Bias in Skin Lesion Classification

IW Talk: Spring 2022

Name: Tanushree Banerjee
Class year: 2024
Advisor: Prof Olga Russakovsky

Motivation

Opportunity to improve health care [1]

Dermatology Has a Problem With Skin Color

Common conditions often manifest differently on dark skin. Yet physicians are trained mostly to diagnose them on white skin.

By Roni Caryn Rabin

Aug. 30, 2020

CONTRIBUTE

STAT

TOPICS OPINION TEAM EVENTS NEWSLETTERS REPORTS VIDEO

TRENDING: CORONAVIRUS FIRST OPINION BIOTECH

HEALTH

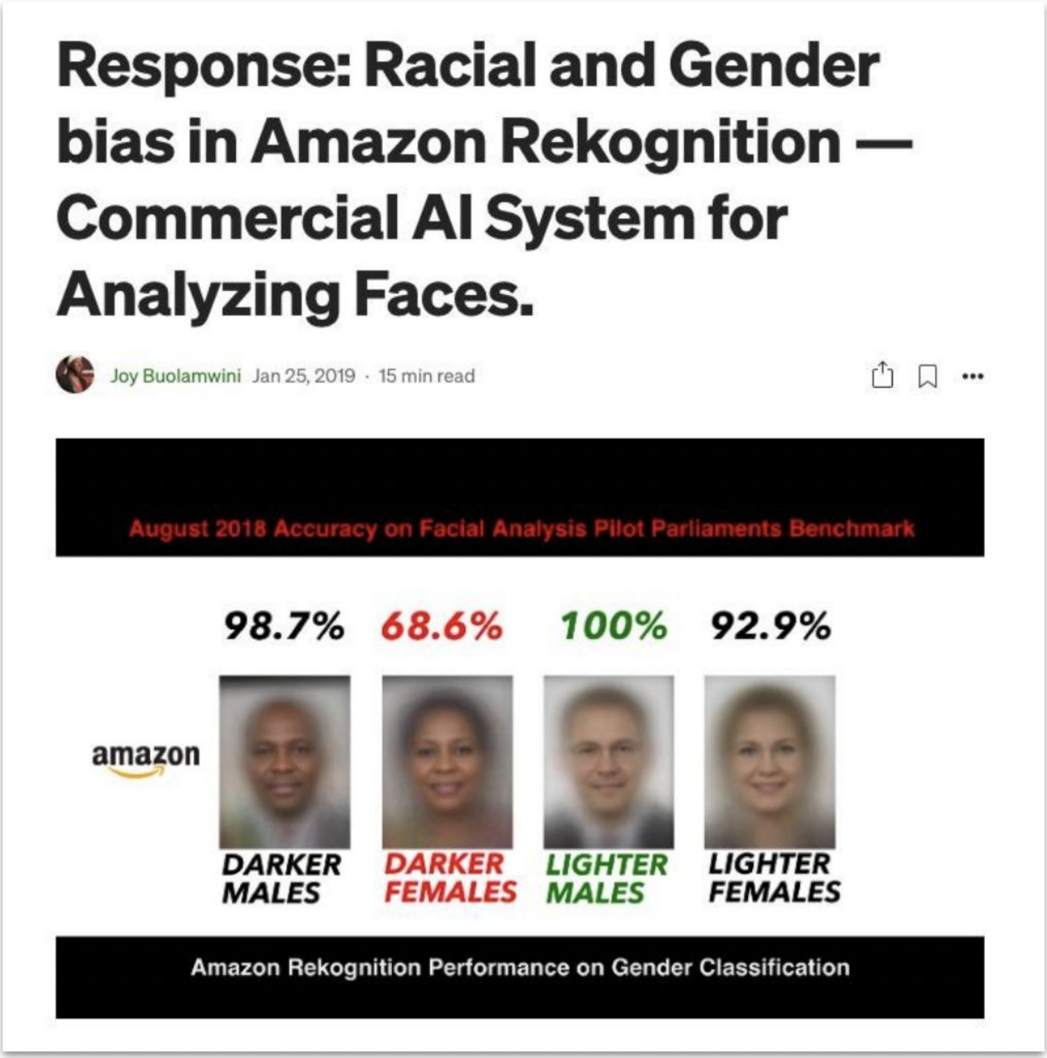
Dermatology faces a reckoning: Lack of darker skin in textbooks and journals harms care for patients of color

By Usha Lee McFarling July 21, 2020

[1] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639):115–118, Feb. 2017.

[2] "Dermatology Has a Problem With Skin Color." The New York Times - Breaking News, US News, World News and Videos. Last modified August 31, 2020. https://www.nytimes.com/2020/08/30/health/skin-diseases-black-hispanic.html.

[3] "Lack of Darker Skin in Textbooks, Journals Harms Patients of Color." STAT. Last modified July 20, 2020. https://www.statnews.com/2020/07/21/dermatology-faces-reckoning-lack-of-darker-skin-in-textbooks-journals-harms-patients-of-color/.



