Avoiding Harms in Classification

Beyond Accuracy, Towards Justice

Accuracy is a Trap

Dr. Ahmed Awais

November 3, 2025

1 Background

The primary goal of a traditional machine learning classification task has been to maximize **accuracy**. We define a model $f : \mathcal{X} \to \mathcal{Y}$ that maps inputs $\mathbf{x}$ to labels y, and we measure its success by the fraction of correct predictions on a test set.

However, in real-world applications, especially those that impact human lives, a myopic focus on accuracy is not just insufficient—it can be **dangerous**. This lecture focuses on the critical practice of **Avoiding Harms in Classification.**

1.1 Core Concepts

Avoiding Harms means proactively identifying, measuring, and mitigating negative, unfair, or damaging consequences that a classification system can have on individuals, groups, and society, even when the system is highly *accurate*.

2 Why Accuracy is a Trap: The Confusion Matrix View

To understand why accuracy is deceptive, we must first deconstruct it using the **Confusion Matrix**.

2.1 The Confusion Matrix

For a binary classification problem (e.g., "Grant Loan" vs. "Deny Loan"), the performance of a classifier can be summarized as follows:

		Actual Class	
		Positive (P)	Negative (N)
Predicted Class	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

From this, we define Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2.2 The Trap Unveiled: A Toy Example

Imagine we are building a model to screen applicants for a prestigious scholarship. The dataset has 1000 applicants, only 20 of whom (2%) truly deserve it (Positive class). Our goal is to classify applicants as "**Deserving**" or "**Not Deserving**".

- **Model A (The "Naive" Model):** This model is lazy. It simply classifies *every single applicant* as "**Not Deserving**".

Let's look at its confusion matrix:

	Actual: Deserving	**Actual: Not Deserving**
Predicted: Deserving	0	0
Predicted: Not Deserving	20	980

$$\text{Accuracy}_{\text{Model A}} = \frac{0 + 980}{1000} = 98\%$$

A **98% accurate** model! But it is utterly useless and **profoundly harmful**. It has a **False Negative Rate (FNR)** of 100%:

$$\text{FNR} = \frac{FN}{FN + TP} = \frac{20}{20 + 0} = 1.0$$

It failed to identify a single deserving student, effectively shutting them out of an opportunity.

Now, consider a smarter model.

- **Model B (The "Fairer" Model):** This model is more sophisticated. It correctly identifies 15 deserving students but also makes some mistakes.

	Actual: Deserving	**Actual: Not Deserving**
Predicted: Deserving	15	40
Predicted: Not Deserving	5	940

$$\text{Accuracy}_{\text{Model B}} = \frac{15 + 940}{1000} = 95.5\%$$

Model B has a **lower accuracy** (95.5%) than Model A, but it is **immeasurably better and fairer**. It successfully allocates the opportunity to 15 deserving students. Its FNR is much lower: $\frac{5}{20} = 25\%$.

2.3 Conclusion of the Example

The pursuit of accuracy alone would have led us to select the **harmful** Model A. We must look **beyond accuracy** into the specific types of errors a model makes.

3 Taxonomy of Harms in Classification

Harms can be categorized into two primary types, which often intersect.

3.1 1. Allocation Harms

These occur when a system unfairly allocates resources, opportunities, or burdens. The harm is in the **denial of a benefit** or **imposition of a cost**.

- **Example 1: Hiring Tool**
  - A model is trained on historical hiring data from a male-dominated tech company to classify resumes as "`Hire`" or "`Reject`".
  - **Harm:** The model learns to associate being male with competence, leading to a high **False Negative** rate for female candidates. Qualified women are systematically denied job opportunities.
  - **Confusion Matrix Impact:** High FN for the protected group (women).
- **Example 2: Loan Application System**
  - A bank uses a model to classify applicants as "`Low-Risk`" or "`High-Risk`".
  - **Harm:** The model uses zip code as a feature, which acts as a *proxy* for race. This leads to a high **False Positive** rate for applicants from minority neighborhoods, incorrectly classifying good candidates as high-risk and denying them loans.
  - **Confusion Matrix Impact:** High FP for the protected group.

3.2 2. Representation Harms

These occur when a system reinforces negative stereotypes, erases social groups, or delivers degrading results. The harm is to a group's **dignity and social standing**.

