



Performance Differentials in Deployed Biometric Systems Caused by Open-Source Face Detectors

Cynthia M. Cook

Identity and Data Sciences Laboratory
Upper Marlboro, Maryland, USA
ccook@idslabs.org

Laurie Cuffney

Identity and Data Sciences Laboratory
Upper Marlboro, Maryland, USA
lcuffney@idslabs.org

John J. Howard

Identity and Data Sciences Laboratory
Upper Marlboro, Maryland, USA
jhoward@idslabs.org

Yevgeniy B. Sirotin

Identity and Data Sciences Laboratory
Upper Marlboro, Maryland, USA
ysirotin@idslabs.org

Jerry L. Tipton

Identity and Data Sciences Laboratory
Upper Marlboro, Maryland, USA
jtipton@idslabs.org

Arun R. Vemury

The U.S. Department of Homeland
Security
Washington, District of Columbia
USA
arun.vemury@hq.dhs.gov

Abstract

Testing of AI systems is important for ensuring accuracy and reliability. In this study, we demonstrate how scenario testing with demographically varied subjects, a form of prospective testing that simulates real-world conditions, revealed significant performance issues in biometric systems prior to broad deployment. Using generalized linear modeling, we show that subjects' measured skin lightness, along with other demographic factors, significantly impacted the probability of failure to detect a face. Failure rates increased from just 0.28% for subjects with the lightest skin in our sample to 24.34% for subjects with the darkest, controlling for other factors. We show that skin lightness, rather than self-reported race, best explained the differences in system performance. We trace these issues to widely used, older methods in open-source packages for face detection. Furthermore, this demographic differential is not observed when testing open-source packages using a different, more curated dataset. Our results highlight the need to evaluate full multi-component, operationally deployed AI systems and the role of scenario testing as a critical component of AI governance. One way to mitigate the likelihood that poor-performing, older open source methods are deployed in an operational system would be to deprecate these functions in favor of higher-performing alternatives. Prospective assessments of AI, in real-world use cases with demographically varied subjects, should be used to identify performance issues before these systems are operationalized.

CCS Concepts

• General and reference → Evaluation; • Human-centered computing → User studies; • Computing methodologies → Biometrics.

Keywords

Face Detection, Demographic Differentials, Scenario Testing



This work is licensed under a Creative Commons Attribution 4.0 International License. *FAccT '25, Athens, Greece*

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1482-5/25/06

<https://doi.org/10.1145/3715275.3732171>

ACM Reference Format:

Cynthia M. Cook, Laurie Cuffney, John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury. 2025. Performance Differentials in Deployed Biometric Systems Caused by Open-Source Face Detectors. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*, June 23–26, 2025, Athens, Greece. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3715275.3732171>

1 Introduction

The rapid growth and adoption of artificial intelligence (AI) and machine learning (ML) technologies has spurred governments and other organizations to study the potential benefits and harms of these technologies and to develop regulations that promote the benefits and mitigate the harms. In 2024 alone, the European Union passed Regulation (EU) 2024/1689, also known as the AI Act [29], the U.S. Commission on Civil Rights released “The Civil Rights Implications of the Federal Use of Facial Recognition Technology” [28] and a joint report was issued by the White House Office of Science and Technology Policy, the U.S. Department of Homeland Security, and the U.S. Department of Justice on the U.S. federal government’s use of biometric technologies [26]. While each of these documents varies in its exact focus and recommended actions, the need for testing of AI systems in operationally relevant scenarios emerged as a common theme. The EU AI Act has explicit exemptions for “testing in real world conditions” for both AI system providers (Article 60) and national regulation authorities (Article 74), while [28] called the “development of an operational testing protocol” a key recommendation for face recognition systems and [26] specified testing “as close to an operational context as possible” as a best practice.

But why this novel emphasis on operationally relevant testing? Technology testing of AI applications typically involves retrospective testing on curated datasets. This allows for benchmarking and performance tracking to identify algorithms with the best potential for utility in the real world (“state of the art”) [12]. Scenario and operational testing of AI applications involves collecting new data, either in simulated real-world conditions (scenario testing) or real-world conditions (operational) to test the full system. Scenario and operational tests of AI systems can reveal new performance issues not encountered in technology tests, including in sample

acquisition [14, 15] because these forms of testing evaluate the full system (e.g. camera systems, user interfaces, signage) in concert with the use-case and user interaction.

Face recognition is a special class of AI systems that is concerned with establishing the identity of individuals. Automated face recognition is often used for access control to secure spaces. A key step in automated face recognition is detecting a face in an image, which is necessary for all subsequent face recognition operations. In this context, if face acquisition fails, a human must manually carry out the task of identity validation, which is inconvenient and potentially prone to error. It is therefore important to ensure that automated face recognition processes work for all individuals, independent of their demographic group membership. Evaluations to establish face recognition performance are socio-technical in nature, since they include both the technology (face recognition system) and social components (humans, signage, etc.).

There is a significant body of research on the topic of challenges faced with automated face detection [3, 13]. Performance differentials including demographic effects are not new. Existing literature exposes performance differentials including those across different demographic groups such as race and skin tone [2, 8, 21, 39]. Emerging research aims to provide better balanced training datasets [20, 31] and explores synthetic data to address ethical and legal challenges with sourcing face data for training and development of modern face detection [16, 23]. While ideally face detectors would be modern variants supported by these research efforts, reality requires that a real-world implementation of the theoretical needs to be cost effective and easy to put into practice. Even though well-known theoretical issues exist with open-source face detection algorithms, they still remain easy to implement, train, and maintain, thus are still used in commercial applications today [25, 34].

This research documents the findings of a scenario test of two face acquisition systems that were being prepared for deployment to facilitate travel processing in real world locations. For this manuscript, we've obfuscated the system names as System A and System B. Both systems had automated face acquisition software to capture face samples without the assistance of a human operator. System A utilized an undisclosed proprietary face detection algorithm. System B utilized a combined OpenCV Haar cascade frontal classifier and Haar Cascade eye classifier for face detection. We examine the accuracy and demographic differentials in accuracy using self-reported and measured demographic variables for these two systems and investigate the face detection component as a potential source of observed variability using three widely utilized open-source face detection algorithms: OpenCV, Dlib, and a DNN-based detector (DNN).

