Who Does Your Algorithm Fail? Investigating Age and Ethnic Bias in the MAMA-MIA Dataset

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Deep learning models aim to improve diagnostic workflows, but fairness evaluation remains underexplored beyond classification, e.g. in image segmentation. Unaddressed segmentation bias can lead to disparities in the quality of care for certain populations, potentially compounded across clinical decision points and amplified through iterative model development. Here, we audit the fairness of the automated segmentation labels provided in the breast cancer tumor segmentation dataset MAMA-MIA. We evaluate automated segmentation quality across age, ethnicity, and data source. Our analysis reveals an intrinsic age-related bias against younger patients that continues to persist even after controlling for confounding factors, such as data source. We hypothesize that this bias may be linked to physiological factors, a known challenge for both radiologists and automated systems. Finally, we show how aggregating data from multiple data sources influences site-specific ethnic biases, underscoring the necessity of investigating data at a granular level.

1 Introduction

Automated segmentation of breast tumors is a critical step in diagnosis, monitoring, treatment planning, and advancing robot-assisted surgeries [Michael et al., 2021, Benjelloun et al., 2018]. Inaccurate segmentation can lead to missed diagnoses, suboptimal treatment plans, and heterogeneous health outcomes across the patient population [Veta et al., 2014]. While recent advancements in deep learning have reached state-of-the-art performance [Isensee et al., 2021], their "fairness"– the principle that a model should not systematically disadvantage certain patient subgroups– remains a critical but often-overlooked aspect in medical image segmentation studies [Larrazabal et al., 2020].

The lack of datasets pairing high-quality imaging with demographic information and clinical metadata has limited the study of bias in medical image segmentation [Suresh and Guttag, 2019]. The MAMA-MIA dataset [Garrucho et al., 2024] addresses this gap, providing a multi-center cohort including detailed demographic information, clinical characteristics, and technical specifications. We perform a careful audit of the automated deep learning segmentations released as part of MAMA-MIA, following the *fairness under unawareness* paradigm [Barocas et al., 2023, Puyol-Antón et al., 2021]. We focus on ethnicity- and age-related disparities, motivated by clinical literature suggesting a correlation between younger patients, breast tissue density, and model performance on tumor detection tasks, caused by challenging tumor delineation for both radiologists and automated systems [Freer, 2015, Kontos et al., 2019, Tiryaki and Kaplanoğlu, 2022].

Our study provides: (i) to our knowledge, a first comprehensive fairness audit examining the intersection of multiple sensitive attributes in the MAMA-MIA breast tumor segmentation dataset, revealing statistically significant age- and ethnicity-based performance disparities; (ii) evidence that aggregating multi-center data and interaction between multiple factors can obscure site-specific ethnic bias; and (iii) a preliminary analysis on whether the bias is caused by lack of representation.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

2 Dataset and Methodology

The MAMA-MIA Dataset is a large, multi-center breast cancer benchmark of dynamic contrast-enhanced magnetic resonance images (DCE-MRI) [Garrucho et al., 2024]. Integrating four cohorts hosted on TCIA[1], it contains 1,507 T1-weighted DCE-MRI cases of female breast cancer patients.

We analyze key demographic and technical attributes with complete data coverage and established clinical relevance to segmentation performance: **Ethnicity:** Caucasian (74.9%), African-American (16.0%), Asian (5.7%), and other minority groups (3.4%); **Age Groups:** Young (<40 years, 23.2%), Middle (40-55 years, 50.1%), Older (>55 years, 26.6%). We discretize the continuous age variable into three bins informed by clinical literature on breast density changes and menopausal status transitions [Boyd et al., 2007, Checka et al., 2012]. The dataset uniquely includes dual annotations: Tumor regions manually segmented by a panel of 16 expert radiologists, serving as ground-truth annotations (gold labels) and automated segmentation masks from a model trained on external data (silver labels). The silver labels come with dual-expert qualitative ratings (Good, Acceptable, Poor, or Missed) assessing their visual quality. This dual annotation structure enables investigation of both model performance disparities and potential label quality biases across subgroups.

