Investigating the Fairness of Deep Learning Models in Breast Cancer Diagnosis Based on Race and Ethnicity

Kuan Huang[1]*, **Yingfeng Wang**[2], **Meng Xu**[1]

[1]Department of Computer Science and Technology, Kean University, Union, NJ, United States
[2]Department of Computer Science and Engineering, University of Tennessee at Chattanooga, Chattanooga, TN, United States
{khuang, mexu}@kean.edu, yingfeng-wang@utc.edu

Abstract

Breast cancer is the leading cancer affecting women globally. Despite deep learning models making significant strides in diagnosing and treating this disease, ensuring fair outcomes across diverse populations presents a challenge, particularly when certain demographic groups are underrepresented in training datasets. Addressing the fairness of AI models across varied demographic backgrounds is crucial. This study analyzes demographic representation within the publicly accessible Emory Breast Imaging Dataset (EMBED), which includes de-identified mammography and clinical data. We spotlight the data disparities among racial and ethnic groups and assess the biases in mammography image classification models trained on this dataset, specifically ResNet-50 and Swin Transformer V2. Our evaluation of classification accuracies across these groups reveals significant variations in model performance, highlighting concerns regarding the fairness of AI diagnostic tools. This paper emphasizes the imperative need for fairness in AI and suggests directions for future research aimed at increasing the inclusiveness and dependability of these technologies in healthcare settings. Code is available at: https://github.com/kuanhuang0624/EMBEDFairModels.

Introduction

Breast cancer remains the most commonly diagnosed cancer among women in the U.S. and is the second leading cause of cancer death among women overall. In 2023, breast cancer was projected to account for 31% of all female cancer cases (Giaquinto et al. 2022). Currently, approximately 4.1 million women in the U.S. live with a history of breast cancer, with about 4% suffering from metastatic disease. Notably, over half of these cases were identified at early stages (1-3) (Gallicchio et al. 2022). According to the American Cancer Society, the survival rates for breast cancer at stages 0 and 1 approached nearly 100% from 2007 to 2013 (DeSantis et al. 2016), underscoring the importance of enhanced screening and treatment facilities to boost women's health both domestically and globally. X-ray mammography and ultrasound are the primary modalities for early breast cancer detection, particularly mammography, which shows promising detection outcomes (Prodan, Paraschiv, and Stanciu 2023). Over the past decade, AI-based Computer-Aided Diagnosis (CAD) systems have been developed for breast cancer diagnosis in mammography (Ricciardi et al. 2021; Atrey et al. 2023). The effectiveness of these AI systems largely depends on the quality of the data used for training, with data fairness being crucial. The key fairness concerns include age and racial/ethnic disparities. Age significantly influences breast density, which is critical for accurate breast cancer diagnosis in mammography (Garrucho et al. 2023). Notably, 83% of breast cancer cases occur in women over 50 years of age, who also account for 91% of breast cancer-related deaths, with half of these deaths occurring in women aged 70 or older (Giaquinto et al. 2022).

Addtionally, there are significant disparities in breast cancer incidence and outcomes among different racial and ethnic groups, according to data from the American Cancer Society (Giaquinto et al. 2022). Table 1 illustrates substantial racial disparities in both incidence and mortality rates in the U.S. Black women, for example, have a lower incidence rate but suffer from a 40% higher death rate compared to White women, along with the lowest five-year relative survival rate among all racial and ethnic groups. These variations underscore the inequities in access to medical resources and financial support across different communities.

	White	***Black***	*API**	*Hispanic*	*AIAN**
Incidence	133.7	**127.8**	101.3	99.2	111.3
Mortality	19.7	**27.6**	11.7	13.7	20.5

**API represents Asian/Pacific Islander. AIAN represents American Indian/Alaska Native. Rates are expressed per 100,000 people.*

Table 1: Female breast cancer incidence and mortality rates by race/ethnicity (2015-2019), the U.S.

