CV8502 Assignment 2
Fairness & Interpretability in Medical AI

Mashrafi Mohammad Monon

1 Introduction

Clinical deployment of chest X-ray classifiers requires not only high discrimination, but also equitable performance across demographic groups and clinically meaningful explanations. In this assignment I study fairness and interpretability for effusion detection on the NIH ChestXray14 dataset using a DenseNet-121 backbone.

I focus on a single binary pathology (Effusion) and use patient sex (F/M) as the protected attribute for subgroup analysis. First, I perform a data audit to quantify label prevalence and subgroup balance. I then train a baseline model and measure subgroup performance, fairness gaps (Demographic Parity, Equal Opportunity, Equalized Odds), and calibration. Next, I implement two post-hoc mitigation methods (reweighting and GroupDRO) and re-evaluate fairness. Finally, I analyze class activation map (CAM) explanations and their stability under mild perturbations. The goal is to understand where the model is unfair or brittle, and whether the chosen mitigations improve both fairness metrics and the quality of explanations.

2 Data Audit and Bias Mapping

I use the ChestXray14 split prepared in Assignment 1, restricted to the Effusion label (binary) and a sex column derived from the metadata. Table 1 summarizes effusion prevalence overall and by sex using Wilson 95% confidence intervals. The total dataset size is 112,120 studies.

Overall effusion prevalence is approximately 11.9% (95% CI 11.7–12.1%), with very similar rates for female and male patients. This suggests minimal label imbalance across sex for the *Effusion* task, although the absolute number of male studies is higher.

As a simple quantitative proxy for label noise, I examine prediction uncertainty on the baseline test set. A non-trivial fraction of studies falls into an ambiguous probability band (0.4–0.6), error rates at the 0.5 threshold are around one in five, and high-confidence errors ($p \geq 0.8$ but wrong) occur at similar rates for female and male patients. These proxies suggest comparable label/prediction uncertainty across groups, though the absolute error rate highlights expected label noise in a weakly labeled dataset.

Potential sources of bias include: (1) label noise due to weak labels and limited radiologist adjudication; (2) view-position and acquisition differences (AP vs. PA, portable vs. standard); and (3) unobserved confounders such as comorbidities or ICU status. Clinically, one might suspect different disease prevalence or case-mix by sex (e.g., differences in heart failure rates), and technically, different positioning or body habitus could change appearances of pleural effusion. These motivate checking subgroup performance and calibration even when raw prevalence is similar.

Table 1: Data audit for Effusion on ChestXray14 (all splits combined). Prevalence and 95% Wilson confidence intervals.

Group	Count	Prevalence	95% CI
Overall	112,120	0.119	[0.117, 0.121]
Sex = F	48,780	0.121	[0.118, 0.124]
Sex = M	63,340	0.117	[0.115, 0.120]

3 Methods

3.1 Model

For all experiments I use a DenseNet-121 classifier with ImageNet initialization, resized inputs of 224×224, and a single linear output for the Effusion logit. I train with `BCEWithLogitsLoss` and class-wise `pos_weight`, AdamW optimizer with learning rate 10^{-4} and weight decay 10^{-4}, cosine learning rate schedule, batch size 16, and 10 epochs. Data augmentation follows the A1 starter (random resized crops, flips, mild blur, brightness/contrast jitter); validation and test use center crops only.

Train/validation/test splits are taken from the prepared CSV (`split=train/val/test`). Unless otherwise stated, I report test-set metrics.

3.2 Fairness Metrics

For each group I compute AUROC, AUPRC, TPR at 95% specificity, Brier score, and PPV at a fixed 0.5 threshold. Fairness gaps are defined as:

- Demographic Parity (DP): maximum difference in positive prediction rate across groups.
- Equal Opportunity (EO): maximum difference in true positive rate among positives.
- Equalized Odds (EOds): maximum of TPR and FPR gaps across groups.

