Mitigating Bias and Improving Expressiveness in Deep Learning-Based MRI Reconstruction

Matteo Wohlrapp[1,2]
Supervised by Niklas Bubeck[1]

[1] AI in Medicine, Technical University of Munich, Munich, Germany
[2]Dana-Farber Cancer Institute, Harvard Medical School, Boston, USA
matteo.wohlrapp@tum.de

Abstract. Medical image reconstruction models can introduce subtle biases that propagates to downstream prediction models, yet bias mitigation at the reconstruction stage remains underexplored. We evaluate three established fairness interventions—sample reweighting, an Equalized Odds (EODD) constraint, and an adversarial "fairness through unawareness" loss—applied to U-Net, Pix2Pix, and a stochastic differential equation (SDE) reconstruction model. Using two datasets (CheXpert chest X-rays and UCSF-PDGM brain MRI), we quantify changes in reconstruction quality (PSNR), downstream classifier performance (AUROC), and group fairness across age, sex, and race. Our results show modest but statistically significant bias reductions for age and sex, with negligible improvements for race. Performance remains largely stable on CheXpert but declines under fairness constraints on USC-PDGM, suggesting dataset-specific sensitivities. Qualitative analyses reveal minimal visible changes in clinically relevant features, underscoring that classifiers are robust to the small perturbations introduced by reconstruction. We therefore argue that reconstruction models have limited "elasticity" to influence downstream bias and recommend focusing bias mitigation efforts on predictive models rather than at the reconstruction stage. Clinically, this emphasizes the need for interventions that balance fairness and diagnostic accuracy within the classification pipeline. The code is provided in our GitHub repository.

Keywords: Bias · Fairness · Mitigation · Reconstruction

1 Introduction

Computer vision AI models capture statistical patterns inherent in their training datasets. While this characteristic benefits certain applications by increasing the overall model performance—such as in scenarios involving homogeneous data distributions—it poses significant challenges in healthcare, where biased models risk exacerbating disparities by underserving particular population subgroups. Consequently, a growing body of literature has emphasized the significance of bias and fairness in medical AI, exploring its origins and impacts.

Bias has been extensively studied within classification [28,9,4,3] and segmentation tasks [11,16,22]. These investigations primarily concentrate on group fairness [15], assessing performance disparities through metrics such as Area Under the Curve (AUC) [26], True Positive Rate (TPR) [9], or specialized fairness metrics [32] across demographic subgroups defined by sensitive attributes, including age, race, and skin tone [26,14]. Across studies, these metrics consistently reveal the presence of bias, with certain population subgroups—often racial or age-based—repeatedly found to be underserved [3,26,28].

Particularly within classification, substantial efforts have been directed toward developing bias mitigation strategies: Data-centric approaches directly modify training datasets to address biases, employing methods such as data redistribution to equalize subgroup representation [21]. Li and Vasconcelos introduced REPAIR [17], a differentiable dataset resampling technique aimed explicitly at balancing representation. Techniques like data harmonization removes sensitive information [2], whereas synthetic generation of diverse samples enriches minority subgroup representation [29]. Liu et al. proposed Just Train Twice (JTT) [18], emphasizing instances previously misclassified to implicitly tackle subgroup biases without explicit subgroup annotations.

Representation-level strategies aim to learn unbiased feature representations through explicit disentanglement or specialized architectures. Creager et al. [5] introduced Flexibly Fair Variational Autoencoders (FFVAE) to disentangle sensitive latent factors from non-sensitive, task-relevant features. Orthogonal disentanglement approaches enforce independence between sensitive attributes and task-specific representations, promoting fairness without compromising accuracy [25,6,4,7]. In contrast, Gong et al. [10] proposed a group-adaptive classifier architecture employing demographic-specific convolutional kernels and attention mechanisms.

Optimization-level methods integrate fairness constraints into the model training phase through adversarial learning, fairness-specific loss functions, or specialized training regimens. Zhang et al. [33] and Adeli et al. [1] developed adversarial training methods discouraging the encoding of protected attributes by jointly training predictors and adversaries. Building upon this, Kim et al. [13] used mutual information-based adversarial training to explicitly minimize biased correlations. Wang et al. [30] systematically compared adversarial training strategies, proposing domain-independent mitigation methods. Distributionally robust optimization (Group DRO), introduced by Sagawa et al. [24], explicitly optimizes for worst-case subgroup distributions. Additionally, Marcinkevičs et al. [20] incorporated fairness-specific constraints directly into training procedures.

Post-processing methods adjust predictions after model training without altering model parameters. Techniques such as model calibration and pruning have effectively achieved fairness post-hoc [31].

Beyond classification and segmentation, emerging research investigates biases introduced by image reconstruction or denoising models [8,27]. Prior studies primarily compared subgroup performance using image quality metrics such as Peak Signal to Noise Ratio (PSNR). Our previous research expanded this

perspective by analyzing the impact of reconstruction-induced biases on downstream tasks, such as classification and segmentation. We were driven by the hypothesis that given the increasing prevalence and accessibility of pre-trained reconstruction models in clinical pipelines, it is plausible that such models will be routinely paired with downstream predictors—intentionally or implicitly. Despite this progress in fairness evaluation, bias mitigation in the context of reconstruction models remains largely unexplored.

Motivated by the gap in existing literature, this paper evaluates the impact of established bias mitigation techniques applied to reconstruction models. Our contributions include: (i) an adaptation of three classification fairness techniques—sample reweighting, Equalized Odds, and adversarial loss—to the reconstruction stage of medical imaging, (ii) an application to multiple architectures (U-Net, Pix2Pix, SDE) and datasets (UCSF-PDGM, CheXpert), and (iii) a comprehensive evaluation of their effects on downstream fairness in classification across age, sex, and race.

We investigate the implications for fairness and performance across two medical imaging modalities and three distinct model architectures.

2 Method

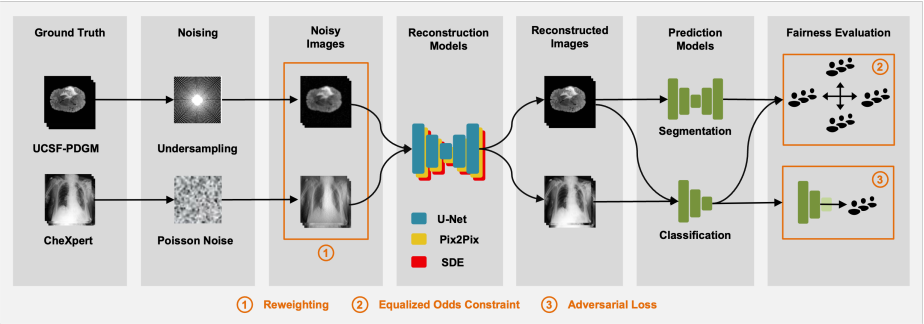


Fig. 1: Overview of the bias mitigation pipeline. Ground-truth MRI and chest X-ray images are corrupted with realistic noise (undersampling, Poisson noise) and reconstructed using U-Net, Pix2Pix, and an SDE model trained under one of three fairness strategies — sample reweighting (1), Equalized Odds (EODDs) constraint (2), or adversarial loss (3) — before evaluation on segmentation and classification networks to assess performance and group fairness.

2.1 Background

Our previous work established an evaluation framework investigating bias introduced by medical image reconstruction models. Specifically, we analyzed their

implications for fairness in downstream classification and segmentation tasks. The framework included systematically introducing artificial noise into medical imaging datasets (MRI and X-ray) to simulate realistic clinical scenarios, such as accelerated MRI scans and low-dose X-ray protocols. We analyzed the performance and fairness influence of several reconstruction techniques, including classical and generative models such as U-Net [23], the Generative Adversarial Network (GAN) Pix2Pix [12], and a stochastic differential equation (SDE)-based diffusion model [19] on downstream classification and segmentation networks. Our findings highlighted modest yet significant biases resulting from reconstruction, particularly evident under specific noise conditions and sensitive attributes, underscoring the importance of exploring bias mitigation in reconstruction models. For more information on the setup, datasets and experiments, please refer to the IDP report: "Exploring Bias in Deep Learning-Based MRI".

2.2 Mitigation

This work examines the interaction between pre-trained reconstruction and classification models, reflecting a common clinical pipeline where reconstruction and downstream predictors are deployed in tandem. We aim to investigate whether fairness mitigation applied solely to the reconstruction stage can meaningfully influence downstream bias without altering the diagnostic model. We fine-tune pre-trained reconstruction models f_θ on unseen validation data and leave the downstream ResNet classifier g_ϕ fixed. We base our implementation on bias mitigation techniques from classification and adapt them to the image reconstruction setting. Specifically, we apply three strategies: (1) a data-level method to reweigh samples and two optimization-level approaches: (2) an Equalized Odds (EODD) constraint and (3) an adversarial "fairness through unawareness" loss. The latter two leverage a frozen, pre-trained classifier g_ϕ to guide optimization and assess fairness impact indirectly via changes in classification outputs. We do not consider representation-level methods, as these typically require architecture-specific modifications and were not feasible to implement consistently across the reconstruction models.

For each noisy input x, the reconstruction model produces $\hat{x} = f_\theta(x)$ and the classifier yields $\hat{y} = g_\phi(\hat{x})$. We jointly optimize a reconstruction loss ℓ_{rec}, a classification loss ℓ_{CE}, and a fairness loss ℓ_{fair} to preserve reconstruction fidelity and classification accuracy while promoting fairness:

$$\ell_{\text{rec}} = \|x - \hat{x}\|_2^2, \quad \ell_{\text{CE}} = -\sum_{c=1}^{C} y_c \log \hat{y}_c,$$

Gradients flow through g_ϕ into f_θ, while g_ϕ itself remains fixed. For the implementation, we stayed close to the original mitigation technique and adjusted the fairness lambda based on a hyperparameter search.

Reweighting We alter sampling frequencies via a weighted random sampler. For K sensitive attributes $(g_i^1, \ldots, g_i^K)$, each sample receives probability

$$p_i = \frac{1/n_{(g_i^1,\ldots,g_i^K)}}{\sum_{j=1}^{n}\big(1/n_{(g_j^1,\ldots,g_j^K)}\big)},$$

where $n_{(g_i^1,\ldots,g_i^K)}$ is the number of training examples sharing exactly that attribute tuple. Training then minimizes

$$\mathcal{L}_{\mathrm{RE}} = \sum_{i=1}^{n} \ell_{\mathrm{rec}}(x_i, \hat{x}_i).$$

Equalized Odds Constraint We base this implementation on prior work of Marcinkevics et al. [20]. However, to enforce parity of true- and false-positive rates across groups, we apply EODD rather than Equal Opportunity (EOP), a relaxation of EODD that only compares true-positive rates. We define the differentiable soft prediction.

$$\hat{y}_i = \sigma\big((f_\theta(x_i) - \tau)/T\big),$$

with Sigmoid σ, threshold τ and temperature T estimated through hyperparameter tuning. We compute the maximum equalized odds difference

$$\mathrm{EODD} = \max_{a,b,c,d \in A} \Big(|\mathbb{E}[\hat{y} \mid A = a, y = 0] - \mathbb{E}[\hat{y} \mid A = b, y = 0]| + |\mathbb{E}[\hat{y} \mid A = c, y = 1] - \mathbb{E}[\hat{y} \mid A = d, y = 1]| \Big)$$

and set $\ell_{\mathrm{fair}} = \mathrm{EODD}^2$. To balance scales, we maintain running averages of ℓ_{CE} and ℓ_{fair} via exponential moving average (EMA). We again EMA-normalize this combined term to match ℓ_{rec}, then weight by λ_{fair}. The final objective is

$$\mathcal{L}_{\mathrm{EODD}} = \ell_{\mathrm{rec}} + \lambda_{\mathrm{fair}}\,\mathrm{EMA}\big(\ell_{\mathrm{CE}} + \ell_{\mathrm{fair}}\big).$$

In the supplementary material 6.1, we show that minimizing EODD^2 between subgroups corresponds to minimizing $\widehat{\mathrm{Cov}}(A, f_\theta(X)|Y = 1) + \widehat{\mathrm{Cov}}(A, f_{\theta(X)}|Y = 0)$.

