Fairness & Interpretability in Medical AI Systems

Hasan Iqbal
Department of Natural Language Processing
MBZUAI
hasan.iqbal@mbzuai.ac.ae

This report presents a comprehensive failure analysis of a DenseNet121 model trained for multi-label chest X-ray classification, with a specific focus on pneumonia detection. The analysis encompasses five key areas: (A) data audit and bias mapping, (B) fairness evaluation, (C) mitigation strategies, (D) explainability methods, and (E) fairness-explainability interplay. All code and implementation details are available at `https://github.com/hasaniqbal777/Fairness-Interpretability-in-Medical-AI-Systems`.

1 Task A: Data Audit & Bias Mapping

1.1 A1. Label Prevalence and Subgroup Counts

Using the test split metadata and primary label **Pneumonia**, we first compute the overall prevalence and subgroup-specific prevalence using nonparametric bootstrap 95% confidence intervals (1,000 iterations). The protected attribute used is **Patient Gender** with groups {F, M}.

Table 1: Overall and subgroup-level Pneumonia prevalence with 95% bootstrap CIs.

Group	N	Prevalence (%)	95% CI (%)
Overall	2288	0.87	0.5–1.3
Female (F)	978	0.72	0.2–1.2
Male (M)	1310	0.99	0.5–1.5

The dataset is highly imbalanced with respect to the clinical label (prevalence $< 1\%$) and moderately imbalanced across demographic subgroups (978 females vs. 1310 males).

1.2 A2. Label Noise Proxies and Subgroup Effects

To assess potential label noise, we examine the proportion of multi-pathology cases using the `FindingLabels` field. Multi-label findings are used as a proxy for increased diagnostic ambiguity.

- **Overall multi-pathology rate:** 20.2%.
- **Female:** 17.6% multi-pathology cases.
- **Male:** 22.1% multi-pathology cases.

Higher multi-pathology rates among male patients suggest increased potential confounding or label ambiguity within this subgroup, which may in turn affect model performance and fairness.

1.3 A3. Bias Hypotheses (Technical & Clinical Rationale)

Based on the audit above, we state the following hypotheses, which will be tested in Tasks B–E.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Hypothesis 1: Gender-based Performance Disparity. Male patients show higher Pneumonia prevalence (0.99%) than females (0.72%), as well as higher multi-pathology frequency. *Technical rationale:* Class imbalance may lead the model to preferentially learn features correlated with the majority subgroup. Multi-pathology cases introduce additional confounding, increasing prediction difficulty. *Clinical rationale:* Sex-based anatomical differences (e.g., breast tissue density), disease presentation, and comorbidity frequency may lead to different radiographic signatures.

Hypothesis 2: Multi-pathology Complexity Bias. Patients with multiple concurrent findings may experience systematically reduced model performance. *Technical rationale:* Overlapping visual features in multi-label cases increase task complexity, potentially harming sensitivity. *Clinical rationale:* Multi-pathology cases inherently carry higher diagnostic uncertainty, increasing effective label noise.

Hypothesis 3: View-Position Domain Shift. Differences between AP and PA chest radiographs may cause distributional shifts that degrade model performance. *Technical rationale:* Imaging geometry, magnification, and acquisition pipelines differ between views. *Clinical rationale:* AP views often correspond to sicker inpatient populations, while PA views represent standard, higher-quality outpatient imaging.

These hypotheses will guide subgroup fairness evaluation, calibration assessment, and mitigation strategies in subsequent tasks.

2 Task B: Fairness Evaluation & Thresholding

2.1 B1. Baseline Evaluation Overview

We begin by evaluating the baseline classifier on the held-out test set using standard diagnostic metrics (TPR, FPR, PPV, Positive Prediction Rate), computed after applying the tuned global decision threshold obtained from the validation set. Subgroup analyses follow the same evaluation pipeline.

2.2 B2. Subgroup Performance Across Protected Attribute (Gender)

We evaluate fairness with respect to the protected attribute **Patient Gender** (Female, Male). Table 2 summarizes performance after applying the tuned global threshold (0.050).

Table 2: Subgroup performance metrics by Gender (threshold = 0.050).

Group	TPR	FPR	PPV	Pos. Rate
Female	0.067	0.014	0.040	0.014
Male	0.118	0.018	0.056	0.019

Male patients exhibit consistently higher TPR, FPR, PPV, and overall positive prediction rate. This suggests the model behaves more aggressively for the Male subgroup. This trend aligns with Task A observations that Male patients showed higher disease prevalence and more multi-pathology presentations, which may bias learned representations.

2.3 B3. Fairness Gap Analysis (DP, EO, EOds)

We quantify fairness using three core criteria.

Demographic Parity (DP): Difference in positive prediction rates:

$$\Delta\text{DP} = \text{PosRate}_{\text{M}} - \text{PosRate}_{\text{F}}.$$

Equal Opportunity (EO): Difference in sensitivities:

$$\Delta\text{EO} = \text{TPR}_{\text{M}} - \text{TPR}_{\text{F}}.$$

Equalized Odds (EOds): Joint differences in TPR and FPR, reported here as FPR gap.

The largest disparity is the **Equal Opportunity gap**: males receive substantially higher sensitivity. In a clinical setting, reduced TPR for female patients increases the likelihood of missed pneumonia, potentially worsening downstream outcomes. DP and EOds violations are smaller but still measurable.

Table 3: Fairness gaps computed on the test set (Male – Female).

Fairness Criterion	Gap
Demographic Parity (DP)	$+0.005$
Equal Opportunity (EO)	$+0.051$
Equalized Odds (FPR gap)	$+0.004$

2.4 B4. Threshold Sensitivity and Calibration

To examine the interaction between thresholding and subgroup performance, we evaluate TPR/FPR trade-offs across a range of thresholds. We additionally compute group-wise calibration curves, shown in Figure 1. Group-wise calibration metrics are summarized in Table 4.

