Targeted Synthetic Impersonation at Decision Points Defending Against Text, Voice, and Video-Based Attacks in Enterprise, Politics, and e-KYC

Skanda Pandith M G
USC ID: 7126339709
Email: pandithm@usc.edu
CSci530 – Computer Security Systems, Fall 2025

Academic Integrity Statement

I have read the Guide to Avoiding Plagiarism published by the USC Student Affairs Office. I understand what is expected of me with respect to properly citing sources and avoiding the representation of the work of others as my own. The material in this paper was written by me, except for such material that is quoted or indented and properly cited to indicate the sources of the material. I understand that using the words of others, even with a citation tag, does not constitute proper citation unless the quoted material is clearly marked. I also understand that overuse of properly cited quotations, while not a violation, may indicate insufficient understanding of the material and may reduce my grade.

Signed: Skanda Pandith M G

I. abstract

Synthetic impersonation attacks have become a systemic security threat as modern text-, voice-, and video-generation models enable highly realistic deepfakes that can subvert biometric systems, decision workflows, and public-facing communication channels. Across the surveyed literature, deepfakes are defined as ML-generated or ML-manipulated media—spanning TTS, VC, replay attacks, and advanced visual synthesis—that mimic real individuals with increasing fidelity. Audio deepfakes remain especially dangerous: modern generators clone voices convincingly, generalize across languages and domains, and operate effectively through noisy, compressed, or telephony channels. Yet detection pipelines—whether based on handcrafted spectral features, deep CNN/GNN/Transformer models, or self-supervised encoders—continue to degrade under unseen generators, codecs, linguistic edits, adversarial perturbations, and realistic environmental shifts. Visual and multimodal systems face similar limitations, with detectors struggling on cross-dataset transfers, partial manipulations, and audio–video desynchronization.

These weaknesses create direct vulnerabilities at critical decision points. Enterprise approval chains, political messaging pipelines, and remote e-KYC verification workflows can all be bypassed by synthetic speech, replay attacks, or joint audio-visual deepfakes, enabling financial fraud, identity takeover, reputational harm, and the liar's dividend. Benchmarks such as ASVspoof, ADD, and DFDC illustrate both the sophistication of modern attacks and the brittleness of current defenses. Complementary protection layers—including content provenance standards (C2PA), classical and deep learning watermarking, and biometric liveness detection—offer partial mitigation, but each comes with structural limitations: provenance requires ecosystem adoption and trusted signers; watermarks remain vulnerable to codec transformations and adaptive attacks; and software-only liveness detection works but varies across sensors and spoofing conditions.

Across policy and UX research, guardrails are needed to prevent misinterpretation, reduce false positives, increase transparency, and balance security with privacy and civil liberties. Legal frameworks provide some accountability through civil and criminal remedies, while analyst-oriented UX guidelines emphasize explainability, contextualized outputs, and workflow integration. Human-in-the-loop systems further strengthen defenses by pairing machine detection with expert review, but they introduce their own risks—cost, annotation bias, workflow complexity, and error propagation.

Taken together, the literature shows that synthetic impersonation attacks exploit gaps across technical, organizational, and societal layers. Defenses remain fragmented and brittle, and no single approach is sufficient. A resilient ecosystem requires a hybrid strategy: robust detection, provenance and watermarking, multimodal biometric safeguards, policy and regulatory guardrails, and carefully designed human-in-the-loop workflows that preserve both trust and accountability.

II. Introduction

Synthetic impersonation attacks have shifted from a niche technical curiosity to a real, operational security threat. Modern generative models now produce highly convincing text, lifelike cloned voices, and hyper-realistic video deepfakes — and these outputs are good enough to fool both humans and current verification technologies. Unlike traditional phishing or generic scams, these attacks target specific individuals and exploit trust at exactly the moment when critical decisions are made, such as approving financial transfers, validating political content, or verifying a user's identity in remote onboarding workflows.

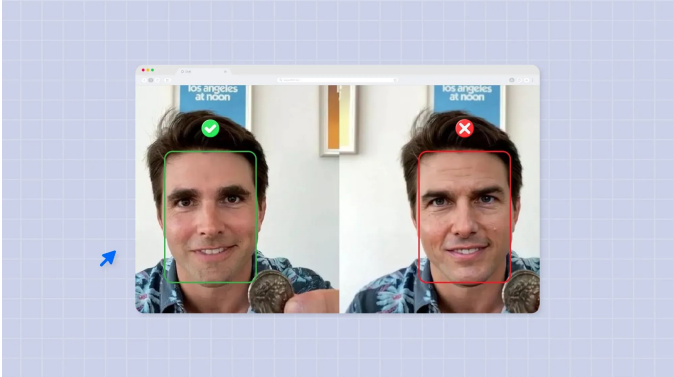

Fig. 1. Examples of deepfake visual manipulations across well-known artworks, demonstrating facial reenactment and expression synthesis.

These attacks succeed because "decision points" in everyday systems depend heavily on human trust cues: a familiar voice, writing style, facial expression, or emotional tone. The literature shows that deepfake audio in particular is a structural weakness — TTS, VC, and replay attacks can now clone voices from short samples, operate through compressed telephony channels, and bypass both human judgment and automated speaker-verification systems. Visual and multimodal deepfakes further amplify this threat by synchronizing synthetic faces and speech, enabling impersonation across biometric checkpoints and public-facing communication pipelines.

Our study and threat model focus precisely on these high-stakes workflows. We examine three real-world settings where synthetic media causes the most damage: enterprise approval chains, political messaging pipelines, and remote e-KYC verification. In each case, we analyze where the workflow breaks, how deepfakes enter, and which trust assumptions make these systems vulnerable. We evaluate four major defense families highlighted across the literature — content provenance (C2PA), digital watermarking, biometric liveness detection, and AI-based multimodal detection — not just by listing them, but by assessing their structural constraints, hidden assumptions, and the conditions under which they fail.

This research delivers five concrete contributions. First, we build a unified threat model that aligns capabilities across text, voice, and video attacks. Second, we map precise vulnerability points within common decision workflows, showing how attackers exploit them. Third, we critically evaluate detection, provenance, and watermarking systems based on their documented failure modes. Fourth, we analyze the human-in-the-loop breakpoints that cause reviewers to misjudge or misinterpret synthetic media. Finally, we propose a practical, multi-layer defense framework and outline research paths toward resilient organizational systems.

Together, these sections establish that synthetic impersonation attacks exploit weaknesses spanning technical infrastructure, organizational workflows, and societal trust. Existing defenses remain fragmented and brittle. A durable solution requires hybrid, layered security: robust technical detection, provenance and watermarking, strong biometric safeguards, policy and UX guardrails, and well-designed HITL workflows that maintain trust without overreliance on automation.

