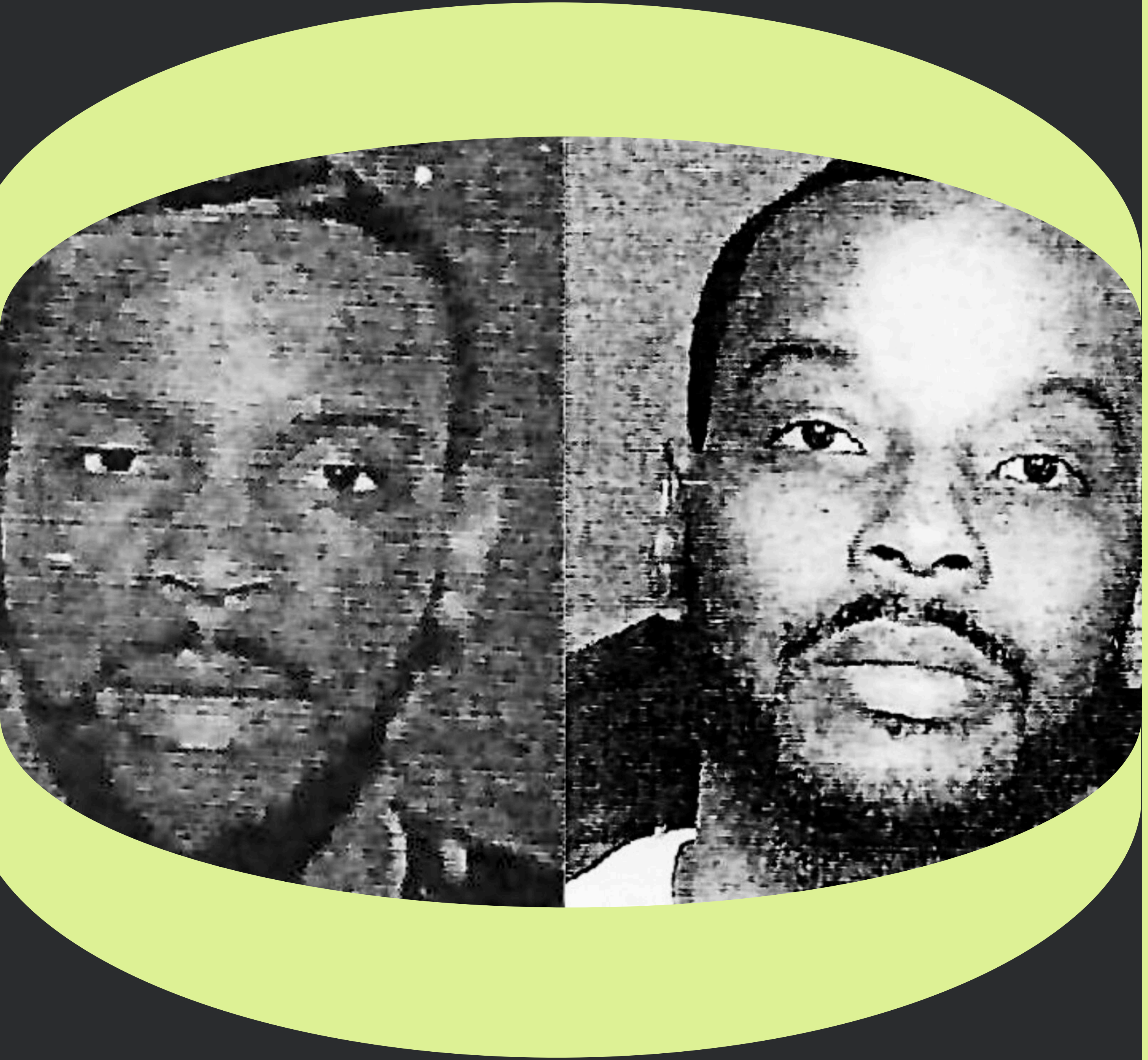
Demographic Disparities in Adversarial Robustness of Face Recognition Systems

By: Keerthana Jayamoorthy, Kaitlin Kartsen, Julian Mariscal, Madalyn Nguyen

A Course Final Project for CS 555: Responsible AI at Worcester Polytechnic Institute

Agenda



“A false facial recognition match sent this innocent Black man to jail” (CNN).

Problem

Face recognition systems exhibit well-documented accuracy disparities across demographic groups .

Facial recognition systems exhibit accuracy disparities across demographic groups:

- Higher False Match Rates (FMR) for Asian and African-American individuals compared to white individuals.
- Higher False Non-Match Rates (FNMR) for underrepresented racial groups, children, and the elderly.

Does adversarial vulnerability compound this existing harm?

Research Question One

H1 (ASR Disparity)

Do attacks (FGSM, PGD, C&W) succeed at different rates across demographics (Race, Gender, Age)?

Measurement	Compare attack success rate (ASR) across groups with statistical significance testing

Research Question Two

H2 (Perturbation Sensitivity)

Are groups more vulnerable at lower perturbation budgets ε?

