Not so Black and White: An Algorithmic Approach to Detecting Colorism in Criminal Sentencing

By Amy Wickett *

It is well documented that individuals of different races have disparate experiences in the criminal justice system in the United States. As most studies focus on Black-white interracial differences, intraracial disparities are often overlooked. Analyzing intraracial disparities is further complicated by the fact that it has been historically difficult to accurately and consistently measure skin tone and Afrocentric features. Utilizing convolutional neural networks and photos as data, this study creates a consistent, running measure of perceived race. Using this new measure and data type, new types of analysis are possible. This study will present photographic summary statistics, show the positive relationship between perceived race and sentence length is robust to the inclusion of various controls and, reevaluate traditional Black-white gaps, measuring both inter- and intraracial disparities.

I. Introduction

Researchers have empirically documented bias against minorities, typically African Americans, at all levels of the criminal justice system (Policing: Anbarci and Lee (2014); Goncalves and Mello (2021); West (2018); Bail Provision: Arnold, Dobbie and Yang (2018); Kleinberg et al. (2017); Sentencing and Prosecutors: Sloan (2019); Tuttle (2019)). The economics literature, and most social sciences, typically focus on discrete racial categories when studying discrimination and disparate outcomes. However, discrete categories are not able to capture how discrimination varies within race. Given the one-drop rule and the historical creation of Black as a race in America, individuals with a wide array of lineages are considered Black. Reece (2016) finds that most individuals today who are mixed race will identify as Black, prompting wide variation in physical appearances of those in the African American community. As a result, assuming homogeneity within the racial identity of Black is likely a faulty assumption. By depending on binary classifications, we thus may be overstating discrimination for some members of the African American community and understating it for others.

Social scientists have shown that colorism, which is bias within group that favors light skinned individuals over dark skinned individuals of the same race, impacts the level of discrimination or institutionalized hardship one will face (Frazier (1957); Keith and Herring (1991); Monk Jr (2015)). Previous scholars

* Harvard University, awickett@fas.harvard.edu

have shown that lighter skin individuals have higher wages (Goldsmith, Hamilton and Darity Jr (2006)), experience better health outcomes (Monk Jr (2015)), and are less likely to receive the death penalty (Eberhardt et al. (2006)). New work has shown that this bias is even present in children's books- as award winning books depict lighter skinned characters more, even with perceived race controls (Adukia et al. (2021)).

There are some studies directly linking criminal justice outcomes and skin tone. Blair et al. use perceived skin tone by correctional officers and find those with lighter skin face shorter sentences and shorter prison time served before parole (Blair, Judd and Chapleau (2004)). Viglione et al. find that inmates with perceived Afrocentric features face longer sentences and Eberhardt et al. find that those inmates who appear more "stereotypically Black" are more likely to get the death penalty (Viglione, Hannon and DeFina (2011); Eberhardt et al. (2006)). While these early studies are important, they are all dependent on individual assessments of skin tone or Afrocentric features. There are no universal standards for what it means to have a certain color of skin or Afrocentric features. Thus individual assessments likely vary widely and contain noise.

The goal of this project is to capture the nuance of the colorism studies but create a more consistent measure of the racial spectrum. In particular, using image processing and deep learning applied to photos of Mississippi inmates, I am able to create a continuous measure of race using a single classifier. The measure is based on a convolutional neural network that is trained to detect whether an individual is characterized as Black or white in the prison roster. I will show evidence that suggests the trained classifier is utilizing both skin tone and facial features in its classification decision. Using this new measure allows for a reexamination of traditional Black-white gaps. The raw Black-white gap in this setting is 777 days. When the sample is restricted to those the model found racially ambiguous (predicted probability of Black from 0.45-0.55), the gap shrinks to 215 days. Even with less restrictive definitions of racially ambiguous (predicted probability of Black from 0.25-0.75), the gap remains significantly lower at 292 days. Looking only at the poles of the distribution, whites with predicted probability of less than 0.05 or Blacks with probabilities greater than 0.95, the Black-white gap rises to 1,648 days. This finding suggests that individuals with more racially distinctive features are the ones driving the Black-white gap. This analysis also suggests that narrowing Black-white gaps may not necessarily be a good indicator for increasing overall equality. While a narrowing gap could be driven by less severe sentence lengths for all African Americans, it could also be driven racially ambiguous or lighter skinned African Americans receiving less severe punishments or racially ambiguous whites receiving harsher sentences, creating even more interracial disparities. I also find evidence of intraracial disparities in regression analysis. Even when controlling for specific crime types and demographics, likely a lower bound estimate on the impact of colorism, a one standard deviation increase in percent predicted Black increases sentence length by approximately 70-90 days.

Racial identity is complicated but given the new proliferation of photos and machine learning techniques, researchers can begin to incorporate a more nuanced version of race into studies of discrimination. This study is one such attempt to utilize new techniques in order to better understand how disparities evolve throughout the racial spectrum.

II. Data

A. Mississippi Prison Roster

After sentencing, soon to be state prisoners in Mississippi are taken to an intake processing center. There is one processing center for all prisoners. Once there, a team of psychologists, doctors, and social workers will assess the individual to determine the risk profile, medical needs, psychological needs, etc. so that the individual can be placed in a proper prison for their risk profile and so that they are given appropriate accommodations for any medical or psychological need. During this intake process, the prison records are formed. These records contain a photo of the inmate, demographic information (race, sex, date of birth, height, weight, complexion, build, eye color and hair color), as well as information about crime (number of sentences, sentence length, county of conviction, crime committed, past crimes details (up to three), prison location and prison unit. This information is posted in the Mississippi prison roster online. The Mississippi prison roster provides complete, contemporaneous information about individuals currently incarcerated in Mississippi. It does not provide information about individuals who were previously incarcerated, those not convicted of a crime, or those who were convicted and have since been paroled and/or released. As a note, this data set contains information on those convicted of more serious crimes. Misdemeanors or smaller infractions that carry short prison times (less than 6 months) are absent from this data.

