Model Performance Tracking

Continuously monitor ML model performance in production to detect degradation

Classification Metrics

Track accuracy, precision, recall, F1-score

Monitor diagnostic accuracy

Inference Latency

P50, P95, P99 response times

Real-time diagnosis < 2s

Throughput

Requests/sec, batch processing rate

Handle peak clinic hours

Error Rates

Failed predictions, timeout errors

Alert if errors > 1%

Calibration

Predicted probabilities vs actual outcomes

80% predictions = 80% accurate?

Fairness Metrics

Performance across demographic groups

Equal accuracy across groups

When to Alert

- Accuracy drops > 5%
- Latency exceeds SLA
- Error rate spikes
- Calibration drift