Adversarial Loss We freeze the classifier's feature extractor h_ϕ and append a small bias predictor on its penultimate features $u_i = h_\phi(\hat{x}_i)$. We measure dependence on the sensitive attribute a_i via squared Pearson correlation,

$$\ell_{\mathrm{fair}} = \mathrm{Corr}^2(u, a).$$

similar to previous work by Adeli et al. [1]. As with EODD, we EMA-normalize ℓ_{CE} and ℓ_{fair}, and EMA-normalize again to ℓ_{rec}'s scale, and weight by λ_{fair}. The full loss is

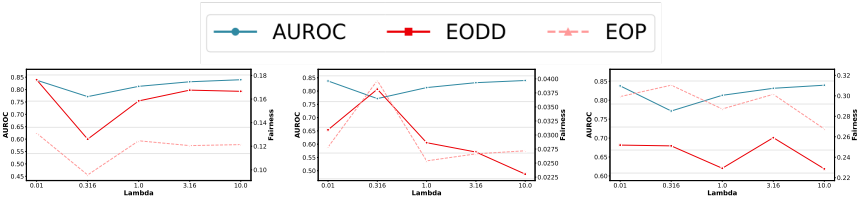
$$\mathcal{L}_{\mathrm{ADV}} = \ell_{\mathrm{rec}} + \lambda_{\mathrm{fair}}\,\mathrm{EMA}\big(\ell_{\mathrm{CE}} + \ell_{\mathrm{fair}}\big).$$

3 Experiments and Results

3.1 Background

Our previous work analyzed bias in reconstruction models and their implications for downstream tasks. Table 3 in the appendix summarizes the initial bias evaluation across the three sensitive attributes: sex, age, and race. These tables report the differences in bias between predictions on ground-truth images and those on reconstructed images. Statistically significant differences are color-coded (yellow for significantly worse, blue for significantly better), while bold entries denote cases where the standard deviation exceeds the effect size. Sex showed better baseline fairness with only minor additional significant bias. Age demonstrated lower initial fairness but slightly improved bias after reconstruction. Race displayed high standard deviation and no consistent significant effects.

3.2 λ_{fair} Selection

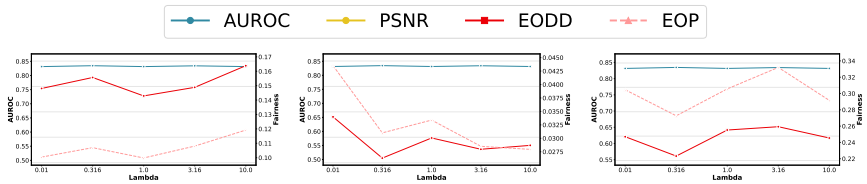


(a) Age (b) Sex (c) Race

Fig. 2: Influence of fairness weighting parameter (λ_{fair}) on classifier AUROC performance and fairness metrics for the Equalized Odds (EODD) mitigation constraint, evaluated with U-Net on the CheXpert dataset. There is minor sensitivity of AUROC to lambda; fairness metrics show greater variance but minimal substantial improvement with increased λ.

We conducted a one-dimensional sweep of λ_{fair} on the log scale using the U-Net model trained on the CheXpert dataset to determine the appropriate weighting for the fairness objective. Figures 2 and 3 show the impact of λ_{fair} (log scale) on AUROC and the equalized-odds (EODD) metric, while Figures 10 and 11 in the appendix show the corresponding effects on PSNR. Across five steps, AUROC and PSNR remained stable, while fairness metrics fluctuated slightly without showing consistent fairness improvement at higher λ_{fair} values. Based on this, we fixed $\lambda_{\text{fair}} = 1$ for all experiments to balance performance and fairness.

3.3 Performance Evaluation



(a) Age (b) Sex (c) Race

Fig. 3: Influence of λ_{fair} on AUROC and fairness metrics for the adversarial fairness loss with U-Net on CheXpert. Similar findings to the EODD loss include minimal AUROC variation and moderate fairness variability without substantial gains.

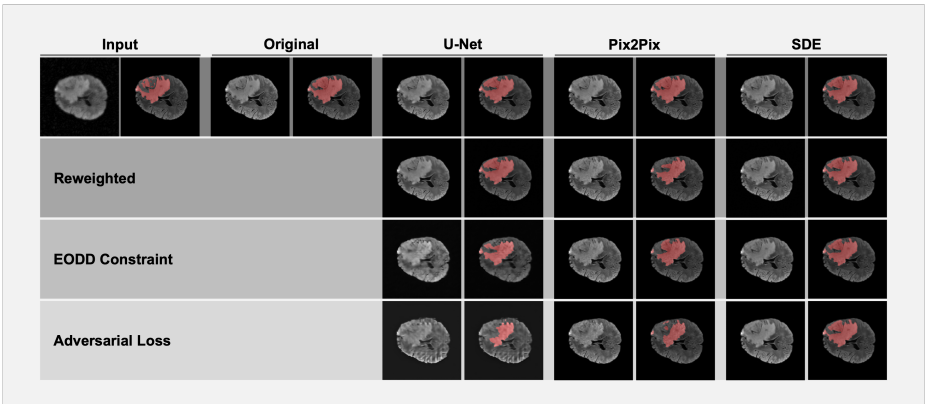


Fig. 4: Reconstructed images and segmentation mask visualizations of models trained on the UCSF-PDGM dataset pre- and post-mitigation. Visually minimal changes; U-Net shows slightly larger deviations from original masks, but no clinically meaningful features are introduced.

Qualitative Analysis Figures 4 (UCSF-PDGM) and 5 (CheXpert) illustrate qualitative comparisons of reconstructions before and after mitigation. While Grad-CAM maps (CheXpert) and segmentation masks (UCSF-PDGM) show only minor visual differences, some subtle deviations are observed—the U-Net reconstructions for UCSF-PDGM exhibit slightly stronger artifacts for the equalized odds and adversarial loss. The SDE model on CheXpert seems to enhance support devices for the adversarial loss not visible in the original images, while Pix2Pix under EODD loss exhibits mild intensity shifts. None of the mitigation strategies introduced changes visibly relevant to clinical interpretation.

Quantitative Analysis Tables 1 and 2 summarize performance metrics (AUROC, PSNR, LPIPS) across all mitigation strategies: sample reweighting (RE), equalized odds (EODD) constraint, and adversarial loss (ADV), compared to

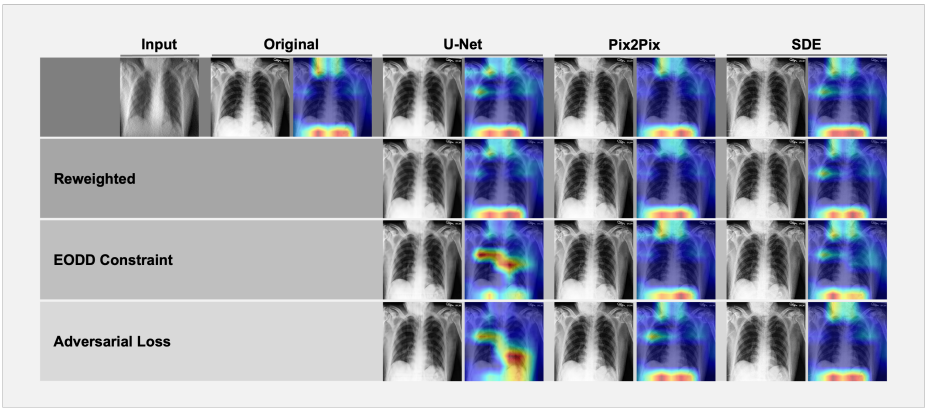


Fig. 5: Reconstructed images and GradCAM visualizations of models trained on the CheXpert dataset pre- and post-mitigation. Minimal overall visual differences, subtle enhancement of support devices by SDE, and minor intensity shifts by GAN were observed.

the standard baseline reconstruction (STD). For CheXpert, all models maintained their original performance, with deviations under 5% relative to the baseline. In contrast, performance on UCSF-PDGM—whose fine-tuning set is significantly smaller—dropped considerably: U-Net showed over 10% decline in PSNR, LPIPS, and Dice under EODD and ADV; Pix2Pix experienced a similar drop in AUROC and LPIPS under ADV.

3.4 Fairness Evaluation

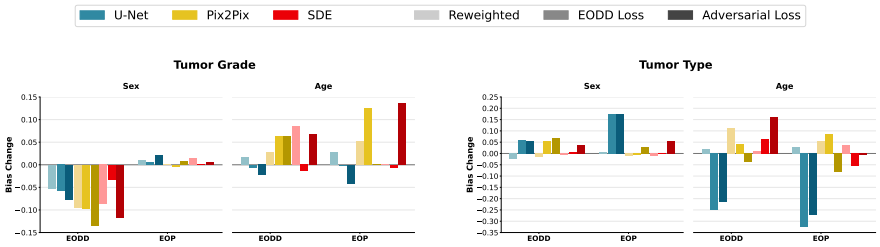


Fig. 6: Bootstrapped bias differences pre- and post-mitigation on the UCSF-PDGM dataset. Lower bars indicate improved fairness. Large variability was observed, with no clear trend associated with specific models, attributes, or mitigation techniques.

