Towards Fairness & Robustness in Machine Learning for Dermatology

Skin-tone representation disparities in dermatology datasets for machine learning applications

Celia Cintas, PhD

IBM Research | Africa

DART
International Workshop on Domain Adaptation & Representation Transfer

Photo: HYACINTH EMPINADO/STAT

The team

IBM Research

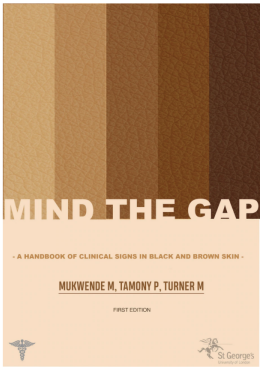
Carnegie Mellon University Africa College of Engineering

Stanford University

IBM Research | Africa

Disparities in Dermatology

- In African American 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 caucasian population, but only 77.2% in african american patients. [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]
- Dermatologists started an international registry to catalog examples of skin manifestations of Covid-19. The registry compiled more than 700 cases, but only 34 of disorders in Hispanic and 13 in Black patients were submitted. [Rab20]



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

How these disparities are reflected in Healthcare Machine Learning models?

1. Are standard **dermatology image datasets** used in ML tasks **biased with respect to skin tone**? Can we quantify this?
2. Are ML models **robust against changes** in the clinical setting or unknown diseases samples?

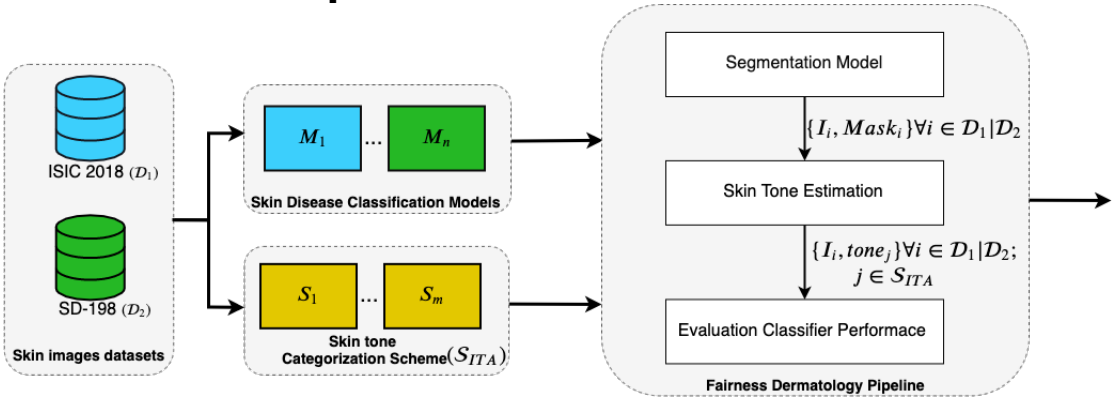
IBM Research | Africa

Machine Learning & Dermatology

- 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]
- Out-of-distribution detection in dermatology [AYAG19, GNS+19, ZZL19, CHP+ss, PAT19, PST+20]

IBM Research | Africa

Overview : Proposed Framework



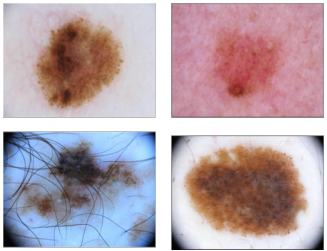
Kinyanjui, et al. "Estimating skin tone and effects on classification performance in dermatology datasets."MICCAI 2020.

IBM Research | Africa

Datasets

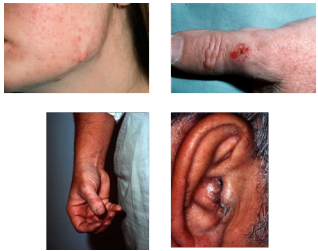
ISIC 2018

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



SD-198

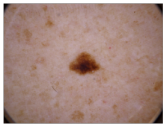
- 6548 clinical images
- 198 disease classes
- No segmentation data