The problem

Gross overrepresentation of light skin in datasets (~11K light skin, ~6K darker skin - Fitzpatrick17k) [4]

Lack of consideration of subgroups within a population – large accuracy disparities across gender and skin color for facial recognition [5]

[4] Groh, Matthew, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badr. "Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset." arXiv:2104.09957 [cs.CV]. Last modified April 20, 2021. https://arxiv.org/pdf/2104.09957.pdf

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency, pages 77–91. PMLR, 2018.

The problem

Can lead to systematic bias against groups of people [6]

Potential to increase healthcare disparities in dermatology [7]

[6] Samaneh Abbasi-Sureshjani, Ralf Raumanns, Britt E. J. Michels, Gerard Schouten, and Veronika Cheplygina. Risk of Training Diagnostic Algorithms on Data with Demographic Bias. arXiv:2005.10050 [cs, stat], June 2020. arXiv: 2005.10050.

[7] Adewole S. Adamson and Avery Smith. Machine Learning and Health Care Disparities in Dermatology. JAMA Dermatology, 154(11):1247, Nov. 2018.

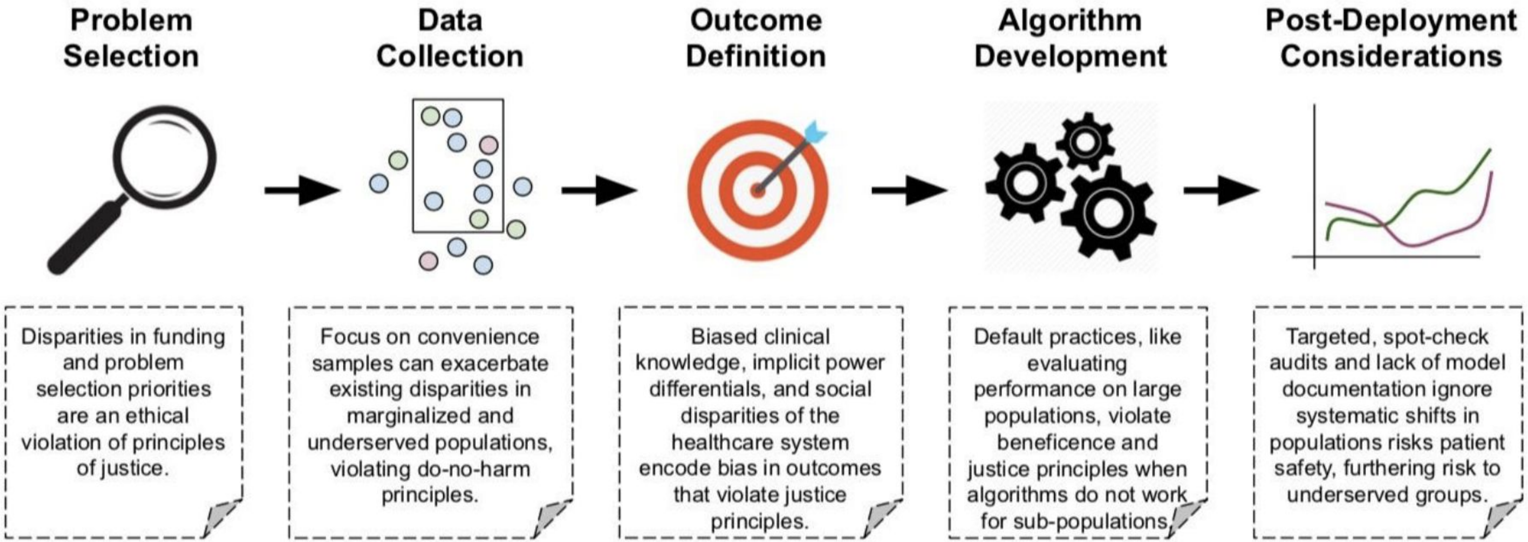
Goal

1. Train a machine learning model to classify skin conditions solely from images
   a. Inputs: An image of skin lesion(s)
   b. Output: Label for the skin condition displayed in the image
2. Examine distribution of images in dataset used: Fitzpatrick17K [4]
3. Examine discrepancies in accuracy across skin types to determine the existence of bias

[4] Groh, Matthew, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badr. "Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset." arXiv:2104.09957 [cs.CV]. Last modified April 20, 2021. https://arxiv.org/pdf/2104.09957.pdf

Significance of goal

1. Develop accurate models that can also serve as discrimination detectors
2. Identify opportunities to address this bias



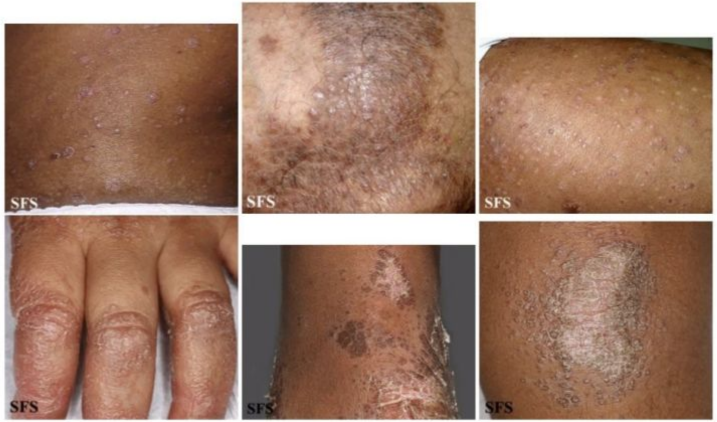
[8] Chen, Irene Y., Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. "Ethical Machine Learning in Health Care." ArXiv.org. Accessed April 19, 2022. https://arxiv.org/abs/2009.10576.

