
Out of Distribution: Does Deepfake Detection Transfer to Real African Faces?¹

Ifihanagbara Olusheye
University of the Witwatersrand, Johannesburg

With
Apart Research

Abstract

Deepfakes are an escalating vector for disinformation, fraud, and image-based abuse, and African information ecosystems are a high-risk setting because such content spreads rapidly through closed, high-trust platforms like WhatsApp during sensitive periods such as contested elections or communal unrest. The technical defence is automated deepfake detection, yet these detectors are trained and validated almost entirely on Western data. We audit three off-the-shelf detectors across two distributions: FaceForensics++, their training distribution, which contains both real and fake faces, and real African faces from FAGE_v2 spanning ten countries, with FairFace as a Western proxy. On FaceForensics++ only the state-of-the-art detector, GenD, works, reaching an AUC of 0.92 and recall of 0.90, while the two freely available commodity detectors perform close to chance at 0.44 and 0.49. On real African faces, however, GenD, the one competent detector, flags 42.6% of genuine faces as fake. Across every detector, real African faces are flagged less often than the Western proxy, so the anti-African bias the literature predicts does not appear in that direction; the deployment hazard is instead a high absolute false-positive rate on real people. The decisive quantity, recall on African deepfakes, cannot be measured at all, because no African deepfake dataset exists. The finding for African AI safety is therefore twofold: the freely deployable detectors do not work even on Western benchmarks, and the one that does cannot be validated for African deployment, because the data needed to validate it has never been built.

¹Research conducted at The Global South AI Safety Hackathon, June 2026.

1. Introduction

Synthetic media has crossed the threshold from novelty to threat. Face-swap and generative video tools are now cheap and convincing enough to fabricate a politician’s confession, a fake official announcement, a fraudulent appeal in a trusted voice, or a compromising clip of a private citizen. In African contexts this is not hypothetical. Fabricated media already circulates during contested elections, periods of communal tension, financial scams that impersonate officials, and campaigns of image-based abuse that target private individuals and disproportionately harm women. News and rumour travel primarily through closed, high-trust channels such as WhatsApp, where a forwarded video carries the weight of personal endorsement and is difficult to fact-check at scale. A convincing deepfake released into that environment can move faster than any correction, and the mere possibility of deepfakes is itself corrosive, since any authentic but inconvenient recording can now be dismissed as fabricated.

The technical defence against this threat is automated deepfake detection, a classifier that takes a face image and decides whether it is real or synthetic. Such detectors are increasingly proposed for content-moderation and fact-checking pipelines. They are trained, benchmarked, and certified almost entirely on Western data, on the faces, cameras, lighting, and manipulation methods of datasets collected in the global North. Whether that competence carries over to the African information ecosystem, to real African faces captured on local devices and degraded by repeated re-compression through messaging apps, is assumed rather than tested.

We test it. We audit three off-the-shelf detectors on two distributions: their training distribution, FaceForensics++, which contains both real and fake faces and so permits full detection metrics, and real African faces from FAGE.v2, with FairFace as a Western proxy, both of which are real-only and so permit measuring the false-positive rate. **Our contributions are:**

1. We show that of three deployable detectors, only the state-of-the-art model, GenD, works on the standard benchmark at all, while the two freely available commodity detectors perform at roughly chance even on FaceForensics++.
2. We show that GenD, the one competent detector, flags 42.6% of real African faces as fake, a high absolute false-positive rate that would translate into real African people being wrongly accused of posting synthetic media.
3. We identify a structural gap. No African deepfake dataset exists, so detector recall on African deepfakes, the single most important quantity for deployment, cannot be measured. The defence cannot be validated for the population it is meant to protect.

2. Related Work

Fairness of deepfake detection has been studied since Trinh and Liu [1], who found detectors performed worst on darker-skinned subjects, with subsequent demographic audits [2, 3] framing the harm along a skin-tone axis using Western-sourced face data. Robustness work shows detectors degrade sharply outside academic benchmarks [4]. GenD [5] is a recent state-of-the-art detector that generalises across benchmarks via parameter-efficient adaptation of a CLIP encoder. Our work differs in *evaluation distribution and framing*: rather than measuring bias within a Western benchmark, we test whether detector competence transfers to real African faces (FAGE.v2 [6]), and we frame the result as a deployment-safety question for the African information ecosystem. The contribution others lack is the cross-distribution juxtaposition — in-domain detection metrics beside real-African false-positive rates and the explicit identification of the missing-data gap that makes African-deepfake recall unmeasurable.