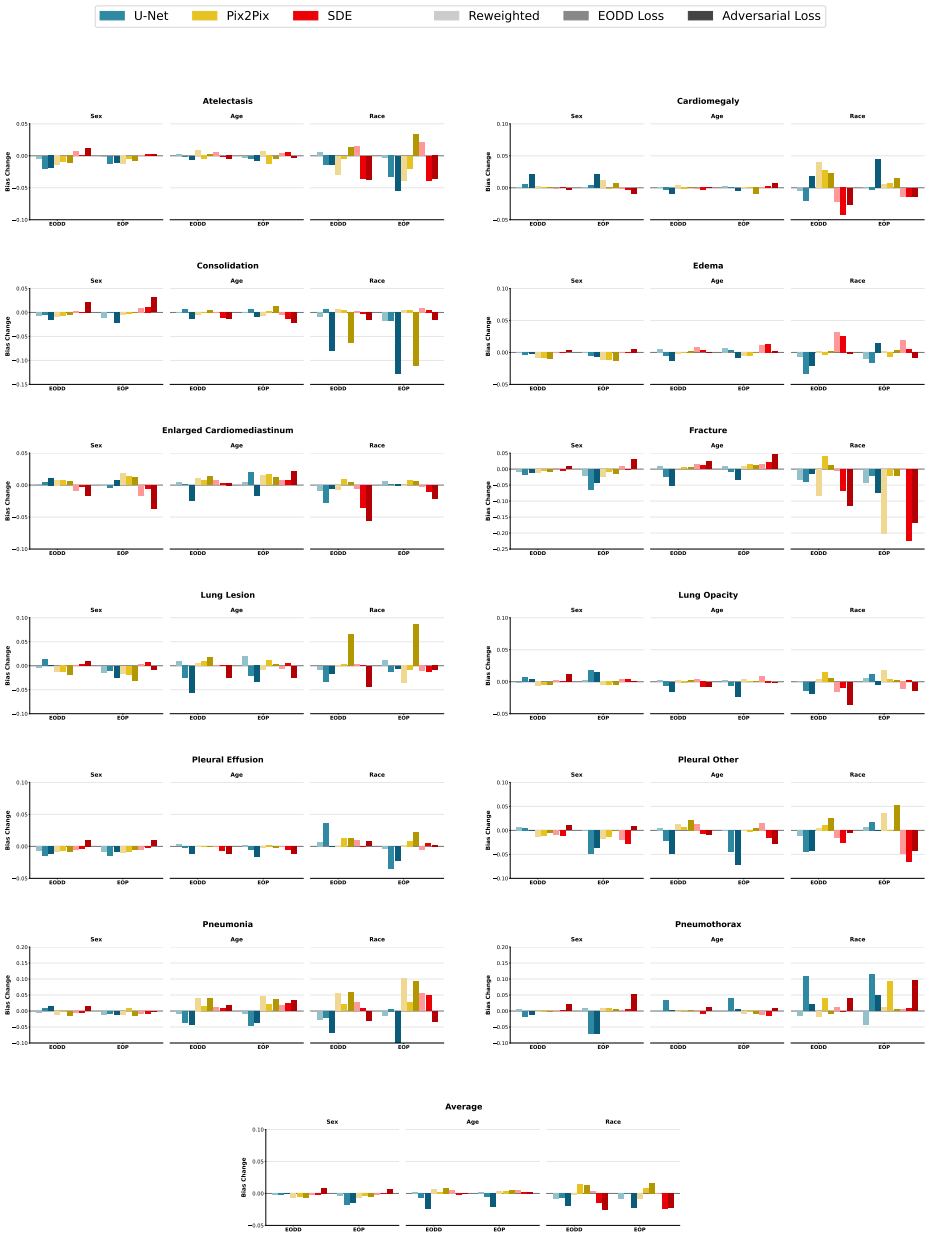


Fig. 7: Bootstrapped bias differences pre- and post-mitigation on the CheXpert dataset. Lower bars indicate improved fairness. Overall, slight bias improvements exist for U-Net and SDE; Pix2Pix worsens bias for specific pathologies.

Table 1: Performance impacts of mitigation methods on the UCSF-PDGM dataset for the different models. Color-coded performance changes to pre-mitigation in percent (yellow=worse, blue=better) with baseline values of predictions on ground truth images. Performance notably declines under equalized odds and adversarial mitigations, particularly affecting U-Net (PSNR) and Pix2Pix (AUROC).

$\downarrow > 10\%$, $5\% \leq \downarrow < 10\%$ $\uparrow < -10\%$ $-10\% < \uparrow \leq -5\%$

(a) U-Net

Metrics		Baseline	U-Net			
			STD	RE	EODD	ADV
AUROC	Tumor Type	0.788	0.773	0.767	0.753	0.807
	Tumor Grade	0.733	0.746	0.745	0.721	0.735
Dice			0.699	0.701	0.606	0.563
PSNR			35.766	36.185	29.660	25.918
LPIPS			0.030	0.029	0.109	0.103

(b) Pix2Pix

Metrics		Baseline	Pix2Pix			
			STD	RE	EODD	ADV
AUROC	Tumor Type	0.788	0.829	0.775	0.763	0.723
	Tumor Grade	0.733	0.777	0.740	0.745	0.710
Dice			0.709	0.697	0.696	0.679
PSNR			35.198	34.204	34.012	31.545
LPIPS			0.022	0.028	0.028	0.049

Figures 7 and 6 present bootstrapped fairness deltas for each mitigation method. Detailed results are shown in Tables 4, 5, and 6 (appendix). On UCSF-PDGM, no clear trend emerges — no mitigation method consistently reduced bias across sensitive attributes or models even though there are substantial changes in the reported bias results of up to 0.3. It seems that for tumor grade, we reduce bias for sex through mitigation while adding to the unfairness for age. For tumor type, the pattern is reversed (Figure 6). On CheXpert, while race again showed no significant changes due to large confidence intervals, small yet statistically significant bias reductions were seen for sex across all mitigation techniques. The most pronounced fairness improvements were observed for age, with EODD and ADV losses yielding significant bias reductions—particularly for U-Net and SDE and, to a lesser extent, for Pix2Pix. Many of these improvements had p-values below 0.05 (Tables 5 and 6). While some pathologies like Lung Lesions or Pneumonia show negative shifts, others, e.g., Consolidation and Fracture, show improvements of up to 0.2 in Figure 7. Averaged across pathologies, this leaves a slight fairness improvement, specifically for U-Net and SDE.

(c) SDE

Metrics		Baseline	SDE			
			STD	RE	EODD	ADV
AUROC	Tumor Type	0.788	0.786	0.780	0.778	0.824
	Tumor Grade	0.733	0.727	0.733	0.737	0.783
Dice			0.707	0.705	0.707	0.662
PSNR			34.654	34.443	34.388	35.035
LPIPS			0.016	0.017	0.017	0.014

Figure 8 (and Figure 12 in the appendix) visualizes fairness versus performance changes in percent. These plots show that fairness metrics fluctuate more than performance metrics, with fairness changes spanning an order of magnitude wider range. UCSF-PDGM data, in particular, exhibited greater variability, as seen by their tendency to lie outside the CheXpert pathologies. Greater variability can also be seen under EODD and ADV compared to RE, and more so for sex and race than age. Most points cluster around zero change, but the few extreme performance outliers were negative.

4 Discussion

Only modest fairness gains were achieved across both datasets and all reconstruction architectures through the applied mitigation strategies (RE, EODD, ADV). Improvements were primarily limited to the age attribute and, to a lesser degree, sex. Race remained unaffected, consistent with earlier findings of high variability and low statistical significance. Performance remained stable on CheXpert, while noticeable degradations occurred on UCSF-PDGM — especially for U-Net under EODD constraint and Pix2Pix under ADV loss. Visual inspection (Figures 4, 5) revealed little perceptible change, reinforcing the conclusion that reconstruction has limited potential for fairness correction.

Dataset Size The two datasets differ quite substantially in size. UCSF-PDGM (495 MRI scans with around 150 scans each) and CheXpert (over 220,000 chest X-rays) are split into 70/20/10 training, validation, and test sets. UCSF-PDGM uses the same data for reconstruction and classification, while CheXpert supports separate partitions. CheXpert, with its larger fine-tuning dataset, tolerated fairness constraints without loss of fidelity or classification accuracy. By contrast, UCSF-PDGM exhibited strong performance degradation under the same constraints, suggesting that larger datasets better accommodate multi-objective training, while smaller datasets force trade-offs. Interestingly, reweighting improved U-Net performance more on UCSF-PDGM than on CheXpert, possibly due to stronger overfitting on the smaller dataset and better sample efficiency from rebalancing. However, across models and datasets, reweighting had a limited effect on performance, indicating that reconstruction models may already

Table 2: Performance impacts of mitigation methods on the CheXpert dataset for different models. Color-coded performance changes in percent (yellow=worse, blue=better) with baseline values of predictions on ground truth images. Minimal variations were observed across all mitigations.

↓> 10%, 5% ≤↓< 10% ↑< −10%, −10% <↑≤ −5%

(a) U-Net

Metrics		Baseline	U-Net			
			STD	RE	EODD	ADV
AUROC	**Atelectasis**	0.872	0.865	0.866	0.864	0.854
	Cardiomegaly	0.909	0.904	0.905	0.902	0.898
	Consolidation	0.914	0.909	0.910	0.904	0.900
	Edema	0.899	0.892	0.892	0.890	0.889
	EC	0.788	0.782	0.782	0.781	0.779
	Fracture	0.757	0.745	0.747	0.749	0.746
	Lung Lesion	0.796	0.780	0.780	0.783	0.765
	Lung Opacity	0.885	0.876	0.877	0.874	0.869
	Pleural Effusion	0.925	0.917	0.917	0.915	0.906
	Pleural Other	0.828	0.813	0.813	0.810	0.796
	Pneumonia	0.833	0.823	0.824	0.822	0.802
	Pneumothorax	0.767	0.747	0.746	0.760	0.765
	Average	0.848	0.838	0.838	0.838	0.831
PSNR			30.521	30.447	29.404	29.153
LPIPS			0.185	0.193	0.178	0.182

generalize well and that additional rebalanced data alone is insufficient for further improvements — unless data volume is significantly increased.

Lambda Sensitivity Despite the observed dataset sensitivity, the fairness weight λ_{fair} had surprisingly little effect. Across multiple values, performance metrics were stable, and fairness fluctuated without showing systematic trends. This might reflect the reconstruction models' inability to introduce semantic-level changes regardless of loss weighting. Alternatively, the fairness signal from the frozen classifier may be too weak to influence the reconstructions meaningfully.

Mitigation Strategies Among the strategies evaluated, reweighting showed the least impact, both in terms of fairness and performance. This may be due to its purely data-centric nature, which leaves the optimization unchanged. EODD and ADV losses were more effective in reducing age bias and, in some cases, sex bias, with relatively low-performance penalties—though this may stem from the overall low magnitude of fairness gains. The similar performance of EODD and ADV losses may also be attributed to the underlying correlation between

(b) Pix2Pix

Metrics		Baseline	Pix2Pix			
			STD	RE	EODD	ADV
AUROC	Atelectasis	0.872	0.858	0.860	0.862	0.862
	Cardiomegaly	0.909	0.902	0.904	0.904	0.905
	Consolidation	0.914	0.905	0.906	0.905	0.907
	Edema	0.899	0.891	0.893	0.891	0.893
	EC	0.788	0.781	0.782	0.782	0.782
	Fracture	0.757	0.736	0.744	0.742	0.743
	Lung Lesion	0.796	0.780	0.781	0.782	0.783
	Lung Opacity	0.885	0.871	0.873	0.873	0.874
	Pleural Effusion	0.925	0.912	0.913	0.913	0.914
	Pleural Other	0.828	0.798	0.807	0.810	0.807
	Pneumonia	0.833	0.818	0.817	0.822	0.820
	Pneumothorax	0.767	0.752	0.757	0.757	0.758
	Average	0.848	0.834	0.836	0.837	0.837
PSNR			28.615	28.797	28.448	28.859
LPIPS			0.109	0.103	0.109	0.103

sensitive attributes and the model's prediction scores. It is worth noting that fine-tuning using the EODD constraint and subsequently evaluating fairness using the same metric introduces a potential bias, as the evaluation criterion aligns with the optimization objective. However, given that fairness improvements under EODD were comparable to those observed under the EOP metric—despite the model not being explicitly trained on it—we consider the EODD loss to be a valid and effective constraint for optimization in this context.

Pathology-Specific Bias Patterns Bias changes were generally without clear patterns across pathologies. For Pneumothorax, Pneumonia, Lung Lesions, and Cardiomegaly, fairness often worsened under Pix2Pix. This could be linked to the instability of GAN training, such as mode collapse or sensitivity to subgroup imbalances, but further targeted analysis is required.

Classical vs. Generative Models Our reconstruction models span classical (U-Net) and generative (Pix2Pix, SDE) architectures. Given their generative capacity, we expected some mitigation effects to manifest as semantically meaningful image changes. However, reconstructions differed mainly at the pixel level, without obvious anatomical relevance. This suggests that the mitigation signal cannot guide the generative process toward producing fairer images with interpretable structural changes.