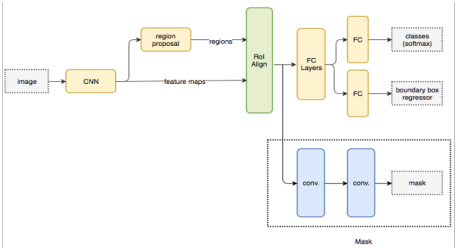


IBM Research | Africa

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

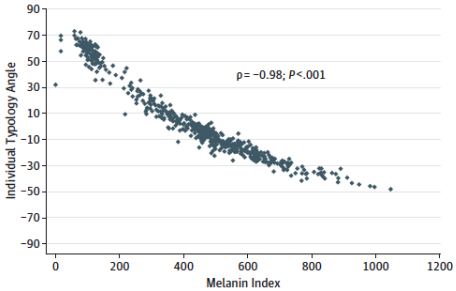


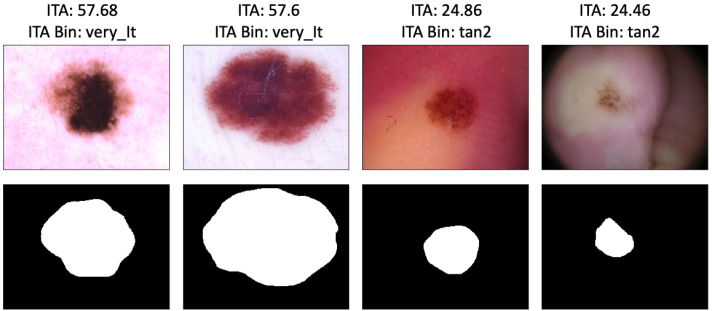
Figure from [WWdPR15].

IBM Research | Africa

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]

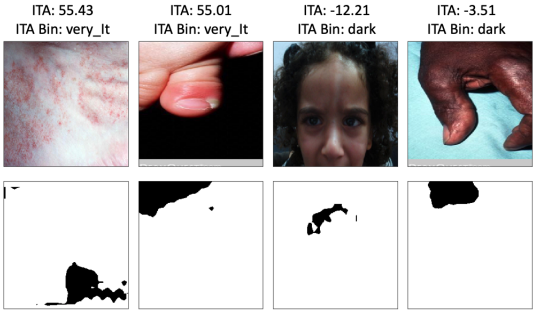


IBM Research | Africa

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]

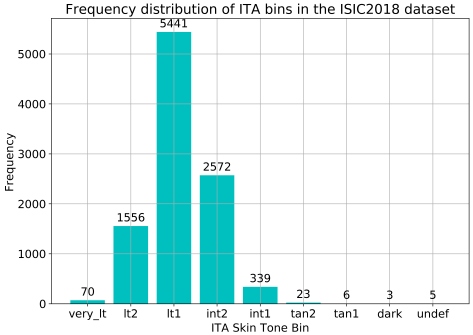


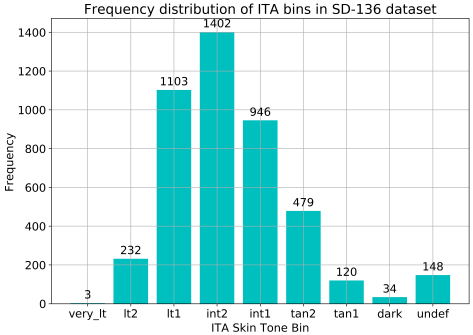
IBM Research | Africa

Results (Cont.)

Skin Tone Distribution

There is underrepresentation of darker skin tones in both datasets



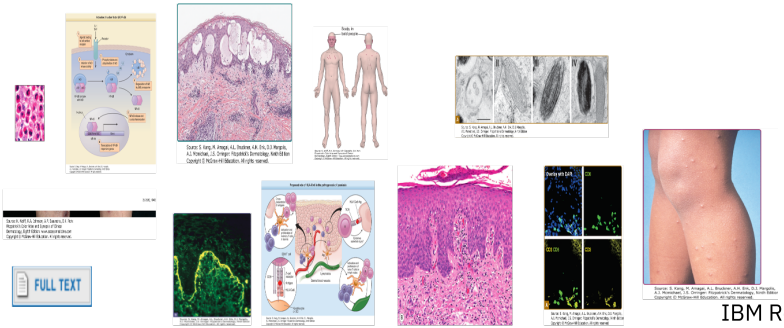


IBM Research | Africa

Extensions

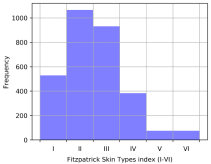
Automatic report of skin tone distributions

Kim et al. are currently working on extending the current segmentation and classification models to work on dermatology textbooks and academic paper images.