B. Racially Distinctive Names

To later test the robustness of the convolutional neural network, I will use inmate last name as a proxy for race by linking naming data from the Decennial Census Surname Files for 2010. The Census data include all surnames reported 100 or more times in the 2010 Census, along with the origin and race category percentages associated with each surname. Eighteen individuals in prison roster have surnames that are not in the dataset (≤ 0.2 % of the sample). The Census suppresses racial percentages if there are not sufficient numbers of individuals to protect confidentiality. For all suppressed values, I impute a value of 0.

III. Measuring Race

This paper is not attempting to develop novel machine learning methods; rather, its aim is to apply established machine learning methods to a new domain. As such, the methodology proposed is purposefully streamlined and minimal.

Using `BeautifulSoup`, I downloaded 17,543 individual prison records from the Mississippi Prison Roster. Using `Tabula-Py` and `OpenCV`, each record was downloaded and the photos were extracted to .PNG files. The photos were all resized to 250 by 250 pixels. Photos of inmates are taken at a central location which ensures that differences in inmate appearance are actually due to differences in appearance, rather than dissimilar photo backgrounds. To standardize the images, I utilized `OpenCV`'s Haar Feature Based Cascades for Frontal Face recognition, a photo algorithm that is trained to isolate faces in photos. Using this algorithm, 16,074 faces were isolated in the sample [92% of the original sample]. In general, the omitted photos were disproportionately from inmates who were processed prior to 2010. Additionally, 2,835 individuals had life in prison/death sentences, had prison records that were missing critical information, or had a race not listed as Black or White ($\leq$ 1% of the data) and thus were excluded from the sample. This leaves a total sample of 13,538 photos. Each photo was cropped to include the area identified by the algorithm as containing a face and resized to 150 by 150 pixels. For this project, the model will take in the cropped photo as an input and outputs a probability that the individual is coded as Black in the prison roster.

A. Transfer Learning

The convolutional neural network is formed using transfer learning. Transfer learning is a popular technique in the machine learning space, where a model that has been trained on another dataset, optimization problem, etc. is adapted to a new setting (Pan and Yang (2009); Weiss, Khoshgoftaar and Wang (2016); Torrey and Shavlik (2010); Shao, Zhu and Li (2014)). Given the limited sample size of my setting, utilizing a model that has been pretrained on similar, larger datasets is a way to ensure that the model is able to pick up complex patterns within photos without requiring large sample sizes from the new setting. FaceNet is a state of the art facial recognition network (Schroff, Kalenichenko and Philbin (2015)). It captures facial characteristics and maps these to a compact Euclidean space. While this method can be used for facial recognition and verification, in this setting it will be used to predict if an individual is likely to be categorized as Black in the prison roster. I use the FaceNet architecture and weights pretrained by Hiroki Taniai on the MS-Celeb-1M dataset, a dataset with over 10 million celebrity photos (Taniai (2018)). It should be noted that traditionally facial recognition algorithms have had poor performance for minorities (Buolamwini and Gebru (2018)). Further development of image generation through GANs has also been shown to exacerbate this issue (Jain et al. (2020)). Despite these limitations, the image classifier used here has relatively high performance.

The training sample in this work is the 3,520 photos of individuals who entered prison prior to 2011 or in 2019. As camera technology has evolved over the past 10 years, solely training the algorithm on older photos lead to an over reliance on lighting conditions and poorer performance on newer photo classifications. To overcome this limitation, training on older photos and the most recent photos

minimizes the reliance of the algorithm on lighting conditions in the photos for race detection. The training sample is further divided into 3,000 photos for training the model and 520 photos as a hold out set for temperature scaling, discussed below.

The photos are reshaped so that they are an appropriate size for the FaceNet architecture but are otherwise unedited. The target values are binary where one represents an individual is recorded as Black and 0 indicates white. It should be noted that I have not entirely escaped the issue of human labeled data. The Mississippi Department of Corrections did not clarify how they determined inmate race, so it is likely either determined by the individual or by the social worker who is in charge of creating the prison record. As with skin tone, there are no universal standards for when someone is considered Black or white, and thus, individual classifications may vary. Despite this limitation, however, there are clear benefits of this strategy. The first benefit is that this model creates a consistent measure for the entire sample. One classifier is used on the entire sample and the features of the model can be tested and robustness checks can be run to ensure the model is using relevant characteristics. This type of analysis is not possible when numerous individuals are making classification decisions. The second advantage is that the classifier is trained on outside data (not the test data) and over thousands of iterations. If we think of each individual determining race as a mini classifier, then using an ensemble of these classifiers should lead to lower variance in the classification, i.e. a less noisy measure of race.

To adapt the FaceNet model to this setting, the original model architecture is appended with one additional layer and a final output layer. To find the optimal model, I use a parameter grid search and three fold cross validation to tune: number of epochs (1,3,5), regularization level (0.01, 0.1), dropout levels (0, 0.1, 0.5) and number of nodes (64, 128, 256) in the final hidden layer. Given the risk for overfitting, I freeze the original weights for the FaceNet network. The final model uses the FaceNet architecture minus the output layer, and appends 1 64 dense layer with ReLU activation and l2 regularization of .01 and one prediction output layer with a sigmoid activation. This model was stopped after training on only 5 epochs. The final model uses an Adam optimizer, a batch size of 32 and binary cross entropy as the loss function. The model was run 1000 separate times and the final 'prediction' variable is the average prediction for each value across all iterations. The final accuracy on the test set is 0.89, the F1-score is 0.94, and the area under the curve is 0.94.

B. Calibration

It is not a guarantee that the output of a neural network can be appropriately interpreted as a probability. In a perfectly calibrated model, if the predicted probability of Black is x%, then x% of the individuals with that probability should be listed as Black and 100-x% should be white in the prison roster. Calibration is important in this setting to be able to interpret the outcomes of the model as

the likelihood that an individual would be labeled as black or white in the prison roster. Much has been written on the importance of calibrating predicted probabilities from neural networks and other models (Guo et al. (2017); Pedregosa et al. (2011); Platt et al. (1999); Tibshirani, Hoefling and Tibshirani (2011); Zadrozny and Elkan (2002)). I use temperature scaling for calibration as advocated by Guo et. al. 2017, using the `netcal` package (Guo et al. (2017)). The Expected Calibration Error (ECE) for the calibrated model is 8.8%.

