Bias and Generalizability of Foundation Models across Datasets in Breast Mammography

Elodie Germani[1], Ilayda Selin-Türk[2], Fatima Zeineddine[3], Charbel Mourad[3], and Shadi Albarqouni[1,2,4]

[1] Clinic for Diagnostic and Interventional Radiology, University Hospital Bonn, Bonn, Germany
[2] TUM School of Computation, Information and Technology, Technical University of Munich, Munich, Germany
[3] Department of Diagnostic Imaging and Interventional Therapeutics, Lebanese Hospital Geitaoui, Beyrouth, Lebanon
[4] Helmholtz AI, Helmholtz Munich, Munich, Germany
shadi.albarqouni@ukbonn.de

Abstract. Over the past decades, computer-aided diagnosis tools for breast cancer have been developed to enhance screening procedures, yet their clinical adoption remains challenged by data variability and inherent biases. Although foundation models (FMs) have recently demonstrated impressive generalizability and transfer learning capabilities by leveraging vast and diverse datasets, their performance can be undermined by spurious correlations that arise from variations in image quality, labeling uncertainty, and sensitive patient attributes. In this work, we explore the fairness and bias of FMs for breast mammography classification by leveraging a large pool of datasets from diverse sources—including data from underrepresented regions and an in-house dataset. Our extensive experiments show that while modality-specific pre-training of FMs enhances performance, classifiers trained on features from individual datasets fail to generalize across domains. Aggregating datasets improves overall performance, yet does not fully mitigate biases, leading to significant disparities across under-represented subgroups such as extreme breast densities and age groups. Furthermore, while domain-adaptation strategies can reduce these disparities, they often incur a performance trade-off. In contrast, fairness-aware techniques yield more stable and equitable performance across subgroups. These findings underscore the necessity of incorporating rigorous fairness evaluations and mitigation strategies into FM-based models to foster inclusive and generalizable AI.

Keywords: Fairness · Mammography · Foundation models.

1 Introduction

Breast cancer is one of the most significant global health challenges, with over 2.3 million new cases and approximately 670,000 deaths reported in 2022 alone [4]. Early and accurate detection is crucial for improving patient outcomes, and

mammographic screening, typically confirmed by biopsy, remains a cornerstone of clinical diagnosis. In recent years, deep learning models have shown promise in aiding radiologists by extracting breast cancer biomarkers with high performance, sometimes even surpassing that of human experts [20]. However, these models are often developed using datasets drawn predominantly from specific populations, which tend to under-represent marginalized groups, potentially leading to biases and reliance on spurious correlations that do not generalize well across populations [27]. This under-representation is particularly problematic in breast cancer detection, as critical risk factors such as age and breast density may vary across different ethnicities, and geographic regions [8, 6, 14].

In response to these challenges, foundation models (FMs) have emerged as a promising solution due to their ability to learn rich and transferable visual representations from diverse large-scale datasets [3, 25]. By working on pre-extracted features rather than raw images, FMs offer the potential for improved generalizability and reduced computational overhead in resource-limited settings [12]. However, recent studies have revealed that FMs are also susceptible to bias, as they can inadvertently capture spurious correlations inherent in their training data [13, 17]. Such biases raise concerns about the equity of AI systems in clinical practice, particularly when deployed across diverse demographic groups.

Motivated by these observations, this work investigates the presence of bias in FMs applied to breast cancer biomarkers detection and explores bias mitigation strategies through domain adaptation and fairness techniques. Unlike previous works primarily assessing FM fairness within individual datasets [17], we extend our analysis to between-dataset biases and domain shifts. To this end, we aggregate a diverse set of mammography datasets sourced from various parts of the world, including underrepresented regions, and supplement them with an in-house dataset from Lebanon (LBMD) with around 3,000 images from 700 patients. Directly sourced from clinical practice, LBMD captures real-world complexities often overlooked in curated public datasets, offering an additional perspective on clinical settings. Our **contributions** are threefold. First, we conduct a comprehensive analysis of bias in FMs by evaluating the risk of spurious correlations when classifiers are trained on different datasets. Second, we assess traditional domain-adaptation and fairness strategies as potential solutions to mitigate these biases. Third, by incorporating the LBMD dataset, we demonstrate the clinical relevance of our results, addressing disparities in breast cancer biomarkers detection, and ultimately advancing the development of more robust and equitable AI tools to support radiologists in diverse clinical settings.