Fitzpatrick17K

16577 clinical images - from 2 online dermatology atlases

Each image: skin condition labels (3 levels) and skin type labels (Fitzpatrick scale)

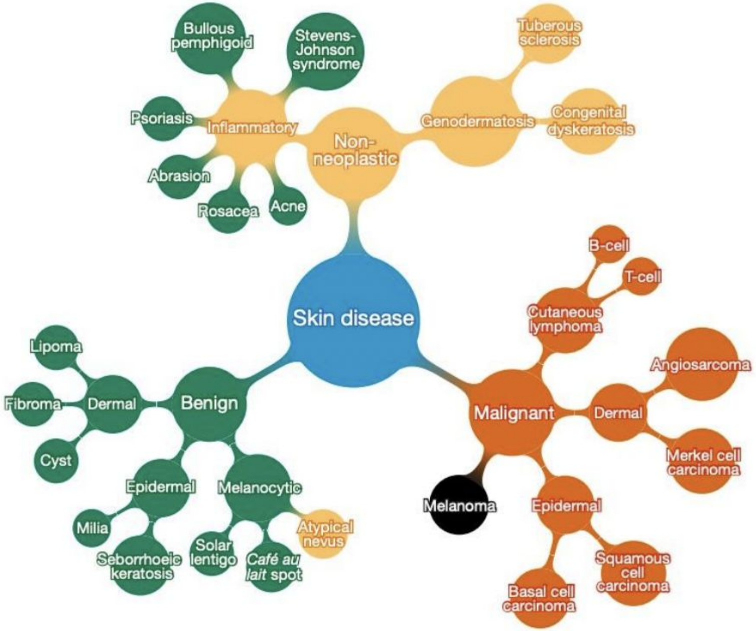
	A	B	C	D	E	F	G
1	md5hash	fitzpatrick	label	nine_partition_label	three_partition_label	qc	url
2	5e82a45bc5d78bd24ae9202d194423f8	3	drug induced pigmentary changes	inflammatory	non-neoplastic		https://www.dermaamin.com/site/images/clinical-pic/m/minocycline-pigmentation/minocycline-pigmentation1.jpg
3	fa2911a9b13b6f8af79cb700937cc14f	1	photodermatoses	inflammatory	non-neoplastic		https://www.dermaamin.com/site/images/clinical-pic/p/photosensitivity/photosensitivity18.jpg
4	d2bac3c9e4499032ca8e9b07c7d3bc40	2	dermatofibroma	benign dermal	benign		https://www.dermaamin.com/site/images/clinical-pic/d/dermatofibroma/dermatofibroma71.jpg
5	0a94359e7eaacd7178e06b2823777789	1	psoriasis	inflammatory	non-neoplastic		https://www.dermaamin.com/site/images/clinical-pic/p/psoriasis/psoriasis38.jpg
6	a39ec3b1f22				tic		https://www.dermaamin.com/site/images/clinical-pic/p/psoriasis-scalp/psoriasis-scalp20.jpg
7	45f7fe0e102						https://www.dermaamin.com/site/images/clinical-pic/k/kaposis-sarcoma/kaposis-sarcoma4.jpg
8	6c395be9325				tic		https://www.dermaamin.com/site/images/clinical-pic/s/sweet-syndrome/sweet-syndrome98.jpg
9	9dc73230c77				tic		https://www.dermaamin.com/site/images/clinical-pic/g/granuloma_annulare/granuloma_annulare41.jpg
10	f23937e86de				tic		https://www.dermaamin.com/site/images/clinical-pic/L/larva-migrans/larva-migrans88.jpg
11	09d46db9589				tic		https://www.dermaamin.com/site/images/clinical-pic/a/allergic_contact_dermatitis/allergic_contact_dermatitis114.jpg
12	9bc21ae9502				tic		https://www.dermaamin.com/site/images/clinical-pic/n/necrobiois-lipoidica-diabeticorum/necrobiois-lipoidica-diabeticorum88.jpg
13	e702b1a7dc4				tic		https://www.dermaamin.com/site/images/clinical-pic/s/sweet-syndrome/sweet-syndrome50.jpg
14	ddcad677b7b				tic		https://www.dermaamin.com/site/images/clinical-pic/h/hidradenitis_suppurativa/hidradenitis_suppurativa50.jpg
15	b87804452f6						https://www.dermaamin.com/site/images/clinical-pic/L/lmm/lmm6.jpg
16	d1fb87ee7ee				tic	1 Diagnostic	https://www.dermaamin.com/site/images/clinical-pic/a/acne_vulgaris/acne_vulgaris150.jpg
17	8438db40abd				tic		https://www.dermaamin.com/site/images/clinical-pic/n/necrobiois-lipoidica-diabeticorum/necrobiois-lipoidica-diabeticorum7.jpg
18	2d57e08861b				tic		https://www.dermaamin.com/site/images/clinical-pic/s/sarcoidosis-of-the-skin-plaque-form/sarcoidosis-of-the-skin-plaque-form15.jpg
19	1e119546f5b				tic		https://www.dermaamin.com/site/images/clinical-pic/X/xeroderma-pigmentosum/xeroderma-pigmentosum13.jpg
20	4c3f795cf8eb						https://www.dermaamin.com/site/images/clinical-pic/m/melanoma/melanoma17.jpg
21	99247c9fe48						https://www.dermaamin.com/site/images/clinical-pic/d/dermatofibroma/dermatofibroma13.jpg
22	b09233673fc						https://www.dermaamin.com/site/images/clinical-pic/a/actinic_keratoses/actinic_keratoses83.jpg
23	449d63bec3a				tic		https://www.dermaamin.com/site/images/clinical-pic/L/localized-scleroderma/localized-scleroderma3.jpg
24	7a06b6baf51				tic		https://www.dermaamin.com/site/images/clinical-pic/h/hidradenitis_suppurativa/hidradenitis_suppurativa18.jpg
25	fb9640a13e0						https://www.dermaamin.com/site/images/clinical-pic/s/syringoma/syringoma33.jpg



