Racial Disparities Persist Beyond Data Representation in Medical Imaging — even Predictive Uncertainty Fails to Capture them

Author Name1[*1,2] ABC@SAMPLE.EDU

Editors: Under Review for MIDL 2025

Abstract

Balanced training sets are often promoted to mitigate racial performance disparities of Deep Learning (DL) models in medical imaging. However, **our preliminary findings on two medical imaging datasets show that while racial training set representation affects model performance, there is more at play, as large racial disparities remain regardless of training set composition.** Moreover, predictive uncertainty is shown to be completely insensitive to these performance disparities. From this, we derive a series of open problems for safe and fair image-guided diagnostics.

Keywords: Algorithmic Fairness, Bias, Uncertainty, Disparities, Representation

1. Introduction

Reliability of machine learning tools is an active area of research (Puyol-Antón et al., 2021; Hussain et al., 2022; Ricci Lara et al., 2022; Jiménez-Sánchez et al., 2023; Ferrante and Echeveste; Lekadir et al., 2025). However, inherent bias in training data is often overlooked, and classification outcomes tend to reinforce biases, replicating current and previous socioeconomic inequalities and produce errors that correlate with demographic variables or even potentially hidden attributes not explicitly available in collected data (Ferrara, 2023; Larrazabal et al., 2020). If these systems are deployed in real-world settings, they could produce inadequate outcomes. Mitigation strategies have been proposed for unfair classifiers (Zong et al., 2022). However, a lack of understanding of the deeper causes of bias will limit performance and could perpetuate or even introduce new unwanted bias (Petersen et al., 2023). This paper investigates how racial composition in training data impacts performance and uncertainty utilising two medical imaging datasets, assessing whether uncertainty and accuracy can reveal any disparities.

2. Methods and Experimental Design

We monitor racial performance disparities using a standard ResNet backbone across two different datasets:

- **RETINAL Dataset**: Consists of 2D retinal nerve fiber layer (RNFL) thickness images (200 × 200 pixels) curated with equal representations of Asian, White, and Black subjects across training (2,100), validation (300), and test (900) groups (Luo et al., 2024). The predictive task was a binary Glaucoma diagnosis classification.

[*] Contributed equally

© 2025 CC-BY 4.0, A. Name1.

- **PASSION Skin Imaging Dataset**: An unbalanced dermatological dataset curated to represent racial groups with darker skin tones (Gottfrois et al., 2024), containing 4,901 images (224 × 224 pixels) from 1,653 patients across five Fitzpatrick skin type (FST) phototypes. The predictive task was a multiclass classification to diagnose eczema, fungal infections, scabies, or other skin diseases.

Training Data Composition: A controlled sampling technique created multiple training subsets to vary the proportions of subjects either based on racial demographics or skin tone, analogous to (Larrazabal et al., 2020). Each model was retrained independently with 10 random seeds for the RETINAL dataset and 5 for the skin dataset, with fixed architecture and parameters to ensure we could isolate the effects of each group's training configuration.

A series of experiments to examined performance for different racial compositions in the training data, using similar hyperparameters as the original papers: Models were trained for 50 and 80 epochs with learning rates of 5×10^{-5} and 1×10^{-4} and batch sizes of 8 and 64 for the RETINAL and PASSION datasets, respectively. Representation of Asian, White, and Black subjects were varied for RETINAL, while for PASSION, due to limited samples, we only varied training representation of the lightest skin tone (FST 3). The configurations of the training sets were 20%, 40%, 60%, 80%, and 100% for each group, whilst the remaining samples were distributed equally across the race groups or skin tones.

Assessing Performance Disparities: To asses the bias in the racial representation during training we measure, for each subgroup, performance using the Area Under the Receiver Operating Characteristic Curve (AUC) in the RETINAL dataset and balanced accuracy (BACC) for the PASSION dataset. Additionally, we quantify the predictive (model) uncertainty as a test of whether these succeed at flagging potential bias. To capture the uncertainty, we employed the popular Monte Carlo (MC) dropout (p=0.3) and the gold standard Ensemble methods (5 models) (Gal and Ghahramani, 2016; Rahaman et al., 2021). As measures of model uncertainties, these should flag performance loss due to samples being out of distribution, as one might expect for groups that are underrepresented during training.

