Fairness-Aware Machine Learning for Social Bias Detection in Healthcare Research Datasets

Ilerioluwakiiye Abolade, Precious Kolawole

ML Collective, Carleton University

Motivation

AI models in healthcare can unintentionally discriminate against vulnerable groups. This work aims to detect and quantify biases in healthcare data and predictive models before deployment.

- How can researchers identify data-level and algorithm-level biases?
- How do neural networks compare to traditional models in balancing accuracy and fairness?

Introducing the Social Bias Detection Tool

We present a lightweight, interactive tool to:

- Detect bias in healthcare data and models
- Compute fairness metrics (SPD, EOD, DD)
- Compare traditional ML vs neural nets on both accuracy and fairness
- Output a "combined score" for model selection

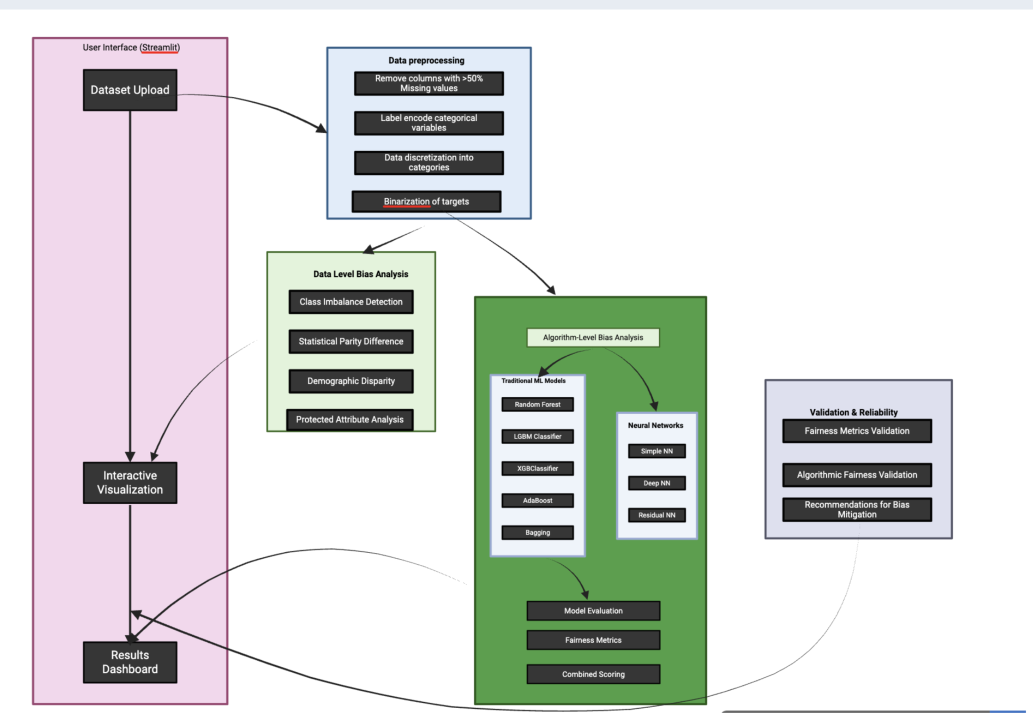


Figure 1:Architectural overview of the Social Bias Detection framework

Data-Level Bias Exists, Even Before Training

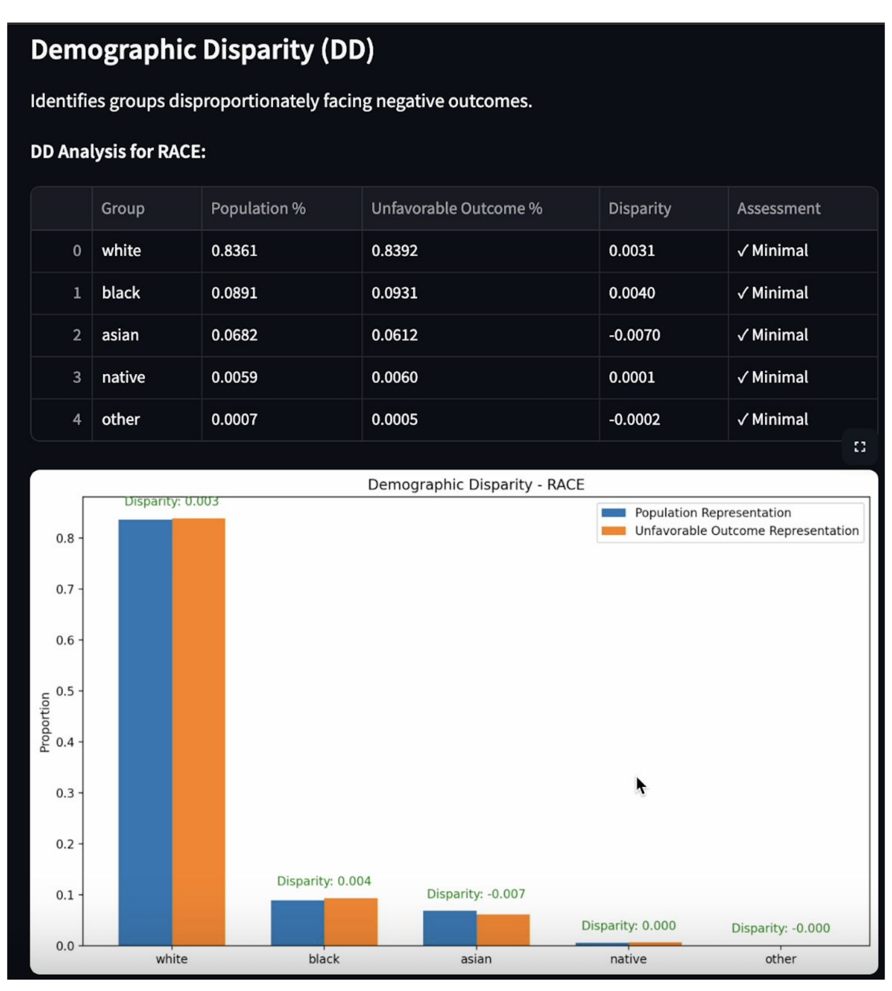
We evaluated two real-world healthcare datasets:

SyntheticMass:

- 83.6% White patients → major racial imbalance
- Age disparity: SPD = 0.82 for 0–35 vs 65+ (substantial)

Brain Stroke Dataset:

- Demographics more balanced overall
- One exception: SPD = 0.10 for 65+ vs 51–65



	Group	Population %	Unfavorable Outcome %	Disparity	Assessment
