AdFair-CLIP: Adversarial Fair Contrastive Language-Image Pre-training for Chest X-rays

Chenlang Yi[1], Zizhan Xiong[1], Qi Qi[2], Xiyuan Wei[1], Girish Bathla[3], Ching-Long Lin[2], Bobak J. Mortazavi[1], and Tianbao Yang[1⋆]

[1] Texas A&M University, College Station, TX, USA
{baileyyeah, tianbao-yang}@tamu.edu
[2] The University of Iowa, Iowa City, IA, USA
[3] Mayo Clinic School of Medicine, Rochester, MN, USA

Abstract. Contrastive Language-Image Pre-training (CLIP) models have demonstrated superior performance across various visual tasks including medical image classification. However, fairness concerns, including demographic biases, have received limited attention for CLIP models. This oversight leads to critical issues, particularly those related to race and gender, resulting in disparities in diagnostic outcomes and reduced reliability for underrepresented groups. To address these challenges, we introduce AdFair-CLIP, a novel framework employing adversarial feature intervention to suppress sensitive attributes, thereby mitigating spurious correlations and improving prediction fairness. We conduct comprehensive experiments on chest X-ray (CXR) datasets, and show that AdFair-CLIP significantly enhances both fairness and diagnostic accuracy, while maintaining robust generalization in zero-shot and few-shot scenarios. These results establish new benchmarks for fairness-aware learning in CLIP-based medical diagnostic models, particularly for CXR analysis.

Keywords: Fairness · CLIP Models · Chest X-ray Diagnosis.

1 Introduction

As artificial intelligence (AI) systems increasingly inform clinical decision-making, concerns have arisen over biases in AI algorithms [23, 19, 15]. These biases can disproportionately disadvantage underrepresented populations, reinforcing healthcare disparities. Ensuring fairness in medical AI is crucial to mitigating demographic biases and promoting equitable healthcare outcomes.

CLIP models, integrating visual and textual modalities, show promise in medical AI for disease diagnosis and report generation [13, 30, 9, 29]. In the CXR domain, models like GLoRIA [12], CXR-CLIP [27], and MedCLIP [25] have achieved strong diagnostic performance via contrastive learning on large image-text datasets. However, prior fairness research has focused on vision-only models [10, 17, 22, 5], while fairness in CLIP-based methods remains largely unexplored. Ensuring fairness in these models is essential to preventing biased

⋆ Corresponding author.

outcomes, underscoring the need for a comprehensive fairness investigation. Existing fairness-aware methods often face trade-offs or limited generalizability. CLIP-clip [24] mitigates bias by pruning bias-associated dimensions but inadvertently removes task-relevant features. FairCLIP [18] introduces similarity-level interventions via Sinkhorn divergence minimization. However, these adjustments remain superficial and fail to fully eliminate biases in feature representations. DebiasCLIP [3] employs a frozen backbone strategy, modifying only cross-modal similarity scores. While effective for zero-shot classification, this approach lacks adaptability to downstream tasks, limiting its debiasing impact.

To overcome these limitations, we propose AdFair-CLIP, an adversarial fairness framework that integrates CLIP [21] with adversarial learning to mitigate demographic biases while preserving multimodal representation learning. Unlike prior methods, AdFair-CLIP operates directly in the feature space, formulates bias mitigation as a minimax optimization problem [8], and reduces spurious correlations between demographic attributes and model predictions, allowing the model to focus on disease-relevant features. We evaluate our approach on CheXpert Plus [6] and MIMIC-CXR [16], assessing fairness across race and gender. Experimental results demonstrate substantial improvements in both fairness and diagnostic accuracy under zero-shot and few-shot conditions.

Our main contributions are as follows: (1) We conduct a systematic fairness investigation of state-of-the-art (SOTA) CLIP models for CXR diagnosis, revealing significant biases across demographics. (2) We propose AdFair-CLIP, a novel framework employing adversarial feature-space intervention to mitigate these biases and address fairness in CLIP-based CXR diagnostic models. (3) We demonstrate AdFair-CLIP's effectiveness through comprehensive evaluations, showing substantial improvements in both fairness and diagnostic accuracy.