2 Methodology

Let $\mathcal{X}$ be the space of mammography images and $\mathcal{Y}$ the label space (*e.g.* $\{0, 1, 2\}$ for diagnosis or $\{1, 2, 3, 4\}$ for breast density classification). Each data point is a triplet (x_i, y_i, d_i), where $x_i \in \mathcal{X}$ is the image, $y_i \in \mathcal{Y}$ its label, and $d_i \in \mathcal{D}$ denotes the domain or dataset source. The complete dataset is given by $\mathcal{S} =$

Table 1: Summary of bias mitigation and domain-adaptation Methods.

Method	**Objective** (Note: $z_i = \phi_z(\phi(x_i))$, and $H(\cdot\|\cdot)$ is conditional entropy)
`DANN` [10]	$\mathcal{L}_{\text{DANN}}(\theta;\psi) = \mathcal{L}_{\text{WCE}} - \ell_d\Big(g_\psi(\phi(x_i)), d_i\Big)$
`FairDisCO` [9]	$\mathcal{L}_{\text{FairDisCO}}(\theta;\psi;\phi_z) = \mathcal{L}_{\text{DANN}} + \alpha\mathcal{L}_{\text{conf}} + \beta\mathcal{L}_{\text{contr}}$ $\mathcal{L}_{\text{conf}}(\theta) = -\sum_{i=1}^{N} \frac{1}{N} \cdot \log(f_\theta(\phi(x_i)))$ $\mathcal{L}_{\text{contr}}(\theta,\phi_z) = \sum_{(i,j)\in P_y} \log \frac{\exp(\text{sim}(z_i,z_j))}{\exp(\text{sim}(z_i,z_j)))+\sum_{k\in N_y}\exp(\text{sim}(z_i,z_k)))}$ $\mathcal{L}_{conf}$ max. equal probabilities across $\mathcal{D}$, $\mathcal{L}_{contr}$ max. $\mathcal{D}$-invariant representations.
`FADES` [16]	$\mathcal{L}_{\text{FADES}}(\theta;\psi;\phi_z) = \mathcal{L}_{\text{DANN}} + \mathcal{L}_{\text{TC}} + \mathcal{L}_{\text{CMI}} + \mathcal{L}_{\text{reg}}$ $\mathcal{L}_{\text{CMI}}(\theta,\psi) = I_\phi(f_\theta(z_i); g_\psi(\phi(z_i))\|d_i)$ and $\mathcal{L}_{\text{TC}}(\phi_z) = D_{\text{KL}}(z_i \parallel \prod_j z_j)$ aim to learn disentangled representations: domain-relevant, task-relevant and irrelevant. $\mathcal{L}_{\text{reg}} = -\left(H(f_\theta(z_i)\|z_R) + H(g_\psi(z_i)\|z_R)\right)$ to regularize training objective.
`GroupDRO` [28]	$\mathcal{L}_{\text{GroupDRO}} = \min_\theta \max_{q\in\Delta^{\|\mathcal{D}\|}} \sum_{d\in\mathcal{D}} q_d\, \mathcal{L}_d(\theta)$ $\mathcal{L}_d(\theta) = \frac{1}{\|S_d\|}\sum_{i:d_i=d} \ell\Big(f_\theta(\phi(x_i)), y_i\Big)$ to minimize empirical worst-group risk.
`MOE` [19]	$\mathcal{L}_{\text{MOE}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell\Big(f_{\text{MOE}}(x_i), y_i\Big)$, with each expert specialized in one domain. $f_{\text{MOE}}(x) = \sum_{e=1}^{E} \alpha_e(x) f_\theta^e(\phi(x))$ and $\alpha_e(x) = \frac{\exp(w_e^T x)}{\sum_{j=1}^{E}\exp(w_j^T x)}$

$\{(x_i, y_i, d_i)\}_{i=1}^N$. Our goal is to learn a classifier $f_\theta : \mathcal{X} \to \mathcal{Y}$ that achieves high predictive performance while mitigating any sort of bias.

Feature Extraction and Classification. We extract a representation $\phi(x) \in \mathbb{R}^m$ using a frozen, pre-trained FM ϕ. A linear probe is then trained over these features: $f_\theta(\phi(x)) = \text{sigmoid}(W\,\phi(x) + b)$, where $\theta = \{W, b\}$. We use a weighted cross-entropy loss: $\mathcal{L}_{\text{WCE}}(\theta;\mathcal{S}) = \frac{1}{N}\sum_{i=1}^{N} w_{y_i} \cdot \ell(f_\theta(\phi(x_i)), y_i)$, with w_{y_i} corresponding to inverse class percentages. Simple minimization of $\mathcal{L}_{\text{WCE}}$ may correct for class imbalance, but might not adequately address domain shift or bias.