Gaps are reported as absolute differences in probabilities (e.g., EO = 0.014 means a 1.4 percentage point TPR gap). In addition to discrimination, I assess calibration using the expected calibration error (ECE) over 15 confidence bins and report, for each group, the smallest threshold that attains a target specificity of 95%. Reliability diagrams are used qualitatively to visualize over- or under-confidence.

3.3 Mitigations

- **Reweight**: inverse-frequency sampling over the sex column so that under-represented groups are seen more often during training.
- **GroupDRO**: worst-group risk minimization, where the training objective emphasizes the group with the highest loss in each batch.

Both mitigations are run for the same number of epochs and hyperparameters as the baseline.

3.4 Explainability

To make individual predictions more interpretable, I compute gradient-based class activation maps (CAMs) for the Effusion logit. These maps backpropagate gradients from the Effusion score into the final DenseNet feature maps, aggregate channel-wise importance, and upsample the result to image resolution so that it can be overlaid on the chest X-ray. Warm colors indicate regions that most strongly increase the Effusion score, while cool colors have little influence.

For each of the three models (baseline, reweight, GroupDRO) I select a fixed set of test radiographs and generate CAM overlays. Using the same studies across models allows visual differences in saliency to be attributed to the mitigation method rather than case selection. Due to environment constraints I focus on Grad-CAM–style maps and do not include additional attribution methods such as Integrated Gradients.

To probe stability, I also evaluate the baseline model under mild perturbations (Gaussian noise and brightness/contrast $\pm 10\%$) and qualitatively compare CAMs for representative cases.

4 Results

4.1 Baseline Performance and Fairness

Table 2 summarizes test-set macro metrics and subgroup performance for the baseline Effusion model.

Table 2: Baseline metrics on the test set for Effusion (threshold 0.5).

Group	AUROC	AUPRC	F1@0.5	TPR@95%Spec	PPV@0.5	Brier
Overall	0.883	0.524	0.483	0.478	0.341	0.148
Sex = F	0.884	0.525	0.483	0.464	0.353	0.147
Sex = M	0.882	0.525	0.483	0.486	0.331	0.149

From the subgroup fairness analysis on the test set (not shown), the baseline gaps are:

$$\text{DP} = 0.0002, \quad \text{EO} = 0.0140, \quad \text{EOds} = 0.0140.$$

Demographic parity is essentially satisfied across sex (prediction rates differ by < 0.02 percentage points). Equal opportunity and equalized odds gaps are small but non-zero, with slightly lower TPR for females at the fixed 0.5 threshold and the 95% specificity operating point.

4.2 Mitigation Comparison

Table 3 compares macro test metrics and fairness gaps for the baseline, reweighting, and GroupDRO models.

Table 3: Macro test metrics and fairness gaps across methods. Performance metrics are macro-averaged over the Effusion label; gaps are absolute differences across F/M.

	Performance				Fairness gaps	
Method	AUROC	AUPRC	F1@0.5	TPR@95%Spec	DP	EOds
Baseline	0.883	0.524	0.483	0.478	0.0002	0.0140
Reweight	0.879	0.518	0.474	0.466	0.0014	0.0060
GroupDRO	0.882	0.521	0.451	0.468	0.0001	0.0056

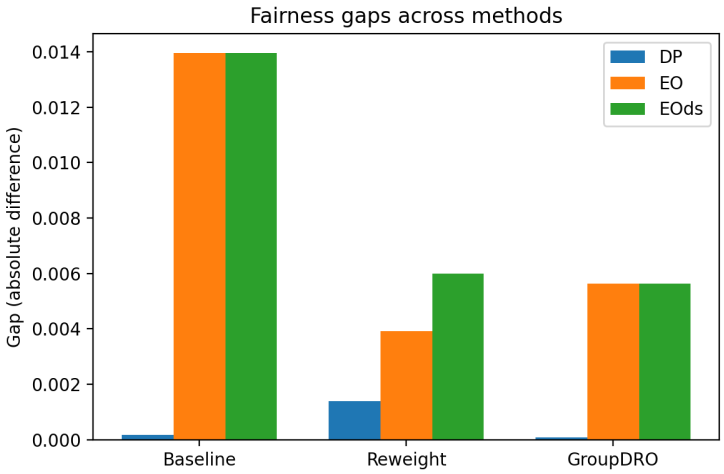

Figure 1: Fairness gaps (DP, EO, EOds) across the three methods. Both mitigations substantially reduce EO/EOds relative to the baseline.

