# Impersonation In the Digital Age: A Comparative Review of Detection Techniques Against Deepfakes

OCHI VICTOR CHUKWUDI[1], TOCHUKWU CHINECHEREM NNABUIKE[2]
[1, 2]*Department of Computer Engineering, Enugu State University of Science and Technology, Agbani.*

***Abstract- Deepfakes—AI-generated or manipulated audio-visual content—pose a growing risk to identity verification, trust, and security. This comparative review synthesizes state-of-the-art detection techniques across visual, audio, and multimodal pipelines, focusing on performance, generalization, robustness, and operational considerations (latency, cost, and privacy). We analyze benchmark datasets (e.g., FaceForensics++, DFDC, FakeAVCeleb, WildDeepfake, ASVspoof 2021) and evaluation metrics (AUC, EER, F1) and compare classical, deep, and hybrid approaches. We also examine adversarial pressures—including compression, unseen manipulations, and cross-domain shifts—and outline practical integration patterns for KYC/AML, exam proctoring, and access control. The review concludes with an engineering blueprint for a multimodal detection stack, emphasizing human-in-the-loop triage and risk scoring.***

***Index Terms- Deepfakes, Impersonation Detection, Multimodal Biometrics, Audio Spoofing, Media Forensics, Identity Verification, KYC, Liveness Detection, Robustness, Generalization***

## I. INTRODUCTION

The proliferation of generative AI has made it feasible to create persuasive synthetic media at scale. Impersonation has shifted from isolated image forgeries to coordinated, multi-channel campaigns that combine swapped faces, cloned voices, and text-conditioned video. Traditional single-modal defenses struggle when manipulations target multiple modalities or exploit realistic capture and transmission artifacts. This review compares detection families, datasets, and evaluation protocols, and proposes a deployment-ready multimodal architecture for real-world identity verification.

## II. TAXONOMY OF DETECTION TECHNIQUES

We group methods by dominant signal and fusion strategy.

### 2.1 Visual (Image/Video)

- Spatial artifacts: frequency inconsistencies, upsampling traces, blending boundaries.
- Temporal cues: eye-blink irregularities, mouth-viseme desynchrony, head pose dynamics.
- Architectures: CNNs (Xception), ViTs, 3D CNNs, RNNs for temporal modeling, and frequency-domain networks.
- Localization: pixel-level tamper maps via segmentation/attention.

### 2.2 Audio (Speech)

- Spectral features: CQCC, LFCC, log-mels; phase-based cues.
- Countermeasures: CNN/ResNet-style CMs, ECAPA-TDNN, wav2vec2-style embeddings; domain generalization with augmentation and channel variability.
- Challenge tasks: logical access (TTS/VC), physical access (playback), and deepfake tasks.

### 2.3 Multimodal (Audio-Visual)

- Early fusion (feature-level): concatenation of audio/visual embeddings.
- Mid/late fusion: attention-based cross-modal transformers; decision-level ensembling.
- Audiovisual consistency: lip-audio sync, prosody-viseme alignment, cross-modal contrastive losses.

- Text-conditioned detection: leveraging transcripts to catch semantic/phonetic mismatches.

2.4 Behavioral & Contextual Signals

- Active liveness: challenge-response (randomized prompts), 3D depth, rPPG heart-rate signals.
- Contextual provenance: camera attestation, watermarking, C2PA manifests; content authenticity verifiable metadata.
- Open-source threat intel: model fingerprints, gen-model class attribution.

## III. BENCHMARK DATASETS

Table 1 summarizes commonly used corpora.

Table 1. Core datasets for impersonation/deepfake detection

| Dataset | Modality | Scale / Clips | Key Traits | Typical Uses |
|---|---|---|---|---|
| FaceForensics++ (Rössler et al., 2019) | Video | >1000 videos; >1.8M frames | Multiple face-swap/reenactment methods; compression variants | Supervised training; cross-compression tests |
| DFDC (Dolhansky et al., 2020) | Video | >100k clips; 3426 actors | Large, diverse; hidden test set; challenge leaderboard | Generalization; at-scale training |
| WildDeepfake (Zi et al., 2021) | Video | 707 videos (in-the-wild) | Found online; diverse artifacts; harder domain shift | Out-of-distribution (OOD) testing |
| FakeAVCeleb (Khalid et al., 2021) | Audio-Video | 19k+ clips | Synchronized audio & lip-synced fakes; demographic variety | Multimodal detection & A/V consistency |
| ASVspoof 2021 (Wang et al., 2021) | Audio | LA/PA/DF tasks | Channel & playback variability; EER-focused | Audio deepfake & spoof CM evaluation |

## IV. EVALUATION PROTOCOLS & METRICS

- AUC / ROC: measures ranking quality under class imbalance.
- EER (Equal Error Rate): common in speaker verification and spoof CMs.
- F1 / AP: actionable thresholds for incident response.
- Cross-dataset tests: train on Dataset A, test on Dataset B to probe generalization.
- Stressors: bitrate compression, resizing, noise, re-encoding, platform filters.