A **Fairness Auditing Framework** based on *fairness under unawareness* is adopted, following Chen et al. [2019]. Notably, [Puyol-Antón et al., 2021] conducted a pioneering work auditing deep learning models for cardiac image segmentation, assessing bias in models trained without explicit knowledge of sensitive attributes. We evaluate bias in automatic segmentation quality by comparing silver labels against gold labels using the Dice Score, 95th percentile Hausdorff Distance (HD95), and expert quality ratings. To formally quantify disparities, we measure the Demographic Parity Difference (DPD) $= |P(\hat{y} = 1|A = a) - P(\hat{y} = 1|A = b)|$ and Disparate Impact Ratio (DIR) $= \frac{\min(P(\hat{y}=1|A=a), P(\hat{y}=1|A=b))}{\max(P(\hat{y}=1|A=a), P(\hat{y}=1|A=b))}$. Here, $\hat{y} = 1$ is the beneficial outcome (e.g., a high-performance segmentation), A is the sensitive attribute (e.g., age or ethnicity), and a, b are distinct subgroups within that attribute [Caton and Haas, 2024, Castelnovo et al., 2022].

For these metrics, samples scoring in the top 25% for each metric were classified as high performers. Fairness gap (§) is the absolute difference in mean performance between the highest- and lowest-performing demographic subgroups [Tran and Woo, 2025].

To isolate the effect of representational imbalance, we conducted a controlled experiment using a setup designed to be representative of the original automated model. We trained a standard nnU-Net model using a 5-fold cross-validation scheme on an age-balanced cohort (n=1,047). This cohort was created by downsampling the 'Middle' and 'Older' groups to match the 'Young' group (n=349) and used only expert-annotated gold labels for training.

Statistical Analysis was used to assess performance differences across demographic subgroups. We first employ Ordinary Least Squares (OLS) regression [Zdaniuk, 2014] to model the relationship between sensitive attributes and performance metrics. Given that the performance metric distributions were non-normal (confirmed by Shapiro-Wilk tests), we used the non-parametric Kruskal-Wallis H-test to identify significant differences in performance across age and ethnicity groups. Where a significant overall difference was found, we conducted post-hoc pairwise comparisons to identify which specific subgroups differed, using Bonferroni correction. For the analysis of categorical expert ratings, the Chi-square test was used.

3 Results and Discussion

Age-Related Performance Disparities: Our analysis reveals that the automated silver labels have lower quality for younger patients, a statistically significant disparity that persists beyond simple representational imbalance. Across the complete cohort, segmentation quality improves with age. A baseline OLS regression (`Performance ~ Age`) demonstrates a significant, although small, relation between age and segmentation performance (Dice score: $R^2 = 0.0104, p = 0.0001$; HD95: $R^2 = 0.0093, p = 0.0009$), as visualized in Fig.1 (Right). These quantitative results are reflected in fairness metrics, which show a notable performance gap; for the Dice score, the DPD was 0.0887, with the 'young' group achieving a high performance at only 70% the rate of the 'older' group (DIR = 0.699).

[1] The Cancer Imaging Archive (TCIA) hosts de-identified cancer imaging datasets. Cohorts include: DUKE, I-SPY1 & 2, and NACT.

To determine if this bias was merely a confounding effect of the data source, we adjusted our OLS model to account for the source dataset, fitting a model of the form `Performance ~ AgeGroup + DataSource`. An ANOVA comparison between the baseline and the source-adjusted models confirmed that the source dataset is a significant factor (Dice: $F = 11.76$, $p = 1.3 \times 10^{-7}$; HD95: $F = 11.07$, $p = 3.4 \times 10^{-7}$). Even after this adjustment, the age effect remained highly significant, indicating the bias is not solely attributable to dataset-specific characteristics. This points towards an **intrinsic bias**. Further analysis revealed an interaction effect between age and dataset (Dice: $p = 1.6 \times 10^{-8}$), suggesting the magnitude of age-related bias varies depending on the data source.

Our controlled experiment on the age-balanced cohort confirmed the age-related bias is **intrinsic**, with a statistically significant fairness gap of 0.0399 (ANOVA p=0.0260) persisting even after eliminating representational imbalance, as shown in the comparative results in Table1.

