Fairness of Classifiers Across Skin Tones in Dermatology

Newton Mwai Kinyanjui, Timothy Odonga, Celia Cintas, Noel C. F. Codella, Rameswar Panda, Prasanna Sattigeri and Kush R. Varshney

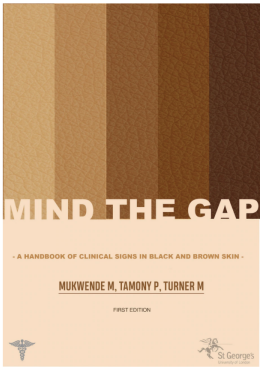
IBM Research | Africa
Carnegie Mellon University
Africa College of Engineering

Background

Disparities with respect to ethnicity exist in dermatology, as in other medical fields

- In black population, melanoma is often diagnosed at an advanced stage with deeper tumors [MSL+17, WEK+11].
- 5 year survival rates for acral lentiginous melanoma (ALM) is 82.6% in whites, but only 77.2% in blacks [MCH15].
- The paucity of images of skin manifestations of COVID-19 in patients with darker skin is a problem, because it may make identification of COVID-19 presenting with cutaneous manifestations more difficult for both dermatologists and the public. [LJZ+20]

Machine learning systems may place certain groups of people at a systematic disadvantage due to dataset bias [BS16].



The cover of the "Mind the Gap" handbook, written by Malone Mukwende, with two of his lecturers, Peter Tamony and Margot Turner.

IBM Research | Africa

Research Questions

1. Are standard **dermatology image datasets** used in ML tasks **biased with respect to skin tone**? Can we quantify this?
2. If so, does the dataset bias lead to **unequal performance** of downstream disease classification?

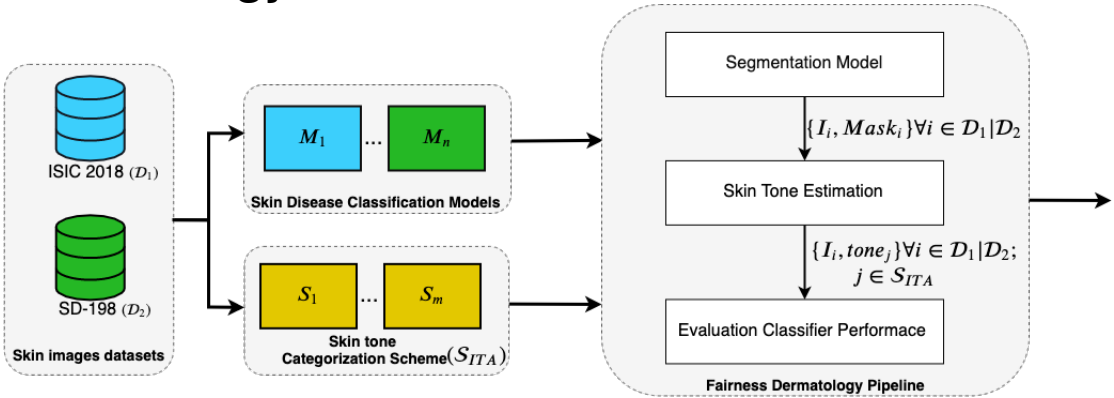
IBM Research | Africa

Related Work

- Skin disease diagnosis using machine learning
  1. Benchmark model for melanoma diagnosis outperforms trained dermatologists [CNP+16]
  2. ISIC challenges (https://www.isic-archive.com/)
- Predictive inequity in computer vision with respect to skin type
  1. Automated face image analysis for gender classification [BG18]
  2. Pedestrian detection systems [WHM19]

IBM Research | Africa

Methodology

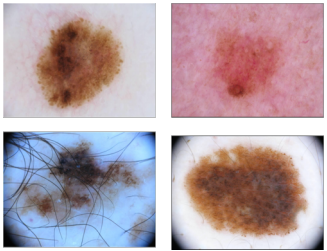


IBM Research | Africa

Datasets

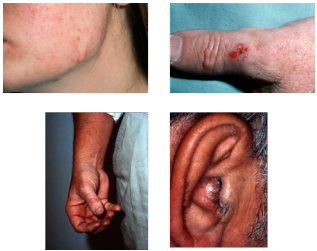
ISIC 2018

- 10015 dermoscopic images
- 7 disease classes
- 2594 images with ground truth segmentation masks for diseased area



