjusted ViT model being more confident in finding the disease for the individual with lighter skin. We also see that the saliency map is overlayed both on the individuals skin and on the infected area itself for both models, which shows that the model was making diagnoses through a combination of skin complexion and disease characteristics. Another interesting thing to consider is that for lighter skinned individuals, the infected area tends to be more red, which seems to not be the case as skin complexion gets darker, this could also be affecting the model's performance on diagnosing conditions across the subgroups. In Figures 2 and 3, we can see comparisons of Saliency maps for individuals with Acne and Rosacea across different skin complexions. In these cases, we also have different camera angles, which could also affect model performance. Lastly, if we compare Figure 2 to Figure 4, we can observe how the weight sampling affects the ViT's classification criterion. In Figure 4, the model for both the light and dark skin complexion is able identify more accurately the actual locations of the individuals' acne, which leads to more confident and correct predictions, whereas in Figure 2, both maps have activated gradients in irrelevant areas, such as around the collar of the darker skin individual and on the top of the image for the lighter skin individual. Ultimately, introducing weight sampling helped the model to identify the characteristics of the disease for both subgroups, which leads to improved performance as indicated by the saliency maps and accuracies.

5 Team Contributions

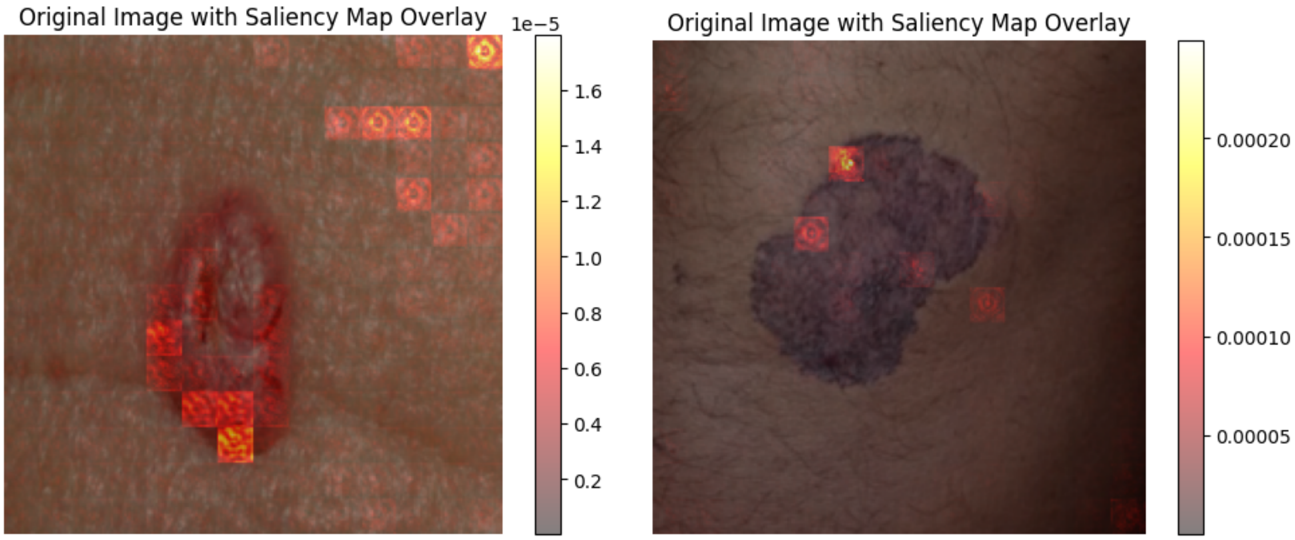
Both members of this partnership were involved in aggregating and preprocessing the datasets, they also worked together to create the slides and each deliverable. Anthony was primarily responsible for fine-tuning the ViT model and producing the saliency maps. Jad focused more on building the CNN and implementing the weight sampling scheme across the models.

