Bias & Fairness Write-up

Putting the Task 2C model into production changes the stakes: predictions now influence how quickly real issues (or patients, tickets, cases) are triaged. That means we must examine not only accuracy but also **who benefits** and **who bears risk** from errors.

Where bias can creep in

1. **Sampling bias.** If the training data over-represents one segment (e.g., cases from a single hospital, geographic region, or device vendor), the model can learn patterns that don't generalize. In our breast-cancer context, many public datasets skew toward particular populations; if underrepresented groups differ in tumor biology or measurement distributions, the model may be less reliable for them.
2. **Label bias from our priority construction.** We converted a binary diagnosis into *low/medium/high* priority by binning predicted malignancy probabilities (tertiles from train set). That choice is *procedural*, not clinical. If baseline risk differs across subgroups (e.g., age bands), fixed global cutoffs may systematically over- or under-prioritize certain groups.
3. **Measurement bias.** Features come from imaging/pathology pipelines. Different machines, staining protocols, or technicians can shift distributions. If some groups are more likely to be processed on older equipment, their feature quality may be lower, inflating error rates.
4. **Historical/operational bias.** If historical "ground truth" incorporated resource constraints (e.g., certain patients historically waited longer), the model can reproduce that inequity when trained to mimic those outcomes.
5. **Missing protected attributes.** Our dataset lacks demographics (race, ethnicity, socioeconomic status). That prevents subgroup auditing; worse, the model might learn proxies (zip code, site ID) that correlate with protected characteristics without us noticing.

Using IBM AI Fairness 360 (AIF360) to check and mitigate bias

1) Frame the fairness question.
Define *protected attributes* (e.g., race, age ≥ 65, insurance type) and *favorable outcomes* (e.g., being assigned **high** priority when clinically warranted). If demographics are missing, work with data governance to collect them ethically and transparently, with consent and minimization.

2) Build AIF360 datasets.
Wrap train/test as BinaryLabelDataset/StructuredDataset, adding protected attributes. For our 3-class target, either (a) analyze pairwise one-vs-rest views (e.g., "high vs not-high") or (b) separate audits per thresholded pathway (e.g., high/medium routing decisions).

3) Diagnose with multiple metrics.
Compute:

- **Statistical Parity Difference** and **Disparate Impact** for "high priority" assignment.
- **Equal Opportunity Difference** (TPR gaps) and **Average Odds Difference** across groups, using a clinically relevant notion of "truly high-need" (e.g., malignant + adverse markers).
- **Calibration within groups** (predicted vs observed risk) and **Theil index** (individual unfairness).
  Report point estimates with confidence intervals (bootstrap) and perform **intersectional** analysis (e.g., age×race), not just single-axis.

4) Investigate sources.
Use feature distribution diagnostics (PSI/KS tests by group), error analysis (per-group confusion matrices), and **counterfactual** checks (e.g., flip protected attributes where feasible or use matched cohorts) to separate data shift from model behavior.

5) Mitigate using AIF360 workflows.

- **Pre-processing:**
  - *Reweighing* to balance group-label associations.
  - *Disparate Impact Remover* (careful with clinical features—avoid removing medically meaningful signal).
- **In-processing:**
  - *Adversarial Debiasing* to reduce group signal in representations.
  - *Prejudice Remover Regularizer* to penalize unfair correlations.
- **Post-processing:**
  - *Equalized Odds* or *Calibrated Equalized Odds* to adjust decision thresholds per group while preserving calibration.
  - *Reject Option Classification* near the boundary to favor the disadvantaged group where uncertainty is highest.

For our pipeline, a pragmatic starting point is: (a) **group-aware calibration** (Platt/Isotonic per group), (b) **group-specific thresholds** chosen to equalize TPR at a fixed FPR band, and (c) **reweighing** during training. Re-evaluate utility trade-offs (overall F1, service level agreements) alongside fairness metrics.

6) Revisit the priority label design.
Replace global tertile bins with **clinically grounded thresholds** (e.g., risk > x% ⇒ high). Validate thresholds per group to ensure comparable sensitivity to serious cases. If prevalence differs by group, consider **constrained optimization** to meet minimum TPR/PPV targets for each group.

7) Governance & operations.

- Publish a **Model Card** and **Data Sheet** documenting populations, known limitations, fairness metrics, and monitoring plans.
- Implement **ongoing monitoring**: drift detection, per-group metrics in production, alerting when gaps exceed agreed limits.
- Run **temporal and site-held-out** validations before deployment.
- Establish a **feedback loop** (clinician/analyst review of escalations and misses) and a **human-in-the-loop** override, especially for borderline cases.

8) Communication & consent.
Be transparent about how priority is assigned, what safeguards exist, and how individuals can contest or seek human review. Ensure access policies don't systematically disadvantage those with limited digital literacy or access.

Bottom line: Fairness isn't a one-time filter. With AIF360 we can quantify gaps, apply targeted mitigations (pre/in/post-processing), and—equally important—rethink our **labels and thresholds** to align with clinical/operational equity goals. Continuous auditing, group-aware calibration, and clear governance turn a high-performing prototype into a **trustworthy** production system.