3. Discussion and conclusion

Our first main finding is that while dataset composition does affect performance, this effect is relatively small, especially on the RETINAL dataset. This does not, however, mean that the models are fair- we see large performance disparities between races and skin color that persist regardless of training set composition. In both datasets, the black or dark skinned population is at a disadvantage.

Our second main finding is that model uncertainty does not flag these performance disparities and is, therefore, not useful for flagging performance drops on racial subgroups. Both these findings support our main conclusion: **Uncertainty methods do not reduce or explain racial performance disparities, and we therefore hypothesize that there are further underlying causes for the differences** (Drukker et al., 2023). A deeper understanding of these underlying causes is, therefore, an important open problem.

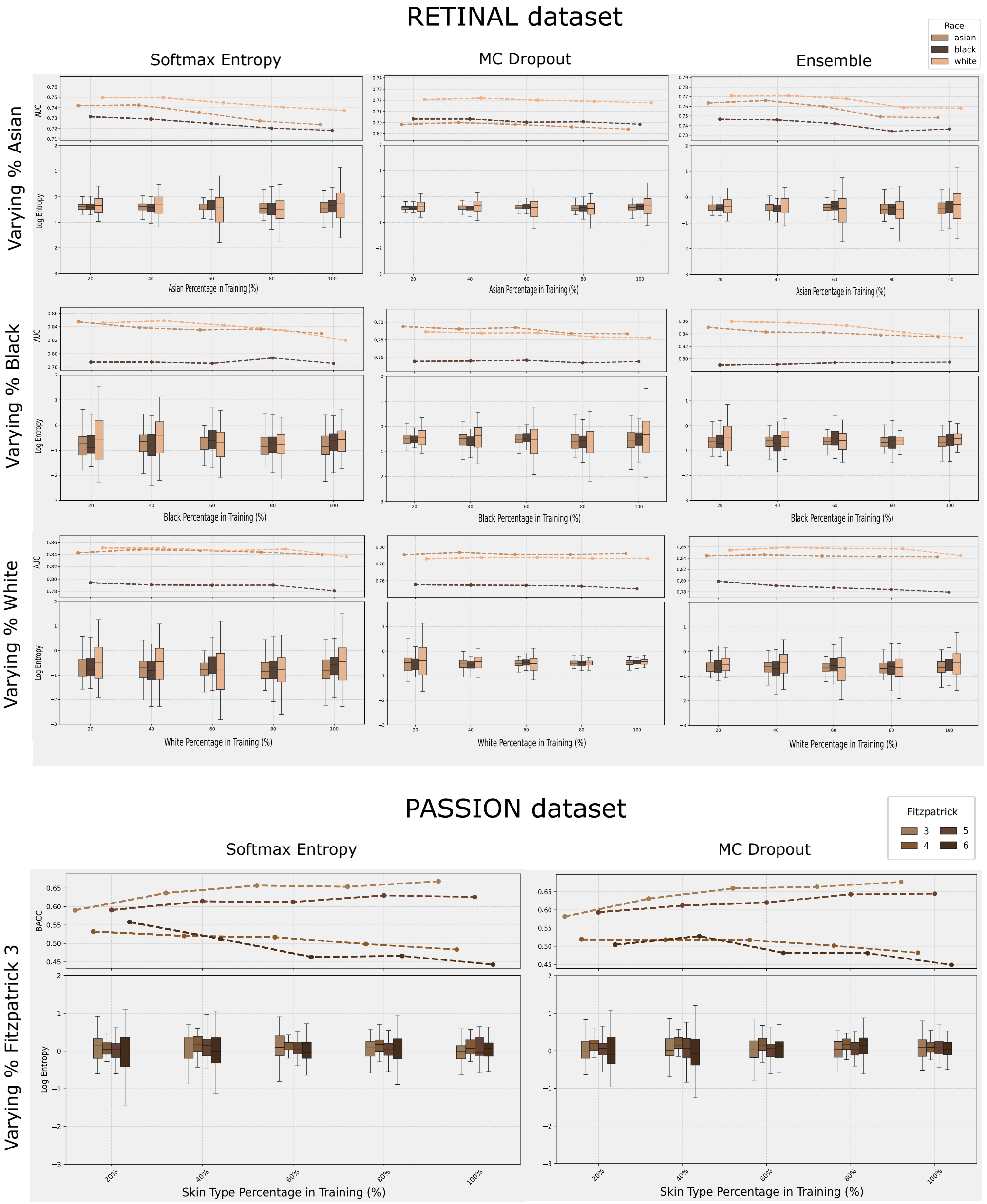

Figure 1: TOP: Retinal AUC trends vs. predictive uncertainty (log entropy) across columns; Asian (top), Black (middle), and White (bottom) demographics for models in each row; Base (left), MC Dropout (middle), and Ensemble (right) models trained on RETINAL datasets. BOTTOM: Comparison on PASSION dataset for Baseline (left) and MC Dropout (right) when trained on Fitzpatrick 3 but using BACC, with Ensemble and alternate methods left for future experiments.

References

Karen Drukker, Weijie Chen, Judy Gichoya, Nicholas Gruszauskas, Jayashree Kalpathy-Cramer, Sanmi Koyejo, Kyle Myers, Rui C Sá, Berkman Sahiner, Heather Whitney, et al. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. *Journal of Medical Imaging*, 10(6):061104–061104, 2023.

Enzo Ferrante and Rodrigo Echeveste. Open challenges on fairness of artificial intelligence in medical imaging applications.

Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, 2023.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Philippe Gottfrois, Fabian Gröger, Faly Herizo Andriambololoniaina, Ludovic Amruthalingam, Alvaro Gonzalez-Jimenez, Christophe Hsu, Agnes Kessy, Simone Lionetti, Daudi Mavura, Wingston Ng'ambi, et al. Passion for dermatology: Bridging the diversity gap with pigmented skin images from sub-saharan africa. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 703–712. Springer, 2024.

