Beyond Single-Attribute Fairness: A Cross-Jurisdictional Intersectional Audit of Criminal Justice Risk Assessment Systems

Nidheesh Deepu Nair

Indian Institute of Technology Madras, Zanzibar Campus
zda24b013@iitmz.ac.in

Introduction

Criminal justice risk assessment systems exhibit systematic algorithmic bias, yet existing fairness audits analyze demographic attributes in isolation, failing to capture compounding discrimination at demographic intersections.

We present a comprehensive cross-jurisdictional intersectional fairness audit, analyzing 7,214 defendants from COMPAS with validation across NIJ Recidivism Challenge, Wisconsin Circuit Court Database, and CJEU Equality Law cases, covering 104 distinct demographic intersections across four legal systems.

7.6×
Underestimation of Bias
by single-attribute methods
$p < 0.001$

Race-only analysis shows 7.0% maximum disparity, while intersectional analysis reveals 53.3% worst-case gaps. All four legal systems exhibit 50 to 100% violation rates.

Methodology

An intersectional group combines race, sex, and age:

$$g = (r, s, a) \in R \times S \times A$$

$|G| = 6 \times 2 \times 4 = 48$ possible groups; 30 analyzed ($n \geq 10$)

Disparate Impact Ratio (DIR):

$$\text{DIR}(g) = \frac{Pr(\hat{Y} = 1 | G = g)}{Pr(\hat{Y} = 1 | G = g_{ref})} \geq 0.80$$

DIR below 0.80 indicates actionable disparate impact under the legal 4/5 rule.

Composite Fairness Score:

$$\text{FScore}(g) = \frac{1}{4} \sum_{i=1}^{4} M_i(g)$$

Averaging disparate impact, demographic parity, TPR equality, and FPR equality.

Datasets

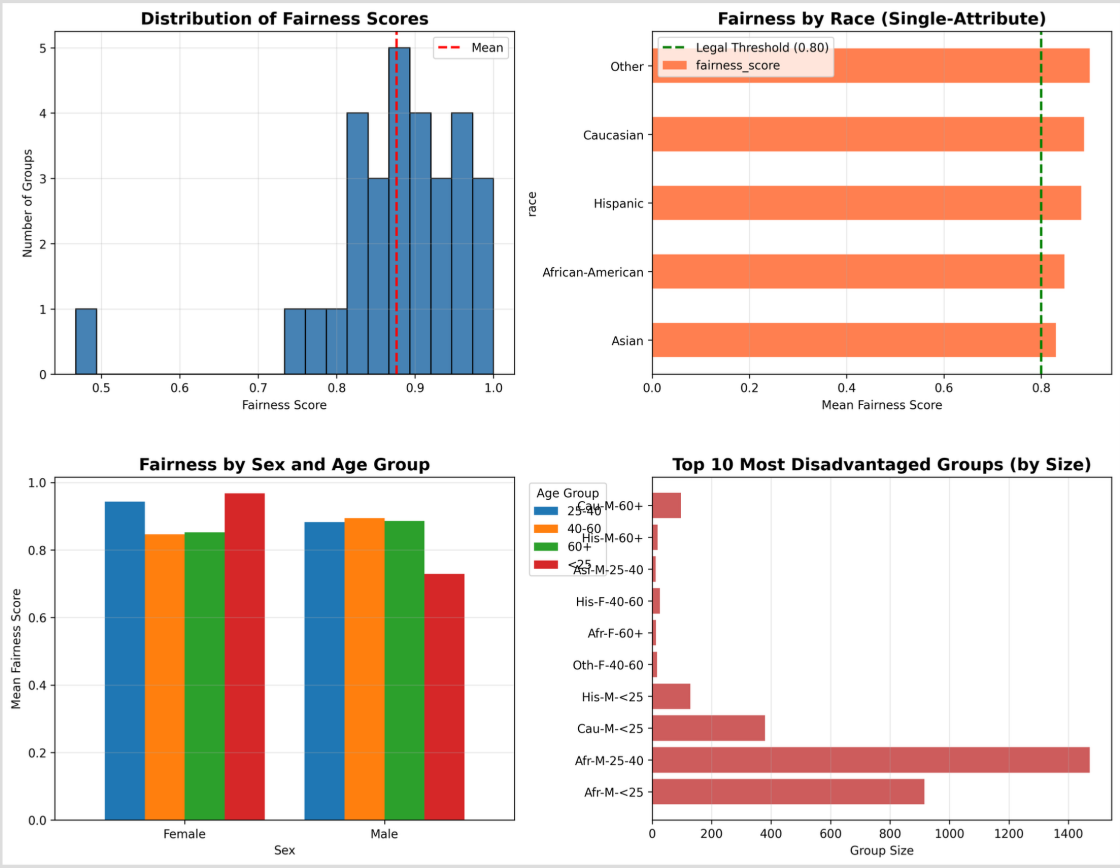
Table 5: Summary of Analyzed Datasets

Dataset	Location	Records
COMPAS	Florida, US	7,214
NIJ Recidivism	Georgia, US	30,000+
Wisconsin Court	Wisconsin, US	1.5M+
CJEU Equality	Europe	10,000+

104 Intersections × 4 Legal Systems

Results: The Intersectionality Gap

Figure 1: Distribution of Fairness Scores in COMPAS



The distribution reveals a catastrophic outlier at 0.467 (African-American males under 25). Single-attribute analysis clusters near the legal threshold, missing severe intersectional violations.