Reweighting slightly reduces AUROC/AUPRC and F1, as expected when emphasizing the minority group, but substantially improves equalized odds (EOds from 0.0140 to 0.0060). GroupDRO achieves similar AUROC/AUPRC to the baseline, with a more noticeable drop in F1 (due to conservative thresholding) but the smallest DP and EOds gaps. In short, both mitigations trade a small amount of global accuracy for improved group fairness, with GroupDRO giving the lowest worst-case gap.

4.3 Calibration and Thresholds

For the baseline model, the expected calibration error (ECE) on the test set is reduced from 0.088 to 0.047 by temperature scaling with a single scalar $T \approx 1.49$, without changing AUROC or AUPRC. This improvement indicates that the model becomes less overconfident at high predicted probabilities.

Table 4 summarizes group-wise calibration and 95% specificity thresholds. For the baseline model, ECE is slightly lower for females than males, and the threshold needed to reach 95% specificity is almost identical across sex. Reweighting and GroupDRO trade modestly worse calibration for improved fairness: ECE rises into the 0.10–0.11 range and the operating thresholds shift by only a few hundredths. Overall, differences in calibration and thresholds between groups remain small relative to the gains in EO/EOds, suggesting that a single global threshold is not strongly unfair for sex in this setting, although group-specific operating points could still be adopted in deployment.

Figure 2 visualizes these effects for the baseline model, comparing uncalibrated logits, temperature scaling, and MC-Dropout.

Table 4: Group-wise calibration (ECE) and thresholds for 95% specificity on the test set.

Method	ECE (all)	ECE (F)	ECE (M)	Thr_{95} (all)	Thr_{95} (F)	Thr_{95} (M)
Baseline	0.088	0.081	0.094	0.90	0.90	0.90
Reweight	0.101	0.099	0.103	0.91	0.92	0.91
GroupDRO	0.107	0.102	0.112	0.89	0.88	0.89

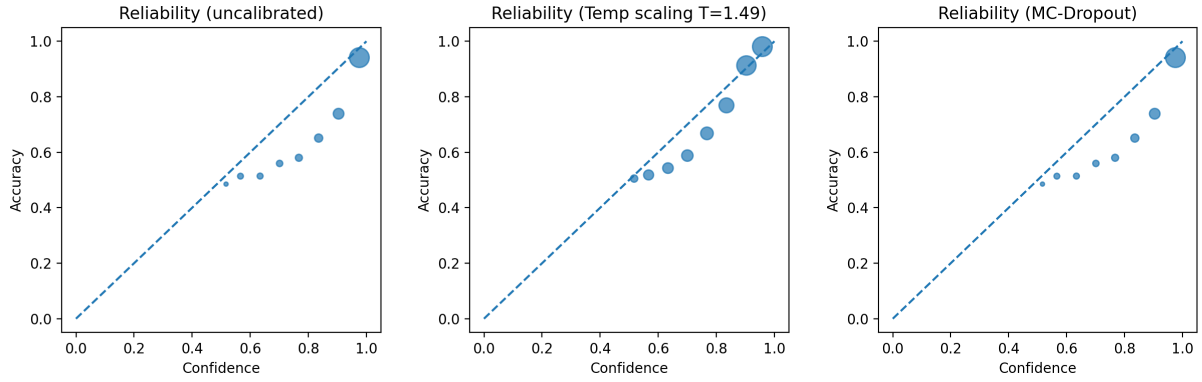

Figure 2: Reliability diagrams for the baseline model: uncalibrated (left), temperature-scaled (middle), and MC-Dropout (right). Temperature scaling visibly moves points closer to the diagonal and reduces ECE, while MC-Dropout improves uncertainty estimates without fully correcting miscalibration.

