Automated Evaluation of Representation in Dermatology Educational Materials

Girmaw Abebe Tadesse[1*], **Hannah Kim**[1*], **Roxana Daneshjou**[3*], **Celia Cintas**[1], **Kush R. Varshney**[2], **Ademide Adelekun**[4], **Jules B. Lipoff**[4], **Ginikanwa Onyekaba**[4], **Veronica Rottemberg**[5], **James Zou**[3]
[1]IBM Research - Africa, [2]IBM Research - T. J. Watson, [3]Standford University
[4]University of Pennsylvania, [5]Memorial Sloan-Kettering Cancer Center

Abstract

Disparities in dermatological outcomes may be related to inequities in dermatological education, particularly the lack of darker skin images in educational materials used to train dermatologists and primary care physicians. To address this issue, we propose a framework to automatically assess bias in skin tone representation in academic documents of dermatology. Given a document, we apply content parsing to extract text, images, and table cells in a structured format. We then select skin images and segment non-disease regions using Mask R-CNN. Individual Typology Angle (ITA) values are computed from non-disease regions and mapped to Fitzpatrick skin indices. The proposed framework is validated with three dermatology textbooks and compared against manually annotated baselines by dermatology experts. Results show encouraging performance in estimating skin tones and discover limited representation of darker skins, i.e., only 10.7%, across these documents. We envision this technology as a tool for dermatology educators to quickly assess their materials.

Introduction

Dermatology textbooks, lecture notes, and published literature lack adequate skin color representation. Because skin disease appears visually different across skin tones, educational materials depicting diverse skin tones are required for a well-trained healthcare workforce (Louie and Wilkes 2018), (Massie et al. 2019), (Massie et al. 2020), (Lester et al. 2020), (Adelekun, Onyekaba, and Lipoff 2020).

The lack of representation in educational materials may translate to the clinical realm, where skin cancer diagnoses (e.g., melanoma, squamous cell carcinoma) are significantly delayed in patients of color, leading to increased morbidity and mortality (Cormier et al. 2006). The COVID-19 pandemic has further highlighted inequities; (Lester et al. 2020) manually annotated published photos of COVID-19 cutaneous findings and found images depicting darker skin lacking.

Unfortunately, manual skin tone annotation is not feasible for a large corpus of dermatology education materials due to its labor-intensive nature, operator visual fatigue, and intra-/inter-observer error related to category assignment for different skin tones (Louie and Wilkes 2018; Lester et al. 2020).

Related Work

Several studies have shown the significant under-representation of skin of color images in medical education (Louie and Wilkes 2018), (Massie et al. 2019), (Massie et al. 2020), (Adelekun, Onyekaba, and Lipoff 2020), (Lester et al. 2020). Unfortunately, all of these studies used manual review and labeling of skin images, thereby suggesting the immediate necessity of automated methods to do such task to facilitate similar studies. (Louie and Wilkes 2018) found over-representation of light skin tones and no images of skin cancer in darker skin across four general medicine textbooks. (Massie et al. 2020) analyzed skin tones in the New England Journal of Medicine from 1992 to 2017 and found that only 18% of the images represented non-white skin. (Adelekun, Onyekaba, and Lipoff 2020) assessed skin tones in the top general dermatology textbooks and found darker skin tones significantly under-represented.

Automatic skin tone representation assessment would significantly aid in identifying bias in educational materials. Previously, (Kinyanjui et al. 2020) used Individual Typology Angle (ITA) to approximate skin tones in curated image datasets (e.g., ISIC 2018 (Codella et al. 2019) and SD-198 (Sun et al. 2016)), and results showed that these datasets under-represent darker skin tones. We extend this work to process images *in the wild* from off-the-shelf textbooks and translate it into representations of skin tone, which could be directly used by domain experts (e.g., dermatologists).

