Science Advances
AAAS

Supplementary Materials for

Disparities in dermatology AI performance on a diverse, curated clinical image set

Roxana Daneshjou *et al.*

Corresponding author: James Zou, jamesz@stanford.edu; Albert S. Chiou, achiou@stanford.edu

Sci. Adv. **8**, eabq6147 (2022)
DOI: 10.1126/sciadv.abq6147

This PDF file includes:

Supplemental Tables and Figures

Supplemental Table 1: Number of images after quality filtering; number in parenthesis indicates the subset of each category that represent “common” diagnosis in dermatology

	Benign	Malignant	Total
Fitzpatrick I-II	159 (148)	49 (42)	208 (190)
Fitzpatrick III-IV	167 (153)	74 (65)	241 (218)
Fitzpatrick V-VI	159 (140)	48 (16)	207 (156)
Total Images	485 (441)	171 (123)	656 (564)

Supplemental Table 2: Performance of ModelDerm, DeepDerm, HAM10000 and an ensemble of dermatologists on the entire DDI dataset (656 images) and the subset of common diagnoses only (564 images) as well as stratification by light (FST I-II) vs dark (FST V-VI) skin tones. ROC-AUC for algorithms was calculated using the probability outputs of each respective algorithm. ROC-AUC for dermatologists was calculated using probabilities generated from ensembling each dermatologist's vote. Sensitivity and specificity of algorithms were calculated by using the cutoffs previously determined for ModelDerm, DeepDerm, and HAM10000 during algorithmic development on their respective test sets. Sensitivity and specificity values of the ensemble of dermatologists were calculated by using the majority vote as the label.

Labeler	Dataset	ROC-AUC			Sensitivity (Recall)			Specificity		
		All	FST I-II	FST V-VI	All	FST I-II	FST V-VI	All	FST I-II	FST V-VI
ModelDerm	DDI	0.65 (0.61-0.70)	0.64 (0.55-0.73)	0.55 (0.46-0.64)	0.36	0.41	0.12	0.83	0.75	0.89
	DDI Common diseases	0.74 (0.69-0.79)	0.68 (0.58-0.77)	0.70 (0.57-0.82)	0.47	0.45	0.25	0.83	0.77	0.88
DeepDerm	DDI	0.56 (0.51-0.61)	0.61 (0.50-0.71)	0.50 (0.41-0.58)	0.53	0.69	0.23	0.53	0.38	0.68
	DDI Common diseases	0.64 (0.58-0.69)	0.64 (0.54-0.75)	0.55 (0.42-0.67)	0.64	0.71	0.31	0.52	0.39	0.68
HAM10000	DDI	0.67 (0.62-0.71)	0.72 (0.63-0.79)	0.57 (0.48-0.67)	0.06	0.02	0.06	0.99	0.99	0.99
	DDI Common diseases	0.71 (0.66-0.76)	0.75 (0.66-0.82)	0.62 (0.47-0.77)	0.07	0.02	0.06	0.99	0.99	0.99
Dermatologist Ensemble	DDI	0.72 (0.68 - 0.77)	0.75 (0.68-0.81)	0.62 (0.54-0.71)	0.71	0.84	0.40	0.67	0.60	0.79
	DDI Common diseases	0.82 (0.79-0.86)	0.80 (0.73-0.85)	0.81 (0.70-0.90)	0.88	0.93	0.62	0.67	0.61	0.79

Supplemental Table 3: DeepDerm overall performance and performance on FST I-II and FST V-VI after using fairness-aware methods to train DeepDerm algorithm.