2 Method

2.1 Fairness Investigation

Dataset imbalances in CXR analysis can introduce biases [5, 28], leading CLIP-based models to develop biased representations. We take racial bias as an example, which stems from two key sources. First, CXR datasets predominantly consist of White patients, with Black and Asian populations underrepresented. As a result, models learn more informative features for the majority group while extracting less meaningful representations for underrepresented groups, resulting in predictive disparities. Second, disease prevalence differs across racial groups. In the CheXpert Plus [6] dataset, Cardiomegaly is more prevalent in Asian patients than in Black and White patients. This imbalance may cause models to associate disease likelihood with racial features rather than pathology, reinforcing biases. These issues affect both zero-shot inference and downstream tasks, producing diagnostic disparities. To systematically evaluate these concerns, we analyze SOTA CLIP-based CXR diagnostic models using multiple fairness metrics, uncovering significant biases (detailed in Experiment & Analysis).

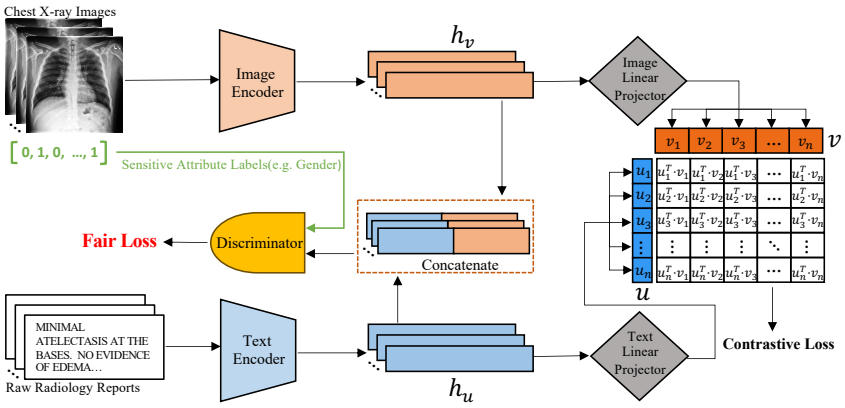

Fig. 1. Overview of the AdFair-CLIP architecture. Image and text encoders extract representations h_v and h_u from chest X-ray images and radiology reports, projecting them to v and u for contrastive alignment. A discriminator predicts sensitive attributes from concatenated representations, enabling adversarial training to mitigate biases.

2.2 AdFair-CLIP

To address fairness challenges, we propose AdFair-CLIP, a min-max adversarial optimization framework that mitigates demographic bias while preserving multimodal alignment. The overview of the method is shown in Fig. 1.

Multi-Modal Alignment. We adopt ResNet50 [11] as the image encoder $E_v(\cdot)$ and BioClinicalBERT [2] as the text encoder $E_t(\cdot)$. Given a chest X-ray image I and a radiology report T, the encoders extract intermediate features $h_v = E_v(I) \in \mathbb{R}^{2048}$ and $h_u = \text{MeanPool}(E_t(T), M) \in \mathbb{R}^{768}$. These are projected into a shared space and ℓ_2-normalized to obtain embeddings $v, u \in \mathbb{R}^{512}$. To align image and text representations, we employ the InfoNCE loss [20]. The loss function $\mathcal{L}_{\text{GCL}}(w_v, w_u)$ is defined as follows:

$$-\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp\left(\text{sim}(v_i,u_i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}(v_i,u_j)/\tau\right)}+\log\frac{\exp\left(\text{sim}(u_i,v_i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}(u_i,v_j)/\tau\right)}\right], \quad (1)$$

where w_v and w_u are the parameters of the vision and text encoders, respectively, $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and τ is a learnable temperature parameter.