Methods

The overview of the proposed framework is shown in Figure 1. Given an academic material ($\mathcal{A}$), e.g., a textbook, we employ a corpus conversion service (CCS) (Staar et al. 2018) to parse text, images, and table cells in a structured format. Skin images are selected ($I = \{i_0, \cdots, i_n\}$) and the non-disease regions are segmented using Mask R-CNN (He et al. 2017), generating pairs of images and masks

---

[*] Equal contributions
Corresponding email: girmaw.abebe.tadesse@ibm.com

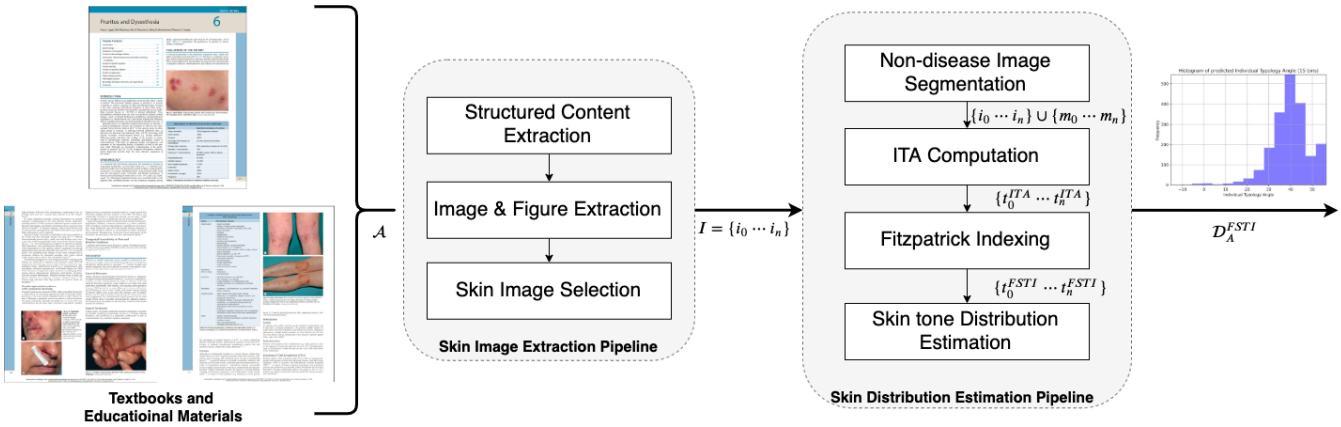


Figure 1: Overview of the proposed framework. Academic materials (e.g., in pdf format) are fed as input to our pipeline, which are parsed using a corpus conversion service (Staar et al. 2018) and document entities (e.g., images and tables) are annotated. Images are then cropped out using the annotations, among which skin images are selected. Segmentation of non-diseased skin regions is applied and skin tone estimation is computed using the luminance and yellow pixel values of these regions.

($\{i_0, \cdots, i_n\} \cup \{m_0, \cdots, m_n\}$). Then, non-disease regions ($\{c_0, \cdots, c_n\}$), where $c_j = i_j \times m_j, \forall j \in [1, n]$, are used to compute ITA values ($\{t_0^{ITA}, \cdots, t_n^{ITA}\}$) that are mapped to Fitzpatrick skin type indices (FSTI) ($\{t_0^{FSTI}, \cdots, t_n^{FSTI}\}$). Lastly, the distribution of predicted skin tone representation ($\mathcal{D}_A^{FSTI}$) is estimated. Manually annotated skin tones ($\{g_0 \cdots g_n\}$) and the distribution ($\mathcal{D}_A^g$) are used to validate the prediction.

Unlike most existing document parsing systems that use rule-based conversion algorithms (e.g., Xpdf[1] and Tabula[2]), CCS is a cloud-based platform that uses a machine learning-based approach. CCS follows a programmable parsing approach that returns document entities (e.g., images and tables) parsed into a structured format (e.g., JavaScript Object Notation - JSON) without preserving the layout. State-of-the-art segmentation techniques, such as Faster-R-CNN (Ren et al. 2015) and the YOLOv2 (Redmon and Farhadi 2017), are integrated in CCS to segment those entities and provide the coordinates in the output JSON file. For this work, only images are extracted and non-skin images (e.g., charts and diagrams) are manually discarded.

The skin tone is estimated from the non-diseased part of the skin images (Kinyanjui et al. 2020). Thus, segmentation of the skin lesion from the non-diseased part is performed using a Mask R-CNN model that achieved state-of-the-art segmentation performance across different domains, e.g., 2018 ISIC Challenge (Codella et al. 2019). The segmentation results in non-disease skin regions ($\{c_0, \cdots, c_n\}$).