Elasticity of Reconstruction Models Across models and mitigation techniques, reconstructions changed minimally — visually and quantitatively — implying limited elasticity. That means reconstruction models shifted pixel inten-

(c) SDE

Metrics		Baseline	SDE			
			STD	RE	EODD	ADV
AUROC	**Atelectasis**	0.872	0.865	0.865	0.867	0.861
	Cardiomegaly	0.909	0.905	0.907	0.905	0.901
	Consolidation	0.914	0.908	0.908	0.910	0.905
	Edema	0.899	0.896	0.895	0.896	0.893
	EC	0.788	0.784	0.784	0.787	0.781
	Fracture	0.757	0.755	0.751	0.752	0.744
	Lung Lesion	0.796	0.790	0.784	0.791	0.788
	Lung Opacity	0.885	0.877	0.877	0.878	0.875
	Pleural Effusion	0.925	0.917	0.918	0.920	0.915
	Pleural Other	0.828	0.819	0.815	0.816	0.810
	Pneumonia	0.833	0.825	0.824	0.825	0.819
	Pneumothorax	0.767	0.770	0.767	0.768	0.757
	Average	0.848	0.843	0.841	0.843	0.837
PSNR			27.121	27.456	27.752	27.112
LPIPS			0.149	0.101	0.110	0.143

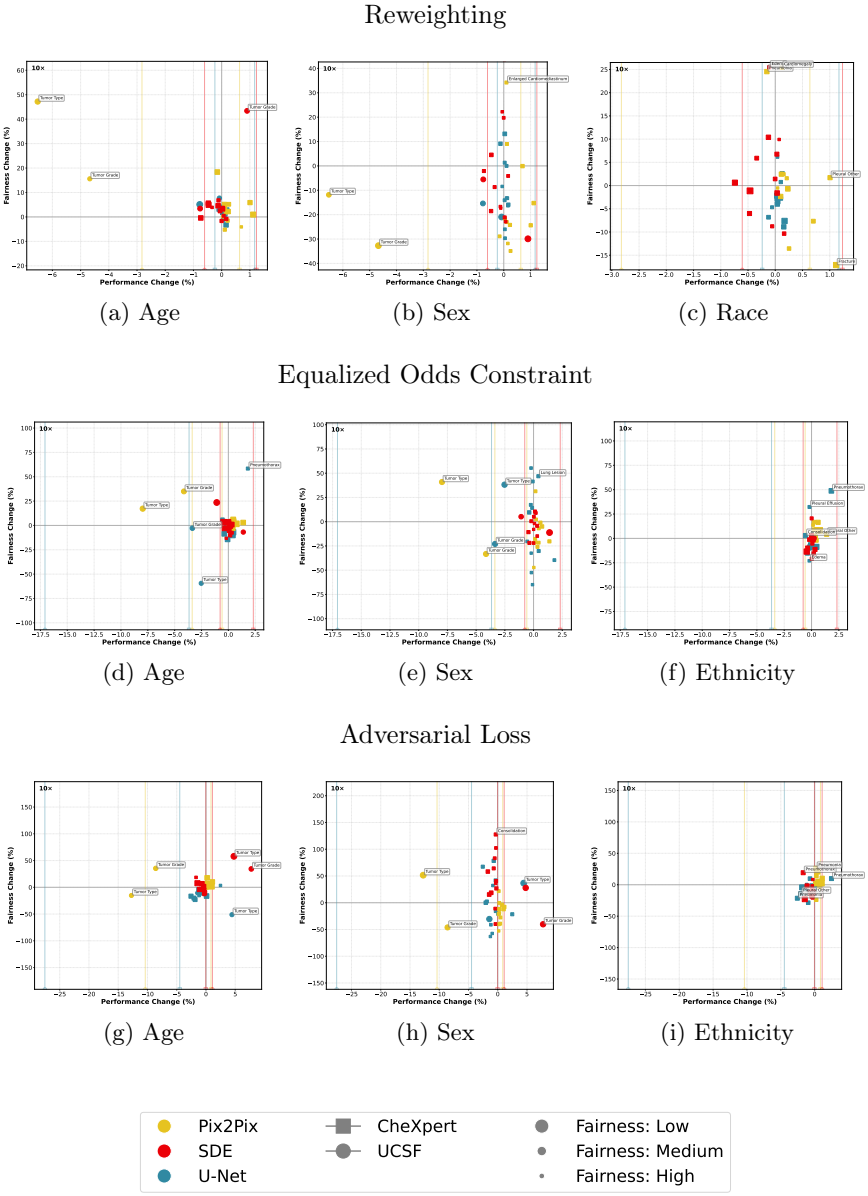
sities but failed to alter higher-level features influencing classifier decisions. This aligns with prior findings demonstrating negligible changes in downstream predictors despite significant variations in PSNR, apparent in Figure 9.

5 Conclusion

We observed modest yet statistically significant improvements in bias, particularly concerning age and, to a lesser extent, sex, through our mitigation techniques. However, the initial bias evaluation and subsequent mitigation analysis demonstrated that the reconstruction step had limited influence on bias reduction. We hypothesize that this limited effect arises because classifiers exhibit robustness against small perturbations, indicating insufficient changes introduced by reconstruction.

Consequently, reconstruction models inherently possess limited capacity to significantly influence downstream bias. When integrating pre-trained models into clinical workflows, efforts to address bias should target the predictive models rather than relying on adjustments at the reconstruction stage.

From a clinical perspective, ensuring unbiased and accurate diagnostics is critical. Therefore, our findings emphasize the necessity of carefully selecting intervention points within medical AI pipelines to effectively balance fairness and performance.



(a) Age (b) Sex (c) Race

(d) Age (e) Sex (f) Ethnicity

(g) Age (h) Sex (i) Ethnicity

Fig. 8: Combined visualization of equalized odds fairness (y-axis) and AUROC performance (x-axis) impacts of mitigation strategies. Shapes represent datasets, colors represent models, and vertical lines indicate PSNR values. Data is generally symmetrical around neutral points; fairness variance notably exceeds performance variance, highlighting UCSF as an outlier.

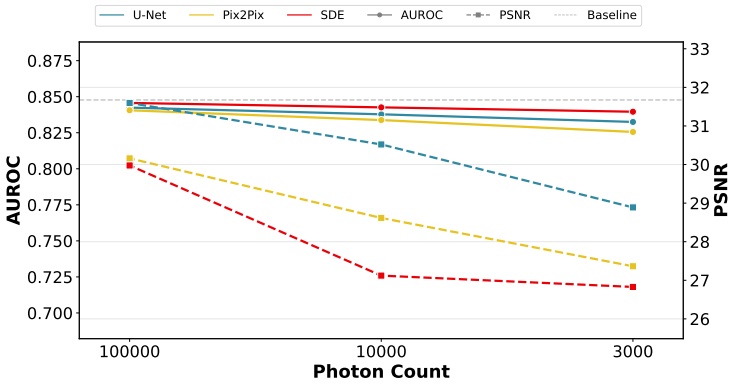


Fig. 9: CheXpert averaged classification performance (AUROC) and PSNR across different noise levels with baseline values of predictions on ground truth images. While PSNR drops with increased noise, the AUROC remains stable.

References

1. Adeli, E., Zhao, Q., Pfefferbaum, A., Sullivan, E.V., Fei-Fei, L., Niebles, J.C., Pohl, K.M.: Representation learning with statistical independence to mitigate bias. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV) pp. 2512–2522 (2019), https://api.semanticscholar.org/CorpusID:211069024
2. Bissoto, A., Fornaciali, M., Valle, E., Avila, S.: (De) Constructing Bias on Skin Lesion Datasets . In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2766–2774. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2019). https://doi.org/10.1109/CVPRW.2019.00335, https://doi.ieeecomputersociety.org/10.1109/CVPRW.2019.00335
3. Burlina, P., Joshi, N., Paul, W., Pacheco, K., Bressler, N.: Addressing artificial intelligence bias in retinal diagnostics. Translational Vision Science and Technology **10**, 13 (02 2021). https://doi.org/10.1167/tvst.10.2.13
4. Chiu, C.H., Chen, Y.J., Wu, Y., Shi, Y., Ho, T.Y.: Achieve fairness without demographics for dermatological disease diagnosis. Medical Image Analysis **95**, 103188 (2024). https://doi.org/https://doi.org/10.1016/j.media.2024.103188, https://www.sciencedirect.com/science/article/pii/S1361841524001130
5. Creager, E., Madras, D., Jacobsen, J.H., Weis, M.A., Swersky, K., Pitassi, T., Zemel, R.S.: Flexibly fair representation learning by disentanglement. In: International Conference on Machine Learning (2019), https://api.semanticscholar.org/CorpusID:174800294
6. Deng, W., Zhong, Y., Dou, Q., Li, X.: On fairness of medical image classification with multiple sensitive attributes via learning orthogonal representations. In: Information Processing in Medical Imaging: 28th International Conference, IPMI 2023, San Carlos de Bariloche, Argentina, June 18–23, 2023, Proceedings. p. 158–169. Springer-Verlag, Berlin, Heidelberg (2023). https://doi.org/10.1007/978-3-031-34048-2_13, https://doi.org/10.1007/978-3-031-34048-2_13
7. Du, S., Hers, B., Bayasi, N., Hamarneh, G., Garbi, R.: FairDisCo: Fairer AI in Dermatology via Disentanglement Contrastive Learning, pp. 185–202 (02 2023). https://doi.org/10.1007/978-3-031-25069-9_13
8. Du, Y., Xue, Y., Dharmakumar, R., Tsaftaris, S.A.: Unveiling fairness biases in deep learning-based brain mri reconstruction. In: Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging: 12th International Workshop, CLIP 2023 1st International Workshop, FAIMI 2023 and 2nd International Workshop, EPIMI 2023 Vancouver, BC, Canada, October 8 and October 12, 2023 Proceedings. p. 102–111. Springer-Verlag, Berlin, Heidelberg (2023). https://doi.org/10.1007/978-3-031-45249-9_10, https://doi.org/10.1007/978-3-031-45249-9_10
9. Glocker, B., Jones, C., Bernhardt, M., Winzeck, S.: Algorithmic encoding of protected characteristics in chest x-ray disease detection models. eBioMedicine **89** (2023), https://api.semanticscholar.org/CorpusID:256858498
10. Gong, S., Liu, X., Jain, A.K.: Mitigating face recognition bias via group adaptive classifier. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3413–3423 (2020), https://api.semanticscholar.org/CorpusID:219687431
11. Ioannou, S., Chockler, H., Hammers, A., King, A.P.: A study of demographic bias in cnn-based brain mr segmentation. In: Machine Learning in Clinical Neuroimaging: 5th International Workshop, MLCN 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings. p. 13–22. Springer-Verlag, Berlin,