The presence of biases in breast cancer incidence across racial and ethnic groups can contribute to biases in the development of AI-based CAD systems. Due to the lower incidence rates in certain racial groups, some datasets may have fewer samples, leading to data imbalances. Several studies have investigated the fairness and bias in AI-based medical imaging systems (Logan, Kennedy, and Catchpoole 2023; Yang et al. 2024; Ueda et al. 2024; Hort et al. 2024). The issues of bias in deep learning models typically fall into two

*Corresponding author.

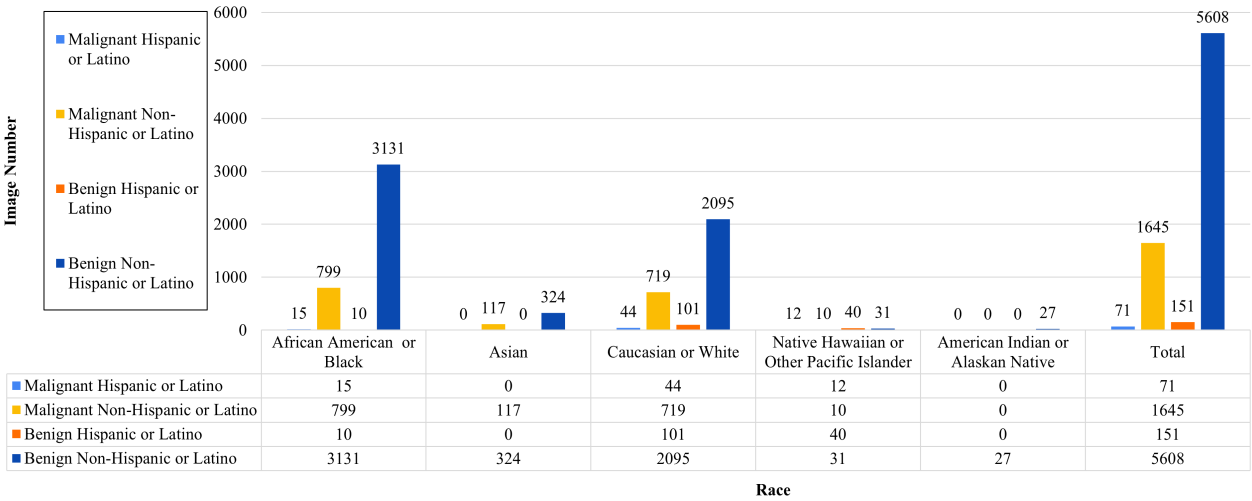

Figure 1: Distribution of selected images across ethnic and racial groups and between malignant and benign classes for training the classification models.

categories: 1) data imbalance and 2) model training. Efforts to mitigate bias in AI models generally focus on these areas, employing strategies such as data augmentation or synthetic data to enhance data representation (Garrucho et al. 2023), and improving model training, for instance, by incorporating regularization terms or using adversarial training approaches (Lahoti et al. 2020). Additionally, post-processing methods are used to refine fairness after model training (Soltan and Washington 2024).

This study investigates how these disparities can lead to biases in AI-based CAD systems. We mainly focus on baseline models like ResNet-50 and Swin Transformer V2, chosen for their widespread use in image classification. We do not choose other complex models because their additional features could introduce new variables, making it harder to isolate and study the specific impacts of algorithmic bias on fairness in classification. Our goal is to highlight concerns about fairness using baseline models. This study utilizes the Emory Breast Imaging Dataset (EMBED) (Jeong et al. 2023) to explore these issues. Our contributions are:

- We assess the fairness of baseline models, ResNet-50 and Swin Transformer V2, in classifying mammography images into benign or malignant across different racial and ethnic groups.
- We highlight significant data imbalances within the EMBED dataset and observe notable declines in model performance across demographic lines, underlining the urgent need for fairness in AI research.

With a focus on refining AI diagnostics through a lens of equity, this research aims to improve technological outcomes and ensure fair medical practices across all populations.