The contributions of this research are four-fold. First, we demonstrate how scenario testing revealed significant differences in failure-to-acquire rates across the two face acquisition systems as tested in their intended high-throughput use case. System B failed-to-acquire over 20 times the number of subjects as System A. Second, we show these failures to acquire do not happen at random and occur at uneven rates across different demographic groups. Using generalized linear modeling, we show that, of the demographic variables controlled for in this study, estimated probabilities of failure-to-acquire on System B increased for volunteers with darker skin, glasses, head coverings, and older ages. These performance issues validate the

recent regulatory focus on the evaluation of operational systems. Third, we reproduce similar failure rates and demographic effects exhibited by System B when running the OpenCV Haar-cascade face detector on images acquired by System A. Finally, we show that another popular face detector from Dlib shows similar, but distinct significant demographic effects on images from System A. Neither detector shows issues handling images from an enrollment use-case with a more controlled environment.

Overall, this socio-technical evaluation shows that both OpenCV and Dlib have significant performance differentials relative to a newer DNN-based detector and demonstrates the need for pre-deployment testing of AI systems as a critical component of AI governance. We discuss the real-world harms that can result from including these popular tools in standard machine vision libraries and propose mitigating the risk of harm by deprecating systems with known performance issues when better performing alternatives are available.

2 Background

2.1 Test Methodologies for Biometric Technologies

Testing of biometric systems is governed by long-standing international standards which define three different test types, each with its own benefits and limitations: technology testing, operational testing, and scenario testing [18]. Technology testing focuses on computing metrics for specific subsystems (e.g., matching algorithms) on large benchmark datasets to report performance metrics useful for comparing the performance of different biometric technologies or tracking technology performance improvement over time. However, benchmark datasets used for technology tests remain fixed over time and may not represent new data captured in specific operational use cases [9]. Therefore, while findings from technology tests can describe how technology performs in general, measured performance may not apply to a specific technology use-case. Additionally, because the datasets are fixed, augmenting these data with new, ground-truthed, demographic information about the data subjects is often not possible.

Operational testing involves testing full operationally deployed systems in a specific operational environment. This type of testing uses data captured by deployed biometric systems within the operational environment and may include data from regular system users to evaluate performance. Operational data is associated with greater privacy concerns and results in limited demographic metadata due to governing regulations, laws, and other consumer protections [1]. Testing in operational settings includes uncontrolled environmental factors and system integration effects that may make it challenging to isolate the impact of specific factors on technology performance. Findings from operational testing may not generalize to other situations or use-cases.

Scenario testing aims to simulate operational deployment of complete biometric systems with demographically varied groups of users, but within a controlled environment. Like operational testing, this type of testing allows assessment of full systems. Scenario testing allows control and evaluation of user interaction and important factors such as variation in the collection environment [9]

which may be impossible or difficult in other forms of testing. Additionally, scenario testing allows the collection of new biometric data with ground truth information about subject demographics and metadata regarding system interactions that can facilitate root cause analysis of any system errors. Scenario tests require appropriate resources and physical test facilities (e.g., The Maryland Test Facility; <https://mdtf.org>).

2.2 Biometric Face Recognition

Face recognition is the process of using the physiological characteristics of a person's face to verify that person against a claimed identity (e.g., matching a live captured face to a face printed on an identity document) or to identify them against a gallery of known subjects (e.g., individuals boarding a plane). In order to perform this function, a face recognition system must first acquire a face sample from the unknown individual. Sample acquisition is defined as successfully completing two steps: 1) capturing a sample (i.e., taking a photograph) and 2) having the resulting sample be declared "suitable" according to the acquisition systems policy [10]. An acquisition of a suitable sample in an operational context could mean many things, including: the sample was captured, a face was detected, the face passed a quality filter, and/or the face was successfully extracted by a face recognition algorithm. Prior scenario testing research has consistently shown the largest source of error in facial recognition to be failure-to-acquire [14, 15], meaning a failure to complete both of the steps previously described. This reinforces the notion that failure-to-acquire a suitable data sample is an important aspect to consider when attempting to estimate the real-world performance of an AI system [11].

2.3 Face Detection

Fully automated acquisition of biometric samples for face recognition requires the use of face detectors. Given an image that contains a facial sample, face detectors return a bounding box around the facial sample. Face detection has been of longstanding interest in machine vision due to the numerous use cases supported by this capability [36]. Because of this long history, popular machine vision libraries like OpenCV and Dlib contain built-in face detectors. These libraries are available in a variety of programming languages, common in computer vision academic curricula and machine learning tutorials, and are active in their respective open-source repositories. For example, the OpenCV python package "opencv-python" is downloaded over half a million times a day [30]. Therefore, it is highly likely these open-source machine vision libraries are regularly deployed in operational machine vision applications.

Face detection systems are benchmarked using retrospective evaluations [37]. Whereas these benchmarks reveal performance differences between different detectors, they generally do not address demographic differentials. Furthermore, face detection systems are likely trained on images included in such datasets which may artificially improve their performance on such benchmarks. Recently, Yang et al. added demographic labels to the wider-face benchmark dataset and found demographic performance differentials in several open-source face detection models [38]. Dooley et al. have demonstrated demographic differentials in face detection robustness to digital perturbations applied to images from several

datasets [7]. Both studies have shown that detection performance is reduced based on several demographic factors, including for people with darker apparent skin lightness.

In some ways, face detection is viewed as a commodity technology with most work on face related demographic performance focused instead on demographic differentials in face recognition [12]. Further, there has been a recent proliferation of face detection tools including open-source projects and services provided via web-based APIs and developers may use whatever detector is most convenient for their application. Developers new to image processing may be biased toward the most readily available systems that are frequently included in tutorials including face detectors in OpenCV [24, 32] and Dlib [33]. Similarly, customers purchasing biometric systems that include face detection components generally do not ask about detection performance. It is important to ensure that popular packages do not introduce avoidable performance issues into machine vision systems relying on face detection. Such issues can go unnoticed even in deployed systems especially when they only affect a minority group or an acceptable level of performance is achieved in aggregate.