Heidelberg (2022). https://doi.org/10.1007/978-3-031-17899-3_2, https://doi.org/10.1007/978-3-031-17899-3_2
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CVPR (2017)
13. Kim, B., Kim, H., Kim, K., Kim, S., Kim, J.: Learning not to learn: Training deep neural networks with biased data. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
14. Kinyanjui, N.M., Odonga, T., Cintas, C., Codella, N.C.F., Panda, R., Sattigeri, P., Varshney, K.R.: Fairness of classifiers across skin tones in dermatology. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. pp. 320–329. Springer International Publishing, Cham (2020)
15. Lara, M.A.R., Echeveste, R., Ferrante, E.: Addressing fairness in artificial intelligence for medical imaging. Nature Communications **13** (2022), https://api.semanticscholar.org/CorpusID:251371910
16. Lee, T., Puyol-Antón, E., Ruijsink, B., Shi, M., King, A.P.: A systematic study of race and sex bias in cnn-based cardiac mr segmentation. In: Statistical Atlases and Computational Models of the Heart. Regular and CMRxMotion Challenge Papers: 13th International Workshop, STACOM 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Revised Selected Papers. p. 233–244. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-23443-9_22, https://doi.org/10.1007/978-3-031-23443-9_22
17. Li, Y., Vasconcelos, N.: Repair: Removing representation bias by dataset resampling. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9564–9573 (2019). https://doi.org/10.1109/CVPR.2019.00980
18. Liu, E.Z., Haghgoo, B., Chen, A.S., Raghunathan, A., Koh, P.W., Sagawa, S., Liang, P., Finn, C.: Just train twice: Improving group robustness without training group information. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 6781–6792. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/liu21f.html
19. Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Image restoration with mean-reverting stochastic differential equations. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 23045–23066. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/luo23b.html
20. Marcinkevics, R., Ozkan, E., Vogt, J.E.: Debiasing deep chest x-ray classifiers using intra- and post-processing methods. In: Lipton, Z., Ranganath, R., Sendak, M., Sjoding, M., Yeung, S. (eds.) Proceedings of the 7th Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research, vol. 182, pp. 504–536. PMLR (05–06 Aug 2022), https://proceedings.mlr.press/v182/marcinkevics22a.html
21. Oguguo, T., Zamzmi, G., Rajaraman, S., Yang, F., Xue, Z., Antani, S.: A comparative study of fairness in medical machine learning. pp. 1–5 (04 2023). https://doi.org/10.1109/ISBI53787.2023.10230368
22. Puyol-Antón, E., Ruijsink, B., Mariscal Harana, J., Piechnik, S., Neubauer, S., Petersen, S., Razavi, R., Chowienczyk, P., King, A.: Fairness in cardiac magnetic resonance imaging: Assessing sex and racial bias in deep learning-based seg-

mentation. Frontiers in Cardiovascular Medicine **9**, 859310 (Apr 2022). https://doi.org/10.3389/fcvm.2022.859310
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. vol. 9351, pp. 234–241 (10 2015). https://doi.org/10.1007/978-3-319-24574-4_28
24. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In: International Conference on Learning Representations (ICLR) (2020)
25. Sarhan, M.H., Navab, N., Eslami, A., Albarqouni, S.: Fairness by learning orthogonal disentangled representations. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX. p. 746–761. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-58526-6_44, https://doi.org/10.1007/978-3-030-58526-6_44
26. Seyyed-Kalantari, L., Zhang, H., McDermott, M., Chen, I., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine **27** (12 2021). https://doi.org/10.1038/s41591-021-01595-0
27. Sheng, Y., Yang, J., Lin, Y., Jiang, W., Yang, L.: Toward fair ultrasound computing tomography: Challenges, solutions and outlook. In: Proceedings of the Great Lakes Symposium on VLSI 2024. p. 748–753. GLSVLSI '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3649476.3660387, https://doi.org/10.1145/3649476.3660387
28. Stanley, E.A.M., Wilms, M., Mouches, P., Forkert, N.D.: Fairness-related performance and explainability effects in deep learning models for brain image analysis. Journal of Medical Imaging **9**, 061102 – 061102 (2022), https://api.semanticscholar.org/CorpusID:251876386
29. Wang, R., Kuo, P.C., Chen, L.C., Seastedt, K.P., Gichoya, J.W., Celi, L.A.: Drop the shortcuts: image augmentation improves fairness and decreases ai detection of race and other demographics from medical images. eBioMedicine **102**, 105047 (2024). https://doi.org/https://doi.org/10.1016/j.ebiom.2024.105047, https://www.sciencedirect.com/science/article/pii/S2352396424000823
30. Wang, Z., Qinami, K., Karakozis, I.C., Genova, K., Nair, P., Hata, K., Russakovsky, O.: Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation . In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8916–8925. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2020). https://doi.org/10.1109/CVPR42600.2020.00894, https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00894
31. Yawen, W., Zeng, D., Xu, X., Shi, Y., Hu, J.: Fairprune: Achieving fairness through pruning for dermatological disease diagnosis. pp. 743–753. Springer Nature (sep 2022)
32. Yuan, C., Linn, K.A., Hubbard, R.A.: Algorithmic fairness of machine learning models for alzheimer disease progression. JAMA Network Open **6**(11), e2342203–e2342203 (11 2023). https://doi.org/10.1001/jamanetworkopen.2023.42203, https://doi.org/10.1001/jamanetworkopen.2023.42203
33. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. p. 335–340. AIES '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3278721.3278779, https://doi.org/10.1145/3278721.3278779

6 Appendix

6.1 Proof of Proportionality

When the protected attribute A takes more than two categories (e.g., multiple races, genders, or age groups), equalized odds typically compares all pairs a_i, a_j of subgroups. Then, one takes the maximum of the pairwise disparities in true positive and false positive rates:

$$EODD = \max_{1 \le i < j \le k} \Big[\big| P(\hat{Y} = 1 \mid Y = 1, A = a_i) \; - \; P(\hat{Y} = 1 \mid Y = 1, A = a_j) \big| + \big| P(\hat{Y} = 1 \mid Y = 0, A = a_i) \; - \; P(\hat{Y} = 1 \mid Y = 0, A = a_j) \big| \Big]$$

Each pairwise comparison is handled exactly as in the binary case by treating a_i, a_j as $0, 1$. Therefore, all the steps below—derived under a binary setup—apply pairwise to any two subgroups. Taking the maximum over these pairwise disparities then yields the multi-group measure.
This proof is based on the derivation by [20], and adjusted for EODD.
EODD measure the disparity in true positive rate (TPR) and false positive rate (FPR) between subgroups. In the binary case:

$$\begin{aligned} EODD = {} & P_{X,Y,A}(\hat{Y} = 1 | Y = 1, A = 1) - P_{X,Y|A}(\hat{Y} = 1 | Y = 1, A = 0) \\ & + P_{X,Y,A}(\hat{Y} = 1 | Y = 0, A = 1) - P_{X,Y,A}(\hat{Y} = 1 | Y = 0, A = 0) \end{aligned}$$

This can be expressed by the following proxy function.

$$EODD = \frac{\sum_{i=1}^{n} f_\theta(x_i) a_i y_i}{\sum_{i=1}^{n} a_i y_i} - \frac{\sum_{i=1}^{n} f_\theta(x_i)(1 - a_i) y_i}{\sum_{i=1}^{n} (1 - a_i) y_i} \tag{1}$$

$$+ \frac{\sum_{i=1}^{n} f_\theta(x_i) a_i (1 - y_i)}{\sum_{i=1}^{n} a_i (1 - y_i)} - \frac{\sum_{i=1}^{n} f_\theta(x_i)(1 - a_i)(1 - y_i)}{\sum_{i=1}^{n} (1 - a_i)(1 - y_i)} \tag{2}$$

To start, let's define the conditional covariance:

$$\begin{aligned} \operatorname{cov}(A, X | Y = y) &= \mathbb{E}[(A - \mathbb{E}[A|Y = y])(X - \mathbb{E}[X|Y = y])|Y = y] \\ &= \mathbb{E}[AX|Y = y] - \mathbb{E}[A|Y = y]\mathbb{E}[X|Y = y] \end{aligned} \tag{3}$$

We can use the law of total covariance to prove the validity:

$$\operatorname{cov}(A, X) = \mathbb{E}\Big[\operatorname{cov}(A, X|Y)\Big] + \operatorname{cov}\Big(\mathbb{E}[A|Y], \mathbb{E}[X|Y]\Big) \tag{4}$$

Expanding the first expectation term with (3):

$$\begin{aligned} \mathbb{E}[\operatorname{cov}(A, X|Y)] &= \mathbb{E}\Big[\mathbb{E}[AX|Y] - \mathbb{E}[A|Y]\mathbb{E}[X|Y]\Big] \\ &= \mathbb{E}[AX] - \mathbb{E}[\mathbb{E}[A|Y]\mathbb{E}[X|Y]] \end{aligned} \tag{5}$$

Expanding the second covariance term:

$$\text{cov}(\mathbb{E}[X|Z], \mathbb{E}[Y|Z]) = \mathbb{E}[\mathbb{E}[X|Z]\mathbb{E}[Y|Z]] - \mathbb{E}[X]\mathbb{E}[Y] \tag{6}$$

Substituting (5) and (6) into (4):

$$\begin{aligned}\text{cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[\mathbb{E}[X|Z]\mathbb{E}[Y|Z]] + \mathbb{E}[\mathbb{E}[X|Z]\mathbb{E}[Y|Z]] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \text{cov}(X, Y)\end{aligned}$$

We want to show that $\Delta_{OOD} \propto \widehat{\text{Cov}}(A, f_\theta(X)|Y = 1) + \widehat{\text{Cov}}(A, f_\theta(X)|Y = 0)$
Let $\sum_i a_i y_i = S_{AY}, \quad \sum_i a_i = S_A, \quad \sum_i y_i = S_Y$.

Expanding EODD:
Expanding (1):

$$\begin{aligned}&\frac{\sum_{i=1}^N f_\theta(x_i) a_i y_i}{\sum_{i=1}^N a_i y_i} - \frac{\sum_{i=1}^N f_\theta(x_i)(1 - a_i) y_i}{\sum_{i=1}^N y_i (1 - a_i) y_i} \\ &= \frac{1}{S_{AY}} \sum_{i=1}^N f_\theta(x_i) a_i y_i - \frac{1}{S_Y - S_A} \sum_{i=1}^N f_\theta(x_i) + \frac{1}{S_Y - S_{AY}} \sum_{i=1}^N f_\theta(x_i) a_i y_i \\ &= \frac{S_Y}{S_{AY}(S_Y - S_{AY})} \sum_{i=1}^N f_\theta(x_i) y_i a_i - \frac{1}{S_Y - S_{AY}} \sum_{i=1}^N f_\theta(x_i) y_i\end{aligned}$$

Note that:

$$\begin{aligned}\widehat{\text{Cov}}(A, f_\theta(X)|Y = 1) &= \frac{\sum_{i=1}^n f_\theta(x_i) a_i y_i}{\sum_{i=1}^n y_i} - \frac{\sum_{i=1}^n a_i y_i}{\sum_{i=1}^n y_i} \frac{\sum_{i=1}^n f_\theta(x_i) y_i}{\sum_{i=1}^n y_i} \\ &= \frac{1}{S_Y} \sum_{i=1}^n f_\theta(x_i) a_i y_i - \frac{S_{AY}}{S_Y^2} \sum_{i=1}^n f_\theta(x_i) y_i.\end{aligned}$$

Showing (5) $\propto \widehat{\text{Cov}}(A, f_\theta(X)|Y = 1)$ with factor $\frac{S_Y^2}{S_{AY}(S_Y - S_{AY})}$, independent of f_θ.

