HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 1 (2023)

Unintended Bias in Artificial Intelligence Driven Diagnosis of Melanoma: A Systematic Review

Kemka Ihemelandu[1] and Chris Albanese[#]

[1]McDonogh School, Owings Mills, MD, USA
[#]Advisor

ABSTRACT

Melanoma remains a public health crisis, with incidence rates increasing rapidly in the past decades. Improving diagnostic accuracy to decrease misdiagnosis using Artificial intelligence (AI) continues to be documented. Unfortunately, unintended racially biased outcomes a product of lack of diversity in the dataset used, with a noted class imbalance favoring lighter vs. darker skin tone have increasingly been recognized as a problem. Resulting in noted limitations of the accuracy of the Convolutional neural network (CNN) models. CNN models are prone to biased output due to biases in the dataset used to train them. Although the incidence of melanoma is lower in patients with darker skin tone, it is associated with a worse prognosis than in Caucasians, underscoring the need for accurate early diagnosis in these patients. Our objective in this systematic review was to assess to what degree race/ethnicity, specifically Black/ African American patient cohort were included in training datasets used in generating machine learning algorithms for automated melanoma diagnosis. Our review documents the fact that there is a remarkable lack of and inconsistent reporting of patient demographics especially race/ethnicity with notable under-representation of patients of color, highlighting a currently unmet critical need of lack of diversity in the publicly available skin image datasets. These publicly available skin image datasets, are an inherently unbalanced unintentionally biased datasets from which AI models created for the diagnosis of melanoma skin cancer have restricted applicability to real life clinical scenarios and limited population representation preventing generalizability to the community as a whole.

Introduction

Melanoma remains a public health crisis, with incidence rates increasing rapidly in the past decades. Improving diagnostic accuracy to decrease misdiagnosis using Artificial intelligence (AI) continues to be documented. Unfortunately, unintended racially biased outcomes a product of lack of diversity in the dataset used, with a noted class imbalance favoring lighter vs. darker skin tone have increasingly been recognized as a problem. Resulting in noted limitations of the accuracy of the Convolutional neural network (CNN) models. CNN models are prone to biased output due to biases in the dataset used to train them. The American Cancer Society estimates that skin cancer is the most predominant of all cancers, with melanoma shown to account for ~ 1% of all cancers of the skin but leading to a greater proportion of skin cancer related deaths[1]. Melanoma as defined by the National Cancer Institute is a "form of cancer that arises in melanocytes or a mole"[2]. Approximately, 99,780 new cases of melanoma are anticipated to be diagnosed in the USA for the year 2022, with ~ 7,650 associated deaths[3]. According to the American Cancer Society Cancer Statistics Center the incidence rates for melanoma of the skin, by race and ethnicity for the time period 2014-2018 was 29.4% for non-Hispanic white, 9.3% for American Indian and Alaska Native, 4.7% for Hispanics, 1.3% for Asian and Pacific Islanders and 1%- for Non-Hispanic black. Mahendraraj et al[4] in a Population-Based Clinical Outcomes Study Involving 1106 African–American Patients from the Surveillance, Epidemiology, and End Result (SEER) Database (1988–2011), found that African–American Patients presented with deeper tumors, more advanced stage of disease, and