III. Background and Definitions

A. Deepfakes and Audio Deepfakes

Across the surveyed papers, deepfakes are ML-generated or ML-altered synthetic media—image, video, or audio—that depict events that never happened.[5] Chesney & Citron frame them as advanced digital doctoring using neural networks and GANs.[1] Firc et al. highlight their spread across entertainment, fake news, fraud, identity theft, and biometric spoofing. Audio surveys define audio deepfakes as speech generated or modified through TTS, VC, or replay, enabling political misinformation, scams, cybercrime, and biometric impersonation.[2], [3]

B. Categories of Audio Deepfakes

The audio literature consistently groups attacks into:

TTS (Synthetic Speech): Models such as WaveNet, Tacotron/Tacotron2, DeepVoice, VoiceLoop, and Parallel WaveNet generate natural, often voice-cloned speech.[2], [3]

VC (Imitated Speech): GMM-, DBLSTM-, autoencoder-, and VAE/GAN-based approaches (CycleGAN, StarGAN, VAW-GAN; non-parallel, zero/one-shot, cross-lingual) convert one speaker's voice to another.[2], [3]

Replay Attacks: Genuine recorded speech is played back to bypass liveness.[4]

Newer analyses also track emotion fakes, background/scene fakes, and partial fakes, which are harder to detect because only segments of the signal are altered.[2]

C. Visual Deepfakes

Firc et al. [1] extend these threats to facial domains, outlining: face synthesis (StyleGAN and online generators), face morphing (used in passport and border-control fraud), face swap (DeepFaceLab, FSGAN, SimSwap, FaceShifter, Reface/Zao), and facial reenactment (driven by source audio or video).[5] They emphasize that combining manipulated audio + video enhances impersonation attacks on biometric systems.

D. Benchmarks, Datasets, Evaluation Metrics

Audio deepfake research relies heavily on benchmark suites such as ASVspoof (2015–2024)—especially 2021—which defines three tasks: LA: digital TTS/VC with telephony distortions; PA: replay in real rooms with varied acoustics/devices; DF: online deepfakes with compression and diverse sources.[4]

ASVspoof 2021 introduced no new training data, forcing generalization.[4] Other datasets include ADD, FoR, WaveFake, HAD, ITW, LibriSeVoc, SceneFake, EmoFake, CVoiceFake, MLAAD.[2] Evaluation relies mainly on EER and t-DCF (combined ASV + countermeasure cost).[4], [3]

E. Detection Pipelines: Features and Models

Detection follows a two-stage structure:

Frontend Features: Handcrafted: LPS, filterbanks, mel-spectrograms, CQT/CQCC, MFCC, LFCC, GTCC, spectral flux/centroid.[2], [3] Learned: SincNet, RawNet2, and self-supervised/transformer models such as Wav2Vec2, WavLM, XLS-R, HuBERT, SSAST, ViT.[2]

Backend Models: Classical: GMM, SVM, logistic regression. Deep learning: CNNs (ResNet, LCNN, SENet), capsule networks, GNNs (GAT/GCN), transformers (SSAST/ViT), Deep-Sonar, Res-/Inc-TSSDNet, TEResNet, LC-GRNN, Spec-ResNet.[2], [3]

Key findings across surveys: deep models outperform handcrafted features but degrade on unseen generators, noise, and codecs; self-supervised features with strong classifiers currently perform best; robustness depends heavily on augmentation (noise, codecs, reverberation, SpecAug, RawBoost).[2], [4]

F. Open Problems

All three surveys report the same persistent challenges: poor generalization to new codecs, languages, environments, and unseen models; dataset limitations and weak cross-dataset robustness;[2], [4] privacy concerns requiring content-obscuring detection (e.g., SafeEar); fairness issues across gender, accent, and language; lack of interpretability for operational or legal contexts; vulnerability to adversarial perturbations and simple signal transformations.[2], [3]

G. Societal and Legal Context

Chesney & Citron warn that deepfakes exploit online dynamics—information cascades, novelty bias, and filter bubbles—leading to extortion, sextortion, reputational damage, election manipulation, the undermining of journalism, and the liar's dividend.[1] Firc et al. note real-world incidents such as political videos and morphed passports, alongside slow-moving legal frameworks.[5] Together, these works motivate a threat model where audio deepfakes pose direct risks to biometric systems and the trustworthiness of digital communication.[2], [5]

IV. Threat Model

A. Assets and Stakeholders

Across the surveyed works, the key assets at risk are biometric authentication systems (especially ASV and voiceprint-based verification)[4], the identity and reputation of individuals or organizations, the integrity of communications and decision workflows, and broader societal trust in news, institutions, and recorded speech. These risks include impersonation attacks (ASVspoof)[4], fake political or corporate instructions, fabricated compromising recordings[5], and erosion of trust via the liar's dividend[5]. Stakeholders span end users, organizations relying on voice verification or recorded evidence, platforms hosting content, and institutions that depend on trustworthy media[5].

B. Adversary Goals

The literature points to several consistent attacker goals:

biometric impersonation to bypass ASV or voice-based authentication using TTS, VC, or replay[4], [2], [3];

financial fraud / CEO- or BEC-style attacks by generating fake executive audio to trigger transfers or disclosures[2], [5];

political or societal manipulation through synthetic speech or audio-visual deepfakes that influence elections or distort public discourse[2], [5];

extortion and blackmail using fake but damaging speech recordings[5];

exploiting the liar's dividend by undermining real evidence or creating plausible deniability[5];

bypassing detection through generators, codecs, or conditions that exploit known weaknesses in modern detectors[4], [2], [3].

C. Adversary Capabilities

Adversaries can use modern TTS/VC and replay methods (WaveNet/Tacotron/DeepVoice-like models, GMM/DBLSTM/AE/VAE-GAN VC, and high-quality replay)[2], [3]. They can gather public target speech data and face data for audio-visual deepfakes. They control channels/codecs across logical access, physical playback, and online distribution—mirroring ASVspoof's LA/PA/DF setups[4]. They can manipulate full, partial, emotional, or background-altered fakes[2], and iterate using different generators, codecs, or distortions until detectors fail. Tools are democratized, so attackers do not need ML expertise[5], [3].

D. Defender Capabilities and Limitations

Defenders can use detection pipelines based on handcrafted features (CQCC, CQT, MFCC, LFCC, GTCC), DL architectures (ResNet, LCNN, SENet, capsule networks, GNNs, transformers like SSAST and ViT, Deep-Sonar, TSSDNet, TEResNet, LC-GRNN, Spec-ResNet), and self-supervised frontends (Wav2Vec2, WavLM, XLS-R, HuBERT)[2], [3]. They can tune using datasets such as ASVspoof[4], ADD, FoR, WaveFake, HAD, ITW, LibriSeVoc, SceneFake, EmoFake, CVoiceFake, and MLAAD[2], plus augmentation and model ensembles.