Expanding (2):

$$
\begin{aligned}
&\frac{\sum_{i=1}^{n} f_\theta(x_i) a_i (1-y_i)}{\sum_{i=1}^{n} a_i (1-y_i)} - \frac{\sum_{i=1}^{n} f_\theta(x_i)(1-a_i)(1-y_i)}{\sum_{i=1}^{n} (1-a_i)(1-y_i)} \\
&= \frac{N - S_Y}{(N - S_Y - S_A + S_{AY})(S_A - S_{AY})} \sum_{i=1}^{N} f_\theta(x_i) a_i \\
&\quad - \frac{N - S_Y}{(N - S_Y - S_A + S_{AY})(S_A - S_{AY})} \sum_{i=1}^{N} f_\theta(x_i) a_i y_i \\
&\quad - \frac{1}{N - S_Y - S_A + S_{AY}} \sum_{i=1}^{N} f_\theta(x_i) y_i \\
&\quad - \frac{N}{N - S_Y - S_A + S_{AY}} \sum_{i=1}^{N} f_\theta(x_i)
\end{aligned}
$$

Similarly:

$$
\begin{aligned}
\widehat{\mathrm{Cov}}(A, f_0(X) | Y = 0) &= \frac{\sum_{i=1}^{N} f_0(x_i) a_i (1-y_i)}{\sum_{i=1}^{N} (1-y_i)} - \frac{\sum_{i=1}^{N} a_i (1-y_i)}{\sum_{i=1}^{N} (1-y_i)} \cdot \frac{\sum_{i=1}^{N} f_0(x_i)(1-y_i)}{\sum_{i=1}^{N} (1-y_i)} \\
&= \frac{1}{N - S_Y} \sum_{i=1}^{N} f_0(x_i) a_i - \frac{N}{N - S_Y} \sum_{i=1}^{N} f_0(x_i) a_i y_i \\
&\quad - \frac{S_A - S_{AY}}{(N - S_Y)^2} \sum_{i=1}^{N} f_0(x_i) - \frac{S_A \cdot S_{AY}}{(N - S_Y)^2} \sum_{i=1}^{N} f_0(x_i) y_i
\end{aligned}
$$

Showing (6) $\propto \widehat{\mathrm{Cov}}(A, f_\theta(X) | Y = 0)$ with factor $\frac{(S_A - S_{AY})(N - S_Y - S_A + S_{AY})}{(N - S_Y)^2}$, independent of f_θ.

Therefore, $EODD \propto \widehat{\mathrm{Cov}}(A, f_\theta(X) | Y = 1) + \widehat{\mathrm{Cov}}(A, f_\theta(X) | Y = 0)$.

Table 3: Bias differences between baseline predictions on ground truth and reconstructed images across sensitive groups. Colored cells denote statistical significance (yellow: significantly worse; blue: significantly better); bold entries indicate cases where the standard deviation exceeds the effect size. For sex, there is better baseline fairness, yet slightly significant additional bias post-reconstruction, while age has a smaller baseline fairness but slightly improved bias after reconstruction. High standard deviation across racial groups, with no clear significant bias trends.

+, $p < 0.05$ +, $0.05 \leq p < 0.1$ −, $p < 0.05$ −, $0.05 \leq p < 0.1$
Bold indicates standard error larger than absolute effect size

(a) Sex

	Baseline		U-Net		Pix2Pix		SDE	
	EODD	**EOP**	**Δ EODD**	**Δ EOP**	**Δ EODD**	**Δ EOP**	**Δ EODD**	**Δ EOP**
EC	0.030	0.047	**-0.005**	**-0.003**	**-0.009**	-0.015	0.010	0.023
Cardiomegaly	0.021	0.039	**0.002**	0.006	**0.005**	**-0.004**	**0.003**	**0.005**
Lung Opacity	0.010	0.011	**0.000**	-0.008	0.011	**-0.001**	**-0.000**	-0.006
Lung Lesion	0.022	0.030	**0.006**	**0.011**	0.028	0.051	0.011	0.020
Edema	0.001	0.001	0.012	0.013	0.018	0.023	0.005	**0.004**
Consolidation	0.020	0.039	0.008	**-0.001**	0.018	**0.007**	**-0.003**	-0.020
Pneumonia	0.006	0.012	0.010	**0.016**	0.027	0.029	**0.008**	**0.014**
Atelectasis	0.014	0.005	0.015	0.017	0.026	0.020	**-0.003**	**-0.001**
Pneumothorax	0.043	0.067	**-0.000**	0.016	**0.005**	0.013	-0.007	-0.022
Pleural Effusion	0.015	0.016	0.013	0.009	0.026	0.021	0.009	0.005
Pleural Other	0.039	0.059	0.017	0.030	0.018	0.035	0.015	0.037
Fracture	0.046	0.061	0.009	0.025	0.016	0.054	0.009	0.021
Tumor Grade	0.287	0.146	-0.056	-0.112	**-0.024**	-0.120	**-0.024**	-0.120
Tumor Type	0.140	0.058	**-0.038**	**-0.020**	**-0.059**	**-0.007**	**-0.069**	**0.029**

	SER	**Δ Dice**	**Δ SER**	**$\Delta\Delta$ Dice**	**Δ SER**	**$\Delta\Delta$ Dice**	**Δ SER**	**$\Delta\Delta$ Dice**
Segmentation	1.099	0.026	**-0.000**	**0.002**	**-0.019**	**-0.004**	**-0.021**	**-0.004**

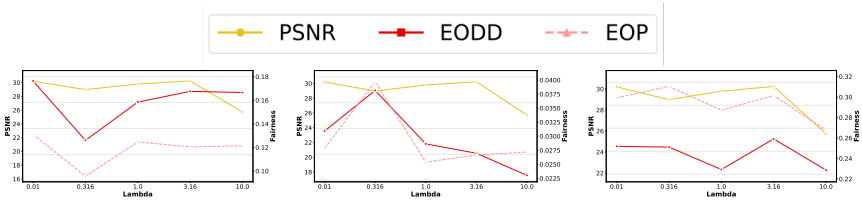
(b) Age

	Baseline		U-Net		Pix2Pix		SDE	
	EODD	**EOP**	**Δ EODD**	**Δ EOP**	**Δ EODD**	**Δ EOP**	**Δ EODD**	**Δ EOP**
EC	0.228	0.196	-0.009	-0.009	-0.018	-0.024	**-0.002**	**-0.008**
Cardiomegaly	0.127	0.081	-0.007	-0.011	**0.000**	**-0.001**	0.009	**0.003**
Lung Opacity	0.157	0.092	-0.009	-0.005	-0.010	-0.009	**-0.000**	-0.006
Lung Lesion	0.261	0.190	-0.017	-0.031	-0.025	-0.018	**-0.006**	**-0.009**
Edema	0.122	0.068	**-0.002**	**-0.004**	**-0.002**	**-0.001**	-0.009	-0.012
Consolidation	0.115	0.069	**-0.002**	**0.006**	**-0.002**	**-0.000**	**0.004**	**0.008**
Pneumonia	0.232	0.190	0.018	0.029	**-0.010**	**0.014**	**0.008**	**0.007**
Atelectasis	0.161	0.096	**-0.003**	**-0.004**	**-0.005**	**-0.002**	**0.002**	**-0.004**
Pneumothorax	0.055	0.016	**-0.004**	**-0.007**	**-0.008**	**0.004**	**0.000**	**0.009**
Pleural Effusion	0.083	0.049	-0.004	-0.006	**-0.002**	-0.005	0.006	**0.000**
Pleural Other	0.222	0.164	0.024	0.048	**-0.012**	**0.019**	**0.012**	0.028
Fracture	0.325	0.282	-0.011	**-0.005**	**-0.005**	**0.002**	-0.018	**-0.012**
Tumor Grade	0.058	0.061	0.043	0.086	**-0.037**	**-0.018**	**-0.011**	**0.033**
Tumor Type	0.167	0.252	0.229	0.167	**-0.049**	**-0.056**	**-0.004**	**0.033**

	SER	**Δ Dice**	**Δ SER**	**ΔΔ Dice**	**Δ SER**	**ΔΔ Dice**	**Δ SER**	**ΔΔ Dice**
Segmentation	1.224	0.057	**-0.016**	**0.000**	**-0.000**	**0.002**	**-0.009**	**0.000**

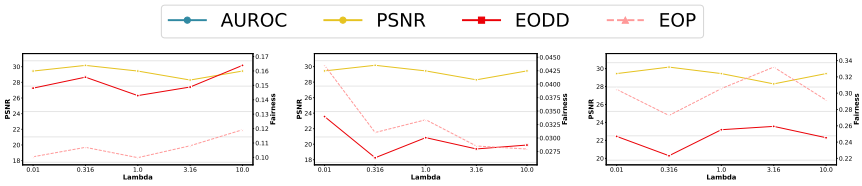
(c) Race

	Baseline		U-Net		Pix2Pix		SDE	
	EODD	**EOP**	**Δ EODD**	**Δ EOP**	**Δ EODD**	**Δ EOP**	**Δ EODD**	**Δ EOP**
EC	0.273	0.346	**0.003**	**0.000**	**-0.006**	**0.003**	**0.018**	**0.011**
Cardiomegaly	0.174	0.133	-0.038	**-0.008**	-0.064	**-0.010**	**-0.007**	**-0.008**
Lung Opacity	0.117	0.116	**0.018**	**-0.009**	**0.007**	**-0.021**	**0.027**	**0.001**
Lung Lesion	0.311	0.431	**0.017**	**-0.035**	**-0.043**	**-0.084**	**0.003**	**-0.022**
Edema	0.116	0.092	**0.010**	**0.005**	**0.013**	**0.015**	**-0.033**	**0.005**
Consolidation	0.149	0.192	0.092	0.159	0.088	0.163	**-0.007**	**-0.006**
Pneumonia	0.120	0.147	0.139	0.180	**-0.000**	**-0.021**	0.060	0.049
Atelectasis	0.180	0.203	**0.016**	**0.000**	**0.008**	**0.000**	**0.017**	**0.023**
Pneumothorax	0.189	0.215	**-0.003**	**0.007**	**0.020**	**0.009**	**-0.029**	**-0.007**
Pleural Effusion	0.061	0.045	**0.016**	0.027	**-0.002**	**0.009**	**0.001**	**0.010**
Pleural Other	0.123	0.100	0.072	**0.024**	**0.010**	**0.005**	**0.037**	0.100
Fracture	0.455	0.723	**-0.040**	**-0.167**	**-0.003**	**0.009**	**0.016**	**-0.001**



(a) Age (b) Sex (c) Race

Fig. 10: Impact of λ_{fair} on reconstruction quality (PSNR) compared to fairness for the EODD constraint mitigation. PSNR remains stable across lambda variations, while fairness shows slight variation without substantial improvement.



(a) Age (b) Sex (c) Race

Fig. 11: Impact of λ_{fair} on PSNR and fairness for the adversarial fairness loss. Stable PSNR across lambda values with minor fairness variations are similar to the EODD loss results.

Table 4: Fairness after fine-tuning with reweighting across age, sex, and race. Each table reports the fairness, along with standard deviations. Statistically significant changes compared to pre-mitigation are color-coded (yellow: significantly worse; blue: significantly better) based on the p-value. Reweighting yields the most limited effect with very little change across attributes.