higher rates of ulceration and lymph node positivity than Caucasians. They also noted that cancer-specific mortality was significantly higher, while 5-year cancer-specific survival was significantly lower among African–American Patients. Although the incidence of melanoma is lower in patients with darker skin tone, it is associated with a worse prognosis than in Caucasians, underscoring the need for accurate early diagnosis in these patients. Melanoma is currently regarded as a public health crisis, with noted rapid increase in the incidence of melanoma cases over the past decades believed to be driven by ultra violet radiation and global warming. Artificial intelligence (AI) has improved the accuracy of melanoma diagnosis, However, there are noted limitations with the accuracy of the models used to test. One such noted limitation is the poor efficiency of current AI models amongst darker skin types. The preponderance of currently used AI algorithms are trained or validated on the International Skin Imaging Collaboration (ISIC) dataset, a repository of histopathological validated dermoscopic images of cutaneous malignancies but plagued by a paucity of images representative of diverse racial skin tones[56]. Daneshjou et al in a study looking at the "Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review", identified three issues with the datasets being used to create and validate clinical AI algorithms for skin disease that should be addressed before clinical translation: (1) sparsity of data set characterization and lack of transparency, (2) nonstandard and unverified disease labels, and (3) inability to fully assess patient diversity used for algorithm development and testing[7]. Kaushal et al[8] in their study looking at Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms, concluded that in clinical applications of deep learning across multiple disciplines, algorithms trained on US patient data were disproportionately trained on cohorts from a few states, that may have economic, educational, social, behavioral, ethnic, and cultural features that are not representative of the entire nation; as such algorithms trained primarily on patient data from these states may generalize poorly, a well-known articulated risk factor when implementing AI algorithms in new geographies[9]. Our objective in this systematic review was to assess to what degree race/ethnicity, specifically Black/ African American patient cohort were included in training datasets used in generating machine learning algorithms for automated melanoma diagnosis.

Methods

A systematic review was conducted in accordance with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)[10]. Literature review was carried out using PubMed, for peer-reviewed original article in English using artificial intelligence to automate diagnosis of melanoma between 1981-2022. All potential terms related to automated diagnosis of melanoma were used including melanomas, bias, racial disparity, artificial intelligence, machine Learning, neural networks, convolutional.

The PubMed database was queried by combining terminology related to these terms, with MeSH terms and key words including ("Nevi and Melanomas"[Mesh] OR "Skin Neoplasms"[Mesh] OR "Bias, Racial"[Mesh] OR "disparity, bias"[Mesh:NoExp] OR melanoma*[tiab] OR melanocyt*[tiab] OR melanotic[tiab] OR nevus[tiab] OR nevi[tiab] OR skin cancer*[tiab] OR skin lesion*[tiab] OR skin neoplasm*[tiab] OR melanoma*[tiab]) AND ("Artificial Intelligence"[Mesh:NoExp] OR "Machine Learning"[Mesh] OR "Neural Networks, Convolutional "[Mesh] OR artificial intelligence[tiab] OR augumentation [tiab] OR transfer learning[tiab] OR deep learning[tiab] OR hierarchical learning[tiab] OR machine learning[tiab] OR neural network*[tiab] OR "dataset, database"[Mesh]).

All peer-reviewed journal studies utilizing a machine learning process for melanoma diagnosis using dermoscopic images were included. Studies detailing dermoscopic image datasets for skin cancer were included. Studies that did not diagnose melanoma by automation via AI were excluded. All manuscript titles and abstract were reviewed for potentially eligible studies, which were then examined for source of the studies database, listing of patient race, ethnicity. Only studies that used dermoscopic or clinical images were included.

Our initial search was performed in June of 2022 and was repeated in December 2022. Two independent reviewers CI and KI independently extracted data. Study title and abstract were used for pre-screening and

then the full text review was carried out for final inclusion, any variance regarding a study's inclusion was resolved by discussion as recommended by the PRISMA guideline. Automated diagnosis via machine learning was defined as an automated algorithm that improves automatically through training on datasets[11]. Appropriate information was abstracted from included studies, specifically identification of racial under-representation in model datasets used.

Results

Of our 1349 identified records, following removal of non-English articles we had 1326 articles. Following title and abstract review 1225 articles were excluded. 101 full text articles were assessed for eligibility (Fig 1). Thirty (29.7%) full-text studies were included in the final analysis. Disclosure of race and ethnicity was found in only a handful of the studies which initially met inclusion criteria. No study specifically detailed Black or African American patients. Study datasets for algorithm training included cohorts from mainly Australia, USA, Europe and East Asian countries[12 13 14 15 16 17 18]. In the studies where ethnicity/race was detailed it was mostly for a single ethnic group that made up the cohort of patients in the database.