Table 1: Age-Stratified Performance

Age	**Balanced Cohort**	**Automated**
Young	0.7304 ± 0.2333	0.8082 ± 0.2193
Middle	0.7333 ± 0.2253	0.8204 ± 0.2139
Older	0.7703 ± 0.1899	0.8612 ± 0.1679
§	**0.0399**	**0.0530**
p-value	**0.0260**	**0.0006**

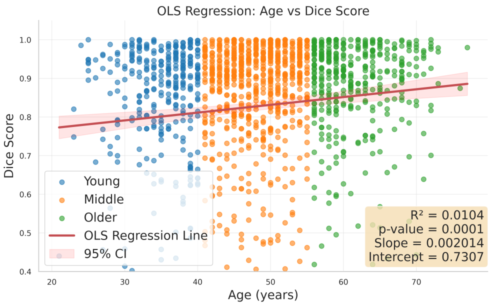

Figure 1: **Analysis of Age-Related Bias.** *(Left)* Comparison between model trained on balanced cohort vs. automated model segmentation i.e., the full (imbalanced) cohort, showing that a significant fairness gap remains even after balancing the training cohort. *(Right)* OLS regression visualizing the significant positive correlation between age and Dice score.

Ethnic Disparities and the Masking Effect of Data Aggregation: A global analysis across the aggregated dataset presents a misleading picture of ethnic fairness. For the Dice score, initial analysis suggests minimal disparity, with a non-significant Kruskal-Wallis test ($H = 5.09$, $p = 0.166$) and a near-equitable DIR of 0.89. Conversely, the HD95 metric indicates a significant disparity against the Asian subgroup ($p = 0.0046$), with a more critical DIR of 0.52. This inconsistency highlights the unreliability of aggregated analysis.

The true extent of bias is only revealed when disaggregating by data source. ANOVA test (`Performance ~ Ethnicity + DataSource`), confirms that the data source is a highly significant variable for both Dice ($F = 11.78$, $p = 1.2 \times 10^{-7}$) and HD95 ($F = 9.10$, $p = 6.0 \times 10^{-6}$). This demonstrates that institutional or cohort-specific factors are major confounders. For example, while the global DPD in Dice scores was only 3.0%, it amplified to 10.0% within the ISPY2 cohort. This disparity, entirely masked by pooling data, reveals that certain ethnic groups face substantial performance degradation in specific clinical contexts.

Additionally, we also find that the interpretation of ethnic bias is dependent on the evaluation metric. For the Dice score, a measure of volumetric overlap, the disparity proved to be an intrinsic bias, as adjusting for the data source only reduced its effect size by 6.2%. In contrast, for the HD95 score, a measure of boundary accuracy, the bias was largely a result of source confounding; adjusting for the data source reduced its effect size by a substantial 64.0%. This divergence suggests the model produces different *types* of segmentation errors for certain ethnic groups.

Outlook: We present a comprehensive fairness audit of a breast tumor segmentation model using the multi-center MAMA-MIA dataset, revealing significant age and ethnic disparities. Our analysis identified a persistent intrinsic bias against younger patients that survives balanced training, and severe, site-specific ethnic biases that are masked by multi-center data aggregation. This audit establishes a foundation for investigating the causal mechanisms underlying these biases. Future work will therefore focus on these origins through controlled training experiments and a systematic examination of annotation quality for evidence of label bias, with the ultimate goal of developing targeted mitigation strategies to ensure equitable model performance.

Potential Negative Societal Impacts

The primary motivation for this work is to mitigate the negative societal impact of biased AI systems in healthcare. Our goal is to promote equity by identifying performance disparities so they can be addressed before deployment. However, we recognize that this research, like any fairness audit, could have unintended negative consequences.

The most direct negative impact stems from the subject of our study itself. If the biases we identify in the segmentation model are not rectified, its deployment in a clinical setting would perpetuate and potentially amplify existing health inequities. Younger patients and certain ethnic minorities would receive a lower quality of diagnostic support, which could lead to delayed diagnoses, suboptimal treatment planning, and ultimately, worse health outcomes. Our work seeks to prevent this exact scenario. Furthermore, there is a risk that our findings could be misinterpreted. A superficial reading might lead to the oversimplified conclusion that "all AI is biased," fostering general distrust in valuable clinical tools.

