

Three Detectors, Three Failure Profiles: Detector-Specific Demographic Bias in Public Deepfake Detection and Its Implications for Content Moderation in Latin America

José Ernesto Ramírez García, Angel Pineda Aguirre, Daniel Alberto Curiel Vargas

Universidad de Guadalajara · Independent researcher · Universidad de Guadalajara

jose.ramirez2861@alumnos.udg.mx · angelpdev@gmail.com · danielcuriel.vargas@gmail.com

With Apart Research

Track: AI Governance — System auditing and accountability · Region: Latin America

Abstract. Removing non-consensual intimate imagery and sexual deepfakes increasingly relies on automated deepfake detectors. If a detector fails more often on some demographic groups, those people may receive weaker protection from automated moderation systems. We run a controlled audit of three popular public detectors (two from prithivMLmods and dima806) on a public dataset of real (FFHQ) and synthetic (StyleGAN) faces, using the same dataset and an identical sampling protocol and labeling each face's apparent demographics with FairFace. The group with the highest false-negative rate changes across detectors - White, Asian and Black depending on the model - and only one detector shows a statistically significant racial difference while another shows a gender difference. Moreover, two models with strong advertised performance metrics (dima806: 0.99 accuracy; prithiv v2: ~0.92 precision) drop to near-chance under distribution shift at the default threshold, and one model's racial gap persists after lowering the threshold as its author recommends. A detector's fairness cannot be assumed or inherited: it must be audited per deployment, a governance need for platforms that operationalize content-removal obligations under frameworks such as Mexico's Ley Olimpia.

1. Introduction

Non-consensual intimate imagery, including sexual deepfakes, is a growing harm of generative AI: fabricating a fake sexual image of a person now takes minutes and costs almost nothing. In Mexico, the Ley Olimpia penalizes the distribution of intimate content without consent, but legal sanction arrives late against the speed of distribution, and an estimated 80%+ of cases go unreported. Actual protection depends on a barely visible link: platforms' automated moderation systems, which decide what gets removed and what does not.

This is where the AI safety problem appears. When a victim requests removal, an algorithm typically decides, analyzing the face without human context. If that detector lets through more deepfakes of one demographic group than another, people in that group are structurally less protected even though the law protects them equally. Facial-analysis research documents worse performance on darker skin tones and non-Western features, but evidence specific to deepfake detection is mixed, and almost none is Latin-America-focused. There is no standardized way to check whether deployed detectors protect the region's people equitably.

Research question: does a public deepfake detector make more errors —a higher false-negative rate (FNR)— on faces from certain demographic groups, and is that pattern consistent across detectors?

Our main contributions are:

1. An open, reproducible instrument that measures the demographic bias (per-group FNR) of any public deepfake detector from automatically labeled images.
2. A comparative audit of three current public detectors using the same dataset and sampling protocol, showing that the highest-FNR group, and the very presence of a significant difference, depend on the detector: the bias profile is detector-specific, not uniform.

3. Evidence that these detectors' advertised metrics do not transfer: two models with strong reported performance drop to near chance off their dataset, and one model's racial gap persists even after adjusting the threshold as its own author recommends.

2. Related Work

Deepfake detection relies on reference datasets such as FaceForensics++ [1]; here we use a public set of 70k real faces (FFHQ) and 70k synthetic ones (StyleGAN) [12]. On demographic bias in facial and detection systems, evidence is mixed: some work reports worse performance by skin tone or age, and recent work finds no systematic bias. To label apparent demographics we use FairFace [2], widely used because it includes a Latino/Hispanic category.

Resources designed to measure fairness with more careful labels exist: GBDF manually annotates gender on deepfake datasets [5]; AI-Face [4] provides million-scale demographic annotation and a multi-detector fairness benchmark; and Casual Conversations [6] contributes annotated skin tone and self-reported attributes. In technical governance, system auditing and accountability are among the field's open problems [3]. On the regional context, several works document digital gender-based violence and Mexico's Ley Olimpia [7,8,9], as well as reliance on each platform's procedures for content removal [10,11]. The gap we address: there is no comparative, Latin-America-focused audit of public detectors' demographic error rates; we provide one and show the pattern is detector-specific rather than stable across models.

3. Methods

Data. We use the public “140k Real and Fake Faces” set (real from FFHQ/Flickr, fakes generated with StyleGAN). Because of limited compute (free Google Colab), we evaluated samples rather than the full set: we drew 400 images per class via a fixed-seed shuffle and kept adults (filtering apparent age under 20), leaving about 778 faces per run after demographic labeling. The ground-truth label comes from the source folder. Importantly, these are whole synthetic faces, not face swaps or intimate content, so this is a controlled audit of detector behavior on synthetic faces rather than a direct simulation of sexual-deepfake moderation; we return to this external-validity limit in the Discussion.