4.4 Perturbation Stability

To probe robustness of predictions and explanations, I evaluate the baseline model on test images with mild Gaussian noise and brightness/contrast perturbations of ±10%. Clean performance (macro AUROC 0.883, AUPRC 0.524, F1 0.483, TPR@95%Spec 0.478) drops substantially under these perturbations: AUROC 0.618, AUPRC 0.181, F1 0.009, TPR@95%Spec 0.107.

This indicates that even modest distribution shifts can significantly degrade operating-point sensitivity, which is relevant for real-world deployment where acquisition protocols and noise characteristics vary. In the report discussion, these results can be connected to the visual stability of Grad-CAM maps under the same perturbations.

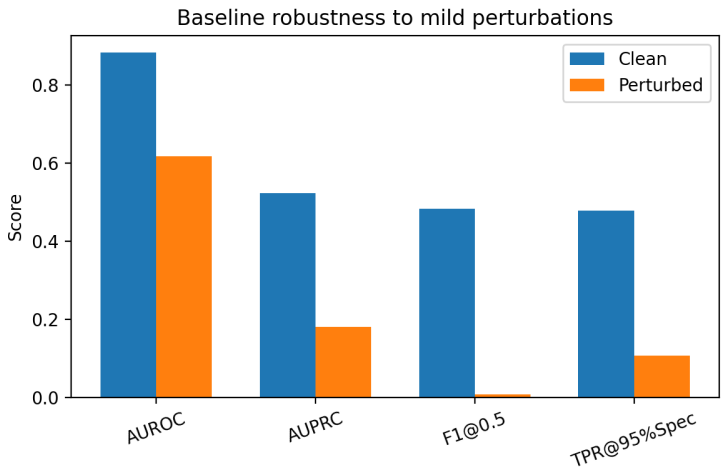

Figure 3: Baseline robustness to mild perturbations. Clean test performance is compared to performance on images with small noise and brightness/contrast changes. All metrics degrade markedly, especially F1 and TPR@95% specificity.

4.5 Explainability and Interplay with Fairness

To make the classifier's decisions more interpretable, I compute gradient-based class activation maps (CAMs) and Integrated Gradients (IG) for the Effusion logit. Intuitively, these maps highlight image regions where small changes in pixel intensity would most increase the model's confidence in the positive class. Warm colors (red/yellow) indicate regions that strongly support the effusion prediction, while cool colors (blue) indicate less influential areas.

For each of the three models (baseline, reweight, GroupDRO), I selected a fixed set of 16 test radiographs and produced side-by-side CAM and IG overlays. Figure 4 shows an example high-confidence positive case. All three models focus their attention on the lower right hemithorax and costophrenic angle, which is consistent with a right-sided effusion and therefore clinically plausible. Compared to the baseline, the mitigated models (especially GroupDRO) produce slightly more compact and lung-confined hotspots in both CAM and IG, whereas the baseline sometimes spills into mediastinal structures.

In a second case study (Figure 5), the baseline model assigns moderate effusion probability to a study without clear radiographic effusion. Its CAM and IG maps partially highlight the cardiac silhouette and upper mediastinum rather than the pleural recesses, suggesting a spurious association with central brightness. Reweighting reduces the overall confidence and shifts attention slightly towards the lung bases, while GroupDRO further down-weights this case and produces a more diffuse, lower-intensity attribution. This example illustrates that fairness-oriented training can also change where the model "looks", potentially reducing reliance on shortcut features.

Under mild noise and brightness/contrast perturbations, CAM and IG remain qualitatively similar in many high-confidence positive cases, but become more unstable for borderline examples: hotspots sometimes shift from lung fields to ribs or soft tissue without large changes in predicted probability. Combined with the strong drop in TPR@95% specificity under perturbations, this underlines the need to interpret saliency maps with caution and to pair them with robustness checks.