Bias and Domain-Adaptation Methods. We investigate several strategies to promote domain invariance and mitigate biases. Specifically, we explored two categories of techniques to balance classifiers' performance across datasets: i) domain-adaptation strategies (`DANN`, `FairDisCO`, and `FADES`) primarily designed to learn domain-invariant representations, and ii) fairness-aware strategies (`GroupDRO`, `MOE`) which explicitly focus on reducing performance disparities across subgroups. Table 1 summarizes the formulations of each method.

Domain Adversarial Neural Network(`DANN`) uses adversarial learning to enforce extraction of domain-invariant features from $\phi_z(\cdot)$ by introducing a domain classifier $g_\psi(\cdot)$ and reversing the gradient of the domain classification loss ℓ_d [10].
Fair Disentanglement with Contrastive Learning (`FairDisCO`) employs adversarial and contrastive learning. It encourages samples from different domains with the same label to be close in a new feature space $\phi_z(\cdot)$ to mitigate bias [9].
Fair Disentanglement with Sensitive Relevance (`FADES`) penalizes $\phi_z(\cdot)$ features predictive of domain while maintaining those correlated with both domain and target tasks. It integrates total correlation (TC), conditional mutual information (CMI), and adversarial loss to minimize domain information leakage [16].
Group Distributionally Robust Optimization (`GroupDRO`) explicitly optimizes for the worst-case performance across domains. By re-weighting the loss based on

Table 2: Overview of mammography datasets with available scans after selection and splitting. The number of samples in the training sets is shown in parentheses.

	CBIS-DDSM	**RSNA**	**INbreast**	**MIAS**	**CMMD**	**VinDR**	**CDD-CESM**	**KAU-BCMD**	**MMD**	**LBMD**
Country	US	US/AU	Portugal	UK	China	Vietnam	Egypt	KSA	Iraq	XXX
Sites	4	2	1		2	2	1	1	4	1
Patients	1,391	1,970	115	165	1,277	930	326	442	745	696
Scans	2,844	9,594	410	322	2,742	3,709	1,003	1,774	745	3,090
Age (y)	*N/A*	59 ± 11	*N/A*	*N/A*	47 ± 11	44 ± 12	50 ± 12	49 ± 7	*N/A*	58 ± 11
Diagnosis	✓	✓	-	✓	✓	-	✓	-	✓	✓
Benign	1,253 (875)	1,487 (1,039)		64 (47)	1,102 (774)		331 (252)		0	1,993 (1,392)
Malignant	1,220 (860)	1,069 (734)		52 (31)	1,640 (860)		331 (239)		125 (88)	2 (2)
Density	✓	✓	✓	✓	-	✓	✓	✓	-	✓
A	396 (273)	529 (377)	136 (97)	106 (82)		12 (12)	8 (8)	577 (399)		375 (231)
B	1,103 (760)	2,789 (1,904)	146 (108)	104 (70)		337 (226)	329 (247)	827 (600)		1,050 (746)
C	879 (633)	2,861 (2,095)	99 (64)	112 (77)		2,852 (1,968)	515 (315)	332 (208)		1,012 (717)
D	464 (325)	343 (209)	28 (17)	0		508 (390)	70 (62)	108 (80)		180 (135)

each domain's $\mathcal{D}$ performance, `GroupDRO` ensures that the model does not favor majority groups at the expense of under-represented ones [28].
Mixture-of-Experts (`MOE`) uses a set of expert classifiers $f_\theta^e(\cdot)$, each specializing in different domains, and combines their outputs through a gating mechanism $\alpha_e(x)$. This allows the model to adaptively leverage domain-specific expertise while benefiting from a shared representation, as described in [19].

3 Experiments and Results

Foundation Models. We consider several FMs drawn from recent reviews [25, 21]. **MammoCLIP** [11] was trained on 25,355 mammograms from the UPMC dataset using contrastive multi-view learning and yields 2,048-dimensional features via its EN-B5 encoder. In contrast, **MedCLIP** [32] and **GLORIA** [15] were developed on 500,000 and 200,000 X-ray images respectively, both employing a ResNet-50 backbone to produce 512-dimensional embeddings. Additionally, **CLIP** [26] was trained on 400 million internet-sourced image-text pairs with contrastive learning, while **DINOv2** [24] uses a self-distillation framework on 142 million images to generate lightweight representations of size 384.