Measurement	Compare ASR across groups at multiple perturbation budgets (ϵ = 0.01, 0.03, 0.05)

Research Question Three

H3 (Clean vs. Adversarial Gap)

Is the adversarial robustness gap larger, smaller, or comparable to the clean accuracy gap?

Measurement	Compare ratio of (worst group / best group) for clean accuracy vs. adversarial robustness

Research Question Four

H4 (Attack Method Consistency)

Do disparities persist across different attack methods?

Measurement	Compare demographic disparity magnitude across attack methods

What do you think will be the most significant factors?

POLL EV LINK

Methodology

Task

Face Verification (1:1 Matching)

Control

Created Impostor Pairs (two different people)

Target Model

FaceNet (Inception-ResNet-v1) from David Sandbergs Github

Dataset

FairFace from Hugging Face

2 genders

7 races

9 age ranges

Each race demographic dataset had approximately 90 photoes, 5 male and female from each age range

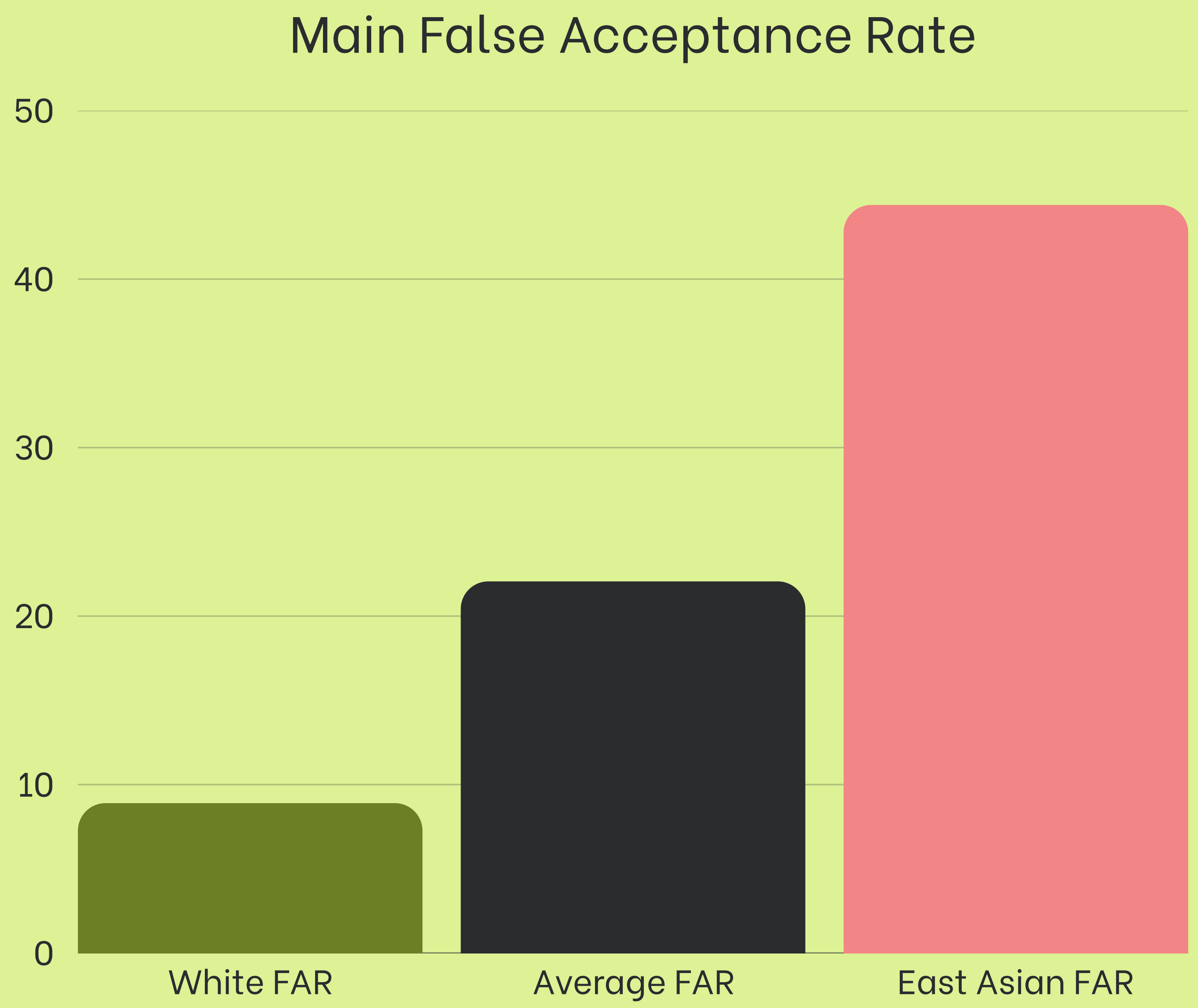
Pairs share the exact same Race, Gender, and Age.


```
pair_id,race,age,gender,image_A_index,image_B_index,image_A_age,image_A_gender,image_B_age,im
East_Asian_000000,East_Asian,0-2,Male,6277,399,0-2,Male,0-2,Male,0.8816535,True,1.0
East_Asian_000001,East_Asian,0-2,Male,6277,233,0-2,Male,0-2,Male,0.75424767,True,1.0
East_Asian_000002,East_Asian,0-2,Male,6277,7562,0-2,Male,0-2,Male,1.0445495,False,1.0
East_Asian_000003,East_Asian,0-2,Male,6277,2271,0-2,Male,0-2,Male,0.70869416,True,1.0
East_Asian_000004,East_Asian,0-2,Male,399,233,0-2,Male,0-2,Male,0.60981965,True,1.0
East_Asian_000005,East_Asian,0-2,Male,399,7562,0-2,Male,0-2,Male,0.7700919,True,1.0
East_Asian_000006,East_Asian,0-2,Male,399,2271,0-2,Male,0-2,Male,0.66403866,True,1.0
East_Asian_000007,East_Asian,0-2,Male,233,7562,0-2,Male,0-2,Male,0.8871743,True,1.0
East_Asian_000008,East_Asian,0-2,Male,233,2271,0-2,Male,0-2,Male,0.8248079,True,1.0
East_Asian_000009,East_Asian,0-2,Male,7562,2271,0-2,Male,0-2,Male,0.70480096,True,1.0
```

