Improving the Fairness of Artificially Intelligent Skin Disease Detectors Using Stable Diffusion

ROBO069

*All images & graphs were created by the student unless otherwise noted

Research Problem
AI misdiagnoses minorities

People of color underserved in healthcare → Imbalanced datasets w/ POC underrepresented → Diagnostic AIs trained on fewer POC images → Diagnostic AIs disproportionately misdiagnose POC

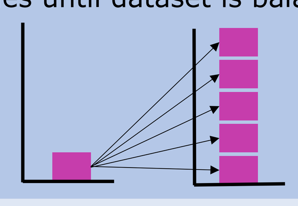
Objective: improve the fairness of a skin disease detector AI for POC

Background
Techniques to increase fairness

Flower image from similarpng.com/purple-flower-isolated-on-transparent-background-png/

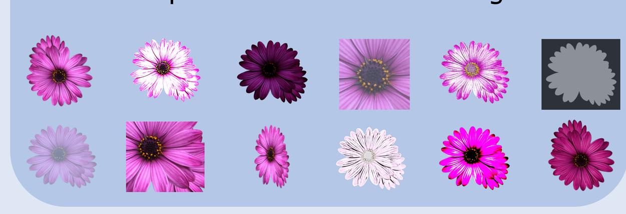
Loss Function
Optimizes fairness & accuracy of performance during training
- No new images added
- Fairness-accuracy trade-off [9]

Random Oversampling
Duplicate randomly selected minority images until dataset is balanced



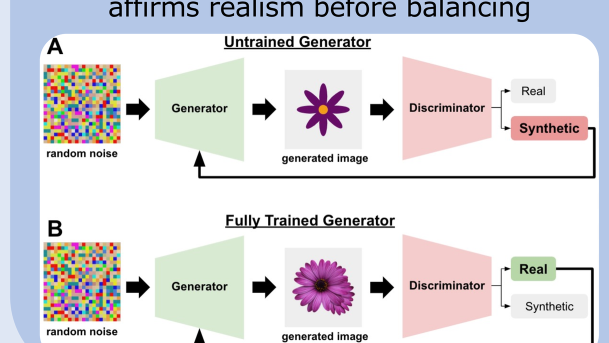
- Increases accuracy [11]
- Decreases **diversity** [13, 14]
- Statistically insignificant increase in fairness [37]

Augmentative Oversampling
Alter (crop/color shift/turn) minority duplicates before balancing



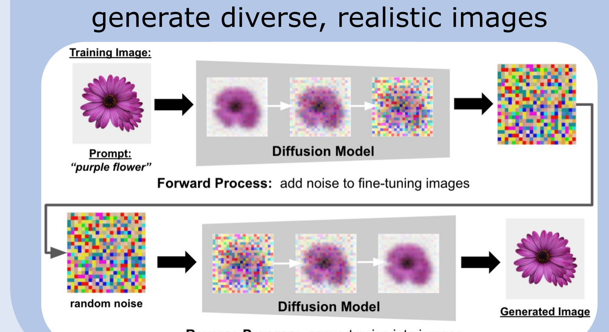
- Increases **diversity** [15]
- Increases accuracy [17]
- **Diversity** still limited
- Statistically insignificant increase in fairness [37]

Generative Adversarial Networks
Generator generates images, discriminator affirms realism before balancing



- Increases **diversity**
- Significant increase in accuracy & fairness (retinal dataset) [21]
- Generator & discriminator must balance/converge → unstable
- Discriminator only recognizes small range of images as realistic (mode collapse) [23]
- **Realism-diversity trade-off**

Stable Diffusion Model
Use controlled statistical variation to generate diverse, realistic images



- Generates **diverse** & **realistic** skin disease images [26]
- Increased accuracy [27]
- Not yet used to balance medical datasets by race **(This work)**

Statistically significant increases in fairness have not yet been achieved for a skin disease dataset

Research Questions

1. Will using stable diffusion-generated images of POC to balance a skin disease detector's training dataset significantly improve its fairness?
2. Will the detector's diagnostic performance be sustained?