Most of the studies utilized patient images which were not individually labelled for ethnicity, skin tone, or nationality. The International Skin Imaging Collaboration dataset appeared to be the most used dataset, with dermoscopic skin images predominantly derived from European or Australian populations[19]. Kinyanjui et al[20] in their study Estimating Skin Tone and Effects on Classification Performance in Dermatology Datasets, found that the majority of the data in the two datasets they used (the ISIC 2018 Challenge dataset, a collection of dermoscopic images of skin lesions for the detection of skin cancer, and the SD-198 dataset, a collection of clinical images capturing a wide variety of skin diseases) have an individual typology angle (ITA) values between 34.5◦ and 48◦, which are associated with lighter skin, and is consistent with under-representation of darker skinned populations in these datasets

Phillips et al[21], in their study looking at Assessment of Accuracy of an Artificial Intelligence Algorithm to Detect Melanoma in Images of Skin Lesions, had a study population of 514 patients including 279 women (55.7%) and 484 white patients (96.8%), a total of 1550 images of skin lesions were included in the analysis. The study concludes that the algorithm demonstrated an ability to identify melanoma from dermoscopic images of selected lesions with an accuracy similar to that of specialists.

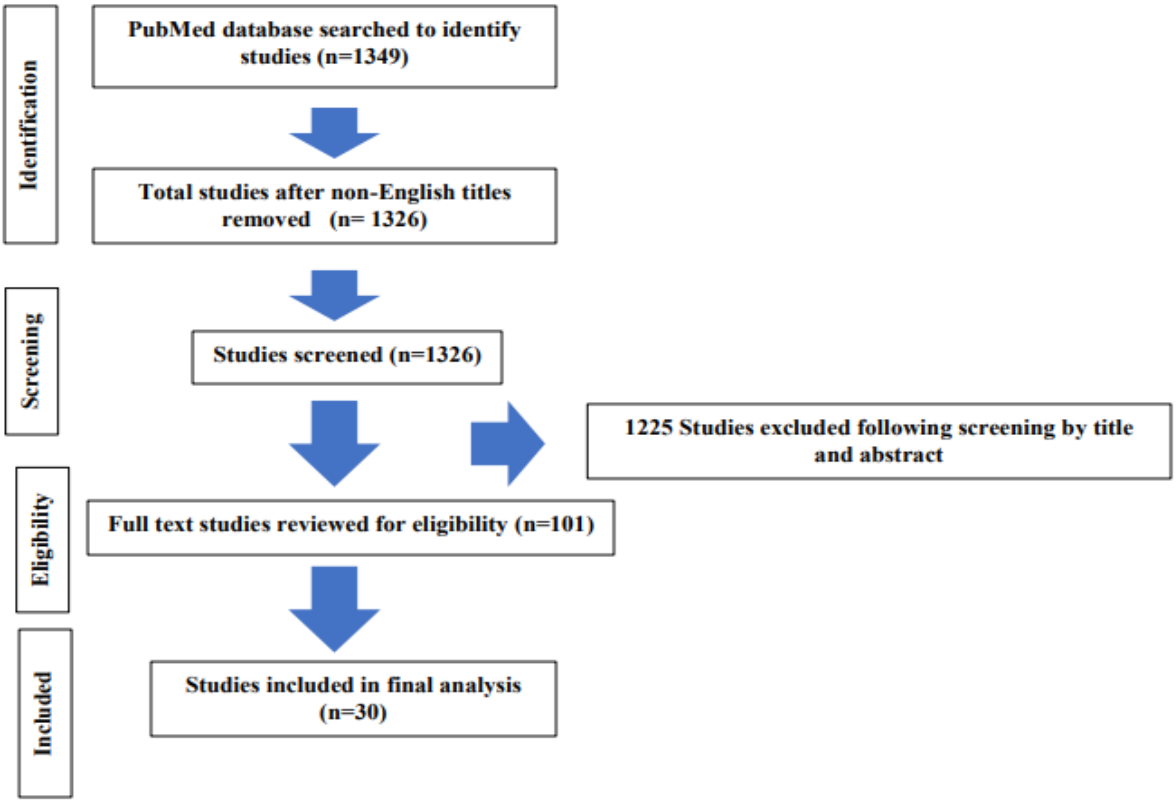

Figure 1. PRISMA flow diagram for study selection

Upon review of our full-text papers we were able to identify a number of datasets including ISIC archive[22], BCN20000[23], HAM10000, Asan dataset[24], Hallym dataset[24], Dermofit image library[25]. As detailed in one of our full-text studies included in our final analysis by Wen et al[18], patient ethnicity data was available for only 1415 images (1·3% of all images), of a total of 106 950 skin lesion images contained in the open access dataset analyzed for the study. Highlighting the staggering inherent unintended bias in the datasets being used for algorithm generation for AI models.

