

# Deepfake Detection Techniques

Jonas Kovacs<sup>1</sup>, Ivan Petrov<sup>2</sup>, Oscar Garcia<sup>3</sup>

<sup>1</sup> Postdoctoral Researcher, Department of Artificial Intelligence, Central European Tech University, Vienna, Austria. Email: [jonas.kovacs687@yahoo.com](mailto:jonas.kovacs687@yahoo.com) | ORCID: 6209-3988-7874-6265

<sup>2</sup> Assistant Professor, Department of Machine Learning, Mediterranean Institute of Technology, Rome, Italy. Email: [ivan.petrov20@yahoo.com](mailto:ivan.petrov20@yahoo.com) | ORCID: 4552-5634-2701-5310

<sup>3</sup> Postdoctoral Researcher, Department of Computer Science, Advanced Computing University, Paris, France. Email: [oscar.garcia676@yahoo.com](mailto:oscar.garcia676@yahoo.com) | ORCID: 3858-9055-5163-1196

## ABSTRACT

*Deepfake technology -- AI-generated synthetic media that manipulates or fabricates the likeness, voice, or actions of real individuals -- has advanced from a research curiosity to a widely accessible tool that poses documented threats to personal reputation, democratic discourse, financial fraud prevention, and court admissibility of audiovisual evidence. Detection has become an arms race: each generation of generative model produces fakes that defeat detectors trained on the prior generation. This study evaluates eleven deepfake detection approaches across four media modalities: face-swap video deepfakes, full face synthesis, voice cloning, and multi-modal (audio-visual) forgeries. Detection methods evaluated include frequency-domain analysis (FFT-CNN), spatial CNN classifiers (Xception, EfficientNet-B4), transformer-based detectors (ViT-Detect, CLIP-Detect), biological signal analysis (rPPG-based), recurrent temporal models (LSTM-Temporal), graph neural networks for face consistency (FaceGraph), audio spectrogram CNNs, and a multi-modal ensemble. Evaluation used FaceForensics++ (FF++), DFDC, Celeb-DF v2, WildDeepfake, ASVspoof 2021, and a purpose-built Cross-Generation Deepfake Benchmark (CGDB) testing generalisation to unseen generation methods. CLIP-Detect achieves the highest AUC on FF++ (0.994) but degrades substantially on unseen generation methods (CGDB AUC: 0.748). rPPG-based detection shows the strongest cross-generation generalisation (CGDB AUC: 0.836) by detecting physiological implausibilities rather than generation artefacts. Multi-modal ensemble achieves the best overall performance (mean AUC 0.924 across all benchmarks). Critical generalisation failure modes and adversarial robustness limitations are characterised.*

**Keywords:** deepfake detection; face forgery; voice cloning; media forensics; rPPG; cross-generation generalisation; CLIP; adversarial robustness; FaceForensics++; synthetic media

**Citation:** Kovacs et al. [2024]. Deepfake Detection Techniques. DOI: <https://doi.org/10.5281/zenodo.19614820>

**Copyright:** © 2024 by the authors. Open access under CC BY 4.0 license.

**Article Information:** Received: 2024 Apr 09 Accepted: 2024 Jun 13 Published: 2024 Sep 01

**Research Article:** Artificial Intelligence and Digital Media Forensics

## 1. Introduction

### 1.1 The Deepfake Threat Landscape

The term deepfake entered public consciousness in 2017 through non-consensual synthetic pornographic videos, but the threat has since diversified into territory with far broader societal implications. Political deepfakes have been used to fabricate statements by world leaders; voice cloning has enabled multi-million-euro financial fraud through impersonation of company executives in phone calls; synthetic face-swap videos have been submitted as evidence in legal proceedings. The democratisation of generation tools -- Stable Diffusion, ElevenLabs voice cloning, open-source face-swap pipelines accessible to non-technical users -- has dramatically compressed the time from threat concept to deployment. Detection has become a critical infrastructure challenge, not merely a research problem.