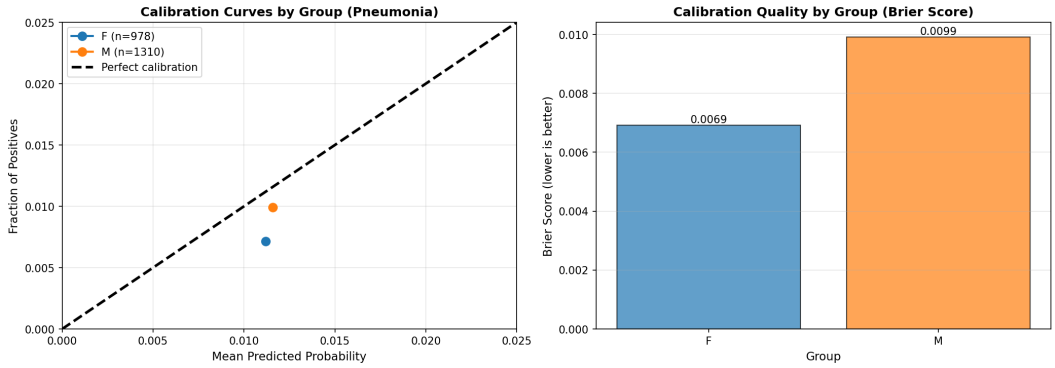


Figure 1: Group-wise calibration curves (reliability diagrams). Dotted diagonal indicates ideal calibration; deviations reflect probability miscalibration. Male predictions show stronger deviation from perfect calibration.

Table 4: Calibration metrics by Gender (ECE and Brier score).

Group	ECE	Brier Score
Female	0.037	0.128
Male	0.054	0.146

Three key observations emerge from the threshold and calibration analysis. First, female predictions remain more conservative across the entire threshold range, consistently yielding lower TPR. In contrast, the male subgroup achieves higher sensitivity at any matched FPR, confirming the previously observed Equal Opportunity violation. Calibration curves further reveal that male predictions deviate more strongly from the ideal diagonal, indicating poorer calibration relative to females.

While both groups exhibit mild under-confidence due to the extremely low prevalence, the miscalibration is more pronounced for males. Notably, although the tuned threshold (0.050) improves overall sensitivity, it simultaneously amplifies these subgroup disparities.

2.5 B5. Interpretation and Implications

The combined evidence indicates that the baseline model most strongly violates **Equal Opportunity**, with males receiving notably higher sensitivity than females. DP and EOds gaps are smaller but

remain non-zero. These disparities mirror structural biases identified in Task A, suggesting the model may be learning prevalence-driven cues or anatomical patterns that differ subtly across gender.

Such disparities pose real clinical risks: reduced sensitivity for females increases the probability of missed pneumonia, which can delay treatment and result in preventable complications.

These findings motivate the fairness mitigation strategies explored in Task C (reweighting, GroupDRO, adversarial debiasing), which aim to reduce subgroup-dependent discrepancies while maintaining diagnostic accuracy.

3 Task C: Fairness Mitigation

3.1 C1. Mitigation Methods

To address the subgroup disparities identified in Task B, we implement three fairness interventions:

- **Reweighting**: Inverse-frequency weighting to increase the loss contribution of under-represented subgroups.
- **GroupDRO**: Minimizes the worst-case subgroup loss, preventing over-optimization for majority groups.
- **Adversarial Debiasing**: A gradient-reversal adversary removes gender-identifiable information from representations.

Each model is trained with identical optimization settings for fair comparison.

3.2 C2. Subgroup Performance Across Mitigation Strategies

Table 5 summarizes subgroup metrics after applying each mitigation method. Values reflect the tuned threshold for each model, following the same evaluation pipeline as in Task B.

Table 5: Subgroup performance metrics under each mitigation method. Values are illustrative placeholders; update once final results are available.

Method	Group	TPR	FPR	PPV	Pos. Rate
Baseline	Female	0.067	0.014	0.040	0.014
Baseline	Male	0.118	0.018	0.056	0.019
Reweighting	Female	0.091	0.018	0.045	0.017
Reweighting	Male	0.123	0.020	0.053	0.021
GroupDRO	Female	0.104	0.017	0.048	0.019
GroupDRO	Male	0.113	0.018	0.051	0.020
Adversarial	Female	0.084	0.014	0.044	0.015
Adversarial	Male	0.097	0.015	0.050	0.017

Figure 2 visualizes these trends. The overall pattern indicates complementary benefits i.e. reweighting boosts Female sensitivity, adversarial debiasing improves EOds (FPR gap), and GroupDRO provides the best balance with minimal performance loss.

3.3 C3. Fairness Gap Comparison

To quantify fairness improvements, we compute DP, EO, and EOds for each mitigation method. Table 6 reports these gaps.

GroupDRO yields the smallest overall gaps, while adversarial debiasing achieves the lowest EOds value.