- **Example 1: Image Recognition**
  - A photo app's classifier labels images of people.
  - **Harm:** The model consistently fails to detect people with dark skin tones or, worse, labels them with offensive terms like "`gorilla`". This is a form of **erasure** and **insult**.
- **Example 2: Language Models**
  - An autocomplete system suggests the next word in a sentence.
  - **Harm:** Given "The nurse is...", it suggests "`kind`" and "`female`"; given "The CEO is...", it suggests "`driven`" and "`male`". This **reinforces harmful social stereotypes**.

4 Fairness Metrics: Looking Beyond Accuracy

To quantify and mitigate these harms, we define metrics based on the confusion matrix, often calculated separately for different protected groups (e.g., Group A and Group B).

4.1 Key Metrics from the Confusion Matrix

- **False Positive Rate (FPR):** $\text{FPR} = \frac{FP}{N}$
  - *Interpretation:* What fraction of truly negative people were incorrectly flagged?
  - *Harm:* An innocent person is punished or denied a benefit.
- **False Negative Rate (FNR):** $\text{FNR} = \frac{FN}{P}$
  - *Interpretation:* What fraction of truly positive people were incorrectly missed? *Harm:* A deserving person is denied an opportunity or help.
- **True Positive Rate (TPR) / Recall / Sensitivity:** $\text{TPR} = \frac{TP}{P} = 1 - \text{FNR}$
- **True Negative Rate (TNR) / Specificity:** $\text{TNR} = \frac{TN}{N} = 1 - \text{FPR}$

4.2 Group Fairness Definitions

Using these rates, we can define statistical fairness criteria:

1. **Demographic Parity (Independence)**

$$P(\hat{Y} = 1 | \text{Group=A}) = P(\hat{Y} = 1 | \text{Group=B})$$

"The selection rate is the same for all groups." Focuses on the outcome, not the correctness.

2. **Equalized Odds (Separation)**

$$\text{TPR}_{\text{Group=A}} = \text{TPR}_{\text{Group=B}} \quad \text{and}$$
$$\text{FPR}_{\text{Group=A}} = \text{FPR}_{\text{Group=B}}$$

"The model has the same error rates across groups." A core metric for avoiding allocation harms. It requires the model to be equally good at identifying positives and negatives in all groups.

3. **Predictive Parity (Sufficiency)**

$$P(Y = 1 | \hat{Y} = 1, \text{Group=A}) = P(Y = 1 | \hat{Y} = 1, \text{Group=B})$$

"Of those predicted to be positive, the same fraction should actually be positive in all groups." This is about the precision of the model being equal across groups.

5 A Framework for Mitigating Harms

Avoiding harm is an active process integrated throughout the ML lifecycle.

1. **Problem Formulation: Ask:** "Should we even build this system?" Consider power dynamics and potential for misuse.
2. **Data Collection & Inspection: Audit the data** for historical biases and representation. Is one group under-represented? Are the labels themselves biased?
3. **Model Training & Selection:**
   - **Pre-processing:** Modify the training data to remove biases (e.g., reweighting, resampling).
   - **In-processing:** Add fairness constraints (e.g., `fairlearn`, `AIF360`) directly to the model's optimization objective to enforce metrics like **Equalized Odds**.
   - **Post-processing:** Adjust decision thresholds for different groups after the model is trained to equalize FPR and FNR.
4. **Evaluation & Testing: Disaggregate evaluation!** Report performance and fairness metrics (Accuracy, FPR, FNR, etc.) for all relevant subgroups. Use "slicing" to find blind spots.
5. **Deployment & Monitoring:** Monitor for **model drift**—the world changes, and a fair model can become unfair over time. Implement a human-in-the-loop for high-stakes decisions and a clear appeals process.

6 Conclusion

- **Accuracy is a dangerously incomplete metric** for evaluating classifiers in sociotechnical systems.
- Harms are real and can be categorized as **Allocation** and **Representation** harms.
- The **Confusion Matrix** is our fundamental tool for diagnosing these harms through metrics like FPR and FNR.
- **Fairness** must be explicitly defined, measured, and optimized for using metrics like **Equalized Odds**.
- Avoiding harm is an **ongoing, proactive responsibility** that requires integrating ethical considerations into every stage of the ML pipeline.

The goal is not just to build accurate models, but to build **just and equitable sociotechnical systems**.