Materials and Methods

Dataset

The EMBED dataset comprises 400,000 de-identified images from around 20,000 patients and includes both 2D and C-view images in this release. In this research, we exclusively use 2D mammography images and do not use 3D or C-view images. In addition to images, the dataset provides detailed metadata for each image, such as the capture date, image type, and the number and location of regions of interest (ROI). It also includes patient identifiers and extensive clinical information about the patients. Notably, the dataset captures the ethnic and racial backgrounds of the patients, encompassing categories such as “African American or Black,” “Asian,” “Caucasian or White,” “Native Hawaiian or Other Pacific Islander,” “American Indian or Alaskan Native,” “Not Recorded,” ”Patient Declines,” “Unknown, Unavailable or Unreported,” and “Multiple” race groups. For ethnic categorization, it includes “Hispanic or Latino,” “Non-Hispanic or Latino,” “Not Recorded,” and “Unknown-Non-Hispanic,” among others. To analyze the fairness of a deep learning-based classification model, we have chosen specific racial groups, including “African American or Black,” “Asian,” “Caucasian or White,” “Native Hawaiian or Other Pacific Islander,” and “American Indian or Alaskan Native.” Additionally, we have selected “Hispanic or Latino” and “Non-Hispanic or Latino” as the ethnic groups for this study. Groups categorized as unknown or unreported have been excluded from our selection.

Approximately 90% of screening mammograms yield normal results, and only a minority of patients undergo pathology testing. To train our deep learning models, we classify patients based on whether they have cancerous conditions or benign lesions in the breast. For this purpose, we use the “path_severity” attribute from the clinical data as the label, indicating the most severe pathology result from a given specimen. The “path_severity” attribute encompasses 7 categories: 0: invasive cancer 1: non-invasive cancer 2: high-risk lesion 3: borderline lesion 4: benign findings 5: negative (normal breast tissue) 6: non-breast cancer. For this study, we select images from patients with “path_severity” values of 0 and 1 to represent malignant cases and those

with a "path_severity" of 4 for benign cases. Images from other categories are excluded from our selection.

Under the abovementioned conditions, the total number of selected samples amounts to 7,475. We analyze the distribution of these selected images across various ethnic and racial groups. The distribution of images among different racial and ethnic groups and between malignant and benign classes is depicted in Figure 1. To train deep learning models, we have divided the 7,475 selected images into an 70% training set and a 30% testing set, maintaining consistent distributions across racial and ethnic groups in both datasets.

Implementation Details and Metric

All experiments are conducted using PyTorch 1.13.1 on an Ubuntu 20.04 system. The hardware setup includes an AMD EPYC 7513 2.60 GHz CPU and eight NVIDIA GeForce RTX 3090 graphics cards, each with 24GB of memory. We assess the training performance of the model using accuracy as the metric. The model is trained on a dataset that includes all racial and ethnic groups. We then evaluate and report the accuracy for different racial and ethnic groups within the test dataset and the entire test dataset. We also employ Equalized Odds (EqOdd) following (Zong, Yang, and Hospedales 2023) as a group fairness metric, which requires that the true positive and false positive rates be equalized across subgroups. We evaluate EqOdd specifically for the Hispanic or Latino and Non-Hispanic or Latino groups.

Methods

Models: We train two widely utilized classification models: ResNet-50 (He et al. 2016) and Swin Transformer V2 (Liu et al. 2022). The implementation of ResNet-50 is sourced from the torchvision library, using an input image size of 224×224 pixels, and the model is configured to output two classes, utilizing a pre-trained model from ImageNet. The Swin Transformer V2 implementation is derived from a pre-trained small model available on the Hugging Face library, with the same input image size of 224×224 pixels.