Datasets. We use mammography datasets from diverse countries and institutions, ensuring a representative analysis. Our collection includes four prominent datasets: the Digital Database for Screening Mammography **(CBIS-DDSM)** [29] from the USA, the RSNA Screening Mammography Breast Cancer Detection Dataset **(RSNA)** [7] from the USA and Australia, **INbreast** [22] from Portugal, and Mammographic Image Analysis Society **(MIAS)** [31] from the UK. To further capture diversity and address the under-representation of certain regions, we integrated datasets including the Chinese Mammography Database **(CMMD)** [5], **VinDr-Mammo** [23] from Vietnam, the Categorized Digital Database for Low Energy and Subtracted Contrast Enhanced Spectral Mammography images **(CDD-CESM)** [18] from Egypt, the King Abdulaziz University Breast Cancer Mammogram Dataset **(KAU-BCMD)** [1] from Saudi Arabia, and the Mammogram Mastery dataset **(MMD)** [2] from Iraq. Additionally, we

incorporated the Lebanese Breast Mammography Dataset **(LBMD)**, an internally curated collection co-developed with our clinical partners at the Lebanese Hospital Geitaoui and assembled exclusively for this project; all cases within the **LBMD** are biopsy-confirmed, and the data collection protocol received full ethical approval for use in this work.

Sample selection. Our combination of datasets was initially highly imbalanced, with some datasets containing over 50,000 samples (*e.g.* RSNA) while others had as few as 300 (*e.g.* MIAS). Additionally, the class imbalance was significant within datasets; for instance, 75% of VinDR samples belong to density class C. To minimize these imbalances and focus on dataset biases, we applied a sample selection strategy to have more balanced classes. Note that in real-world clinical settings, datasets are often highly imbalanced, potentially amplifying the observed biases. First, we dropped samples with no labels for diagnosis or density class, *i.e.* the two classification tasks investigated. All labels were aggregated from the original metadata, where benign and malignant classes were biopsy-confirmed in most datasets. We categorized patients into three diagnosis classes: healthy, benign, and malignant. We capped each class at 1,000 patients, randomly sampling when necessary while retaining all available patients in smaller classes. For VinDR, as this dataset did not contain the diagnosis information, we applied our sampling selection strategy at the density level. Finally, datasets were split at the patient level into training (70%) and test (30%) sets, ensuring no data leakage. Table 2 provides more details on the composition of each dataset after the sample selection strategy.

Implementation details. Images were preprocessed using the framework proposed by [11]. We used a rule-based approach to crop images according to the breast ROI. We set values less than 40 to 0 and eliminated consistently identical rows and columns, supposing these denote background. The final images had a size of 1,520×912. Experiments were implemented in Python v3.10 using Pytorch v2.4.1. Individual classifiers were trained on each dataset, and `Unified` on the aggregated datasets for two tasks: diagnosis and breast density. We searched for optimal batch size (8, 16, or 32) and learning rate (1e-3, 1e-4, or 1e-5) using 3-fold cross-validation within the training set. Hyperparameters giving the best accuracy after 20 epochs were then used for training on the whole set for 50 epochs. We used the same hyperparameter optimization for all mitigation strategies, except for `FADES`. Due to its computational cost, we fixed the batch size to 32, the learning rate to 1e-4, and trained for 30 epochs. Technical details specific to each strategy are reported in the code and will be publicly available upon acceptance. We computed differences between F1 score distributions across datasets using a one-sided Wilcoxon test for statistical significance. To evaluate the classifiers' fairness, we computed Equal Opportunity Difference (EOD) and Average Odds Difference (AOD) across subgroups $g_1, g_2 \in \mathcal{G}$ and labels $k \in \mathcal{Y}$: $\text{EOD} = \Big(P(\hat{Y} = k \mid Y = k, G = g_1) - P(\hat{Y} = k \mid Y = k, G = g_2)\Big)$, and $\text{AOD} = \text{EOD} + (P(\hat{Y} = k \mid Y \neq k, G = g_1) - P(\hat{Y} = k \mid Y \neq k, G = g_2)$.

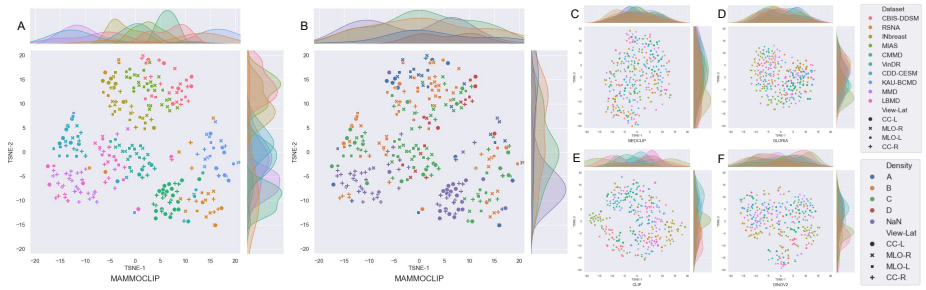

Fig. 1: t-SNE visualization of MammoCLIP, color-coded by dataset (**A**) and density (**B**), and FMs: GLORIA (**C**), MedCLIP (**D**), CLIP (**E**), DINOv2 (**F**).