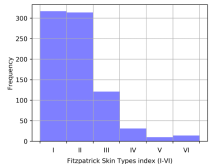


IBM Research | Africa

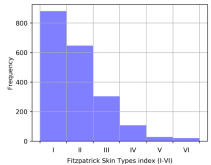
Preliminary Results



(a) Bolognia 4e



(b) Atlas 8e



(c) General Med 9e

Validation across three dermatology textbooks labeled by dermatology professionals.

Bolognia 4e

Atlas 8e

General Med 9e

We can observe the lack of darker skin tones across all the evaluated books

	Bolognia	Atlas	General Med
GT	11.92%	9.56%	11.78%
Proposed	17.42%	6.82%	7.84%

Tadesse ,A. G, Kim ,H., Daneshjou ,R., Cintas ,C., Varshney ,K., Adelekun, A., Lipoff ,J., Onyekab, G., Rotemberg, V., Zou, J. Automated Evaluation of Representation in Dermatology Educational Materials. In AAAI 2021 Workshop: Trustworthy AI for Healthcare.

IBM Research | Africa

How these disparities are reflected in Healthcare Machine Learning models?

1. Are standard dermatology image datasets used in ML tasks biased with respect to skin tone? Can we quantify this?
2. Are ML models **robust against changes** in the clinical setting or unknown diseases samples?

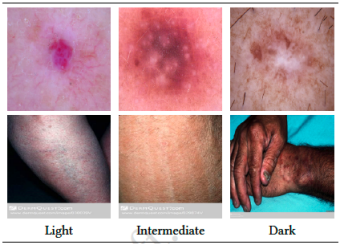
IBM Research | Africa

OOD for Skin Disease Classifiers

Recent advances in deep learning have led to breakthroughs in the development of automated skin disease classification. As we observe an **increasing interest** in these models in the **dermatology space**, it is crucial to address aspects such as the **robustness** and fairness of these solutions.

We validated our approach in two use cases:

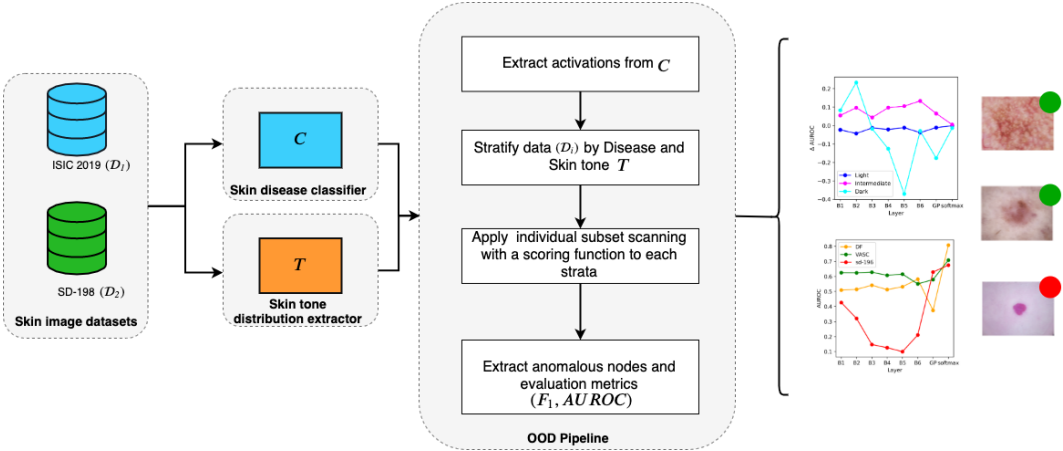
1. Different clinical settings.
2. Unknown disease classes.



Example images from unknown disease case (top) and clinical setting changes (bottom).