[4] Groh, Matthew, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badr. "Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset." arXiv:2104.09957 [cs.CV]. Last modified April 20, 2021. https://arxiv.org/pdf/2104.09957.pdf

Skin condition labels

3 high-level categories, 9 mid-level categories, 114 low-level categories



[8] Esteva et al 2017 Dermatologist-level classification of skin cancer

Fitzpatrick skin type labels

-1: labelled as unknown skin type

[4] Groh, Matthew, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badr. "Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset." arXiv:2104.09957 [cs.CV]. Last modified April 20, 2021. https://arxiv.org/pdf/2104.09957.pdf

Related work

Groh et al.:

- present the Fitzpatrick 17k: dataset consisting of 16,577 clinical images of 114 different skin conditions annotated with Fitzpatrick skin type labels
  - Reveal underrepresentation of dark skin images in online dermatology atlases
  - Reveal accuracy disparities that arise from training a neural network on a subset of skin types
- Train a deep neural network model to classify 114 skin conditions
  - Find that the model is most accurate on skin types similar to those it was trained on

[4] Groh, Matthew, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badr. "Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset." arXiv:2104.09957 [cs.CV]. Last modified April 20, 2021. https://arxiv.org/pdf/2104.09957.pdf

Related Work

Bissoto et al.:

- Bias analysis on other skin lesion datasets: Atlas, ISIC
- Datasets not labelled by skin type
- Analysis focused on discrepancies due to spurious correlations, but not based

[12] Bissoto, Alceu, Michel Fornaciali, Eduardo Valle, and Sandra Avila. "(De)Constructing Bias on Skin Lesion Datasets." ArXiv.org. Accessed April 19, 2022. https://arxiv.org/abs/1904.08818.

Approach and key idea

Analysis in previous work is done on 114 low-level categories classification and 3-high level categories classification, but not on the 9 mid-level categories classification

- Mid-level categories have more images per label. Hence, conclusions drawn might be less noisy than 114 way classification
- More likely that there is sufficient data to fully evaluate accuracy discrepancies
- More fine grained information than 3-way classifier

Extract image features from Alexnet [9] pretrained on ImageNet [10], with the penultimate layer removed

- reduces runtime compared to previous paper, and removes the need for access to a GPU

[9] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. Last modified December 2012. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[10] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

