MANKIND QUARTERLY
Mankind Quarterly, 66(3), 677-705 2026 Spring Edition
ISSN: 0025-2344

Detecting Systematic Bias in Criminal Racial Assignment

Daniel Lee Van Pelt*†

Abstract

We analysed racial classification in U.S. Department of Corrections databases across 14 states comprising 1.5 million criminal records. An accurate linear model trained on biased data learns the underlying signal rather than the bias itself; we interpret systematic deviations between model predictions and official classifications as evidence of mislabelling by authorities rather than model error. Using facial recognition algorithms on mugshots and name-based demographic data, we achieved 92.76% agreement with assigned race labels. We identified substantial misclassification: 29% of predicted Hispanics were officially assigned as White. This pattern persisted among high-confidence model predictions (median confidence 91%). Correcting for misclassification increased Hispanic criminal count rates by 31%, decreased White rates by 6%, and decreased Black rates by 1%. Simulation studies confirmed the pattern resembled random rather than deliberate bias. State-level analysis (n = 14) revealed no statistically significant association with political ideology (r = .21, 95% CI: -0.36 to 0.67, p = .473). The proportion of predicted Hispanics assigned White and the proportion of predicted Whites assigned Hispanic both correlated with Native American ancestry among Latinos (r = -.80, 95% CI: -0.95 to -0.38, p = .003, n = 11; r = .74, 95% CI: 0.26 to 0.93, p = .009, n = 11).

Keywords: Racial classification, Criminal justice, Data quality, Hispanic, Facial recognition, Administrative records, Measurement error

1 Introduction

Racial differences have been studied across every stage of America's criminal justice system, from initial police contact to final sentencing (Figuereo et al., 2025). This research investigates whether observed differences reflect systematic discrimination or other factors. It showed that after controlling for IQ, impulsivity, and criminal history, race becomes unrelated to arrest probability (Beaver et al., 2013; Schwartz & Beaver, 2019). Similarly, controlling for offense severity and criminal history eliminates racial bias in sentencing across 43 U.S. states (thelawofaverages, 2023). Experimental and observational studies with rigorous controls find either no racial bias against minorities or bias favouring minorities over Whites:

1. The first randomized controlled experiment on prosecutorial bias tested nationwide prosecutors with realistic case vignettes and found no statistically significant relationship between defendant race and charging decisions, with some analyses showing pro-black treatment (Robertson et al., 2019).

2. Experimental research using realistic deadly force simulators found officers were slower to shoot armed Black suspects than armed White suspects, and less likely to shoot unarmed Black suspects despite showing implicit bias (James et al., 2016).

3. The largest jury study ever conducted analysed 300,000+ felony cases over 32 years and found no taste-based or statistical discrimination against Black defendants, with similar disparate impact when race was unknown to jurors (Hoekstra et al., 2023).

†Independent researcher
*Corresponding author: DanielLeeVanPelt@pm.me

4. Comprehensive analysis of sentencing data across 43 U.S. states found either no racial bias in sentence length or bias favouring minorities, with studies consistently showing that legal factors (offense severity, criminal history) rather than race determine sentencing outcomes (thelawofaverages, 2023).

5. Race becomes unrelated to arrest probability after controlling for IQ, impulsivity, and criminal history in samples of 1,331 ex-convicts and other populations (Beaver et al., 2013; Schwartz & Beaver, 2019).

6. Direct measurement using cameras found that the proportion of speeding drivers who were Black mirrored the proportion of Black drivers stopped by police (Lange et al., 2005).

7. Using the rate of attacks on police as a benchmark, Black Americans were 40% less likely to be shot by police than White Americans (Shjarback & Nix, 2020).

8. Multiple studies comparing arrest rates to incident reports found either no racial bias or pro-black bias in arrests, with consistent findings across 22 crime types (Beck, 2021; D'Alessio & Stolzenberg, 2003; Rubenstein, 2016).

9. Analysis using violent crime rates as a benchmark found White people over-represented among police killings, with no evidence of anti-black bias in most estimates (Cesario et al., 2019).

This pattern extends across numerous additional studies spanning multiple domains of criminal justice. In police stops, Black and White Americans report similar contact rates while Hispanics report lower rates; White officers prove less likely than Black officers to stop Black citizens; citizen observers document equal treatment across races; and New York's stop-and-frisk data revealed identical arrest rates when stopped. Search patterns show no racial bias post-2010 according to recent meta-analyses, with minimal hit rate differences between races. Police killings data reveal that Black officers kill Black citizens at equivalent rates to White officers, and publication bias correction eliminates apparent force disparities. Incarceration rates in Pennsylvania and nationally align with arrest patterns. Sentencing meta-analyses spanning dozens of studies find no bias after controlling for legal factors, with Black and White judges imposing similar sentences. Even in death penalty cases, 11 of 14 post-1981 studies found no racial effects, with psychological controls revealing pro-Black bias and historical execution rates 7% lower than expected for Blacks (Figuereo et al., 2025).

These studies rely on racial classifications recorded in administrative data. We examine whether racial classifications in corrections databases are accurate. Systematic misclassification would distort estimated crime rates by race, artificially inflating rates for the group individuals are misclassified into and deflating rates for their actual racial group.

We obtained records for 5.5 million criminals from 39 U.S. states through web scraping. After applying data quality requirements, our final dataset comprises 1.5 million criminal records with complete mugshot, demographic, and naming information. We developed predictive models incorporating 21 variables derived from facial recognition algorithms, demographic name databases, and census data, achieving 92.76% agreement with DOC-assigned race labels for three categories (Black, White, Hispanic). By comparing model predictions to official racial assignments, we identify systematic patterns of misclassification.

2 Data

2.1 Quantity and Quality

Every U.S. state maintains a Department of Corrections, and most operate websites allowing public access to criminal records. Using web scraping technologies, we attempted to collect complete databases from all 51 jurisdictions. We successfully scraped 5.5 million distinct criminal records from 39 states. The 12 unsuccessful states (shown in red in Figure 1) either employed sophisticated anti-scraping measures, operated dysfunctional websites, or outsourced their databases to third-party systems that proved inaccessible. Our

bias analysis required specific data elements from each state. We established minimum requirements for inclusion:

1. Mugshots
2. Racial classifications, with White, Hispanic and Black as distinct categories
3. Sex classifications
4. Complete names (first, middle, last) and suffixes

Most states and individual records failed to meet these criteria. Data missingness occurred at two levels: Some states entirely lacked required fields (such as race or mugshot data), while others had individuals with missing information in otherwise complete datasets. Individual records frequently had race recorded as "Unknown" or lacked mugshots entirely. We excluded 16 states for lacking either race or mugshot data, and an additional 9 states for not recording Hispanic as a distinct racial category. This filtering process reduced our dataset from 5.5 million to 1.5 million records. Figure 1 displays the exclusion categories by state. The final list of valid states we included in our subsequent analyses is as follows: Arizona, Arkansas, Colorado, Florida, Illinois, Michigan, Mississippi, Oregon, Tennessee, Oklahoma, Nebraska, New Jersey, Nevada and Kentucky.

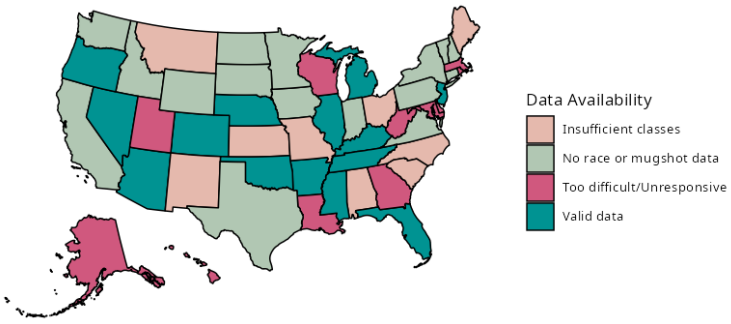

Figure 1: Disqualification of states from bias dataset

Furthermore, data quality varied dramatically across states in terms of quantity, completeness, and accessibility. North Carolina[1] exemplifies comprehensive data availability, allowing full database downloads with detailed records reaching back to the 1970s, yielding 1.2 million unique criminal records. In contrast, California[2], despite having four times North Carolina's population and similar incarceration rates, produced nearly ten times fewer accessible records after scraping. Figure 2 shows the final representation of criminal records by state. From our curated dataset, we extracted the following variables for our bias detection methodology:

[1] North Carolina criminal database https://webapps.doc.state.nc.us/opi/offendersearch.do?method=view

[2] California criminal database https://apps.cdcr.ca.gov/ciris/search

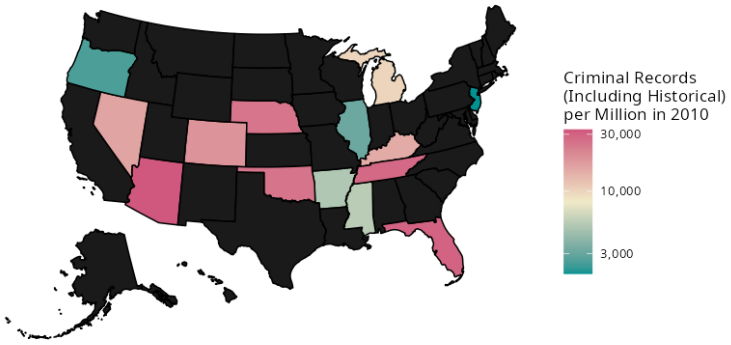

Figure 2: Representation of criminal figures by state

2.2 Race

We identified seven consistent racial classifications across our state datasets: Black, White, Hispanic, Asian or Pacific Islander, Native American, Multiracial, and "Unknown or Other".[3] States used varying nomenclature (e.g., "Hispanic or Latin American", "Mexican American", "Mexican National" for Hispanic categories), which we standardized for analysis. All qualifying states recorded Black, White, Hispanic, Asian or Pacific Islander, and Native American categories, with the sole exception of Florida, which did not assign Native American classifications.

2.3 Mugshots

We processed mugshots using (Serengil et al., 2024)[4], a comprehensive facial recognition[5] and classification framework. The system processes each image through facial detection, cropping, and alignment[6] before applying neural network classifiers. DeepFace's internal model generates predictions across six racial categories: Asian, Indian, Black, White, Middle Eastern, and Latino Hispanic. We incorporated these predictions as variables in our custom classification model rather than using DeepFace's outputs directly, because its racial categories did not align perfectly with the classifications used by Departments of Corrections. Direct application would have been ideal, providing an independent baseline trained outside our potentially biased dataset, but the category mismatch necessitated our custom modelling approach.

2.4 Sex

Sex classifications proved significantly cleaner and more consistent than racial data across states. Only New Mexico lacked sex classifications entirely. Since this state provided mugshots, we imputed missing sex data using DeepFace's built-in gender classification model.[7]

Preliminary analysis showed that including sex as a predictor variable provided no significant improvement to classification accuracy. We excluded sex from subsequent models to maintain parsimony.