Fair Loss. We design a discriminator to predict sensitive attributes. It is a multi-layer perceptron denoted as $D(\cdot)$ that processes the concatenated intermediate representations $h_v \in \mathbb{R}^{2048}$ and $h_u \in \mathbb{R}^{768}$. The combined vector $h_{\text{concat}} = [h_v; h_u] \in \mathbb{R}^{2816}$ is passed through $D(\cdot)$, producing an output $s = D(h_{\text{concat}}) \in \mathbb{R}^C$, where C is the number of sensitive attribute categories. To address sample imbalance in sensitive attributes, particularly for underrepresented groups, a weighted fair entropy loss is applied to ensure fair classification across all categories. The fairness loss is computed over a batch of N samples, based on the discriminator's output logits $\mathbf{s}^{(k)} = [s_1^{(k)}, \ldots, s_C^{(k)}] \in \mathbb{R}^C$

for each sample k. Let the ground-truth one-hot vector for the k-th sample be $\mathbf{y}^{(k)} = [y_1^{(k)}, \ldots, y_C^{(k)}] \in \mathbb{R}^C$. The class weights $\mathbf{w} = [w_1, \ldots, w_C]$, where $w_i = 1/n_i$ and n_i denotes the number of samples in category i in a batch, are used to balance sample disparities. The weighted fair loss is defined as:

$$\mathcal{L}_{\text{Fair}}(w_v, w_u, w_{\text{d}}) = -\frac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{C} w_i y_i^{(k)} \log \left(\frac{\exp(s_i^{(k)})}{\sum_{j=1}^{C} \exp(s_j^{(k)})} \right), \tag{2}$$

where w_{d} denotes the parameter of the discriminator network.

Adversarial Optimization. Our method is based on adversarial training, formulated as a minimax optimization problem [8]. The discriminator aims to maximize the accuracy of predicting sensitive attributes from the learned representations, while the encoders seek to minimize the contrastive loss and reduce the accuracy of predicting sensitive attributes. This encourages the encoder to generate representations invariant to sensitive attributes. The final optimization objective is expressed as:

$$\min_{w_v, w_u} \max_{w_{\text{d}}} \mathcal{L}_{\text{GCL}}(w_v, w_u) - \alpha \cdot \mathcal{L}_{\text{Fair}}(w_v, w_u, w_{\text{d}}), \tag{3}$$

where the hyperparameter α balances the trade-off between multimodal alignment and fairness regularization.

3 Experiment & Analysis

3.1 Datasets

We use CheXpert Plus [6] and MIMIC-CXR [16] for training, and evaluate exclusively on a demographically balanced test set sampled from CheXpert Plus.

CheXpert Plus and MIMIC-CXR. CheXpert Plus extends CheXpert [14] for multimodal learning and fairness research, providing CXR images with radiology reports. We use the "impression" section as textual input, pre-train on frontal image-text pairs, and reserve expert-annotated samples for validation. We focus on frontal chest radiographs with 191,071 image-text pairs, using 3,000 randomly sampled training images for validation. MIMIC-CXR is a large-scale CXR dataset with radiology reports. We extract the "findings" and "impression" sections as text, pre-train on the training split, and retain a subset for validation. Further details are available in [30]. Both datasets include demographic metadata, enabling fairness assessment in model performance.

Fair Test Dataset. Previous CLIP-based models for CXR diagnosis have primarily used the CheXpert 5x200 [12] dataset for evaluation, which underrepresents certain demographic groups. To ensure a fair evaluation, we construct a demographically balanced test set sampled from CheXpert Plus, explicitly stratified by race and gender. However, balancing demographic representation while adhering to the strict criteria in [12], which require each image to contain positive labels for only one condition, limits the test dataset to 450 images across five disease categories (90 per class). This balanced test dataset enables a more rigorous assessment of models' fairness and is named FCXP 5x90.

3.2 Baseline

Our baselines span multiple model categories. To assess fairness in CLIP-based CXR diagnostic methods, we compare GLoRIA [12], an attention-based framework for learning global and local representations, CXR-CLIP [27], mitigating data scarcity via synthetic image-text generation and multi-view contrastive supervision, and MedCLIP [25], leveraging unpaired data and medical knowledge to reduce false negatives. Additionally, we include the original CLIP [21] for comparison. To validate AdFair-CLIP's effectiveness, we benchmark it against two fairness-aware CLIP methods: FairCLIP [18] and DebiasCLIP [3].