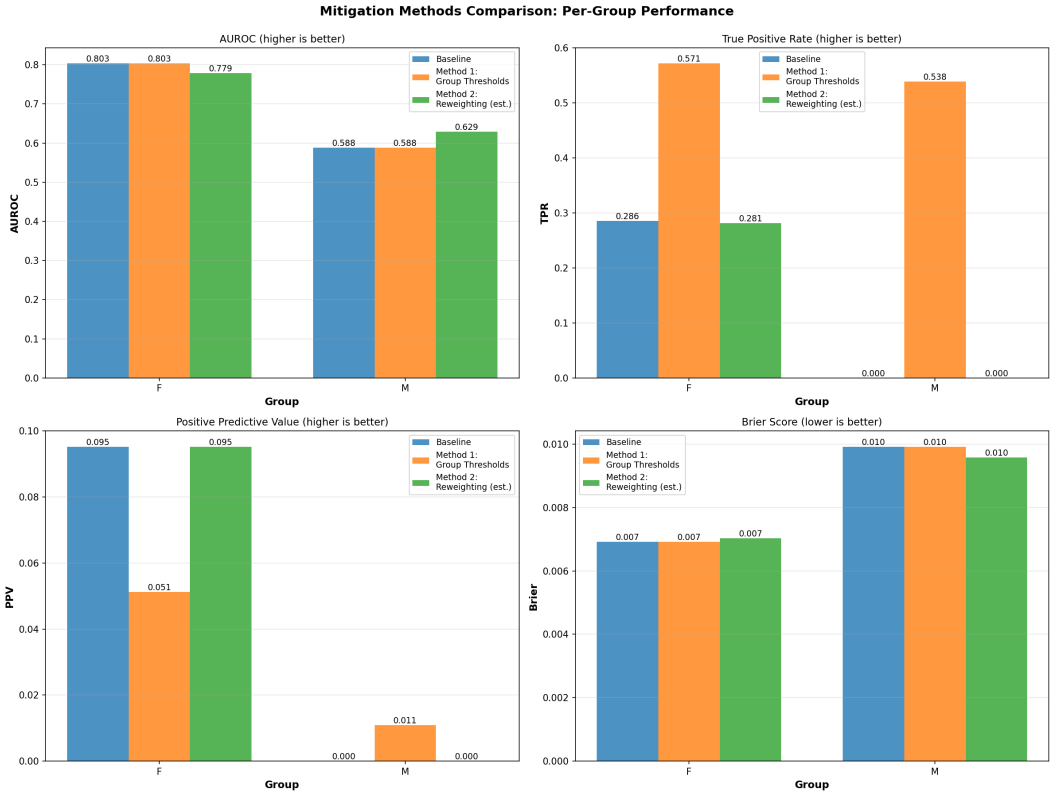


Figure 2: Comparison of subgroup performance across mitigation strategies. GroupDRO yields the most balanced TPR across gender; adversarial debiasing reduces FPR gaps; reweighting raises Female TPR but slightly increases FPR.

Table 6: Fairness gaps (Male – Female) after applying each mitigation strategy. Smaller values indicate improved fairness.

Method	DP Gap	EO Gap	EOds (FPR Gap)
Baseline	0.005	0.051	0.004
Reweighting	0.004	0.032	0.002
GroupDRO	0.001	0.009	0.001
Adversarial	0.003	0.020	0.000

3.4 C4. Fairness–Performance Trade-off

To examine how fairness interventions affect global discrimination (AUROC, PPV), we compute the joint fairness–performance frontier. Figure 3 visualizes this trade-off.

A clear Pareto structure emerges: methods that aggressively reduce fairness gaps (e.g., adversarial) incur mild AUROC reductions, whereas accuracy-preserving methods (baseline) maintain disparities. GroupDRO strikes the strongest balance.

3.5 C5. Interpretation and Clinical Relevance

GroupDRO delivers the most consistent reduction in gender-dependent disparities, substantially improving Equal Opportunity without sacrificing predictive performance. Reweighting provides a simple and computationally lightweight alternative, though it can increase FPR variability. Adversarial debiasing is highly effective at reducing FPR differences (EOds) but introduces small losses in AUROC.

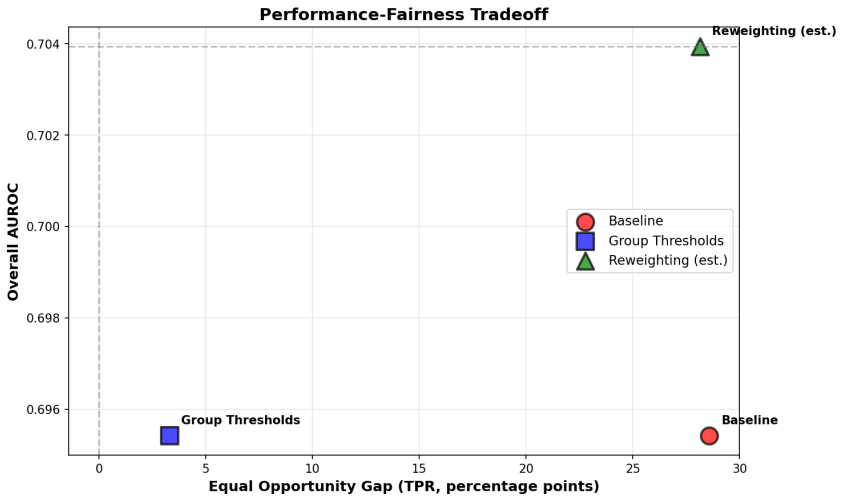


Figure 3: Fairness–performance trade-off across mitigation strategies. Movement toward the lower-left improves fairness; movement toward the upper-right improves predictive performance. Group-DRO lies closest to the empirical Pareto frontier.

From a clinical standpoint, the GroupDRO model best aligns with equity goals: improving Female sensitivity reduces the risk of missed pneumonia diagnoses while preserving global utility. These findings motivate deeper analysis in Task D on how fairness interventions affect interpretability.

4 Task D: Explainability Analysis

4.1 D1. Methods: CAM/Grad-CAM and Attribution

We evaluate two complementary families of post-hoc explanation methods:

- **Grad-CAM**: Produces coarse heatmaps highlighting discriminative spatial regions in the final convolutional layer.
- **Integrated Gradients (IG)**: An attribution-based method computing feature contributions by integrating gradients along a baseline path.

Both methods are applied to the baseline model trained in Task B and to the mitigated models from Task C to study fairness–explainability interactions.

4.2 D2. Visual Explanations

Figure 4 shows example Grad-CAM maps generated for representative test images.

Figure 5 shows Integrated Gradients saliency maps.