3.1 Exploring the feature embeddings

Fig. 1 presents a t-SNE-based analysis of features extracted from FMs. The results illustrate distinct clustering behaviors, reflecting how each FM encodes mammography-specific characteristics. There is no visible clustering according to view, aligning with MammoCLIP's multi-view learning strategy. For other FMs, this suggests their ability to learn view-invariant features, likely through data augmentation. MammoCLIP exhibits well-defined clusters, with features from the same dataset tightly grouped, suggesting strong dataset-specific encoding. Smoother patterns emerge for CLIP and DINOv2, where at least one t-SNE component captures dataset-specific information. Features from MedCLIP and GLORIA are widely dispersed, with no clear dataset-specific clustering. These models seem to learn more generalized feature representations, likely due to pre-training on diverse medical images. Interestingly, breast density attributes seem to impact feature distributions along the t-SNE components, with a smooth transition from low- to high-dense breasts in MammoCLIP features.

3.2 Robustness of classifiers to domain-shift

Tab. 3 shows the performance of classifiers trained using features extracted from each FM for diagnosis and breast density classification. For both tasks, `Unified` classifiers trained on features from MammoCLIP outperform those based on GLORIA ($p < 0.05$, average relative improvement of $+15.3\%$), CLIP ($p < 0.01$, $+9.7\%$), and DinoV2 ($p < 0.01$, $+13.3\%$), highlighting the advantages of pre-training on modality-specific data compared to domain-related (*i.e.* X-rays) or natural images. It is worth mentioning that MedCLIP-based classifiers exhibit notably poor performance across all tasks, suggesting that the extracted features may be predominantly noisy. Overall, classifiers have high performance when tested on the same dataset they were trained on (*Indiv. (internal)*), with average F1 scores of 0.73 and 0.53 for MammoCLIP on diagnosis and breast density, respectively. However, when tested on other datasets (*Indiv. (external)*), a drastic drop in F1 scores (up to -50% for diagnosis) was remarkable, potentially due to overfitting of dataset-specific characteristics encoded in MammoCLIP's features.

Table 3: Weighted F1-score of classifiers: mean ± standard deviation across datasets, overall best performance is in **bold**. Stars indicate significantly different performance: red stars for lower than MammoCLIP, green stars for higher than Unified, blue stars for lower than Unified. * = $p < 0.01$, ** = $p < 0.05$.

Diagnosis					
	MammoCLIP	MedCLIP	GLORIA	CLIP	DinoV2
Indiv. (internal)	0.73 ± 0.11	0.37 ± 0.21	0.68 ± 0.13	0.61 ± 0.14	0.62 ± 0.12
Indiv. (external)	0.32 ± 0.11	0.18 ± 0.07	0.28 ± 0.08	0.24 ± 0.10	0.32 ± 0.10
Indiv. (overall)	0.37 ± 0.22*	0.25 ± 0.18*	0.38 ± 0.19*	0.34 ± 0.2*	0.39 ± 0.19*
`Unified`	**0.65 ± 0.14**	0.32 ± 0.27**	0.58 ± 0.13*	0.56 ± 0.14**	0.57 ± 0.16**
`DANN` [10]	0.54 ± 0.14	0.18 ± 0.16	0.42 ± 0.14	0.46 ± 0.11	0.47 ± 0.07
`FairDisCO` [9]	0.63 ± 0.14	0.18 ± 0.16	0.43 ± 0.14	0.56 ± 0.15	0.54 ± 0.18
`FADES` [16]	0.62 ± 0.16	0.39 ± 0.17	0.51 ± 0.17	0.54 ± 0.19	0.56 ± 0.15
`MOE` [19]	0.64 ± 0.13	0.38 ± 0.04	0.43 ± 0.13	0.49 ± 0.15	0.55 ± 0.16
`GroupDRO` [28]	0.57 ± 0.14	0.32 ± 0.27	0.49 ± 0.15	0.47 ± 0.16	0.54 ± 0.16
Density					
	MammoCLIP	MedCLIP	GLORIA	CLIP	DinoV2
Indiv. (internal)	0.53 ± 0.19	0.32 ± 0.18	0.41 ± 0.19	0.41 ± 0.19	0.50 ± 0.13
Indiv. (external)	0.42 ± 0.08	0.2 ± 0.04	0.31 ± 0.10	0.33 ± 0.12	0.4 ± 0.07
Indiv. (overall)	0.41 ± 0.16*	0.22 ± 0.15*	0.32 ± 0.16*	0.33 ± 0.16*	0.4 ± 0.14*
`Unified`	0.59 ± 0.17	0.19 ± 0.08**	0.5 ± 0.15*	0.5 ± 0.16**	0.51 ± 0.15**
`DANN` [10]	0.56 ± 0.16	0.19 ± 0.08	0.47 ± 0.14	0.5 ± 0.14	0.48 ± 0.11
`FairDisCO` [9]	0.54 ± 0.17	0.19 ± 0.08	0.47 ± 0.14	0.49 ± 0.14	0.48 ± 0.12
`FADES` [16]	0.57 ± 0.16	0.19 ± 0.08	0.49 ± 0.15	0.54 ± 0.16*	0.51 ± 0.15
`MOE` [19]	0.55 ± 0.17	0.19 ± 0.08	0.42 ± 0.13	0.46 ± 0.13	0.5 ± 0.16
`GroupDRO` [28]	**0.66 ± 0.08***	0.19 ± 0.08	0.52 ± 0.11	0.57 ± 0.06	0.56 ± 0.08