0	white	0.8361	0.8392	0.0031	✓ Minimal
1	black	0.0891	0.0931	0.0040	✓ Minimal
2	asian	0.0682	0.0612	-0.0070	✓ Minimal
3	native	0.0059	0.0060	0.0001	✓ Minimal
4	other	0.0007	0.0005	-0.0002	✓ Minimal

Figure 2:Population vs. unfavorable outcomes by race, showing minimal to moderate pre-training bias.

This shows that bias can be embedded in the data itself, independent of modeling.

Fairness Metrics Used

We assess bias using industry-standard metrics:

1. Statistical Parity Difference (SPD)

Difference in favorable outcomes across groups:

$$\text{SPD} = P(\hat{Y} = 1 \mid A = a) - P(\hat{Y} = 1 \mid A = b)$$

2. Equal Opportunity Difference (EOD)

Gap in true positive rates across groups:

$$\text{EOD} = \text{TPR}_a - \text{TPR}_b$$

3. Average Odds Difference (AOD)

Average gap in true *and* false positive rates:

$$\text{AOD} = \frac{1}{2}\Big[(\text{FPR}_a - \text{FPR}_b) + (\text{TPR}_a - \text{TPR}_b)\Big]$$

4. Demographic Disparity (DD)

Difference between a group's outcome share and its population share:

$$\text{DD} = P(A = a \mid Y = 1) - P(A = a)$$

Interpretation thresholds:
0.00–0.05: Minimal **0.05–0.10**: Small > 0.10: Substantial bias

How We Compare Models Fairly

High accuracy $\neq$ fair predictions. We introduce the **Combined Score** to balance both.

$$\text{Combined Score} = 0.5 \times \text{Normalized Accuracy} + 0.5 \times \text{Fairness Score}$$

where:

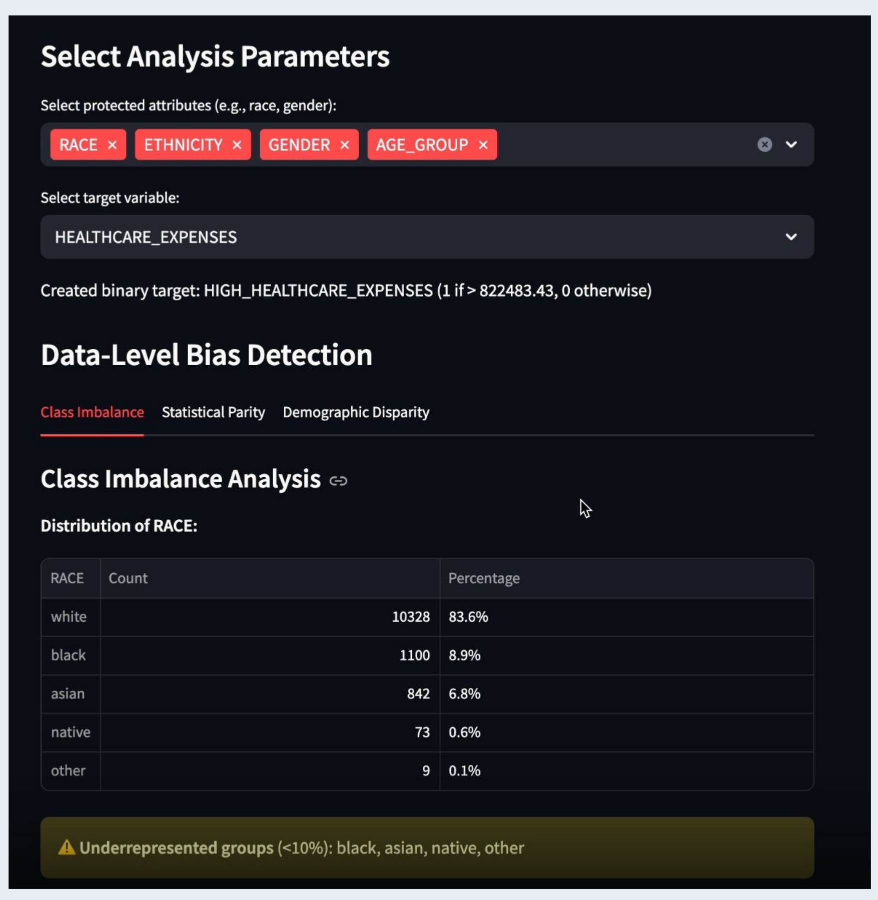
$$\text{Fairness Score} = 1 - \frac{|\text{SPD}|}{|\text{SPD}_{\max}|} - \frac{|\text{EOD}|}{|\text{EOD}_{\max}|}$$

Lower SPD and EOD values increase the fairness score, rewarding models that treat groups more equally.

How the Tool Works

1. Data-Level Analysis:

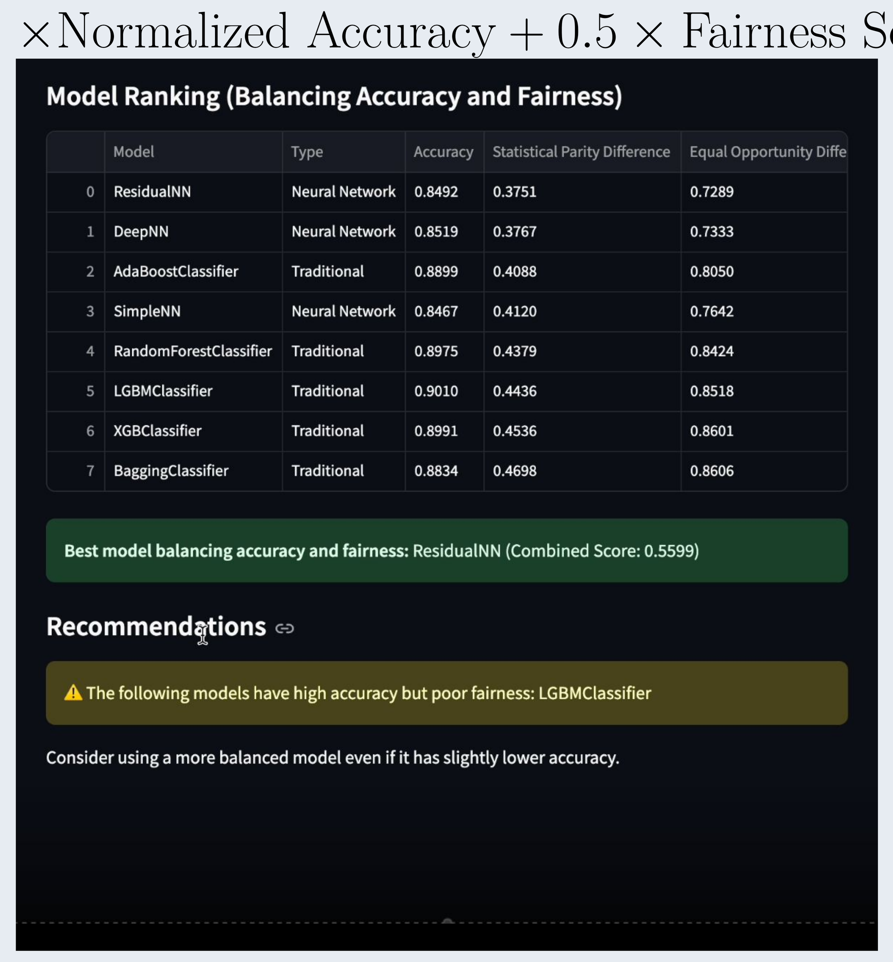
- Class imbalance across demographic groups
- Statistical Parity Difference (SPD) between protected groups
- Demographic Disparity (DD) in outcomes vs. population share