Baseline False Acceptance Rate

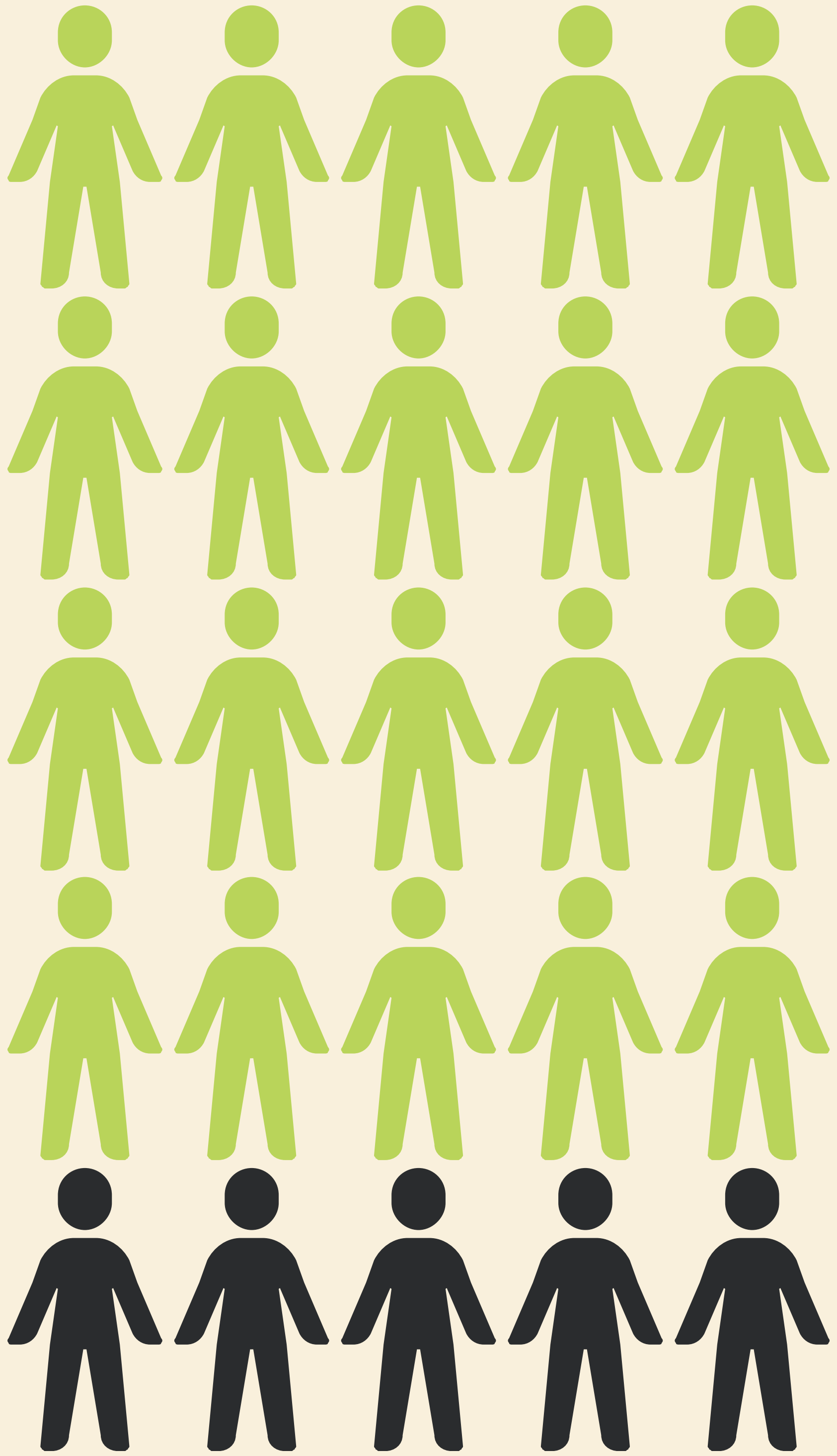
Racial Group	Mean False Acceptance Rate (FAR)	Standard Deviation	Vulnerability Status
East Asian	44.40%	0.335	Highest Error Rate
Southeast Asian	30.00%	0.298	High Error
Black	23.30%	0.295	Moderate Error
Indian	23.30%	0.313	Moderate Error
Middle Eastern	13.80%	0.28	Low Error
Latino_Hispanic	9.50%	0.212	Low Error
White	8.90%	0.241	Lowest Error Rate