3. Methods

Detectors. We evaluate three off-the-shelf image-level detectors at inference only, with no training. The first is GenD [5], a CLIP-Large detector that represents the current state of the art. The other two are freely available commodity detectors, a ViT model and a SigLIP model from the `prithivMLmods` collection on Hugging Face. Each detector maps a face image to a probability p_{fake} . Exact model identifiers are listed in the Code and Data section.

Distributions. We use two. The in-domain distribution is FaceForensics++ (c23), the detectors’ training distribution, which contains real videos (the `original` set) and five manipulation methods: Deepfakes, Face2Face, FaceShifter, FaceSwap, and NeuralTextures. We sample 100 real and 100 fake videos and extract face-cropped frames, giving a balanced real and fake set on which full metrics, namely accuracy, precision, recall, and AUC, are computable. The deployment distribution is FAGE_v2, 5,000 real African faces across ten countries, with FairFace and its seven race groups as a Western proxy. Both are real-only, so only the false-positive rate, the fraction of real faces wrongly flagged as fake, can be computed.

Why we cannot measure African recall. Recall, the fraction of real deepfakes a detector catches, requires fake examples. No public African deepfake dataset exists, and generating one was ruled out on ethical and time grounds. This absence is not a limitation of our method but a finding in itself: recall on African deepfakes is structurally unmeasurable with available data.

Pipeline. Every image passes through uniform MTCNN face detection and cropping, then each detector, then metrics. FaceForensics++ frames are extracted at fixed positions per video. The audit uses fixed thresholds and a released script, and the FaceForensics++ source is a public Kaggle mirror of the dataset.

4. Results

Only GenD works in-domain (Table 1, Figure 1). On FaceForensics++, GenD achieves accuracy 0.785, recall 0.903, and AUC 0.921. The two commodity detectors are close to chance: siglip_v1 AUC 0.436, vit_v2 AUC 0.493. The freely-available detectors a non-specialist would most plausibly deploy do not reliably detect deepfakes even on the standard academic benchmark.

Table 1: In-domain detection metrics on FaceForensics++ (balanced real/fake).

Detector	accuracy	precision	recall	AUC
GenD	0.785	0.730	0.903	0.921
SigLIP (v1)	0.452	0.455	0.483	0.436
ViT (v2)	0.498	0.497	0.303	0.493

GenD flags 43% of real African faces as fake (Table 2, Figure 2). On real faces, GenD’s false-positive rate is 0.480 on the Western proxy and **0.426 on real African faces**. It wrongly flags nearly half of genuine African faces as synthetic. Across all detectors the African FPR is *lower* than the proxy (all gaps negative), so the anti-African bias the literature predicts does not appear in that direction; the deployment hazard is the high absolute false-positive rate of the one competent detector.