But the surveys highlight major limits: detectors fail on unseen generators/codecs[4], [2], [3], generalize poorly across languages and wild conditions[2], [3], plateau in realistic performance, raise fairness concerns, are vulnerable to adversarial manipulation, and—when privacy-preserving—operate with further data constraints. Overall, defenders have useful but brittle tools.

E. Attack Surfaces

Main attack surfaces include:

ASV-protected services—digital injection (LA) and physical playback (PA) attacks[4];

voice-based customer service/call centers—deepfake TTS/VC over compressed or noisy telephony channels[4];

online platforms and public media—deepfake distribution for misinformation, reputational harm, and leveraging human + platform detector weaknesses[5], [2];

combined audio–visual biometric systems—joint face–voice deepfakes exploiting reenactment + synthetic speech[5];

evidence and archival systems—legal, journalistic, or compliance recordings that can be disputed or polluted[5].

F. Scope of Our Analysis

The surveyed literature consistently shows that audio deepfake generation (TTS, VC, replay) is rapidly improving and already accessible to non-experts[2], [3], [5], while detection remains fragile—especially under realistic, unseen, or shifted conditions[4], [2], [3]. Speaker verification systems are structurally vulnerable since every modality of audio deepfake enables impersonation[4]. At the societal level, false positives, false negatives, and the liar's dividend jointly threaten trust in evidence[5]. Thus, subsequent sections reasonably assume powerful but imperfect attackers and defenders using detection that is helpful but not robust, matching the reality depicted across the technical surveys and legal analyses.

V. Vulnerabilities in Decision-Point Workflows

A. Enterprise Approval Workflows

Enterprises increasingly rely on voice-driven or communication-driven approval chains — everything from payment authorizations to password resets and internal escalations. Deepfake audio fits into this workflow like a master key: TTS/VC systems trained on publicly available speech (as described in the surveys) can convincingly clone executives or managers and issue instructions that employees perceive as legitimate. These workflows rarely have multi-factor validation or provenance checks, making them structurally exposed to impersonation attacks and the liar's dividend — where even genuine instructions can later be disputed. Because detection systems degrade under new codecs, unseen generators, noise, and telephony distortions (ASVspoof LA/PA/DF findings), an attacker can inject synthetic audio that bypasses both human judgment and automated ASV, subverting high-stakes decisions such as financial transfers or access approvals.

B. Political Messaging Pipelines

Political communication pipelines — campaign messaging, press statements, public speeches, and grassroots mobilization content — are acutely vulnerable because they operate at scale and depend on rapid, trust-based dissemination. Deepfake audio allows attackers to fabricate inflammatory statements, mimic political leaders, or distort policy positions, aligning with the harms Chesney & Citron describe (manipulation of elections, degradation of democratic discourse, and reputational sabotage). Modern tools enable partial edits, emotion fakes, and scene fakes that are hard to detect and easy to distribute through online platforms, exploiting information cascades and filter bubbles. Detection systems struggle across domains, languages, and noisy real-world channels, meaning that once synthetic audio enters the pipeline, it can influence opinion long before fact-checking or forensic tools can respond — and afterward, the liar's dividend enables genuine content to be dismissed as fabricated.

C. e-KYC Remote Verification

Remote identity verification workflows — especially e-KYC processes that rely on speaker verification, liveness checks, or audio-visual matching — inherit the full spectrum of deepfake risks outlined in the audio and visual surveys. Attackers can use TTS/VC systems, replay attacks, or joint audio-visual deepfakes to bypass biometric checks, since ASV systems remain vulnerable to unseen generators, new codecs, and partial fakes.[1] Firc et al.'s discussion of facial morphing, face swaps, and reenactment further expands the threat surface: attackers can combine manipulated audio with synthetic or reenacted faces to defeat multimodal checks that expect synchronized lips and speech. Because e-KYC pipelines often operate in compressed channels (mobile apps, telephony, WebRTC), the adversary gains additional leverage — many detectors perform poorly under such conditions. The result: account takeovers, fraudulent onboarding, and large-scale identity fabrication that corrupt the integrity of entire digital onboarding systems.

VI. Defense Approaches

A. Content Provenance / C2PA

Content provenance systems make media origins and edit history transparent instead of filtering deepfakes directly. The leading standard, C2PA (from CAI + Adobe, Microsoft, BBC, etc.), attaches cryptographically signed Content Credentials to images, audio, video, and documents.[7] A C2PA manifest can record the creator/publisher, capture time/location, device or software used, and editing steps. This manifest is bound to the media via hashes and signatures, so any change invalidates the provenance chain.

C2PA tools present this metadata as a "nutrition label," showing whether provenance is intact, who signed it, and whether anything was tampered with. Importantly, C2PA does not judge truth—it only certifies that a specific entity signed specific content and that neither content nor metadata changed afterward.[7]

This approach fits platform/enterprise workflows: cameras and editing tools can auto-attach credentials, while newsrooms and platforms can require valid chains from trusted issuers. But limitations include low adoption, gaps when metadata is missing or stripped, and reliance on trust in signers. C2PA is metadata-centric and complementary to watermarking, not a replacement.[6], [7] Overall, it strengthens transparency for legitimate content but labels everything else as "provenance unknown."

B. Digital Watermarking

Watermarking embeds information inside audio/visual signals. The two referenced papers survey deep-learning watermarking for images/video and benchmark audio watermark robustness (AudioMarkBench). Together, they outline how watermarking functions as an in-band defense across modalities.

1) Classical Image & Video Watermarking: Traditional watermarking includes embedding (DCT, DWT, FFT, etc.) and detection (blind/non-blind extraction). Key requirements: invisibility (PSNR/SSIM), capacity, and robustness (BER/NC). Methods span spatial (LSB, MIDSB), transform (DCT/DFT/DWT/SVD), and hybrid (DWT+DCT, DWT+SVD) domains. For video, frame-by-frame watermarking is fragile; robust designs use temporal redundancy or embed in compressed domains (MPEG, H.264, HEVC).

2) Deep Learning–Based Watermarking: DL watermarking uses CNNs, autoencoders, or GANs in an encoder → attack layer → decoder workflow. CNN/autoencoder models deliver high imperceptibility and add enhancements such as strength scaling, preprocessing networks, or Octave Convolution. GAN-based methods improve robustness via discriminators, attention masks, robust pixel selection, and WGAN/CycleGAN variants.