[3] Classification methods vary by jurisdiction. Some states use inmate self-identification (e.g., New York: https://publicapps.doccs.ny.gov/ILookup/fpmsdoc.html) while others rely on officer assignment at arrest. Florida's exceptionally high Hispanic-to-White misclassification rate, combined with documented Cuban self-identification as White, suggests self-reporting plays a role in at least some states. We did not systematically investigate classification source across jurisdictions.

[4] https://github.com/serengil/deepface

[5] For the facial recognition engine we used Retinaface, with MTCNN and OpenCV set as fallbacks in case of detection failure

[6] The facial recognition system estimates the bounding box of the face, cropping it, the face is then aligned. This prepares it for the CNN racial and sexual classifier.

[7] We did not use name-based sex prediction since only New Mexico required imputation

2.5 Names

We extracted complete name information (first, middle, last names, and suffixes) for each individual from Department of Corrections records. To predict race from names, we utilized the comprehensive dataset from "Race and ethnicity data for first, middle, and surnames" (Rosenman et al., 2023), which provides racial breakdowns for each name segment across five categories: White, Black, Hispanic, Asian Pacific Islander, and "other" race.

This primary dataset yielded 15 predictive variables (5 races $\times$ 3 name segments). We supplemented this with U.S. 2010 census data (US Census Bureau, 2010) breakdowns by race for surnames. The census data included two categories not present in the Rosenman dataset: "American Indian" and "2 or more races", adding two more variables specifically for last names. Because no comprehensive suffix dataset exists[8], we created Boolean indicators for the most common suffixes: Jr/Sr designations, and Roman numerals.

Preliminary multinomial logistic regression analysis revealed that both middle name and suffix variables contributed insignificantly to classification accuracy. Middle names alone achieved 41.84% accuracy in five-race classification and 49.03% in three-race classification, worse than the naive baseline of predicting all individuals as White (54.46% and 55.7% respectively). Suffix variables added to first and last names provided no significant improvement in either classification scheme. When both variable sets were removed from the full model (including mugshot and census features), the effect on accuracy was insignificant. Additionally, principal component analysis showed middle name variables introduced spurious double clustering artefacts without improving racial separability. We excluded all middle name and suffix variables from subsequent analyses.

Our final name-based prediction system comprised 12 variables: 5 from first names and 7 from last names. Combined with DeepFace's 6 mugshot-derived variables, our complete model incorporated 18 predictive features for racial classification.

3 Method

Our goal is to identify systematic bias in racial assignment by corrections authorities. We train predictive models on DOC-assigned race labels, then interpret systematic deviations between model predictions and official assignments as evidence of mislabelling rather than model error.

This approach requires that our model learns true race rather than replicating the bias present in DOC labels. A model trained on biased labels will still learn the underlying signal (true race) rather than the bias itself when two conditions hold: (1) the predictors capture genuine racial differences (facial features, name demographics), and (2) the model achieves sufficiently high accuracy that it fits the dominant signal in the data rather than noise. Linear models are particularly well-suited for this task because they fit overall linear relationships between predictors and outcomes, effectively averaging through systematic bias patterns rather than overfitting to them. Under these conditions, systematic deviations between model predictions and official classifications indicate bias in the original labels.

Given this theoretical framework, our modelling approach prioritized maximizing multi-class classification accuracy. We recognized that including all racial categories might compromise overall accuracy due to insufficient representation or inherent classification difficulty with certain groups. We adopted a systematic approach to calibrate which racial groups to include, using multinomial logistic regression across three progressively focused classification schemes:

1. *Classification of all races, including the "unknown" category*

2. *Classification of all races while excluding the "unknown" category*

3. *Classification restricted to the three primary racial groups with sufficient representation: White, Black, and Hispanic*

[8] Implicitly, it is included in the Harvard Dataverse 'middle names' but the choice was made to separate these, as suffixes exist in addition to middle names

This tiered approach allowed us to empirically determine the optimal balance between classification accuracy and racial group coverage, ensuring our bias detection method operates under favourable conditions for accurate race prediction.

Additionally, models must be optimized for multi-class classification accuracy. Without this optimization, models fit to maximize simple accuracy, which automatically defaults towards predicting classes with the greatest proportion in the criminal dataset (Whites, Blacks) and away from less represented races (Hispanic, Native American). To correct for this, we used inverse weighting to the sample size of each race.

We focused bias detection on multinomial logistic regression (MLR) rather than non-linear models. Linear models fit overall linear relationships between predictors and outcomes, effectively averaging through systematic bias patterns rather than overfitting to them. We also tested XGBoost as an alternative modelling approach to assess whether non-linear patterns might improve predictive accuracy.[9] XGBoost improved BWH classification accuracy by only 0.28% (from 92.76% to 93.04%), confirming that MLR captured the essential relationships in the data. Given this minimal improvement and our theoretical preference for linear approaches in bias detection, we proceeded with MLR for all subsequent analyses (see Appendix for full XGBoost implementation details and comparative results).

3.1 Separability and dimensionality reduction

A critical methodological concern involves disentangling genuine bias from natural classification difficulties arising from differential phenotypic distinctness between racial groups. Certain racial categories exhibit closer genetic ancestry, physical appearance, and naming patterns than others; most notably, the relatively modest differences between Hispanics and Whites compared to the more pronounced distinctions separating these groups from Blacks.

This natural variation in inter-group distinctness generates a testable prediction: If misclassification patterns simply reflect inherent classification difficulty rather than systematic bias, we should observe higher accuracy rates for more phenotypically distinct groups (such as Blacks) and proportionally higher error rates for phenotypically similar groups (Hispanics and Whites). Under this scenario, elevated misclassification rates between similar groups could be attributed to reduced Euclidean distance in feature space rather than systematic bias.

This analysis is complemented by classification models, as they are fundamentally designed to maximize separability between groups, trained explicitly to find the boundaries that best distinguish categories even when those boundaries are subtle. If our model achieves high overall accuracy, it demonstrates that the racial groups are sufficiently separable in our feature space. Persistent misclassification of a specific pairing at high confidence would then suggest something other than mere phenotypic similarity—it would indicate cases where the model has detected strong, convergent evidence across multiple predictive features (mugshot analysis, name demographics) pointing toward one classification while official records show another.

When such high-confidence predictions contradict official classifications at rates exceeding random error, this indicates consistent patterns of misclassification rather than model uncertainty. We explore the relationship between model confidence and racial classification to quantify whether these patterns reflect random errors or systematic bias.

3.2 Simulations

To validate our methodological assumptions and develop a framework for interpreting real-world bias patterns, we constructed controlled simulation studies using synthetic datasets with known bias characteristics. These simulations enable us to test whether our analytical approach can successfully detect and characterize different types of systematic bias under controlled conditions. Our simulation framework employed a simplified three-group structure (Red, Blue, Green) designed to mirror the essential characteristics of our

[9] XGBoost was considered under the assumptions that bias was mostly linear and that misclassifications represent ground truth. The difference in accuracy between XGBoost and MLR illuminates the degree of non-linearity present in the data.

real-world three-race analysis while maintaining interpretive clarity. This design facilitates direct comparison between simulated and observed bias patterns.

We constructed our synthetic data within a two-dimensional feature space. Because our primary methodology relies on linear modelling techniques that do not assume complex nonlinear interactions between variables, higher-dimensional representations would introduce unnecessary complexity without enhancing our ability to detect the linear bias patterns central to our analysis. The two-dimensional approach also enables clear visualization of bias patterns and model performance. Within this simplified space, we positioned each group with equal separation distances of three standard deviations between all group centroids, ensuring balanced distinctness across categories. We then introduced three distinct bias types affecting only the Blue and Green groups, leaving Red as an unbiased control:

- *Random bias: Greens randomly assigned to Blue..*
- *Strategic bias: Greens closest to the Blue mean assigned to Blue.*
- *Obvious bias: Greens furthest from the Blue mean assigned to Blue.*

For each bias scenario, we reassigned 10% of Greens to Blue classification, with the selection mechanism varying according to the specific bias type being simulated. The mathematical details are in the appendix.

3.3 State analysis

To test whether any detected bias reflects deliberate discrimination versus random administrative error, we examined correlations between state-level misclassification rates and political ideology, measured through Republican vote share in recent elections. Systematic correlations with political variables would suggest intentional bias, while their absence would support alternative explanations such as administrative inconsistency or measurement error.

We also investigated whether genetic ancestry composition influences misclassification patterns. Using ancestry data from (Bryc et al., 2015), we tested whether misclassification rates correlate with population-level genetic distinctness between racial groups within states. This analysis helps distinguish between bias arising from genuine classification difficulty due to genetic similarity versus systematic administrative bias.

4 Results

4.1 Model results

We implemented our systematic modelling approach across three progressively focused classification schemes, each employing eight distinct feature combinations to identify the optimal predictor set for racial classification. Our results demonstrate clear patterns in agreement that guided our modelling strategy.

4.2 Classification performance summary

Table 1 summarizes the performance of the best model across the three classification approaches.

4.3 Classification Including "Unknown or Other" Category

Our initial approach included all seven racial categories present in the criminal database. Across eight model specifications, accuracy ranged from 38.73% to 83.68%, with the naive baseline (predicting all individuals as White) achieving 52.84% accuracy. The mugshot-derived race variables alone achieved 74.53% accuracy.

Table 1: Model performance across classification schemes

Classification scheme	Categories	Naive baseline	Best model accuracy	Improvement
All races + Unknown[a]	7	52.84%	83.68%	+30.84%
All races (clean)	5	54.46%	84.14%	+29.68%
Black-White-Hispanic	3	55.70%	92.76%	+37.06%

[a] Best model for unknown + other was the use of sex in addition to all features. Including unknown + other seemed to break the usual best model as MLR defaulted to the catch-all of predicting nearly all Whites as Unknown.

Table 2: Feature set performance comparison (five-race vs BWH classification)

Feature set	Five-race accuracy	BWH accuracy	Improvement
Mugshot-derived only	77.17%	81.76%	+4.59%
Last name features	66.31%	68.92%	+2.61%
First name features	61.15%	68.18%	+7.03%
All name features (including middle, suffixes, etc.)	71.75%	75.67%	+3.92%
All features	84.14%	92.76%	+8.62%

Examining the confusion matrices revealed a critical limitation: The “Unknown or Other” category exhibited highly unpredictable classification patterns. The model consistently misallocated substantial portions of clearly identifiable racial groups into this ambiguous category, while simultaneously failing to reliably predict when individuals should legitimately be classified as “Unknown or Other”. This pattern suggested that the “Unknown or Other” designation functioned more as an administrative convenience than a meaningful racial category for bias detection purposes. We filtered this category out of our dataset.