Method	Overall ROC-AUC mean (95% CI)	FST I-II ROC-AUC mean (95% CI)	FST V-VI ROC-AUC mean (95% CI)
CDANN[16]	0.59 (0.56-0.63)	0.61 (0.54-0.68)	0.48 (0.43-0.53)
CORAL[15]	0.58 (0.54-0.61)	0.61 (0.56-0.67)	0.45 (0.42-0.47)
GROUP-DRO[14]	0.60 (0.58-0.62)	0.63 (0.60-0.65)	0.50 (0.48-0.53)

Supplemental Table 4: Performance on FST III, which was not matched to FST I-II or FST V-VI

Labeler	**DDI ROC-AUC**	**DDI Common Diseases ROC-AUC**
ModelDerm	0.74 (0.66-0.80)	0.78 (0.71-0.84)
DeepDerm	0.60 (0.53-0.68)	0.63 (0.56-0.71)
HAM10000	0.69 (0.62-0.76)	0.71 (0.63-0.78)
Dermatologist Ensemble	0.78 (0.72-0.84)	0.82 (0.77-0.87)

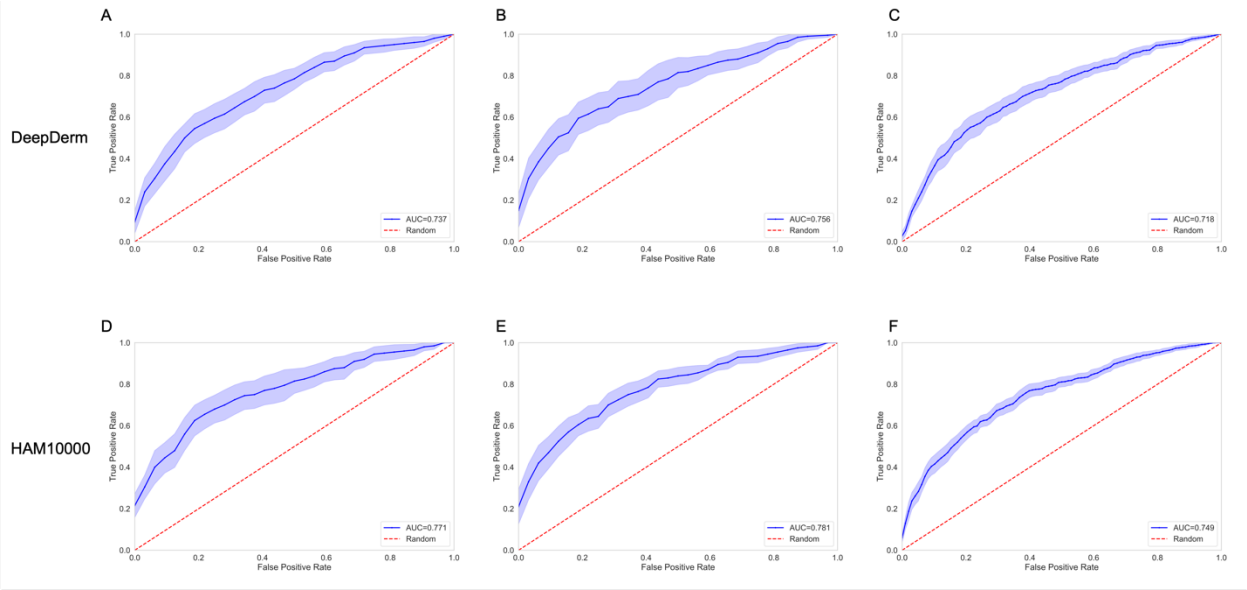
Supplemental Table 5: Performance on uncommon disease subset

Labeler	**ROC-AUC**
ModelDerm	0.48 (0.37-0.59)
DeepDerm	0.40 (0.28-0.52)
HAM10000	0.55 (0.43-0.67)
Dermatologist Ensemble	0.45 (0.34-0.57))