Deep video watermarking is still emerging (approximately 12 papers since 2019). Methods work on frames (CNN/GAN + attention, repetition, multiscale) or on compressed domains (HEVC/H.265). Some use mosaic-based reversible conversions. Performance typically hits 34–44 dB PSNR and improves robustness to compression, cropping, blur, and frame dropping. Key gaps remain: weak temporal modeling, missing differentiable codecs, limited collusion resistance, and no dominant architecture.

3) Audio Watermarking Robustness (AudioMarkBench): AudioMarkBench evaluates watermark robustness for synthetic speech. It uses LibriSpeech + a 20k multilingual dataset, three SOTA systems (AudioSeal/AudioSeal-B, Timbre, WavMark), and 15 perturbations: no-box transformations (codecs, filters, noise, stretching, echo, quantization), black-box adversarial attacks (HopSkipJump, Square), and full white-box attacks. Two threat models are tested: watermark removal and watermark forgery.

Results: performance is near-perfect when clean, but modern codecs (EnCodec, SoundStream, Opus) can break watermarks while keeping audio quality high. No-box forgery rarely works, but black-box removal improves with many queries, and white-box attacks defeat all systems (FNR$\approx$ 1 or high FPR). Fairness tests show variation across sex/language and minimal age trends. Overall, audio watermarking remains vulnerable to adaptive adversaries and new codecs.

4) Takeaways & Relation to C2PA: DL watermarking improves imperceptibility and robustness for images and video, but video lags in temporal and collusion resilience. In audio, watermarks survive everyday distortions but fail under aggressive codecs and white-box/high-query attacks, with demographic variance.

Compared with C2PA, watermarking embeds information in-band and can persist across reposts, but can also be removed or overwritten. C2PA provides out-of-band cryptographic provenance that stays intact in trusted workflows but disappears outside compliant ecosystems. Neither is sufficient alone: watermarking marks the signal, while C2PA secures provenance, and together they form complementary layers in a broader defense stack.

C. Biometric Liveness Detection

This paper presents a high-performance, software-only fingerprint liveness detection method designed to distinguish real fingerprints from fake ones created using silicone, gelatin, and playdoh[8]. The system requires only a single fingerprint image and avoids any additional hardware, making it low-cost, fast, and easy to deploy across existing sensors[8]. Because biometric systems are vulnerable to direct spoofing attacks—known as gummy fingers—the authors motivate a purely image-based approach as a cheaper and more user-friendly alternative to intrusive hardware solutions[8]. The proposed pipeline segments the fingerprint using Gabor filters, extracts ten quality-related features capturing ridge strength, continuity, and clarity, performs exhaustive feature-subset selection, and classifies samples as real or fake using Linear Discriminant Analysis, with feature subsets tuned to each sensor type[8]. The method is evaluated on six datasets, including the LivDet 2009 sensors (Biometrika, CrossMatch, Identix) and the ATVS database (Biometrika, Precise, Yubee), covering more than 10,500 images and multiple spoofing conditions[8]. Experiments follow a clear two-stage protocol—training on development sets and validating through cross-dataset swaps—with performance measured using ACE, defined as the average of fake-to-real and real-to-fake error rates[8]. Results show that the system achieves approximately 90% accuracy overall, with ACE values sometimes as low as 1.47%, although dataset inconsistencies (notably in LivDet Biometrika) can produce higher validation errors[8]. The study further notes that non-cooperative spoofs tend to be easier to detect than cooperative ones due to fabrication artifacts[8]. Overall, the paper's contributions include a novel quality-driven fingerprint representation, sensor-specific optimization, broad multi-dataset evaluation, and a demonstration that software-based image analysis can effectively support biometric liveness detection[8].

D. AI-Based Detection

1) Audio Deepfake Detection: AI-based audio deepfake detection spans TTS and VC generation families, with the survey "Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead"[2] detailing models such as WaveNet, DeepVoice 3, Tacotron 2, FastSpeech 2 for TTS and CycleGAN-VC, AutoVC, MulliVC, VQVC, and FreeVC for VC, along with emotion fakes, partially fake audio, and scene fakes. Datasets include ASVspoof (LA, PA, DF, and ASVspoof 5 with adversarial attacks), ADD (Low-Quality, Partially Fake, Fake Game, Region Localization, Algorithm Recognition), and corpora such as FoR, WaveFake, HAD, ITW, LibriSeVoc, SceneFake, EmoFake, CVoiceFake, and MLAAD (38 languages, 82 models). Evaluation metrics include EER and t-DCF. Detection pipelines combine handcrafted spectral features (LPS, LFB, Mel, CQT, C-CQT, Spec), cepstral features (MFCC, LFCC, CQCC, MPE), raw-waveform models (SincNet, RawNet/RawNet2), and self-supervised encoders (Wav2Vec2, XLSR, XLS-R, WavLM, HuBERT), with transformer/VGG embeddings and fusion systems (MFA, MTB, SLIM). Back-end classifiers range from GMM, SVM, Logistic Regression, MLP to CNNs (ResNet, LCNN, SENet), CapsNet / ABC-CapsNet, ViT, SSAST, and GNN models such as GCN, GAT, RawGAT-ST, and AASIST. SSL + augmentation dominate performance, although cross-domain generalization, lack of multilingual/partially-fake datasets, noise/codec vulnerability, adversarial fragility, and fairness/explainability remain unresolved challenges.Table I offers a brief snapshot of each defense and its practical limits.

"Spoofing Speech Detection using Temporal CNN"[14] shows that long-range temporal convolution helps against unit-selection spoofing: using five STFT-derived 256-dim features

TABLE I
COMPARATIVE EVALUATION OF DEFENSES AND GUARDRAILS FOR SYNTHETIC IMPERSONATION ATTACKS AT CRITICAL DECISION POINTS.