CIELAB color space is known for its robustness across imaging devices, and hence the conversion of non-disease regions to CIELAB yields $\{\hat{c}_0, \cdots, \hat{c}_n\}$. Then the ITA value, t_j^{ITA}, for each transformed image, $\hat{c}_j$, is computed as

$$t_j^{ITA} = \arctan\left(\frac{L_\mu^j - 50}{b_\mu^j}\right) \times \left(\frac{180^\circ}{\pi}\right), \forall j \in [1, n] \qquad (1)$$

[1]https://www.xpdfreader.com
[2]http://tabula.technology/

ITA Range	FSTI
$ITA > 41^\circ$	I
$34.5^\circ < ITA \leq 41^\circ$	II
$28^\circ < ITA \leq 34.5^\circ$	III
$19^\circ < ITA \leq 28^\circ$	IV
$10^\circ < ITA \leq 19^\circ$	V
$ITA \leq 10^\circ$	VI

Table 1: Summary of skin tone categorization from Individual Typology Angle (ITA) values to Fitzpatrick Skin Type Indices (FSTI). The lower ITA values (higher FSTI indices - e.g., V and VI) represent darker skin tones.

Where L_μ^j and b_μ^j are the mean luminance and yellow pixel values of $\hat{c}_j$. To avoid outliers, we only consider those non-diseased skin pixels within one standard deviation from L_μ^j and b_μ^j. The ITA values are then mapped to Fitzpatrick Skin Type Indices (FSTI) as shown in Table 1.

Experimental Setup

This section describes the dataset and evaluation metrics used to validate the framework proposed to automatically characterize dermatology educational materials.

Datasets

We evaluate our framework using images extracted from three dermatology textbooks: Bolognia 4e (Bolognia, Schaffer, and Cerroni 2018), containing 160 chapters and 2861 images; Fitzpatrick Color Atlas 8e (Wolff et al. 2017), containing 35 Chapters and 868 annotated images; and Fitzpatrick Dermatology in General Med 9e (Kang et al. 2019), containing 217 chapters and 1528 annotated images. See Figure 2 for example images extracted from each of these textbooks.

The images were labeled by two trained medical stu-

Bolognia 4e Atlas 8e General Med 9e

Figure 2: Example images extracted from three dermatology textbooks: Dermatology Bolognia 4e ((Bolognia, Schaffer, and Cerroni 2018)), Fitzpatrick Color Atlas 8e ((Wolff et al. 2017)), and Fitzpatrick Dermatology in General Med 9e ((Kang et al. 2019).

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

Table 2: Percentage of images with dark skins in the ground truth (GT) (Adelekun, Onyekaba, and Lipoff 2020) and those predicted by the proposed approach for each of the three textbooks, Dermatology Bolognia 4e (Bolognia, Schaffer, and Cerroni 2018), Fitzpatrick Color Atlas 8e (Wolff et al. 2017), and Fitzpatrick Dermatology in General Med 9 (Kang et al. 2019).

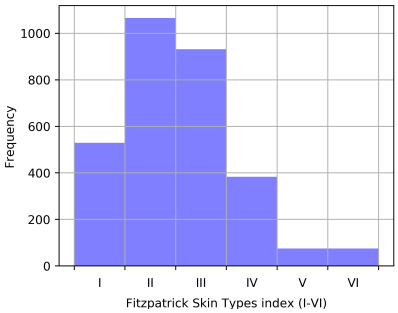
dents under the supervision of a dermatologist as either dark or light using previously described procedures (Adelekun, Onyekaba, and Lipoff 2020). A subset of 93 images from Bolognia 4e were annotated into the six FSTI labels by a single dermatologist.

Evaluation metrics

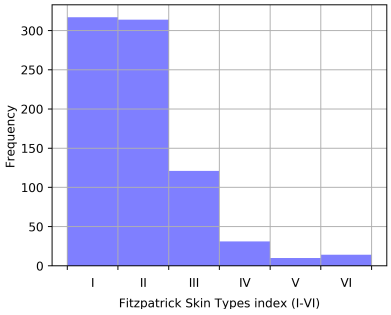
We employ the following error rates to evaluate the skin tone prediction performance: sum of absolute difference (SAD), mean absolute error (MAE), mean squared error (MSE), and root mean squared error (rMSE). These are computed only on the subset of Bolognia 4e with the six FSTI labels, treating each of the indices as numeric. The confusion matrix is also computed to show the misclassification between skin tones. The distribution of skin tones in the prediction and ground truth are compared to provide graphical evaluation.