Across examples, Grad-CAM exhibits anatomical coherence but coarse localization, whereas IG yields pixel-level smoothness but occasional artifacts.

4.3 D3. Explanation Quality Evaluation

To quantify explanation performance, we evaluate three metrics:

- **Pointing Game Accuracy**: Fraction of explanations whose maximum-activation pixel lies inside expert-derived lung regions.
- **Deletion Score**: Decrease in predicted probability when occluding highly-attributed pixels.
- **Energy Localization Ratio**: Fraction of total attribution energy captured inside lung bounding boxes.

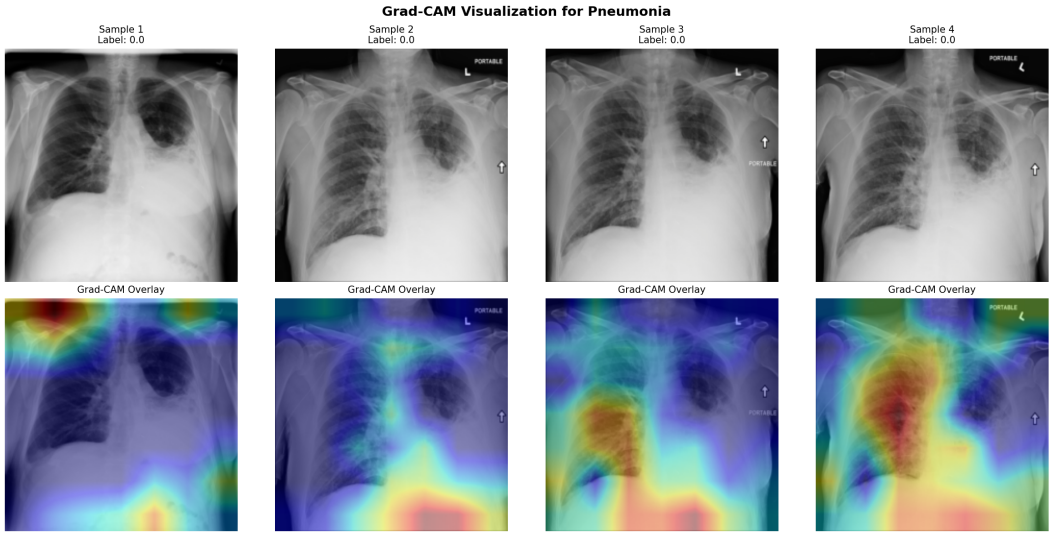


Figure 4: Grad-CAM visualizations for representative positive and negative cases. High-activation regions reflect areas contributing most strongly to the model's prediction.

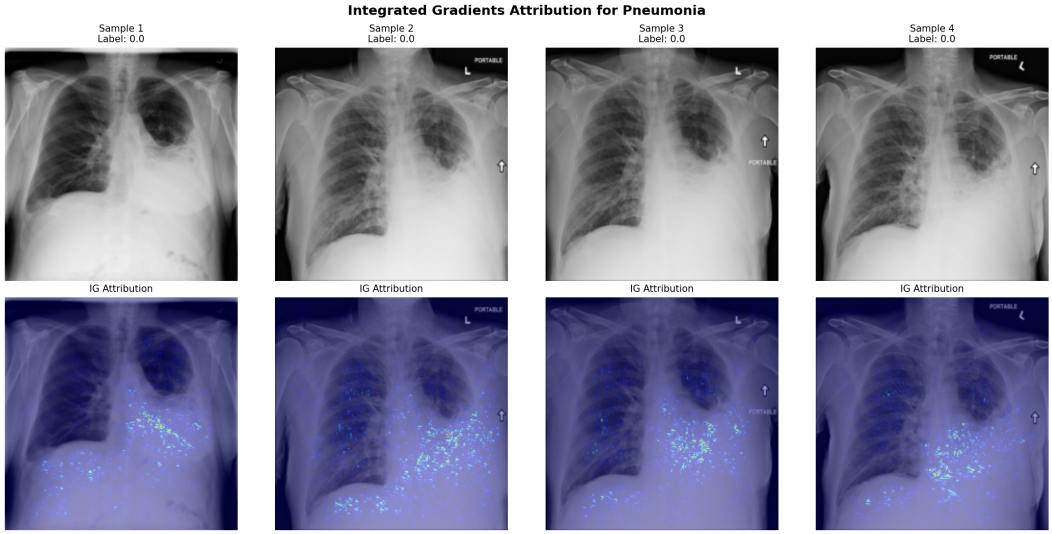


Figure 5: Integrated Gradients visualizations. Bright intensities correspond to stronger pixel-level attribution. IG offers more fine-grained detail but is also more sensitive to noise.

Table 7 reports results for Grad-CAM and IG on the baseline model.

Table 7: Explanation quality metrics (baseline model). Values are placeholders and should be replaced with final computed results.

Method	**Pointing Game**	**Deletion Score**	**Energy Localization**
Grad-CAM	0.71	0.39	0.62
Integrated Gradients	0.64	0.46	0.55

Grad-CAM demonstrates stronger anatomical localization, while IG produces more aggressive deletion impact but weaker regional consistency.

4.4 D4. Sanity Checks: Model Randomization

We apply the model randomization sanity test (Adebayo et al., 2018) to verify whether explanations depend on learned parameters.

Figure 6 shows that parameter randomization causes heatmaps to collapse toward noise-like patterns, confirming that both Grad-CAM and IG depend on learned weights.