4.4 Classification on All Races

Removing the “Unknown or Other” category and focusing on the five substantive racial groups yielded more interpretable results, though with a modest reduction in overall accuracy. The naive baseline increased to 54.46%, reflecting the higher proportion of White individuals in the cleaned dataset. Model performance ranged from 41.84% to 84.14%, with our comprehensive model achieving 84.14% accuracy. The mugshot-derived variables alone achieved 77.17% accuracy, while name-based predictors showed more variable performance (Table 2). The combination of all name-based predictors achieved 71.75% accuracy, 5.4 percentage points below the mugshot-derived approach.

Despite the improved interpretability, five-category classification accuracy remained at 84.14%, insufficient for reliable bias detection. Asian or Pacific Islander and Native American categories showed persistent misclassification patterns due to small sample sizes and phenotypic similarity to other groups that DeepFace’s six-category classification system cannot fully distinguish (Figure 3).

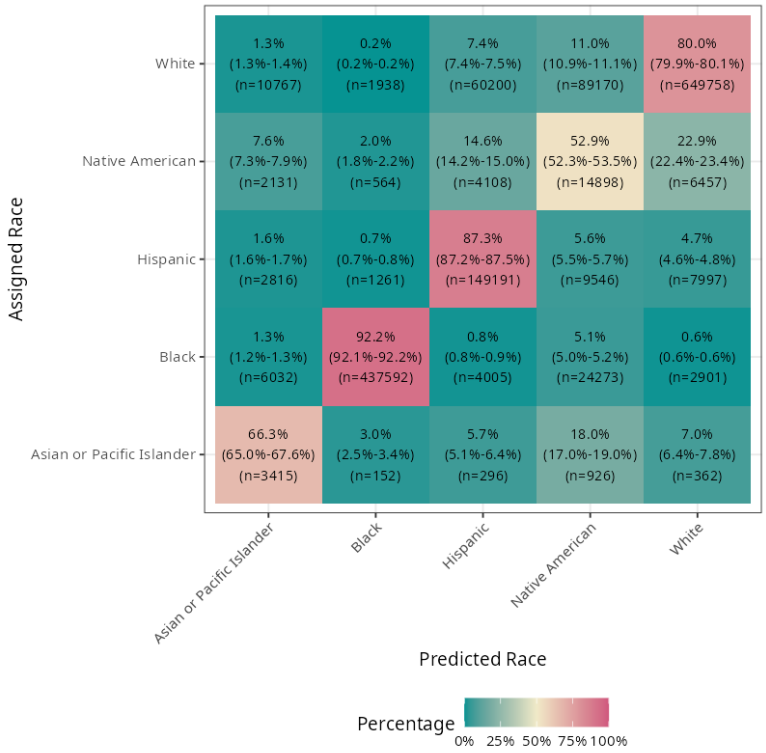

Figure 3: Confusion matrix, default (not inverted), for classifying all races

4.5 Three-race classification: Black, White, and Hispanic

Restricting our analysis to the three most populous and best-represented racial groups increased accuracy from 84.14% to 92.76%. The naive baseline increased to 55.7%. The confusion matrices for the three-race classification revealed much cleaner patterns. Figure 4 presents the confusion matrix for our optimal BWH model.

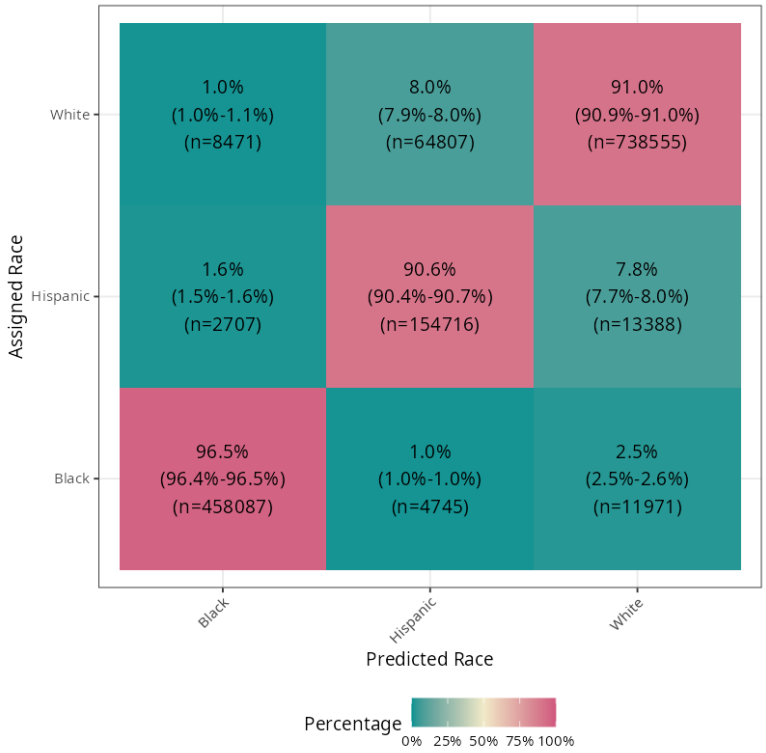

Figure 4: Confusion matrix - BWH classification (comprehensive model, 92.76% accuracy)

4.6 Model selection rationale

Based on these systematic results, we selected the multinomial logistic regression (MLR) model with three-race classification focusing on Black, White, and Hispanic individuals for our bias detection analysis. This decision was driven by three key considerations.

First, the “Unknown or Other” category proved fundamentally unpredictable and likely reflected administrative rather than demographic realities. Including this category introduced noise that compromised our ability to detect systematic bias patterns in our MLR models.

Second, while the five-race MLR classification was more comprehensive, the 84.14% accuracy level remained insufficient for reliable bias detection. Our theoretical framework requires near-maximal accuracy to distinguish between random error and systematic bias, a threshold not met by the more inclusive approach.

Third, the three-race MLR classification achieved 92.76% accuracy while covering approximately 89% of our total sample (excluding Unknown/Other cases and minority ethnicities). This combination provided the optimal foundation for bias detection, ensuring that apparent misclassifications likely reflected genuine discrepancies rather than accuracy limitations.

The 92.76% accuracy level of our MLR model represents a sufficiently high threshold to support our core assumption: that systematic deviations from predicted classifications indicate bias in the original racial assignments rather than random classification error. This accuracy level, combined with our class-weighted approach to prevent majority-class bias, establishes the methodological foundation for our subsequent bias detection analysis. While XGBoost achieved marginally higher overall accuracy (93.04%), we prioritized MLR for its theoretical alignment with our bias detection framework. Analysis of the confusion matrix produced by this model (flipped, consistent with our assumption) indicates that 28.9% of Hispanics were incorrectly assigned as White, as seen in Figure 5.

Table 3: Euclidean distance between racial centroids

Race comparison	Asian or Pacific Islander	Black	Hispanic	Native American
Asian or Pacific Islander		12.01	11.67	11.14
Black	12.01		5.45	4.17
Hispanic	11.67	5.45		4.34
Native American	11.14	4.17	4.34	
White	11.92	3.41	4.53	3.31

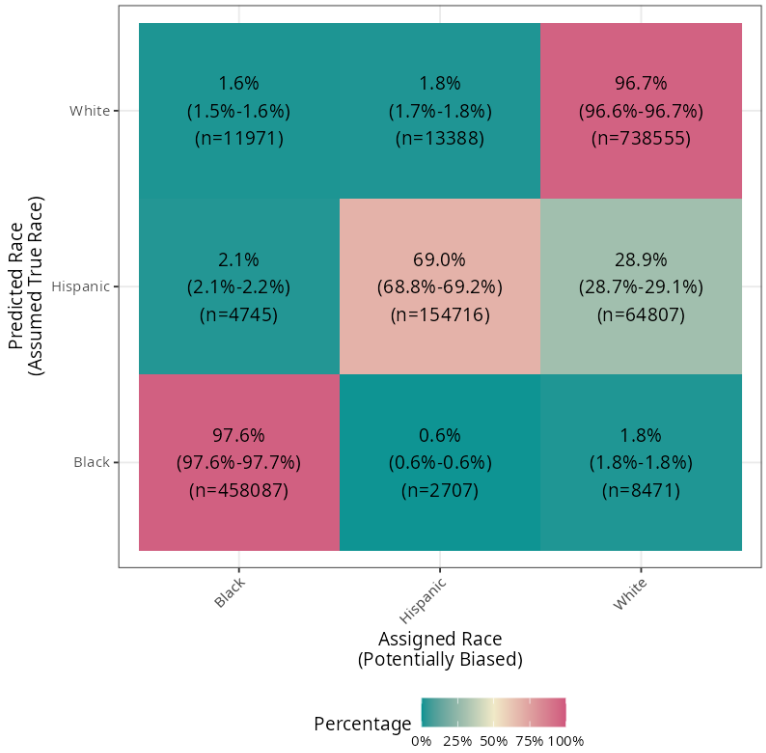

Figure 5: Inverted confusion matrix for MLR on Black-White-Hispanic classifications

4.7 Separability and dimensionality reduction

Across the first 16 PCs (16 was chosen as the threshold for 90% of the variance), the Euclidean distance between Whites and Blacks was smaller than the distance between Whites and Hispanics (Table 3). The centroids roughly conform to the PC1/PC2 map (Figure 6), with Native Americans closest to Whites, followed by Blacks, Hispanics, then Asians. On the PC1/PC2 map with just the three main races, as seen in Figure 7, individuals classified as White are distributed throughout regions associated with Hispanics.