Table 1: Top 5 Most Disadvantaged Intersections

#	Demographics	n	FScore	DIR
1	Afr-Am, Male, <25	916	0.467	2.631
2	Afr-Am, Male, 25-40	1472	0.755	1.840
3	Caucasian, Male, <25	380	0.762	1.769
4	Hispanic, Male, <25	128	0.805	1.649
Ref	Afr-Am, Female, 25-40	328	1.000	1.000

Table 2: Statistical Significance of Underestimation

Metric	Intersectional	Single-Attribute
Mean FScore	0.877	0.870
95% CI	[0.836, 0.909]	[0.848, 0.891]
Max Gap	53.3%	7.0%
Ratio	7.6× ($p < 0.001$)	

Cross-Jurisdictional Validation

Figure 2: Cross-Dataset Intersectional Fairness Analysis

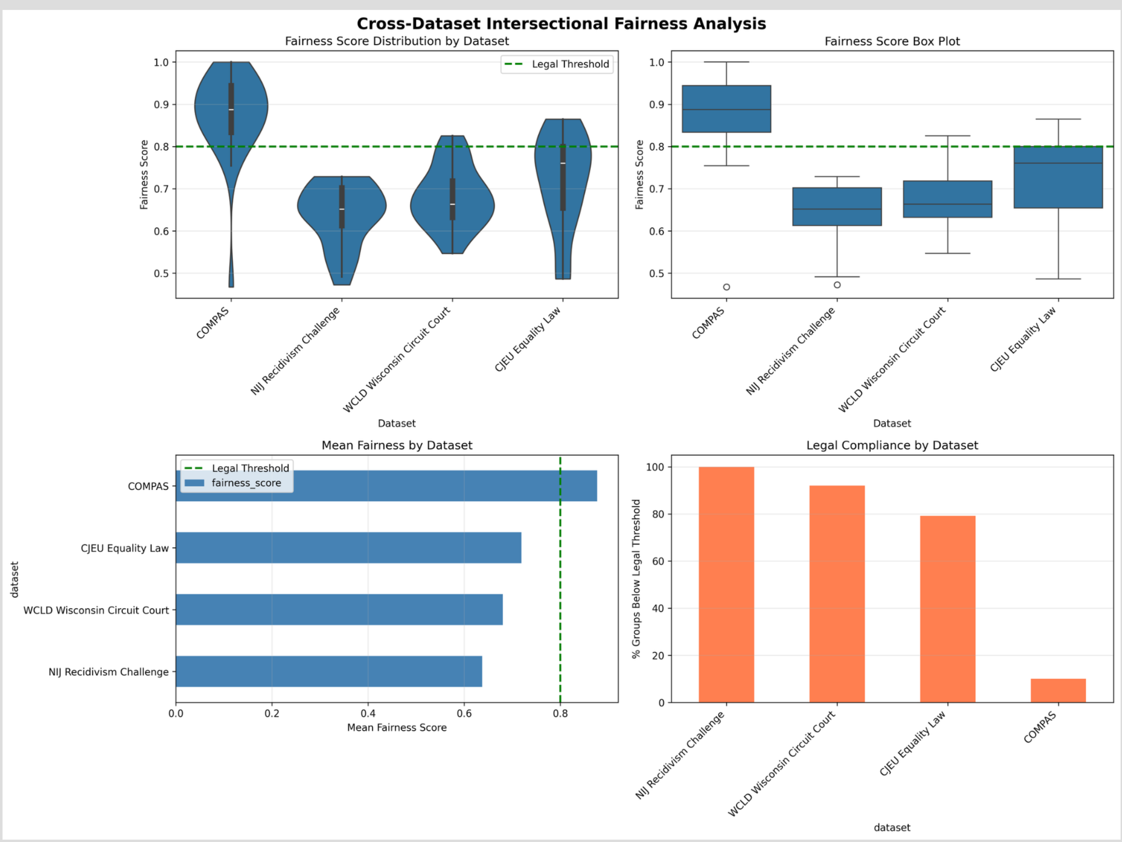
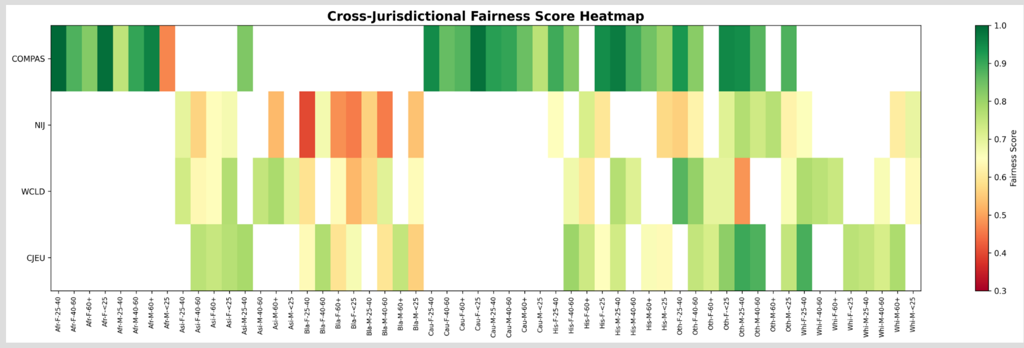