3.3 Experimental Setup

In this section, we outline the pre-training and evaluation strategies, as well as the metrics used to thoroughly assess predictive accuracy and fairness.

Pre-Training and Evaluation. All models are independently pre-trained on CheXpert Plus and MIMIC-CXR, with image encoders initialized from ResNet-50 [11] and text encoders from BioClinicalBERT [2], following the official checkpoints. To systematically evaluate model performance, we assess each pre-trained image encoder under two settings: zero-shot classification and few-shot classification. In the zero-shot setting, classification is framed as an image-text similarity task, where labels are converted into textual prompts following [12]. For few-shot classification, we freeze the pre-trained image encoder and train a linear classifier on top of it using only 10% of the labeled training data from CheXpert Plus. All evaluations are conducted on FCXP 5x90, assessing fairness across race (Black, White, Asian) and gender (Male, Female). This setup allows us to assess both the models' performance on the in-domain and out-of-domain generalization.

Metrics. To thoroughly evaluate the effectiveness of our proposed method, we employ both performance and fairness metrics. For performance evaluation, we use the accuracy and Area Under the Curve (AUC). To assess fairness, we incorporate multiple fairness-aware metrics. Demographic Parity Difference (DPD) [1] measures the maximum absolute difference in the probabilities of positive predictions across groups. Difference in Equalized Odds (DEOdds) [1] quantifies the maximum disparity in True Positive Rates (TPR) and False Positive Rates (FPR) separately across groups. Additionally, Group-wise AUC (GAUC) [26] measures the probability that a randomly selected sample from one demographic group receives a higher predicted score than a randomly selected sample from another. Inter-Group AUC (Inter-AUC) [4] quantifies the degree to which the ranking of positive samples from one demographic group relative to negative samples from another group remains independent of group membership. Intra-Group AUC (Intra-AUC) [4] evaluates the consistency of ranking within each demographic group by comparing the ranking of positive and negative samples within the same group.

Implementation Details. Chest X-ray images are resized to 256×256 and cropped to 224×224. Radiology reports are tokenized using BioClinicalBERT [2]. Pre-training uses the Adam optimizer [7] with a learning rate of 1×10^{-5}, batch

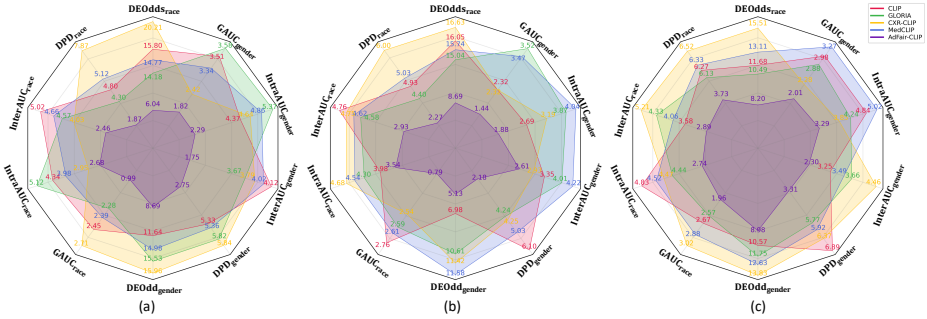

Fig. 2. Fairness assessment of SOTA CLIP-based CXR diagnostic methods across race and gender in three scenarios, with all scores presented in percentage.

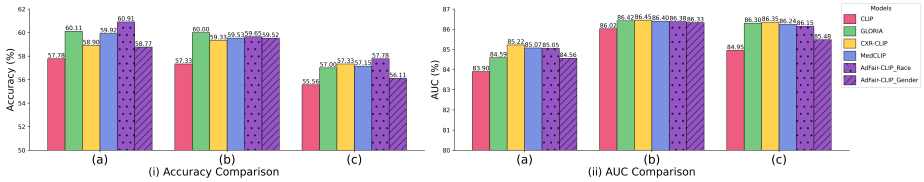

Fig. 3. Performance assessment of SOTA CLIP-based CXR diagnostic methods across race and gender in three scenarios, with all scores presented in percentage.

size 36, and weight decay 1×10^{-6}, adjusted by ReduceLROnPlateau. The contrastive loss temperature τ is 0.1, and the fairness regularization coefficient α is 0.3. For linear evaluation, the learning rate is 1×10^{-4} with batch size 64. All experiments are implemented in PyTorch on six NVIDIA A100 GPUs.

