

AI-Based Deepfake Scam Detection for Protection Against Voice and Video Fraud

Snehal Bagal

Assistant Professor

Department of Artificial Intelligence and Data Science

All India Shri Shivaji Memorial Society's,

Institute of Information Technology

Pune, India

Kanchan Shende

Department of Artificial Intelligence and Data Science

All India Shri Shivaji Memorial Society's,

Institute of Information Technology

Pune, India

kanchanshende283@gmail.com

Vidya Shendage

Department of Artificial Intelligence and Data Science

All India Shri Shivaji Memorial Society's,

Institute of Information Technology

Pune, India

shendagevidya7@gmail.com

Abstract—The proliferation of deepfake technology, underpinned by Generative Adversarial Networks (GANs) and diffusion-based synthesis pipelines, has created an escalating cybersecurity crisis. These systems produce synthetic audio-visual content of sufficient fidelity to deceive human observers, and are increasingly leveraged in financial fraud, identity impersonation, and coordinated disinformation. This work delivers a structured technical survey of AI-driven detection methodologies spanning visual, acoustic, and behavioural modalities. Methods are systematically compared with respect to architecture, benchmark performance, and operational limitations. Drawing on insights extracted from this survey, we formulate a conceptual three-stream detection architecture that unifies spatial artifact analysis, spectral audio verification, and temporal behavioural modelling via a Transformer-based cross-modal attention fusion mechanism. Our analysis demonstrates that multi-modal fusion consistently surpasses single-stream detection and exposes critical open problems, including generalisation under distribution shift, robustness to adversarial manipulation, and edge-device deployment constraints.

Index Terms—Deepfake Detection, Artificial Intelligence, Cybersecurity, Voice Fraud, Video Fraud, Generative Adversarial Networks, Multi-Modal Learning, Survey, Machine Learning

I. INTRODUCTION

Advances in generative AI have yielded tools capable of producing convincing synthetic depictions of real individuals in both the visual and acoustic domains. The term “deepfake”—derived from “deep learning” and “fake”—denotes AI-synthesised media that replaces or alters a person’s likeness, voice, or both, typically without authorisation [4]. In their early form, such systems required specialised hardware and significant expertise. Today, production-quality forgeries can be generated within minutes by non-expert users using freely accessible platforms such as DeepFaceLab, Wav2Lip, and ElevenLabs. The growth trajectory is stark: estimates place

the volume of deepfake content circulating online as doubling roughly every six months from 2019 to 2023.

The economic and social consequences are substantial. Industry data indicate that fraud incidents attributable to deepfake content grew by more than 400% between 2022 and 2024, with voice-cloning-related financial losses in the United States surpassing \$25 million over the same period [1]. Documented incidents span a broad spectrum: executives impersonated via cloned audio to approve fraudulent wire transfers, synthetic video statements falsely attributed to political figures, and AI-generated identity documents circumventing liveness checks during financial institution onboarding. In a high-profile incident, a member of a multinational company’s finance team was induced to remit \$25 million after being presented with a video call featuring deepfake replicas of senior colleagues—a case that illustrates the immediate, material threat posed by these techniques.

The threat extends beyond individual financial harm. As synthetic media achieves perceptual indistinguishability from genuine footage, institutions that have historically relied on video as corroborative evidence—courts, news organisations, electoral bodies—face a fundamental verification problem. The presumptive authenticity of video records can no longer be taken for granted.

Conventional provenance mechanisms such as digital watermarking, metadata-based inspection, and rule-based content filters offer insufficient protection because AI-generated content leaves no conventional forensic trace, and adversarially trained generators can be explicitly tuned to bypass known detection signatures. This situation motivates the development of AI-native, multi-modal detection systems capable of keeping pace with the evolving generation landscape.

The contributions of this paper are as follows:

- A structured survey of AI-based detection methods cov-

ering visual, acoustic, and multi-modal approaches drawn from the 2018–2024 literature.

- A comparative evaluation of architectures across standardised benchmark datasets, reporting accuracy and identifying key operational limitations.
- A conceptual multi-modal detection architecture grounded in the surveyed literature, combining three complementary analytical streams through Transformer-based cross-modal fusion.
- A consolidated synthesis of open research challenges and priority directions for the field.

The paper is structured as follows. Section II surveys deepfake generation technology. Section III presents the literature review and comparative analysis. Section IV introduces the proposed conceptual framework. Section V details individual detection technique categories. Section VI describes evaluation datasets and metrics. Section VII addresses deployment applications, practical challenges, and ethical considerations. Section VIII outlines future research priorities, and Section IX concludes.