3 Methods

3.1 Biometric Systems

Two kiosk-based face acquisition systems (System A and System B) were tested in a scenario test to determine if they were ready for broad deployment. The systems were already operational in limited deployments. Both systems returned a single face image for each subject standing in front of the system. System A directed all aspects of the subject interaction automatically. System B was staffed by an operator required only to initiate image capture and inform the subject when the process completed. The operator did not provide any information to the subject or answer any of the subject's questions. Markers on the ground indicated the appropriate standing location in front of the systems. Each subject used System A followed by System B.

3.2 Subjects and Demographics

Systems were tested with a demographically varied population of 624 volunteer subjects (Figure 1). All subjects consented to participate in the study under an established Institutional Review Board (IRB) protocol. Age, gender, height, and race were self-reported during study enrollment. We use the term sex to refer to male and female subjects. Presented race categories included: "American Indian or Alaska Native", "Asian", "Black or African American", "Hispanic or Latino", "Multi", "Other", and "White" in accordance with the U.S. Office of Management and Budget categories (pre-2024 update) [27]. Due to limited sample sizes in some self-reported race categories, race was recategorized to "Black" (B), "Other" (O), and "White" (W).

Skin lightness was measured using a calibrated colorimeter (cyberDERM, DSM III). Measurements were collected from the left and right temples of each subject and averaged. Skin lightness was extracted as the L^* component of CIELAB color (Figure 1E).

Two binary covariates known to affect biometric systems, head coverings and glasses, were analyzed. Head covering was defined to be yes (Y) if a subject's head was obstructed in any fashion when

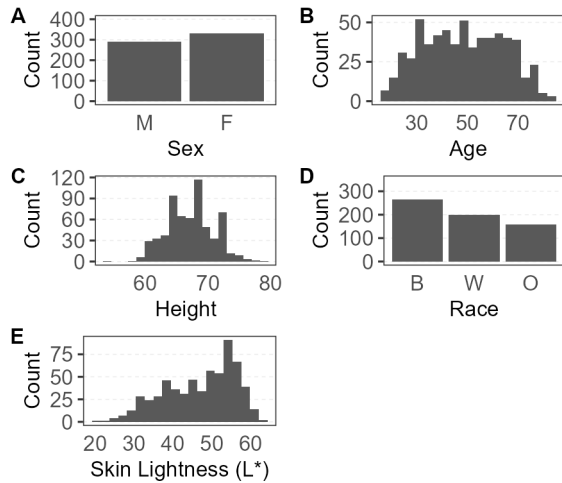


Figure 1: Distributions of subject self-reported and same-day measured demographics. A. Counts of subject self-report sex: (M) Male; (F) Female. B. Distribution for subject self-reported age in years. C. Distribution of subject self-reported height in inches. D. Counts of subject self-reported race. Due to low sample sizes, subjects with responses other than (B) Black or (W) White are grouped into a general (O) "Other" category for analysis. E. Distribution of subject same-day measured skin lightness (L^*).

Table 1: Counts of annotated variables: glasses and head covering. Counts of subjects wearing glasses or a head covering (Y) and not wearing glasses or not wearing a head covering (N) while interacting with System A and System B.

| System | Variable | Y | N |
|----------|---------------|-----|-----|
| System A | Glasses | 436 | 188 |
| System B | Glasses | 429 | 195 |
| System A | Head Covering | 545 | 79 |
| System B | Head Covering | 545 | 79 |

standing in front of the system; and glasses was defined to be yes (Y) if a subject was wearing a set of glasses while in front of the system. Two reviewers annotated images from each system with these two variables. If the subject's head was not fully in view in the image, videos of the subject's interaction with the system were reviewed. Reviewer annotation discordance for glasses and head covering were minimal. Less than 1% of images had annotation differences between reviewer 1 and reviewer 2 after initial review. All reviewer disagreements were resolved after a second joint review. Table 1 shows the counts for head covering and glasses. Ten subjects were excluded from analysis due to missing demographic data. Missing data was a random event unrelated to the performance of the tested systems.

3.3 Enrollment and Scenario Test

Prior to the start of testing, an enrollment system was used to collect high quality subject face images using quality criteria suitable for identity documents (enrollment transactions) [17]. Subjects stood in front of a gray background with a frontal pose relative to the camera. They were asked to remove any glasses or head coverings and maintain a neutral expression. These images represent an attended human-adjudicated enrollment use-case.

The scenario test was designed to replicate the intended high-throughput biometric identification use-case for the two acquisition systems. Subjects were briefed about the operational use-case (high-throughput biometric identification), but not about the technical details of each system. Subjects were asked to comply with all instructions presented by the systems. Each subject interacted with both systems. The systems attempted to collect a single face image for each subject (System A and System B transactions). The images had to pass an image quality check built into each system. Each subject transaction was recorded by the test infrastructure and associated with subject demographics, captured images, and the outcome of the quality check.

3.4 Datasets

Analyses were performed on three datasets of outcomes, images, and demographics associated with subject transactions with each biometric system: Enrollment, System A, and System B. The datasets contained a single image for each subject with the exception of System A, which failed to obtain an image for one subject. System B returned numerous images which failed to pass the built-in image quality check resulting in a failure to acquire outcome. All datasets used for modeling had complete observations.

3.5 Face Detectors

We examined three distinct face detectors. The first detector (OpenCV) utilizes the pre-trained¹ Haar cascade classifier implemented in OpenCV [36]. The second detector (Dlib) is Dlib's face detector based on the Histogram of Oriented Gradients feature descriptor combined with a linear Support Vector Machine [6]. This detector is accessible through Dlib's `get_frontal_face_detector()` function. The third detector (DNN) employs OpenCV's Deep Neural Network module, utilizing a pre-trained Single Shot Multibox Detector model with a ResNet-10 backbone. Model files² were sourced from a GitHub repository (<https://github.com/sr6033/face-detection-with-OpenCV-and-DNN>; last commit in 2018).