Methodology

Dataset

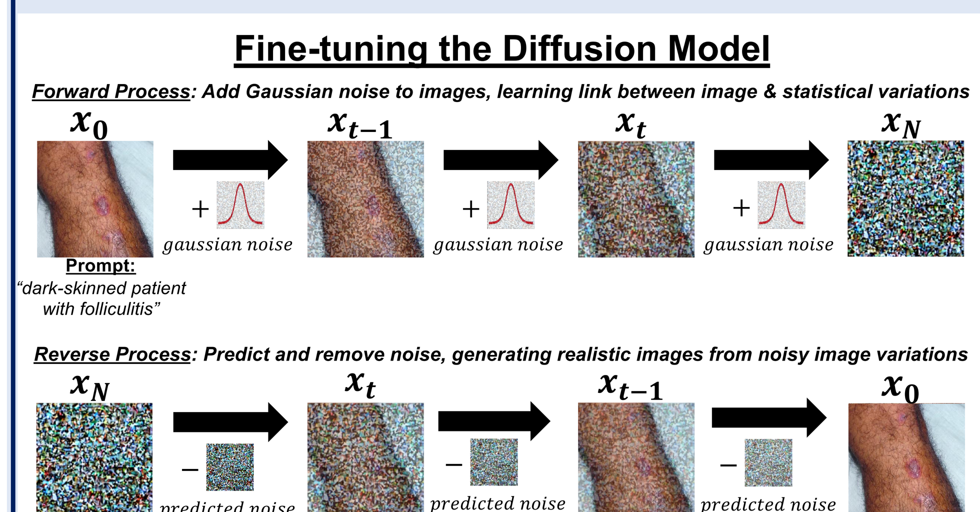
Disease	Lightest Skin Tone	Darkest Skin Tone
Folliculitis	145	42
Lichen planus	160	117
Squamous cell carcinoma	298	51

- Images from Fitzpatrick17K dataset
- 3 diseases, 2 skin tones
- Lightest skin tone: 74.2%, Darkest skin tone: 25.8%
- 210 testing images (35 from each skin tone-disease combination)
- 603 training images (remaining images)

Diffusion model-based augmented oversampling

Fine tuning the diffusion model

Diffusion model: Stability AI's Stable Diffusion 2.0
- Pre-trained on LAION-5B
  - Few skin disease images
- Fine-tune on 5 images + descriptive prompts from each skin tone-disease combination w/ DreamBooth → generate realistic skin disease images

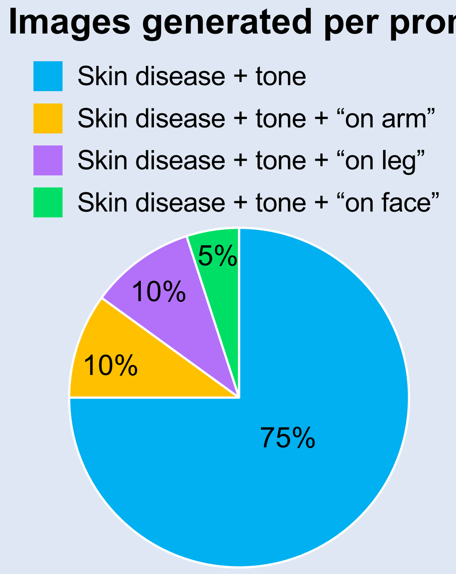


Normal curve image from alamy.com/stock-photo/gaussian-bell-normal-distribution-curve.html?cutout=1&sortBy=relevant
Gaussian noise image from onceuponanalgorithm.org/guide-what-is-a-stable-diffusion-seed-and-how-to-use-it/

Prompt engineering

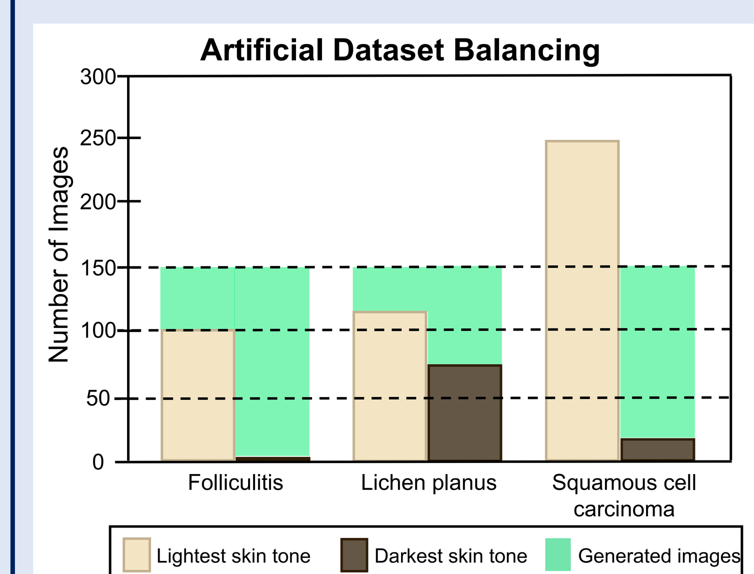
Input prompts into fine-tuned model → model creates images as prompts describe via reverse process
- Vary disease location in prompts → generate diverse images
- Prompt distribution described below

% Images generated per prompt



Oversampling

Each skin tone/disease combination should have at least N images
- # images to generate = N - # dataset images
- Sweep 4 values for threshold N: 0 (control), 50, 100, 150