Despite these risks, we firmly believe that the benefit of transparently reporting these biases far outweighs the potential for misuse. The greatest harm comes from allowing such disparities to remain hidden, where they can silently influence patient care. By bringing these issues to light, we intend to spur corrective action and encourage the development of more robust and equitable medical AI systems.

References

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

Mohammed Benjelloun, Mohammed El Adoui, Mohamed Amine Larhmam, and Sidi Ahmed Mahmoudi. Automated breast tumor segmentation in dce-mri using deep learning. In *2018 4th International Conference on Cloud Computing Technologies and Applications (Cloudtech)*, pages 1–6. IEEE, 2018.

Norman F Boyd, Helen Guo, Lisa J Martin, Limei Sun, Jennifer Stone, Eve Fishell, Roberta A Jong, Greg Hislop, Anna Chiarelli, Salomon Minkin, et al. Mammographic density and the risk and detection of breast cancer. *New England journal of medicine*, 356(3):227–236, 2007.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific reports*, 12(1):4209, March 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-07939-1. URL https://europepmc.org/articles/PMC8913820.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56 (7), April 2024. ISSN 0360-0300. doi: 10.1145/3616865. URL https://doi.org/10.1145/3616865.

Cristina M Checka, Jennifer E Chun, Freya R Schnabel, Jiyon Lee, and Hildegard Toth. The relationship of mammographic density and age: implications for breast cancer screening. *American Journal of Roentgenology*, 198(3):W292–W295, 2012.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. FAT* '19, page 339–348, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287594. URL https://doi.org/10.1145/3287560.3287594.

Phoebe E Freer. Mammographic breast density: impact on breast cancer risk and implications for screening. *Radiographics*, 35(2):302–315, 2015.

Lidia Garrucho, Claire-Anne Reidel, Kaisar Kushibar, Smriti Joshi, Richard Osuala, Apostolia Tsirikoglou, Maciej Bobowicz, Javier del Riego, Alessandro Catanese, Katarzyna Gwoździewicz, et al. Mama-mia: A large-scale multi-center breast cancer dce-mri benchmark dataset with expert segmentations. *arXiv e-prints*, pages arXiv–2406, 2024.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Despina Kontos, Stacey J Winham, Andrew Oustimov, Lauren Pantalone, Meng-Kang Hsieh, Aimilia Gastounioti, Dana H Whaley, Carrie B Hruska, Karla Kerlikowske, Kathleen Brandt, et al. Radiomic phenotypes of mammographic parenchymal complexity: toward augmenting breast density in breast cancer risk assessment. *Radiology*, 290(1):41–49, 2019.

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.

Epimack Michael, He Ma, Hong Li, Frank Kulwa, and Jing Li. Breast cancer segmentation methods: current status and future potentials. *BioMed research international*, 2021(1):9962109, 2021.

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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 413–423. Springer, 2021.

Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2(8):73, 2019.

V.M. Tiryaki and V. Kaplanoğlu. Deep learning-based multi-label tissue segmentation and density assessment from mammograms. *IRBM*, 43(6):538–548, 2022. ISSN 1959-0318. doi: https://doi.org/10.1016/j.irbm.2022.05.004. URL `https://www.sciencedirect.com/science/article/pii/S1959031822000562`.

Khoa Tran and Simon S. Woo. Fairness and robustness in machine unlearning. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 1336–1340, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713316. doi: 10.1145/3701716.3715598. URL `https://doi.org/10.1145/3701716.3715598`.

Mitko Veta, Josien PW Pluim, Paul J Van Diest, and Max A Viergever. Breast cancer histopathology image analysis: A review. *IEEE transactions on biomedical engineering*, 61(5):1400–1411, 2014.

Bozena Zdaniuk. *Ordinary Least-Squares (OLS) Model*, pages 4515–4517. Springer Netherlands, Dordrecht, 2014. ISBN 978-94-007-0753-5. doi: 10.1007/978-94-007-0753-5_2008. URL `https://doi.org/10.1007/978-94-007-0753-5_2008`.