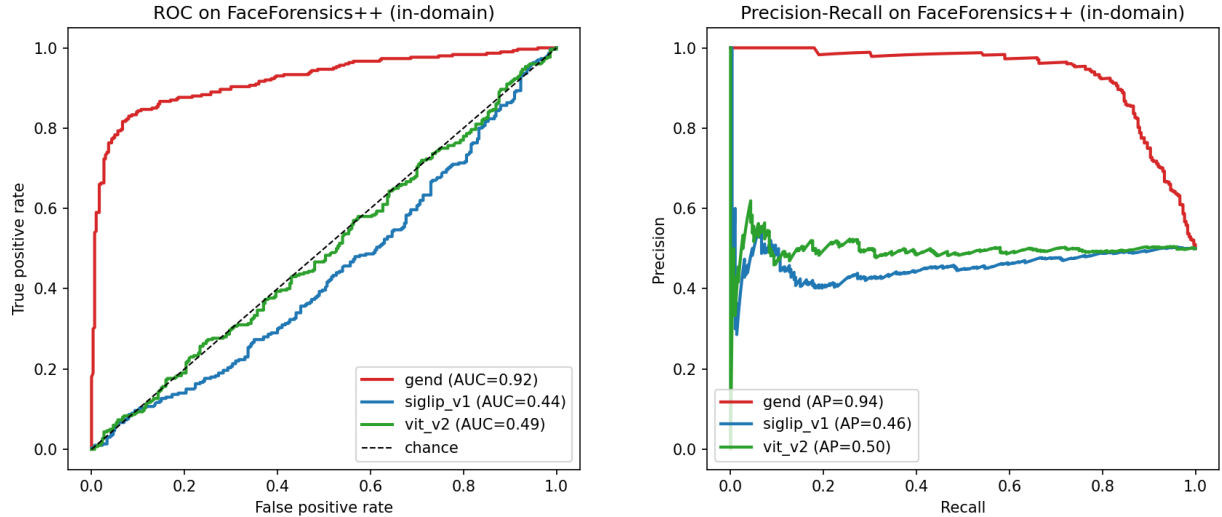


Figure 1: ROC (left) and precision-recall (right) on FaceForensics++. Only GenD separates real from fake; the commodity detectors track the chance diagonal.

Table 2: False-positive rate on real faces (real wrongly flagged as fake). Negative gap = African flagged less than the Western proxy.

Detector	proxy FPR	African FPR	gap
GenD	0.480	0.426	−0.054
SigLIP (v1)	0.090	0.038	−0.051
ViT (v2)	0.314	0.227	−0.086

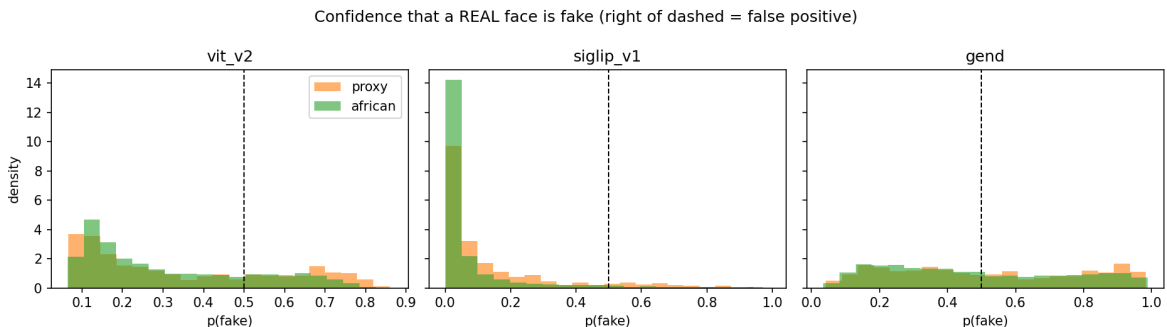


Figure 3: Distribution of $p(\text{fake})$ on *real* faces, by detector. Mass to the right of the dashed line (0.5) is real faces wrongly flagged as fake. GenD (right) spreads its confidence across the full range, placing large mass above 0.5; the commodity detectors concentrate near zero — they rarely flag anything, consistent with their near-chance in-domain performance. African and proxy distributions largely overlap.

The decisive number is the one we cannot compute. For deployment, the question that matters is recall on African deepfakes: if a real synthetic video of an African public figure circulated, would the detector catch it? That number is *unmeasurable* — no African deepfake dataset exists. We can establish that GenD catches 90% of *Western* fakes and mis-flags 43% of *real* African faces,

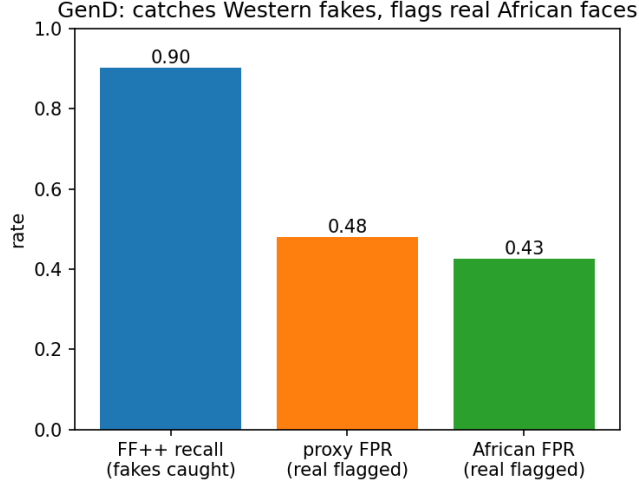


Figure 2: In-domain recall (fakes caught on FF++) beside real-face false-positive rates. GenD catches 90% of Western fakes yet flags 43% of real African faces.

but its ability to catch *African* fakes is unknown and currently unknowable.