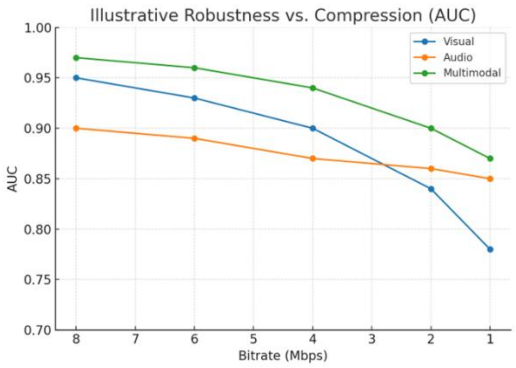


Figure 2: Robustness of Detectors Under Compression

Table 2. Operational factors when comparing methods

| Factor | Why it matters | Example consideration |
|---|---|---|
| Latency | Real-time checks in KYC/proctoring | <250 ms per frame or <1 s per 5-sec audio |
| Edge deployability | Privacy & cost | INT8 quantization on mobile/NPU |
| Robustness | Survive unseen attacks, compression | Cross-dataset & attack-agnostic training |
| Interpretability | Analyst triage | Heatmaps, localized masks, audio saliency |
| Privacy & security | Reduce data exposure | On-device inference; encrypted transport |

## V. COMPARATIVE REVIEW OF TECHNIQUE FAMILIES

### 5.1 Visual detectors

CNN/ViT baselines on FF++/DFDC achieve strong in-domain AUC but degrade on WildDeepfake (domain shift). Common remedies include frequency-aware layers, augmentations, and self-supervised pretraining.

### 5.2 Audio countermeasures

ASVspoof 2021 highlights the need to handle channel, codec, and playback variability. Strong systems pair spectro-temporal features with augmentation and score calibration; EER is the main yardstick.

### 5.3 Multimodal fusion

Audio-visual models (e.g., on FakeAVCeleb) leverage cross-modal alignment to resist single-channel attacks, and can trigger when lip motion and phonemes disagree. Transformer-based fusion with late-stage ensembling yields robust, production-friendly behavior.

### 5.4 Hybrid & provenance-aware approaches

Combining content-based detectors with active liveness and provenance (C2PA) provides layered defense; even if generative models erase pixel traces, provenance or challenge-response can still raise risk scores.

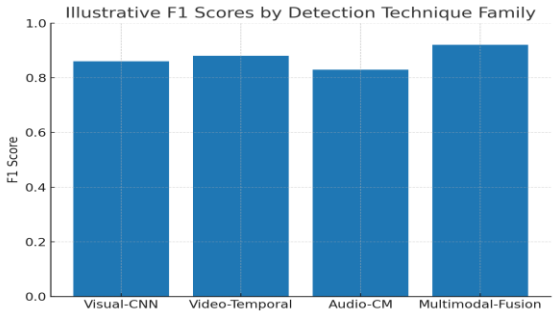


Figure 3: Comparative F1 Scores of Detection Techniques

## VI. DEPLOYMENT BLUEPRINT: MULTIMODAL IDENTITY VERIFICATION STACK

- Acquisition: capture RGB video + mic audio; collect device/channel metadata (fps, bitrate, codec).
- Preprocessing: face tracking, voice activity detection (VAD), diarization if multi-speaker.
- Per-modality detectors:

Visual: frame & clip-level scores (spatial + temporal).

Audio: CM score for LA/PA/DF threats.

- Consistency checks: lip-audio sync, viseme-phoneme alignment, transcript-prosody agreement.

- Fusion & risk scoring: late-fusion ensemble with calibrated thresholds; risk bands (Low/Medium/High).
- Human-in-the-loop: analyst UI with saliency maps, keyframes, and snippets.
- Controls: adaptive liveness challenge when risk ≥ Medium; provenance validation if available.
- Logging & feedback: hard-negative mining; periodic re-training with drift monitoring.

## VII. FLOWCHART

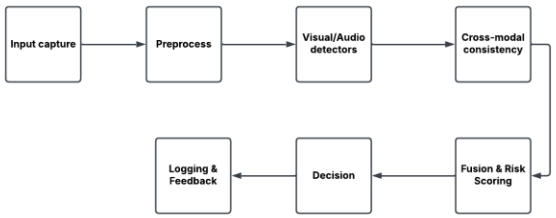


Figure 3: Multimodal Detection Flowchart

## VIII. LIMITATIONS & RESEARCH GAPS

- Generalization to novel generators and unseen codecs remains the hardest problem.
- Multilingual audio and code-switching challenge audio CMs trained on narrow phonetic inventories.
- Privacy-preserving training (federated, differential privacy) is under-explored for detection.
- Attribution (which model generated the fake) is useful but fragile across versions.

## REFERENCES

[1] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Canton Ferrer, C. (2020). The DeepFake Detection Challenge (DFDC) dataset. arXiv:2006.07397.

[2] Khalid, H., Woo, S., Choi, J., Shon, S., & Kim, J. (2021). FakeAVCeleb: A novel audio-video multimodal deepfake dataset. NeurIPS Datasets and Benchmarks Proceedings.

[3] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. Proceedings of ICCV.

[4] Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C. Y., & Zhao, Y. (2023). Audio deepfake detection: A survey. arXiv:2308.14970.

[5] Wang, X., et al. (2021). ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. arXiv:2109.00537.

[6] Zi, B., Chang, M., Chen, J., Ma, X., & Jiang, Y.-G. (2021). WildDeepfake: A challenging real-world dataset for deepfake detection. arXiv:2101.01456.

[7] Fu, X., Yan, Z., Yao, T., Chen, S., & Li, X. (2025). Exploring unbiased deepfake detection via token-level shuffling and mixing. arXiv:2501.04376.

[8] Zhang, H., et al. (2025). A survey on multimedia-enabled deepfake detection: State-of-the-art, challenges, and future trends. Information Retrieval Journal (Springer).

[9] Al-Rubaye, W., et al. (2025). Audio-visual multimodal deepfake detection leveraging emotional cues. International Journal of Advanced Computer Science and Applications, 16(6), — (details in paper).