3.6 Performance Measures

We consider two measures for our analysis; failure to acquire rate (FTAR) and failure to detect rate (FTDR). FTAR was measured on imagery collected from three systems during the scenario test: System A, System B, and Enrollment. Each system consisted of both a face detector, f and a dataset, d ; FTAR for system $\{d, f\}$, $\gamma_{d,f}$, was computed as the proportion of subjects for which the system failed to capture or captured an image that did not pass the

¹We used the Haar cascade frontal classifier stored in `haarcascade_frontalface_default.xml`

²OpenCV DNN model files included `res10_300x300-ssd_iter_140000.caffemodel` and `deploy.prototxt.txt`

system's quality check, in accordance with the ISO/IEC 2382-37 convention.

FTDR was measured for three open-source face detectors using images from System A and Enrollment datasets. Each image was labeled face detected true if a valid bounding box was returned by the detector. Valid face bounding boxes were required to include both of the eyes, the nose and the mouth. Validity was established by two reviewers that independently reviewed the position of the bounding box relative to the image. Reviewers agreed on 100% of all reviewed images. Based on these results, FTDR for face detector f and dataset d , $\eta_{f,d}$, was computed as the proportion of images in each dataset for which the detector failed to return a valid face bounding box.

3.7 Statistical Analysis

To estimate demographic effects on $\gamma_{f,d}$ and $\eta_{f,d}$ we apply generalized linear modeling techniques using a logit link function, $g(\pi) = \ln(\frac{\pi}{1-\pi})$, to estimate the log odds of a failure to acquire/detect as a linear combination of our demographic covariates. We consider seven demographic variables: race, age, sex, height, skin lightness or L^* , head covering, and glasses. We normalized the continuous variables age, height, and L^* prior to fitting according to $z = (x - \mu_x)/\sigma_x$. We estimated model parameters β , using iteratively reweighted least squares (IRLS). The inclusion of higher order terms or interaction terms, which could lead to over-fitting, were not considered in this analysis. Given our set of subjects $j \in 1, \dots, n$, a dataset d (subsection 3.4), a face detection algorithm f (subsection 2.3), a full model for failures to acquire, $\gamma_{f,d,j}$, or failures to detect, $\eta_{f,d,j}$ is as follows. Let $\theta_{f,d,j} \in \{\gamma_{f,d,j}, \eta_{f,d,j}\}$, then

$$\theta_{f,d,j} = \beta_0 + \beta_1 \text{race}_j + \beta_2 \text{age}_j + \beta_3 \text{sex}_j + \beta_4 \text{height}_j + \beta_5 L^*_j + \beta_6 \text{headcovering}_{d,j} + \beta_7 \text{glasses}_{d,j} + \epsilon_{f,d,j} \quad (1)$$

We define an optimal model for a given d or given f as one that minimizes the Bayesian Information Criterion or $BIC = k \ln(n) - 2 \ln(\hat{L})$, where k represents the number of estimated parameters in the model, n represents the sample size, and \hat{L} represents the maximum value of the model's fitted likelihood. BIC measures the goodness of fit of the model while discouraging over-fitting with a penalty for increasing the number of model parameters. To obtain our optimal model, we apply a step wise procedure in both directions using the *step()* function in the R package MASS.

Model assumptions were checked for each model as follows. The linearity assumption for each model was checked graphically; multicollinearity was checked using the generalized variance inflation factors; outliers were found using Cook's distance. In the presence of any outliers or potential influential point(s), partial models were fit without the outlier(s) to compare any deviations from the full model. Modeling assumptions held for all models and no outlier was found to be influential in this analysis.

Models fit to the data were assessed by using both Tjur's pseudo R^2 or the coefficient of discrimination [35] and the common calculation for the area under the curve or the AUC. An additional sensitivity analysis of the optimal model coefficients was performed through the use of a Boruta feature selection algorithm. The algorithm provides an unbiased and stable selection of important

Table 2: Number of acquired and not acquired subjects on System A and System B and McNemar's test results.

| | Not Acquired (System B) | Acquired (System B) |
|---|----------------------------|------------------------|
| Not Acquired (System A) | 1 | 1 |
| Acquired (System A) | 43 | 579 |
| $\chi^2 = 38.205, df = 1, p = 6.37 \times 10^{-10}$ | | |

Table 3: System B maximum and minimum observed FTARs and the observed change in FTAR, $\Delta_{FTAR} = \max(FTAR) - \min(FTAR)$. Variables head covering (HC) and skin lightness (L^*) are abbreviated for spacing.

| Variable | Min FTAR Category | Max FTAR Category | Δ_{FTAR} |
|-----------|------------------------|------------------------|-----------------|
| Age Group | 4.40% 31-45 | 11.31% 61+ | 6.91 |
| Sex | 4.82% F | 9.59% M | 4.77 |
| Glasses | 3.26% N | 15.38% Y | 12.10 |
| HC | 4.59% N | 24.05% Y | 19.50 |
| Height | 5.73% [54,64] | 9.23% (70,79] | 3.50 |
| Race | 4.02% W | 9.40% B | 5.38 |
| L^* | 0.00% (53.7,55.2] | 17.74% [21.1,33.4] | 17.74 |

attributes from an information system [22]. All optimal model coefficients were found to be the most important coefficients selected in the Boruta algorithms. All statistical tests used a significance level of $p \geq 0.05$.

Net change in estimated probability, $\Delta_{\hat{\pi}}$, for a covariate was measured as the absolute difference between estimated probabilities, $\hat{\pi}$, for minimum and maximum observed covariate values holding other covariates constant: at average values for numeric covariates (age, height, skin lightness); and false for glasses and head covering.

4 Results

The FTAR for System A was significantly lower than for System B ($\gamma_A = 0.32\%$; $\gamma_B = 7.05\%$; Table 2) despite both systems operating on the same subjects. Just 2 subjects failed to be acquired on System A as compared with 44 on System B, a 20-fold difference in the performance of an operational face acquisition system.

4.1 Demographic Effects in Face Acquisition

4.1.1 FTAR Disaggregation. To understand the factors responsible for increased FTAR for System B, we disaggregated its FTAR based on available demographic factors (Figure 2). This analysis revealed large demographic differentials in FTAR for some demographic factors. Table 3 lists each demographic variable and Δ_{FTAR} the difference between the maximum and minimum observed FTAR. The largest differentials were related to head coverings and skin lightness.