Baseline Stats



Disparity Ratio

1 : 4.99



Approximately 22.06% of pairs of people are being incorrectly considered as a match

Adversarial Attack Setup

Metric Measured

Attack Success Rate

Attack Method	ϵ=0.01	ϵ=0.03	ϵ=0.05
FGSM	Dataset 1	Dataset 2	Dataset 3
PGD	Dataset 4	Dataset 5	Dataset 6
C&W	Dataset 7	Dataset 8	Dataset 9

FGSM

Fast Gradient Sign Method

applies a single, small perturbation in the direction that most increases similarity between two face, used as a computationally efficient, weak baseline.

PGD

Projected Gradient Descent

An iterative attack (10-20 steps), generally considered stronger and more robust than FGSM

C&W

Carlini & Wagner

An optimization-based attack designed to find the smallest possible perturbation that causes a misclassification, representing the strongest available threat mode

Designing a Three Way Analysis of Variance (ANOVA) Test to Measure Significance

Select Target Dataset

Used the PGD Attack dataset with ε=0.03.

the standard, most robust, and realistic benchmark for adversarial testing

Define Factors

Established three independent factors: Race (7 groups), Gender (2 groups), and Age (9 groups).

Calculate ANOVA Statistics

Calculated Sum of Squares (SS), Mean Squares (MS), and the F-ratio for all factors.

Example of p-value calculations for Race

Mean Square Residual($\mathbf{MS}_{\text{Residual}}$):

$$MS_{\text{Residual}} = \frac{SS_{\text{Residual}}}{df_{\text{Residual}}} = \frac{16.000}{124.0} \approx \mathbf{0.129}$$

$$\mathbf{MS}_{\text{Race}} = \frac{SS_{\text{Race}}}{df_{\text{Race}}} = \frac{0.8393}{6.0} \approx \mathbf{0.1399}$$

$$\mathbf{F}_{\text{Race}} = \frac{MS_{\text{Race}}}{MS_{\text{Residual}}} = \frac{0.1399}{0.1290} \approx \mathbf{1.084}$$

$$\text{p-value} = \mathbf{P}(\text{F-distribution with} df_1 = 6, df_2 = 124 \geq \mathbf{1.084})$$

p-value = 0.317

To determine the variance contribution and statistical significance (p-value) of each factor

Determine ASR Significance (H1)

Compared the calculated p-values to the alpha=0.05 threshold.

Baseline Significance Level Using ANOVA

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F (Ratio)	p-value	Statistical Conclusion (α=0.05)
Race	0.839	6	0.14	1.084	0.376	Not Significant
Gender	0.022	1	0.022	0.169	0.682	Not Significant
Age	11.524	8	1.441	11.164	8.5 x 10^-12	HIGHLY Significant
Race x Gender	0.706	6	0.118	0.912	0.488	Not Significant
Race x Age	8.329	48	0.174	1.345	0.099	Not Significant
Gender x Age	0.872	8	0.109	0.844	0.566	Not Significant
Race x Gender x Age	6.819	48	0.142	1.101	0.331	Not Significant
Residual (Error)	16	124	0.129			
Total (Model + Residual)	45.11	249				

Results of H1 (ASR Disparity)

Age HIGHLY Significant ($p < 10^{-11}$)

