Enhancing Skin Disease Detection Accuracy and Fairness

Sophia Li, Qianyi Li, Xinran Zhang
Department of Computer Science, University of Toronto

Motivation

Challenge: Suspicious skin conditions can escalate into serious cancers. Catching them early significantly boosts survival chances. But how do we speed up and improve diagnosis when there's a shortage of dermatologists?

Innovative Solution: Hospitals are now using Convolutional Neural Networks (CNNs) for faster and more accurate skin condition diagnosis. This tech advancement is a game-changer, aiding experts in delivering timely care.

But There's a Catch: The Bias Problem

The Bias Issue: While beneficial, CNN models predominantly rely on datasets that represent lighter skin tones. This imbalance leads to less effective diagnoses for those with darker skin, creating a gap in care quality.

Our Goal: We're determined to bridge this gap. Our research focuses on identify and reduce imbalance in the dataset. Emphasizing fairness, we aim to reduce biases and ensure accurate diagnoses for all skin tones.

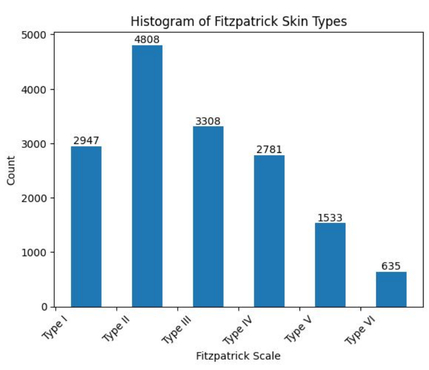

Fig.1 Distribution of Fitzpatrick Scales

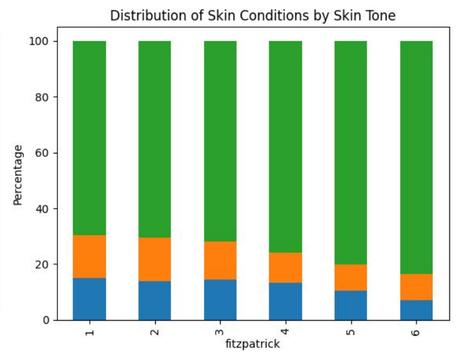

Fig.2 Distribution of Skin Condition by Skin Tone

Related Work

Shankar et al.'s research [2] highlighted a critical issue in image classification tasks: a notable drop in performance for underrepresented subgroups due to a lack of diversity in datasets. Addressing this challenge, various studies have turned to Generative Adversarial Networks (GANs) [3], leveraging their sophisticated capabilities in synthetic image generation and noise detection. Furthermore, augmenting datasets with GAN-generated images [4] has been shown to enhance classifier performance, primarily by expanding the training data size, thereby providing a more diverse and representative dataset.

In our project, we integrated the Deep Convolutional Generative Adversarial Networks (DCGANs) framework, as proposed in Radford, Metz, and Chintala's work [1]. This integration was pivotal for extracting features from a wide range of dermatological images. However, we encountered a notable limitation of the DCGANs framework: its optimal performance is seen with uniform image inputs. Our dataset, comprising diverse skin disease images from various body regions, presented a significant challenge. This heterogeneity in the dataset led to inconsistencies in the quality of synthetic images generated by DCGANs, highlighting a critical area for improvement in applying DCGANs to medical datasets that feature a high degree of input variability.

References

[1] Radford, A., Metz, L., & Chintala, S. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", 2016
[2] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world, 2017.
[3] Dipali Pattanayak and Kuntal Patel. Generative adversarial networks: Solution for handling imbalanced datasets in computer vision. In 2022 International Conference for Advancement in Technology (ICONAT), pages 1–6, 2022.
[4] Hubert Cecotti and Ganesh Jha. Training dataset extension through multiclass generative adversarial networks and k-nearest neighbor classifier. In K. C. Santosh and Ravindra S. Hegadi, editors, Recent Trends in Image Processing and Pattern Recognition, pages 596–610, Singapore, 2019. Springer Singapore.