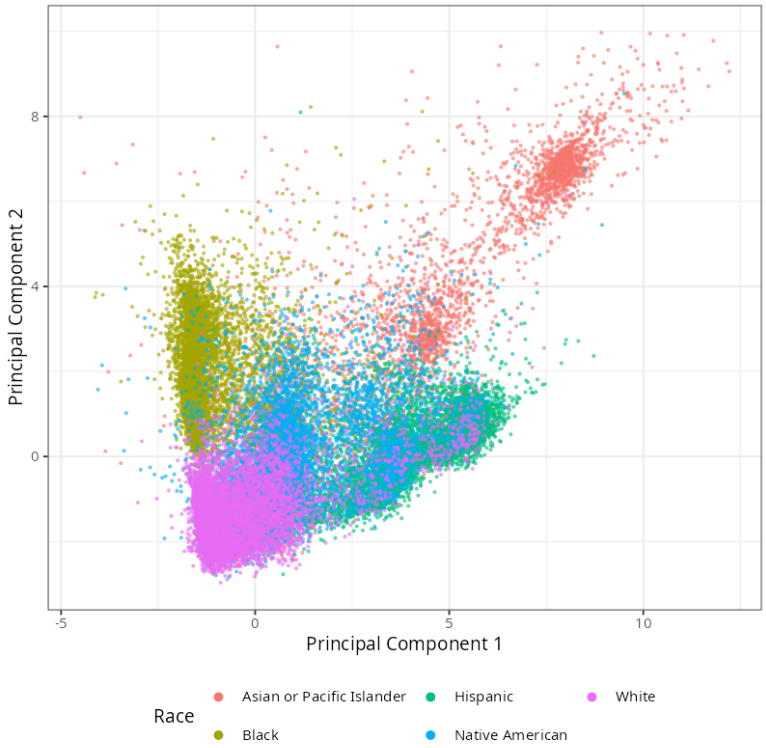

Figure 6: PC1, PC2 map of all races

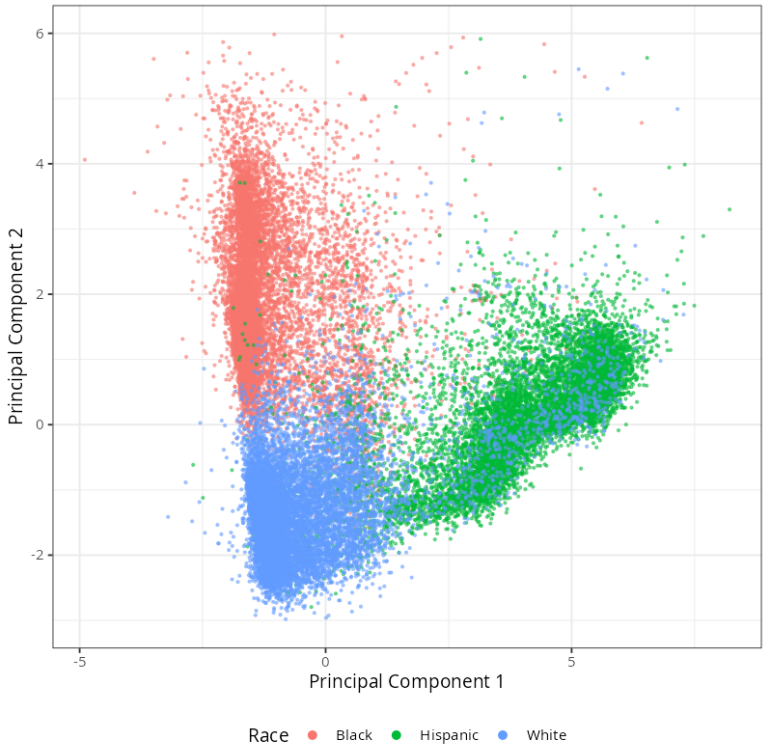

Figure 7: PC1, PC2 map of the three main races

Setting Hispanics to a density contour and vice versa, visual inspection shows a large number of Whites located in the PC1/PC2 region associated with Hispanics, as observed in Figure 8. The reverse

is true only for the tail. Crucially, the mix of Whites within the Hispanic distribution parallels the Hispanic distribution, whereas this pattern does not hold for Hispanics located in the White region.

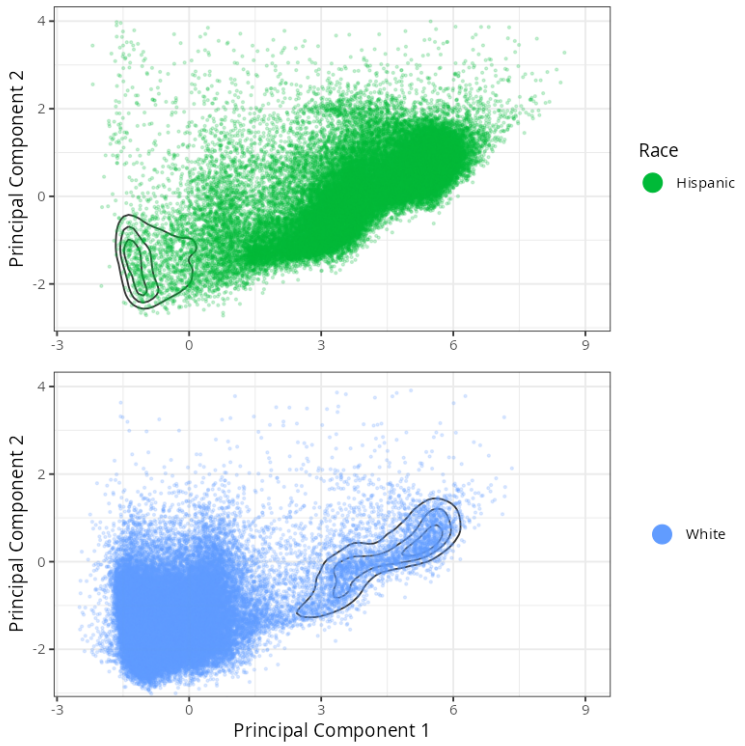

Figure 8: Density of Hispanics over Whites and vice versa

The main races show similar distances from each other (Table 3). This provides a foundation for further analysis, indicating the races are sufficiently separable for accurate classification and bias detection. Despite this separability, Whites are distributed throughout the Hispanic region in PC space. Using the supervised approach with the Black-White-Hispanic model, as observed in Figure 9, for both predicted Blacks and Whites, model confidence was proportional to model agreement. For these two races, as confidence approaches 100%, so does model agreement.

However, for those predicted to be Hispanic, this is not the case. Despite the multinomial model being 95-100% confident, 22.4% of these Hispanics (assumed Hispanics given our supposition of predictions as ground truth) are labelled as White. Furthermore, as indicated by Figure 10 and Table 4, the plurality of Hispanics classified by our model belong in the 90%+ confidence range. Almost all misclassification directions had a spike in misclassifications near the 90% range. The bias distribution for predicted Hispanics also indicated misclassification of Blacks as Hispanic.

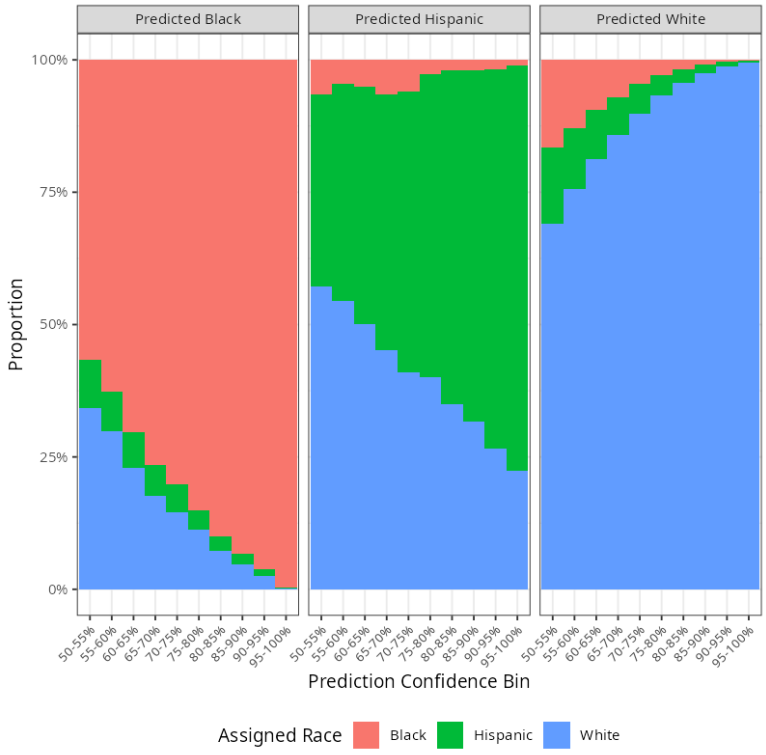

Figure 9: Breakdown of assigned race within predicted race groups by confidence bin. Bars show the proportion of actual 'race' within each predicted group for different prediction probability bins.

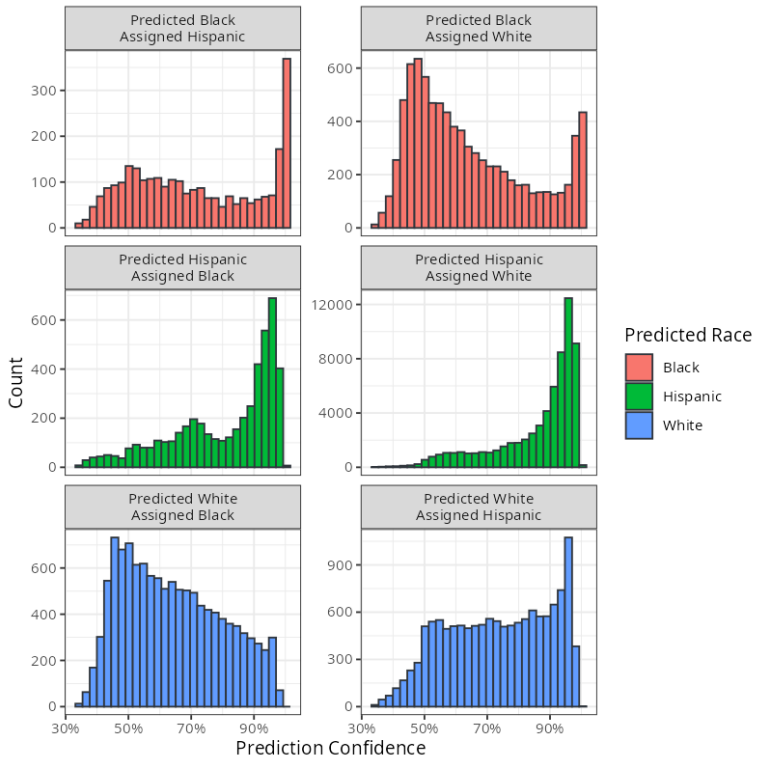

Figure 10: Each facet represents the predicted race and the count of its assigned races by confidence bin.

Table 4 presents confidence statistics for all misclassification types, showing the count, mean and median confidence levels for each predicted-assigned race combination. Hispanic predicted as White

Table 4: Misclassification confidence statistics

Predicted race	Assigned race	Count	Mean confidence	Median confidence	Q25	Q75
Hispanic	White	64,807	86.1%	91.7%	80.2%	95.9%
Hispanic	Black	4,745	80.7%	87.4%	69.4%	94.4%
White	Hispanic	13,388	73.5%	74.3%	59.7%	88.3%
Black	Hispanic	2,707	71.4%	69.1%	53.4%	92.6%
White	Black	11,971	64.2%	62.4%	50.6%	76.5%
Black	White	8,471	63.9%	58.9%	48.8%	76.2%

showed the highest median confidence (91.7%) among all misclassification types, with 64,807 cases. This evidence aligns with the principal component and separability analyses. A fraction of Hispanic individuals are often classified as White despite their distance and distinction from the main distribution of White persons. This finding suggests the existence of bias in this dataset.

4.8 Simulations

After applying multinomial logistic regression to the simulated biased datasets and comparing the results to the real-world dataset, the “random” bias appeared most similar, as seen in Figure 11 (see Appendix for other bias types). For this simulated set, it appears that, as is with the real-world dataset, model confidence is proportional to accuracy for all racial classifications. That is, except for where the Green-Blue bias was introduced, confidence is still commensurate with accuracy, but the randomness introduces a high ceiling equal to the prevalence of the bias.