SD-198

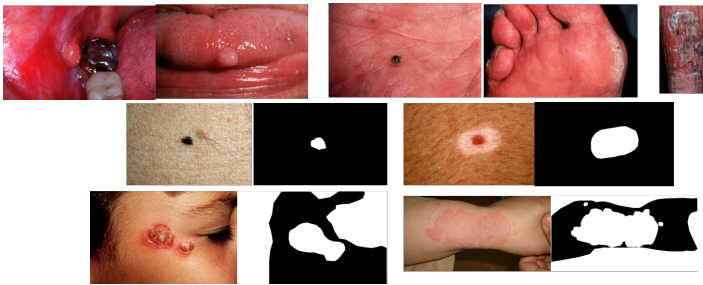
- 6548 clinical images
- 198 disease classes
- No segmentation data



IBM Research | Africa

Pre-processing: Creating SD-136 Dataset

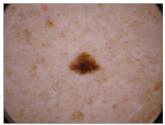
1. Manually remove:
   - Classes with images containing no observable non-diseased skin.
   - Images of palms, soles of the feet, inside of the mouth.
   - E.g., Fibroma, Arsenical Keratosis, Stasis Ulcer
   - Classes reduced from 198 to 136: SD-136
2. Manually segment a subset of 343 images to create segmentation masks.
3. Segment regions of non-diseased skin from other regions.

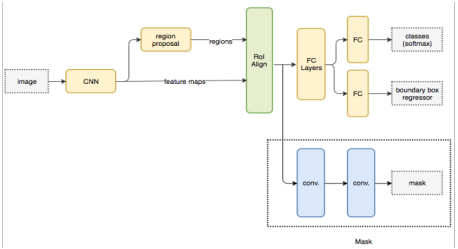


IBM Research | Africa

Segmentation to Obtain Non-Diseased Region

1. Finetune Mask R-CNN model [HGDG17]
   - Adjust pretrained classifier with a FastRCNNPredictor with 2 classes (background and mask)
   - Adjust mask predictor with new MaskRCNNPredictor with 2 classes and 512 hidden neurons
2. Further apply thresholding techniques on predicted grayscale mask including contour extraction for ISIC2018 and grid search for optimal binary thresholding for SD-136





IBM Research | Africa

Skin Tone Metric of Non-Diseased Region

1. Given non-diseased pixels, characterize them with a skin tone metric
   1. Use individual typology angle (ITA) [WWdPR15], Highly correlated with melanin index
   2. $\text{ITA} = \tan^{-1}\left(\frac{L-50}{b}\right) \times \frac{180}{\pi}$ Where L is luminance and b quantifies amount of yellow.
   3. Use pixels with L and b values within 1 standard deviation to deal with outliers.
2. Bin into categories [CSD$^+$15]

ITA Range	Skin Tone Category	Abbreviation
ITA $> 55^\circ$	Very Light	very_lt
$48^\circ <$ ITA $\leq 55^\circ$	Light 2	lt2
$41^\circ <$ ITA $\leq 48^\circ$	Light 1	lt1
$34.5^\circ <$ ITA $\leq 41^\circ$	Intermediate 2	int2
$28^\circ <$ ITA $\leq 34.5^\circ$	Intermediate 1	int1
$19^\circ <$ ITA $\leq 28^\circ$	Tanned 2	tan2
$10^\circ <$ ITA $\leq 19^\circ$	Tanned 1	tan1
ITA $\leq 10^\circ$	Dark	dark

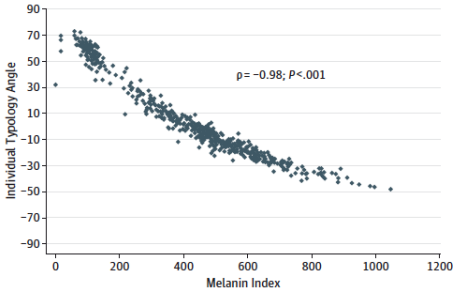


Figure from [WWdPR15].