Demographic labeling. For each image we estimate apparent demographics with FairFace via the DeepFace library: ethnicity (includes Latino/Hispanic), gender and age. Demographics are model-perceived, not self-reported; this is a documented limitation.

Detectors. We audit three public Hugging Face detectors without retraining, via the transformers image-classification pipeline at the default 0.5 threshold: Deep-Fake-Detector-Model (prithivMLmods, ViT; prithiv v1), deepfake_vs_real_image_detection (dima806, ViT) and Deep-Fake-Detector-v2-Model (prithivMLmods, ViT; prithiv v2).

Metrics. The primary metric is per-group false-negative rate (FNR): of the images that are deepfakes, the proportion the detector classified as real. A high FNR means more deepfakes slip through for that group. We also report per-group accuracy and AUROC, the FNR gap between the highest- and lowest-FNR group with a 95% bootstrap confidence interval, and a chi-square test of correctness by group. Because some groups have few deepfakes (Table 1), these tests are fragile; we treat them as exploratory and recommend exact tests (Fisher, permutation), stratified bootstrap and multiple-comparison correction for confirmatory work. As a robustness check, we re-ran dima at a 0.1 threshold, which its author recommends (at the user's own risk) due to concept drift.

Reproducibility and sampling. We release the scripts (extraction, labeling, detection and analysis) and the Colab notebooks. The three detectors were evaluated with the same dataset and an identical sampling procedure (fixed seed, 400 per class). However, file ordering is not guaranteed stable across sessions, so the exact image subset can vary between runs; we therefore treat the cross-detector comparison as exploratory and read per-group differences descriptively rather than as precise model-versus-model contrasts.

4. Results

Overall performance. On this data, prithiv v1 reaches about 81% accuracy (AUROC ~ 0.95), while dima and prithiv v2 fall to roughly 54% and 50% —near chance. The two models with the strongest advertised performance metrics (dima806: 0.99 accuracy; prithiv v2: ~ 0.92 precision) are the ones that generalize worst here at the default threshold; v1, with more modest advertised performance, holds up better. Published metrics do not predict out-of-distribution behavior.

Bias by ethnicity. The central finding is in Figure 1 and Table 1: the group with the highest observed FNR changes across detectors. For prithiv v1 it is the Black group (FNR 0.67), for dima the Asian group (0.745), and for prithiv v2 the White group (0.48). Only dima shows a statistically significant racial difference (gap 0.517, 95% CI [0.293, 0.715]; chi-square $p=0.0004$); in v1 ($p=0.10$) and v2 ($p=0.11$) it is not significant. The Latino/Hispanic group's position is equally inconsistent: mid-table in v1, among the lowest FNR in dima, and second-highest in v2. Given the small per-group counts in Table 1, we read these as observed differences, not definitive rankings.