Approach	Primary layer / integration point	Main strengths at decision points	Structural limitations / failure modes	Best suited role in overall defense stack
Content provenance (C2PA)	Metadata and publishing workflow (capture devices, editing tools, platforms)	Cryptographically signed manifests bind media to creator, device, time, and edit history; provides a "nutrition label" for transparency; fits newsroom/platform workflows and can be required for high-stakes channels.	Depends on ecosystem adoption and trusted signers; fails open when metadata is stripped or missing; does not judge truth — only that specific content was signed; everything outside the ecosystem is "provenance unknown."	Anchor of authenticity for vetted enterprise and political communications; baseline provenance layer for approval chains and public messaging.
Digital watermarking (image, video, audio)	Signal layer (in-band marks embedded into pixels or audio samples)	Works even without metadata; deep-learning watermarking improves imperceptibility and robustness; video watermarks leverage temporal redundancy; audio watermarks can tag synthetic/protected speech across reposts and channels.	AudioMarkBench shows modern codecs (EnCodec, SoundStream, Opus) break watermarks while preserving perceptual quality; all systems fail under white-box and high-query black-box attacks; video watermarking has weak temporal modeling and collusion resistance; demographic fairness variation.	Complementary authenticity signal outside controlled ecosystems; good for tagging executive voices and sensitive assets, but must be treated as soft evidence.
Biometric liveness detection (fingerprint exemplar)	Sensor / biometric subsystem (e.g., fingerprint readers in e-KYC or access control)	Purely software-based, single-image method; no hardware changes; achieves $\approx 90\%$ accuracy with ACE as low as 1.47%; quality-driven features detect many spoof attempts.	Scoped to fingerprints only; performance varies across sensors/datasets; does not defend against audio or multimodal deepfakes; vulnerable to high-quality cooperative spoof fabrication.	Hardening point biometric checkpoints in e-KYC and access workflows; strengthens one modality but insufficient against cross-modal impersonation.
AI-based detection (audio, visual, multimodal)	Algorithmic detection layer (pre-screening calls, media uploads, A/V sessions)	State-of-the-art accuracy in controlled datasets (ASVspoof, ADD, FF++, DFDC, Celeb-DF); uses rich features (spectral, raw waveform, SSL encoders) and diverse back-ends (CNNs, GNNs, Transformers, temporal models, sync models).	Poor cross-domain generalization; highly sensitive to codecs, noise, and partially fake content; vulnerable to adversarial perturbations and linguistic edits; fairness and explainability issues limit high-stakes deployment; high false positives destroy trust.	Front-line triage filter for high-volume pipelines (calls, uploads, onboarding); must be combined with provenance, watermarking, and human review for critical decisions.
Policy, platform behavior, and UX guardrails	Organizational/platform UX layer (moderation policies, interfaces, workflows)	Supports labeling, demotion, contextual warnings; legal guardrails (civil/criminal) create accountability; UX research shows ontology-guided tools and explainable interfaces reduce misinterpretation and prioritize low false-positive rates.	Over-enforcement chills legitimate speech; under-enforcement enables disinformation; strong monitoring raises privacy concerns; bad UX (bare scores, fragmented tools) causes misuse and workflow fragmentation.	Governance layer determining how technical signals are surfaced and acted upon; essential for balancing security, privacy, and civil liberties and preventing misuse of detectors.
Human-in-the-loop (HITL) systems	Human–machine decision layer (analysts, operators, domain experts)	Improves accuracy when humans focus on difficult slices; explanations increase interpretability; workflow-aligned interfaces (ontology-guided analytics, report tools) reduce fragmentation and support calibrated trust.	Expert time is costly; human labels are noisy/bias-prone; bad explanations cause illusion of understanding; unclear division of labor leads to rubber-stamping or ignoring the system; scaling and privacy constraints remain difficult.	Final arbitration for high-risk decisions (large transfers, political content, sensitive onboarding); best when detectors + provenance raise flags requiring contextual human interpretation.

(LMS, IF, BPD, GD, MGD), a Temporal CNN with 512 filters (256×11), max-pooling (100-frame window), and a 2048-sigmoid FFN outperforms MLP baselines in S10 (e.g., LMS MLP 35.24% EER vs. CNN 13.65%), giving 21.59% gains; fusion gives 2.87%→1.41% EER in mismatched setups, and 3.60%→0.41% in matched settings. "What You Read Isn't What You Hear"[13] demonstrates linguistic sensitivity: synonym-level perturbations (WordNet, PWWS, BERT MLM via BAE/BERTAttack, TextFooler) can break detectors across TTS engines (Kokoro, Coqui clones, F5-TTS, OpenAI TTS). One-word edits reduce AASIST-2 bona-fide probability by 67.9%; attack success rates reach 60–82%; commercial API

accuracy collapses from 100%→32%; in a Brad Pitt scam case, edits shift API-A from 0.2%→90.3%. Encoder similarity is the strongest robustness predictor; an XGBoost predictor achieves 76.5% F1.

2) Deep Learning for Visual and Multimodal Deepfake Detection: Visual detection is represented by "Deep Fake Detection for Preventing Audio and Video Frauds…"[11], which uses MTCNN for face-cropping, ResNeXt + ResNet for frame-level feature extraction, and LSTM for temporal modeling. The model is evaluated on FaceForensics++ (590 real, 5,639 fake, ∼2M frames), Celeb-DF (same structure), and DFDC (3,426 participants, 48,190 videos, 25 TB). The literature review covers eye-blink detection (DeepVision, 87.5%), transformer-based detectors (AVFakeNet), Shallow-FakeFaceNet, optical-flow temporal-inconsistency models, multimodal detectors, Haar-wavelet and PRNU approaches, XceptionNet, DeepFake Dissection Network (DFDN, 92.3% talking-head AUC), teeth-based models, FFR-FD, hybrid CNN/RNN pipelines, and lightweight DefakeHop. Metrics include accuracy, F1, precision, AUC-ROC, with the MTCNN+ResNeXt+ResNet+LSTM system reporting very strong performance. "Transformer-Based Feature Compensation and Aggregation (Trans-FCA)"[12] addresses ViT limitations using a ViT-Base backbone (12 blocks, 196 tokens, D=768) with Locality Compensation Blocks (GLCA global–local fusion), Multi-head Clustering Projection (token clustering, best at 16 clusters), Frequency-guided Fusion (FFT→learnable filter→iFFT), and Self-Mutual Learning (KL-divergence). Results: FF++ C23 (AUC 99.85%, ACC 98.60%), C40 (AUC 95.91%, ACC 91.43%); cross-dataset: Celeb-DF 78.57%, TIMIT HQ 89.74%, LQ 89.05%. Ablations confirm every module's contribution, with GLCA and 2D FFM outperforming weaker fusion methods.

3) Joint Audio–Visual Detection and Sync Modeling: "Joint Audio-Visual Deepfake Detection"[10] introduces the first framework capable of identifying manipulation in audio-only, video-only, or both, based on the principle that real speech preserves strong viseme–phoneme synchrony. The architecture uses a 2+1-stream design: an R(2+1)D-18 video network, a 1D CNN for raw-waveform audio, and a sync stream that aligns cross-modal representations using convolution + tiling and optional inter-/intra-attention mechanisms (Eqs. 7–10). Because existing datasets lack aligned AV fakes, the authors build a custom dataset using MTCNN face crops, 10 FPS video, 22.05 kHz audio, 3-second clips, Mel-spectrograms, and vocoders (Griffin-Lim, WORLD, WaveNet, WaveRNN, MelGAN, WaveGlow, PWGAN), including random blurring to mimic unclear TTS artifacts. Trained with Adam (lr 0.0002), the joint model outperforms strong baselines: on FF++ whole-sequence — independent streams 95.83%, late fusion 94.25%, 2+1-stream inter-attention 97.62%; on DFDC — independent 89.36% vs. joint 91.01%. It generalizes better to unseen manipulations, surpassing lip-reading-based detectors like LipForensics. Audio-shuffle ablations collapse audio accuracy to 0.39%, proving that the system learns genuine sync constraints, and t-SNE visualizations show clean separation among real+real, fake+real, real+fake, and fake+fake clusters.