Training Details: We employ a consistent training approach for both models, spanning 40 epochs with a batch size of 32, utilizing the Adam optimizer. The optimizer is configured with a learning rate of 1e-4 and a weight decay of 1e-4. We use cross-entropy loss as the loss function. The learning rate is scheduled to decay every 20 epochs, reducing to 10% of its original value.

Results

The breast mammography classification results for ResNet-50 and Swin Transformer V2 models are presented in Tables 2 and 3, respectively. These results show the test accuracy across different races and ethnicities.

ResNet-50: Table 2 presents the test accuracy for the ResNet-50 model. The overall test accuracy is 0.8012. When breaking down by race and ethnicity, the accuracies are as follows: African American or Black: 0.5000 (Hispanic or Latino), 0.8183 (Non-Hispanic or Latino), and a combined accuracy of 0.8172. Asian: Data for Hispanic or Latino is unavailable, while non-Hispanic or Latino shows an accuracy of 0.8182. Caucasian or White: 0.6667 (Hispanic or Latino), 0.7829 (Non-Hispanic or Latino), and a combined accuracy of 0.7770. Native Hawaiian or Other Pacific Islander: 0.8000 (Hispanic or Latino), 0.6923 (Non-Hispanic or Latino), and a combined accuracy of 0.7500. American Indian or Alaskan Native: Data for Hispanic or Latino is unavailable, while non-Hispanic or Latino shows an accuracy of 1.0000. Overall, Hispanic or Latino groups had a lower test accuracy (0.6875) compared to non-Hispanic or Latino groups (0.8045). The fairness analysis between Hispanic or Latino and Non-Hispanic or Latino groups yielded an EqOdd score of 0.9387. The EqOdd scores for Hispanic or Latino and Non-Hispanic or Latino groups were 0.8056 for African American or Black, 0.9323 for Caucasian or White, and 0.9231 for Native Hawaiian or Other Pacific Islander, respectively.

Swin Transformer V2: Table 3 shows the test accuracy for the Swin Transformer V2 model, with an overall test accuracy of 0.7704. The breakdown by race and ethnicity is as follows: African American or Black: 0.5000 (Hispanic or Latino), 0.8115 (Non-Hispanic or Latino), and a combined accuracy of 0.8104. Asian: Data for Hispanic or Latino is unavailable, while non-Hispanic or Latino shows an accuracy of 0.7576. Caucasian or White: 0.5778 (Hispanic or Latino), 0.7248 (Non-Hispanic or Latino), and a combined accuracy of 0.7173. Native Hawaiian or Other Pacific Islander: 0.8667 (Hispanic or Latino), 0.6154 (Non-Hispanic or Latino), and a combined accuracy of 0.7500. American Indian or Alaskan Native: Data for Hispanic or Latino is unavailable, while non-Hispanic or Latino shows an accuracy of 1.0000. The Swin Transformer V2 model also displayed a lower accuracy for Hispanic or Latino groups (0.6406) than non-Hispanic or Latino groups (0.7742). The fairness analysis between Hispanic or Latino and Non-Hispanic or Latino groups yielded an EqOdd score of 1.0.

Discussion

The results of our study demonstrate significant disparities in model performance across different racial and ethnic groups when classifying breast mammography images using the ResNet-50 and Swin Transformer V2 models. As shown in Figure 1, the dataset distribution provides essential context for interpreting these results.

Dataset Analysis: The dataset distribution reveals several important aspects that impact the performance and fairness of the models: 1) Race imbalance: The number of images is significantly higher for African American or Black (3,955) and Caucasian or White (2,959) compared to other races combined (561). 2) Ethnic imbalance: The dataset is heavily skewed towards non-Hispanic or Latino cases (1,645 malignant, 5,608 benign), making up the vast majority of the data (7,253 out of 7,475 total cases). Hispanic or Latino cases are significantly underrepresented (71 malignant, 151 benign), making it challenging for the model to learn and generalize accurately for this group. 3) Class imbalance: A substantial imbalance exists between benign (7,253) and malignant (222) images.