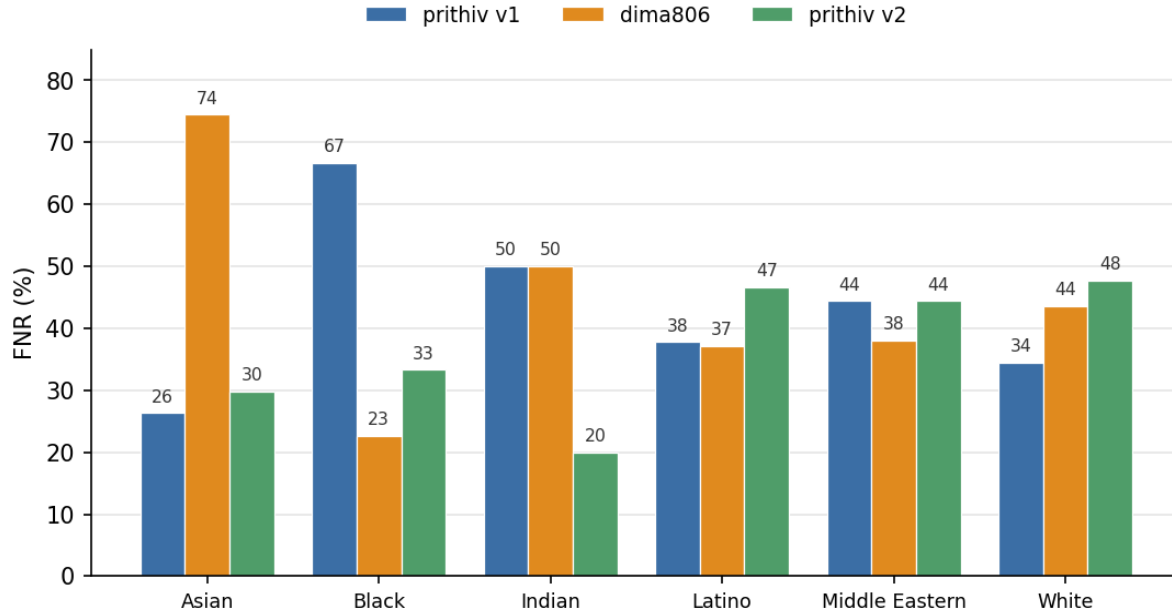


Figure 1. False-negative rate (FNR) by ethnic group for the three detectors (default 0.5 threshold). The group with the highest FNR changes across models: Black (prithiv v1), Asian (dima806), White (prithiv v2). Deepfakes per group (prithiv v1/v2 run): Asian 57, Black 12, Indian 10, Latino 45, Middle Eastern 27, White 235; the dima806 run used an independently drawn sample, so its per-group counts differ slightly. Small groups carry wide uncertainty.

Group	prithiv v1	dima806	prithiv v2
Asian	26% (57)	74% (47)	30% (57)
Black	67% (12)	23% (22)	33% (12)
Indian	50% (10)	50% (6)	20% (10)
Latino	38% (45)	37% (43)	47% (45)
Middle Eastern	44% (27)	38% (21)	44% (27)
White	34% (235)	44% (243)	48% (235)
Highest-FNR group	Black	Asian	White
FNR gap	0.40	0.52	0.28
95% CI	[.10, .68]	[.29, .72]	[.00, .49]
p (chi²)	0.10	0.0004	0.11
Overall accuracy	0.81	0.54	0.50

Table 1. FNR by ethnic group for the three detectors (default 0.5 threshold); in parentheses, the number of deepfakes per group. Counts are small for several groups, so we read these as observed differences rather than definitive rankings. Only dima806 shows a statistically significant racial difference; the highest-FNR group differs across detectors.

Bias by gender. Figure 2 shows that only prithiv v1 has a significant gender difference: men have a higher FNR (0.42 vs 0.30 for women; gap 0.125, 95% CI [0.034, 0.218]; $p=0.014$). In dima ($p=0.67$) and v2 ($p=0.39$) the gender difference is not significant. Again, the profile depends on the model: the detector that is significant by ethnicity is not significant by gender, and vice versa.

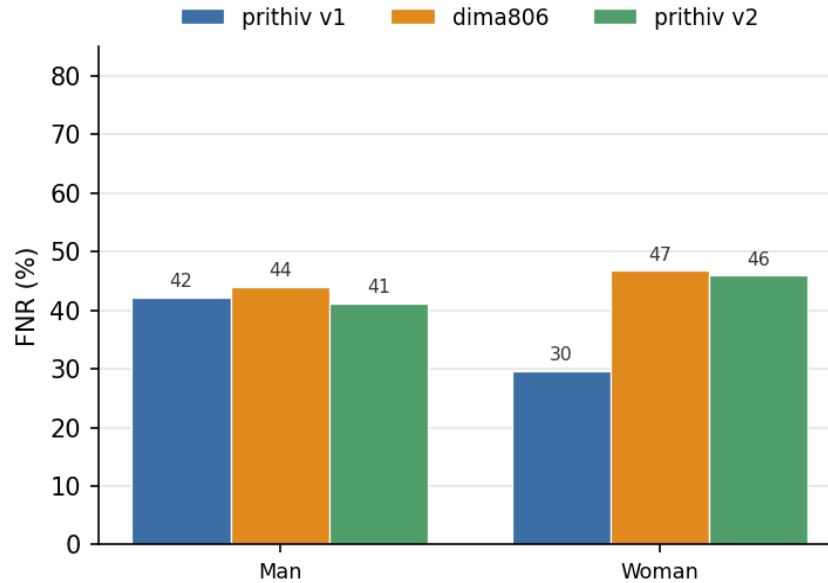


Figure 2. FNR by gender. Only prithiv v1 shows a significant difference (men higher FNR, $p=0.014$). Deepfakes per group (prithiv v1/v2 run): Man 190, Woman 196.

Threshold robustness. We re-ran dima at a 0.1 threshold, as its author recommends (Figure 3). At the 0.1 threshold, the model flags more inputs as fake and FNR is lower for most groups, although not all; however, accuracy still does not recover, remaining between 0.39 and 0.66 by group. The racial difference remains statistically significant in the 0.1-threshold run (chi-square $p=0.0001$), with the Asian group again among the highest FNR and the Black group the lowest. Dima's racial difference is therefore not a threshold artifact, and its advertised accuracy does not transfer even with the recommended fix.