IBM Research | Africa

Classification of Skin Disease

We replicated state-of-the-art skin disease classification neural networks from ISIC2018:

1. Finetune Densenet 201 model pretrained on ImageNet [HLVDMW17].
2. Regularization methods: Dropout and early stopping.

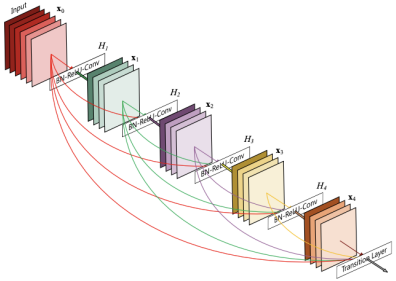


Figure from [HLVDMW17]

Overall accuracy on validation data

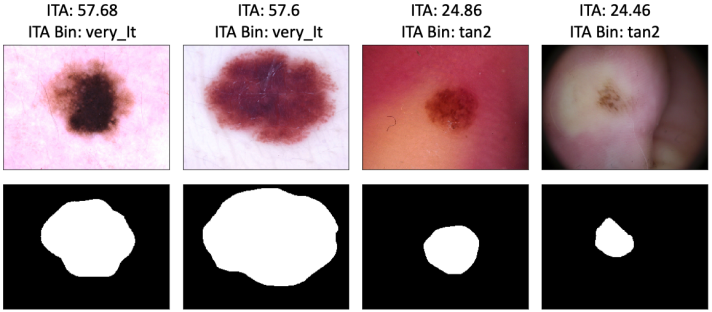
- ISIC2018: balanced accuracy 0.884 (benchmark model 0.885)
- SD-136: accuracy score 0.567 (benchmark model 0.52)

IBM Research | Africa

Results

Metrics for segmentation on ISIC 2018

The Mask R-CNN model yields an accuracy of **0.956**, a false negative rate of **0.024**, and a mean absolute error in ITA computation of **0.428** degrees. [KOC+19]

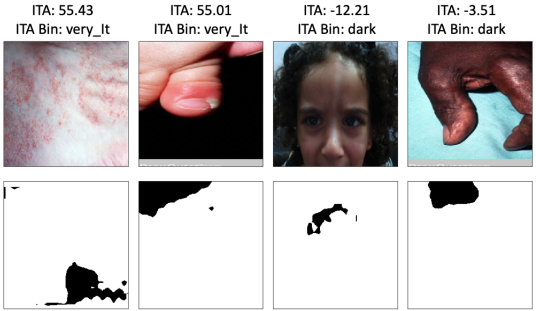


IBM Research | Africa

Results (Cont.)

Metrics for segmentation on SD-136

The segmentation model on the SD-136 dataset yield an accuracy of **0.802**, a false negative rate of **0.076**, and a mean absolute error in ITA computation of **3.572** degrees. [KOC+19]

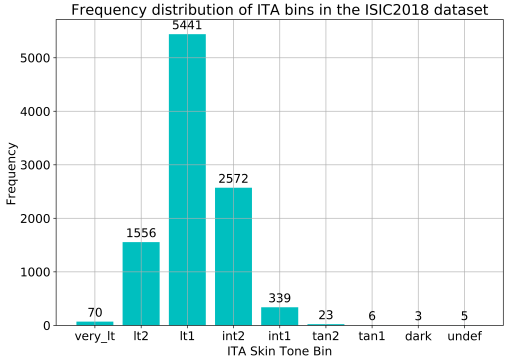


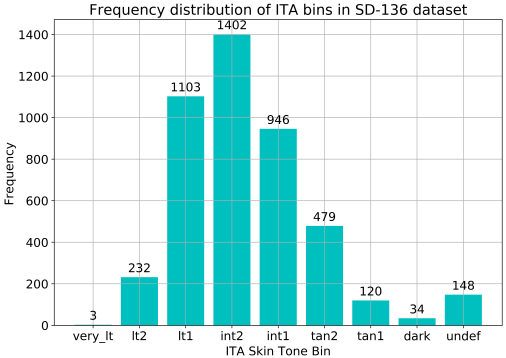
IBM Research | Africa

Results (Cont.)

Skin Tone Distribution

There is underrepresentation of darker skin tones in both datasets



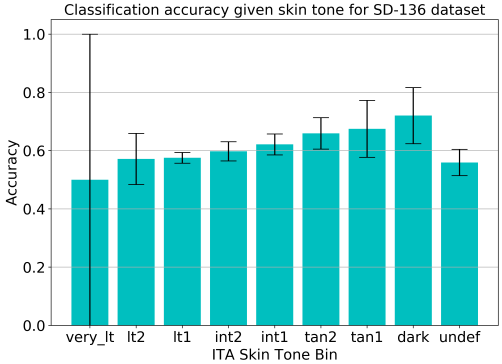


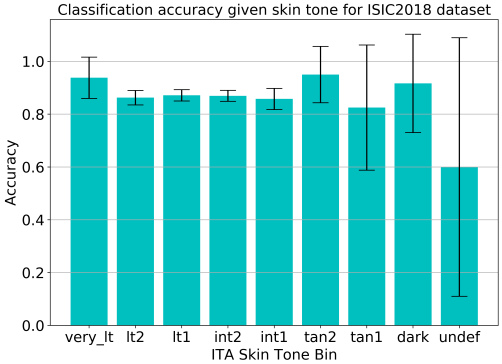
IBM Research | Africa

Results (cont.)