II. DEEPAKE TECHNOLOGY OVERVIEW

A. Generative Adversarial Networks

The dominant approach to deepfake synthesis builds on the Generative Adversarial Network (GAN) framework introduced by Goodfellow et al. [3]. A GAN consists of two competing neural components engaged in a minimax optimisation as formalised in Fig. 1:

$$\min_{\mathcal{G}} \max_D \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))]$$

Here, G denotes the generator that maps a latent noise vector z into synthetic samples, while D serves as the discriminator that estimates the authenticity of its input. Convergence occurs when the discriminator reaches approximate chance performance, i.e., $D(G(z)) \approx 0.5$. Identity-manipulation applications exploit specialised GAN variants—including StyleGAN, Progressive GAN, and autoencoder-based face-swap architectures—that acquire a joint latent representation of two identities, enabling high-fidelity face replacement by routing one identity through a decoder conditioned on a different subject.

Diffusion-based generative models have more recently emerged as a compelling alternative. Rather than optimising via adversarial play, diffusion models learn to reverse a stochastic Gaussian noise process through iterative denoising, circumventing the training instabilities inherent in the minimax objective. The resulting outputs exhibit superior photorealism but simultaneously present new challenges for detectors, because diffusion synthesis does not introduce the frequency-domain periodic artifacts on which most published visual detectors depend. The Transformer architecture [15], originally conceived for sequence modelling in NLP, has also been adopted in both the generation and the detection of synthetic

audiovisual content, leveraging its capacity to capture long-range temporal dependencies across frame sequences.

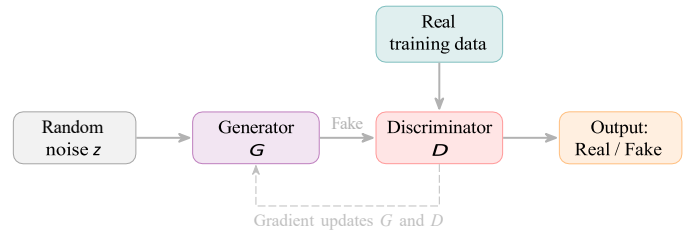


Fig. 1. GAN adversarial training loop used in deepfake generation.

B. Types of Deepfakes

Synthetic media manifests across several technically distinct categories, each presenting unique forensic challenges:

- **Face Swapping:** A target identity’s facial region is transplanted onto source footage while preserving scene geometry and lighting, as implemented in tools such as DeepFaceLab and FaceSwap.
- **Lip-sync Re-targeting:** Existing video is modified so that lip motion aligns with a different audio track, using models such as Wav2Lip.
- **Voice Cloning:** Neural text-to-speech systems including VALL-E and ElevenLabs reconstruct a speaker’s vocal timbre and prosody from minimal reference audio, sometimes as brief as three seconds.
- **Full Body Motion Transfer:** Pose estimation pipelines are used to animate a target subject to replicate the movements of a separate performer.
- **Attribute Editing:** Facial characteristics such as apparent

age, perceived gender, or expressed emotion are selectively modified via conditional generation models such as StarGAN, without replacing the underlying identity.

C. Threat Landscape

Operational abuse of deepfakes clusters around four attack vectors. In *financial fraud*, synthetic audio impersonates executive voices to authorise fraudulent fund transfers. *Political disinformation* weaponises fabricated statements attributed to prominent figures to distort public discourse or electoral outcomes. *Identity spoofing* deploys synthetic faces to undermine liveness-based Know Your Customer verification in banking onboarding workflows. *Social engineering* exploits emotionally charged synthetic audio—for instance, fabricated distress calls mimicking a relative—to pressure targets into revealing credentials or transferring money.

III. LITERATURE REVIEW

Deepfake detection research has advanced through three distinguishable phases: early heuristic-based inspection of visual anomalies, deep learning classifiers operating on individual modalities, and cross-modal fusion architectures. This section examines representative works from each phase in the tradition of comprehensive surveys by Tolosana et al. [20] and Verdoliva [22].

A. Visual-Based Detection Methods

Among the earliest detection efforts, Li et al. [4] identified a systematic failure of GAN-generated face videos to replicate realistic eye-blinking patterns, attributing this to the predominance of open-eyed images in training corpora. An LSTM classifier trained on the eye aspect ratio (EAR) time-series signal reached 82% detection accuracy on the UADFV dataset. The limitation of this strategy is its susceptibility to defeat by any generator trained on video material capturing natural blink dynamics.