Discussion

Our systematic analysis showed that race/ethnicity was rarely detailed separately in the reviewed machine learning models detailed in the papers, and as such the reliability of the performance of the algorithms on patients of color is largely unknown. Most of the skin lesion datasets are populated by lighter skinned patients, with dark or brown skinned patients making up a very small cohort. This maybe reflective of the incidence rates among the different racial groups, unfortunately people of color even though they may have a lower incidence rate, are frequently diagnosed at a more advanced stage[26]. Given the noted imbalances in the dataset being used for dataset training, automated diagnosis of melanoma based on deep learning algorithms validated for the automated diagnosis of melanoma in lighter skinned patients is prone to a higher degree of misdiagnosis amongst patients of color[27]. Han et al[28], in their study stated that "to improve the performance of convolutional neural network, additional images with a broader range of ages and ethnicities should be collected". This recommendation was borne out of observations in their study where a convolutional neural network using images from the training section of the Asan dataset, MED-NODE dataset, and atlas site images (19,398 images in total) was fine-tuned, with the trained model validated with the testing section of the Asan, Hallym and Edinburgh datasets. With the Asan dataset they reported an area under the curve for the diagnosis of melanoma as 0.96, this

dropped to 0.88 with the Edinburgh dataset. The Asan datasets represents skin lesions from a predominantly Asian population for which the study reported an accuracy of ~ 80% when their trained deep learning algorithm using the Asan training dataset was used on the Asian testing set, this number dropped drastically to 56% on the Dermofit dataset which is populated by skin lesions from Caucasian patients. It is obvious that learned features of deep learning algorithms are not interchangeable amongst patients of a different ethnicity or race. An eloquent study by Wen et al[18] details that in their review of publicly available dermatology datasets for skin cancer, that 106 950 images from 21 open access datasets were identified, along with 17 open access atlases that are freely accessible to researchers and the public. They conclude that although this represents a rich data resource for innovation, a lack of transparency in metadata reporting for clinically essential characteristics (such as ethnicity) limits the clinical utility of these images alone. Guo et al[29] in their study Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection—A scoping review, reported that only one study definitively included Black or African American individuals or American Indian or Alaska Native individuals, respectively, and only 2 studies definitively included Hispanic patients. In their study images were predominantly of lighter skin types in 3 studies with no other specifics provided. They conclude that their findings underscore the need for greater transparency and full disclosure of phototypes or race/ethnicity of subjects in dermatologic ML studies. In their study when demographics were reported, darker skin types were starkly underrepresented.

It is obvious that race/ethnicity, specifically Black/ African American patient cohorts are under-represented in training datasets used in generating machine learning algorithms for automated melanoma diagnosis. In other for machine learning algorithms to perform efficiently amongst diverse patient groups, they must be generated using unbiased datasets, therefore a concerted effort is required to ensure that all the various patient demographics are represented, especially skin of color. The American Academy of Dermatology in their Diversity, Equity and Inclusion statement of intent, state that "The American Academy of Dermatology recognizes diversity, equity, and inclusion as a pathway to excellence in delivering dermatologic care, the Academy has committed to increasing diversity, equity, and inclusion in the dermatology specialty as a priority." Efforts such as this are required if we are to see ethical and judicious integration of AI into skin cancer care in the broader population. Narla et al[30], in their study on Racial disparities in dermatology, conclude that "We must continue to work toward increasing the diversity of the dermatology workforce, increasing diversity education of current dermatologists in practice, including a diverse range of skin tones in images used in dermatology training, and teaching trainees how diseases may present differently in different skin tones".