6 References

1. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118. https://doi.org/10.1038/nature210

2. Buolamwini, J. Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research, 77-91 Accessed May 04, 2024.

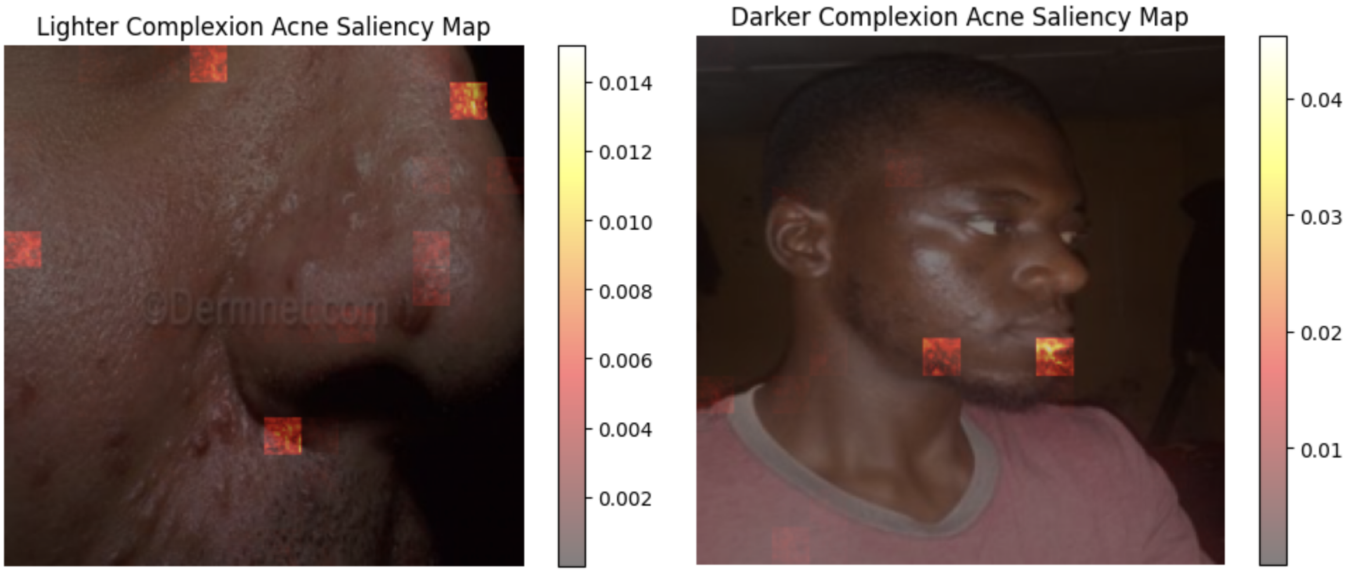
7 Appendix


Original Image with Saliency Map Overlay
1e−5
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
Original Image with Saliency Map Overlay
0.00020
0.00015
0.00010
0.00005

(a) Lighter Skin Complexion

(b) Darker Skin Complexion

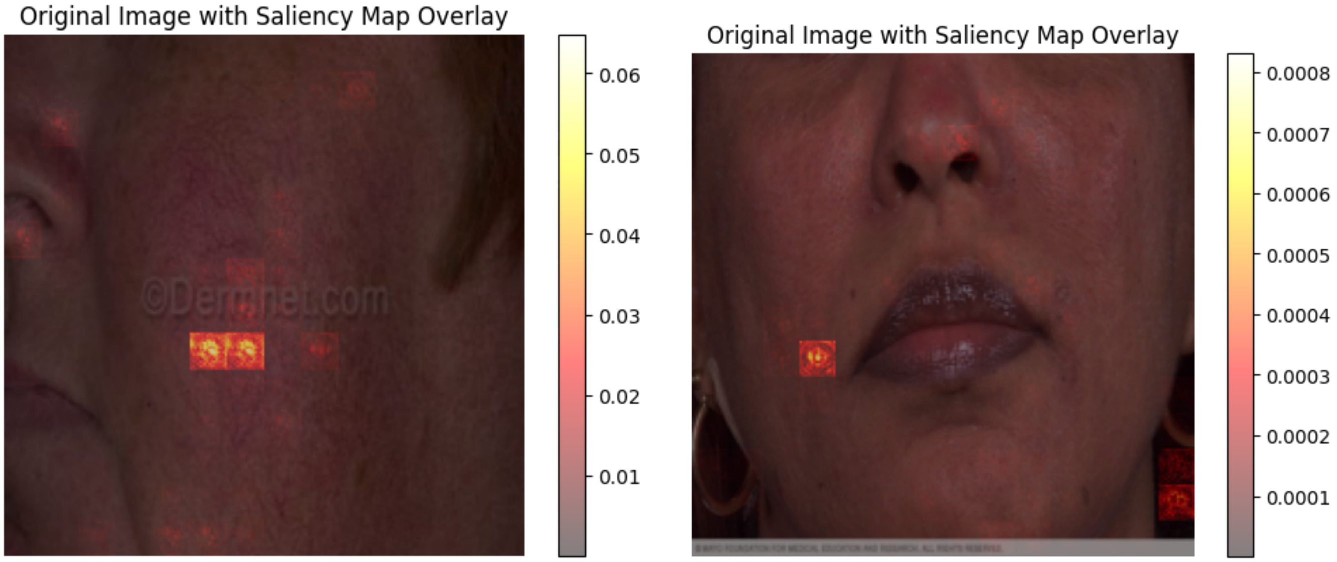
Figure 1: Non-Adjusted ViT Model Basil Cell Carcinoma Saliency Maps