C. Mechanisms

Convolutional neural networks are able to pick up complex trends but can also be sensitive to noise. In this setting, it is important to ensure that the model is focusing more on facial characteristics and skin tone, traits that humans use in classifying race, rather than any background or tangential characteristics.

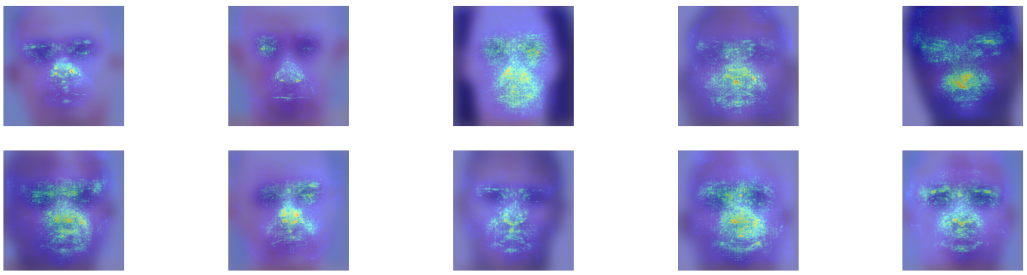
First, I use saliency maps to test if the model appears to be focusing on facial characteristics. Saliency maps are used to determine which pixels are particularly important in classifier decisions. Figure 1 shows the saliency map superimposed on blurred photos for a random selection of inmates in the test set. The saliency map is formed using SmoothGrad with 20 samples and noise at 0.2. The original photos are blurred for publication purposes, in order to ensure privacy for inmates, but the original, non-blurred photos are used in the formation of the saliency maps. In a saliency map, brighter colors indicate that a given pixel has relatively more influence over the final classification. In the figure, several things are clear. First, the most influential pixels are disproportionately concentrated on the nose, mouth and eye portion of the prison photos. This is reassuring that the classifier is primarily using facial characteristics in its classification. Additionally, these points of interest correspond to regions that are used in classifying Afrocentric vs. Eurocentric facial features (Blair, Judd and Chapleau (2004)).

Next, to test if skin tone appears to be an influential trait for the classifier, I created a skin tone only measure. This measure is found by finding the average individual typology angle, ITA, of the cropped prison photos. ITA is a skin tone measure used by dermatologists and uses the lightness (L^*) of the photos and the difference in coloration between yellow and blue tones (b^*) to detect darkness of skin tones. The L^* and b^* are from the CIELAB color space. The formula for ITA is below.

$$ITA^\circ = \arctan(\frac{L^* - 50}{b*}) \times \frac{180}{\pi} \tag{1}$$

In this measure, higher values indicate lighter skin tones. I find that the correlation between the predicted percent Black and this measure is -0.6 for the restricted sample and -0.5 for the full sample. Figure 2 shows a binned scatter plot, with predictive measure on the X-axis and the ITA skin tone on the Y-axis. The blue values correspond to the full sample and the red correspond to the restricted sample. The two measures are clearly negatively correlated through-

Figure 1. : Saliency Maps



Images are saliency maps with the highlighted pixels disproportionately concentrated on the nose and mouths of the photos

out the distribution. The relatively high magnitude of correlation between these two measures indicate that the model is likely using skin tone coloration in its determination.

Finally, Figure 3 shows a binned scatter plot of the predicted percent Black against a measure for racially distinctive Black names. The percent last name variable measures what share of individuals with the same name as the inmate were classified as Black in the Census. This is shown on the y-axis. There is a clear upward trend in both the restricted sample (in red) and the full sample (in blue). Individuals with higher predicted Black values are much more likely to have a more racially distinctive Black last name. As a note, one may be concerned about the presence of other racial groups or Hispanics in this sample. Based on the population makeup of Mississippi, this is unlikely to be a large threat to validity; less than one percent of Mississippi is Asian and approximately three percent are Hispanic. However, using the naming data, I can test if there are a large share of individuals with last names that disproportionately belong to individuals who report being Hispanic or Asian. Using the last name data, the average percent of Hispanic last names is 2.88 and Asian last names is 0.84. Thus, the trends shown here are unlikely to be influenced by individuals who are not white or Black.

The above analysis suggests that the classifier is capturing relevant characteristics in its decision, both facial characteristics and skin tone. It's difficult to disentangle the importance of skin tone versus facial features for the classifier; this distinction is further complicated as the two features can be correlated. Since the correlations with the new measurement are more or less the same as for the standard approach, I do not expect facial features to be a driving factor to a more severe degree than they already are in the status quo.

Table 1—: Summary Statistics

	Mean	St.Dev	Min	Median	Max
Panel A: Full Sample					
Year of Entry	2015.90	2.12	2011.00	2017.00	2018.00
Sentence Length 1st	3684.81	2582.11	180.00	2920.00	21900.00
Percent Black Predicted	0.62	0.31	0.02	0.72	1.00
Black	0.63	0.48	0.00	1.00	1.00
Age at Sentence	32.98	10.76	15.00	31.00	80.00
Male	0.93	0.26	0.00	1.00	1.00
Assault/Murder	0.19	0.39	0.00	0.00	1.00
Violent Property Crime	0.15	0.36	0.00	0.00	1.00
Drug Possession	0.08	0.27	0.00	0.00	1.00
DUI	0.02	0.14	0.00	0.00	1.00
Intent to Distribute	0.12	0.32	0.00	0.00	1.00
Burglary	0.25	0.43	0.00	0.00	1.00
Sex Crime	0.12	0.33	0.00	0.00	1.00
Other Crime	0.04	0.18	0.00	0.00	1.00
Weapon Crime	0.03	0.17	0.00	0.00	1.00
Panel B: Restricted Sample					
Year of Entry	2017.64	0.83	2011.00	2018.00	2018.00
Sentence Length 1st	3510.36	2925.01	188.00	2555.00	21900.00
Percent Black Predicted	0.57	0.31	0.02	0.65	1.00
Black	0.61	0.49	0.00	1.00	1.00
Age at Sentence	33.55	10.70	16.00	32.00	73.00
Male	0.89	0.31	0.00	1.00	1.00
Assault/Murder	0.25	0.43	0.00	0.00	1.00
Violent Property Crime	0.08	0.27	0.00	0.00	1.00
Drug Possession	0.12	0.32	0.00	0.00	1.00
DUI	0.01	0.07	0.00	0.00	1.00
Intent to Distribute	0.09	0.29	0.00	0.00	1.00
Burglary	0.34	0.47	0.00	0.00	1.00
Sex Crime	0.04	0.21	0.00	0.00	1.00
Other Crime	0.06	0.23	0.00	0.00	1.00
Weapon Crime	0.01	0.11	0.00	0.00	1.00