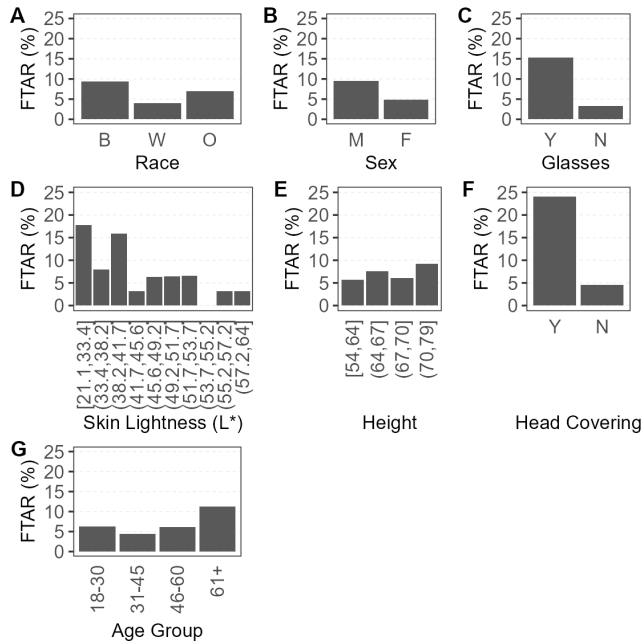


Figure 2: System B FTARs disaggregated across seven demographic factors. A. Disaggregated by self-reported race: (B) Black; (W) White, (O) Other. B. Disaggregated by self-reported sex: (F) Female; (M) Male. C. Disaggregated by wearing glasses (Y) and not wearing glasses (N) while interacting with System B. D. Disaggregated by same-day measured skin lightness L^* . E. Disaggregated by self-reported height. F. Disaggregated by wearing a head covering (Y) and not wearing a head covering (N) while interacting with System B. G. Disaggregated by self-reported age.

4.1.2 Modeling FTAR. The results presented in Table 3 do not account for the interdependence between factors. Categorical variables generalize several quantitative aspects of a subject. For example, sex tends to indicate average height, race tends to indicate average skin lightness, and glasses tend to indicate average age. In practice, people are more complicated. We estimated the distinct demographic effects on System B's FTAR using generalized linear modeling (subsection 3.7). This estimates the distinct impact of demographic factors on error rates while controlling for demographic effects from other factors.

We modeled the effects of seven demographic factors on FTAR. Starting with a full model including all seven covariates (1), we used a BIC-based model selection approach to find an optimal model including only those demographic covariates that improved model fit while minimizing the number of parameters. The optimal model retained four of the seven covariates: skin lightness L^* , head covering, glasses, and age. Selection of skin lightness, rather than race in the optimal model indicates that skin lightness was a better predictor of FTAR than self-reported race, similar to prior results observed in face recognition performance [4, 5]. Neither sex nor height were retained suggesting these factors did not independently predict FTAR for System B.

We computed the net change in estimated probability, or $\Delta\hat{\pi}$ (subsection 3.7) for each retained covariate. Table 4 shows the parameter estimates and $\Delta\hat{\pi}$. This analysis showed that, holding all else constant, the predicted FTAR for system B increased for subjects wearing glasses or head coverings respectively. Similarly, FTAR increased for older people and people with darker skin lightness values. It is important to note that the effects of glasses and head coverings could be mitigated in operational settings by directing users to remove those articles. On the other hand, the effects of skin lightness and age could not as these cannot be changed. Indeed, $\Delta\hat{\pi}$ for skin lightness was the largest observed differential, predicting that the system would fail for 8.03% of individuals with the darkest skin lightness as compared with only 0.45% of individuals with the lightest.

4.2 Face Detection Algorithm Performance

We hypothesized that a major contributor to the realized difference in performance between System A and System B is the face detector used by these systems. To test this hypothesis, we simulated 6 notional systems using the acquired Enrollment ($n = 624$) and System A ($n = 622$) images as a proxy for two acquisition systems and combined these with three face detectors (OpenCV, Dlib, DNN) to create the 6 simulated systems.

We reasoned that, if part of System B's performance issues lie with the face detector, we would see elevated FTAR using System A's images, with demographic effects similar to the measured FTAR for System B. Enrollment images were used to determine whether the effect was specific to the high-throughput use-case – a potential explanation for why acquisition issues were not addressed during earlier developmental testing. System B images were not included in this analysis because of the demographic imbalance in the acquired images (Figure 2). The images acquired by this system would comprise a sanitized dataset [38]. This was not an issue for the other datasets: images from Enrollment included all subjects; System A had a face detector with no apparent demographic differentials.

Table 5 shows the observed overall FTDR across the six simulated systems. All three simulated systems based on Enrollment images had near zero FTDR and thus no observed demographic differentials for these standardized, high-quality acquired face images (subsection 3.4). Two of the three simulated systems based on System A images had substantial FTDRs. The System A images with OpenCV face detection had the highest observed FTDR at 8.36% closely mirroring the observed System B scenario test performance of 7.05%. No failures to detect were observed using the DNN face detector.

4.2.1 Modeling FTDR. We repeated our linear modeling analysis (subsection 3.7) to understand the demographic factors that contributed to FTDR for our simulated systems. Of the six simulated systems, two systems: OpenCV and Dlib with System A images, had enough failures for this analysis. We modeled the effects of seven demographic factors on FTDR for both OpenCV and Dlib with System A's images. For both systems, the optimal model retained skin lightness and head covering. Glasses were uniquely retained for OpenCV whereas height was uniquely retained for Dlib. Interestingly, age was not selected by either model, despite being selected for System B's FTAR results, suggesting that age

Table 4: Parameter estimates for the optimal model, fitting failure to acquire from System B as discussed in subsection 3.7. Parameters not included in the optimal model are not shown (see Equation 1).