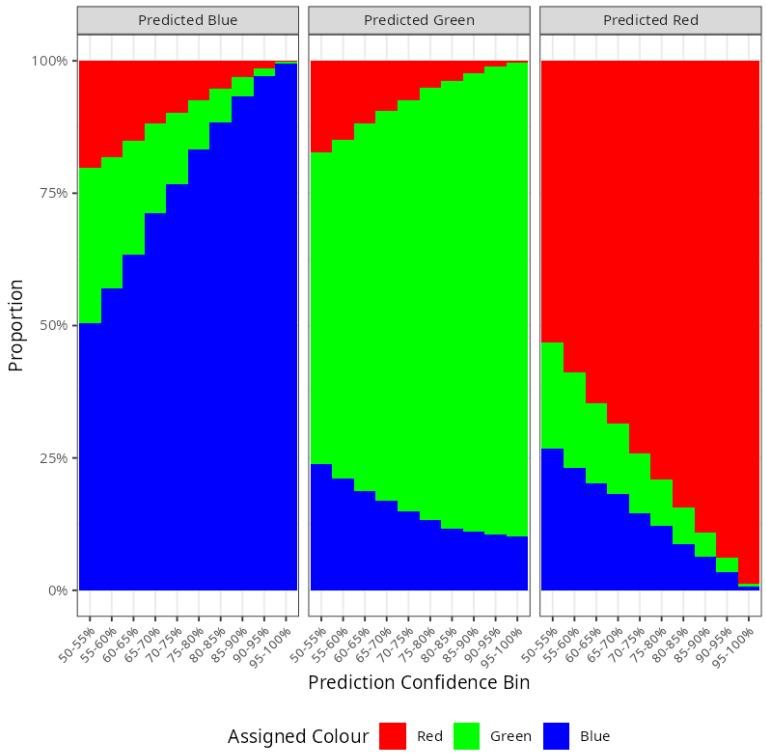

Figure 11: Random bias simulation classification outcome

4.9 State analysis

Misclassification of Hispanics as White varies greatly by state, with Florida showing exceptionally high rates (>60%). The correlation between Republican vote share and Hispanic bias was not statistically significant (r = .21, 95% CI: -.36 to .67, p = .473), indicating that this bias was random rather than deliberate and motivated by ideology. Breakdown by state can be seen in Figure 12.

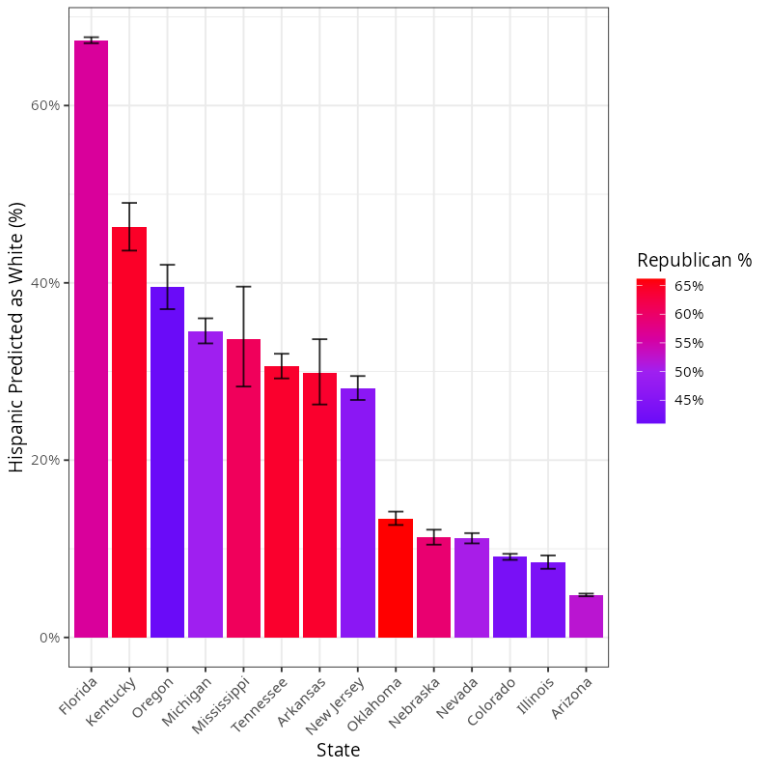

Figure 12: Proportion of Hispanics classified as White by state

We found a strong negative correlation (r = -.799, 95% CI: -.95 to -.38, p = .00315, n = 11) between Native American ancestry among Latinos and misclassification of Hispanics as White. The reverse pattern was also statistically significant: White-to-Hispanic misclassification showed a strong positive correlation with Native American ancestry among Latinos (r = .741, 95% CI: .26 to .93, p = .009). Both relationships are visualized in Figure 13.

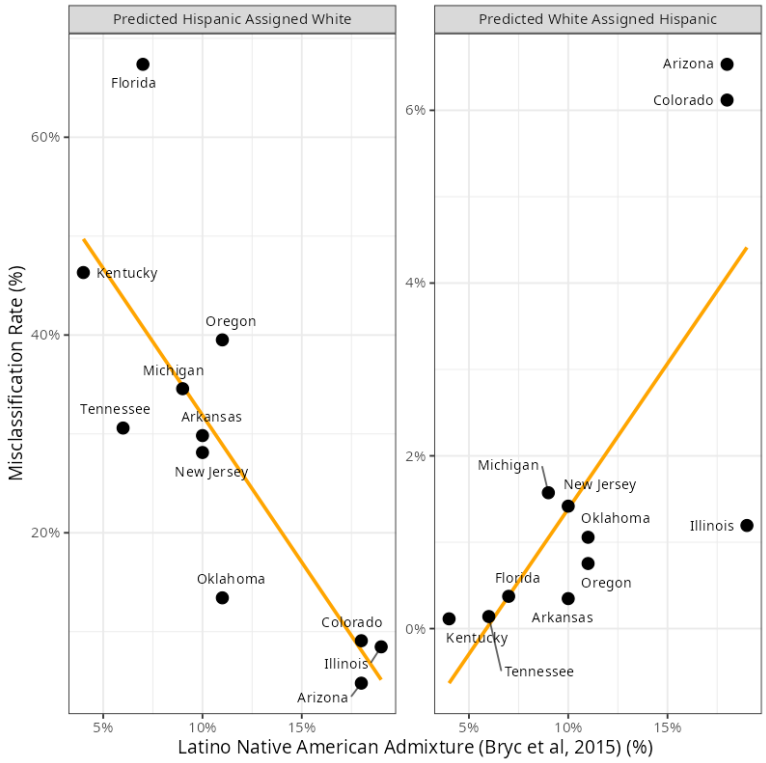

Figure 13: Relationship of Native American ancestry in Latinos with misclassification of Hispanics as White.

These findings suggest two competing interpretations. First, genetic similarity: States where Latinos have higher European ancestry show higher misclassification rates because Hispanics phenotypically resemble Europeans. Second, self-identification effects: Where Native American ancestry makes Hispanic identity more visible and distinctive, both Hispanics and classification personnel are more likely to recognize and assign Hispanic identity.

To test whether state-level self-identification norms explain the misclassification pattern, we added state as a variable to the MLR classification model. If self-identification drove the pattern, state should meaningfully improve model performance by capturing regional norms (e.g., Cubans in Florida identifying as White). However, adding state increased classification accuracy by only ~1%, and the mass of Hispanic misclassification remained unchanged. This suggests that state-level self-identification patterns do not substantially explain the observed misclassification.

4.10 Individual inspection

Following our analysis, we used Python to randomly and programmatically select and render mugshot collages to verify our approach. First, we inspected low confidence and high confidence predictions (Figures 14 and 15). Contrasting these, we found that model confidence was commensurate with the predicted race being the likelier true racial classification.

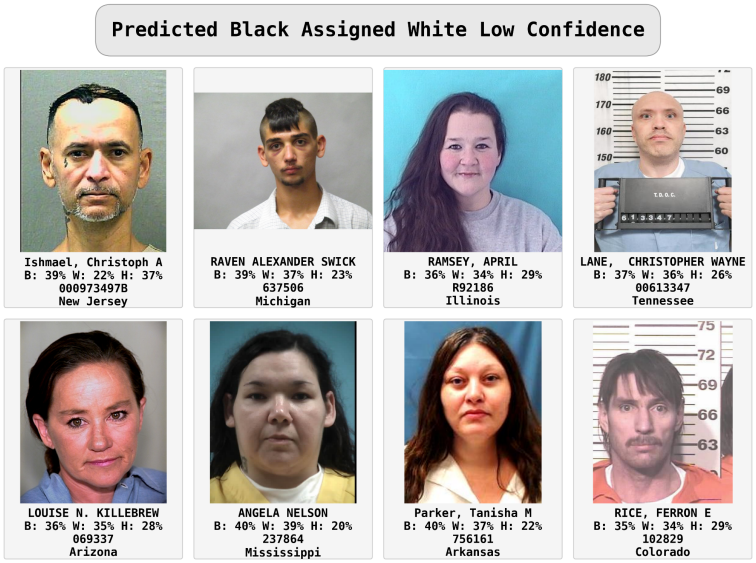

Figure 14: Low confidence classifications Black Assigned White

Figure 15: High confidence classifications Black Assigned White

We then generated mugshots for the strongest and weakest misclassification directions: predicted Hispanics assigned White and predicted Blacks assigned White respectively (confidence seen in Table 4). To ensure adequate representation of mugshots by confidence and state, eight examples from eight states were randomly selected above and below the median model confidence, yielding sixteen mugshots. Observing the two, it is apparent from the surnames and physical appearances in Figure 16 that the predicted Hispanics are indeed Hispanic rather than White, except for one or two low confidence cases, e.g. “Davila Autumn Dawn”. In contrast, Figure 17 shows that the discrepancy between predicted and assigned race results from model error, not racial misassignment, once again excluding the few high confidence cases, e.g “Mark Young”. This again shows the importance of model confidence.