The full sample is composed of of 9,719 individuals and the restricted sample has 2,544 individuals. The full sample is the complete contemporaneous set of individuals incarcerated in June 2019. The restricted sample is composed of only those individuals who were convicted of crimes such that they would be illegible for parole and/or release in June 2019. More detailed information about the restricted sample composition can be found in Section III.

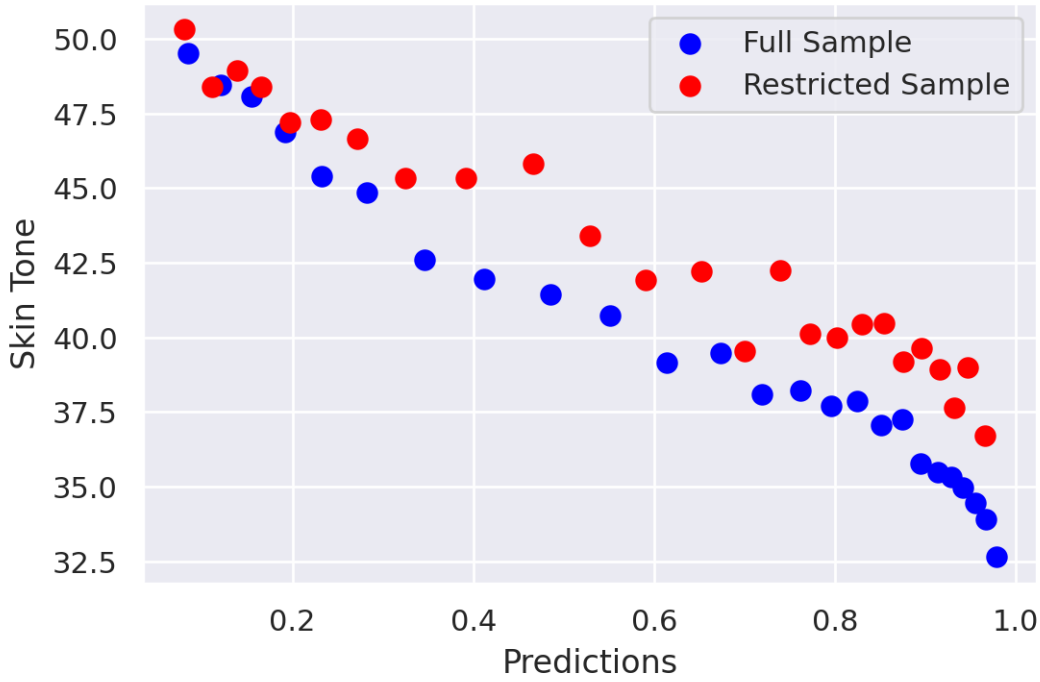


Figure 2. : Prediction vs. Skin Tone

The graph shows a scatter plot, depicting a negative relationship between predictions and skin tone for both the full and restricted sample

IV. Summary Statistics

The analysis that follows will use two different datasets: the complete dataset, which contains the contemporaneous prison record i.e. only individuals who were in prison in June 2019 when the data was scraped, and a restricted sample.

The full dataset will provide a snapshot of the prison population at June 2019. Individuals who have been convicted prior to June 2019 but have since been paroled or released are no longer included in the prison roster. So for a given sentence year, the roster will be over-representative of individuals with longer prison sentences, as those individuals with shorter prison sentences are more likely to be paroled or released and absent from the data. If darker skinned individuals are more likely to serve longer sentences, they will be disproportionately represented in earlier sentence years. To account for this sample composition, the restricted sample has been created to minimize differential levels of attrition, and thus will contain a more representative sample of individuals convicted of a given crime in a given year.