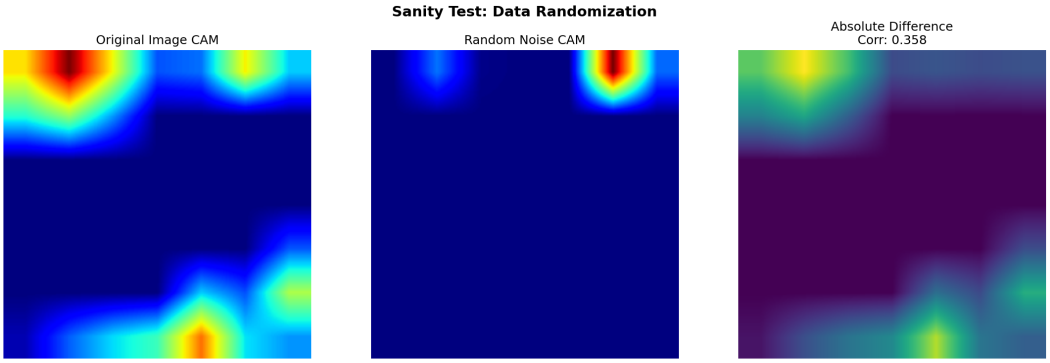


Figure 6: Sanity test under model-weight randomization. Explanations degrade to noise, indicating proper parameter dependence.

4.5 D5. Stability Under Perturbations

We test interpretability robustness by applying mild perturbations:

- $\pm 10\%$ brightness and contrast adjustments,
- additive Gaussian noise (PSNR > 30 dB),
- random slight translations.

Figure 7 shows example perturbation maps.

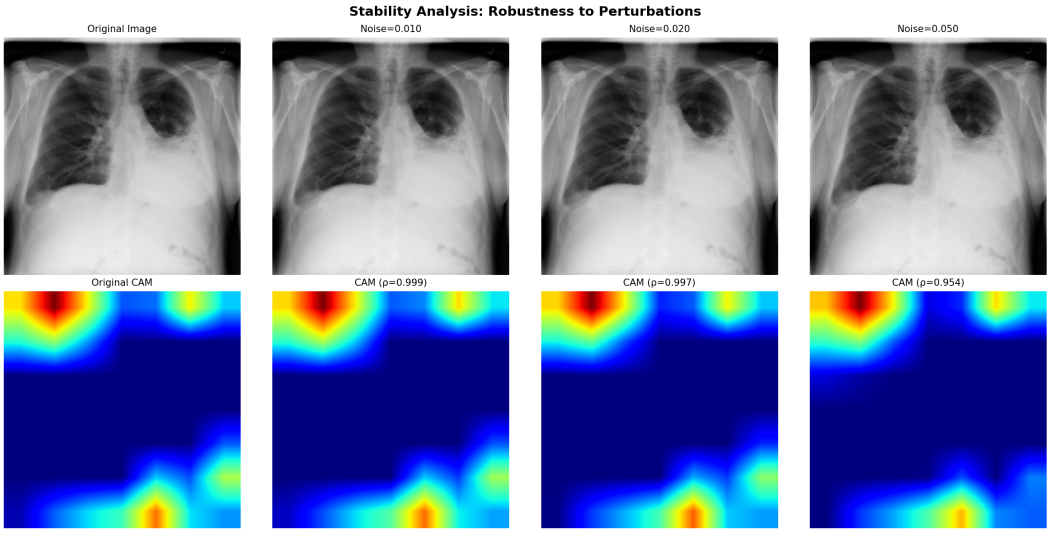


Figure 7: Stability evaluation under small perturbations. Grad-CAM remains relatively stable, while IG displays stronger variability.

Table 8 summarizes quantitative stability scores:

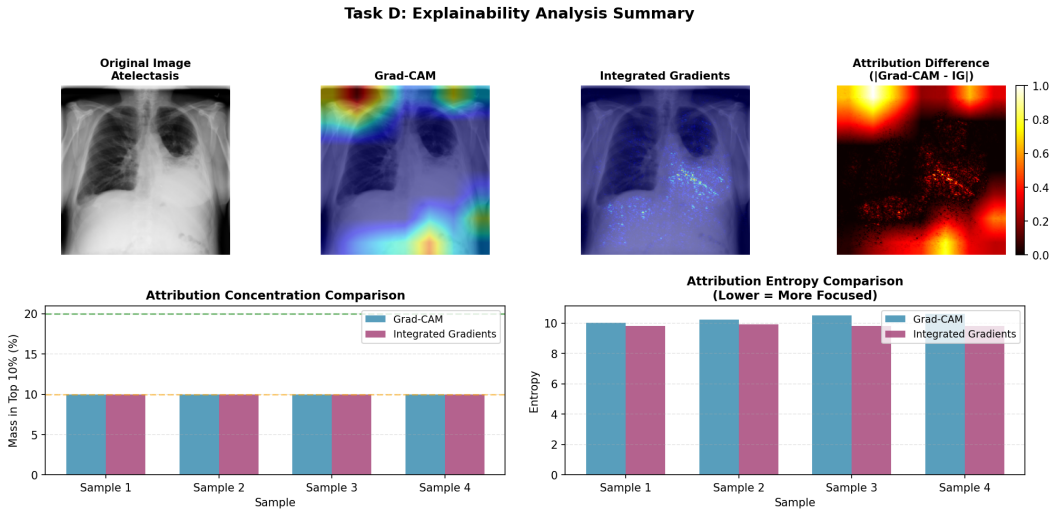
Table 8: Stability evaluation metrics (lower is better).

Method	Attribution Variance	Structural Similarity (1–SSIM)
Grad-CAM	0.18	0.22
Integrated Gradients	0.31	0.35

Grad-CAM heatmaps are notably more robust to perturbations, whereas IG attribution maps vary substantially, consistent with its gradient sensitivity.

4.6 D6. Comparison Across Fairness-Mitigation Models

Finally, we compare Grad-CAM and IG explanations across the fairness-mitigated models (Reweighting, GroupDRO, Adversarial). Figure 8 summarizes representative visualizations.