3.4 Results Analysis

We evaluate three scenarios: (a) Zero-shot: models pre-trained on CheXpert Plus, (b) Few-shot: models pre-trained on CheXpert Plus and fine-tuned on 10% of its data, and (c) Transfer learning: models pre-trained on MIMIC-CXR and fine-tuned on 10% of the labeled data from CheXpert Plus. All results are averaged over three random seeds, with standard deviations ranging from 0.22% to 5.18%.

AdFair-CLIP vs. CLIP-based CXR Diagnostic Methods. The fairness results in Fig. 2 reveal significant and persistent biases in CLIP-based CXR diagnostic models across race and gender groups. Each axis represents a fairness metric for race or gender, where lower values indicate better fairness and a larger enclosed area signifies greater disparities. GLoRIA, CXR-CLIP, MedCLIP, and CLIP exhibit substantial fairness gaps in multiple metrics, demonstrating limited effectiveness in ensuring equitable predictions. These biases persist across all evaluation scenarios, underscoring the need for stronger fairness interventions in medical AI. In contrast, AdFair-CLIP consistently achieves lower disparities across all metrics, effectively addressing fairness limitations in existing models.

Table 1. Zero-shot comparison of AdFair-CLIP with baselines, pre-trained on CheXpert Plus, with all scores presented in percentage.

Attribute	Method	Acc	AUC	DPD	DEOdds	GAUC	Inter-AUC	Intra-AUC
Race	CLIP	57.78	83.90	4.80	15.80	2.45	5.02	4.34
	FairCLIP	60.89	**85.08**	3.20	8.99	1.64	3.40	3.15
	DebiasCLIP	59.33	84.61	2.40	9.65	1.73	3.41	2.87
	AdFair-CLIP	**60.91**	85.05	**1.87**	**6.04**	**0.99**	**2.46**	**2.68**
Gender	CLIP	57.78	83.90	5.33	11.64	3.51	4.12	4.37
	FairCLIP	58.34	84.37	3.12	10.28	2.51	3.29	3.27
	DebiasCLIP	56.00	84.01	3.67	10.00	**1.67**	1.83	2.34
	AdFair-CLIP	**58.77**	**84.56**	**2.75**	**8.69**	1.82	**1.75**	**2.29**

Table 2. Few-shot (10%) comparison of AdFair-CLIP with baselines, pre-trained and fine-tuned on CheXpert Plus, with all scores presented in percentage.

Attribute	Method	Acc	AUC	DPD	DEOdds	GAUC	Inter-AUC	Intra-AUC
Race	CLIP	57.33	86.02	4.93	16.05	2.76	4.76	3.98
	FairCLIP	59.55	86.24	3.03	10.47	2.02	3.81	**3.36**
	AdFair-CLIP	**59.65**	**86.38**	**2.27**	**8.69**	**0.79**	**2.93**	3.54
Gender	CLIP	57.33	86.02	6.10	6.98	2.32	3.35	2.69
	FairCLIP	59.34	86.17	3.48	6.12	2.01	3.21	2.47
	AdFair-CLIP	**59.52**	**86.33**	**2.18**	**5.13**	**1.44**	**2.61**	**1.88**

Moreover, Fig. 3 compares accuracy and AUC. AdFair-CLIP with racial bias mitigation performs competitively with SOTA baselines and even surpasses them in some cases, while its gender bias counterpart shows slightly lower performance yet remains superior to standard CLIP.

Table 3. Few-shot (10%) comparison of AdFair-CLIP with baselines, pre-trained on MIMIC-CXR and fine-tuned on CheXpert Plus, with all scores presented in percentage.