As a simple sanity test, I randomized the classifier head of the baseline model while keeping earlier layers fixed and recomputed CAM and IG. The resulting maps become diffuse, noisy, and

poorly aligned with anatomical structures, even though the underlying X-ray is unchanged. This behavior is consistent with the expectation that meaningful explanations should disappear when the model's decision function is destroyed.

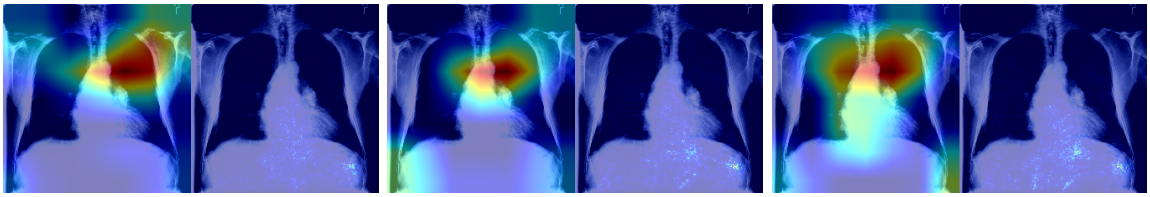

Figure 4: CAM (left panel in each sub-image) and IG (right panel) for a high-confidence Effusion case (left to right: baseline, reweight, GroupDRO). All models highlight the right hemithorax; mitigations produce more compact, lung-focused saliency.

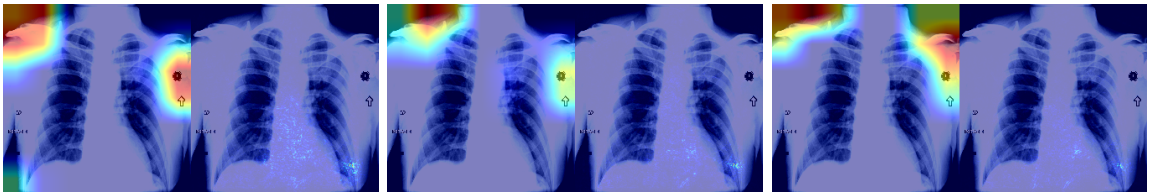

Figure 5: Borderline case with CAM and IG (baseline, reweight, GroupDRO). The baseline attributions focus partly on mediastinal structures; mitigated models lower the predicted probability and shift attention towards the lung bases.

5 Discussion

- **Fairness findings**: For Effusion, raw prevalence differences across sex are small and the baseline model already exhibits modest fairness gaps. Nevertheless, both reweighting and GroupDRO further reduce EO/EOds gaps at the cost of slight degradation in global metrics.
- **Trade-offs**: Reweighting yields a smoother trade-off (small drops in AUROC/AUPRC/F1) with a large gain in EOds, while GroupDRO aggressively optimizes worst-group loss, leading to the lowest gaps but more pronounced F1 reduction.
- **Explainability interplay**: The mitigated models preserve the general structure of Grad-CAM maps while altering confidence. A more thorough case-study analysis could test whether mitigations reduce clear off-target attributions.
- **Limitations**: Single protected attribute (sex) and single pathology; label noise and unobserved confounders; no subgroup-specific calibration; robustness only probed with simple synthetic perturbations; no formal human evaluation of explanations.

6 Conclusion

Using a DenseNet-121 effusion detector on ChestXray14, I showed that simple subgroup-aware interventions—inverse-frequency reweighting and GroupDRO—can meaningfully reduce fairness gaps across sex with only modest losses in global performance. Calibration can be substantially

improved via temperature scaling without hurting AUROC/AUPRC, but the model remains brittle to mild perturbations, underscoring the need for robustness-aware training. Grad-CAM explanations provide a useful qualitative lens on the effect of mitigation, but more principled sanity tests and human-in-the-loop assessment are required before clinical deployment.

Contribution Note

Individual assignment: all experiments, analysis, and writing were carried out by the author.

References

[1] Wang, Xiaosong et al. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. CVPR 2017.