4) Common Themes and Limitations: Across all six papers[2], [10], [11], [12], [13], [14], common trends emerge: pipelines blend handcrafted spectral/phase features, raw-waveform encoders, and self-supervised representations with CNNs, GNNs, Transformers, and temporal/multimodal models (e.g., LSTMs, R(2+1)D, sync-streams). Temporal modeling is crucial (Temporal CNNs for unit-selection spoofing; LSTM visual modeling; joint A/V sync modeling). In-dataset performance is consistently strong across FF++, DFDC, Celeb-DF, and ASVspoof. But all works identify limitations: poor cross-domain and cross-lingual generalization; high vulnerability to unit-selection audio, emerging deepfake architectures, and linguistic adversarial attacks; sensitivity to noise, codec, and channel effects; difficulty with partially fake content; and the need for privacy-preserving, fair, explainable detectors alongside more realistic datasets and continual/adaptive training strategies.

E. Policy, Platform Behavior, and User-Experience Design

Chesney & Citron make a blunt point[5]: if platforms don't step up with real policy guardrails, deepfakes won't just embarrass a few people — they'll rattle institutions and shake public trust. They walk through the policy levers platforms already try to use: limiting or down-ranking harmful synthetic media, slapping on labels or warnings, and plugging in provenance and forensic detection features.[5] But every lever has a cost. Push too hard and you end up smothering satire, art, and harmless creativity; go too soft and you open the door to mass manipulation and political chaos. Add too much surveillance and suddenly you're trampling privacy.[5]

They also dive into the legal safety net[5]: victims of deepfakes can lean on civil liability (defamation, privacy torts like false light or intrusion, emotional distress, economic harms) to fight back. Criminal statutes (fraud, impersonation, extortion, threats, harassment) add teeth for intentional abuse. And regulators — FTC for misleading practices, FCC for political ads, FEC for election integrity — draw the outer boundaries.[5] When deepfakes jump from mischief to national-security threats, governments may escalate to sanctions, covert operations, or even military options.[5] Finally, they point to technological and market guardrails: forensic tools, classifiers, and end-to-end authenticity logs ("life-logs") help create accountability, though life-logs raise their own privacy alarms. The second paper, Wu et al. (CHI '25)[15], zooms in on the people actually fighting these threats — intelligence analysts — and the UX pitfalls that can quietly derail their work. Their warning is crystal clear: don't hand analysts a mysterious "fakeness score" and expect them to magically intuit what it means. Analysts *hate* opaque probabilities; they want explanations they can defend.[15] Explainability becomes a guardrail, not a luxury: reasoning traces, structured output, and clear justification reduce misinterpretation and help analysts write defensible reports.[15]

Wu et al. also expose a very human problem[15]: analysts juggle tools that don't talk to each other. This fragmentation leads to copy-paste mistakes, lost metadata, and broken reasoning chains. Their "Why / Where / What" ontology acts like a

mental GPS, steering analysts toward the right tools, reducing cognitive overload, and improving report clarity.[15] One theme comes up again and again: false positives are deadly. A single incorrect accusation can tank credibility, escalate political tensions, or mislead senior decision-makers.[15] Their prototype UX system tries to minimize this risk by bundling report-generation guardrails, standardized terminology (e.g., ICD-203), and a clean path from analysis → explanation → policy relevance.[15]

When you put both papers together, a set of shared principles jumps out.[5], [15] Guardrails have to steer users away from both extremes — blind trust in deepfake detectors and blind dismissal of real evidence (the liar's dividend).[5] Both papers demand transparency at every step: explain why a detection happened, show provenance, expose limitations, and avoid black-box magic tricks.[5], [15] Every flag or label needs context: what modality was analyzed, what the confidence actually means, why the detector thinks something is off, and where its blind spots lie. And everything has to balance security with civil liberties — Chesney & Citron warn repeatedly that authenticity systems and life-logs can slip into surveillance if not designed carefully.[5]

Taken together, the two works argue that defending against deepfakes is not a one-layer problem.[5], [15] It needs platform policies (labeling, moderation, authenticity standards), legal frameworks (civil and criminal accountability), regulatory boundaries (election rules, advertising oversight, consumer protection), and analyst-side UX that reduces false positives and improves interpretability.[15] Ontology-guided interfaces help analysts make fewer mistakes, while transparency and contextualization help everyone avoid misusing either deepfakes or the detection tools themselves.[5], [15] Both papers converge on the same reality: deepfakes create systemic risk, and the only sustainable defense is a hybrid of strong policy plus thoughtful, human-centered UX design.[5], [15]

VII. Human-in-the-Loop (HITL) Systems

Human-in-the-loop systems combine ML models with human expertise across data selection, labeling, model steering, explanation, validation, and downstream decisions. The surveyed work applies this paradigm in general ML (active learning, RL, XAI)[16], security (human-guided intrusion detection)[17], and media forensics (deepfake analysis tools)[15]. Collectively, these studies show HITL's value but also its cost, fragility, and dependence on equally rigorous human-side design.

A. Strengths

1) Improved accuracy, safety, and decision quality: The HITL surveys show that injecting human feedback into active learning, reinforcement learning, and explainable-AI processes boosts accuracy and safety in domains like autonomous driving, healthcare, robotics, NLP, and visual tasks[16]. Humans correct difficult cases, refine segmentations, choose informative samples, and provide domain knowledge unavailable to purely data-driven systems.

2) Efficient use of scarce expert time: Active learning focuses annotation on the most informative unlabeled samples, reducing cost without degrading performance[16]. In the cyber-intrusion detection study, alert prioritization strategies (Max KL, Max Ratio) direct analysts to high-value cases, cutting detection time by up to 79% while preserving the required false-alarm rate[17]. Analysts avoid low-value noise and intervene where their judgment meaningfully shifts posterior beliefs.