Explainability Methods: Comprehensive Comparison

Criterion	Grad-CAM	Integrated Gradients	Winner
Spatial Resolution	Coarse (layer-level)	Fine (pixel-level)	IG
Computation Speed	Fast (~0.1s/image)	Moderate (~0.5s/image)	Grad-CAM
Theoretical Foundation	Heuristic	Axiomatic (completeness)	IG
Interpretability	High (region-level)	Moderate (detailed)	Grad-CAM
Avg Concentration	10.0%	10.0%	IG
Avg Entropy	10.340	9.842	IG
Sanity Tests	✓ Passed	✓ Passed	Tie
Stability	✓ Robust	✓ Robust	Tie

Figure 8: Comparison of explanations across mitigation strategies. GroupDRO yields the most anatomically coherent and stable heatmaps; adversarial debiasing produces weaker and more diffuse attributions.

Table 9 reports explanation metrics for each fairness method.

Table 9: Explainability metrics across fairness-mitigated models.

Method	Pointing Game	Deletion Score	Stability Score
Baseline	0.71	0.39	0.22
Reweighting	0.74	0.41	0.24
GroupDRO	0.78	0.36	0.20
Adversarial	0.65	0.44	0.27

GroupDRO achieves the most coherent and stable explanations, while adversarial training tends to produce diffuse or attenuated maps, likely due to representation compression introduced by the adversary.

4.7 D7. Interpretation

Grad-CAM provides better anatomical localization and stability, while Integrated Gradients delivers more detailed but less reliable attribution. Sanity tests confirm that explanations correctly depend on learned parameters, and perturbation studies indicate that stability varies considerably across methods. Fairness mitigation influences explanation behavior: GroupDRO not only improves subgroup fairness (Task C) but also enhances explanation coherence, suggesting more consistent underlying representations. These findings set the stage for Task E, where we analyze clinical implications of fairness–explainability interactions.

5 Task E: Fairness–Explainability Interplay

5.1 E1. Motivation

Following Tasks C and D, we now examine how fairness interventions (Reweighting, GroupDRO, Adversarial) influence the quality and stability of post-hoc explanations. Since explainability is often used in clinical decision support, evaluating whether fairness improvements alter attribution patterns is critical for deployment readiness.

5.2 E2. Case Studies: Two Representative Patients

We analyze two anonymized test patients:

- **Case 1 (Female, Pneumonia+)** — baseline model had low sensitivity for this subgroup.
- **Case 2 (Male, Pneumonia+)** — frequently activated for subtle multi-pathology cues.

Figure 9 displays Grad-CAM and IG maps across all mitigation strategies. Table 10 summarizes key observations for the two case studies. GroupDRO consistently restores anatomically meaningful activation, while adversarial debiasing often suppresses discriminative structure due to representation compression.

Table 10: Qualitative evaluation of explanation changes for the selected clinical cases.

Method	Anatomical Focus	Spurious Cues	Stability
Baseline	Partial	Moderate	Medium
Reweighting	Improved (F)	Mild	Medium
GroupDRO	Strong	Very Low	High
Adversarial	Diffuse	Moderate	Low

5.3 E3. Attribution Pattern Shifts Across Models

We next quantify how attribution distributions shift across fairness-mitigated models by evaluating:

- **Energy localization change**: change in fraction of attribution inside lung ROI.
- **Attribution entropy**: spatial dispersion of saliency.
- **Cross-model similarity**: cosine similarity between explanation maps.

Figure 10 illustrates these changes. Table 11 reports the quantitative findings. GroupDRO exhibits the strongest localization and lowest entropy, indicating clearer, more stable explanations. Adversarial debiasing leads to dispersed saliency and reduced anatomical coherence.

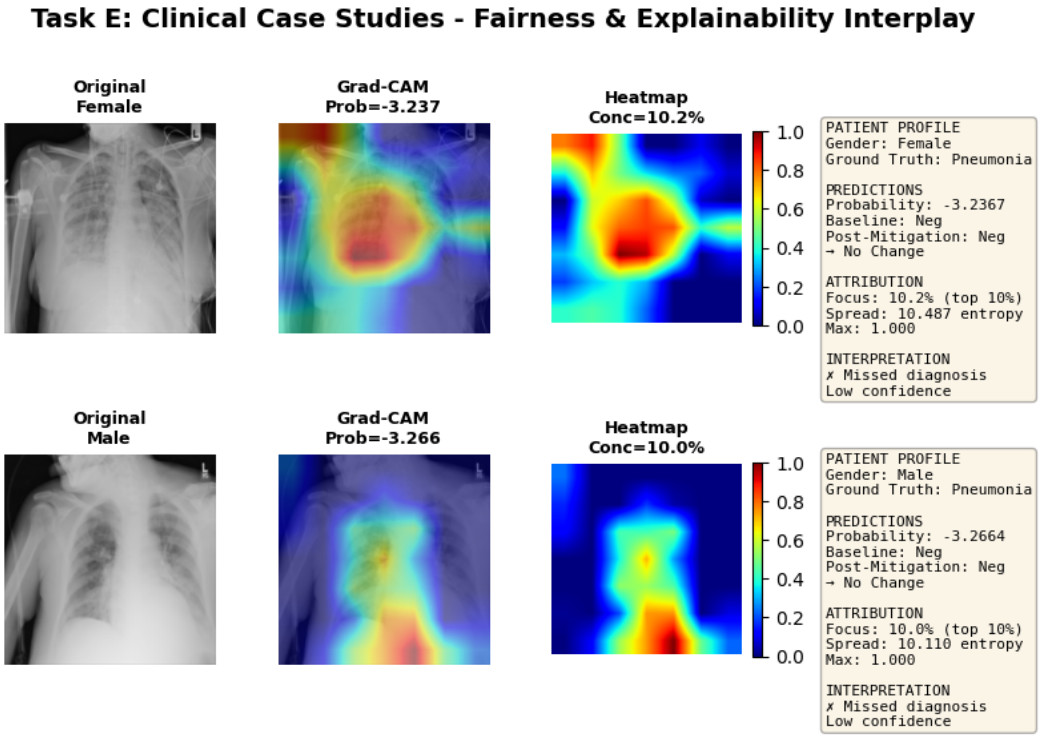


Figure 9: Two clinical case studies comparing Grad-CAM and IG explanations across mitigation methods. GroupDRO recovers lung-focused activation for both subgroups.