5. Discussion and Limitations

Implications for the African information ecosystem. A fact-checking or moderation pipeline that deploys these tools in an African context faces two distinct failures. The first is availability. The detectors that are easiest to obtain do not work, so a moderator relying on a commodity detector receives near-random verdicts even on clean Western fakes. The second is reliability on real people. The one detector that does work mislabels real African faces at scale, and a false-positive rate of 43% means that for any genuine photo or video of an African citizen the system is nearly as likely as not to brand it synthetic. That outcome is itself a disinformation hazard, because declaring authentic material to be a deepfake is a well-known way to discredit real evidence. Underlying both failures is a third and deeper problem. The protective question, whether the system can catch a real African deepfake before it spreads, cannot be answered with existing data. The defence is therefore unavailable in its working form, unreliable on real people where it is available, and unvalidated against the threat it exists to stop.

Limitations. Several constraints shape how far these results generalise. FAGE_v2 faces are lower in resolution than FairFace, and resolution can move p_{fake} for reasons unrelated to a face being African, so the African false-positive rate should not be read as a pure demographic effect. We used a public Kaggle mirror of FaceForensics++ given the time available, and the official gated release should be used for any formal publication. On the African side we measure only the false-positive rate rather than recall, because no African deepfakes exist, which is the central limitation and at the same time the core finding. FAGE_v2 is web and celebrity imagery rather than deployment-captured media. Finally, because the two commodity detectors sit close to chance in-domain, their false-positive behaviour is harder to interpret than GenD’s, so we centre the deployment claims on GenD.

Future Work. The clearest next step is the dataset the field is missing, a responsibly constructed African deepfake benchmark with both real and synthetic faces and with capture-condition metadata, so that recall on African deepfakes can finally be measured. Beyond that, a resolution-matched evaluation would separate the effect of a face being African from the effect of image quality,

extension to diffusion-based manipulations would broaden the threat coverage, and integrating a confidence-aware referral step into a real moderation pipeline would test whether these detectors can be deployed safely with a human in the loop.

6. Conclusion

We audited three deepfake detectors across their Western training distribution and a set of real African faces. Only the state-of-the-art detector works in-domain, and it goes on to flag 43% of real African faces as fake, while the freely available detectors fail even on the standard benchmark. The most important deployment number, recall on African deepfakes, cannot be measured at all, because no African deepfake dataset exists. For the African information ecosystem this means the defence against deepfake-driven misinformation is at present either unavailable, unreliable on real people, or unvalidated against the threat. Closing the gap begins with building the African deepfake data the field has never had.

Code and Data

- Code: Public GitHub repository. [\[link\]](#)
- Data: FaceForensics++; FairFace; FAGE_v2.
- Models: GenD_CLIP_L_14; Deep-Fake-Detector-v2; deepfake-detector-v1.

References

- [1] L. Trinh and Y. Liu. An Examination of Fairness of AI Models for Deepfake Detection. *arXiv:2105.00558*, 2021.
- [2] Y. Ju et al. Improving Fairness in Deepfake Detection. *WACV*, 2024.
- [3] L. Lin et al. AI-Face: A Million-Scale Demographically Annotated AI-Generated Face Dataset and Fairness Benchmark. *CVPR*, 2025. *arXiv:2406.00783*.
- [4] N. Chandra et al. Deepfake-Eval-2024. *arXiv:2503.02857*, 2025.
- [5] A. Yermakov, J. Cech, J. Matas, M. Fritz. Deepfake Detection that Generalizes Across Benchmarks. *WACV*, 2026.
- [6] O. Ajewole et al. Unmasking the Uniqueness: Age-Invariant Face Recognition of Indigenous African Faces. *arXiv:2408.06806*, 2024.

LLM Usage Statement

We used Claude to help edit and debug the code pipeline. All results were generated by the released code and independently verified against the output CSVs; all claims were checked against the figures by the author.