Accuracy by Skin Tone

There is no statistically significant pattern of unequal performance by skin tone





Conclusions & Future Work

- The two skin disease datasets are biased towards lighter skin with majority of the samples between ITA values $[34.5°, 48°]$.
- The bias in the datasets does not seem to affect the classification results for each skin tone category. Caveat: this conclusion is from datasets with significantly fewer dark-skinned samples.

We're currently working on:

1. Creation of datasets with more evenly distributed samples by skin tone categories.
2. Creation of ground truth segmentation masks for the entire SD-198 dataset.
3. Implementation of better segmentation models for clinical images.

IBM Research | Africa

Asante, Thanks, Gracias!

todonga@alumni.cmu.edu
nkinyanj@alumni.cmu.edu

IBM Research | Africa

References I

- Joy Buolamwini and Timnit Gebru, *Gender shades: Intersectional accuracy disparities in commercial gender classification*, Proc. Conf. Fair. Account. Transp., February 2018, pp. 77–91.
- Solon Barocas and Andrew D. Selbst, *Big data's disparate impact*, Calif. Law Rev. 104 (2016), no. 3, 671–732.
- Noel C. F. Codella, Quoc-Bao Nguyen, Sharath Pankanti, David A. Gutman, Brian Helba, Allan C. Halpern, and John R. Smith, *Deep learning ensembles for melanoma recognition in dermoscopy images*, IBM J. Res. Dev. 61 (2016), no. 4/5, 5.
- Giuseppe R. Casale, Anna Maria Siani, Henri Diémoz, Giovanni Agnesod, Alfio V. Parisi, and Alfredo Colosimo, *Extreme UV index and solar exposures at Plateau Rosà (3500 m a.s.l.) in Valle d'Aosta Region, Italy*, Sci. Total Environ. 512–513 (2015), 622–630.

IBM Research | Africa

References II

- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B Girshick, *Mask r-cnn. corr abs/1703.06870 (2017)*, arXiv preprint arXiv:1703.06870 (2017).
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, *Densely connected convolutional networks*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.
- Newton M Kinyanjui, Timothy Odonga, Celia Cintas, Noel CF Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R Varshney, *Estimating skin tone and effects on classification performance in dermatology datasets*, arXiv preprint arXiv:1910.13268 (2019).
- JC Lester, JL Jia, L Zhang, GA Okoye, and E Linos, *Absence of skin of colour images in publications of covid-19 skin manifestations*, British Journal of Dermatology (2020).

IBM Research | Africa

References III

- Michael A. Marchetti, Esther Chung, and Allan C. Halpern, *Screening for acral lentiginous melanoma in dark-skinned individuals*, JAMA Dermatol. 151 (2015), no. 10, 1055–1056.
- Krishnaraj Mahendraraj, Komal Sidhu, Christine S. M. Lau, Georgia J. McRoy, Ronald S. Chamberlain, and Franz O. Smith, *Malignant melanoma in African–Americans: A population-based clinical outcomes study involving 1106 African–American patients from the surveillance, epidemiology, and end result (SEER) database (1988—2011)*, Medicine 96 (2017), no. 15, e6258.

IBM Research | Africa

References IV

- Xiao-Cheng Wu, Melody J. Eide, Jessica King, Mona Saraiya, Youjie Huang, Charles Wiggins, Jill S. Barnholtz-Sloan, Nicolle Martin, Vilma Cokkinides, Jacqueline Miller, Pragna Patel, Donatus U. Ekwueme, and Julian Kim, *Racial and ethnic variations in incidence and survival of cutaneous melanoma in the United States, 1999-2006*, J. Am. Acad. Dermatol. 65 (2011), no. 5, S26.e1–S26.e13.
- Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern, *Predictive inequity in object detection*, arXiv:1902.11097, February 2019.
- Marcus Wilkes, Caradee Y. Wright, Johan L. du Plessis, and Anthony Reeder, *Fitzpatrick skin type, individual typology angle, and melanin index in an African population*, JAMA Dermatol. 151 (2015), no. 8, 902–903.

IBM Research | Africa