Attribute	Method	Acc	AUC	DPD	DEOdds	GAUC	Inter-AUC	Intra-AUC
Race	CLIP	55.56	84.95	6.27	11.68	2.67	3.58	4.83
	FairCLIP	57.47	86.07	5.96	9.25	**1.86**	3.71	3.95
	AdFair-CLIP	**57.78**	**86.15**	**3.73**	**8.20**	1.96	**2.89**	**2.74**
Gender	CLIP	55.56	84.95	6.89	10.57	2.98	3.25	4.84
	FairCLIP	56.02	85.18	4.81	10.02	**1.77**	3.31	3.69
	AdFair-CLIP	**56.11**	**85.48**	**3.31**	**8.98**	2.01	**2.30**	**3.29**

AdFair-CLIP vs. CLIP Fairness Benchmarks. Tables 1, 2, and 3 demonstrate that AdFair-CLIP consistently achieves the highest fairness and diagnostic performance improvements across most metrics in all evaluated scenarios for race and gender. Notably, DebiasCLIP is only compared in scenario (a) because its frozen strategy makes it equivalent to CLIP in scenarios (b) and (c). To illustrate AdFair-CLIP's superiority in in-domain and out-of-domain generalization, we compute fairness improvements over CLIP by averaging percentage gains across five fairness metrics. In the zero-shot setting, AdFair-CLIP achieves the

Table 4. Vision-only vs. multimodal (vision + text) features for adversarial learning.

Zeroshot								
Attribute	**Method**	**Acc**	**AUC**	**DPD**	**DEOdds**	**GAUC**	**Inter-AUC**	**Intra-AUC**
Race	Vision-only	**61.06**	84.83	2.13	8.16	1.36	3.15	2.98
	Multimodal	60.91	**85.05**	**1.87**	**6.04**	**0.99**	**2.46**	**2.68**
Gender	Vision-only	58.44	84.17	3.09	8.80	**1.73**	1.77	2.87
	Multimodal	**58.77**	**84.56**	**2.75**	**8.69**	1.82	**1.75**	**2.29**
Linear Evaluation								
Race	Vision-only	59.11	86.25	2.53	9.47	1.54	3.45	**3.41**
	Multimodal	**59.65**	**86.38**	**2.27**	**8.69**	**0.79**	**2.93**	3.54
Gender	Vision-only	59.13	86.15	3.26	5.82	1.72	3.17	2.18
	Multimodal	**59.52**	**86.33**	**2.18**	**5.13**	**1.44**	**2.61**	**1.88**

highest improvements (54.33% for race and 45.40% for gender), outperforming DebiasCLIP (36.85% and 39.94%) and FairCLIP (33.84% and 25.39%). Although fine-tuning and transfer learning introduce additional biases that degrade fairness across all methods, AdFair-CLIP remains robust, achieving 44.14% (race) and 36.18% (gender) in the few-shot setting and 31.89% (race) and 32.16% (gender) in transfer learning, whereas FairCLIP suffers a marked decline, with race fairness dropping from 33.84% in the zero-shot setting to 14.14% in transfer learning and a similar trend observed for gender fairness. These results confirm that AdFair-CLIP's feature-space intervention is more effective and generalizable than similarity-space debiasing and the backbone-frozen fine-tuning strategy.

3.5 Ablation

Vision-only vs. Multimodal Features. To examine whether bias is propagated or amplified during the image-text pairing process, we conduct an ablation study comparing the performance of our AdFair-CLIP model on CheXpert Plus when employing vision-only features (h_v) versus multimodal features ($[h_v; h_u]$) in the min-max adversarial learning framework. As shown in Table 4, the multimodal approach achieves superior diagnostic performance and improved fairness metrics compared to the vision-only setting. These results indicate that sensitive information can be implicitly encoded in vision-text alignments, even when textual inputs do not explicitly contain demographic information. Furthermore, joint optimization over multimodal features enables the adversarial framework to more effectively suppress biases introduced through image-text pairing, leading to enhanced fairness and more robust debiased representations.