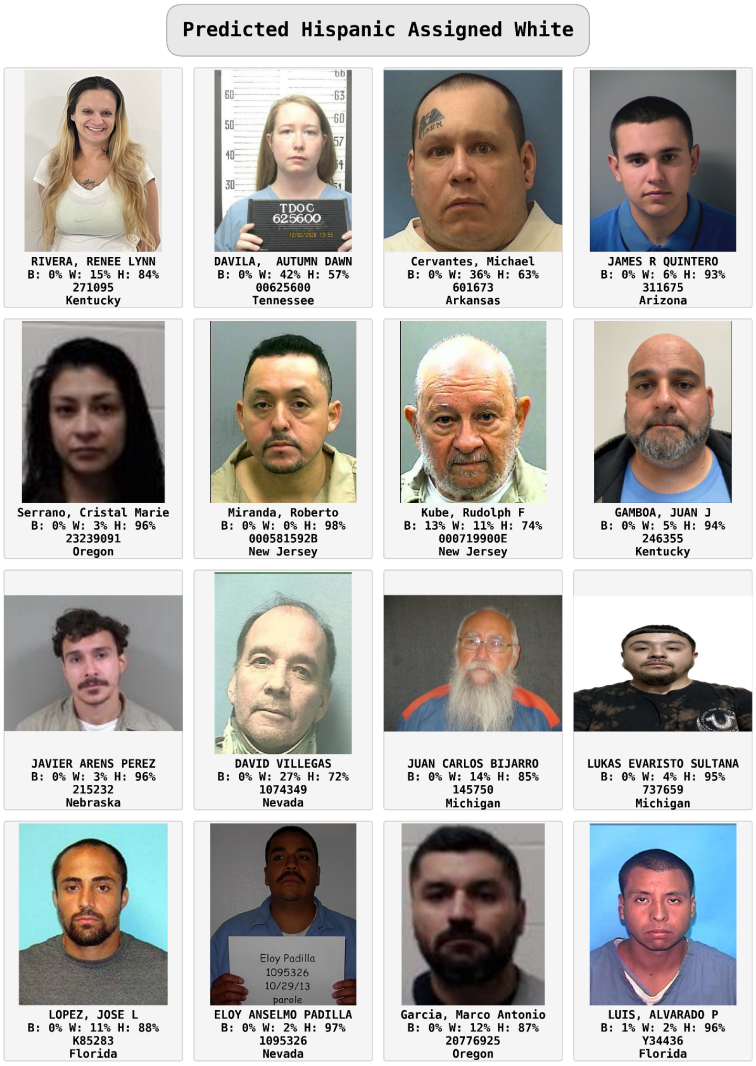

Figure 16: Predicted Hispanic assigned White

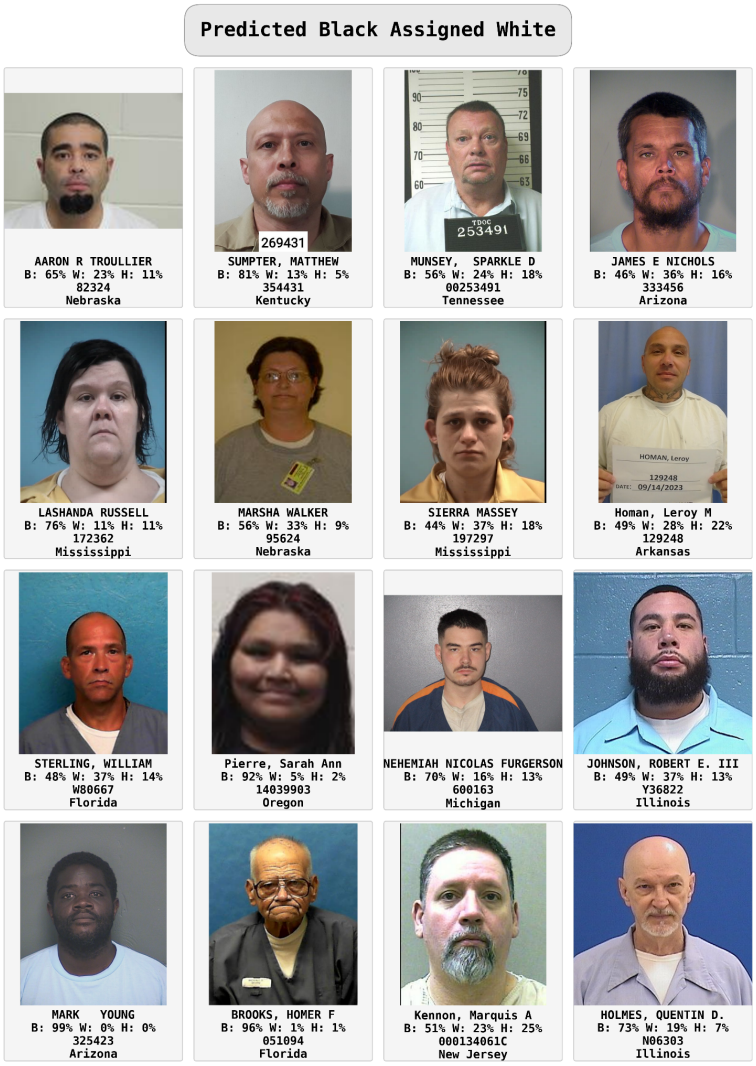

Figure 17: Predicted Black assigned White

4.11 Effects on crime rates

To quantify the impact of racial misclassification on reported criminal record rates, we corrected the assigned racial counts in each state's criminal database. We calculated DOC criminal records per 100,000 population using census data, normalizing all rates relative to the assigned White rate within each state. We applied two corrections:

First, a high-confidence correction. We assumed that the predicted race was the true race where the model classified with >90% confidence; for cases where model confidence was <90%, the assumed race was the assigned race.

Second, a more generous reclassification. We assumed that the predicted race reflected the true race rather than the assigned race for all cases.

In this second adjustment, correcting for misclassification increases Hispanic criminal record rates by 31% among the states analysed while decreasing Black rates by 1% and White rates by 6%. The high-confidence adjustment was more modest. Hispanic criminal record rates increased by 20%, Black criminal rates fell an indistinguishable 0.2%, and White rates fell by 4%. Figures 18 and 19 show the combined criminal record rates across all states. The effect varied considerably by state, as seen in Figure 19. These results demonstrate that racial misclassification in criminal databases artificially deflates Hispanic criminal record rates while mildly inflating White rates. The magnitude of the effect is substantial: The corrected Hispanic rate is 20% higher even in the conservative case.

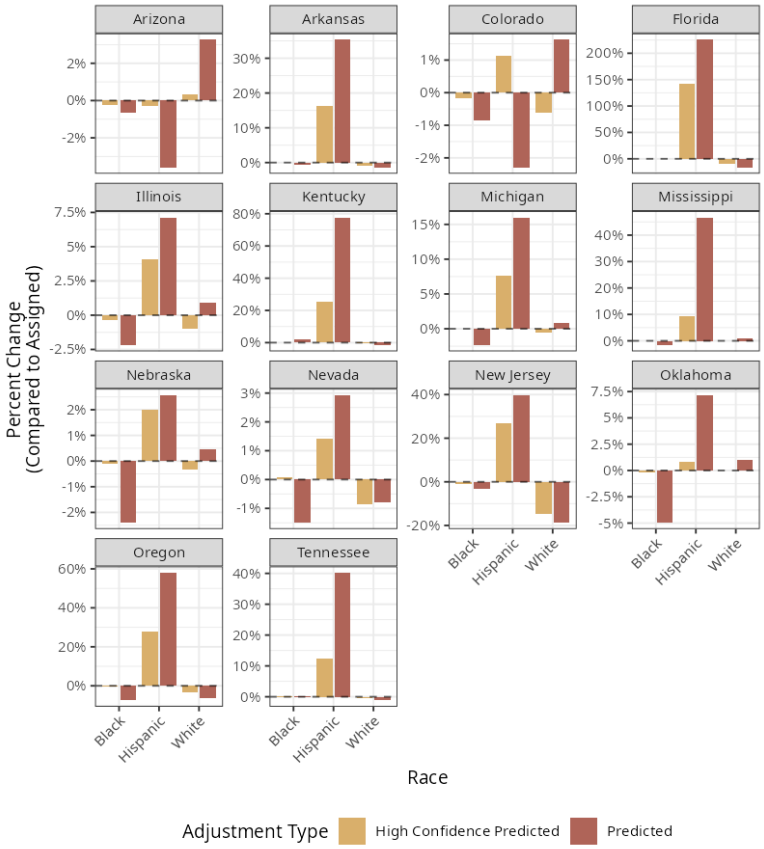

Figure 18: Adjustment by state and race

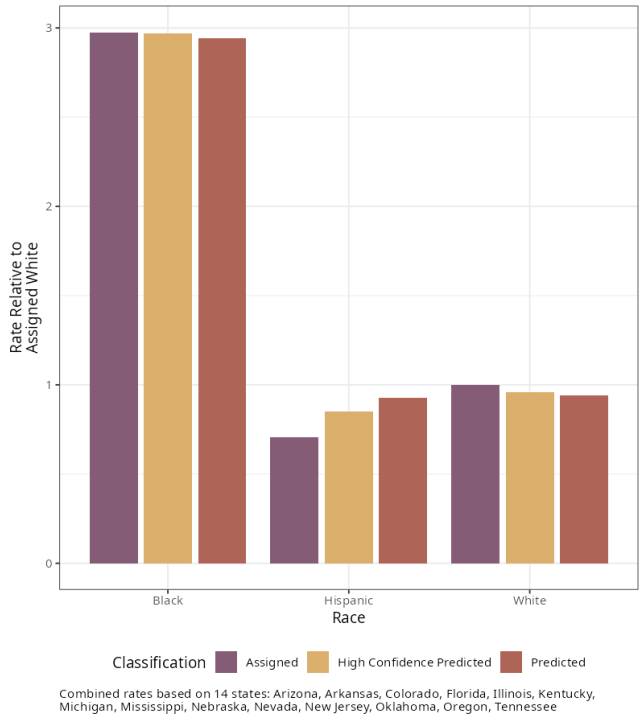

Figure 19: Percent change in criminal record rates for all states

5 Discussion

Our analysis identified systematic racial misclassification in U.S. Department of Corrections databases. Using a model with 92.76% agreement in three-race classification covering 89% of the total sample, we found that 28.9% of individuals predicted to be Hispanic were officially assigned as White. This pattern persisted among high-confidence predictions (95%+ model certainty), where 22.4% of predicted Hispanics received White classifications. Simulation studies indicated the pattern resembles random rather than deliberate bias.

5.1 Evidence for genuine misclassification

Several lines of evidence support genuine Hispanic-to-White misclassification rather than model error. Visual inspection of high-confidence cases showed individuals with Hispanic phenotypes and surnames despite White classifications. Principal component analysis revealed a separability paradox: Whites and Hispanics showed greater Euclidean distance (4.53) than Whites and Blacks (3.41) in PC space, yet Hispanics were misclassified as White at high rates while Blacks were not. While PCA separability differs from predictive separability, this pattern contradicts explanations based solely on phenotypic similarity. Additionally, Whites were distributed throughout the Hispanic region in PC space, but Hispanics were not distributed throughout the White region, suggesting that individuals classified as White within Hispanic PC space represent misclassified Hispanics.

Model confidence predicted agreement for Black and White classifications: As confidence approached 100%, agreement approached 100%. For Hispanic predictions, this relationship broke down. At 95-100% confidence, 22.4% of predicted Hispanics were still assigned as White. This pattern was asymmetric, occurring only for Hispanic-to-White misclassification at high rates (>20% in most states), not for other racial combinations. When we simulated datasets with known random label bias (randomly reassigning 10% of one group to another), models trained on the biased labels reproduced this exact pattern: normal confidence-agreement relationships for unbiased groups, but a high error ceiling equal to the bias prevalence for the biased group. The real-world data showed the same signature, indicating label bias rather than model limitations. This pattern was not an artifact of class imbalance because we used inverse frequency weighting to ensure the model optimized for balanced performance across all racial groups.