Conclusion

Our review documents the fact that there is a remarkable lack of and inconsistent reporting of patient demographics especially race/ethnicity with notable under-representation of patients of color, highlighting a currently unmet critical need of lack of diversity in the publicly available skin image datasets. These publicly available skin image datasets, are an inherently unbalanced unintentionally biased datasets from which AI models created for the diagnosis of melanoma skin cancer have restricted applicability to real life clinical scenarios and limited population representation preventing generalizability to the community as a whole. A global wide concerted effort to broaden image data collection to include a representative percentage of patients of color will lend itself to supporting the generation of more inclusive AI models for melanoma diagnosis in the future.

Limitations

Our study was restricted to assessing investigator-reported demographics, actual images used were not assessed. Secondly, our search was limited to PubMed, and even though we ran our search at two different time points, recent publications that were published but not yet PubMed indexed might have been missed.

Acknowledgments

I would first like to thank my family for believing in me and being my biggest supporters. Your endless encouragement made this possible. I would like to acknowledge Dr. Parsa Akbar for guiding me through the research process. A special thanks to Mr. Thompson at McDonogh School for his continuous support and mentorship

References

[1] https://www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html

[2] https://www.cancer.gov/publications/dictionaries/cancer-terms/def/melanoma

[3] American Cancer Society. Facts & Figures 2022. American Cancer Society. Atlanta, Ga. 2022.

[4] Mahendraraj, K., Sidhu, K., Lau, C. S. M., McRoy, G. J., Chamberlain, R. S., & Smith, F. O. (2017). 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(15), e6258. https://doi.org/10.1097/MD.0000000000006258

[5] Codella N, Rotemberg V, Tschandl P, et al. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). arXiv; 2018

[6] Kinyanjui NM, Odonga T, Cintas C, et al. Fairness of Classifiers Across Skin Tones in Dermatology. Medical Image Computing and Computer Assisted Intervention -- MICCAI 2020.

[7] Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V., & Zou, J. (2021). Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review. JAMA dermatology, 157(11), 1362–1369. https://doi.org/10.1001/jamadermatol.2021.3129

[8] Kaushal, A., Altman, R., & Langlotz, C. (2020). Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. JAMA, 324(12), 1212–1213. https://doi.org/10.1001/jama.2020.12067

[9] Wang, X., Liang, G., Zhang, Y., Blanton, H., Bessinger, Z., & Jacobs, N. (2020). Inconsistent Performance of Deep Learning Models on Mammogram Classification. Journal of the American College of Radiology : JACR, 17(6), 796–803. https://doi.org/10.1016/j.jacr.2020.01.006

[10] Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ (Clinical research ed.), 339, b2700. https://doi.org/10.1136/bmj.b2700

[11] Topol E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

[12] Tomatis, S., Bono, A., Bartoli, C., Carrara, M., Lualdi, M., Tragni, G., & Marchesini, R. (2003). Automated melanoma detection: multispectral imaging and neural network approach for classification. Medical physics, 30(2), 212–221. https://doi.org/10.1118/1.1538230

[13] Fujisawa, Y., Otomo, Y., Ogata, Y., Nakamura, Y., Fujita, R., Ishitsuka, Y., Watanabe, R., Okiyama, N., Ohara, K., & Fujimoto, M. (2019). Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. The British journal of dermatology, 180(2), 373–381. https://doi.org/10.1111/bjd.16924

[14] Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H., & Binder, M. (2001). A comparison of machine learning methods for the diagnosis of pigmented skin lesions. Journal of biomedical informatics, 34(1), 28–36. https://doi.org/10.1006/jbin.2001.1004

[15] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data, 5, 180161. https://doi.org/10.1038/sdata.2018.161

[16] Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A. B. H., Thomas, L., Enk, A., Uhlmann, L., Reader study level-I and level-II Groups, Alt, C., Arenbergerova, M., Bakos, R., Baltzer, A., Bertlich, I., Blum, A., Bokor-Billmann, T., Bowling, J., … Zalaudek, I. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of oncology : official journal of the European Society for Medical Oncology, 29(8), 1836–1842. https://doi.org/10.1093/annonc/mdy166