| Covariate | Estimate [SE] | Range | $\hat{\pi}(\min)$ [SE] | $\hat{\pi}(\max)$ [SE] | $\Delta_{\hat{\pi}}$ |
|------------------------------------|---------------|---------------|------------------------|------------------------|----------------------|
| Intercept | -4.23 [0.36] | NA | NA | NA | NA |
| Skin Lightness (L^*) | -0.63 [0.18] | (-2.87, 1.89) | 8.03% [0.04] | 0.45% [0.00] | 7.58% |
| Head Covering | 1.90 [0.38] | {0, 1} | 1.43% [0.01] | 8.89% [0.03] | 7.45% |
| Glasses | 1.92 [0.37] | {0, 1} | 1.43% [0.01] | 9.04% [0.02] | 7.60% |
| Age | 0.52 [0.19] | (-1.94, 2.23) | 0.53% [0.00] | 4.40% [0.02] | 3.86% |
| Tjur's $R^2 = 0.19$, $AUC = 0.84$ | | | | | |

Table 5: FTDR for the 6 simulated notional systems.

| Dataset | OpenCV | Dlib | DNN |
|------------|--------|-------|-------|
| Enrollment | 0.32% | 0.00% | 0.00% |
| System A | 8.36% | 4.02% | 0.00% |

effects for System B may not have been robustly due to the face detector. Finally, race was not retained by either model, reinforcing the observation that detection issues are better explained by skin lightness (likely due to its interaction with camera optics) rather than self-reported race categories. Figure 3 shows face images of subjects from the highest 10 and lowest 10 model predicted FTDR for OpenCV and Dlib along with each subject's covariate values.

Table 6 shows the parameter estimates for the factors retained in each model. This analysis revealed further differences between OpenCV and Dlib detectors. For the OpenCV system, skin lightness was the factor with the largest net effect ($\Delta_{\hat{\pi}}$) whereas for Dlib, the factor with the largest net effect was height. With the exception of age, the pattern of effect sizes for the OpenCV FTDR was similar to that observed for System B FTAR.

5 Discussion

This study shows how widely used, older methods in open-source packages can introduce demographic differentials into operational face recognition systems. Face detection is a common operation in computer vision applications that interact with humans, so much so that commodity face detection algorithms exist in popular computer vision packages. However, as shown here, these face detection algorithms may exhibit poor performance and significant demographic differentials. This highlights the need to test models including in open-source packages and to provide developers with results so they can make informed decisions regarding which model is appropriate for their intended use-case.

Our findings suggest two reasons why performance issues for System B were not detected in technology testing. First, the system may not have produced high error rates during technology testing, which can occur in a more controlled environment. Applying this detector to images from our enrollment use-case, for example, yielded error rates below 1%. Second, the system may not have produced high aggregate error rates during technology testing if the demographic makeup of users did not include varied demographic groups. For example, the system produced error


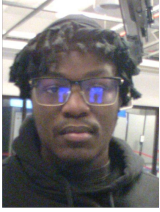

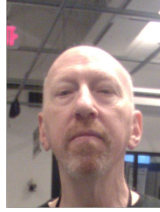
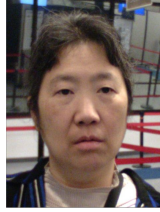
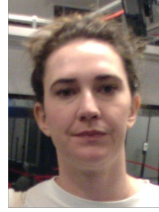
rates below 1% for people with the lightest skin tones. If people with lighter skin made up the majority of the population using the system during technology testing, aggregate performance may be high even if error rates were elevated for people with darker skin. Other demographic factors also impacted performance, such as the presence of head coverings. However, whereas people can be asked to remove head coverings, in some situations, to facilitate a face recognition process, they cannot change their skin tone or other innate characteristics. To maximize the likelihood of detecting issues prior to operations, prospective scenario tests should be conducted to simulate the real-world use-cases while also ensuring a sample of different demographic groups sufficient to measure performance differentials. In the case of biometrics, requirements for such testing have recently been standardized by the International Organization for Standardization [19]. Similar standardization for broader AI use-cases will benefit the ability to detect performance issues through testing.

In the biometric domain, most studies of AI performance focus on the face matcher, which compares two face images to determine if they show the same person or different people. This focus is understandable as face recognition is often considered the more challenging core problem to be solved in biometric systems. However, the main source of error in high-throughput biometric systems is often failure-to-acquire an image [4]. Errors and demographic differentials in face detection affect all downstream functions [14]. Our study shows that face detection can introduce and did introduce large demographic differentials into an operational system. This reinforces the notion that, in multi-component AI systems, errors and demographic differentials can be introduced by any one component, even if it is considered simple relative to others.

These observed differences between simulated systems highlight the need to evaluate operationally deployed systems as full systems and the need for scenario testing of AI systems as a critical component of AI governance. In our study, Dlib and OpenCV were well performing face detectors when assessed in combination with the study's enrollment process. However, changing the acquisition source to a deployed operational system, System A, we observed a large increase in the overall FTDR. Error rates rose from 320 instances out of 100,000 people to 8,360 instances out of 100,000 people, a 26-fold increase, when assessing OpenCV on a curated dataset versus the imagery produced by System A.

Inclusion of older methods in popular machine vision packages confers important benefits including backward compatibility, pedagogical value, enabling comparative studies, and encouraging rapid

Example Subjects with Estimated FTDR for OpenCV

| | | | | | | |
|---|---|---|---|--|---|---|
| |  |  |  |  |  |  |
| Detected | N | N | N | Y | Y | Y |
| Estimated FTDR: $\hat{\pi}$ | 79.26% | 76.86% | 76.62% | 0.42% | 0.37% | 0.34% |
| Skin Lightness L^* | 30.01 | 31.28 | 31.39 | 60.15 | 61.45 | 62.01 |
| Head Covering | Y | Y | Y | N | N | N |
| Glasses | Y | Y | Y | N | N | N |

Example Subjects with Estimated FTDR for Dlib

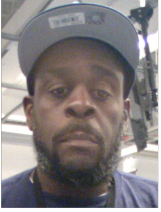
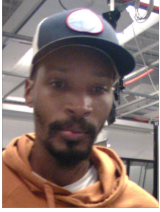