Race \ Ethnicity	Hispanic or Latino	Non-Hispanic or Latino	Total
African American or Black	0.5000	0.8183	0.8172
Asian	-	0.8182	0.8182
Caucasian or White	0.6667	0.7829	0.7770
Native Hawaiian or Other Pacific Islander	0.8000	0.6923	0.7500
American Indian or Alaskan Native	-	1.0000	1.0000
Total	0.6875	0.8045	0.8012

Table 2: Test Accuracy for the ResNet-50 Model

Race \ Ethnicity	Hispanic or Latino	Non-Hispanic or Latino	Total
African American or Black	0.5000	0.8115	0.8104
Asian	-	0.7576	0.7576
Caucasian or White	0.5778	0.7248	0.7173
Native Hawaiian or Other Pacific Islander	0.8667	0.6154	0.7500
American Indian or Alaskan Native	-	1.0000	1.0000
Total	0.6406	0.7742	0.7704

Table 3: Test Accuracy for the Swin Transformer V2 Model

Model Performance and Fairness: 1) Both models show higher accuracy for non-Hispanic or Latino groups than for Hispanic or Latino groups. This disparity is due to the significant imbalance between the two groups, which makes it challenging for the models to learn and generalize effectively for Hispanic or Latino individuals, resulting in poorer performance. The models perform better in the more represented non-Hispanic or Latino groups due to the larger volume of training data. 2) Among different races, having a larger number of images does not necessarily result in higher accuracy. For example, the number of images is significantly higher for African American or Black (3,955) and Caucasian or White (2,959) individuals compared to all other races combined (561). However, despite having a large number of images, the Caucasian or White group has the lowest accuracy in Swin Transformer V2 and the second to the lowest accuracy in the ResNet-50. This suggests that merely having more images does not guarantee better model performance within this group. 3) The imbalance between malignant and benign cases affects the models' ability to achieve higher performance. The overwhelming number of benign cases compared to malignant ones can lead to a bias in the models, making them less effective at correctly identifying malignant cases. 4) The Swin Transformer V2 model yields an EqOdd score of 1.0 because it performs poorly on both Hispanic or Latino and Non-Hispanic or Latino groups, rendering the fairness metric meaningless in this context.

Conclusion and Future Works

In this study, we evaluate the performance and fairness of two widely used baseline classification models, ResNet-50 and Swin Transformer V2, in classifying mammography images from the EMBED dataset. Our results demonstrated significant disparities in model accuracy across different racial and ethnic groups in two baseline models. The key findings of our study are as follows: 1) Imbalance in Data Representation: The dataset has significant imbalances in race, ethnicity, and class distributions. The number of images is disproportionately higher for African American or Black and Caucasian or White individuals compared to other races. Non-Hispanic or Latino cases are vastly overrepresented compared to Hispanic or Latino cases. There is a substantial imbalance between benign and malignant images, with benign cases overwhelmingly dominating the dataset. 2) Model Performance and Fairness: The ResNet-50 and Swin Transformer V2 models exhibit higher accuracy for non-Hispanic or Latino groups compared to Hispanic or Latino groups due to the imbalance in representation. The imbalances in the number of samples among different racial groups also contribute to varying performance across these groups. Additionally, the significant imbalance between benign and malignant cases introduces a bias in the models, reducing their effectiveness in accurately identifying malignant cases.

Future Works: To address these disparities and enhance the fairness and performance of deep learning models in breast cancer diagnosis, we will focus on the following areas: 1) Data balancing techniques: Implement methods to balance the representation of different demographic groups in the dataset. 2) Data augmentation: Use augmentation techniques to increase the diversity and quantity of training data, especially for underrepresented groups. 3) Bias mitigation strategies: Develop and apply strategies to reduce model bias arising from data imbalances.

Acknowledgements

This work was performed with partial support from the National Science Foundation under Grants Nos. 2430746 and 2430747.