$+$, $p < 0.05$ $+$, $0.05 \leq p < 0.1$ $-$, $p < 0.05$ $-$, $0.05 \leq p < 0.1$

(a) Sex

	U-Net		Pix2Pix		SDE	
	EODD	**EOP**	**EODD**	**EOP**	**EODD**	**EOP**
EC	0.027±0.005	0.044±0.009	0.031±0.008	0.052±0.010	0.032±0.006	0.053±0.010
Cardiomegaly	0.027±0.004	0.046±0.006	0.029±0.005	0.047±0.007	0.025±0.004	0.042±0.006
Lung Opacity	0.010±0.003	0.007±0.004	0.014±0.006	0.006±0.008	0.013±0.004	0.010±0.003
Lung Lesion	0.025±0.008	0.029±0.012	0.038±0.010	0.064±0.016	0.034±0.009	0.055±0.015
Edema	0.012±0.005	0.014±0.009	0.010±0.005	0.011±0.008	0.007±0.003	0.007±0.004
Consolidation	0.019±0.006	0.027±0.009	0.029±0.006	0.041±0.010	0.021±0.006	0.030±0.009
Pneumonia	0.018±0.009	0.022±0.013	0.024±0.011	0.030±0.016	0.018±0.008	0.024±0.014
Atelectasis	0.026±0.006	0.021±0.010	0.026±0.005	0.012±0.007	0.021±0.005	0.011±0.006
Pneumothorax	0.050±0.007	0.092±0.011	0.048±0.007	0.089±0.011	0.033±0.006	0.047±0.011
Pleural Effusion	0.021±0.003	0.017±0.004	0.032±0.003	0.027±0.004	0.019±0.003	0.016±0.004
Pleural Other	0.064±0.011	0.089±0.021	0.044±0.013	0.076±0.023	0.045±0.012	0.077±0.023
Fracture	0.046±0.008	0.067±0.013	0.054±0.009	0.089±0.015	0.058±0.008	0.091±0.014
Tumor Grade	0.199±0.090	0.091±0.130	0.195±0.078	0.089±0.105	0.204±0.071	0.100±0.064
Tumor Type	0.129±0.094	0.101±0.065	0.121±0.101	0.083±0.064	0.131±0.098	0.100±0.034

	SER	**Δ Dice**	**SER**	**Δ Dice**	**SER**	**Δ Dice**
Segmentation	1.122±0.027	0.033±0.006	1.109±0.032	0.030±0.008	1.111±0.029	0.030±0.007

(b) Age

	U-Net		Pix2Pix		SDE	
	EODD	**EOP**	**EODD**	**EOP**	**EODD**	**EOP**
EC	0.224±0.005	0.192±0.008	0.221±0.006	0.187±0.010	0.234±0.005	0.197±0.009
Cardiomegaly	0.120±0.004	0.071±0.006	0.131±0.004	0.079±0.006	0.135±0.005	0.085±0.006
Lung Opacity	0.151±0.004	0.088±0.003	0.151±0.004	0.088±0.003	0.160±0.004	0.094±0.003
Lung Lesion	0.253±0.008	0.178±0.013	0.241±0.009	0.163±0.014	0.254±0.009	0.174±0.014
Edema	0.124±0.003	0.071±0.005	0.118±0.004	0.063±0.005	0.121±0.003	0.067±0.004
Consolidation	0.111±0.005	0.074±0.009	0.107±0.006	0.062±0.010	0.117±0.005	0.072±0.008
Pneumonia	0.242±0.009	0.210±0.015	0.263±0.010	0.249±0.017	0.252±0.009	0.214±0.016
Atelectasis	0.160±0.005	0.090±0.006	0.164±0.006	0.101±0.007	0.168±0.005	0.095±0.006
Pneumothorax	0.057±0.008	0.013±0.015	0.046±0.007	0.013±0.013	0.058±0.006	0.014±0.012
Pleural Effusion	0.082±0.003	0.046±0.004	0.080±0.003	0.042±0.004	0.089±0.003	0.048±0.004
Pleural Other	0.252±0.012	0.214±0.021	0.223±0.013	0.183±0.023	0.248±0.012	0.209±0.023
Fracture	0.322±0.008	0.286±0.012	0.323±0.009	0.293±0.014	0.322±0.008	0.285±0.013
Tumor Grade	0.227±0.059	0.177±0.088	0.209±0.054	0.143±0.079	0.284±0.125	0.114±0.054
Tumor Type	0.441±0.130	0.442±0.080	0.350±0.125	0.256±0.055	0.288±0.046	0.317±0.045

	SER	**Δ Dice**	**SER**	**Δ Dice**	**SER**	**Δ Dice**
Segmentation	1.223±0.032	0.059±0.007	1.218±0.031	0.059±0.007	1.222±0.029	0.058±0.006

(c) Race

	U-Net		Pix2Pix		SDE	
	EODD	**EOP**	**EODD**	**EOP**	**EODD**	**EOP**
EC	0.288±0.028	0.354±0.011	0.276±0.021	0.348±0.011	0.299±0.031	0.357±0.012
Cardiomegaly	0.180±0.033	0.175±0.010	0.201±0.039	0.170±0.012	0.183±0.035	0.166±0.012
Lung Opacity	0.165±0.030	0.132±0.031	0.159±0.034	0.137±0.044	0.155±0.037	0.129±0.028
Lung Lesion	0.374±0.040	0.492±0.042	0.305±0.116	0.354±0.218	0.376±0.031	0.485±0.035
Edema	0.140±0.018	0.115±0.033	0.154±0.017	0.131±0.027	0.154±0.015	0.145±0.027
Consolidation	0.269±0.079	0.385±0.155	0.270±0.083	0.392±0.157	0.202±0.012	0.271±0.016
Pneumonia	0.282±0.075	0.370±0.144	0.278±0.078	0.377±0.143	0.279±0.072	0.360±0.138
Atelectasis	0.227±0.042	0.210±0.050	0.186±0.038	0.170±0.056	0.239±0.033	0.250±0.020
Pneumothorax	0.207±0.057	0.228±0.099	0.219±0.031	0.277±0.026	0.218±0.032	0.267±0.030
Pleural Effusion	0.117±0.029	0.100±0.024	0.096±0.025	0.089±0.023	0.103±0.026	0.080±0.031
Pleural Other	0.285±0.035	0.320±0.023	0.258±0.056	0.342±0.092	0.249±0.034	0.306±0.034
Fracture	0.406±0.115	0.543±0.218	0.397±0.116	0.540±0.220	0.485±0.035	0.728±0.059

Table 5: Fairness after fine-tuning with the equalized odds constraint across age, sex, and race. Each table reports the fairness, along with standard deviations. Statistically significant changes compared to pre-mitigation are color-coded (yellow: significantly worse; blue: significantly better) based on the p-value. The EODD loss achieves moderate fairness gains for age and sex and no meaningful change for race, likely due to high variance in that attribute.

$+,\ p < 0.05$ $+,\ 0.05 \leq p < 0.1$ $-,\ p < 0.05$ $-,\ 0.05 \leq p < 0.1$

(a) Sex

	U-Net		Pix2Pix		SDE	
	EODD	**EOP**	**EODD**	**EOP**	**EODD**	**EOP**
EC	0.031±0.007	0.041±0.012	0.030±0.009	0.048±0.010	0.039±0.005	0.065±0.009
Cardiomegaly	0.032±0.007	0.049±0.008	0.027±0.006	0.035±0.006	0.028±0.005	0.041±0.006
Lung Opacity	0.018±0.008	0.023±0.005	0.016±0.006	0.005±0.008	0.011±0.003	0.010±0.003
Lung Lesion	0.043±0.018	0.033±0.034	0.037±0.011	0.062±0.015	0.038±0.009	0.060±0.016
Edema	0.009±0.005	0.009±0.006	0.010±0.005	0.012±0.009	0.008±0.004	0.006±0.004
Consolidation	0.022±0.006	0.036±0.011	0.030±0.006	0.043±0.009	0.019±0.005	0.031±0.008
Pneumonia	0.031±0.015	0.024±0.024	0.032±0.012	0.051±0.022	0.017±0.007	0.023±0.013
Atelectasis	0.011±0.009	0.010±0.010	0.032±0.006	0.020±0.009	0.016±0.005	0.013±0.007
Pneumothorax	0.028±0.010	0.013±0.018	0.048±0.008	0.089±0.011	0.036±0.006	0.051±0.011
Pleural Effusion	0.014±0.004	0.010±0.005	0.034±0.003	0.028±0.004	0.021±0.003	0.019±0.004
Pleural Other	0.062±0.017	0.040±0.031	0.046±0.015	0.081±0.027	0.043±0.012	0.068±0.022
Fracture	0.038±0.012	0.023±0.022	0.061±0.009	0.106±0.014	0.050±0.008	0.080±0.014
Tumor Grade	0.194±0.105	0.086±0.104	0.194±0.084	0.085±0.107	0.259±0.056	0.085±0.113
Tumor Type	0.211±0.102	0.271±0.116	0.193±0.040	0.087±0.049	0.146±0.067	0.112±0.033

	SER	**Δ Dice**	**SER**	**Δ Dice**	**SER**	**Δ Dice**
Segmentation	1.107±0.063	0.039±0.018	1.111±0.032	0.031±0.008	1.109±0.030	0.029±0.007

(b) Age

	U-Net		Pix2Pix		SDE	
	EODD	**EOP**	**EODD**	**EOP**	**EODD**	**EOP**
EC	0.221±0.008	0.208±0.012	0.218±0.006	0.189±0.010	0.229±0.006	0.197±0.009
Cardiomegaly	0.117±0.006	0.069±0.008	0.125±0.004	0.079±0.006	0.132±0.004	0.087±0.006
Lung Opacity	0.142±0.006	0.081±0.005	0.145±0.005	0.083±0.003	0.148±0.004	0.084±0.003
Lung Lesion	0.219±0.012	0.138±0.018	0.246±0.009	0.184±0.015	0.257±0.009	0.185±0.014
Edema	0.113±0.005	0.067±0.006	0.119±0.004	0.062±0.005	0.116±0.004	0.068±0.005
Consolidation	0.119±0.007	0.083±0.011	0.112±0.005	0.072±0.008	0.107±0.005	0.064±0.009
Pneumonia	0.214±0.013	0.174±0.022	0.236±0.010	0.223±0.016	0.249±0.009	0.222±0.014
Atelectasis	0.157±0.007	0.089±0.009	0.151±0.006	0.082±0.006	0.161±0.005	0.098±0.006
Pneumothorax	0.087±0.008	0.056±0.015	0.051±0.012	0.021±0.023	0.048±0.007	0.013±0.012
Pleural Effusion	0.077±0.004	0.038±0.005	0.081±0.003	0.046±0.004	0.082±0.003	0.044±0.004
Pleural Other	0.226±0.018	0.170±0.032	0.217±0.014	0.182±0.026	0.228±0.013	0.177±0.024
Fracture	0.292±0.011	0.269±0.018	0.325±0.008	0.300±0.014	0.317±0.008	0.291±0.013
Tumor Grade	0.205±0.082	0.146±0.106	0.244±0.058	0.215±0.090	0.185±0.028	0.107±0.055
Tumor Type	0.170±0.126	0.089±0.132	0.279±0.038	0.288±0.047	0.344±0.124	0.224±0.025

	SER	**Δ Dice**	**SER**	**Δ Dice**	**SER**	**Δ Dice**
Segmentation	1.090±0.102	0.033±0.024	1.213±0.031	0.057±0.007	1.227±0.031	0.059±0.007

(c) Race

	U-Net		Pix2Pix		SDE	
	EODD	**EOP**	**EODD**	**EOP**	**EODD**	**EOP**
EC	0.269±0.033	0.350±0.017	0.291±0.022	0.356±0.012	0.270±0.026	0.349±0.013
Cardiomegaly	0.166±0.058	0.171±0.014	0.187±0.038	0.173±0.013	0.161±0.046	0.166±0.011
Lung Opacity	0.150±0.038	0.137±0.044	0.170±0.031	0.124±0.037	0.161±0.044	0.143±0.059
Lung Lesion	0.349±0.059	0.469±0.052	0.311±0.118	0.381±0.226	0.373±0.029	0.483±0.020
Edema	0.113±0.041	0.109±0.041	0.147±0.017	0.122±0.030	0.148±0.015	0.130±0.026
Consolidation	0.286±0.087	0.386±0.156	0.268±0.079	0.392±0.153	0.195±0.012	0.267±0.019
Pneumonia	0.289±0.080	0.391±0.129	0.244±0.036	0.301±0.044	0.260±0.079	0.353±0.147
Atelectasis	0.207±0.063	0.180±0.048	0.210±0.040	0.190±0.041	0.189±0.040	0.190±0.047
Pneumothorax	0.330±0.100	0.382±0.189	0.277±0.052	0.358±0.082	0.203±0.027	0.272±0.025
Pleural Effusion	0.146±0.047	0.068±0.040	0.110±0.032	0.095±0.034	0.094±0.026	0.090±0.021
Pleural Other	0.252±0.073	0.331±0.114	0.265±0.052	0.303±0.093	0.239±0.065	0.290±0.090
Fracture	0.402±0.097	0.566±0.180	0.518±0.053	0.720±0.061	0.425±0.113	0.507±0.206

Table 6: Fairness after fine-tuning with the adversarial loss across age, sex, and race. Each table reports the fairness, along with standard deviations. Statistically significant changes compared to pre-mitigation are color-coded (yellow: significantly worse; blue: significantly better) based on the p-value. The adversarial loss achieves the most consistent fairness gains for age, moderate improvements for sex, and no meaningful change for race, likely due to high variance in that attribute.