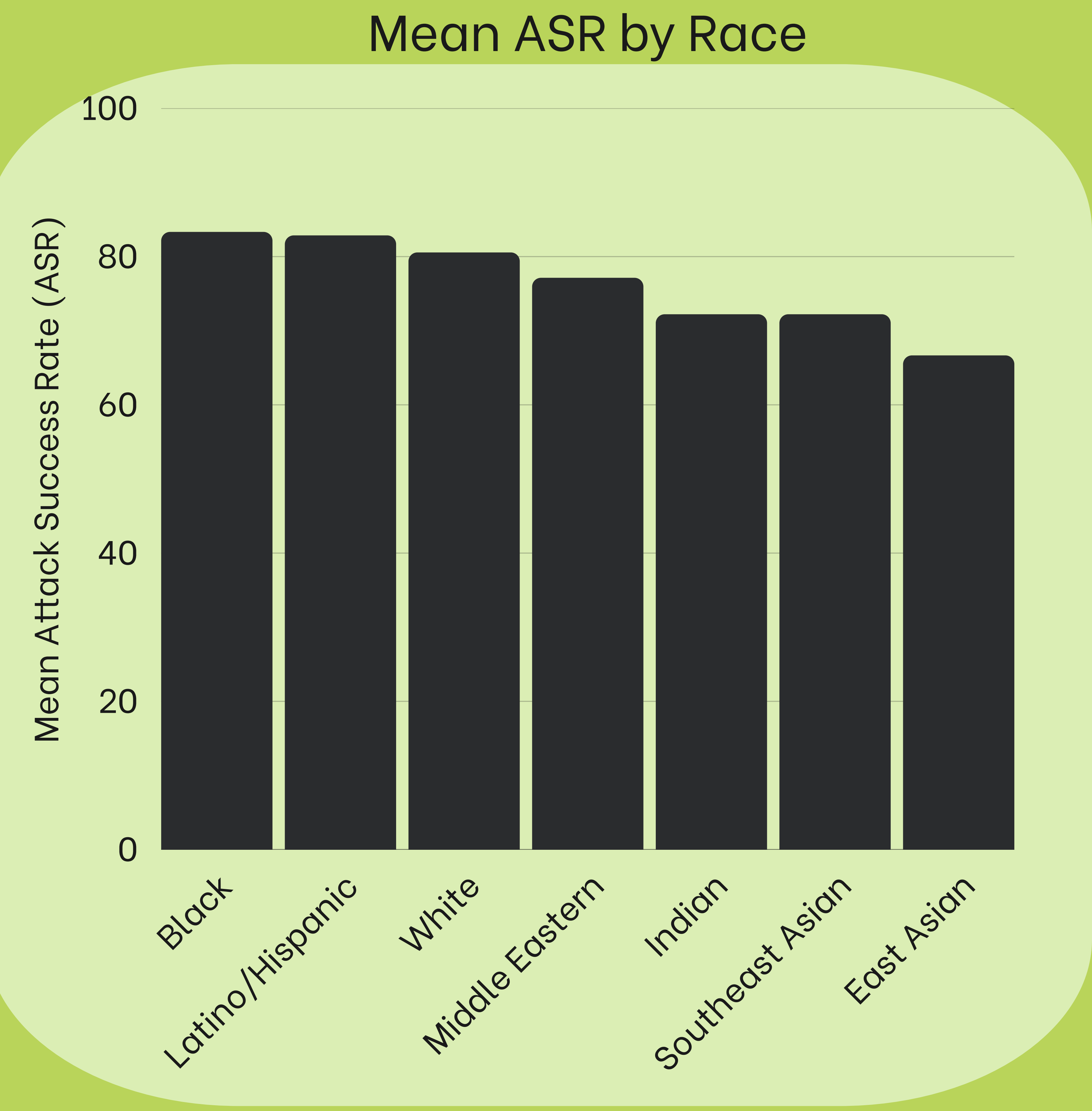
Race Not Significant ($p=0.376$)

Gender Not Significant ($p=0.682$)

Analysis

Though the model appears to perform disparately, race is not necessarily a statistically significant factor.

*Graph shows results for Projected Gradient Descent, ε=0.03



We assumed race would be our most significant factor

Results of H2 (Perturbation Sensitivity)

ASR Disparity Gap

The gap remains stable across all ε values. (Approx. 16.67% at ε=0.01 and 16.66% at ε=0.05).

Vulnerability Inversion

The most vulnerable group flips from East Asian (ε=0.01) to White (ε=0.05).