IBM Research | Africa

Overview: Proposed Approach



Out-of-Distribution Detection in Dermatology using Input Perturbation and Subset Scanning [KTC+21]

IBM Research | Africa

Subset Scanning for Anomalous Pattern Detection

💡 Treat Neural Networks as data-generating systems and apply anomalous pattern detection methods to activation data.

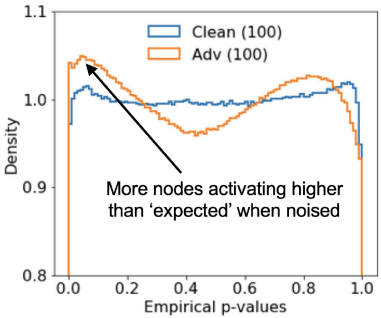
🔍 Subset Scanning efficiently searches over a large combinatorial space in order to find groups of records that differ the most from ‘expected’ behavior.

Some goodies about this type of approach:

1. We can provide detection improvements **at run time**.
2. We can **abstract from domains** and focus only on the deep representation of the input.
3. **No** need to **re-train** or have **labeled examples** of the anomalies.

IBM Research | Africa

Subset Scanning for Anomalous Pattern Detection (Cont.)



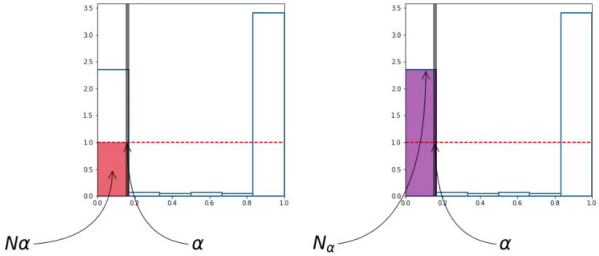
Assumption

Activations from abnormal images have a different distribution of p-values than normal samples.

p-value is the proportion of background activations (H_0), drawn from the same node for several clean samples, greater than the activation from a test sample.

IBM Research | Africa

Subset Scanning for Anomalous Pattern Detection (Cont.)



$$\max_{\alpha} \varphi(\alpha, N_\alpha, N) = \frac{N_\alpha - N\alpha}{\sqrt{N}} \quad (1)$$

Where N_α is the number of p-values less than α
N is the number of p-values
α is the level of significance
φ is a scoring function

How we score a test sample?

Scoring functions operate on a test sample in order to measure how much the p-values deviate from uniform.

IBM Research | Africa

Subset Scanning for Anomalous Pattern Detection (Cont.)

NPSS maximization

Scoring functions may be viewed as set functions that operate on subsets of nodes. We search for the highest scoring subset of nodes that maximize the deviance from uniform.

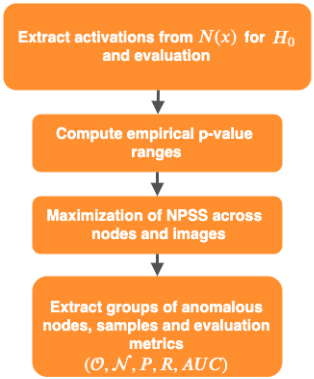
$$F(S) = \max_{\alpha} F_{\alpha}(S) = \max_{\alpha} \varphi(\alpha, N_{\alpha}(S), N(S)) \tag{2}$$

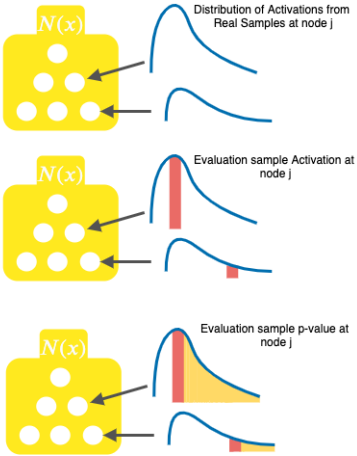
Group vs. Individual Scanning

For group-based scanning our search space is: $S = X_{\hat{S}} \times O_{\hat{S}}$, where $X_{\hat{S}}$ is a subset of test samples and $O_{\hat{S}}$ is a subset of nodes' activations.
For individual scanning we work with only one X_i.

IBM Research | Africa

Subset Scanning for Anomalous Pattern Detection (Cont.)