5.2 Nature of the misclassification

Several findings indicate the misclassification pattern reflects administrative inconsistency rather than deliberate bias. No correlation existed between state-level Republican vote share and Hispanic misclassification rates (r = .21, 95% CI: -0.36 to 0.67, p = .473). XGBoost improved classification accuracy by only 0.28% over multinomial logistic regression (93.04% vs 92.76%), indicating the racial classification relationships are fundamentally linear. If deliberate bias existed, we would expect complex non-linear decision rules that XGBoost would capture. Simulation studies showed the observed pattern most closely resembled random rather than strategic or obvious bias.

5.2 Nature of the misclassification

Several findings indicate the misclassification pattern reflects administrative inconsistency rather than deliberate bias. No correlation existed between state-level Republican vote share and Hispanic misclassification rates (r = .21, 95% CI: -0.36 to 0.67, p = .473). XGBoost improved classification accuracy by only 0.28% over multinomial logistic regression (93.04% vs 92.76%), indicating the racial classification relationships are fundamentally linear. If deliberate bias existed, we would expect complex non-linear decision rules that XGBoost would capture. Simulation studies showed the observed pattern most closely resembled random rather than strategic or obvious bias.

5.3 Mechanisms of misclassification

The correlation between Native American ancestry among Latinos and misclassification rates admits two competing interpretations. The strong negative correlation (r = -0.799, 95% CI: -0.95 to -0.38, p = 0.003) between Native American ancestry and Hispanic-to-White misclassification, coupled with the positive correlation for White-to-Hispanic misclassification (r = 0.741, 95% CI: 0.26 to 0.93, p = 0.009), creates an interpretive challenge.

One interpretation centres on phenotypic similarity. States where Latinos have higher European ancestry show higher misclassification rates because Hispanics phenotypically resemble Europeans. Given substantial European ancestry (averaging 75.1% across analysed states) and genetic recombination, many Latinos phenotypically resemble Europeans despite Hispanic surnames. The bidirectional correlation pattern supports this: More Native American ancestry makes Hispanics more phenotypically distinct, reducing misclassification in both directions.[10]

An alternative interpretation emphasizes self-identification effects. As demonstrated by Florida's exceptionally high misclassification rate (>60%), self-identification plays a role. Latino racial self-identification varies substantially by national origin: Approximately 91% of Cuban Americans self-identify as White compared to 56% of Puerto Ricans and 49% of Mexicans (Figuereo et al., 2025; Michael & Timberlake, 2007). Under this interpretation, where Native American ancestry makes Hispanic identity more visible and distinctive, group effects emerge. When Hispanic identity becomes more salient, both individuals and classification personnel are more likely to recognize and assign it.

However, our model already incorporates phenotypic distinctiveness through DeepFace's facial classification, and darker skin among Latinos robustly predicts Hispanic identification (Michael & Timberlake, 2007). If self-identification norms drove state-level variation, adding state as a model variable should improve classification agreement. It did not. Agreement increased only from 92.76% to 93.3% and Hispanic misclassification remained unchanged. This suggests phenotypic similarity, rather than self-identification norms, primarily drives the observed pattern, though data quality limitations prevent definitive conclusions.

Administrative inconsistency also contributes. The U.S. government treats Hispanic as an ethnicity rather than a race, and nine states meeting our data quality requirements did not classify Hispanic as a distinct category, presumably assigning these individuals as White. States that do distinguish Hispanics apply inconsistent criteria, with classification personnel treating Hispanic and White as interchangeable categories in ambiguous cases

5.4 Implications

Correcting for racial misclassification increases Hispanic criminal record rates by 31% nationally while decreasing Black rates by 1% and White rates by 6%. A more conservative high-confidence adjustment (using predicted race only where model confidence exceeded 90%) produced more modest effects: Hispanic rates increased by 20%, Black rates decreased by 0.2%, and White rates decreased by 4%. Visual inspection confirmed that many misclassified individuals display Hispanic phenotypes and surnames, indicating the misclassification occurs despite observable Hispanic characteristics. This distorts published estimates of racial differences in criminal justice involvement, systematically understating Hispanic representation and overstating White representation in criminal databases.

Data and code availability

All data and code required to replicate this analysis are publicly available. The complete dataset (1.5 million criminal records with facial recognition features, name demographics, and official classifications) and replication code are available at:

[10] 23andMe users skew toward higher socioeconomic status and European ancestry, so these figures likely overestimate European ancestry relative to actual state Latino populations.

GitHub Repository: https://github.com/uncorrelated1/Detecting-Systematic-Bias-in-Criminal-Racial-Assignment

OSF Repository: https://osf.io/7fqkp

The GitHub repository contains all analysis scripts, replication instructions, and documentation. The OSF repository hosts the data files required for replication. All code is released under the MIT License.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. https://doi.org/10.48550/ARXIV.1907.10902

Beaver, K. M., DeLisi, M., Wright, J. P., Boutwell, B. B., Barnes, J. C., & Vaughn, M. G. (2013). No evidence of racial discrimination in criminal justice processing: Results from the National Longitudinal Study of Adolescent Health. *Personality and Individual Differences*, *55*(1), 29–34. https://doi.org/10.1016/j.paid.2013.01.020

Beck, A. J. (2021). Race and ethnicity of violent crime offenders and arrestees, 2018 [NCJ 255969, U.S. Department of Justice Statistics].

Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D., & Mountain, J. L. (2015). The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *The American Journal of Human Genetics*, *96*(1), 37–53. https://doi.org/10.1016/j.ajhg.2014.11.010

Cesario, J., Johnson, D. J., & Terrill, W. (2019). Is there evidence of racial disparity in police use of deadly force? analyses of officer-involved fatal shootings in 2015–2016. *Social Psychological and Personality Science*, *10*(5), 586–595. https://doi.org/10.1177/1948550618775108

D'Alessio, S. J., & Stolzenberg, L. (2003). Race and the probability of arrest. *Social Forces*, *81*(4), 1381–1397. https://doi.org/10.1353/sof.2003.0051

Figuereo, V., Rosales, R., Takeuchi, D. T., & Calvo, R. (2025). What race am i?: Factors associated with racial self-classification among U.S. Latinx adults. *Du Bois Review: Social Science Research on Race*, *22*(2), 215–236. https://doi.org/10.1017/s1742058x25000050

Hoekstra, M., Oh, S., & Tangvatcharapong, M. (2023). Are American juries racially discriminatory? evidence from over a quarter million felony grand jury cases. *Working Paper*.

James, L., James, S. M., & Vila, B. J. (2016). The reverse racism effect: Are cops more hesitant to shoot Black than White suspects? *Criminology & Public Policy*, *15*(2), 457–479. https://doi.org/10.1111/1745-9133.12187

Lange, J. E., Johnson, M. B., & Voas, R. B. (2005). Testing the racial profiling hypothesis for seemingly disparate traffic stops on the New Jersey Turnpike. *Justice Quarterly*, *22*(2), 193–223. https://doi.org/10.1080/07418820500088952

Michael, J., & Timberlake, J. M. (2007). Are Latinos becoming White? determinants of Latinos' racial self-identification in the United States. *Social Forces*, *86*(2).

Robertson, C., Baughman, S. B., & Wright, M. S. (2019). Race and class: A randomized experiment with prosecutors. *Journal of Empirical Legal Studies*, *16*(4), 807–847. https://doi.org/10.1111/jels.12235

Rosenman, E. T. R., Olivella, S., & Imai, K. (2023). Race and ethnicity data for first, middle, and surnames. *Scientific Data*, *10*(1), 299. https://doi.org/10.1038/s41597-023-02202-2

Rubenstein, E. S. (2016). *Race, Crime, and Justice in America*. New Century Foundation.

Schwartz, J. A., & Beaver, K. M. (2019). A longitudinal examination of the association between intelligence and rearrest using a latent trait–state–occasion modeling approach in a sample of previously adjudicated youth. *Developmental Psychology*, *55*(12), 2678–2691. https://doi.org/10.1037/dev0000838

Serengil, S., şefik, & Özpınaŗ, A. (2024). A benchmark of facial recognition pipelines and co-usability performances of modules. *Bilişim Teknolojileri Dergisi*, *17*(2), 95–107. https://doi.org/10.17671/gazibtd.1399077

Shjarback, J. A., & Nix, J. (2020). Considering violence against police by citizen race/ethnicity to contextualize representation in officer-involved shootings. *Journal of Criminal Justice*, *66*, 101653. https://doi.org/10.1016/j.jcrimjus.2019.101653

thelawofaverages (2023). Do minorities get longer sentences? an analysis of every U.S. state. https://thelawofaveragesblog.wordpress.com

US Census Bureau (2010). Frequently occurring surnames from the 2010 census. Retrieved December 18, 2025. https://www.census.gov/topics/population/genealogy/data/2010_surnames.html

Appendix

Figures A1, A2, and A3 show the MLR classification outcomes for the three bias simulation types. Figure A4 displays the original and biased datasets.