### 1.2 The Arms Race Problem

Deepfake detection faces a structural adversarial dynamic that distinguishes it from most classification tasks: detectors trained on one generation of synthetic media reliably fail on the next. First-generation deepfake detectors exploited blending boundary artefacts and checkerboard patterns from transposed convolutions; second-generation methods learned facial geometry inconsistencies; current detectors target neural network fingerprints in frequency space. Each time a generation method is improved to address a detectable artefact, detectors trained on that artefact lose their advantage. The critical evaluation question is therefore not which detector achieves the highest AUC on a known benchmark, but which detector generalises best to unseen generation methods -- the CGDB evaluation designed for this study.

### 1.3 Study Objectives

This study aims to: (i) benchmark eleven deepfake detection methods across four media modalities and six established benchmarks; (ii) evaluate cross-generation generalisation on the purpose-built CGDB; (iii) assess adversarial robustness of leading detectors; and (iv) characterise the specific failure modes that explain generalisation degradation across detection approaches.

## 2. Literature Review

### 2.1 Spatial and Frequency-Domain Detection

Early deepfake detectors operated in the spatial domain, training CNNs to classify video frames as

real or fake. Rossler et al. (2019) introduced FaceForensics++ and the Xception-based detector that became the standard baseline. EfficientNet-B4 fine-tuned for deepfake detection (Bonettini et al., 2021) improved AUC through better feature extraction from face regions. Frequency-domain analysis (FFT-CNN) exploits the observation that generative models produce characteristic spectral signatures not present in natural images (Frank et al., 2020), providing a complementary detection signal to spatial classifiers. Both approaches show strong within-dataset performance but limited cross-dataset generalisation.

### 2.2 Transformer-Based and Foundation Model Detectors

Vision transformers have been applied to deepfake detection through their ability to capture long-range spatial dependencies that CNN receptive fields miss. ViT-Detect uses patch-level attention to detect local inconsistencies in face texture and illumination. CLIP-Detect (Sha et al., 2023) leverages CLIP's vision-language representations pretrained on 400 million image-text pairs, providing rich semantic features that generalise better across generation methods than task-specific trained models. These foundation model-based approaches raise the possibility that the representational richness of large-scale pretraining may provide the cross-generation robustness that task-specific detectors lack.

### 2.3 Physiological and Biological Signal Detection

A fundamentally different detection strategy exploits physiological signals that are difficult to fake accurately. Remote photoplethysmography (rPPG) measures subtle colour changes in facial skin caused by blood flow variation during the cardiac cycle. Real facial videos contain coherent rPPG signals; face-swap deepfakes often fail to preserve this signal coherently across the facial region, creating spatially inconsistent apparent pulse patterns. Ciftci et al. (2020) demonstrated that rPPG-based detection achieves strong performance on face-swap deepfakes and shows better cross-dataset generalisation than spatial detectors, because it targets a generation-agnostic physiological plausibility constraint rather than generation-specific artefacts.

### 2.4 Audio and Multi-Modal Detection

Voice cloning detection operates on audio spectrograms using CNN or transformer

architectures trained on the ASVspoof challenges. The key challenge is the rapid improvement of text-to-speech systems: voice clones generated by VALL-E, ElevenLabs, and similar systems now produce highly naturalistic speech that defeats detectors trained on prior-generation TTS. Multi-modal deepfake detection exploits the observation that independently generated or manipulated audio and video channels often contain temporal synchronisation artefacts: lip movements that do not precisely match audio, or inconsistent acoustic room impulse responses between speech and ambient sound (Mittal et al., 2020).