The restricted sample is composed of inmates who were convicted of crimes that

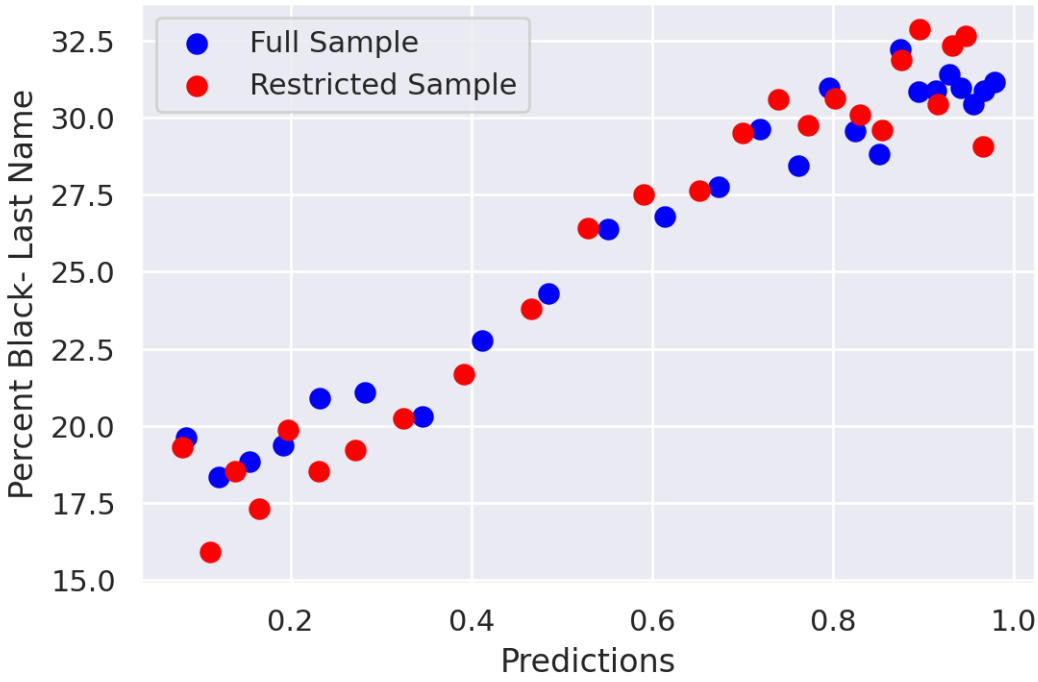


Figure 3. : Prediction vs. Last Name Black

The graph shows a scatter plot depicting a positive relationship between predictions and percent last name for both the full and restricted sample

carry sentences that would make them illegible for parole in June 2019, when the data was collected. Under the 2014 Mississippi prison reforms, all individuals who are convicted of violent crimes are required to serve at least 50% of their sentence in prison and all individuals who are convicted of nonviolent crimes are required to serve at least 25% of their sentence. To determine the sentence length for a given type of crime, I use the first quartile sentence lengths for individuals convicted of the same crime in 2019. As a reminder, those convicted in 2019 are not part of the test sample in this study. By using this out of sample and recent sentencing data, I have a more complete representation of all individuals and sentences for a given crime type. Given the classification of nonviolent and violent offenses, which was also outlined in the 2014 legislation, the sample is formed of individuals who would not have been eligible for parole in June 2019, even if they were given a lenient sentence. To illustrate, the first quartile of sentence length for aggravated assault is 5 years. Aggravated assault is a violent offense so individuals must serve at least 2.5 years, thus only individuals who were convicted of aggravated assault on or after January 6, 2017 (2.5 years prior) will be included in the restricted sample. For given time periods, this assures that

the complete set of individuals convicted of this crime are included in the sample. Given the formation of this restriction, the restricted sample will include relatively more individuals sentenced recently and for more serious crimes - although the full sample is already weighted towards individuals convicted of felony offenses and those sentenced more recently.

The summary statics are shown in Table 1. The individuals in the full sample have an average first sentence length of 3,685 days, a little over 10.5 years. The sample is 63% Black and is relatively young, with the average age at first sentence at 33 years old. The most common crime type is burglary (25%) which is a nonviolent property crime, followed by assaults/murders (19%) and violent property crimes (15%), which are property crimes where some form of force or weapon was used. In the restricted sample, the average sentence length falls to 9.6 years and percent Black falls to 61%. The crime composition is slightly different, burglary makes up a larger share (34%), followed by assault/murder (25%) and drug possession (12%). Sex crimes and intent to distribute crimes, drug crimes where there was some intent to distribute or manufacture an illicit substance, make up a smaller share of the restricted subset. In addition to standard summary statistics, one of the innovations of this paper is to incorporate new media types to document social phenomenon. Figure 4 shows the average prison roster photo by sentence length. This was created by averaging each RGB (red, green, blue) value for each pixel for every prisoner within a certain sentencing length band. The first row shows the average photo for the entire sample, where the first column is all individuals who were sentenced to 0-5 years in prison, the second column shows the average photo of individuals sentenced for 5-10 years, the third column is individuals sentenced to 10-20 years and the final column is those sentenced to 20+ years in prison. There is a clear trend where as sentence length increases, the average photos appear to darken. It is possible that the trends in the first row do not show the true change in skin color but rather reflects the fact that African Americans are sentenced to longer sentences on average and thus make up a higher percentage of each subsequent binned sentence length group. While this is true, the same trends can be seen in the the second row which shows Black prisoners (as classified by the MS prison roster), and in the third row which shows only white prisoners. As sentence length increases, the average complexion darkens. This is seen most clearly in the jump from 5-10 year sentences to 10-20 year sentences. This trend is most pronounced in the second row (Black) but can also be seen for white prisoners as well. This visual evidence suggests that there may be a connection between darker skinned individuals and longer sentences.

V. Results

The photo analysis techniques are themselves one of the main contributions of this paper. However, it is important to analyze if the new measure of race has value added over the traditional measure of race. The following sections will show two applications using both the new measure and the traditional Black/white