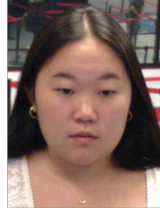

| | | | | | | |
|---|---|---|---|--|---|---|
| |  |  |  |  |  |  |
| Detected | Y | N | Y | Y | Y | Y |
| Estimated FTDR: $\hat{\pi}$ | 68.79% | 56.66% | 49.99% | 0.06% | 0.05% | 0.05% |
| Skin Lightness L^* | 32.11 | 39.63 | 35.29 | 53.47 | 56.80 | 61.91 |
| Height (inches) | 75 | 75 | 73 | 59 | 59 | 60 |
| Head Covering | Y | Y | Y | N | N | N |

Figure 3: Example subject face images from System A for subjects in the 10 highest and 10 lowest model estimated FTDR for OpenCV and Dlib. Subjects shown consented to have their images shared in scientific publications. Detected indicates actual detection outcome: Y - valid face detected; N - no valid face detected. The model estimated FTDR ($\hat{\pi}$) and corresponding covariate values are provided below the image of each subject. Note that model estimated FTDR is the predicted likelihood that subjects with similar values of the covariates would fail to be detected.

Table 6: Parameter estimates for the optimal models, fitting failure to detect images from System A as discussed in subsection 3.6 and subsection 3.7. Parameters not included in the optimal model are not shown (see Equation 1).

| Covariate | Estimate [SE] | Range | $\pi(\min)$ [SE] | $\pi(\max)$ [SE] | $\Delta_{\hat{\pi}}$ |
|------------------------------------|---------------|---------------|------------------|------------------|----------------------|
| OpenCV | | | | | |
| Intercept | -4.00 [0.33] | NA | NA | NA | |
| Skin Lightness (L^*) | -1.00 [0.17] | (-2.87, 1.89) | 24.34% [0.08] | 0.28% [0.00] | 24.10% |
| Head Covering | 1.46 [0.37] | {0, 1} | 1.79% [0.01] | 7.27% [0.03] | 5.48% |
| Glasses | 2.00 [0.35] | {0, 1} | 1.79% [0.01] | 11.88% [0.03] | 10.10% |
| Tjur's $R^2 = 0.22$, $AUC = 0.84$ | | | | | |
| Dlib | | | | | |
| Intercept | -4.56 [0.43] | NA | NA | NA | |
| Skin Lightness (L^*) | -0.63 [0.23] | (-2.87, 1.89) | 5.91% [0.04] | 0.32% [0.00] | 5.59% |
| Head Covering | 2.11 [0.46] | {0, 1} | 1.03% [0.00] | 7.94% [0.03] | 6.91% |
| Height | 1.07 [0.26] | (-3.52, 3.12) | 0.02% [0.00] | 22.80% [0.12] | 22.78% |
| Tjur's $R^2 = 0.13$, $AUC = 0.90$ | | | | | |

development and innovation. Indeed, using these older methods in some use-cases is sufficient to achieve a high level of performance. However, our analysis shows that some pre-trained DNNs can lead to better face detection performance.

Our results show that the pre-trained Haar cascade frontal face detector can be replaced by the pre-trained DNN detector which can also be readily implemented in OpenCV. A pre-trained DNN-based face detector is also available in Dlib. Here, we do not take a position regarding the specific merits of DNNs as a replacement for Haar cascades as each method has its own limitations. For example, DNNs require significant amounts of training data, which can have copyright and privacy implications, particularly in biometric applications, and can be more computationally intensive than other methods. However, when methods with superior performance are available and have a comparable ease of implementation, we suggest deprecating older and lower performing methods while ensuring that the higher-performing alternatives are included in standard packages. As DNNs and other pre-trained machine learning models proliferate for a variety of applications, a robust program is needed to test these systems and update standard packages with new higher performing methods. This will help system developers avoid preventable performance issues and demographic differentials in deployments.

Acknowledgments

This research was sponsored by the United States Department of Homeland Security's Science and Technology Directorate on contract number 70RSAT23CB00000003.

The paper authors acknowledge the following author contributions: Cynthia M. Cook, Laurie Cuffney, Yevgeniy B. Sirotin and John J. Howard conceived the work, generated face detection bounding boxes, developed and performed analyses, and wrote the paper; Jerry Tipton and Arun Vemury conceived the work and edited the paper.

This research was also supported by the multi-disciplinary staff of the Identity and Data Sciences Laboratory at the Maryland Test Facility and the Biometric and Identity Technology Center who planned and executed the scenario test.