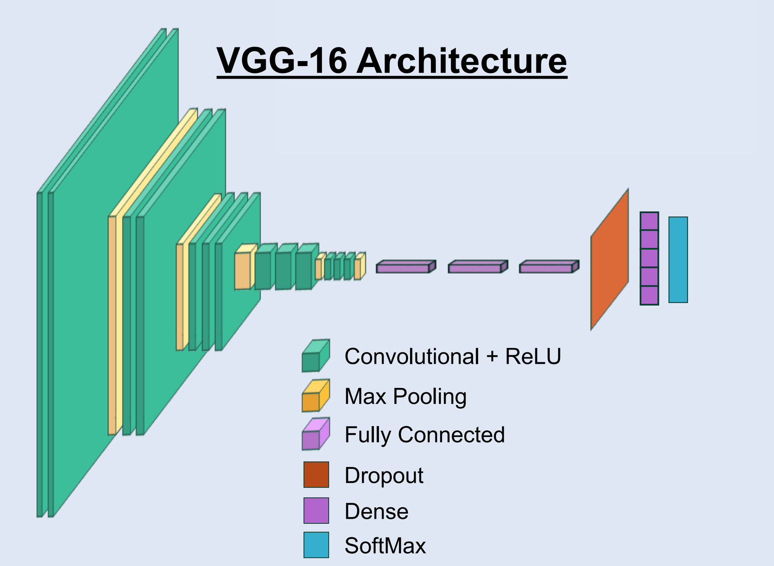


Add generated images to dataset until each skin tone/disease combination has N images → balanced dataset

Evaluating fairness using diagnostic AIs

Diagnostic AI architecture

Modified VGG-16 neural network pretrained on ImageNet
- As per Fitzpatrick17K creators
- Train 17 networks on each of the 4 balanced datasets (4 values of threshold N)
  - Trained using Keras & TensorFlow packages on Google Colaboratory platform



Fairness evaluation

Minimize difference between performance on dark & light skin tones → fair AI

Statistical Parity Difference (SPD) $= |PR_{dark} - PR_{light}|$
- PR – positive diagnosis rate
- Doesn't consider accuracy of diagnosis

Equal Opportunity Difference (EOD) $= |TPR_{dark} - TPR_{light}|$
- TPR – true/correct positive diagnosis rate
- Doesn't consider inaccurate diagnoses

Average Odds Difference (AOD) $= \frac{|TPR_{dark} - TPR_{light}| + |FPR_{dark} - FPR_{light}|}{2}$
- FPR – false/incorrect positive diagnosis rate
- Considers both correct & incorrect diagnosis

Overall performance should be sustained
- Measured using **Area Under Receiver Operating Curve** (AUROC)

Results

Image generation with stable diffusion

- Fine-tuning & generated images for folliculitis on dark skin shown on right
- Generated images diverse due to prompt engineering
- Generated images realistic due to diffusion model & DreamBooth efficacy

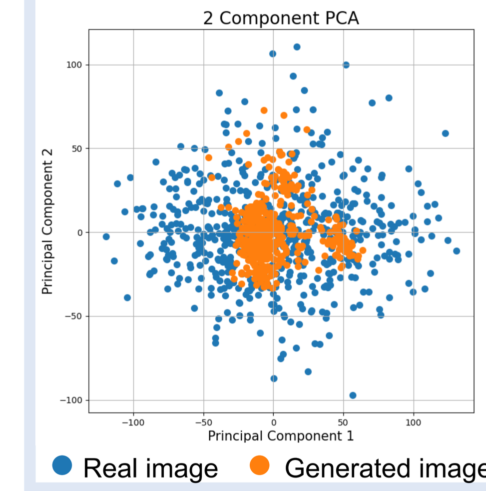
Fine-tuning Images for "folliculitis on a dark skin tone"

Images from [7]

Sample images generated of "folliculitis on a dark skin tone"