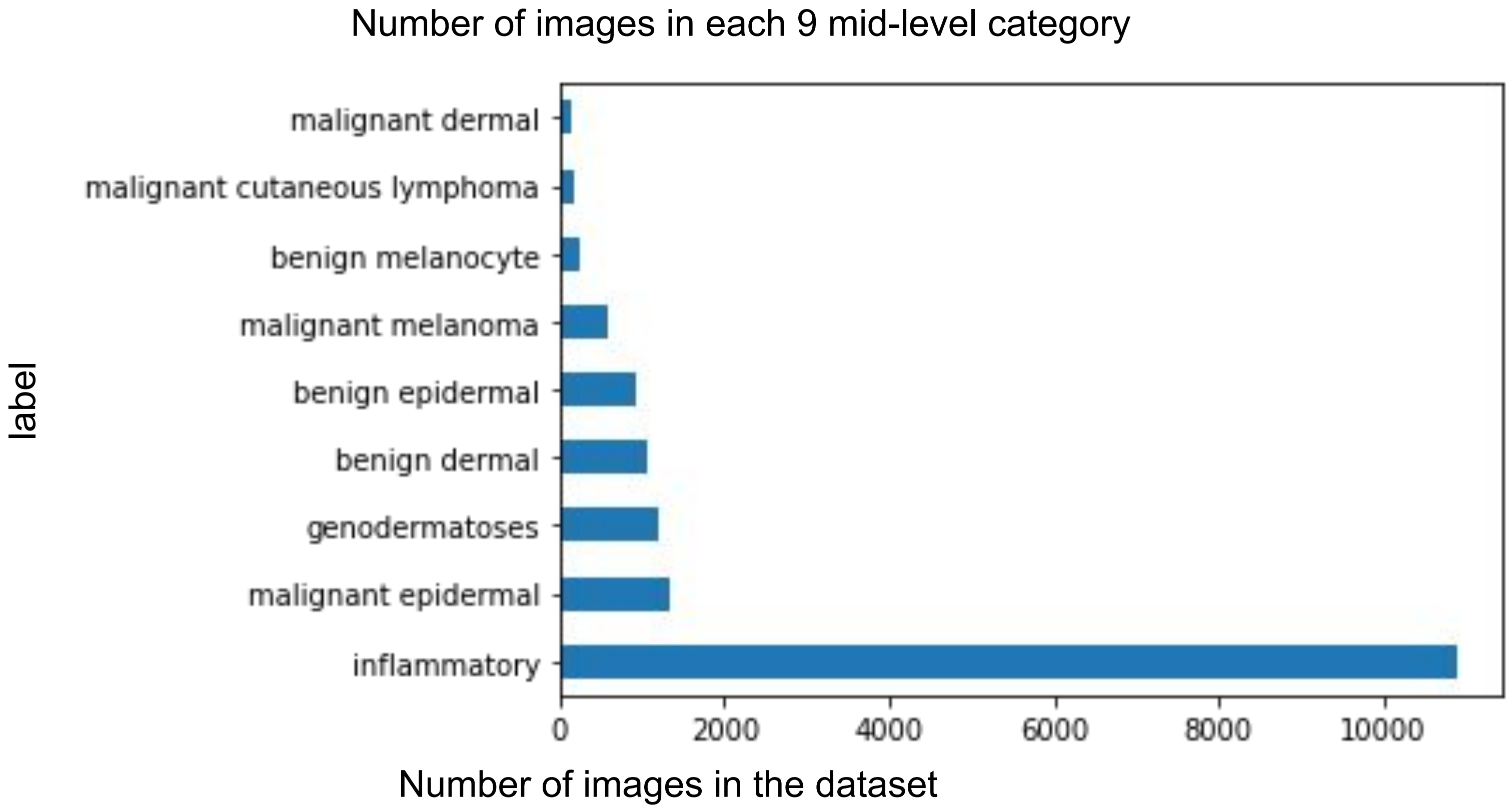
Implementation

1. Randomly shuffle dataset, then split into the train/validation set and test set
2. Use the Alexnet architecture with the final classification layer removed
3. Pretrain the Alexnet [9] architecture on Imagenet [10]
4. Perform one forward pass for each image, storing the values obtained at the penultimate classification layer as the feature vector for the image
5. Using 9-fold cross validation on a linear classifier to obtain the best regulariser value
6. Train the linear classifier using the best regularisation value obtained from the previous step
7. Evaluate accuracies across skin types on the test set

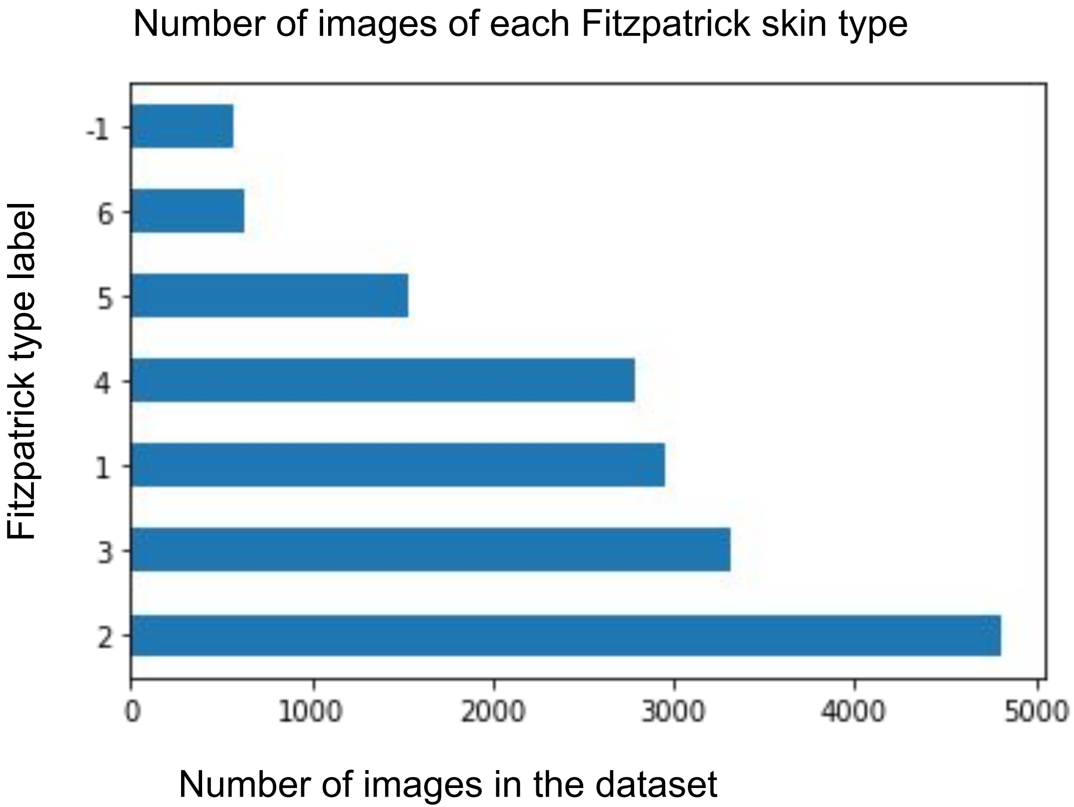
[9] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. Last modified December 2012. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[10] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