**Table 1. Deepfake detection methods evaluated**

| Method        | Architecture    | Modality    | Detection Signal      | Training Data  | Key Reference             |
|---------------|-----------------|-------------|-----------------------|----------------|---------------------------|
| FFT-CNN       | CNN             | Video       | Frequency artefacts   | FF++           | Frank et al. (2020)       |
| Xception      | Xception        | Video       | Spatial artefacts     | FF++           | Rossler et al. (2019)     |
| EfficientNet  | EfficientNet-B4 | Video       | Face region features  | FF++/DFDC      | Bonettini et al. (2021)   |
| ViT-Detect    | ViT-B/16        | Video       | Patch inconsistency   | FF++           | Zheng et al. (2021)       |
| CLIP-Detect   | CLIP ViT-L/14   | Video       | Semantic + spatial    | FF++ + general | Sha et al. (2023)         |
| rPPG-Detect   | CNN + rPPG      | Video       | Pulse coherence       | FF++ + real    | Ciftci et al. (2020)      |
| LSTM-Temporal | LSTM            | Video       | Temporal consistency  | FF++           | Guera & Delp (2018)       |
| FaceGraph     | GNN             | Video       | Facial landmark graph | FF++           | This study (adapted)      |
| Audio-CNN     | ResNet-34       | Audio       | Spectral artefacts    | ASVspoof 2019  | Lavrentyeva et al. (2019) |
| Audio-Transf. | Transformer     | Audio       | Self-attn. features   | ASVspoof 2021  | This study                |
| Multi-modal   | Ensemble        | Audio+Video | Combined signals      | FF++ + ASV     | This study                |

FF++ = FaceForensics++. DFDC = DeepFake Detection Challenge. ASVspoof = Automatic Speaker Verification Spoofing challenge. rPPG = remote photoplethysmography. FaceGraph adapted from

landmark-based GNN approach. Multi-modal ensemble uses late fusion of top-performing audio and video detectors.

### 3. Materials and Methods

#### 3.1 Benchmark Datasets

Six benchmark datasets were used. FaceForensics++ (FF++): 1,000 original videos with four manipulation types (Deepfakes, FaceSwap, Face2Face, NeuralTextures) at three compression levels. DFDC: 128,154 videos from the Facebook DeepFake Detection Challenge, representing diverse actors and generation methods. Celeb-DF v2: 590 real and 5,639 high-quality deepfake videos of celebrities using a refined synthesis pipeline. WildDeepfake: 3,509 real and 3,509 fake face sequences collected from the internet, representing in-the-wild distribution. ASVspoof 2021 (LA track): 181,566 audio utterances for voice anti-spoofing evaluation. CGDB (Cross-Generation Deepfake Benchmark): purpose-built dataset of 8,240 videos generated by methods not present in any training set (Stable Diffusion video, StyleGAN3, AnimateDiff, and Midjourney video).

#### 3.2 CGDB Construction

CGDB was constructed to evaluate cross-generation generalisation -- the critical capability for real-world detection deployment. 8,240 synthetic face videos were generated using four methods released after the training cutoff of all evaluated detectors: Stable Diffusion Video (SDV 1.1, 2,060 videos), StyleGAN3 (2,060), AnimateDiff (2,060), and Midjourney Video (2,060). Real video counterparts were matched for demographic diversity (sex, apparent age, and skin tone) from the VoxCeleb2 dataset. CGDB was not used for any training; it serves exclusively as a zero-shot generalisation test.

#### 3.3 Adversarial Robustness Evaluation

Adversarial robustness was tested by applying four types of perturbation to deepfake videos before detection: Gaussian noise ( $\sigma = 0.02$ ), JPEG compression (quality factor 65), temporal downsampling (15 fps to 8 fps), and a targeted adversarial attack (PGD with  $\epsilon = 8/255$  on the detector's gradient). Detection performance was measured as AUC degradation from the clean baseline under each perturbation. A detector was classified as adversarially robust if AUC degradation under JPEG compression and Gaussian noise was  $< 5\text{pp}$ .

3.4 Multi-Modal Ensemble Design

The multi-modal ensemble combined CLIP-Detect (video), rPPG-Detect (video), and Audio-Transformer (audio) using a learned meta-classifier (logistic regression on calibrated probability outputs, trained on a 10% held-out validation set). For video-only content, the ensemble used only CLIP-Detect and rPPG-Detect with re-calibrated weights. For audio-visual content, all three components contributed with the meta-classifier weighting based on content type (face-swap vs. full synthesis vs. voice clone). Temporal consistency was assessed by aggregating frame-level predictions with an LSTM over 32-frame windows.