Methods

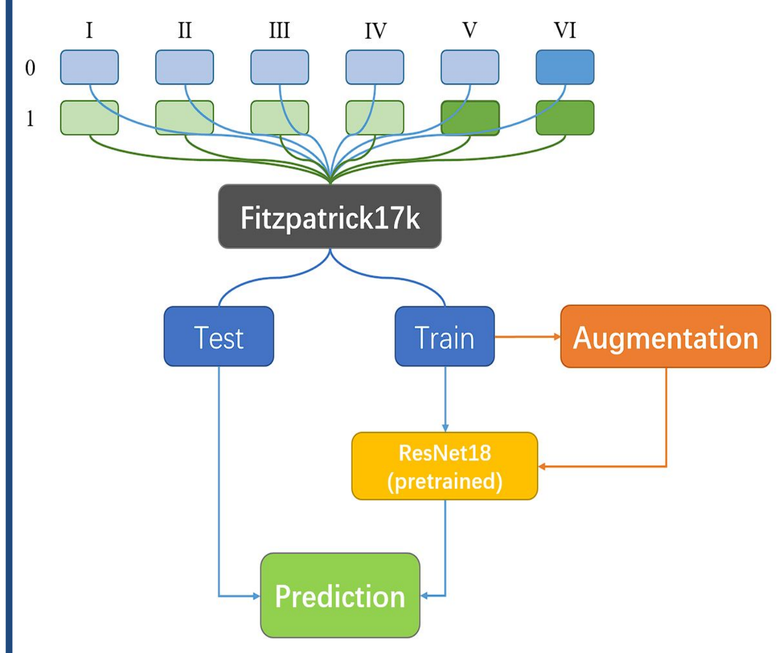

Fig.3 Flow Chart

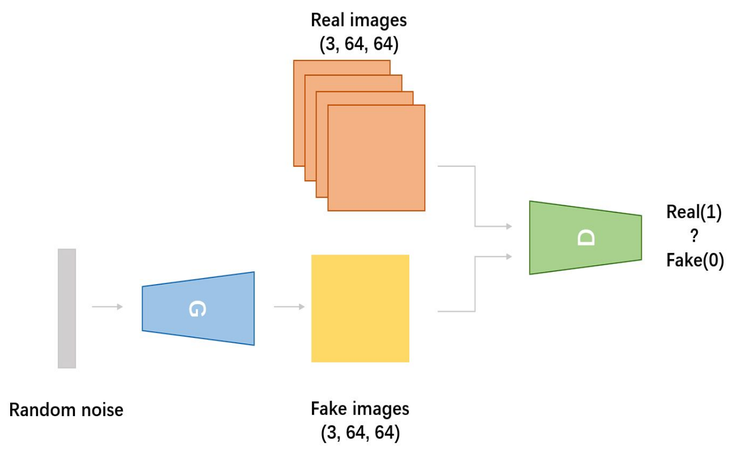

Fig.4 GAN architecture

Fitzpatrick17k

The dataset was divided into training and testing sets in an 80:20 ratio. Five models were fine-tuned using a pretrained ResNet18:

Model 1 Trained on the entire training.
Model 2 Balanced in Fitz scale by undersampling Fitz I-V subgroups
Model 3 Balanced in disease label by undersampling the non-neoplastic subgroup
Model 4 Balanced in both Fitz scale and label
Model 5 Balanced in both by undersampling the large subgroups and augmenting the smaller ones

HAM10000

same as Fitzpatrick17k
Model 1 Entire Balanced Training set
Model 2 Imbalance dataset constructed by random removing 50% male data points from the training
Model 3 DCGAN-generated fake images to re-balance the dataset used in Model 2

Experimental Results

	Fitzpatrick I	Fitzpatrick II	Fitzpatrick III	Fitzpatrick IV	Fitzpatrick V	Fitzpatrick VI
model1: baseline	0.7066	0.7141	0.7224	0.7629	0.8053	0.8067
Sampling						
model2: Balance in scale	0.6820	0.6885	0.7066	0.7594	0.8020	0.8151
model3: Balance in label	0.7459	0.7357	0.7476	0.7825	0.8119	0.8319
model4: Balance in both	0.7082	0.6988	0.7224	0.7415	0.7789	0.7893
Data Augmentation						
model5: simple augmentation	0.7672	0.7408	0.7697	0.7736	0.8284	0.8152