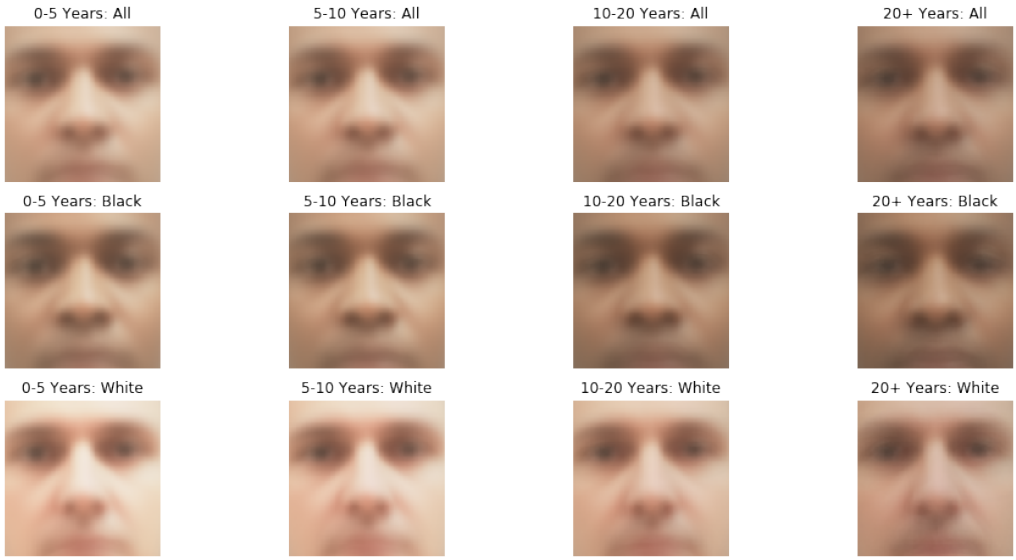


Figure 4. : Average Photos by Binned Sentence Length

Photos included in this analysis are from inmates in the 2019 prison roster. Section II includes details on the photo cleaning process. The top row contains all individuals, the second row is restricted to individuals classified as Black in the prison roster and the third row is restricted to individuals classified as white in the prison roster. The first column shows the average photo, found by taking the average RGB measure for each pixel, for individuals sentenced to 5 years or less. The second column shows the average photo for individuals sentenced to 5-10 years. The third column shows individuals sentenced for 10-20 years and the fourth column shows individuals sentenced to 20+ years.
The first column shows the average photo, found by taking the average RGB measure for each pixel, for individuals sentenced to 5 years or less. The second column shows the average photo for individuals sentenced to 5-10 years. The third column shows individuals sentenced for 10-20 years and the fourth column shows individuals sentenced to 20+ years

measure of race.

VI. Black-white Gaps

Black-white gaps are a key metric used in quantifying racial disparities. While this is a helpful metric, focusing on interracial differences misses the nuance of how racial disparities may vary for those with more or less racially distinctive features. Using a running measure of race can help illuminate inter- and intraracial differences, a more informative metric in evaluating disparate outcomes.

For all analysis, the determination of Black and white will be sourced from the Mississippi prison roster. The first column in Table 2, the "Raw" column, shows the Black-white gap in the full sample is 311 days and 777 days in the restricted sample. The difference between the two Black-white gaps is almost solely driven by the shorter average sentences for whites in the restricted sample. The average white sentence is 3,489 days in the full sample and 3,038 days in the restricted sample. The levels of attrition likely differ by perceived race in the full

Table 2—: Black-white gaps in sentencing length by prediction probability

	Raw	0.45-0.55	0.25-0.75	≤0.05, ≥0.95
	(1)	(2)	(3)	(4)
Panel A: Full Sample				
Black	3800.5	3687.8	3761.9	4064.9
White	3489.0	3684.9	3669.5	3140.1
Black-white Gap	311.4	2.9	92.3	924.8
Panel B: Restricted Sample				
Black	3814.4	3718.7	3746.6	4409.6
White	3037.6	3503.8	3454.5	2761.3
Black-white Gap	776.7	214.9	292.1	1648.3

Column 1 shows the average sentence length in days, subdivided by stated race in the prison roster. Columns 2-4 show the average sentence length for individuals who have predicted percent black values in various ranges, subdivided by stated race in the prison roster. Panel A shows the results for the full sample and Panel B shows the results for the restricted sample.

sample, with whiter individuals more likely to be given lesser sentences and thus absent from the data. The restricted sample has been created to minimize overall attrition, and thus has a higher relative representation of white individuals with shorter sentence lengths.

To understand how the Black-white gap may vary across the racial spectrum, I limit the sample to individuals who have a predicted probability of Black from 0.45 to 0.55. These individuals are likely the most racially ambiguous in this sample. When the sample is restricted to individuals with a predicted measure of 0.45 to 0.55, the Black-white gaps shrink significantly. In the full sample, the gap shrinks to only 3 days and the average sentence is approximately 3,685 days. This value is almost directly between the average white (3,489) and Black sentence (3,800) length in the raw full sample (column 1). In the restricted sample, the gap shrinks by over two-thirds to 215 days, where the average sentence length increases to 3,503 for whites and falls to 3,718 for Blacks. Most of the difference results from a large increase in average white sentence length for more racially ambiguous whites, although there is also decline in average sentence length for those who were racially ambiguous and recorded as Black.

Limiting the samples to only those with predictions ranging from 0.45-0.55 may be seen as too restrictive, so column 3 shows the range expanded to individuals with predicted probability of 0.25 to 0.75. The Black-white gap for individuals who have a predicted probability of Black between 0.25 and 0.75 is 93 days in the full sample. The Black-white gap is 292 days in the restricted sample. This is the result of both slightly shorter sentences for African Americans with more

racially distinctive features but again is largely driven by much longer sentences for whites that are racially ambiguous.

Finally, to fully demonstrate that much of the Black-white gap is driven by those at the poles, column 4 shows the average sentence length for those at the very poles of the distribution, those with predicted probability of less than 0.05 or greater than 0.95. Given the model is properly calibrated, the vast majority of individuals with a probability of 0.05 or less are coded as white (2% Black) and the vast majority of individuals with a probability of 0.95 or greater are coded as Black (98% Black). Therefore, I will restrict the row value to only include Black individuals with predicted probability of 0.95 or above and the white row will include white individuals with a predicted probability of 0.05 or below. Using this sub-sample, the Black-white gaps grow to 925 days for the full sample and 1,648 for the restricted sample. In the full sample, the average Black sentence length is approximately 250 days longer and the average white sentence length approximately 350 days shorter at the poles. In the restricted sample, the average Black sentence length is approximately 600 days longer and the average white sentence length is approximately 275 days shorter than in the raw sample.

This analysis demonstrates the variation in Black-white gaps throughout the distribution of the race measure. The gaps shrink significantly when the sample is restricted to those with more racially ambiguous features but balloon when only including those at the poles of the distribution. It follows that individuals who are more racially ambiguous tend to have sentence lengths between those who are perceived to have more racially distinctive white and Black features. This analysis also suggests that whites who appear racially ambiguous face penalties much more similar to those faced by racially ambiguous African Americans than those who are perceived to have more distinct white features, which actually drives down the Black-white gap. Racially ambiguous African Americans also have much lower sentence lengths than African Americans at the poles of the distribution. This analysis suggests that Black-white gaps alone are an incomplete measure of disparities and may understate disparities for those African Americans who are perceived to have more Afrocentric features and overstate disparities for those with less racially distinctive features.