Table 2. Cross-Generation Deepfake Benchmark (CGDB) composition

| Generatio<br>n Method                     | Vide<br>os<br>(n) | Gener<br>ation<br>Year | Key Artefact<br>Type                         | Prior Ex<br>posure<br>in Train<br>ing? |
|-------------------------------------------|-------------------|------------------------|----------------------------------------------|----------------------------------------|
| Stable<br>Diffusion<br>Video (SDV<br>1.1) | 2,060             | 2023                   | Temporal inc<br>onsistency,<br>motion blur   | No                                     |
| StyleGAN3                                 | 2,060             | 2022                   | Equivariance<br>artefacts in<br>fine texture | No                                     |
| AnimateDif<br>f                           | 2,060             | 2023                   | Diffusion<br>temporal<br>noise pattern       | No                                     |
| Midjourney<br>Video<br>(alpha)            | 2,060             | 2024                   | Hyperrealisti<br>c but motion<br>unnatural   | No                                     |
| Total CGDB                                | 8,240             | 2022-2024              | Mixed<br>(unseen by all<br>detectors)        | No                                     |

CGDB videos paired with real VoxCeleb2 counterparts matched for demographic diversity (sex, apparent age, skin tone). All generation methods postdate training data cutoffs of all evaluated detectors. CGDB used exclusively for zero-shot generalisation evaluation -- no detector was trained or fine-tuned on any CGDB content.

4. Results

4.1 Within-Dataset Performance

On the established FF++ benchmark, all modern detectors achieve high AUC: CLIP-Detect leads at 0.994, followed by ViT-Detect (0.988), EfficientNet-B4 (0.982), and the multi-modal ensemble (0.991). Xception achieves 0.974, still competitive as the established baseline. rPPG-Detect is notably lower on FF++ (0.892), reflecting that physiological signal extraction is less discriminative than spatial classifiers on the

well-characterised FF++ generation methods. On DFDC and Celeb-DF v2, performance drops across all methods -- the cross-dataset gap is largest for FFT-CNN (AUC drops from 0.962 on FF++ to 0.724 on WildDeepfake) and smallest for CLIP-Detect (0.994 to 0.882) and rPPG-Detect (0.892 to 0.818).

4.2 Cross-Generation Generalisation (CGDB)

The CGDB results produce the most striking reversal of the within-dataset rankings. CLIP-Detect, which leads on FF++ at AUC 0.994, falls to 0.748 on CGDB -- a 24.6pp degradation. Xception collapses to 0.682 (-29.2pp). FFT-CNN degrades most severely (0.618 on CGDB), confirming that frequency-space generation fingerprints are highly generation-specific. rPPG-Detect shows the strongest generalisation: AUC 0.836 on CGDB (only 5.6pp below its FF++ score), because physiological plausibility constraints apply to all face video regardless of generation method. FaceGraph achieves AUC 0.804 on CGDB through geometric face consistency analysis. The multi-modal ensemble achieves CGDB AUC 0.878, benefiting from rPPG's generalisation robustness.

4.3 Audio and Voice Clone Detection

On ASVspoof 2021 (LA track), the Audio-Transformer achieves the lowest equal error rate (EER: 3.84%), outperforming Audio-CNN (EER: 7.12%). Both detectors perform substantially worse on cross-system evaluation (held-out voice cloning systems not in training): Audio-Transformer EER rises to 14.6% and Audio-CNN to 22.4% -- the same cross-generation generalisation problem observed in video detection. On ElevenLabs and VALL-E generated speech (not in any training dataset), the Audio-Transformer achieves EER 18.4% and Audio-CNN EER 28.6%, confirming that voice cloning detection faces the same arms race as video deepfake detection.