References

- [1] Mehrdad Amini and Laleh Javidnejad. 2024. Legal Regulation of Biometric Data: A Comparative Analysis of Global Standards. *Legal Studies in Digital Age* 3, 1 (2024), 26–34.
- [2] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.
- [3] Christoph Busch. 2024. Challenges for automated face recognition systems. *Nature Reviews Electrical Engineering* 1, 11 (2024), 748–757.
- [4] Cynthia M. Cook, John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury. 2019. Demographic Effects in Facial Recognition and their Dependence on Image Acquisition: An Evaluation of Eleven Commercial Systems. *Transactions on Biometrics, Behavior, and Identity Science* 1, 1 (2019).
- [5] Cynthia M. Cook, John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury. 2023. *Demographic Effects Across 158 Facial Recognition Systems*. Technical Report. Technical report, DHS.
- [6] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 1. Ieee, 886–893.
- [7] Samuel Dooley, George Z. Wei, Tom Goldstein, and John P. Dickerson. 2022. Robustness Disparities in Face Detection. *arXiv:2211.15937 [cs.CV]* <https://arxiv.org/abs/2211.15937>
- [8] Paweł Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. 2020. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society* 1, 2 (2020), 89–103.
- [9] Belén Fernández Saavedra, Raul Sanchez-Reillo, Raúl Alonso Moreno, and Óscar Miguel Hurtado. 2010. Environmental testing methodology in biometrics. (2010).
- [10] International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC). 2022. Information technology – Vocabulary – Part 37: Biometrics.
- [11] Judy Wawira Gichoya, Kaesha Thomas, Leo Anthony Celi, Nabile Safdar, Imon Banerjee, John D Banja, Laleh Seyyed-Kalantari, Hari Trivedi, and Saptarshi Purkayastha. 2023. AI pitfalls and what not to do: mitigating bias in AI. *The British Journal of Radiology* 96, 1150 (2023), 20230023.
- [12] Patrick Grother, Mei Ngan, and Kayee Hanaoka. 2019. *Face recognition vendor test (fvrt): Part 3, demographic effects*. National Institute of Standards and Technology Gaithersburg, MD.
- [13] HL Gururaj, BC Soundarya, S Priya, J Shreyas, and Francesco Flammini. 2024. A Comprehensive Review of Face Recognition Techniques, Trends and Challenges. *IEEE Access* (2024).
- [14] Jacob A. Hasselgren, John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury. 2020. *A scenario evaluation of high-throughput face biometric systems: Select results from the 2019 department of homeland security biometric technology rally*. Technical Report, DHS.
- [15] John J. Howard, Andrew J. Blanchard, Yevgeniy B. Sirotin, Jacob A. Hasselgren, and Arun R. Vemury. 2018. An investigation of high-throughput biometric systems: Results of the 2018 department of homeland security biometric technology rally. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 1–7.
- [16] Marco Huber, Anh Thi Luu, Fadi Boutros, Arjan Kuijper, and Naser Damer. 2024. Bias and diversity in synthetic-based face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 6215–6226.
- [17] International Civil Aviation Organization (ICAO). 2018. *Portrait Quality – Reference Facial Images for Machine Readable Travel Documents (MRTDs)*. International Civil Aviation Organization. <https://www.icao.int/Security/FAL/TRIP/Documents/TR%20-%20Portrait%20Quality%20v1.0.pdf> Accessed: 2025-01-21.
- [18] International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC). 2006. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and framework. <https://www.iso.org/standard/81223.html> Accessed: 2025-04-14.
- [19] International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC). 2024. Information Technology – Biometric Performance Testing and Reporting – Part 10: Quantifying Biometric System Performance Variation Across Demographic Groups. <https://www.iso.org/standard/81223.html> Accessed: 2025-01-21.
- [20] Kimmo Kärkkäinen and Jungseock Joo. 2019. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913* (2019).
- [21] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on information forensics and security* 7, 6 (2012), 1789–1801.
- [22] Miron B. Kursa and Witold R. Rudnicki. 2010. Feature Selection with the Boruta Package. *Journal of Statistical Software* 36, 11 (2010), 1–13. doi:10.18637/jss.v036.i11
- [23] Pietro Melzi, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, Dominik Lawatsch, Florian Domin, and Maxim Schaubert. 2023. Gandifface: Generation of synthetic datasets for face recognition with realistic variations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3086–3095.
- [24] Adarsh Menon. 2019. Face Detection in 2 Minutes using OpenCV & Python. <https://towardsdatascience.com/face-detection-in-2-minutes-using-opencv-python-90f89d7c0f81> Accessed: 2025-01-21.
- [25] Onyemachi Joshua Ndubuisi, Gift Adene, Belonwu Tochukwu Sunday, Chinedu Emmanuel Mbonu, and Adannaya Uneke Gift-Adene. 2024. Digital Criminal Biometric Archives (DICA) and Public Facial Recognition System (FRS) for Nigerian criminal investigation using HAAR cascades classifier technique. *World Journal of Advanced Engineering Technology and Sciences* 11, 2 (2024), 029–043.
- [26] U.S. Department of Homeland Security, U.S. Department of Justice, White House Office of Science, and Technology Policy. 2024. Biometric Technology Report. https://www.dhs.gov/sites/default/files/2024-12/24_1230_st_13e-Final-Report-2024-12-26.pdf
- [27] Office of Management and Budget (OMB). 1997. Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity. <https://www.govinfo.gov/content/pkg/FR-1997-10-30/pdf/97-28653.pdf> Federal Register, Vol. 62, No. 210, October 30, 1997. Accessed: 2025-01-21.
- [28] U.S. Commission on Civil Rights. 2024. The Civil Rights Implications of the Federal Use of Facial Recognition Technology. <https://www.usccr.gov/reports/2024/civil-rights-implications-federal-use-facial-recognition-technology>

- [29] European Parliament. 2024. Artificial Intelligence Act. Regulation (EU) 2024/1689. Published in the Official Journal of the European Union on July 12, 2024.
- [30] PyPI Stats. 2025. OpenCV-Python Statistics. <https://pypistats.org/packages/opencv-python> Accessed: 2025-01-21.
- [31] Parsa Rahimi, Christophe Ecabert, and Sébastien Marce. 2023. Toward responsible face datasets: modeling the distribution of a disentangled latent space for sampling face images from demographic groups. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 1–11.
- [32] Natassha Selvaraj. 2024. Face Detection with Python Using OpenCV. <https://www.datacamp.com/tutorial/face-detection-python-opencv> Accessed: 2025-01-21.
- [33] Sukanya Singh. 2023. A Step-by-Step Guide to Face Detection with the dlib Library. <https://medium.com/@sukanyasingh303/a-step-by-step-guide-to-face-detection-with-the-dlib-library-2e8f6429e632> Accessed: 2025-01-21.
- [34] Abhishek Tiwari, Suhail Manzoor, Jiya Sehgal, and Ashutosh Mishra. 2024. A Comprehensive Review of Face Detection Technologies. In *2024 Second International Conference on Advances in Information Technology (ICAIT)*, Vol. 1. IEEE, 1–6.
- [35] Tue Tjür. 2009. Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination. *The American Statistician* 63, 4 (2009), 366–372.
- [36] Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, Vol. 1. Ieee, 1–I.
- [37] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. 2016. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5525–5533.
- [38] Yu Yang, Aayush Gupta, Jianwei Feng, Prateek Singhal, Vivek Yadav, Yue Wu, Pradeep Natarajan, Varsha Hedau, and Jungseock Joo. 2022. Enhancing Fairness in Face Detection in Computer Vision Systems by Demographic Bias Mitigation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) (*AIES '22*). Association for Computing Machinery, New York, NY, USA, 813–822. doi:10.1145/3514094.3534153
- [39] Seyma Yucer, Furkan Tektas, Noura Al Moubayed, and Toby Breckon. 2024. Racial bias within face recognition: A survey. *Comput. Surveys* 57, 4 (2024), 1–39.