A. Regression Results

The results in the Black-white gaps are important in assessing raw disparities. However other attributes, such as crime type, age of arrest, gender, etc., may also influence sentence length and correlate with perceived race. Using regression analysis with a measure for the racial spectrum thus offers an opportunity to evaluate how sentencing disparities vary throughout the distribution while controlling for confounding factors.

The main specification will estimate:

$$SentLength_i = \beta_0 + \beta_1 Prediction_i + \lambda_i + \epsilon_i \tag{2}$$

$$SentLength_i = \beta_0 + \beta_1 Black_i + \lambda_i + \epsilon_i \tag{3}$$

Table 3—: Regression Results

	Sentence Length for First Sentence in Days			
	(1)	(2)	(3)	(4)
Panel A: Full Sample, Prediction				
Prediction	473.50	495.20	505.57	244.44
	[314.24, 629.92]	[321.17, 661.28]	[350.51, 656.5]	[77.40, 399.58]
Panel B: Full Sample, Black				
Black	273.25	294.11	348.79	140.55
	[173.6, 370.86]	[194.7, 395.66]	[248.89, 455.93]	[36.06, 251.10]
Panel C: Restricted Sample, Prediction				
Prediction	1033.38	991.80	579.53	312.19
	[695.72, 1391.84]	[626.88, 1352.65]	[260.31, 915.94]	[9.04, 599.38]
Panel D: Restricted Sample, Black				
Black	759.90	729.68	463.23	242.30
	[516.76, 971.07]	[488.69, 962.86]	[246.30, 687.17]	[23.46, 450.57]
Additional Controls	Reform Year	Demographics	Crime Category	Detailed Crime

Coefficient listed is the mean coefficient estimate on the measure of race after 1,000 bootstrapped iterations. The outcome variable for all regression is sentence length, measured in days. The brackets show the 2.5 and 97.5 percentiles of the coefficients, respectively. Panel A and B use the full sample (N=9,719) and Panel C and D use the restricted sample (N=2,544). Details about sample composition can be found in Section III. Panel A and C use the created measure of predicted percent Black as the racial measure. Panel B and D use the binary designation of Black, sourced from the prison roster, as the measure of race. Column one controls for whether an individual was sentenced pre/post sentencing reform in March 2014. Column 2 controls for both the sentencing reform indicator, as well as demographics of age and sex. Column three controls for demographics, sentencing reform, and broad crime categories of violent, drug, property, or other. Column four controls for demographics, sentencing reform, and specific crime type.

Where $SentLength_i$ is the first sentence length in days, $Prediction_i$ is the running measure of predicted race, $Black_i$ is 0/1 indicator for being listed as black in the prison roster and λ_i is a set of controls. Table 3 shows the results for both the entire sample (Panel A and B) and the restricted sample described above (Panel C and D). The coefficients shown are the mean of the coefficient value after 1,000 bootstrapped iterations. The brackets show the 2.5 and 97.5th percentiles of the coefficients, respectively.

Column 1 shows the regression of sentence length on the metric of race and a single control indicating whether the individual was sentence pre- or post the 2014 sentencing reforms. Column 2 adds in additional controls of age at sentencing and sex. Column 3 adds in controls for collapsed crime categories of drug, violent, property or other. Finally column 4 has controls for demographics, an indicator for pre/post reform and detailed crime types, i.e. intent to distribute, murder/assault, etc. Panel A and C use the running measure of predicted percent Black as the metric of race and Panel B and D use a binary indicator for Black.

For the full sample and using the running measure of race (Panel A column 1), the raw coefficient for predicted percent Black is 473.5 days. To interpret this, for a one standard deviation increase in predicted percent Black, the predicted sentence length increases by 146.6 days. The magnitude is relatively unchanged when adding in additional controls for demographics and broad crime categories with implied increases in sentence length of 153.5 days and 156.7 days, respectively, for one standard deviation increases in predicted percent Black. When controlling for specific crime types, however, the magnitude of the coefficient falls almost by half to 244.4 or a 75.8 day increase for a one standard deviation increase in predicted percent Black.

The magnitude for the restricted sample (Panel C and D) is larger for all measures, although the results are more similar to those for the full sample (Panel A and B) once crime type controls are added. This may in part be due to the relatively higher shares of individuals convicted for second degree murder and manslaughter, offenses that carry significant sentences, in the restricted sample. Similar to the full sample measures, in Panel C, the inclusion of demographics has little impact on the magnitude of the coefficient (1,033 vs. 992 days). Controlling for broad crime category drops the magnitude from 992 to 580, or a 307 day increase in sentence length to a 180 day increase in sentence length for a one standard deviation increase in predicted percent Black. Similar to the full sample, the magnitude almost halves when controlling for detailed crime type to 312 or a 97 day increase for a one standard deviation increase. Panel D uses the same restricted data as in Panel C but uses Black as the racial measure. The magnitude of the coefficients are all smaller than the coefficients in Panel C. The trend in coefficients in Panel D also mirrors the trends in Panel C, where each additional control decreases the magnitude of the coefficient on race.

Before analyzing the differences between the two measures of race, various trends should be noted. Irregardless of which measure for race is used, in all regressions the coefficient on the racial measure are large and significant at the 5 percent level. This is consistent with the vast literature documenting racial disparities in the criminal justice system. Beyond statistical significance, there are other noteworthy trends in this analysis. The coefficients on the metrics of race are much smaller with the inclusion of detailed crime type, even when compared to the regressions that control for broad crime type. The large decrease in magnitude on the coefficient suggests that perceived race is correlated with

conviction offense severity. I cannot rule out the explanation that perceived race may influence criminal behavior nor can I rule out that discrimination on behalf of prosecutors or other members of the criminal justice system may lead to this difference. However, this analysis is suggestive that both increased sentences for similar crimes and being charged for more serious crimes play a role in the longer sentences for those with higher likelihood of being perceived as Black.