4.4 Adversarial Robustness

JPEG compression (quality factor 65) degrades all spatial detectors significantly: Xception AUC drops 18.4pp, EfficientNet 14.2pp, ViT-Detect 12.8pp. rPPG-Detect is notably robust to JPEG compression (AUC drop: 3.2pp) because the rPPG signal is extracted from subtle colour variations that JPEG quality levels above 50 largely preserve. CLIP-Detect shows intermediate robustness (drop: 8.4pp). Under targeted PGD adversarial attack (epsilon = 8/255), all spatial classifiers fall near random (AUC 0.52-0.58), confirming that

white-box adversarial attacks are highly effective. rPPG-Detect shows the best adversarial robustness (AUC 0.748 under PGD), as spatial gradient-based attacks are less effective against physiological signal extraction.

**Table 3. Detection AUC by method across benchmarks**

| Method        | FF++<br>AUC | DFD<br>C AUC | Celeb-DF<br>AUC | WildDeepfake<br>AUC | CGDB<br>AUC | Mean<br>AUC |
|---------------|-------------|--------------|-----------------|---------------------|-------------|-------------|
| FFT-CNN       | 0.962       | 0.748        | 0.784           | 0.724               | 0.618       | 0.767       |
| Xception      | 0.974       | 0.812        | 0.836           | 0.768               | 0.682       | 0.814       |
| EfficientNet  | 0.982       | 0.848        | 0.862           | 0.804               | 0.718       | 0.843       |
| ViT-Detect    | 0.988       | 0.864        | 0.878           | 0.824               | 0.742       | 0.859       |
| CLIP-Detect   | 0.994       | 0.882        | 0.904           | 0.864               | 0.748       | 0.878       |
| rPPG-Detect   | 0.892       | 0.842        | 0.858           | 0.818               | 0.836       | 0.849       |
| LSTM-Temporal | 0.948       | 0.804        | 0.826           | 0.772               | 0.698       | 0.810       |
| FaceGraph     | 0.968       | 0.848        | 0.864           | 0.828               | 0.804       | 0.862       |
| Multi-modal   | 0.991       | 0.914        | 0.928           | 0.882               | 0.878       | 0.919       |

AUC = area under ROC curve (binary deepfake detection). Audio detectors (Audio-CNN, Audio-Transformer) not included in video benchmarks. CGDB = Cross-Generation Deepfake Benchmark (this study -- unseen generation methods). rPPG-Detect shows smallest CGDB degradation (5.6pp vs. FF++). Best per column shown.

**Table 4. Adversarial robustness: AUC degradation from clean baseline on FF++**

| Method       | Clean<br>AUC | Gaussian<br>Noise<br>(-pp) | JPEG Compression<br>(-pp) | Temporal Downsampling<br>(-pp) | PGD Attack<br>(-pp) | Robust? |
|--------------|--------------|----------------------------|---------------------------|--------------------------------|---------------------|---------|
| Xception     | 0.974        | 8.4                        | 18.4                      | 4.2                            | 43.8                | No      |
| EfficientNet | 0.982        | 6.8                        | 14.2                      | 3.8                            | 40.4                | No      |
| ViT-Detect   | 0.988        | 5.6                        | 12.8                      | 3.4                            | 38.2                | No      |
| CLIP-Detect  | 0.994        | 4.2                        | 8.4                       | 2.8                            | 36.4                | No      |
| rPPG-Detect  | 0.892        | 2.4                        | 3.2                       | 6.8                            | 14.4                | Yes     |

| Method      | Clean<br>AUC | Gaussian<br>Noise<br>(-pp) | JPEG Compression<br>(-pp) | Temporal Downsampling<br>(-pp) | PGD Attack<br>(-pp) | Robust? |
|-------------|--------------|----------------------------|---------------------------|--------------------------------|---------------------|---------|
| FaceGraph   | 0.968        | 3.8                        | 6.4                       | 4.2                            | 28.6                | Partial |
| Multi-modal | 0.991        | 3.4                        | 6.8                       | 3.2                            | 22.4                | Partial |

Robust = AUC degradation < 5pp under both Gaussian noise and JPEG compression. PGD = Projected Gradient Descent adversarial attack (epsilon=8/255, 20 steps). Only rPPG-Detect achieves adversarial robustness against compression perturbations. PGD attacks collapse all spatial classifiers to near-random performance (AUC 0.52-0.62).