Results and Discussion

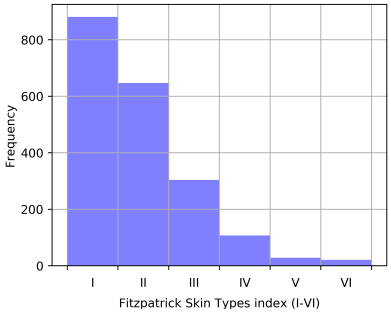
The predicted skin tone distributions are shown in Figure 3. A lack of darker skin tone representation (indices V and VI) is discovered consistently across all textbooks. Table 2 shows the competitive performance of the proposed approach in predicting the percentage of darker skin tones in the three textbooks studied, compared with the ground truth annotation by domain experts (Adelekun, Onyekaba, and Lipoff 2020). The absolute differences of percentages of dark skins between the predicted and ground truth are 5.5% in Bolognia 4e (Bolognia, Schaffer, and Cerroni 2018), 2.74% in Atlas 8e (Wolff et al. 2017) and 3.94% in General Med 9e ((Kang et al. 2019).



(a) Bolognia 4e



(b) Atlas 8e



(c) General Med 9e

Figure 3: Distribution of predicted Fitzpatrick skin tone indices (FSTI) across the three textbooks studies, where under-representation of darker skin tones, i.e., FTSI indices of V and VI, is consistently observed across these textbooks.

The evaluation metrics for the subset of Bolognia 4e (Bolognia, Schaffer, and Cerroni 2018) annotated with six FTSI labels are shown in Table 3 and achieve encouraging performance. The confusion matrix in Figure 4 shows plausible misclassifications between adjacent skin tones, i.e., the confusion matrix shows encouraging detection performance of higher-level skin tones (white vs. dark).

We attribute much of the error to the segmentation step. The segmentation of non-diseased skin was challenging for darker skin tone with a tendency of segmenting the whole image as diseased in our validation (see Figure 5). This is partly due to the lack of proportional samples with dark skin tones during the training of the segmentation step. Thus, we aim to address this issue by incorporating more manually-segmented images with dark skin tones for re-training.

SAD	MAE	MSE	rMSE
61.0	0.6559	0.8710	0.9333

Table 3: Evaluation metrics of skin tone predictions compared to manually annotated ground-truth. Metrics are SAD: sum of absolute difference; MAE: mean absolute error; MSE: mean squared error; and rMSE: root mean squared error.

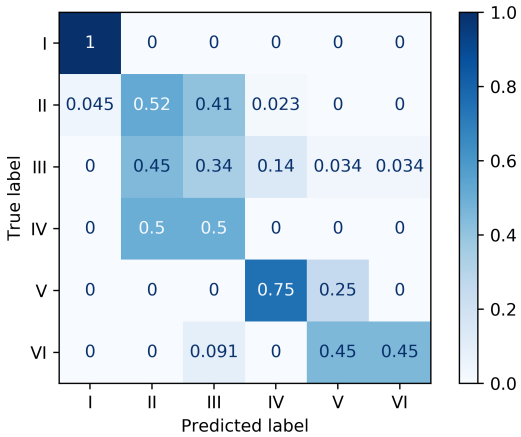


Figure 4: Normalized confusion matrix of the predicted FSTI in Bolognia 4e (Bolognia, Schaffer, and Cerroni 2018). While encouraging performance is achieved to distinguish dark vs. white skin tones, expected misclassifications between adjacent FSTI indices are observed.

Conclusion

We proposed a framework that automatically assesses bias in representation of darker skin tones in dermatology documents. The framework is an end-to-end pipeline of ingesting academic documents, extracting image contents and automatically quantifying representation across groups. To this end, Mask R-CNN is employed to segment non-diseased skin regions upon which Individual Typology Angle (ITA) values are extracted from the luminance and yellow pixel values after CIELAB color space conversion. These ITA values are mapped to Fitzpatrick skin tone indices.

We validated the framework using three dermatology textbooks (Bolognia 4e, Atlas 8e and General Med 9e), using ground truth annotation performed by domain experts. Prediction performance is evaluated using multiple performance metrics, such as distribution of indices, confusion matrices. The results show encouraging performance of our proposed approach that aims to predict skin tones of images directly form academic materials. We also observed, relatively, poor segmentation results of darker skin tones, which could be addressed using more representative training samples.