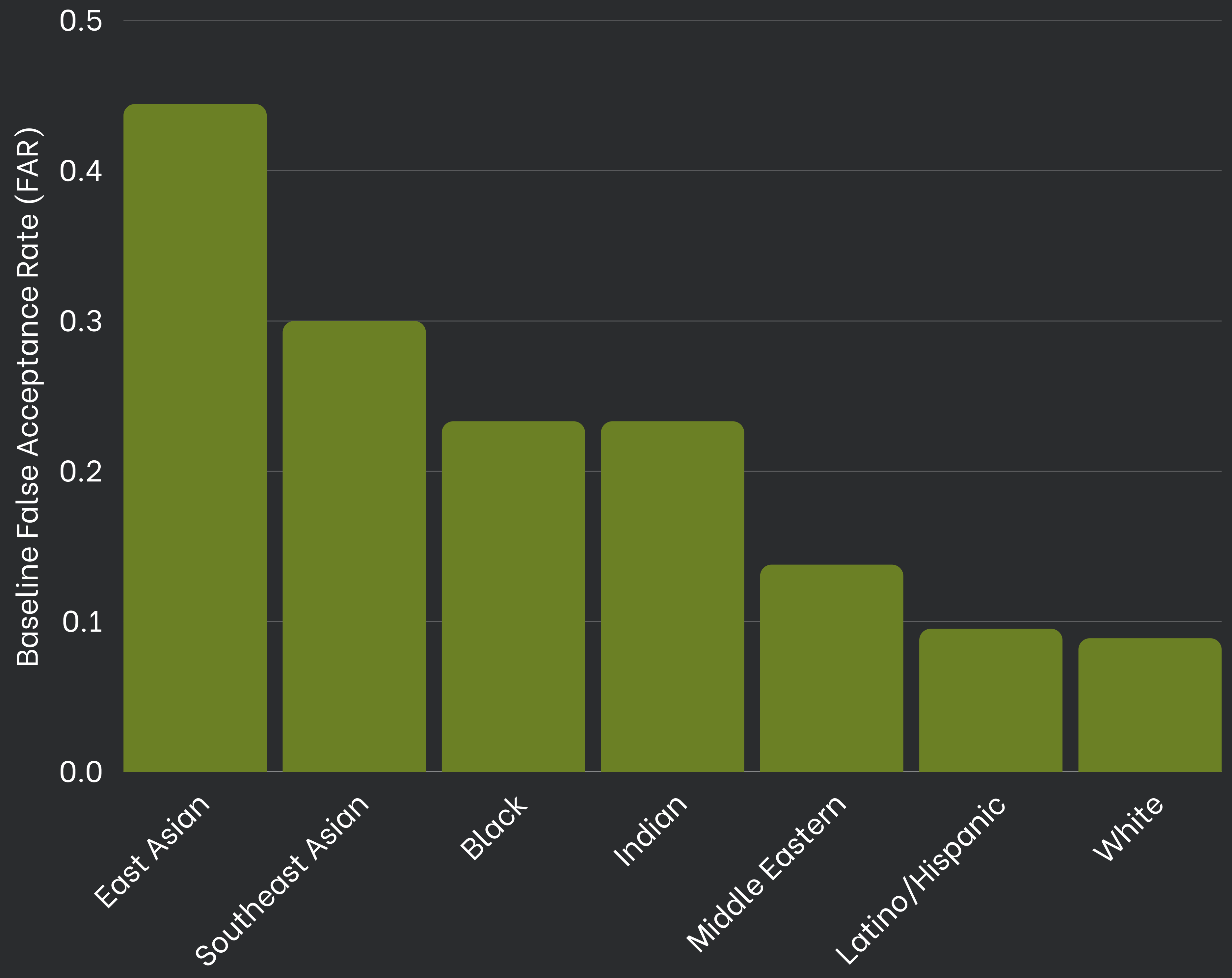
Analysis

These findings suggest that a model's robustness must be evaluated across a range of adversarial budgets to fully uncover hidden biases.

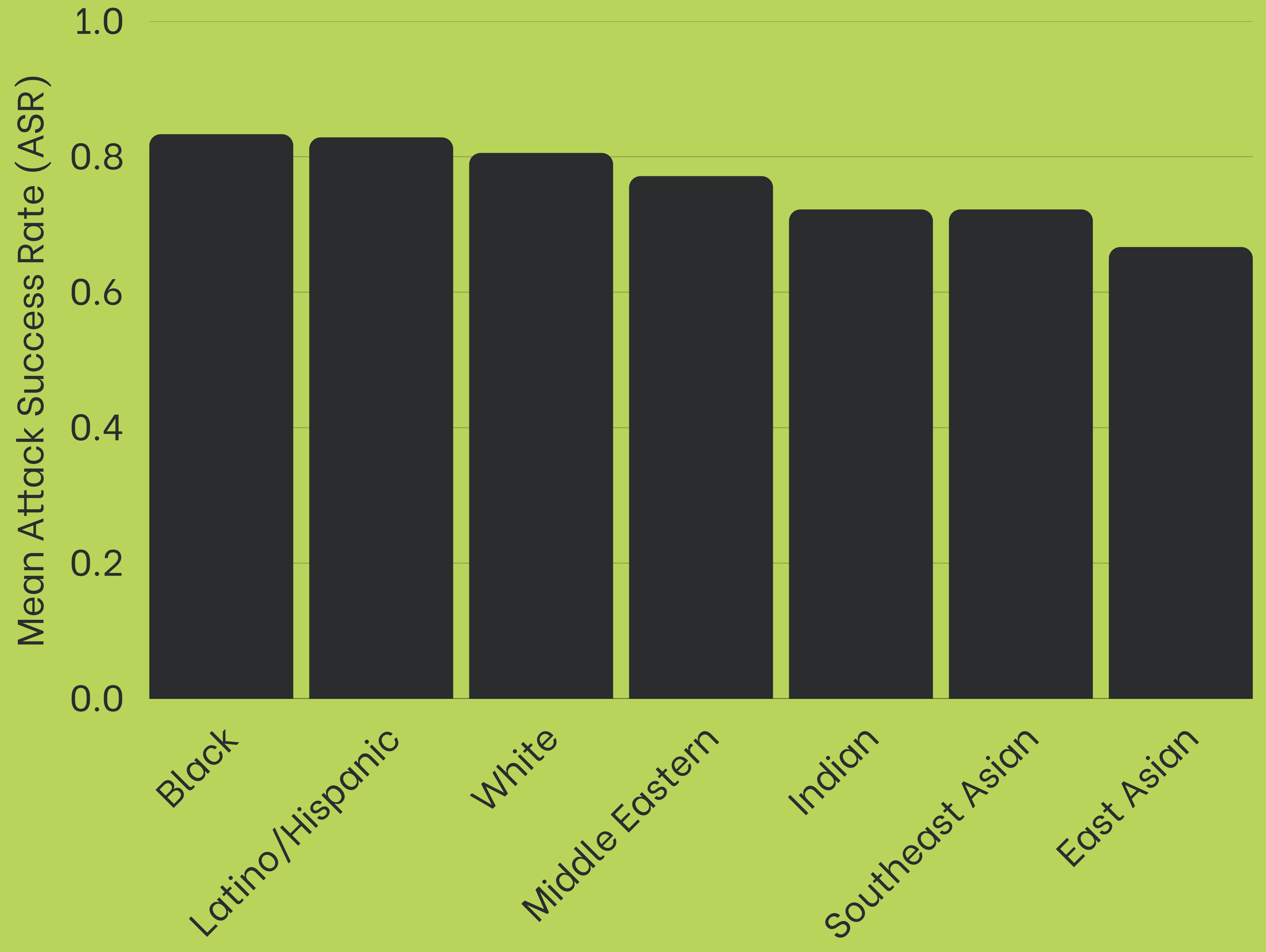
*Graph shows results for Projected Gradient Descent