3.3 Effectiveness of bias and domain-adaptation strategies

`Unified` classifiers show performance similar to individual classifiers on their test sets (internal), with F1 score variations of ±15%. Aggregating datasets effectively improves generalization compared to individual classifiers (overall F1 score improved by +75% and +49%, $p < 0.05$ with `Unified` vs. Indiv.). However, `Unified` classifiers exhibit performance disparities, with F1 score standard deviations of 0.15 across test datasets, suggesting that such aggregation cannot fully mitigate biases. `DANN` shows slightly lower overall F1 scores than `Unified` classifiers, especially for the diagnosis task. While this technique aims to learn domain-invariant representations, it seems to do so at the cost of overall performance. The mutual information (MI) between features and labels ($MI = 0.07$) is lower than that between features and datasets ($MI = 0.26$). By disentangling dataset-related features, `DANN` may thus inadvertently discard task-relevant information. In addition, this technique might fail in case of severe dataset shift: here, the mean pairwise Wasserstein distance between datasets' features increases from 5.75 before adaptation to 101.39 after, and a similar trend is observed between labels' features, suggesting the assumption of covariate shift is not met. Similar observations can be made for `FairDisCO`, `FADES`, and `MOE`. For Breast density, `GroupDRO` produces consistently tighter F1 score distributions than `Unified` classifiers with standard deviations of 0.08, 0.06, and 0.08 for MammoCLIP, CLIP, and DinoV2, respectively, indicating reductions of disparities across test datasets. Additionally, it outperforms the `Unified` classifier across FMs, *e.g.* MammoCLIP with a relative improvement of 12% ($p < 0.01$). For diagnosis,

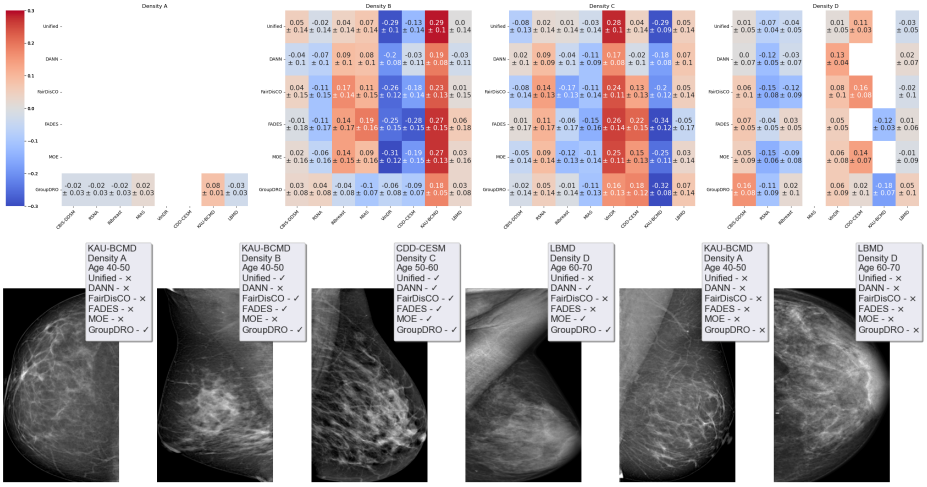

Fig. 2: (*Top*) AOD scores across datasets for MammoCLIP on breast density. (*Bottom*) From left to right: samples from different density groups from A to D, and from under-represented subgroups (density A, age < 40 and density D, age >70)). ✓ indicates correctness, × represents misclassification.

mitigation strategies do not improve overall performance nor reduce disparities compared to `Unified`, likely due to variability in diagnostic label availability across datasets, *e.g.*, no benign samples for MMD.