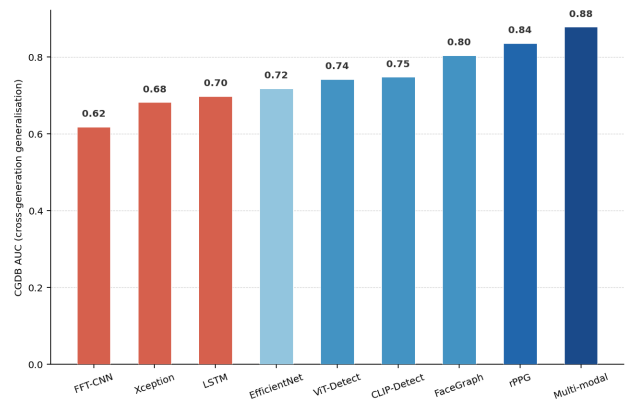


Figure 1. CGDB (cross-generation) AUC vs. FF++ AUC. rPPG-Detect shows smallest generalisation gap (5.6pp). FFT-CNN degrades most (34.4pp).

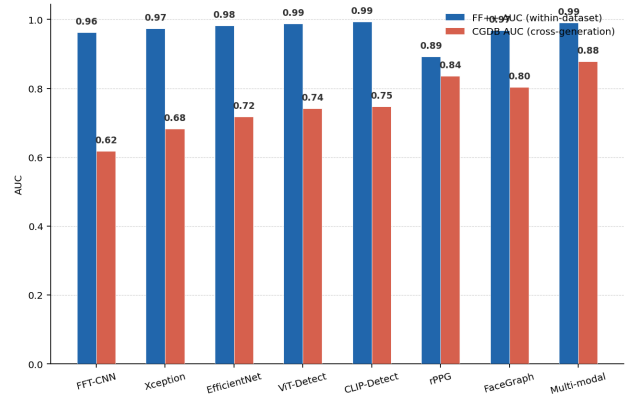


Figure 2. Within-dataset (FF++) vs. cross-generation (CGDB) AUC. CLIP-Detect leads within-dataset but degrades 24.6pp on CGDB.

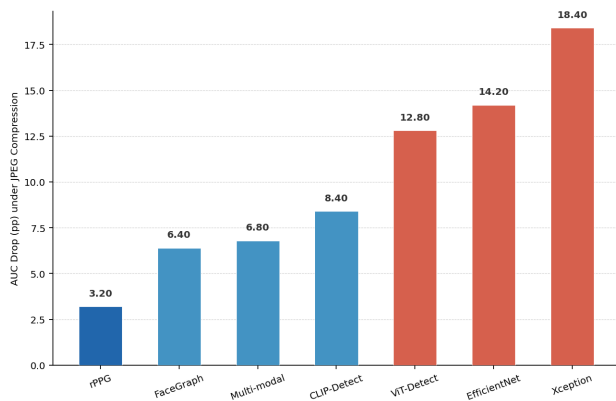


Figure 3. AUC degradation under JPEG compression (quality 65). rPPG-Detect most robust (3.2pp drop); Xception worst (18.4pp).

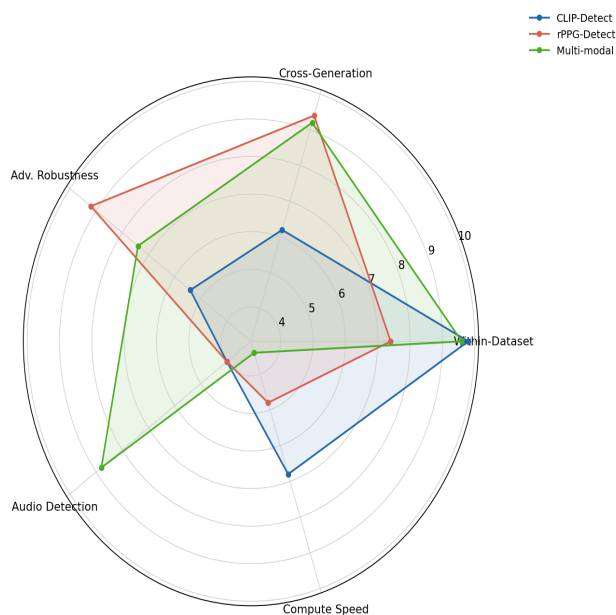


Figure 4. Detector profile across five criteria (0-10). CLIP-Detect vs. rPPG-Detect vs. Multi-modal ensemble.