Skin condition label distribution in dataset

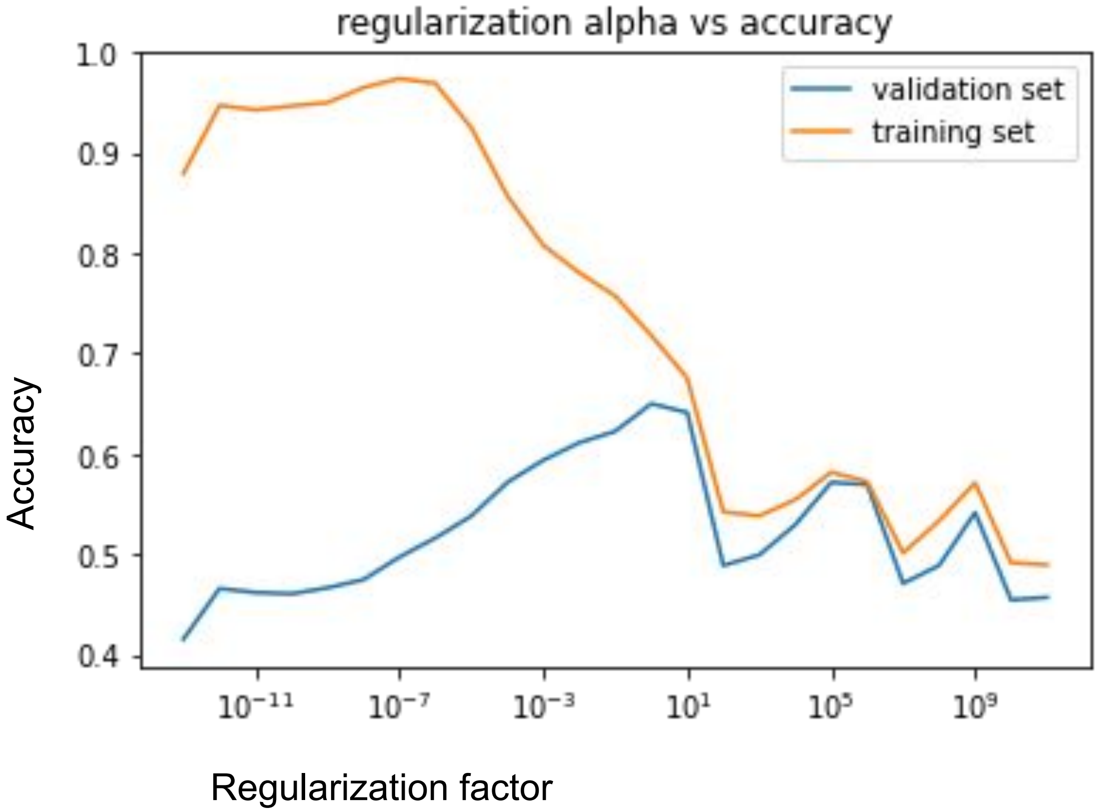


Skin type label distribution in dataset



Note: -1 is unknown skin type

9-fold cross validation to find optimal regularisation factor



Best regularisation value = 1e-1
Best validation set accuracy = 0.65

Confusion matrix on test set (9-way classification)

	Predicted label									
True label		**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
	0	11	3	0	0	78	0	0	4	0
	1	1	9	0	0	75	0	0	4	0
	2	1	0	0	0	20	0	0	0	1
	3	0	1	0	18	89	0	0	0	0
	4	5	3	1	5	1115	0	0	7	3
	5	0	0	0	0	28	0	0	0	0
	6	0	0	0	0	17	0	1	0	0
	7	1	3	0	0	73	0	0	36	2
	8	0	2	0	0	26	0	0	4	11

Confusion matrix (Normalised)