RACE	Count	Percentage
white	10328	83.6%
black	1100	8.9%
asian	842	6.8%
native	73	0.6%
other	9	0.1%

2. Algorithmic Bias Analysis:

- Trains both traditional ML and neural networks
- Evaluates SPD, Equal Opportunity Difference (EOD), and Average Odds Difference (AOD)
- Computes a **Combined Score** $= 0.5 \times \text{Normalized Accuracy} + 0.5 \times \text{Fairness Score}$



	Model	Type	Accuracy	Statistical Parity Difference	Equal Opportunity Diffe
0	ResidualNN	Neural Network	0.8492	0.3751	0.7289
1	DeepNN	Neural Network	0.8519	0.3767	0.7333
2	AdaBoostClassifier	Traditional	0.8899	0.4088	0.8050
3	SimpleNN	Neural Network	0.8467	0.4120	0.7642
4	RandomForestClassifier	Traditional	0.8975	0.4379	0.8424
5	LGBMClassifier	Traditional	0.9010	0.4436	0.8518
6	XGBClassifier	Traditional	0.8991	0.4536	0.8601
7	BaggingClassifier	Traditional	0.8834	0.4698	0.8606

Main Findings

- Neural networks generally achieved higher fairness without sacrificing accuracy.
- ResidualNN scored highest in the SyntheticMass dataset for balancing both metrics.
- DeepNN achieved near-perfect fairness in Brain Stroke predictions.

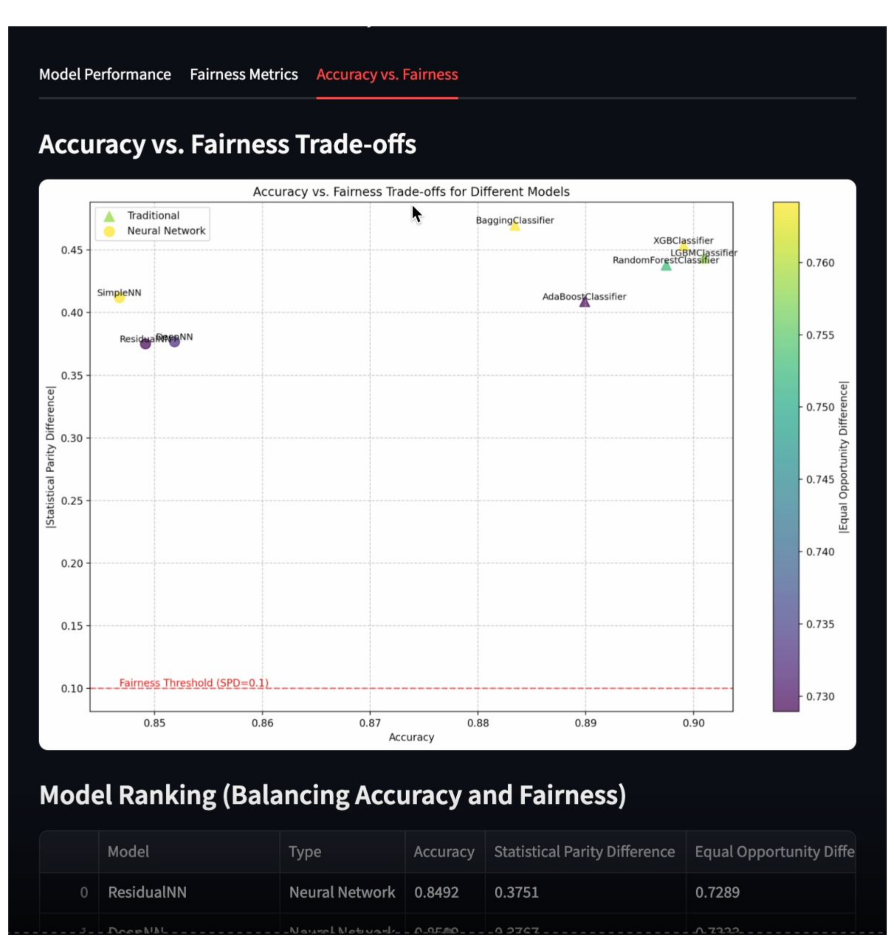


Figure 3:Accuracy vs Fairness trade-offs across models

Key Takeaways

- Bias is present in many healthcare datasets before training.
- Our tool enables quick, transparent bias assessment.
- Neural networks can be both accurate *and* fair.
- The Combined Score metric helps balance ethics with performance.

Acknowledgements

We thank Professor Matthew S. Holden for mentorship and feedback.

Project Repository

github.com/precillieo/social-bias-detection-tool

DEEP LEARNING INDABA