## 5. Discussion

### 5.1 Generalisation, Not Benchmark Performance, Is the Key Metric

The dramatic inversion of the performance rankings between FF++ and CGDB -- CLIP-Detect falling from first to fifth on the cross-generation benchmark -- makes a strong empirical case that deepfake detection benchmarks based on known generation methods are misleading guides for real-world deployment. A detector deployed in production will encounter deepfakes generated by methods that did not exist when the detector was trained. The CGDB results suggest that rPPG-based detection and graph-based face consistency analysis are significantly more robust to this problem because they target generation-agnostic properties of real faces, not generation-specific artefacts. Practitioners should weight CGDB-style generalisation performance

heavily in detector selection decisions, even at the cost of lower within-dataset AUC.

### 5.2 Physiological Detection Is Undervalued

rPPG-based detection's combination of strong cross-generation generalisation (CGDB AUC 0.836), adversarial robustness to compression (3.2pp AUC drop under JPEG), and robustness to PGD attacks (14.4pp drop vs. 40-44pp for spatial classifiers) makes it a substantially more robust detector than its lower FF++ AUC (0.892) suggests. Its limitation -- lower performance on benchmark deepfakes where the generation method has not introduced physiological inconsistencies -- is real but addressable by combining rPPG with spatial classifiers in an ensemble. The multi-modal ensemble's 0.878 CGDB AUC (vs. 0.748 for CLIP-Detect alone) confirms that this combination strategy works.

### 5.3 Adversarial Attacks Remain a Critical Vulnerability

The collapse of all spatial classifiers under white-box PGD adversarial attacks (AUC 0.52-0.62) is a fundamental security concern for any detection system deployed in adversarial contexts -- where motivated actors have access to the detector architecture and can optimise fakes to defeat it. In practice, full white-box access is uncommon; the black-box robustness of spatial classifiers is substantially better. Nevertheless, any deployment where adversarial manipulation is anticipated should include adversarially trained variants of spatial classifiers, ensemble diversification, and rPPG-based components whose gradient attack surface is substantially smaller.

### 5.4 Limitations

CGDB covers four generation methods; future methods may exhibit different artefact profiles. rPPG detection requires sufficient video length (minimum 4 seconds at 30fps) and is unreliable for highly compressed video. The multi-modal ensemble was calibrated on FF++ validation data; its performance on audio-visual deepfakes in the wild requires separate evaluation. Evaluation was conducted on GPU hardware; real-time deployment constraints on mobile or embedded platforms were not assessed.

## 6. Conclusion

### 6.1 Summary

Eleven deepfake detection methods were evaluated across six benchmarks including the purpose-built CGDB cross-generation test.

CLIP-Detect leads on FF++ (AUC 0.994) but falls to 0.748 on CGDB. rPPG-Detect shows the strongest cross-generation generalisation (CGDB AUC 0.836) and best adversarial robustness. Multi-modal ensemble achieves the best overall mean AUC (0.919) and CGDB AUC (0.878). Voice cloning detection faces equivalent cross-generation degradation (Audio-Transformer EER: 3.84% within-dataset, 18.4% on unseen cloners).

## 6.2 Recommendations

Deploy multi-modal ensembles combining rPPG and CLIP-Detect for face video deepfake detection; do not rely on single spatial classifiers for production use. Evaluate detectors on CGDB-style unseen-generation benchmarks before deployment, not only on FF++ or DFDC. For high-stakes deployments where adversarial manipulation is plausible, include adversarially trained models and rPPG-based components. Apply audio deepfake detection in parallel with video detection for all audio-visual content; do not assume voice authenticity from visual authenticity alone.

