Bias and Fairness

Bias Detection and Mitigation Pipeline

Bias Sources

Historical Bias
Past practices reflected
in training data

Representation Bias
Underrepresented groups:
Age, gender, ethnicity, geography

Measurement Bias
Different equipment quality
across hospitals/regions

Aggregation Bias
One-size-fits-all model
ignores subgroup differences

Detection

Subgroup Performance

Male	Female	White	Black	Asian
92%	87%	95%	78%	82%

Fairness Metrics

Demographic Parity: 0.72
Equalized Odds: 0.81
Equal Opportunity: 0.85
Calibration Gap: 0.12
⚠ Significant disparities

Mitigation

Pre-processing
- Balanced sampling
- Data augmentation for minorities
- Synthetic data generation

In-processing
- Fairness constraints in loss
- Adversarial debiasing
- Multi-task learning

Post-processing
- Threshold optimization
- Calibration per subgroup
- Output recalibration

After Mitigation
Performance

	Before	After
Black	78%	86%
Asian	82%	89%
Female	87%	90%

Trade-offs
Overall: 91% → 89%
✓ Disparity: 17% → 4%
Fairness-accuracy balance

Continuous Monitoring
Track fairness over time

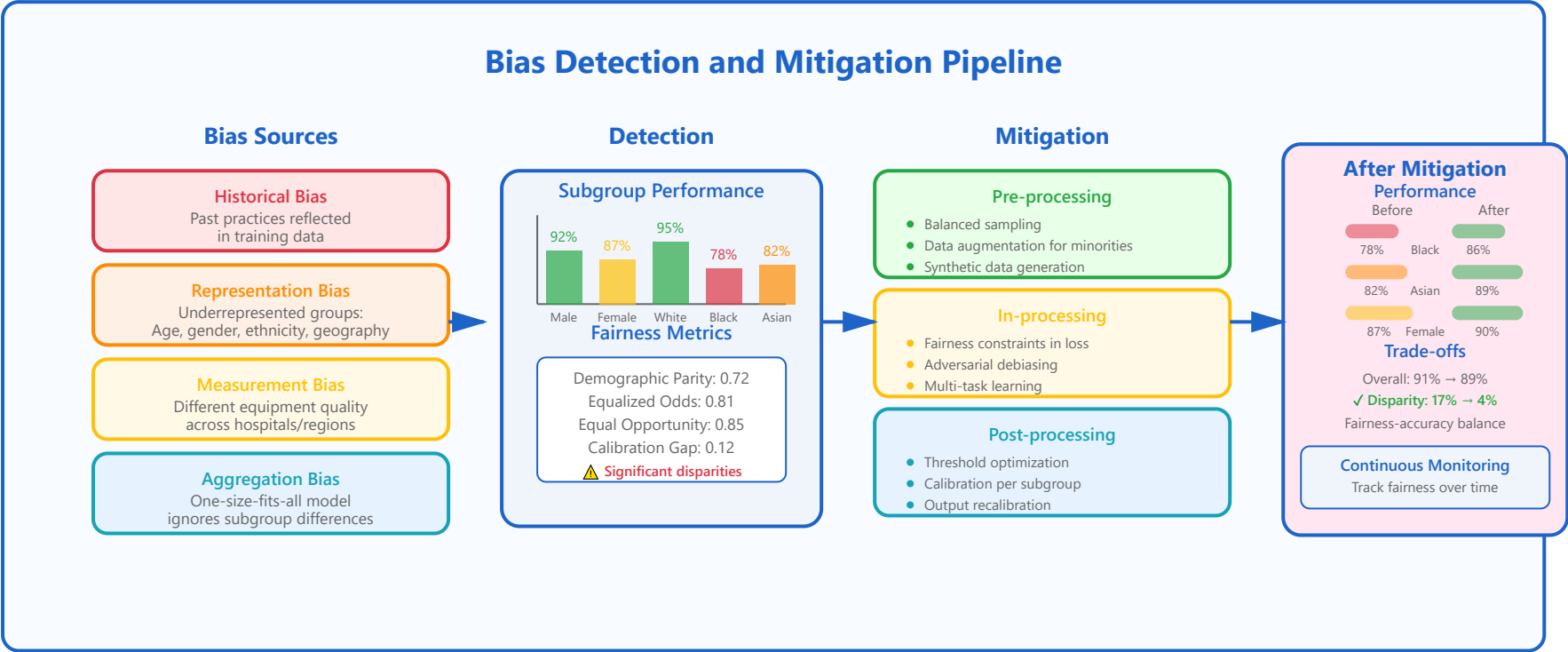## Dataset Bias

Selection bias, label bias. Underrepresentation of certain demographics

Demographic Disparities

Performance gaps across age, sex, race. Different disease prevalence and presentation

Fairness Metrics

Equalized odds, demographic parity. Group fairness vs individual fairness

Mitigation Strategies

Diverse training data, reweighting, adversarial debiasing. Fairness constraints

Continuous Monitoring

Track performance by subgroup. Detect and address emerging disparities