$+,\ p < 0.05$ $+,\ 0.05 \leq p < 0.1$ $-,\ p < 0.05$ $-,\ 0.05 \leq p < 0.1$

(a) Sex

	U-Net		Pix2Pix		SDE	
	EODD	**EOP**	**EODD**	**EOP**	**EODD**	**EOP**
EC	0.038±0.008	0.052±0.013	0.028±0.008	0.047±0.010	0.025±0.006	0.033±0.010
Cardiomegaly	0.048±0.008	0.067±0.009	0.026±0.004	0.043±0.006	0.023±0.004	0.035±0.006
Lung Opacity	0.015±0.006	0.021±0.005	0.017±0.006	0.005±0.008	0.022±0.004	0.006±0.003
Lung Lesion	0.030±0.012	0.019±0.021	0.031±0.010	0.050±0.015	0.044±0.009	0.044±0.015
Edema	0.011±0.006	0.007±0.006	0.009±0.004	0.010±0.007	0.012±0.003	0.013±0.004
Consolidation	0.012±0.007	0.016±0.013	0.033±0.007	0.045±0.010	0.040±0.005	0.051±0.009
Pneumonia	0.037±0.012	0.022±0.017	0.021±0.009	0.028±0.016	0.035±0.008	0.030±0.014
Atelectasis	0.011±0.008	0.012±0.011	0.029±0.006	0.017±0.009	0.026±0.005	0.012±0.008
Pneumothorax	0.036±0.011	0.013±0.020	0.047±0.007	0.086±0.011	0.057±0.006	0.095±0.010
Pleural Effusion	0.017±0.004	0.016±0.005	0.032±0.003	0.030±0.004	0.034±0.003	0.031±0.004
Pleural Other	0.056±0.016	0.053±0.030	0.053±0.014	0.095±0.025	0.066±0.011	0.104±0.020
Fracture	0.044±0.012	0.044±0.020	0.056±0.009	0.099±0.015	0.064±0.009	0.113±0.016
Tumor Grade	0.175±0.094	0.102±0.092	0.156±0.110	0.098±0.132	0.174±0.057	0.092±0.085
Tumor Type	0.209±0.101	0.270±0.104	0.207±0.131	0.123±0.091	0.177±0.050	0.168±0.079

	SER	**Δ Dice**	**SER**	**Δ Dice**	**SER**	**Δ Dice**
Segmentation	1.074±0.073	0.030±0.019	1.108±0.037	0.032±0.009	1.091±0.050	0.029±0.014

(b) Age

	U-Net		Pix2Pix		SDE	
	EODD	**EOP**	**EODD**	**EOP**	**EODD**	**EOP**
EC	0.195±0.008	0.172±0.013	0.224±0.006	0.185±0.010	0.230±0.005	0.211±0.009
Cardiomegaly	0.111±0.006	0.065±0.008	0.127±0.004	0.069±0.006	0.136±0.004	0.091±0.006
Lung Opacity	0.133±0.006	0.063±0.005	0.149±0.004	0.083±0.003	0.149±0.004	0.084±0.003
Lung Lesion	0.188±0.012	0.125±0.018	0.253±0.009	0.176±0.014	0.229±0.009	0.155±0.014
Edema	0.105±0.005	0.055±0.007	0.122±0.003	0.067±0.005	0.114±0.003	0.057±0.005
Consolidation	0.100±0.007	0.067±0.012	0.118±0.006	0.082±0.009	0.106±0.005	0.056±0.008
Pneumonia	0.209±0.013	0.183±0.022	0.261±0.009	0.239±0.015	0.257±0.009	0.231±0.016
Atelectasis	0.152±0.008	0.085±0.009	0.157±0.006	0.091±0.006	0.158±0.005	0.089±0.006
Pneumothorax	0.057±0.008	0.020±0.014	0.051±0.009	0.014±0.017	0.066±0.005	0.034±0.010
Pleural Effusion	0.067±0.005	0.027±0.006	0.079±0.003	0.042±0.004	0.077±0.003	0.038±0.004
Pleural Other	0.197±0.018	0.143±0.032	0.233±0.013	0.189±0.023	0.225±0.012	0.166±0.023
Fracture	0.263±0.012	0.245±0.019	0.325±0.009	0.294±0.014	0.330±0.009	0.315±0.014
Tumor Grade	0.189±0.069	0.107±0.087	0.245±0.126	0.090±0.089	0.266±0.042	0.249±0.068
Tumor Type	0.205±0.142	0.141±0.112	0.201±0.096	0.121±0.104	0.440±0.051	0.274±0.088

	SER	**Δ Dice**	**SER**	**Δ Dice**	**SER**	**Δ Dice**
Segmentation	1.073±0.113	0.030±0.027	1.168±0.042	0.048±0.009	1.326±0.032	0.094±0.006

(c) Race

	U-Net		Pix2Pix		SDE	
	EODD	**EOP**	**EODD**	**EOP**	**EODD**	**EOP**
EC	0.291±0.034	0.350±0.025	0.287±0.025	0.355±0.017	0.249±0.038	0.339±0.010
Cardiomegaly	0.203±0.049	0.219±0.070	0.182±0.049	0.181±0.010	0.178±0.034	0.166±0.017
Lung Opacity	0.145±0.039	0.121±0.038	0.161±0.027	0.121±0.031	0.135±0.049	0.127±0.025
Lung Lesion	0.366±0.060	0.475±0.050	0.372±0.037	0.477±0.031	0.330±0.048	0.486±0.029
Edema	0.125±0.036	0.139±0.042	0.154±0.017	0.133±0.029	0.120±0.018	0.115±0.033
Consolidation	0.199±0.032	0.275±0.031	0.201±0.028	0.276±0.025	0.184±0.015	0.246±0.022
Pneumonia	0.243±0.051	0.285±0.062	0.281±0.075	0.365±0.141	0.222±0.042	0.273±0.044
Atelectasis	0.207±0.050	0.159±0.076	0.227±0.043	0.243±0.036	0.187±0.048	0.193±0.048
Pneumothorax	0.244±0.084	0.318±0.155	0.230±0.033	0.272±0.015	0.245±0.023	0.358±0.026
Pleural Effusion	0.111±0.041	0.081±0.045	0.110±0.037	0.109±0.057	0.102±0.015	0.087±0.023
Pleural Other	0.254±0.045	0.313±0.048	0.280±0.054	0.358±0.089	0.260±0.052	0.312±0.083
Fracture	0.425±0.093	0.513±0.149	0.492±0.038	0.721±0.023	0.378±0.096	0.565±0.173

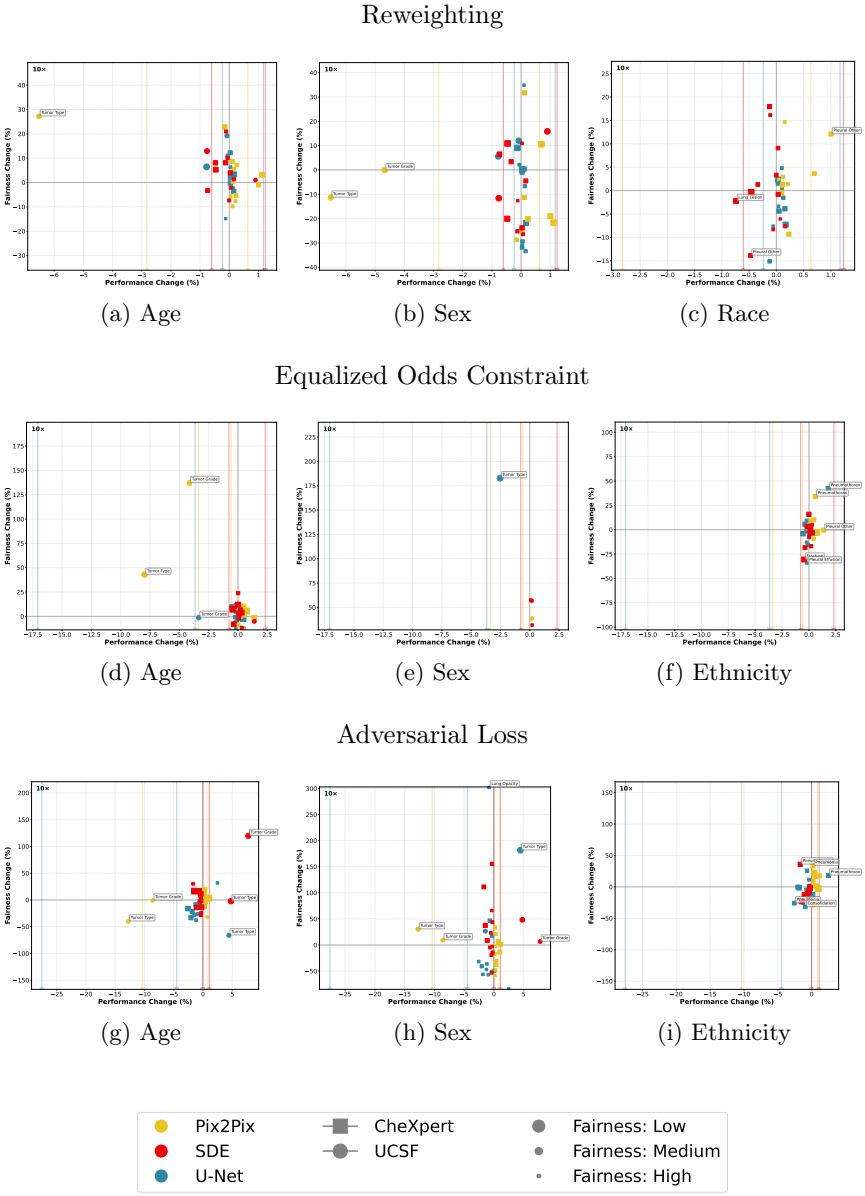


Fig. 12: Analogous combined visualization of equalized opportunity fairness (y-axis) and AUROC performance (x-axis) impacts of mitigation strategies. Shapes represent datasets, colors represent models, and vertical lines indicate PSNR values. Similar trends to EODD fairness were observed, with higher fairness variability than performance.