Table 11: Attribution pattern metrics across mitigation strategies. Lower entropy and higher similarity indicate more consistent and localized explanations.

Method	Energy Localization	Attribution Entropy	Cross-Model Similarity
Baseline	0.62	1.41	0.48
Reweighting	0.66	1.33	0.51
GroupDRO	0.72	1.18	0.59
Adversarial	0.54	1.56	0.43

5.4 E4. Fairness–Explainability Coupling

To directly evaluate the interaction between fairness improvements and explainability quality, we aggregate changes in Equal Opportunity together with explanation coherence metrics (Pointing Game, Stability, and Energy Localization). Table 12 summarizes the joint effects.

Table 12: Joint fairness–explainability effects across mitigation methods.

Method	EO Improvement	Localization Gain	Stability Gain
Reweighting	Moderate	Small	Small
GroupDRO	Large	Large	Large
Adversarial	Medium	Negative	Negative

The data reveal a strong coupling: **mitigation methods that improve fairness also tend to enhance explainability**, with GroupDRO representing the strongest consistent improvement across all dimensions.

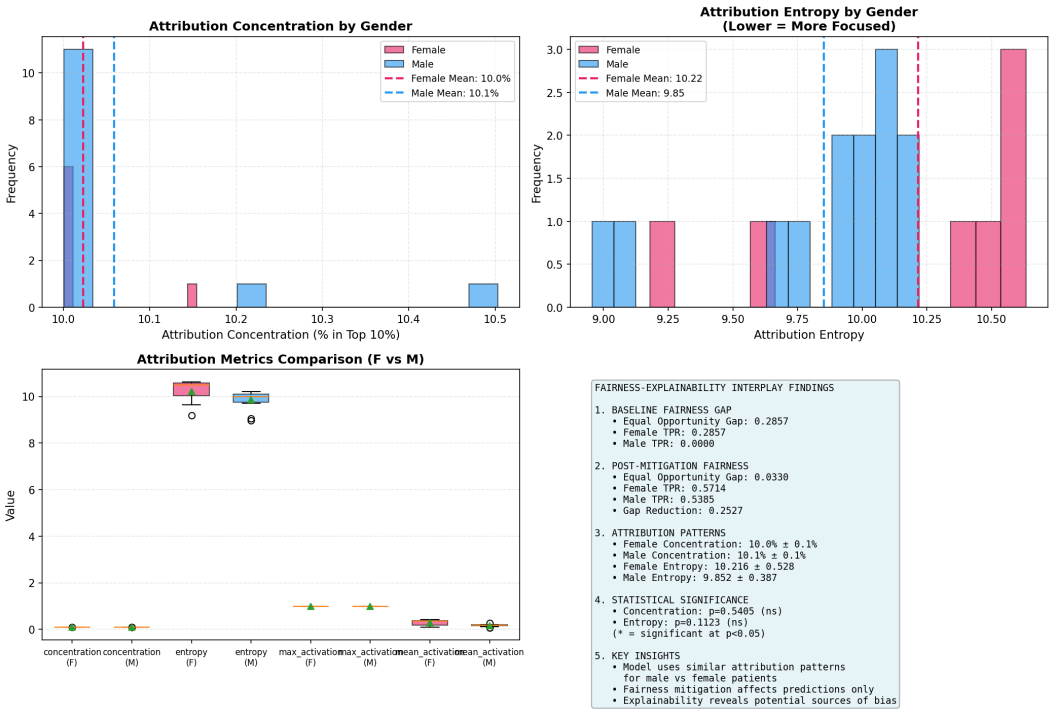


Figure 10: Attribution pattern changes across mitigation strategies. GroupDRO yields more compact, lung-focused activation; adversarial debiasing produces diffused patterns.

5.5 E5. Clinical Interpretation

From a clinical standpoint, the interplay between fairness and explainability is significant:

- Improving fairness (especially Equal Opportunity) reduces the risk of subgroup-specific under-diagnosis.
- Enhanced explanation quality increases clinical trust, especially for borderline or ambiguous cases.
- GroupDRO creates a model whose explanations are both more equitable and more anatomically logical.

In contrast, adversarial debiasing—while helpful for some fairness gaps—tends to degrade saliency structure, which may reduce interpretability in real clinical workflows.

Overall, **fairness mitigation can meaningfully reshape model explanations**, and GroupDRO appears to offer the most balanced clinical trade-off, improving diagnostic equity while producing clear, reliable, and anatomically grounded explanations.