3.4 Bias and domain-adaptation in under-represented subgroups

Certain subgroups are under-represented in our datasets, *e.g.* breast density classes A and D representing 11 and 9% of the dataset and age <40 (10%) and >70 (10%), and are unequally represented across datasets (see Tab. 2). Fig.2 (top) illustrates prediction disparities across datasets and breast density classes, where $AOD \simeq 0$ indicates fair performance. `GroupDRO` and `DANN` achieve the most fair performance for all breast density classes ($AOD_{max} \simeq 0.2$, $AOD_{min} \simeq -0.3$ and $AOD_{avg.} \simeq 0$ across breast density classes and datasets), aligning with `DANN`'s domain-invariant feature learning strategy. However, for `DANN`, this fairness comes at the cost of performance (see Tab. 3). `GroupDRO` stabilizes performance across breast density classes, notably improving prediction for class A, which other classifiers struggled with. This ability to learn across domains, while favoring under-represented subgroups, is critical for extreme breast densities (A, D) and age groups due to their strong interplay [8], their association with breast cancer risk and their influence on the sensitivity of mammography [30]. Fig. 2 (bottom) presents samples from different subgroups and classifiers' successes and failures in breast density classification. Variations in contrast, texture, and patterns across classes and datasets may introduce spurious correlations, un-

derscoring the need for fairness-aware strategies.`GroupDRO` seems to effectively mitigate biases and could be further refined with more fine-grained attributes.

Conclusion

This paper explores biases in FM for breast mammography classification. Our analysis reveals that modality-specific pre-training of FM is beneficial for performance, but individual classifiers still fail to generalize well beyond their training data. Aggregating datasets enhances overall performance, emphasizing the need for broader dataset contributions. However, this strategy is insufficient to mitigate biases, resulting in disparities across underrepresented subgroups. Domain-adaptation strategies address these disparities, but often at the cost of performance, suggesting that general-purpose FMs may encode more image- than task-related information. On the other hand, fairness-aware techniques ensure equitable performance across underrepresented subgroups and classes. These findings have significant implications for deploying AI-driven mammography analysis in clinical practice. Such shortcuts can lead to biased models, over-predicting each dataset's majority class, a critical situation in case of high negative rates or poor accuracy on underrepresented populations. Future works will investigate the root causes of these shortcuts, paving the way for new fairness constraints and bias mitigation strategies for FMs training and fine-tuning.

Acknowledgments SA would like to acknowledge the partial support from the German Academic Exchange Service (DAAD) with funds from the Federal Foreign Office (AA). It was also developed within the interdisciplinary framework of the Arab-German Young Academy of Sciences and Humanities (AGYA), which is funded by the German Federal Ministry of Education and Research (BMBF) under grant 01DL20003. The authors thank their colleagues at the Lebanese Hospital Geitaoui, Beyrouth, Lebanon for their support.

Disclosure of interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alsolami, A.S., Shalash, W., Alsaggaf, W., et al.: King Abdulaziz University Breast Cancer Mammogram Dataset (KAU-BCMD). Data **6**(11), 111 (2021). https://doi.org/10.3390/data6110111
2. Aqdar, K.B., Abdalla, P.A., Mustafa, R.K., et al.: Mammogram Mastery: A Robust Dataset for Breast Cancer Detection and Medical Education **1** (2024). https://doi.org/10.17632/fvjhtskg93.1
3. Bommasani, R., Hudson, D.A., Adeli, E., et al.: On the opportunities and risks of foundation models (2022)
4. Bray, F., Laversanne, M., Sung, H., et al.: Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **74**(3), 229–263 (2024)