3) Better interpretability and trust calibration: HITL systems built around XAI allow humans to understand model behavior, identify errors, and decide when to override automation[16]. In healthcare, tools like TCAV and case-retrieval improve clinicians' grasp of model concepts and, when combined with human confidence, raise diagnostic accuracy. The deepfake-analyst study shows that textual explanations and ontology-guided analytic selection make detector outputs easier to interpret and justify in formal reports[15].

4) Workflow alignment and analyst empowerment: The deepfake-analyst work demonstrates that aligning HITL tools with real analyst workflows — integrated interfaces, multimodal support, batch processing, report generation — reduces fragmentation and improves confidence[15]. The Digital Media Forensics Ontology helps analysts select analytics by capability, media element, and technical feature, strengthening analytic choice and report clarity[15]. This reflects HITL at both the model and organizational-workflow levels.

B. Failure Modes and Risks

1) High cost and operational complexity: HITL requires expensive expert time, training, and coordination. The survey notes that reconciling inconsistent feedback, ensuring high-quality labels, and managing expert availability creates major overhead[16]. Even when active learning reduces label counts, the remaining samples often require the most skilled annotators. Poor processes can erase theoretical efficiency gains.

2) Biased, noisy, and inconsistent human signals: Human feedback is not ground truth. Surveys report biased, inconsistent annotations and low-quality crowd labels[16]. The intrusion-detection model explicitly includes human investigation error (probability ω), showing performance degradation as ω approaches 0.5[17]. Systematic bias or noise can mislead the model faster — and with more confidence.

3) Overreliance on explanations and misleading XAI: HITL assumes interpretability, but explanation quality directly affects decision-making. The deepfake-analyst study shows that visual explanations (heatmaps, frequency or noise maps) often feel "too technical" without strong baselines or training[15]. In healthcare, bad XAI can actively harm decisions. A key failure mode is the illusion of understanding — users trust or reject outputs for the wrong reasons because the explanations appear convincing but are not informative.

4) Ambiguous human–machine role allocation: The surveys highlight unclear division of labor, difficulty integrating human and machine features, and poorly designed human+machine decision rules[16]. Healthcare studies show that the order of human and AI review affects accuracy. Analysts

in the deepfake study report frustration with fragmented tools and black-box outputs[15]. Without explicit protocols for who acts first, who overrides, and how conflicts are resolved, HITL collapses into either blind rubber-stamping or ignoring the system.

5) Validation, scalability, and ethical/legal constraints: Because human input is subjective and expensive, validating HITL systems at scale is difficult. The surveys note evolving validation criteria, the need for multiple annotators, sampling challenges, and persistent fairness and privacy issues[16] — especially when annotators see sensitive data. In security and deepfake forensics, high false positives are particularly damaging; analysts explicitly rate high FP rates as more harmful than some false negatives[15], [17]. HITL must therefore be tuned for organizational risk, not just raw detection metrics.

C. Overall Summary

Across all studies, HITL succeeds when humans are treated as strategic, limited, and noisy sensors, not perfect oracles. The gains come from targeted human effort — active learning queries, prioritized alerts, ontology-guided analytics — and from providing interpretable outputs and coherent workflows[15], [16], [17]. The failure modes appear when human input is assumed to be cheap, perfectly accurate, or automatically beneficial, without addressing cost, bias, interpretability limits, and validation.

VIII. Conclusion and Future Directions

Synthetic impersonation attacks are no longer rare or hypothetical; they appear repeatedly in real systems that blend people, software, and platforms. The surveyed work shows that convincing synthetic media can be generated from very little data, can transfer across languages and channels, and can be deployed at scale against biometric checks, internal workflows, and public communication pipelines. At the same time, current detection systems—even strong self-supervised and deep models—remain fragile when the data distribution shifts, new generators emerge, codecs change, or attackers apply simple adversarial or linguistic tweaks. This creates an uneven landscape in which attackers move quickly with accessible tools, while defenders work to keep brittle systems robust, fair, and interpretable.

This paper anchors the threat model in three concrete decision-point workflows rather than broad notions of "AI risk": enterprise approval chains, political messaging pipelines, and remote e-KYC verification. In all three settings, attackers target moments when humans or institutions must make fast, high-stakes decisions based on limited cues such as a phone call, a short clip, or a recorded message. Benchmarks like ASVspoof, ADD, FF++, and DFDC capture important parts of this reality, but they still miss critical elements such as noisy real-world channels, partially manipulated content, multilingual and low-resource contexts, and adaptive adversaries who probe detectors until they fail. These gaps matter most at decision points, where one misclassification can have disproportionate consequences.

Our comparison of defense mechanisms shows that no single layer is sufficient. Content provenance standards like C2PA help establish authenticity within cooperating ecosystems but lose effectiveness when manifests are stripped or never attached. Digital watermarking tags the media itself, yet AudioMarkBench results show how modern codecs and strong adversaries can weaken or remove audio watermarks. Biometric liveness detection strengthens individual sensors but does not protect against cross-modal impersonation. AI-based detection is useful for triage but remains brittle, hard to interpret, and politically sensitive at scale. As a result, policy, user-experience guardrails, and human-in-the-loop processes are not optional extras—they determine how technical signals are interpreted, escalated, and challenged in real organizational workflows.

Several research directions follow from this analysis. First, benchmarks should move closer to operational conditions by incorporating realistic platform and telephony codecs, multilingual and low-resource content, partial manipulations, and iterative attack cycles rather than one-shot evaluations. Second, provenance and watermarking should be studied together, examining how in-band and out-of-band signals can be combined, how conflicting signals should be resolved, and how these systems interact with privacy and interoperability requirements. Third, multimodal detection needs to prioritize robustness to linguistic edits, distribution shift, and cross-dataset transfer, building directly on the limitations identified in existing audio, visual, and joint audio–visual detection work. Fourth, human-in-the-loop pipelines and analyst tools need evaluation metrics that reflect organizational risk, including the burden of false positives, escalation costs, and the documented impact of explanations on human decision-making.

Beyond technical progress, a broader governance challenge remains. As synthetic impersonation attacks become more routine, institutions must find ways to handle both genuine and fabricated media without defaulting to blind trust or blanket skepticism. This requires not only stronger detection and authentication mechanisms, but also clear policies, liability structures, transparency norms, and training for end users and analysts. The literature reviewed here shows that synthetic impersonation attacks exploit weaknesses across technical infrastructure, organizational workflows, and societal trust. Building systems that can withstand these pressures will require coordinated advances in machine learning, security engineering, human–computer interaction, law, and organizational design, rather than isolated improvements in any single domain.