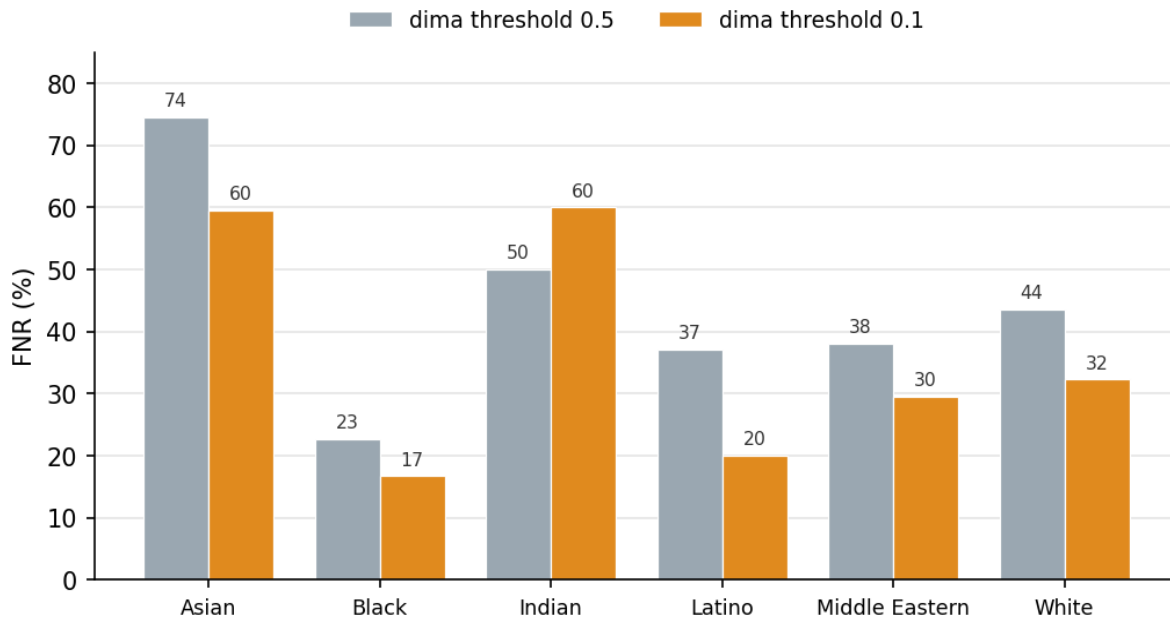


Figure 3. dima806 threshold robustness: lowering from 0.5 to 0.1 reduces FNR for most groups but does not recover accuracy, and the racial difference remains statistically significant in the 0.1-threshold run ($p=0.0001$). The two runs use independently drawn samples, so they are not directly comparable image by image.

5. Discussion

Two readings coexist. First, two of the most popular detectors are operationally unreliable at the default threshold under distribution shift - near-chance performance - a deployment risk in itself. Second, and more central: the demographic error is detector-specific. The group with the highest FNR flips (Black, Asian, White) and the presence of a significant difference depends on the model. Because prithiv v1/v2 and dima806 were not evaluated on an identical image manifest, we treat the cross-detector comparison as exploratory; even so, the magnitude of the observed differences and the fact that they point to different highest-FNR groups make the instability worth further investigation.

The implication for victim protection is direct: there is no “safe” detector one can adopt assuming fairness, and auditing one detector tells you little about the next. This creates a governance need for deployment-specific auditing in platforms that operationalize content-removal obligations under frameworks such as Mexico's Ley Olimpia. Platforms and authorities cannot rely on advertised metrics or a single benchmark; each deployed detector should be evaluated with demographically disaggregated error rates under its real operating conditions. Our theory of change: require disaggregated performance reporting before adopting a detector as a protection tool.

Limitations

Demographics are model-estimated (FairFace), not self-reported, and may contain errors. Because of limited Google Colab compute, we evaluated small samples rather than the full 140k set and filtered faces under 20; several groups therefore have few deepfakes (some fewer than 15), which makes the chi-square fragile and the per-group rates unstable, particularly Black and Indian. Exact tests (Fisher, permutation) and multiple-comparison correction are the appropriate next step. Sampling is not perfectly reproducible across sessions, so the runs used independently drawn samples and the cross-detector comparison is exploratory. The set uses whole synthetic faces (StyleGAN), not face swaps, which limits external validity to the non-consensual intimate imagery case; this is a controlled audit of detector behavior, not a simulation of moderation. Finally, the near-chance performance of dima and v2 means their per-group differences rest on weak discrimination, and everything is based on a single dataset.

