CV8502 — Assignment: Failure Analysis of Medical AI Systems

Work mode: Individual

Deliverables: Report (PDF), analysis notebook(s) + code, contribution note

Length (report): ≤ 6 pages main text (excl. refs/appendix).

Suggested time: 3 weeks

---

1) Motivation & Learning Outcomes

This assignment trains you to **find, categorize, and mitigate failures** in medical AI systems. You will:

- Build a **taxonomy of failure modes** (data, model, evaluation, human-in-the-loop).
- Design **evaluations that surface brittle behavior** (domain shift, corruptions, rare cases, subgroup gaps).
- Quantify **risk-aware metrics** (sensitivity, calibration, coverage-risk, uncertainty quality).
- Propose and test **mitigations** (e.g., TTA, calibration, selective prediction).
- Communicate findings for clinical stakeholders (concise visuals; actionable guidance).

---

2) Assets & Scope

- You may re-use models from Week(1-2) in Lab (e.g., **SAM/MedSAM** for segmentation or a classifier baseline such as **DenseNet121** on ChestX-ray14). Choose **one primary task**:
    - **Segmentation** (Dice-optimized) on a small medical set (e.g., NuInsSeg-style, FLARE22 subset).
    - **Classification** (AUROC-optimized) on a curated subset (e.g., 2–4 target pathologies).
- Use your **existing trained baseline** or train light-weight versions (≤ 10–15 epochs). No extra private data.

> **Privacy:** Use only course-permitted datasets. De-identify any images and avoid patient-identifying metadata.

---

3) Tasks (what to do)

Task A — Build a Failure Taxonomy (15%)

Create a one-page taxonomy tailored to your task. Include at least these buckets and give 1–2 concrete examples each:

- **Data**: acquisition/contrast, artifacts, class imbalance, label noise, rare morphologies.
- **Model**: overconfidence, calibration drift, boundary errors, spurious cues.

- **Evaluation**: metric blind spots, threshold sensitivity, poor subpopulation coverage.
- **Human factors**: ambiguous labels, prompt misplacement (for SAM), annotation variability.

Task B — Stress Tests & Slicing (45%)

Design evaluations that deliberately expose brittle behavior. Implement **≥ 3** of the following:

- **Common corruptions** (e.g., gaussian noise/blur, JPEG artifacts, brightness/contrast shifts) at **3 severities**.
- **Domain shift** (scanner/site, modality windowing, histology stain variants, different dataset split).
- **Prompt perturbations** (segmentation): jitter point/box prompts; simulate user error.
- **Class/size slices**: small lesions vs large; hollow vs solid; central vs peripheral.
- **Subgroup slices**: acquisition protocol or anatomical region subsets (avoid sensitive personal attributes).

For each slice/severity, compute at minimum:

- **Segmentation**: Dice, Jaccard, **Boundary-F1@2px**; report mean±SD and Δ vs. clean.
- **Classification**: AUROC, AUPRC, sensitivity @ 95% specificity, F1; report Δ vs. clean.

Task C — Uncertainty & Calibration (30%)

Quantify the reliability of confidence scores:

- **Calibration**: reliability diagrams; **ECE**/**Brier** score.
- **Selective prediction**: risk-coverage curve (defer low-confidence cases).
- For segmentation, calibrate per-pixel or per-region (e.g., mean mask probability).
- Compare at least **two** methods: temperature scaling, TTA-variance, MC-Dropout, or entropy thresholding.

Task D — Mitigation Experiments (Optional)

Implement mitigations and measure improvement on the worst slices from Task B:

- **Post-hoc**: temperature scaling, TTA, uncertainty filtering (coverage control), morphology-aware post-processing.

Task E — Case Study & Checklist (10%)

Pick **two concrete failure cases**. For each: show image(s), prediction, GT, confidence map; diagnose cause; state mitigation and residual risk. Conclude with a **one-page "deployment checklist"** (data assumptions, monitoring signals, safe-use guidance).

---

4) Deliverables

1. **Report (PDF, ≤ 6 pages)**
    - **Problem & models** (≤ 0.5 p): task, dataset, train/eval split, model(s).
    - **Taxonomy** (≤ 1 p): table or diagram with examples.
    - **Stress tests & slices** (1–2 p): figures and Δ-metric tables; which slices break the model?

    - **Uncertainty & calibration** (≤ 1 p): reliability plots; risk-coverage; key numbers.
    - **Mitigations** (≤ 1 p): what helped, how much, at what cost(**Optional**).
    - **Case studies & checklist** (≤ 1 p): two detailed failures; a concise deployment checklist.
    - **Repro notes**: hardware, time/epoch, seeds, exact commands; link to code.

2. **Code & notebooks**: scripts to reproduce all figures/tables; environment file (`.yml` or `requirements.txt`).

3. **Contribution note (≤ 0.5 p)**: who did what.

File naming: `CV8502_FA_<your name>_report.pdf`, `CV8502_FA_<your name>_code.zip`.

---

5) Evaluation Rubric (100%)

- **Taxonomy quality (15%)**: Coverage, clarity, medical relevance.
- **Stress tests & slicing (45%)**: Experimental design, completeness, and insights (not just numbers).
- **Uncertainty & calibration (30%)**: Correct methodology, interpretation, and plots.
- **Mitigations (Optional)**: Soundness, measurable gains, honest trade-off analysis.
- **Communication & reproducibility (10%)**: Clear figures/tables, checklist, seeds/configs, runnable code.

Bonus (up to +5): A small **internal validation tool** (CLI or notebook) that loads a model and runs your failure suite on a new folder of images.

---

6) Constraints & Integrity

- **Compute**: Keep runs reasonable; log wall-clock.
- **Reproducibility**: Fix seeds; report versions.
- **Integrity**: Cite code/data sources; follow university policy.