References

[1] Firc, A., Malinka, K., & Hanáček, P. (2023). Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. *Frontiers in Artificial Intelligence*, 6, 1125633. https://pmc.ncbi.nlm.nih.gov/articles/PMC10114207/

[2] Zhang, B., Cui, H., Nguyen, V., & Whitty, M. (2025). Audio deepfake detection: What has been achieved and what lies ahead. *Sensors*, 25(7), 1989. https://www.mdpi.com/1424-8220/25/7/1989

[3] Shaaban, O. A., Yildirim, R., & Alguttar, A. A. (2023). Audio Deepfake Approaches. https://ieeexplore.ieee.org/abstract/document/10320354

[4] Liu, X., Wang, X., Sahidullah, M., Patino, J., Delgado, H., & Kinnunen, T. (2021). ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. https://ieeexplore.ieee.org/abstract/document/10155166

[5] Chesney, R., & Citron, D. K. (2018). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753–1819. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954

[6] Liu, H., Guo, M., Jiang, Z., Wang, L., & Gong, N. Z. (2024). AudioMarkBench: Benchmarking robustness of audio watermarking. *NeurIPS 2024*. https://neurips.cc/virtual/2024/poster/97471

[7] Ben Jabra, S., & Ben Farah, M. (2024). Deep learning–based watermarking techniques challenges: A review of current and future trends. *Circuits, Systems, and Signal Processing*, 43(8), 4339–4368. https://link.springer.com/article/10.1007/s00034-024-02651-z

[8] Galbally, J., Alonso-Fernandez, F., Fierrez, J., & Ortega-Garcia, J. A high performance fingerprint liveness detection method based on quality related features. https://www.sciencedirect.com/science/article/pii/S0167739X1000244X

[9] Babu, A., Paul, V., & Baby, D. E. An Investigation of Biometric Liveness Detection Using Various Techniques. https://www.researchgate.net/publication/320652679_An_investigation_of_biometric_liveness_detection_using_various_techniques

[10] Zhou, Y., & Lim, S.-N. (2021). Joint audio-visual deepfake detection. *ICCV 2021*. https://openaccess.thecvf.com/content/ICCV2021/html/Zhou_Joint_Audio-Visual_Deepfake_Detection_ICCV_2021_paper.html

[11] Vaidya, A. O., Dangore, M., Borate, V. K., Raut, N., Mali, Y. K., & Chaudhari, A. Deep Fake Detection for Preventing Audio and Video Frauds Using Advanced Deep Learning Techniques. https://ieeexplore.ieee.org/abstract/document/10689785

[12] Tan, Z., Yang, Z., Miao, C., & Guo, G. (2022). Transformer-Based Feature Compensation and Aggregation for DeepFake Detection. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9919359

[13] Nguyen, B., Shi, S., Ofman, R., & Le, T. (2025). What you read isn't what you hear: Linguistic sensitivity in deepfake speech detection. arXiv:2505.17513. https://arxiv.org/abs/2505.17513

[14] Tian, X., Xiao, X., Chng, E. S., & Li, H. Spoofing speech detection using temporal convolutional neural network. https://ieeexplore.ieee.org/abstract/document/7820738

[15] Wu, Y. K., Sohrawardi, S. J., Gerstner, C. R., & Wright, M. (2025). Understanding and empowering intelligence analysts: User-Centered Design for Deepfake Detection Tools *CHI 2025*. https://dl.acm.org/doi/10.1145/3706598.3713711

[16] Kumar, S., Datta, S., Singh, V., Datta, D., Singh, S. K., & Sharma, R. Applications, Challenges, and Future Directions of Human-in-the-Loop Learning. https://ieeexplore.ieee.org/document/10530996

[17] Kim, Y., Dán, G., & Zhu, Q. Human-in-the-Loop Cyber Intrusion Detection Using Active Learning. https://ieeexplore.ieee.org/document/10613858

Supplementary References

[1] R. San Roman, P. Fernandez, A. Défossez, T. Furon, T. Tran, and H. Elsahar, "Proactive detection of voice cloning with localized watermarking," *arXiv:2401.17264*, 2024. [Online]. Available: https://arxiv.org/abs/2401.17264

[2] Z. Khanjani, G. Watson, and V. P. Janeja, "How deep are the fakes? Focusing on audio deepfake: A survey," *arXiv:2111.14203*, Nov. 2021. [Online]. Available: https://arxiv.org/abs/2111.14203

[3] N. M. Müller, F. Dieckmann, and J. Williams, "Attacker attribution of audio deepfakes," *arXiv:2203.15563*, 2022. [Online]. Available: https://arxiv.org/abs/2203.15563

[4] Y. Wang, H. Guo, G. Wang, B. Chen, and Q. Yan, "VSMask: Defending against voice synthesis attack via real-time predictive perturbation," *arXiv:2305.05736*, 2023. [Online]. Available: https://arxiv.org/abs/2305.05736

[5] J. Sedlmeir, R. Smethurst, A. Rieger, and G. Fridgen, "Digital identities and verifiable credentials," *Business & Information Systems Engineering*. [Online]. Available: https://link.springer.com/article/10.1007/s12599-021-00722-y

[6] Z. Wang, Z. Cheng, J. Xiong, X. Xu, T. Li, B. Veeravalli, and X. Yang, "A timely survey on Vision Transformer for deepfake detection," *arXiv:2405.08463*, 2024. [Online]. Available: https://arxiv.org/abs/2405.08463

[7] P. Neekhara, S. Hussain, X. Zhang, K. Huang, J. McAuley, and F. Koushanfar, "FaceSigns: Semi-fragile neural watermarks for media authentication and countering deepfakes," *arXiv:2204.01960*, 2022. [Online]. Available: https://arxiv.org/abs/2204.01960

[8] M. Pawelec, "Deepfakes and democracy (theory): How synthetic audio-visual media for disinformation and hate speech threaten," *AI and Ethics*. [Online]. Available: https://link.springer.com/article/10.1007/s44206-022-00010-6

[9] J. C. Simmons and J. M. Winograd, "Interoperable provenance authentication of broadcast media using open standards-based metadata, watermarking and cryptography," *arXiv:2405.12336*, 2024. [Online]. Available: https://arxiv.org/abs/2405.12336

[10] J. Lee, X. Zhu, G. Karadzhov, T. Stafford, A. Vlachos, and D. Lee, "Collaborative evaluation of deepfake text with deliberation-enhancing dialogue systems," *arXiv:2503.04945*, 2025. [Online]. Available: https://arxiv.org/abs/2503.04945

[11] M. El Harras and M. A. Salahddine, "Tracking financial crime through code and law: A review of regtech applications in anti-money laundering and terrorism financing," *arXiv:2511.15764*, 2025. [Online]. Available: https://arxiv.org/abs/2511.15764

Note: Online resources and AI-based language tools were used solely for formatting assistance, grammar refinement, and improving clarity of presentation.