Shah Hussain, Iqra Mubeen, Niamat Ullah, Syed Shahab Ud Din Shah, Bakhtawar Abduljalil Khan, Muhammad Zahoor, Riaz Ullah, Farhat Ali Khan, and Mujeeb A Sultan. Modern diagnostic imaging technique applications and risk factors in the medical field: a review. *BioMed research international*, 2022(1):5164970, 2022.

Amelia Jiménez-Sánchez, Dovile Juodelyte, Bethany Chamberlain, and Veronika Cheplygina. Detecting shortcuts in medical images-a case study in chest x-rays. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.

Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23): 12592–12594, 2020.

Karim Lekadir, Alejandro F Frangi, Antonio R Porras, Ben Glocker, Celia Cintas, Curtis P Langlotz, Eva Weicken, Folkert W Asselbergs, Fred Prior, Gary S Collins, et al. Future-ai: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *bmj*, 388, 2025.

Yan Luo, Yu Tian, Min Shi, Louis R Pasquale, Lucy Q Shen, Nazlee Zebardast, Tobias Elze, and Mengyu Wang. Harvard glaucoma fairness: a retinal nerve disease dataset for fairness learning and fair identity normalization. *IEEE Transactions on Medical Imaging*, 2024.

Eike Petersen, Sune Holm, Melanie Ganz, and Aasa Feragen. The path toward equal performance in medical machine learning. *Patterns*, 4(7), 2023.

Esther Puyol-Antón, Bram Ruijsink, Stefan K Piechnik, Stefan Neubauer, Steffen E Petersen, Reza Razavi, and Andrew P King. Fairness in cardiac mr image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 413–423. Springer, 2021.

Rahul Rahaman et al. Uncertainty quantification and deep ensembles. *Advances in neural information processing systems*, 34:20063–20075, 2021.

María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante. Addressing fairness in artificial intelligence for medical imaging. *nature communications*, 13(1):4581, 2022.

Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. Medfair: benchmarking fairness for medical imaging. *arXiv preprint arXiv:2210.01725*, 2022.