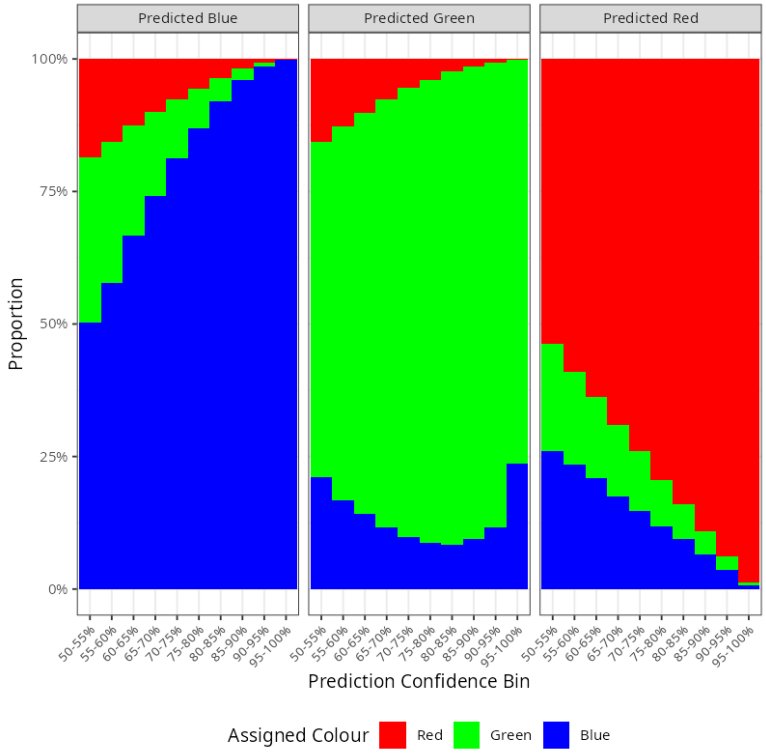

Figure A1: Obvious bias simulation MLR results by prediction confidence bin

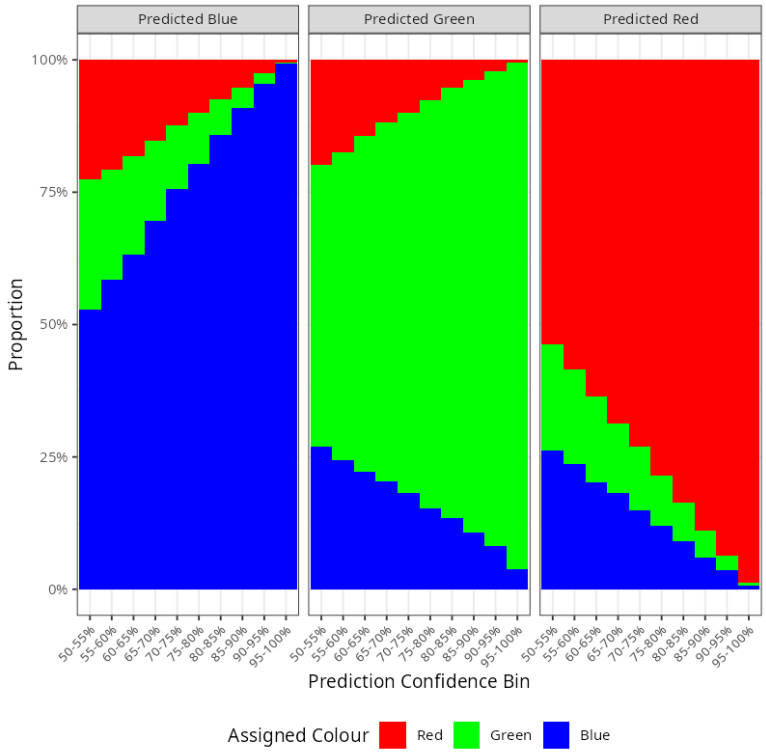

Figure A2: Strategic bias simulation MLR results by prediction confidence bin

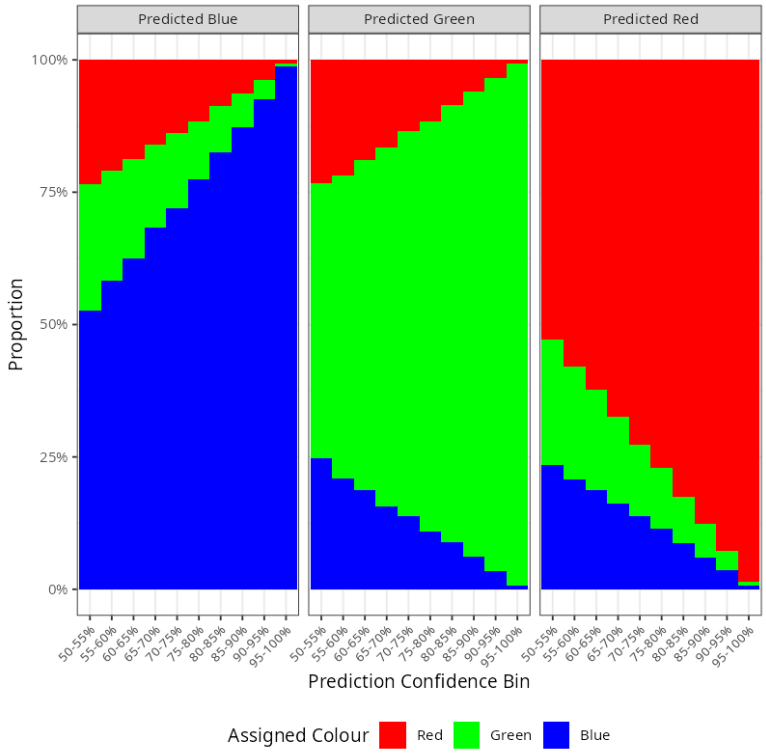

Figure A3: Original simulated dataset, MLR results by prediction confidence bin

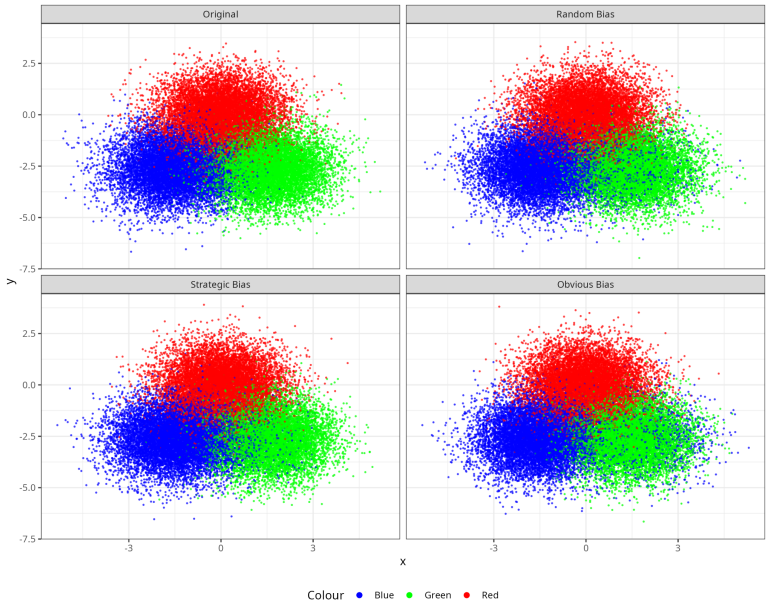

Figure A4: Maps of the four simulated bias variants

XGBoost implementation and results

We implemented XGBoost models for both the complete racial classification task (all five races) and the focused three-race classification (Black, White, Hispanic).

For model development, we employed a three-way data split strategy. We used an initial training set of 80%, with 10% reserved for validation and early stopping, and the final 10% for testing. This approach enabled hyperparameter optimization while preserving an unbiased evaluation framework.

Hyperparameter optimization was conducted using Optuna (Akiba et al., 2019), a Bayesian optimization framework employing the Tree-structured Parzen Estimator (TPE) sampler. We explored key parameters including regularization terms (L1 and L2), learning rate, maximum tree depth, minimum child weight, gamma, column and row subsampling ratios, and tree growth policy. Each optimization run consisted of 25 trials, with balanced accuracy serving as the objective function to ensure optimal performance across all racial categories rather than defaulting to majority class predictions.

Following hyperparameter optimization, final models were trained on the combined training and validation sets using the optimal parameters, with performance evaluated on the held-out test set.

For generating predictions across the entire dataset, we implemented 20-fold cross-validation using the optimized hyperparameters. This approach ensured that every individual received predictions from a model that had not encountered their data during training, providing out-of-sample predictions essential for bias detection. Within each fold, we maintained the same training-validation split strategy (90%-10%) with early stopping to preserve model quality across all iterations.

For the five-race classification, XGBoost achieved an improvement in overall accuracy (+1.44%) but losses in balanced accuracy (-2.74%) and macro F1 score (-12.21%). This indicates that while XGBoost slightly improved overall classification, it offered limited improvement for minority racial groups (Table A1).

In the BWH classification, XGBoost provided only small gains in overall accuracy (+0.28%) and balanced accuracy (+0.97%) while performing slightly worse on macro F1 score (-1.68%). This suggests that for the three-race problem, the MLR model already captured the essential relationships in the data, with non-linear approaches offering little additional benefit.

The class-specific performance metrics revealed similar patterns across both models. XGBoost achieved high precision for Black individuals (98%) with strong recall (97%). For Hispanic individuals, XGBoost achieved moderate precision (69%) but strong recall (93%). White classification showed high precision (98%) with good recall (91%).

Table A1: MLR vs XGBoost performance comparison

Model	Classification scheme	Accuracy	Balanced accuracy	Macro F1
MLR	All races (clean)	84.14%	81.27%	74.21%
XGBoost	All races (clean)	85.58%	78.53%	62.00%
MLR	Black-White-Hispanic	92.76%	92.60%	91.68%
XGBoost	Black-White-Hispanic	93.06%	93.57%	90.00%

Given our theoretical framework requiring that bias detection models fit underlying patterns rather than non-linear bias patterns, and the minimal accuracy improvements with XGBoost, we selected the more interpretable and theoretically sound MLR approach for our subsequent bias detection analysis.

Mathematical formulation of bias

For each bias scenario, we systematically reassigned 10% of Greens to Blue classification, with the selection mechanism varying according to the specific bias type being simulated. Let the Blue centroid be represented by coordinates $(c_x,\ c_y)$ in our two-dimensional feature space. For each Green point g_i with coordinates $(x_i,\ y_i)$, we calculated the Euclidean distance to the Blue centroid:

$$d_i = \sqrt{(x_i - c_x)^2 + (y_i - c_y)^2}$$

The bias assignment probabilities were then defined using exponential functions that create distinct selection patterns for Strategic and Obvious bias types:

$$P_{\text{strategic}}(i) = \exp(-d_i)$$

$$P_{\text{obvious}}(i) = \exp(d_i)$$

These formulations ensure that Strategic bias preferentially selects Green individuals closest to the Blue centroid (higher probability for smaller distances), while Obvious bias preferentially selects those most distant from the Blue centroid (higher probability for larger distances). We then used weighted sampling on these probabilities (higher values more likely to be sampled) for each respective scenario to produce our end-product simulated datasets:

- Strategic bias: weights proportional to $P_{\text{strategic}}(i) = \exp(-d_i)$
- Obvious bias: weights proportional to $P_{\text{obvious}}(i) = \exp(d_i)$
- Random bias: uniform weights (weights = 1 for all individuals)

Model training and evaluation on simulations

Following bias introduction, we trained multinomial logistic regression models on each simulated dataset using the simple formula race ~ x + y, where x and y represent the two-dimensional coordinates. This straightforward approach mirrors our linear modeling strategy for the real-world data while maintaining interpretive clarity. To address class imbalances created by the reassignment process, we implemented inverse frequency weighting:

$$w_j = \frac{N_{\text{total}}}{3 \times N_j}$$

Where w_j represents the weight for class j, N_{total} is the total sample size, and N_j is the number of observations in class j after bias introduction. This weighting scheme ensures that our models optimize

for balanced performance across all three groups. This was done to correct for the imbalance produced by reassignment, and for consistency with the method used on the real dataset. The simulation process generates four distinct datasets: the original unbiased dataset plus three variants incorporating Random, Strategic, and Obvious bias patterns, respectively. Figure 4A visualizes these four scenarios, illustrating how each bias type creates characteristic distortions in the group assignment patterns.