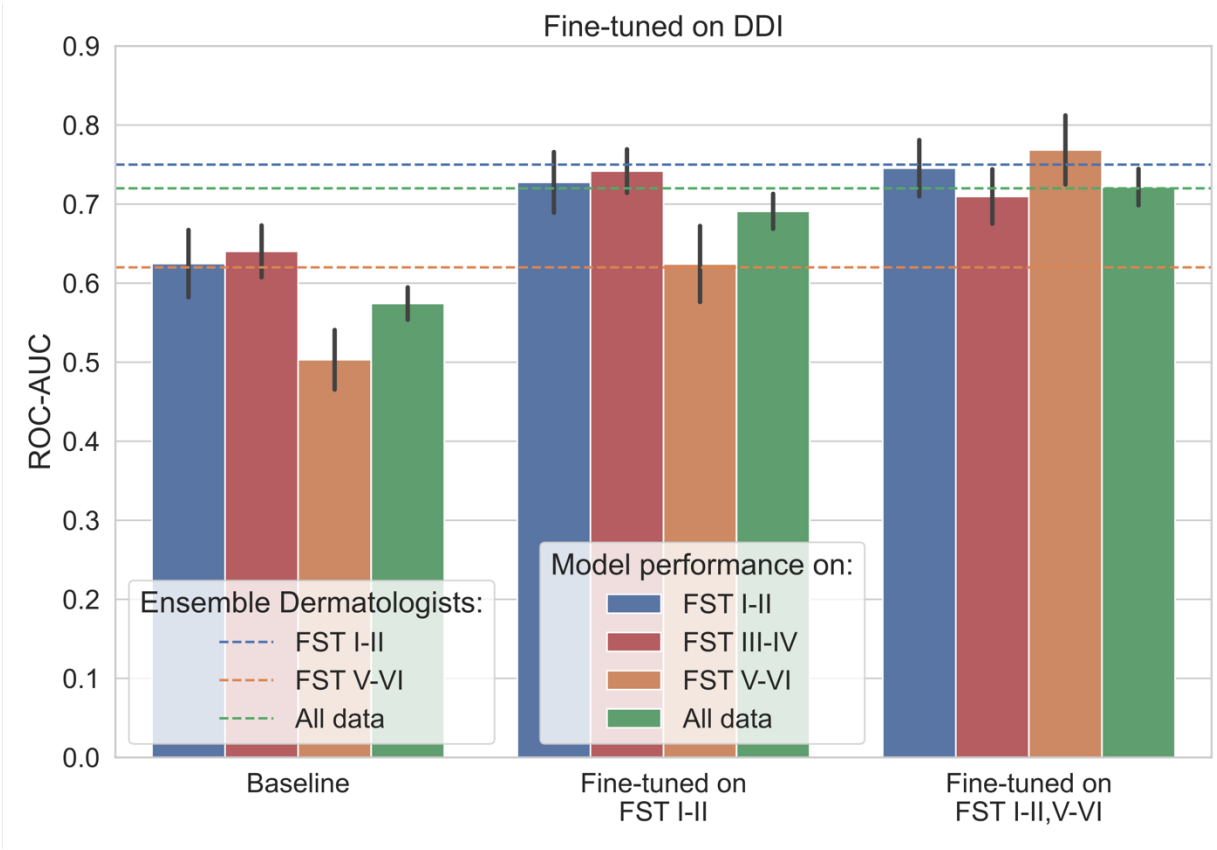
Supplemental Table 6: Data augmentation for fine-tuning models

Augmentation	**Details of augmentation**
Rotation	Random rotation of image, cropping to the largest inscribed rectangle.
Vertical Flip	Flip with 50% probability,
Resize and Cropping	Resize to 299x299 pixels, maintaining aspect ratio and cropping the longer side,
Color Jitter	Using torchvision.transforms.ColorJitter with parameters (brightness=0.1, contrast=0.1, saturation=0.1)
Blurring	Using torchvision.transforms.GaussianBlur with parameters (kernel_size=(5,9), sigma=(0.1,5)).

Supplemental Figure 1: ROC-AUC plots for fine-tuned DeepDerm and HAM10000 algorithm evaluated on FST I-II (A, D), FST V-VI (B,E), and all the data (C,F). Shaded area indicates 95% confidence intervals.



Supplemental Figure 2: DeepDerm fine-tuned on just FST I-II data and on FST I-II and V-VI data compared to DeepDerm baseline. FST V-VI test performance gap indicates the performance gain is largely due to fine-tuning on diverse skin data. 95% confidence interval is calculated using bootstrapping across the 20 seeds for both baseline and fine-tuned models to allow direct comparison.



Supplemental Figure 3: Inclusion and exclusion of DDI images.