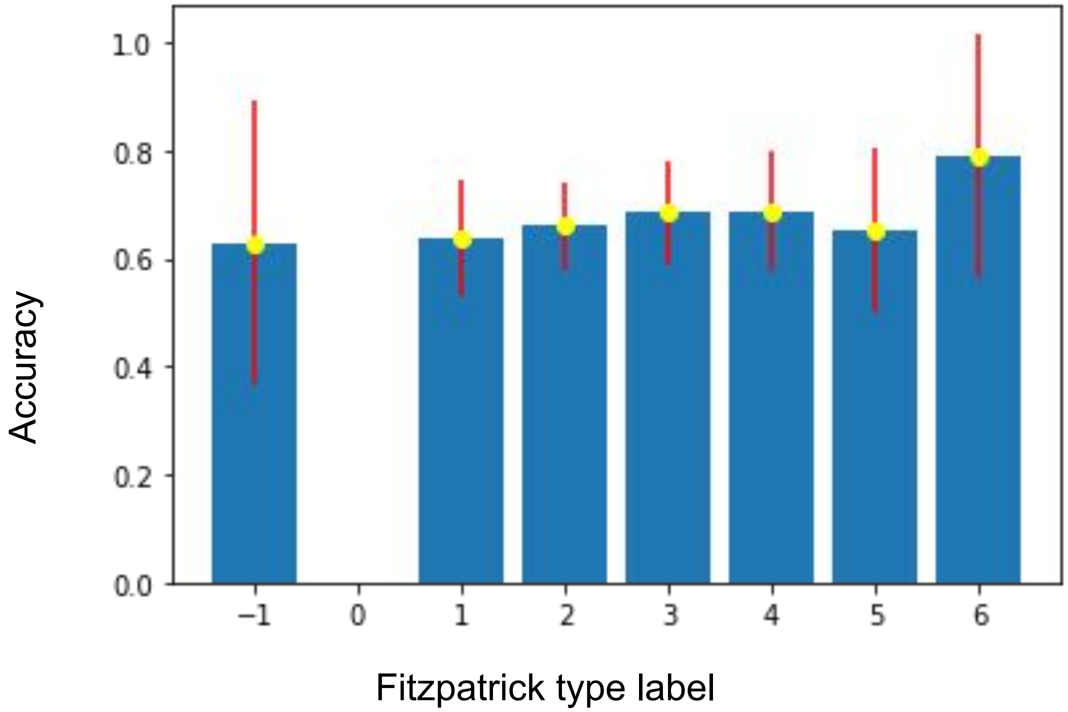
	Predicted label									
True label	0.1146	0.0313	0	0	0.8125	0	0	0.0417	0	0.1146
	0.0112	0.1011	0	0	0.8427	0	0	0.0449	0	0.0112
	0.0455	0	0	0	0.9091	0	0	0	0.0455	0.0455
	0	0.0093	0	0.1667	0.8241	0	0	0	0	0
	0.0044	0.0026	0.0009	0.0044	0.9789	0	0	0.0061	0.0026	0.0044
	0	0	0	0	1.0000	0	0	0	0	0
	0	0	0	0	0.9444	0	0.0556	0	0	0
	0.0087	0.0261	0	0	0.6348	0	0	0.3130	0.0174	0.0087
	0	0.0465	0	0	0.6047	0	0	0.0930	0.2558	0
	0.1146	0.0313	0	0	0.8125	0	0	0.0417	0	0.1146

Evaluation of accuracy differences via Bootstrapping [11]