This chart correctly compares the ASR for two extreme groups across different ε values.

Results of H3 (Clean vs. Adversarial Gap)



Baseline FAR by Race (Clean Bias)



Attack FAR by Race (PGD Attack, ε=0.03)

Results of H3

(Clean vs. Adversarial Gap)

De-Amplification Effect

Scenario	Disparity Ratio	Most Harmed Group
1. Clean Data Bias (FAR)	5.0:1	East Asian
2. Adversarial Vulnerability (ASR)	1.25:1	Black

Analysis

- The adversarial attack significantly DE-AMPLIFIES the racial disparity gap.
- The large 5.0:1 clean bias ratio collapses to a small 1.25:1 adversarial ratio.
- The group most harmed by clean bias (East Asian) is not the group most vulnerable to the attack (Black)
- Therefore, adversarial vulnerability does not compound the existing racial harm.

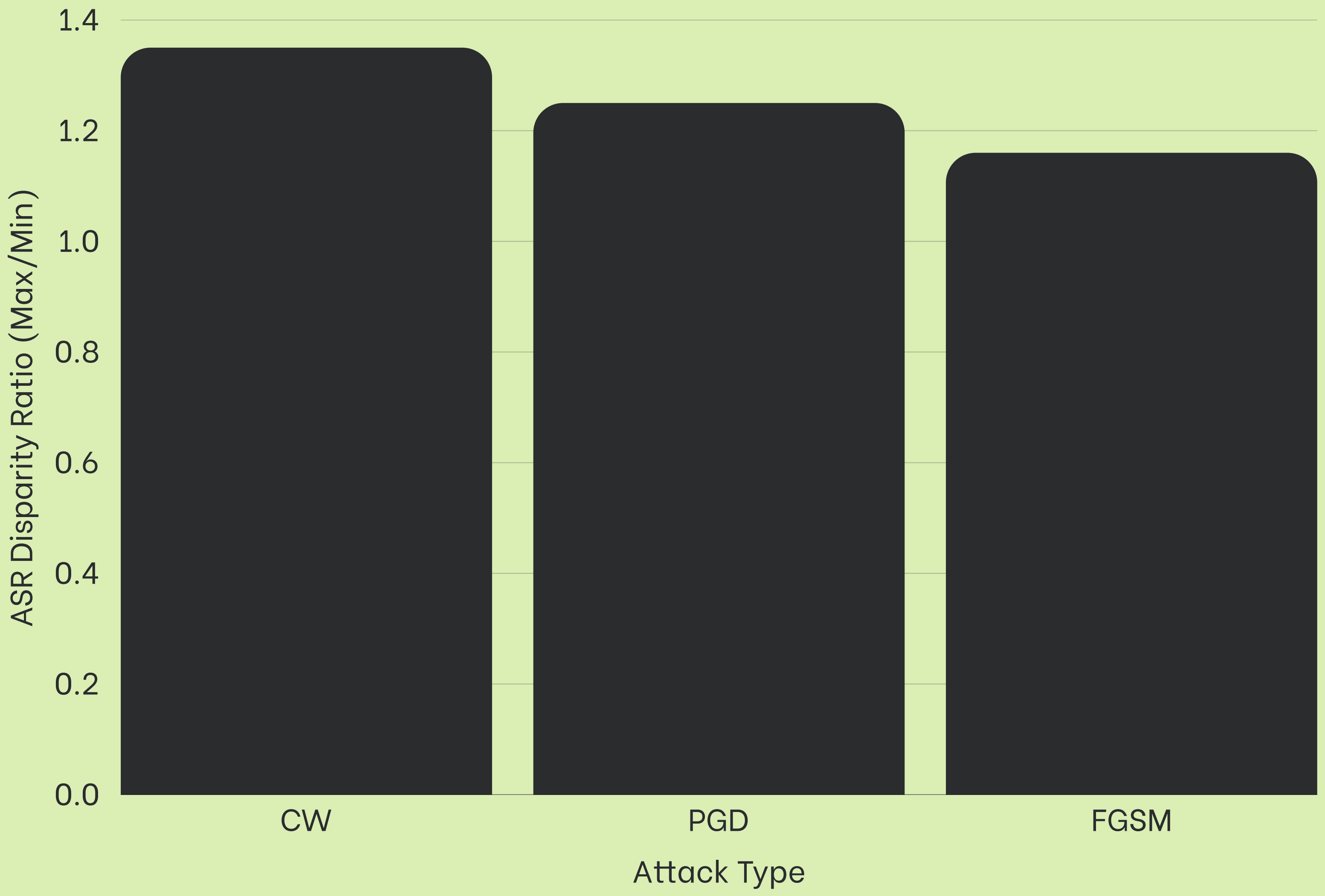
Results of H4 (Attack Method Consistency)