From an interpretation standpoint, it is clear that having a running measure for race, as compared to a categorical measure, holds some advantages. Rather than comparing group level means, one can use running measures to evaluate how predicted changes in sentence length vary for small changes in perceived racial identity. Additionally, one can analyze changes in the poles of the distribution. In this sample, the predicted change in sentence length from a value of predicted Black from 0 to 1 is larger in magnitude than the coefficient on Black. This trend holds for all regression specifications and in both samples. This is consistent with larger interracial gaps at the poles of the distribution and smaller gaps in the middle of the distribution and with the findings in the Black-white gap analysis. This analysis demonstrates that intraracial disparities are persistent and deserve further study and quantification in the criminal justice system.

VII. Conclusion

This paper has presented one novel application of machine learning to the detection of bias. Utilizing photos of individuals rather than human labeled classifications of skin tone or Afrocentric features creates a more consistent measure of the racial spectrum. This project utilizes this improved measure to study how discrimination varies throughout this spectrum in several ways.

First, it shows that while there are significant Black-white gaps, this is largely driven by those at the poles of the distribution. If colorism is occurring but racial measures are only at a binary level, then disparities for more racially ambiguous African Americans will be overstated and disparities for African Americans with more racially distinctive features will be understated. In this setting, the Black-white gap shrunk significantly when only including individuals in the middle of the distribution, those most likely to be racially ambiguous. These trends are also found in the regression analysis, even after controlling for confounding variables.

This study demonstrates one potential application of machine learning in the detection of bias; however, it is not able to evaluate the root of the difference in sentence length. Future work can begin to evaluate the causal mechanisms behind this relationship. Additionally, actors in the social sciences, and increasingly in applied computation, test for forms of discrimination in hiring, education and criminal justice outcomes. This study suggests that in addition to racial bias, machine learning, and specifically the inclusions of photos as data, may allow us to detect more complex discrimination in the form of colorism in these various applications.

REFERENCES

Adukia, Anjali, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz. 2021. "What We Teach About Race and Gender: Representation in Images and Text of Children's Books." *University of Chicago, Becker Friedman Institute for Economics Working Paper*, , (2021-44).

Anbarci, Nejat, and Jungmin Lee. 2014. "Detecting racial bias in speed discounting: Evidence from speeding tickets in Boston." *International Review of Law and Economics*, 38: 11–24.

Arnold, David, Will Dobbie, and Crystal S Yang. 2018. "Racial bias in bail decisions." *The Quarterly Journal of Economics*, 133(4): 1885–1932.

Blair, Irene V, Charles M Judd, and Kristine M Chapleau. 2004. "The influence of Afrocentric facial features in criminal sentencing." *Psychological science*, 15(10): 674–679.

Buolamwini, Joy, and Timnit Gebru. 2018. "Proceedings of Machine Learning Research." Vol. 81, 77–91.

Eberhardt, Jennifer L, Paul G Davies, Valerie J Purdie-Vaughns, and Sheri Lynn Johnson. 2006. "Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital-sentencing outcomes." *Psychological science*, 17(5): 383–386.

Frazier, Edward Franklin. 1957. *Race and culture contacts in the modern world.* Knopf New York.

Goldsmith, Arthur H, Darrick Hamilton, and William Darity Jr. 2006. "Shades of discrimination: Skin tone and wages." *American Economic Review*, 96(2): 242–245.

Goncalves, Felipe, and Steven Mello. 2021. "A Few Bad Apples? Racial Bias in Policing." *American Economic Review*.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. "On calibration of modern neural networks." 1321–1330, JMLR. org.

Jain, Niharika, Alberto Olmo, Sailik Sengupta, Lydia Manikonda, and Subbarao Kambhampati. 2020. "Imperfect imaganation: Implications of gans exacerbating biases on facial data augmentation and snapchat selfie lenses." *arXiv preprint arXiv:2001.09528*.

Keith, Verna M, and Cedric Herring. 1991. "Skin tone and stratification in the Black community." *American journal of sociology*, 97(3): 760–778.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "Human decisions and machine predictions." *The quarterly journal of economics*, 133(1): 237–293.

Monk Jr, Ellis P. 2015. "The cost of color: Skin color, discrimination, and health among African-Americans." *American Journal of Sociology*, 121(2): 396–444.

Pan, Sinno Jialin, and Qiang Yang. 2009. "A survey on transfer learning." *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. "Scikit-learn: Machine learning in Python." *the Journal of machine Learning research*, 12: 2825–2830.

Platt, John, et al. 1999. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods."

Reece, Robert L. 2016. "What are you mixed with: The effect of multiracial identification on perceived attractiveness." *The Review of Black Political Economy*, 43(2): 139–147.

Schroff, Florian, Dmitry Kalenichenko, and James Philbin. 2015. "Facenet: A unified embedding for face recognition and clustering." 815–823.

Shao, Ling, Fan Zhu, and Xuelong Li. 2014. "Transfer learning for visual categorization: A survey." *IEEE transactions on neural networks and learning systems*, 26(5): 1019–1034.

Sloan, CarlyWill. 2019. "Racial bias by prosecutors: Evidence from random assignment."

Taniai, Hiroki. 2018. "keras-facenet."

Tibshirani, Ryan J, Holger Hoefling, and Robert Tibshirani. 2011. "Nearly-isotonic regression." *Technometrics*, 53(1): 54–61.

Torrey, Lisa, and Jude Shavlik. 2010. "Transfer learning." In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. 242–264. IGI Global.

Tuttle, Cody. 2019. "Racial disparities in federal sentencing: Evidence from drug mandatory minimums."

Viglione, Jill, Lance Hannon, and Robert DeFina. 2011. "The impact of light skin on prison time for black female offenders." *The Social Science Journal*, 48(1): 250–258.

Weiss, Karl, Taghi M Khoshgoftaar, and DingDing Wang. 2016. "A survey of transfer learning." *Journal of Big data*, 3(1): 9.

West, Jeremy. 2018. "Racial bias in police investigations."

Zadrozny, Bianca, and Charles Elkan. 2002. "Transforming classifier scores into accurate multiclass probability estimates." 694–699.