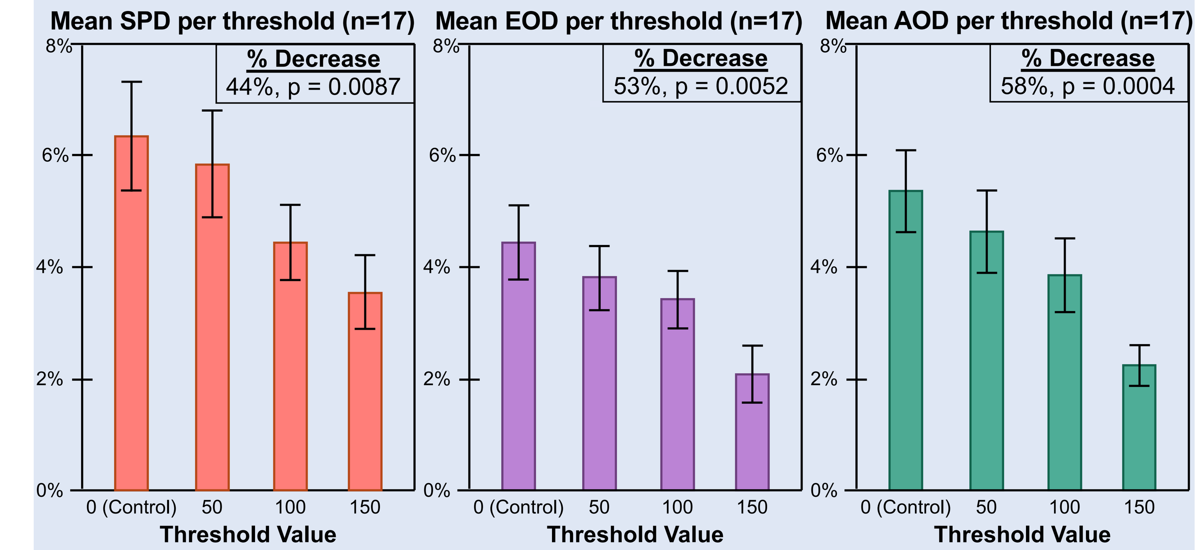
Results cont.

Evaluating Synthetic Images



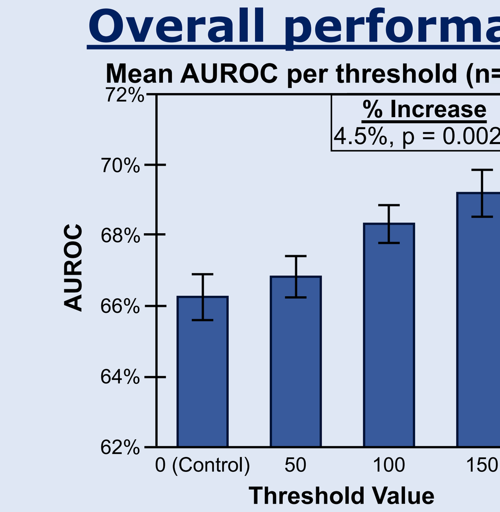
- Principal Component Analysis on real & generated images
  - Projects images onto dataset's 2 largest eigenvectors to represent in 2D space
- Generated images overlap real
  - Indicates realism

Fairness of diagnostic AIs



- Threshold inc. → SPD, AOD, EOD dec. → **Fairness increases**
- Statistically significant
- Due to diversity of images generated

Overall performance of diagnostic AIs



- AUROC scores increase as image threshold increases → **overall performance improves**
- Due to realism of generated images

Comparison to other works (Fitzpatrick17K)

Approach	Technique	Fairness		Performance	Analysis of Results
		% dec EOD	% dec AOD	% inc Accuracy	
Cfair [34]	Loss function	5.6% 5.70% to 5.38%	3.5% 10.5% to 10.2%	0.43% 87.5% to **87.9%**	No new images → fairness-accuracy trade-off
FairAdaBN [9]		29% 5.70% to 4.08%	27% 10.5% to 7.7%	-3.21% 87.5% to 84.7%	
GroupDRO [35]		21.2% 5.70% to 4.49%	22% 10.5% to 8.23%	-1.04% 87.5% to 86.6%	
EnD [36]		10.3% 5.70% to 5.12%	13% 10.5% to 9.20%	-0.83% 87.5% to 86.8%	
Resampling [37]	Oversampling	-1.23% 5.70% to 5.77%	-2.3% 10.5% to 10.8%	0.23% 87.5% to 84.7%	Decreases diversity → decreased fairness
Zhang et al. [22]	GAN	generated unrealistic images (Image from [22])			Small, diverse dataset → overfit discriminator → mode collapse
This work	Stable diffusion	**53%** 4.43% to **2.07%**	**58%** 5.38% to **2.24%**	0.76% 73.8% to 74.3%	Highest fairness, no accuracy trade-off

Conclusion

- Balanced training dataset with new, realistic, diverse diffusion-generated images → increased AI fairness & performance

Contributions
- First to apply stable diffusion to address race imbalance in datasets to the best of our knowledge
- First to achieve a statistically significant increase in the fairness of a skin disease detector to the best of our knowledge
  - Largest increases in fairness & accuracy among Fitzpatrick17K works

Limitations
- Diffusion models require lots of memory to train

Future Work
- Implement algorithm to confirm realism of generated images
- Test on 6 skin tones for more realistic diversity representation
- Test on other applications with underrepresented minorities (e.g. facial recognition, criminal justice)