Analysis

The demographic disparity in adversarial vulnerability is low and consistent across all three attack methods.

- Maximum racial disparity: 1.35:1 (C&W)
- Minimum racial disparity: 1.16:1 (FGSM)

The small range of the disparity ratios suggests that the system's security flaw is broadly distributed across racial lines, regardless of which standard attack type is used.



This chart shows the ratio between maximum and minimum ASR for each type of attack.

Future Implications

Fairness Research

- Robustness should be treated as a dimension of fairness.
- Need multidimensional fairness evaluations.
- Opportunity for fairness-aware robustness metrics.

Security Evaluation & Threat Modeling

- Security benchmarks should include demographic breakdowns.
- Attackers could exploit demographic-specific weaknesses, especially in high-risk settings.

Policy, Governance & Deployment

- Current regulations miss disparities in adversarial vulnerability.
- Need benchmark datasets with demographic labels and robustness annotations.
- Certification should evaluate robustness across demographic groups before deployment.

Strengths

- **Robust Attack Methodology:** successfully applied and compared multiple state-of-the-art adversarial attacks (PGD, CW, FGSM) to test the model's security.
- **Comprehensive Disparity Analysis:** provided a thorough comparison of racial disparity across three key metrics: clean data bias (FAR), adversarial vulnerability (ASR), and attack type consistency.
- **Clear Evidence of Non-Compounding Harm (RQ3):** The finding that the attack significantly de-amplified the clean-data racial bias (the **5.0:1** ratio collapsing to **1.25:1**) is a strong, counter-intuitive result.

Weaknesses

- **Limited Statistical Significance of Race (H1):** The ANOVA test indicated that Race was not a statistically significant factor (**p=0.376**), which undercuts the primary focus on racial disparity in the paper.
- **Overwhelming Effect of Age:** ANOVA results showed that Age was the overwhelmingly significant factor ($p < 8.5 \times 10^{-12}$), suggesting a potential misfocus on race over a stronger demographic factor.
- **Disparity Stability (H2):** The analysis showed the disparity gap itself remained stable (around 16.67% difference) across all ε values, limiting the strength of the claim regarding gap dynamics.

Citations

Al-Dujaili, A., & Nwogu, I. (2022). FairSA: Sensitivity Analysis for Fairness in Face Recognition. Proceedings of Machine Learning Research.

Benedict, T. J. (2022). The Computer Got It Wrong: Facial Recognition Technology and Establishing Probable Cause to Arrest. Washington and Lee Law Review, 79.

Grother, P., Ngan, M., & Hanaoka, K. (2019). Face Recognition Vendor Test (FRVT), Part 3: Demographic Effects. NIST Interagency Report 8280, National Institute of Standards and Technology.

Kotwal, K., & Marcel, S. (2025). Review of Demographic Bias in Face Recognition. arXiv preprint.

Manis, E., Cahn, A. F., Akyol, N., & Magee, C. (2021). SCAN CITY: A Decade of NYPD Facial Recognition Abuse. Surveillance Technology Oversight Project (S.T.O.P.).

Najjar, H., Ronen, E., & Sharif, M. (2025). Sy-FAR: Symmetry-based Fair Adversarial Robustness. arXiv preprint.

Ngan, M., Grother, P., & Hanaoka, K. (2023). Face Recognition Vendor Test (FRVT) Part 8: Summarizing Demographic Differentials. NIST Interagency Report 8429, National Institute of Standards and Technology.

Sun, C., Xu, C., Yao, C., Liang, S., Wu, Y., Liang, D., Liu, X., & Liu, A. (2023). Improving Robust Fairness via Balance Adversarial Training. Proceedings of the AAAI Conference on Artificial Intelligence.

Thank You For Your Time!

GitHub Repo

Research Paper