4 Conclusion

Our investigation reveals significant fairness issues in existing SOTA CLIP-based chest X-ray models, highlighting the need for fairness-aware multimodal learning. To address this, we propose AdFair-CLIP, an adversarial framework designed to mitigate demographic biases in CLIP-based chest X-ray diagnosis. By

incorporating a minimax adversarial objective in the feature space, our method reduces spurious correlations between sensitive attributes and diagnostic predictions while preserving robust multimodal alignment. Extensive evaluations demonstrate that AdFair-CLIP improves both fairness and diagnostic performance while exhibiting strong generalization in zero-shot and few-shot settings. These findings establish a new benchmark for fairness-aware medical AI in chest X-ray and encourage further research into equitable healthcare applications.

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

References

1. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: International conference on machine learning. pp. 60–69. PMLR (2018)
2. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)
3. Berg, H., Hall, S.M., Bhalgat, Y., Yang, W., Kirk, H.R., Shtedritski, A., Bain, M.: A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning (2022), https://arxiv.org/abs/2203.11933
4. Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E.H., Goodrow, C.: Fairness in recommendation ranking through pairwise comparisons (2019), https://arxiv.org/abs/1903.00780
5. Brown, A., Tomasev, N., Freyberg, J., Liu, Y., Karthikesalingam, A., Schrouff, J.: Detecting and preventing shortcut learning for fair medical ai using shortcut testing (short). arXiv preprint arXiv:2207.10384 **2** (2022)
6. Chambon, P., Delbrouck, J.B., Sounack, T., Huang, S.C., Chen, Z., Varma, M., Truong, S.Q., Chuong, C.T., Langlotz, C.P.: Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats (2024), https://arxiv.org/abs/2405.19538
7. Diederik, P.K.: Adam: A method for stochastic optimization. (No Title) (2014)
8. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of machine learning research **17**(59), 1–35 (2016)
9. Ghosh, A., Acharya, A., Jain, R., Saha, S., Chadha, A., Sinha, S.: Clipsyntel: clip and llm synergy for multimodal question summarization in healthcare. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 22031–22039 (2024)
10. Glocker, B., Jones, C., Bernhardt, M., Winzeck, S.: Algorithmic encoding of protected characteristics in chest x-ray disease detection models. EBioMedicine **89** (2023)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

12. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: 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. pp. 3942–3951 (2021)
13. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)
14. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
15. Jin, R., Xu, Z., Zhong, Y., Yao, Q., QI, D., Zhou, S.K., Li, X.: Fairmedfm: fairness benchmarking for medical imaging foundation models. Advances in Neural Information Processing Systems **37**, 111318–111357 (2024)
16. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
17. Khan, M.O., Afzal, M.M., Mirza, S., Fang, Y.: How fair are medical imaging foundation models? In: Machine Learning for Health (ML4H). pp. 217–231. PMLR (2023)
18. Luo, Y., Shi, M., Khan, M.O., Afzal, M.M., Huang, H., Yuan, S., Tian, Y., Song, L., Kouhana, A., Elze, T., et al.: Fairclip: Harnessing fairness in vision-language learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12289–12301 (2024)
19. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019)
20. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
22. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I.Y., Ghassemi, M.: Chexclusion: Fairness gaps in deep chest x-ray classifiers. In: BIOCOMPUTING 2021: proceedings of the Pacific symposium. pp. 232–243. World Scientific (2020)
23. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature medicine **27**(12), 2176–2182 (2021)
24. Wang, J., Liu, Y., Wang, X.E.: Are gender-neutral queries really gender-neutral? mitigating gender bias in image search (2021), https://arxiv.org/abs/2109.05433
25. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
26. Yao, Y., Lin, Q., Yang, T.: Stochastic methods for auc optimization subject to auc-based fairness constraints. In: International Conference on Artificial Intelligence and Statistics. pp. 10324–10342. PMLR (2023)
27. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)

28. Zhang, H., Dullerud, N., Roth, K., Oakden-Rayner, L., Pfohl, S., Ghassemi, M.: Improving the fairness of chest x-ray classifiers. In: Conference on health, inference, and learning. pp. 204–233. PMLR (2022)
29. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M.P., Naumann, T., Wang, S., Poon, H.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs (2025), https://arxiv.org/abs/2303.00915
30. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference. pp. 2–25. PMLR (2022)