[17] Marchetti, M. A., Codella, N. C. F., Dusza, S. W., Gutman, D. A., Helba, B., Kalloo, A., Mishra, N., Carrera, C., Celebi, M. E., DeFazio, J. L., Jaimes, N., Marghoob, A. A., Quigley, E., Scope, A., Yélamos, O., Halpern, A. C., & International Skin Imaging Collaboration (2018). Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. Journal of the American Academy of Dermatology, 78(2), 270–277.e1. https://doi.org/10.1016/j.jaad.2017.08.016

[18] Wen, D., Khan, S. M., Ji Xu, A., Ibrahim, H., Smith, L., Caballero, J., Zepeda, L., de Blas Perez, C., Denniston, A. K., Liu, X., & Matin, R. N. (2022). Characteristics of publicly available skin cancer image datasets: a systematic review. The Lancet. Digital health, 4(1), e64–e74. https://doi.org/10.1016/S2589-7500(21)00252-1

[19] The International Skin Imaging Collaboration. https://www.isic-archive.com/

[20] Kinyanjui NM, Odonga T, Cintas C, et al. Estimating Skin Tone and Effects on Classification Performance in Dermatology Datasets. 2019. https://arxiv.org/abs/1910.13268

[21] Phillips, M., Marsden, H., Jaffe, W., Matin, R. N., Wali, G. N., Greenhalgh, J., McGrath, E., James, R., Ladoyanni, E., Bewley, A., Argenziano, G., & Palamaras, I. (2019). Assessment of Accuracy of an Artificial

Intelligence Algorithm to Detect Melanoma in Images of Skin Lesions. JAMA network open, 2(10), e1913436. https://doi.org/10.1001/jamanetworkopen.2019.13436

[22] Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Langer, S., Lioprys, K., Malvehy, J., Musthaq, S., Nanda, J., Reiter, O., … Soyer, H. P. (2021). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Scientific data, 8(1), 34. https://doi.org/10.1038/s41597-021-00815-z

[23] Combalia, M., Codella, N., Rotemberg, V., Carrera, C., Dusza, S., Gutman, D., Helba, B., Kittler, H., Kurtansky, N. R., Liopyris, K., Marchetti, M. A., Podlipnik, S., Puig, S., Rinner, C., Tschandl, P., Weber, J., Halpern, A., & Malvehy, J. (2022). Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge. The Lancet. Digital health, 4(5), e330–e339. https://doi.org/10.1016/S2589-7500(22)00021-8

[24] Han, S. S., Kim, M. S., Lim, W., Park, G. H., Park, I., & Chang, S. E. (2018). Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. The Journal of investigative dermatology, 138(7), 1529–1538. https://doi.org/10.1016/j.jid.2018.01.028

[25] Edinburgh Innovations, Dermofit image library. https://licensing.edinburgh-innovations.ed.ac.uk/i/software/dermofit-image-library.html

[26] Gloster, H. M., Jr, & Neal, K. (2006). Skin cancer in skin of color. Journal of the American Academy of Dermatology, 55(5), 741–764. https://doi.org/10.1016/j.jaad.2005.08.063

[27] G. Marcus, E. Davis, A.I. Rebooting, Building Artificial Intelligence We Can Trust, 2019. Pantheon.

[28] Han, S. S., Kim, M. S., Lim, W., Park, G. H., Park, I., & Chang, S. E. (2018). Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. The Journal of investigative dermatology, 138(7), 1529–1538. https://doi.org/10.1016/j.jid.2018.01.028

[29] Guo, L. N., Lee, M. S., Kassamali, B., Mita, C., & Nambudiri, V. E. (2022). Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection-A scoping review. Journal of the American Academy of Dermatology, 87(1), 157–159. https://doi.org/10.1016/j.jaad.2021.06.884

[30] Narla, S., Heath, C. R., Alexis, A., & Silverberg, J. I. (2022). Racial disparities in dermatology. Archives of dermatological research, 1–9. Advance online publication. https://doi.org/10.1007/s00403-022-02507-z