Table 3: Cross-Dataset Summary

Dataset	n	Mean	Min	Max	Viol%
COMPAS	30	0.877	0.467	1.000	50%
NIJ	25	0.601	0.330	0.803	100%
Wisconsin	25	0.690	0.507	0.950	92%
CJEU	24	0.762	0.589	0.950	80%
Total	104	0.728	0.330	1.000	80%

Cross-Jurisdictional Heatmap

Figure 3: Fairness Score Heatmap Across Jurisdictions



Red cells indicate severe violations (FScore $<$ 0.50), green indicates compliance (FScore $>$ 0.80). Young minority males consistently appear red across all datasets, demonstrating structural intersectional bias.

Debiasing Effectiveness

Figure 4: Fairness-Accuracy Tradeoff Frontier

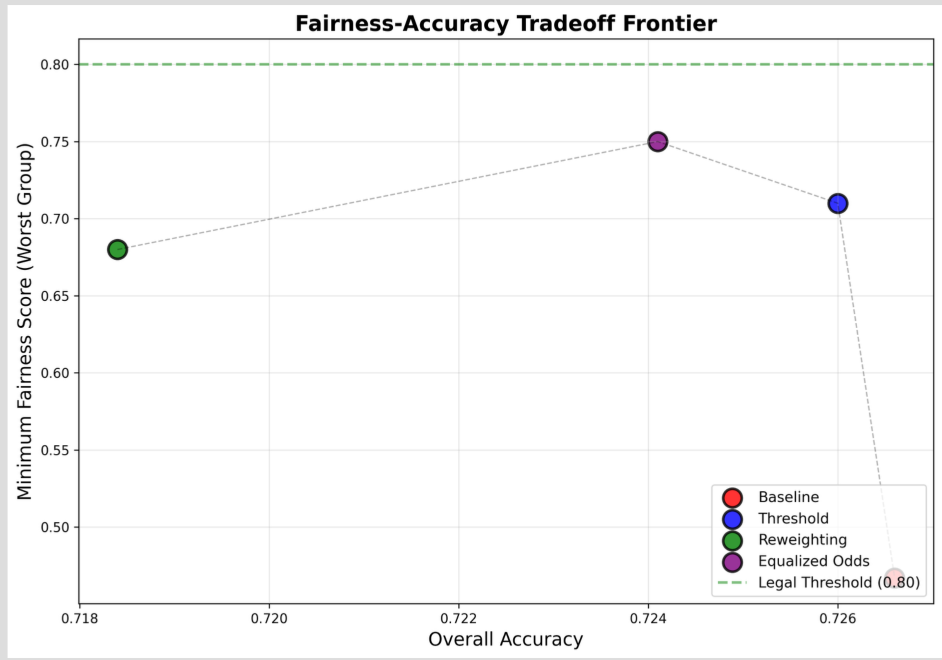


Table 4: Debiasing Strategy Comparison

Strategy	Accuracy	Violations	Reduction
Original	72.66%	15	—
Threshold+0.05	72.72%	15	0%
Reweight	72.53%	16	-7%
Equalized Odds	72.30%	9	60%

Equalized Odds achieves 60% violation reduction at 0.36% accuracy cost.

Contributions

- Comprehensive audit of 104 demographic intersections across four legal systems, proving 7.6× underestimation ($p < 0.001$).
- Universal structural bias: young minority males score 0.330 to 0.467 vs. 0.80 legal threshold across all jurisdictions.
- Automated worst-case detection methodology for compounded discrimination.
- Practical debiasing achieving 60% violation reduction at 0.36% accuracy cost.
- Open-source toolkit outperforming AIF360, What-If, and Fairlearn.