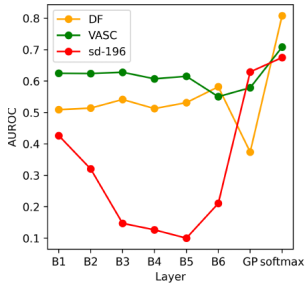


IBM Research | Africa

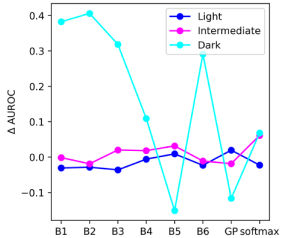
Preliminary results

Results across settings

The layers for detecting new class are different from the ones for OOD



[KTC+21]



Fairness of OOD detectors

We see varying performances for samples of Dark skin tones. This instability of performance for samples of Dark skin tones may be partially because network is trained on the ISIC 2019 dataset that **heavily lacks** samples of **Dark skin tones**.

Conclusions and future work

1. The two skin disease datasets are biased towards lighter skin with majority of the samples between ITA values $[34.5^{\circ}, 48^{\circ}]$.
2. We can provide a single OOD detection for multiple scenarios (clinical setting change or unknown disease)
3. Implementation of better representations for all skin tones extracted from clinical images.
4. Experiments around stratification of skin tone by disease.
5. How a fair distribution looks like in this case?

IBM Research | Africa

My two cents :)

1. It is crucial that the groups/institutions that develop & research technological solutions for sectors such as education, health, etc., are interdisciplinary.
2. The operational constraints and limitations of production models must be clearly and explicitly define.
3. The models to be used in production must make explicit in which context they work and their limitations, be transparent, clarify what biases were evaluated and what are the mitigation techniques used.

IBM Research | Africa

Asante, Thanks, Gracias!

IBM Research

Carnegie Mellon University Africa College of Engineering

Stanford University

@RTFMCelia @ celia.cintas@ibm.com IBM Research | Africa

References I

- **Sara Atito Ali Ahmed, Berrin Yanikoglu, Erchan Aptoula, and Ozgu Goksu, *Skin lesion classification with deep learning ensembles in isic 2019*, 2019.**
- **Joy Buolamwini and Timnit Gebru, *Gender shades: Intersectional accuracy disparities in commercial gender classification*, Proc. Conf. Fair. Account. Transp., February 2018, pp. 77–91.**
- **Marc Combalia, Ferran Hueto, Susana Puig, Josep Malvehy, and Verónica Vilaplana, *Uncertainty estimation in deep neural networks for dermoscopic image classification*, CVPR 2020, ISIC Skin Image Analysis Workshop, 2020 In Press.**

IBM Research | Africa

References II

- 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.
- Nils Gessert, Maximilian Nielsen, Mohsin Shaikh, René Werner, and Alexander Schlaefer, *Skin lesion classification using loss balancing and ensembles of multi-resolution efficientnets*.
- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B Girshick, *Mask r-cnn. corr abs/1703.06870 (2017)*, arXiv preprint arXiv:1703.06870 (2017).

IBM Research | Africa

References III

- 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).
- Hannah Kim, Girmaw Abebe Tadesse, Celia Cintas, Skyler Speakman, and Kush Varshney, *Out-of-distribution detection in dermatology using input perturbation and subset scanning*, arXiv preprint arXiv:2105.11160 (2021).
- 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 IV

- **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.**
- **Andre G. C. Pacheco, Abder-Rahman Ali, and Thomas Trappenberg, *Skin cancer detection based on deep learning and entropy to detect outlier samples*, 2019.**

IBM Research | Africa

References V

- Andre G. C. Pacheco, Chandramouli S. Sastry, Thomas Trappenberg, Sageev Oore, and Renato A. Krohling, *On out-of-distribution detection algorithms with deep neural skin cancer classifiers*, CVPR Workshops, June 2020.
- Roni Caryn Rabin, *Dermatology has a problem with skin color*, Aug 2020.
- 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.

IBM Research | Africa

References VI

- **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.**
- **Pengyi Zhang, Yunxin Zhong, and Xiaoqiong Li, *Melanet: A deep dense attention network for melanoma detection in dermoscopy images*.**

IBM Research | Africa