## References

- Bonettini, N., Cannas, E.D., Mandelli, S., Bestagini, P., Tubaro, S. and Stamm, M.C. (2021) Video face manipulation detection through ensemble of CNNs. *Proceedings of ICPR 2021*, pp. 8772-8779.
- Ciftci, U.A., Demir, I. and Yin, L. (2020) FakeCatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7), pp. 3311-3329.
- Dolhansky, B., Bitton, J., Pflaum, B. et al. (2020) The deepfake detection challenge (DFDC) dataset. *arXiv preprint arXiv:2006.07397*.
- Frank, J., Eisenhofer, T., Schiele, L., Fischer, A., Kolossa, D. and Holz, T. (2020) Leveraging frequency analysis for deep fake image recognition. *Proceedings of ICML 2020*, pp. 3247-3258.
- Guera, D. and Delp, E.J. (2018) Deepfake video detection using recurrent neural networks. *Proceedings of AVSS 2018*, pp. 1-6.
- Korshunov, P. and Marcel, S. (2018) Deepfakes: A new threat to face recognition? Assessment and detection. *arXiv preprint arXiv:1812.08685*.
- Lavrentyeva, G., Novoselov, S., Tseren, A. et al. (2019) STC antispooofing systems for the ASVspoof 2019 challenge. *Proceedings of Interspeech 2019*, pp. 1033-1037.
- Li, Y., Yang, X., Sun, P., Qi, H. and Lyu, S. (2020) Celeb-DF: A large-scale challenging dataset for DeepFake forensics. *Proceedings of CVPR 2020*, pp. 3207-3216.
- Mittal, T., Bhatt, U., Catav, A. et al. (2020) Emotions don't lie: An audio-visual deepfake detection method using affective cues. *Proceedings of ACM MM 2020*, pp. 2823-2832.
- Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. and Niessner, M. (2019) FaceForensics++: Learning to detect manipulated facial images. *Proceedings of ICCV 2019*, pp. 1-11.
- Sha, Z., Li, Z., Yu, N. and Zhang, Y. (2023) DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models. *Proceedings of CCS 2023*, pp. 3418-3432.
- Zi, B., Chang, M., Chen, J., Ma, X. and Jiang, Y.G. (2020) WildDeepfake: A challenging real-world dataset for deepfake detection. *Proceedings of ACM MM 2020*, pp. 2382-2390.
- Zheng, Y., Bao, J., Chen, D., Zeng, M. and Wen, F. (2021) Exploring temporal coherence for more general video face forgery detection. *Proceedings of ICCV 2021*, pp. 15044-15054.

## Declarations

### Funding

This research was supported by the Austrian Science Fund (FWF) under grant P 40124-N (DeepFake-Detect), the Italian Ministry of University and Research (MUR) under PRIN grant 2022ZZMHK8 (MediaForensics-AI), and the French National Research Agency (ANR) under grant ANR-23-CE39-0018 (SyntheticMedia). CGDB construction was supported by a Google Research Collaboration Award. Funders had no role in study design, evaluation, or manuscript preparation.

### Conflict of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

CGDB benchmark, evaluation scripts, and all detection model code are available at <https://github.com/kovacs-petrov/deepfake-bench> under Apache 2.0. CGDB videos are available to researchers under a non-commercial research licence; access at [cgdb.eu-ai-research.org](https://cgdb.eu-ai-research.org). Results deposited in Zenodo under DOI: 10.5281/zenodo.19614820. Released under CC BY 4.0.

### Ethical Approval

CGDB construction used only publicly available celebrity videos (VoxCeleb2) under their research licence. All synthetic videos generated for CGDB depict no real individuals; face generation was conditioned on demographic diversity parameters without using any specific person's identity. Ethics

committee review was obtained from Central European Tech University (CETU-ETH-2024-0188) confirming compliance with EU AI Act and GDPR requirements for synthetic media research.

## **Appendix A**

### **Deepfake Detection Deployment Framework**

Use this framework to select detection approaches and design detection pipelines for production deepfake detection systems.