Future Work

The natural next steps: fix an identical, reproducible sample across all detectors; scale to the full dataset and use larger, skin-tone-balanced data with more ethnic groups (e.g., AI-Face, GBDF, or Casual Conversations); add face-swap and in-the-wild deepfakes; manually validate a portion of the demographic labels and report agreement (kappa); sweep thresholds per model; and expand the number of audited detectors. This study is a fast, resource-limited first pass and is meant to be enriched with larger datasets.

6. Conclusion

Auditing three public deepfake detectors on the same dataset, the group with the highest observed false-negative rate is inconsistent and advertised metrics do not transfer. Fairness should be audited per deployment, not assumed. For Latin America, this means platforms that operationalize content-removal obligations under frameworks such as the Ley Olimpia need independent, disaggregated, deployment-specific auditing rather than reliance on a detector's advertised performance.

Code and Data

- Repository (prithiv v1 and v2): <https://github.com/angelpineda-clients/ley-olimpia>
- Repository (dima806): <https://github.com/angelpineda-clients/ley-olimpia-2>
- Colab prithiv v1: <https://colab.research.google.com/drive/1Jgb83eaRxEgjGrf5JHhMAaoyxnrhOOem>
- Colab dima806: <https://colab.research.google.com/drive/1xrff0lZKoLnEoaw-V7zQ4m7lBo9vHiMj>
- Colab prithiv v2: <https://colab.research.google.com/drive/179WbaqYInP8adlIKFFCz6QmVJTt5Q978>
- Colab dima806 (threshold 0.1): https://colab.research.google.com/drive/1oLddMnALSdF_Jg6YXv8Z0T57RcW8_nkp
- Dataset (Kaggle): <https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces>

Author Contributions

J.E.R.G. led the project, framed the question, and analyzed results. A.P.A. implemented the pipeline and ran the Colab experiments. D.A.C.V. supported demographic labeling, data verification, and writing. All three authors reviewed and approved the final report.

LLM Usage Statement

We used Claude (Anthropic) to structure the workflow, debug the pipeline, verify data integrity from the notebooks, and draft this report. All results were generated with our own code in Colab and verified by the team.

References

- [1] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. ICCV. arXiv:1901.08971.
- [2] Kärkkäinen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. WACV. arXiv:1908.04913.
- [3] Reuel, A., Bucknall, B., et al. (2025). Open Problems in Technical AI Governance. Transactions on Machine Learning Research (TMLR).
- [4] Lin, L., et al. (2025). AI-Face: A Million-Scale Demographically Annotated AI-Generated Face Dataset and Fairness Benchmark. CVPR.
- [5] Nadimpalli, A. V., & Rattani, A. (2022). GBDF: Gender Balanced DeepFake Dataset Towards Fair DeepFake Detection. arXiv:2207.10246.
- [6] Hazirbas, C., Bitton, J., Dolhansky, B., et al. (2021). Towards Measuring Fairness in AI: the Casual Conversations Dataset. IEEE Transactions on Biometrics, Behavior, and Identity Science.
- [7] Huang, W., Arora, P., & Zarzycka, M. (2026). Geographies of hope: Rethinking deepfake harms and gender AI safety in the Global South. Media and Communication, 14, Art. 10969. <https://doi.org/10.17645/mac.10969>
- [8] Escalera Silva, et al. (2024). Digital gender-based violence from the experience of actors in the delivery of justice. Sapienza, 5(1). <https://doi.org/10.51798/sijis.v5i1.741>

- [9] de la Vega, & Escalera Silva. (2025). Substantive analysis of digital violence in Mexico: Olimpia Law. *TransAmerica Review*, 3(1). <https://doi.org/10.62910/transame25003>
- [10] Bioni, B., Garrote, M., & Guedes, P. (2023). Key themes in AI regulation: The local, regional, and global. Data Privacy Brazil Research Association.
- [11] Li, et al. (2024). The sociotechnical stack: Opportunities for social computing research in non-consensual intimate media. *Proc. ACM Human-Computer Interaction (CSCW2)*, Art. 375. <https://doi.org/10.1145/3686914>
- [12] xhlulu. (2020). 140k Real and Fake Faces [dataset]. Kaggle.
- [13] Hugging Face models: prithivMLmods/Deep-Fake-Detector-Model; prithivMLmods/Deep-Fake-Detector-v2-Model; dima806/deepfake_vs_real_image_detection.