References

Atrey, K.; Singh, B. K.; Bodhey, N. K.; and Pachori, R. B. 2023. Mammography and ultrasound based dual modality classification of breast cancer using a hybrid deep learning approach. *Biomedical Signal Processing and Control*, 86: 104919.

DeSantis, C. E.; Fedewa, S. A.; Goding Sauer, A.; Kramer, J. L.; Smith, R. A.; and Jemal, A. 2016. Breast cancer statistics, 2015: Convergence of incidence rates between black and white women. *CA: a cancer journal for clinicians*, 66(1): 31–42.

Gallicchio, L.; Devasia, T. P.; Tonorezos, E.; Mollica, M. A.; and Mariotto, A. 2022. Estimation of the number of individuals living with metastatic cancer in the United States. *JNCI: Journal of the National Cancer Institute*, 114(11): 1476–1483.

Garrucho, L.; Kushibar, K.; Osuala, R.; Diaz, O.; Catanese, A.; Del Riego, J.; Bobowicz, M.; Strand, F.; Igual, L.; and Lekadir, K. 2023. High-resolution synthesis of high-density breast mammograms: Application to improved fairness in deep learning based mass detection. *Frontiers in Oncology*, 12: 1044496.

Giaquinto, A. N.; Sung, H.; Miller, K. D.; Kramer, J. L.; Newman, L. A.; Minihan, A.; Jemal, A.; and Siegel, R. L. 2022. Breast cancer statistics, 2022. *CA: a cancer journal for clinicians*, 72(6): 524–541.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hort, M.; Chen, Z.; Zhang, J. M.; Harman, M.; and Sarro, F. 2024. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 1(2): 1–52.

Jeong, J. J.; Vey, B. L.; Bhimireddy, A.; Kim, T.; Santos, T.; Correa, R.; Dutt, R.; Mosunjac, M.; Oprea-Ilies, G.; Smith, G.; et al. 2023. The EMory BrEast imaging Dataset (EMBED): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. *Radiology: Artificial Intelligence*, 5(1): e220047.

Lahoti, P.; Beutel, A.; Chen, J.; Lee, K.; Prost, F.; Thain, N.; Wang, X.; and Chi, E. 2020. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33: 728–740.

Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. 2022. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12009–12019.

Logan, J.; Kennedy, P. J.; and Catchpoole, D. 2023. A review of the machine learning datasets in mammography, their adherence to the FAIR principles and the outlook for the future. *Scientific Data*, 10(1): 595.

Prodan, M.; Paraschiv, E.; and Stanciu, A. 2023. Applying deep learning methods for mammography analysis and breast cancer detection. *Applied Sciences*, 13(7): 4272.

Ricciardi, R.; Mettivier, G.; Staffa, M.; Sarno, A.; Acampora, G.; Minelli, S.; Santoro, A.; Antignani, E.; Orientale, A.; Pilotti, I.; et al. 2021. A deep learning classifier for digital breast tomosynthesis. *Physica Medica*, 83: 184–193.

Soltan, A.; and Washington, P. 2024. Challenges in Reducing Bias Using Post-Processing Fairness for Breast Cancer Stage Classification with Deep Learning. *Algorithms*, 17(4): 141.

Ueda, D.; Kakinuma, T.; Fujita, S.; Kamagata, K.; Fushimi, Y.; Ito, R.; Matsui, Y.; Nozaki, T.; Nakaura, T.; Fujima, N.; et al. 2024. Fairness of artificial intelligence in healthcare: review and recommendations. *Japanese Journal of Radiology*, 42(1): 3–15.

Yang, Y.; Zhang, H.; Gichoya, J. W.; Katabi, D.; and Ghassemi, M. 2024. The limits of fair medical imaging AI in real-world generalization. *Nature Medicine*, 1–11.

Zong, Y.; Yang, Y.; and Hospedales, T. 2023. MEDFAIR: Benchmarking Fairness for Medical Imaging. In *The Eleventh International Conference on Learning Representations*.