Table 1: Accuracy Across Different Fitzpatrick Scales for Various Models

Simple Augmentation

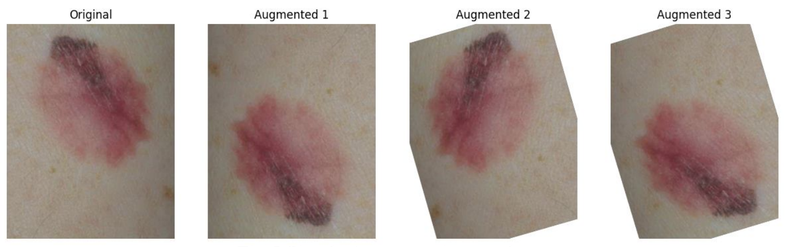

Fig.5 Simple Augmentation results

DCGAN DEMO

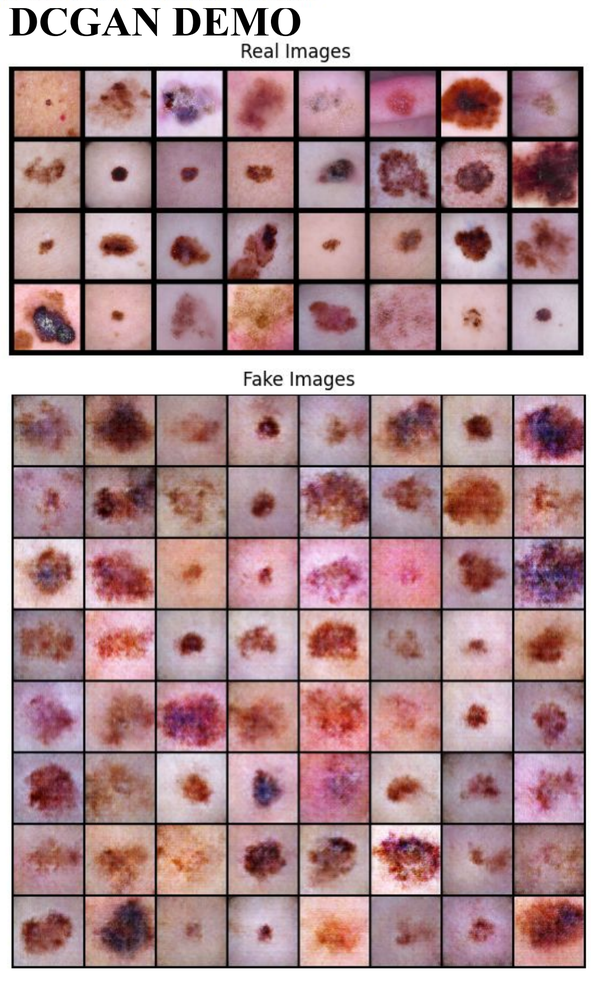

Fig 5. Augmentation on HAM10000

Fitzpatrick17k:

Model 5 demonstrates a **modest enhancement** in overall performance compared to the baseline model, particularly in achieving a **more balanced** classification accuracies across the six Fitz scales.
Takeaway: simple augmentation contributes to mitigating the imbalances present in the original dataset.

However, it shows lower accuracy for Fitz VI than Model 3
Takeaway: simple augmentation techniques are less effective for smaller sub-groups like Fitz VI.

HAM10000 Demo:

Model 3 **surpasses** Model 2 in performance and **gender-balanced accuracy**, yet falls behind Model 1.
Takeaway: GAN-generated images enhance accuracy in imbalanced datasets.

	model1	model2	model3
female	0.7029	0.6514	0.6971
male	0.7105	0.5451	0.6616

Table 2: Augmentation on HAM10000