CV8502 — Assignment 2

Fairness & Interpretability in Medical AI

Course: CV8502 — Trustworthy Medical Vision
Institution: Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)

Handout: Week 4 **Due: End of Week 6 (23:59 local time)**
Theme: Fairness & Explainability
Deliverables: Report (PDF), Code + Notebooks, Repro Config, Contribution Note

---

Learning Outcomes

- Diagnose dataset bias and quantify disparities relevant to healthcare.
- Measure and compare multiple fairness criteria (**DP, EO, EOds**) with calibration.
- Implement post-hoc explainability (CAM/Grad-CAM, attribution) and evaluate quality.
- Explore interplay between fairness interventions and interpretability quality.
- Communicate results for clinical stakeholders (clear visuals, uncertainty, actionable insights).

1. Motivation

Clinical deployment requires not only accuracy, but *equitable* performance and *clinically useful* explanations. This assignment guides you to evaluate fairness, interpretability, and their interaction.

2. Assets & Scope

You will use a **publicly available medical imaging dataset** for a binary classification task that includes at least one demographic attribute (e.g., sex at birth, age group, or scanner site). Examples of suitable datasets include:

- **NIH ChestX-ray14** — over 100,000 chest radiographs labeled for 14 thoracic pathologies with patient sex and age metadata.
- **CheXpert (Stanford)** — 220,000 chest X-rays with uncertainty labels and demographic attributes.
- **ISIC 2020 / 2024 Skin Lesion Dataset** — dermoscopic images for melanoma detection with patient age and sex.
- **RSNA Pneumonia Detection Challenge** — annotated chest X-rays for pneumonia vs. normal classification.

- **HAM10000** — dermoscopic images of pigmented skin lesions with diagnostic labels and demographic metadata.

You may also use other **open-access medical datasets** (e.g., MIMIC-CXR, COVIDx) provided they include relevant demographic attributes for fairness analysis and comply with ethical and privacy requirements.

- **Model:** Start with a decent baseline CNN (DenseNet121) or a lightweight variant.
- **Splits:** Train/val/test splits provided; OOD site test optional.
- **Privacy:** Only de-identified, course-approved data.

3. Tasks Overview

Each section includes deliverables and weightage.

Task A — Data Audit & Bias Mapping (15%)

A1. Report label prevalence and subgroup counts (with 95% CIs).

A2. Examine label noise proxies and subgroup effects.

A3. State bias hypotheses (technical/clinical rationale).

Task B — Fairness Evaluation & Thresholding (30%)

Train baseline model, then analyze:

- Subgroup metrics (AUROC, AUPRC, TPR@95SP, PPV, Brier).
- Fairness gaps: DP, EO, EOds.
- Group-wise calibration and threshold analysis.

Task C — Mitigation (30%)

Apply two methods (e.g., reweighting, GroupDRO, adversarial debiasing) and report subgroup effects, regressions, and compute cost.

Task D — Explainability (20%)

- Apply CAM/Grad-CAM + one attribution method.
- Evaluate localization quality, sanity tests, and stability.

Task E — Fairness–Explainability Interplay (5% + Bonus)

Reassess explanations post-mitigation; include two clinical case studies.

4. Deliverables

Submission Summary

Report (8 pages) — organized by tasks, clear visuals.
Code & Repro. — Notebooks/scripts, `requirements.txt`, and one-command run.
Contribution Note (0.5p). — Summarize each member's role.

5. Evaluation Rubric (100%)

Category	Criteria (Points)
Data audit	Correct distributions, CIs, clear hypotheses (15)
Fairness eval	Metrics & fairness gap analysis, calibration (30)
Mitigation	Implementation, measured gains/regressions (30)
Explainability	Quality, sanity tests, stability (20)
Interplay	Link between fairness & explainability (5)
Communication	Figures, clarity, reproducibility (10)
Bonus	Evaluation CLI (+5)

6. Integrity & Hints

- Keep runs lightweight and reproducible.
- Use bootstrap CIs; avoid overstating small gaps.
- You may use AI tools for boilerplate code only — cite sources.
- For stability, test mild perturbations (contrast ±10%, noise for PSNR ¿ 30 dB).

7. Submission

Upload by **Week 6 (23:59 local time)**:

- `CV8502_A2_<Name>_report.pdf`
- `CV8502_A2_<Name>_code.zip`
- `CV8502_A2_<Name>_contrib.pdf`

Note: This assignment spans Week 4 (*Fairness*) and Week 5 (*Interpretability*). Build on labs, but aim for analytical reasoning and clinical relevance.