5. Cai, H., Wang, J., Dan, T., et al.: An Online Mammography Database with Biopsy Confirmed Types. Scientific Data **10**(1), 123 (2023). https://doi.org/10.1038/s41597-023-02025-1
6. del Carmen, M.G., Halpern, E.F., Kopans, D.B., et al.: Mammographic Breast Density and Race. American Journal of Roentgenology **188**(4), 1147–1150 (2007). https://doi.org/10.2214/AJR.06.0619
7. Carr, C., Kitamura, F., Kalpathy-Cramer, J., et al.: Rsna screening mammography breast cancer detection. 2022
8. Checka, C.M., Chun, J.E., Schnabel, F.R., et al.: The Relationship of Mammographic Density and Age: Implications for Breast Cancer Screening. American Journal of Roentgenology **198**(3), W292–W295 (2012). https://doi.org/10.2214/AJR.10.6049
9. Du, S., Hers, B., Bayasi, N., et al.: Fairdisco: Fairer ai in dermatology via disentanglement contrastive learning. In: European Conference on Computer Vision. pp. 185–202 (2022)
10. Ganin, Y., Ustinova, E., Ajakan, H., et al.: Domain-adversarial training of neural networks. Journal of machine learning research **17**(59), 1–35 (2016)
11. Ghosh, S., Poynton, C.B., Visweswaran, S., et al.: Mammo-CLIP: A Vision Language Foundation Model to Enhance Data Efficiency and Robustness in Mammography. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 632–642 (2024). https://doi.org/10.1007/978-3-031-72390-2_59
12. Girdhar, R., El-Nouby, A., Liu, Z., et al.: ImageBind One Embedding Space to Bind Them All. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15180–15190 (2023). https://doi.org/10.1109/CVPR52729.2023.01457
13. Glocker, B., Jones, C., Roschewitz, M., et al.: Risk of Bias in Chest Radiography Deep Learning Foundation Models. Radiology: Artificial Intelligence **5**(6), e230060 (2023). https://doi.org/10.1148/ryai.230060
14. Heller, S.L., Hudson, S., Wilkinson, L.S.: Breast density across a regional screening population: effects of age, ethnicity and deprivation. The British Journal of Radiology **88**(1055), 20150242 (2015). https://doi.org/10.1259/bjr.20150242
15. Huang, S.C., Shen, L., Lungren, M.P., et al.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3942–3951 (2021)
16. Jang, T., Wang, X.: Fades: Fair disentanglement with sensitive relevance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12067–12076 (2024)
17. Jin, R., Xu, Z., Zhong, Y., et al.: FairmedFM: Fairness benchmarking for medical imaging foundation models. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024)
18. Khaled, R., Helal, M., Alfarghaly, O., et al.: Categorized contrast enhanced mammography dataset for diagnostic and artificial intelligence research. Scientific Data **9**(1), 122 (2022). https://doi.org/10.1038/s41597-022-01238-0
19. Li, B., Shen, Y., Yang, J., et al.: Sparse mixture-of-experts are domain generalizable learners. In: The Eleventh International Conference on Learning Representations (2023)
20. McKinney, S.M., Sieniek, M., Godbole, V., et al.: International evaluation of an ai system for breast cancer screening. Nature **577**(7788), 89–94 (2020)

21. Moor, M., Banerjee, O., Abad, Z.S.H., et al.: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023). https://doi.org/10.1038/s41586-023-05881-4
22. Moreira, I.C., Amaral, I., Domingues, I., et al.: INbreast: Toward a Full-field Digital Mammographic Database. Academic Radiology **19**(2), 236–248 (2012). https://doi.org/10.1016/j.acra.2011.09.014
23. Nguyen, H.T., Nguyen, H.Q., Pham, H.H., et al.: VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. Scientific Data **10**(1), 277 (2023). https://doi.org/10.1038/s41597-023-02100-7
24. Oquab, M., Darcet, T., Moutakanni, T., et al.: DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research (2023)
25. Paschali, M., Chen, Z., Blankemeier, L., et al.: Foundation Models in Radiology: What, How, Why, and Why Not. Radiology **314**(2), e240597 (2025). https://doi.org/10.1148/radiol.240597
26. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8748–8763 (2021)
27. Ricci Lara, M.A., Echeveste, R., Ferrante, E.: Addressing fairness in artificial intelligence for medical imaging. Nature Communications **13**(1), 4581 (2022). https://doi.org/10.1038/s41467-022-32186-3
28. Sagawa, S., Koh, P.W., Hashimoto, T.B., et al.: Distributionally robust neural networks. In: International Conference on Learning Representations (2019)
29. Sawyer-Lee, R., Gimenez, F., Hoogi, A., et al.: Curated breast imaging subset of digital database for screening mammography (cbis-ddsm) [data set] (2016). https://doi.org/https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY
30. Sinclair, N., Littenberg, B., Geller, B., Muss, H.: Accuracy of Screening Mammography in Older Women. American Journal of Roentgenology **197**(5), 1268–1273 (2011). https://doi.org/10.2214/AJR.10.5442
31. Suckling, J., Parker, J., Dance, D., et al.: Mammographic Image Analysis Society (MIAS) database v1.21 (2015). https://doi.org/10.17863/CAM.105113
32. Wang, Z., Wu, Z., Agarwal, D., et al.: MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing **2022**, 3876–3887 (2022). https://doi.org/10.18653/v1/2022.emnlp-main.256