Lighter Complexion Acne Saliency Map
©Dermnet.com
0.014
0.012
0.010
0.008
0.006
0.004
0.002
Darker Complexion Acne Saliency Map
0.04
0.03
0.02
0.01

(a) Lighter Skin Complexion

(b) Darker Skin Complexion

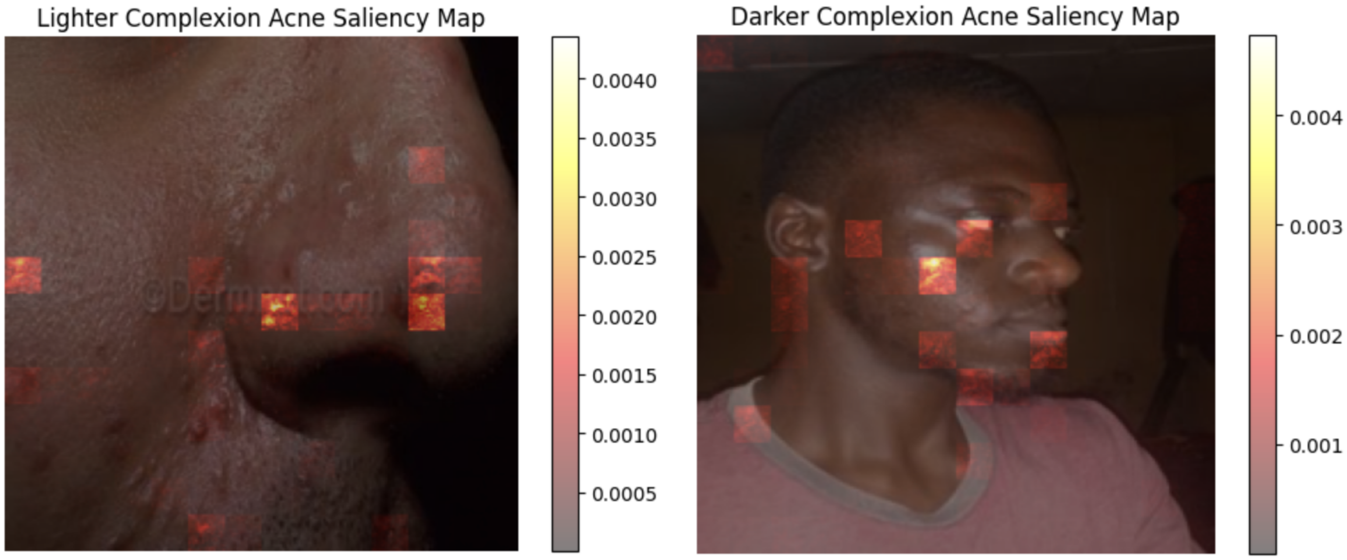
Figure 2: Non-Adjusted ViT Model Acne Saliency Maps



(a) Lighter Skin Complexion

(b) Darker Skin Complexion

Figure 3: Non-Adjusted ViT Model Rosacea Saliency Maps


Lighter Complexion Acne Saliency Map
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
Darker Complexion Acne Saliency Map
0.004
0.003
0.002
0.001

(a) Lighter Skin Complexion

(b) Darker Skin Complexion

Figure 4: Adjusted ViT Model Acne Saliency Maps