Most importantly, the tool described here could be automatically implemented prospectively to evaluate textbooks, lectures, and journal articles for fair representation prior to publication; thereby facilitates the development of trustworthy AI systems in field of dermatology. Moreover, the proposed approach is scalable and could be applied to evaluate academic materials beyond dermatology and image modality.

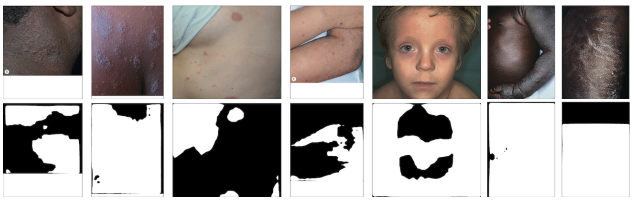


Figure 5: Top row: skin images, and bottom row: segmentation of non-diseased regions (black). Darker skin tones tend to produce poor segmentation results, e.g., the last two columns, partly due to the lack of enough training instances with similar skin tones.

References

Adelekun, A.; Onyekaba, G.; and Lipoff, J. B. 2020. Skin color in dermatology textbooks: An updated evaluation and analysis. *JAAD*.

Bolognia, J. L.; Schaffer, J. V.; and Cerroni, L. 2018. Dermatology: 2-volume set , 4th edition.

Codella, N. C. F.; Rotemberg, V.; Tschandl, P.; Celebi, M. E.; Dusza, S. W.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M. A.; Kittler, H.; and Halpern, A. 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *CoRR* abs/1902.03368.

Cormier, J. N.; Xing, Y.; Ding, M.; Lee, J. E.; Mansfield, P. F.; Gershenwald, J. E.; Ross, M. I.; and Du, X. L. 2006. Ethnic Differences Among Patients With Cutaneous Melanoma. *Archives of Internal Medicine* 166(17):1907–1914.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. B. 2017. Mask R-CNN. *CoRR* abs/1703.06870.

Kang, S.; Amagai, M.; Bruckner, A. L.; Enk, A. H.; Margolis, D. J.; McMichael, A. J.; and Orringer, J. S. 2019. Fitzpatrick's dermatology, 9e.

Kinyanjui, N. M.; Odonga, T.; Cintas, C.; Codella, N. C. F.; Panda, R.; Sattigeri, P.; and Varshney, K. R. 2020. Fairness of classifiers across skin tones in dermatology. In *Medical Image Computing and Computer Assisted Intervention*.

Lester, J.; Jia, J.; Zhang, L.; Okoye, G.; and Linos, E. 2020. Absence of skin of colour images in publications of covid-19 skin manifestations. *British Journal of Dermatology*.

Louie, P., and Wilkes, R. 2018. Representations of race and skin tone in medical textbook imagery. *Social Science & Medicine* 202:38–42.

Massie, J. P.; Cho, D. Y.; Kneib, C. J.; Burns, J. R.; Crowe, C. S.; Lane, M.; Shakir, A.; Sobol, D. L.; Sabin, J.; Sousa, J. D.; et al. 2019. Patient representation in medical literature: Are we appropriately depicting diversity? *Plastic and Reconstructive Surgery Global Open* 7(12).

Massie, J. P.; Cho, D. Y.; Kneib, C. J.; Sousa, J. D.; Morrison, S. D.; and Friedrich, J. B. 2020. A picture of modern medicine: Race and visual representation in medical literature. *Journal of the National Medical Association*.

Redmon, J., and Farhadi, A. 2017. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7263–7271.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. Advances in Neural Information Processing Systems*, 91–99.

Staar, P. W.; Dolfi, M.; Auer, C.; and Bekas, C. 2018. Corpus conversion service: A machine learning platform to ingest documents at scale. In *Proc. of International Conference on Knowledge Discovery & Data Mining*, 774–782.

Sun, X.; Yang, J.; Sun, M.; and Wang, K. 2016. A benchmark for automatic visual classification of clinical skin disease images. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *Computer Vision – ECCV 2016*, 206–222. Cham: Springer International Publishing.

Wolff, K.; Johnson, R. A.; Saavedra, A. P.; and Roh, E. K. 2017. Fitzpatrick's color atlas and synopsis of clinical dermatology, 8e.