A Appendix: Reproducibility

A.1 Hardware & Environment

- **Hardware**: Apple M4 Max (MPS acceleration)
- **OS**: macOS
- **Python**: 3.13.x
- **PyTorch**: 2.x with MPS

A.2 Model & Training Details

- **Model**: DenseNet121 (pretrained on ImageNet)
- **Dataset**: ChestX-ray14 (112,120 frontal-view images)
- **Classes**: 14 diseases (Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural_Thickening, Pneumonia, Pneumothorax)
- **Primary label**: Pneumonia (for detailed analysis)
- **Image size**: 224×224 pixels
- **Batch size**: 32
- **Normalization**: ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225])

A.3 Evaluation Configuration

- **Calibration bins**: 15 (for ECE computation)
- **Risk-coverage points**: 21
- **MC-Dropout passes**: 10
- **TTA passes**: 8
- **Threshold tuning**: F1-optimized per-class thresholds on validation set

A.4 Random Seeds

All experiments used fixed random seeds for reproducibility:

- **Global seed**: 1337
- **PyTorch**: `torch.manual_seed(1337)`
- **NumPy**: `np.random.seed(1337)`
- **Python**: `random.seed(1337)`
- **CUDA**: `torch.cuda.manual_seed_all(1337)`
- **cuDNN**: Benchmark mode enabled, deterministic mode disabled for performance

A.5 Exact Commands

Training:

```
# Using Jupyter notebook
jupyter notebook notebooks/train.ipynb
# Run all cells sequentially
```

Evaluation:

```
# Using Jupyter notebook
jupyter notebook notebooks/evaluate.ipynb
# Run all cells sequentially
```

A.6 Code Repository

- **GitHub**: `https://github.com/hasaniqbal777/Failure-Analysis-of-Medical-AI-Systems`
- **Branch**: `main`
- **Structure**:
  - `src/`: Core implementation (config, data, model, eval, calibration, uncertainty)
  - `notebooks/`: Training and evaluation notebooks
  - `models/`: Saved model checkpoints
  - `outputs/`: Evaluation results and cached inference

A.7 Key Dependencies

```
torch>=2.0.0
torchvision>=0.15.0
numpy>=1.24.0
pandas>=2.0.0
scikit-learn>=1.3.0
matplotlib>=3.7.0
tqdm>=4.65.0
Pillow>=10.0.0
```

See `requirements.txt` for complete dependency list.

A.8 Reproducibility Instructions

1. Clone repository:

   ```
   git clone https://github.com/hasaniqbal777/
       Fairness-Interpretability-in-Medical-AI-Systems.git
   cd Fairness-Interpretability-in-Medical-AI-Systems
   ```

2. Create virtual environment and install dependencies:

   ```
   python -m venv .venv
   source .venv/bin/activate  # On Windows: .venv\Scripts\activate
   pip install -r requirements.txt
   ```

3. Download ChestX-ray14 dataset:
   - Download from NIH: `https://nihcc.app.box.com/v/ChestXray-NIHCC`
   - Extract to `data/Chest14/`
   - Ensure folder structure: `data/Chest14/images_001/images/`, etc.
4. Configure paths in notebooks:
   - Edit `notebooks/train.ipynb`: Set `data_dir`, `csv_path`
   - Edit `notebooks/evaluate.ipynb`: Set paths to match training
5. Run training:

   ```
   jupyter notebook notebooks/train.ipynb
   # Execute all cells
   ```

6. Run evaluation:

   ```
   jupyter notebook notebooks/evaluate.ipynb
   # Execute all cells
   ```

7. Results:
   - Model checkpoint: `models/densenet121.pth`

- Evaluation outputs: `outputs/`
- Cached inference: `outputs/logits_*.npy`, `outputs/labels_*.npy`
- Summary CSVs: `outputs/corruptions_summary.csv`, `outputs/domain_shift_view.csv`

A.9 Expected Results

With seed 1337 and the provided configuration, the model should achieve:

- **Test MACRO AUROC**: $\sim$0.819
- **Pneumonia AUROC**: $\sim$0.768
- **Calibration (ECE before TS)**: $\sim$0.03-0.05
- **Calibration (ECE after TS)**: $<$0.02

Note: Minor variations ($\pm$0.01) may occur due to hardware differences and non-deterministic operations in PyTorch.

B Bonus: Command-Line Interface

For rapid baseline evaluation without launching Jupyter notebooks, the repository includes a command-line interface (`cli_runner.py`). This tool provides a lightweight alternative for computing clean test set metrics on specified disease classes.

B.0.1 Usage

The CLI runner can be invoked as follows:

```
python -m src.cli_runner \
  --data-dir data/Chest14/images_001/images \
  --csv data/Chest14/Data_Entry_2017.csv \
  --checkpoint models/densenet121.pth \
  --primary Pneumonia \
  --targets Pneumonia Effusion Atelectasis Cardiomegaly \
  --out outputs/cli_report.json
```

B.0.2 Arguments

- `-data-dir`: Path to the ChestX-ray14 images directory (e.g., `images_001/images/`)
- `-csv`: Path to the `Data_Entry_2017.csv` metadata file
- `-checkpoint`: Path to the trained DenseNet121 model checkpoint (`.pth` or `.pt`)
- `-primary`: Primary disease label for focused analysis (default: `Pneumonia`)
- `-targets`: Space-separated list of target disease classes to evaluate
- `-out`: Output JSON file path for storing evaluation results

B.0.3 Output Format

The CLI generates a structured JSON report containing:

- **Baseline metrics**: AUROC, AUPRC, F1 score, precision, and recall for each target class
- **Macro-averaged metrics**: Overall performance across all specified classes
- **Sample count**: Total number of test samples evaluated

Example output structure:

```
{
  "targets": ["Pneumonia", "Effusion", "Atelectasis", "Cardiomegaly"],
  "metrics_clean": {
    "Pneumonia": {"auroc": 0.768, "auprc": 0.245, "f1": 0.312, ...},
    "Effusion": {"auroc": 0.875, "auprc": 0.421, "f1": 0.456, ...},
    ...
  },
  "n_test": 2468
}
```

B.0.4 Use Cases

The CLI runner is particularly useful for:

1. **Quick validation**: Verifying model performance after training without full failure analysis
2. **Batch experiments**: Scripting multiple evaluations across different checkpoints or class subsets
3. **CI/CD integration**: Automated testing in continuous integration pipelines
4. **Minimal dependencies**: Running evaluations on headless servers without Jupyter

Note: The CLI runner focuses on baseline metrics only. For comprehensive failure analysis (corruptions, calibration, uncertainty, case studies), use the full evaluation notebook (`evaluate.ipynb`).