To obtain error bounds over the accuracies by skin type

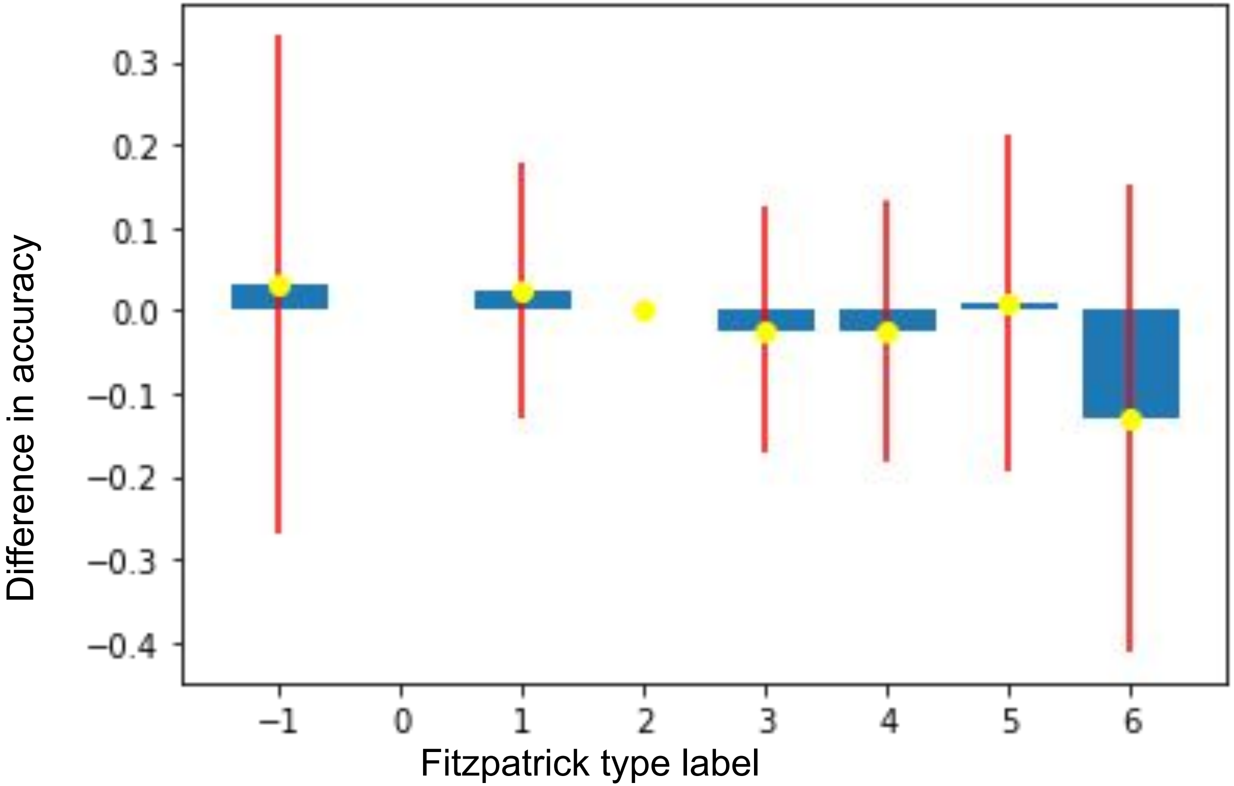
[11] Everingham, Mark, S. M. Ali Eslami, Luc Van Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. "Assessing the Significance of Performance Differences on the PASCAL VOC Challenges via Bootstrapping." The PASCAL Visual Object Classes. Last modified October 18, 2013. https://host.robots.ox.ac.uk/pascal/VOC/pubs/bootstrap_note.pdf.

Accuracy on test set by skin type label



Difference in accuracy from that of skin type 2

0 is contained in each of the confidence intervals for each skin type



Conclusions

Discrepancies in accuracies across skin types are not statistically significant - i.e. there is not enough evidence to conclude that there is a statistically significant difference in accuracy across skin types at the 95% confidence level

Negative bias result does not imply the model is unbiased, just that current analysis did not reveal it

Future work

Next projects: Analysis of accuracy by skin condition, analysis by changing the distribution of images by skin type in the training set versus the test set

Long term: Encourage evaluation accuracy across subpopulations where classification accuracy is suspected to be heterogeneous

Acknowledgments

I cannot thank **Prof Olga Russakovsky** enough for the opportunity to conduct independent work under her guidance as well as her advice on the best way engage with current research in computer vision and more generally in the field of computer science. This project could not have been conducted without her support.

Thank you!

References

Reference code: https://github.com/mattgroh/fitzpatrick17k/blob/main/train.py

[1] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639):115–118, Feb. 2017.

[2] "Dermatology Has a Problem With Skin Color." The New York Times - Breaking News, US News, World News and Videos. Last modified August 31, 2020. https://www.nytimes.com/2020/08/30/health/skin-diseases-black-hispanic.html.

[3] "Lack of Darker Skin in Textbooks, Journals Harms Patients of Color." STAT. Last modified July 20, 2020. https://www.statnews.com/2020/07/21/dermatology-faces-reckoning-lack-of-darker-skin-in-textbooks-journals-harms-patients-of-color/.

[4] Groh, Matthew, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badr. "Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset." arXiv:2104.09957 [cs.CV]. Last modified April 20, 2021. https://arxiv.org/pdf/2104.09957.pdf

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency, pages 77–91. PMLR, 2018.

[6] Samaneh Abbasi-Sureshjani, Ralf Raumanns, Britt E. J. Michels, Gerard Schouten, and Veronika Cheplygina. Risk of Training Diagnostic Algorithms on Data with Demographic Bias. arXiv:2005.10050 [cs, stat], June 2020. arXiv: 2005.10050.

References

[7] Adewole S. Adamson and Avery Smith. Machine Learning and Health Care Disparities in Dermatology. JAMA Dermatology, 154(11):1247, Nov. 2018.

[8] Esteva et al 2017 Dermatologist-level classification of skin cancer

[9] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. Last modified December 2012. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[10] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[11] Everingham, Mark, S. M. Ali Eslami, Luc Van Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. "Assessing the Significance of Performance Differences on the PASCAL VOC Challenges via Bootstrapping." The PASCAL Visual Object Classes. Last modified October 18, 2013. https://host.robots.ox.ac.uk/pascal/VOC/pubs/bootstrap_note.pdf.

[12] Bissoto, Alceu, Michel Fornaciali, Eduardo Valle, and Sandra Avila. "(De)Constructing Bias on Skin Lesion Datasets." ArXiv.org. Accessed April 19, 2022. https://arxiv.org/abs/1904.08818.