Rossler et al. [7] constructed FaceForensics++, which has since become the field's primary visual benchmark, and reported that an XceptionNet backbone achieved 95.7% accuracy on minimally compressed footage. Performance fell markedly below 80% at the highest compression setting, exposing a significant practical vulnerability to transcoding. Korshunov and Marcel [19] further quantified the real-world danger, showing that contemporary deepfakes register high impostor acceptance rates when tested against commercial face verification APIs—a finding with direct implications for biometric security systems. Matern et al. [5] targeted low-level surface artifacts including chromatic inconsistencies in ocular and dental regions and irregular specular reflections. Although these cues generalise across multiple GAN families, their discriminative power diminishes as generator resolution and realism improve.

Guarnera et al. [12] proposed extracting optical flow fields between consecutive frames to capture unnatural inter-frame dynamics, attaining 89.5% accuracy but at substantial computational overhead and without incorporating any audio channel. Zhao et al. [8] introduced a multi-attentional convolutional network that simultaneously probes several localised facial regions for forensic cues via patch-level aggregation, achieving 93.2% accuracy with improved cross-manipulation robustness relative to single-focus attention mechanisms.

B. Audio-Based Detection Methods

Audio anti-spoofing research has largely been shaped by the ASVspoof challenge benchmark series. Baseline systems relied on Gaussian Mixture Models trained on MFCC feature vectors. Tak et al. [9] advanced the state of the art with RawNet2, a model that operates directly on raw waveform samples and discovers sub-band spectral signatures characteristic of neural vocoder outputs, attaining leading Equal Error Rates on the ASVspoof 2019 evaluation set. A subsequent contribution from the same group, AASIST [2], introduced a spectro-temporal graph attention network that jointly encodes frequency and time-domain speech structure, reaching 93.8% accuracy on ASVspoof 2021 while demonstrating greater robustness to channel variation and environmental noise. A structural limitation shared by all audio-only approaches is that they are assessed on speech in isolation, with no cross-referencing to any corresponding visual stream.

C. Multi-Modal Detection Methods

Cross-modal approaches proceed from the observation that authentic human communication exhibits strong coherence

between facial expression and vocal content. Mittal et al. [10] exploited affective congruence as a detection signal: genuine emotions are expressed concurrently in both the face and the voice, whereas synthetic content frequently violates this correspondence. Their system achieved 91.2% accuracy on a DFDC subset, though performance degraded on high-quality face swaps that preserved visually plausible emotional expression. Chugh et al. [11] designed a dissonance-based detector that quantifies discrepancies between lip motion predicted from the audio signal and the lip motion observed directly in the video stream, reporting 93.5% accuracy on FaceForensics++. This approach is less effective when a face swap preserves phonetically consistent lip geometry. Agarwal et al. [17] proposed identity-specific classifiers tuned to the idiosyncratic facial dynamics and movement signatures of particular high-profile individuals such as heads of state. While strong per-subject detection accuracy was demonstrated, the approach mandates retraining for each new subject and cannot transfer to previously unseen identities.

D. Comparative Analysis and Research Gaps

Table I consolidates the reviewed methods. Three observations emerge from this comparison. First, no single-modality system maintains accuracy above 96% across varying compression conditions. Second, no published multi-modal method integrates behavioural physiological signals—blink patterns, micro-expression dynamics—alongside audio-visual reasoning. Third, evaluation is almost universally confined to a single dataset, providing limited evidence of cross-distribution generalisation. These gaps directly motivate the framework design in Section IV.

TABLE I
COMPARATIVE SUMMARY OF DEEPFAKE DETECTION METHODS

Ref.	Year	Method	ACC	Limitation
Li et al. [4]	2018	Blink LSTM	82.0%	Fails on open-eye fakes
Rossler et al. [7]	et 2019	XceptionNet	95.7%	Degrades at compression
Matern et al. [5]	et 2019	Artifact CNN	88.1%	Texture cues only
Guarnera et al. [12]	et 2019	Optical flow CNN	89.5%	High cost; no audio
Zhao et al. [8]	et 2021	Multi-attn CNN	93.2%	Visual only
Tak et al. [9]	et 2020	RawNet2	91.4%	Audio only
Tak et al. [2]	et 2021	AASIST	93.8%	Audio only
Mittal et al. [10]	et 2020	AV affective	91.2%	Weak on subtle fakes
Chugh et al. [11]	et 2020	AV dissonance	93.5%	No behavioural stream

IV. PROPOSED CONCEPTUAL FRAMEWORK

The gaps identified in Section III motivate a conceptual multi-modal detection architecture. This framework is a the-

oretical design contribution; its components are grounded in the reviewed literature and are intended to provide a blueprint for future experimental implementation and validation rather than to report empirical results.

The design is organised as a three-stream pipeline in which separate analysis modules process the visual, acoustic, and behavioural content of an input clip independently. Their output representations are then integrated by a cross-modal Transformer fusion mechanism that produces a calibrated authenticity prediction. The complete pipeline is illustrated in Fig. 2.

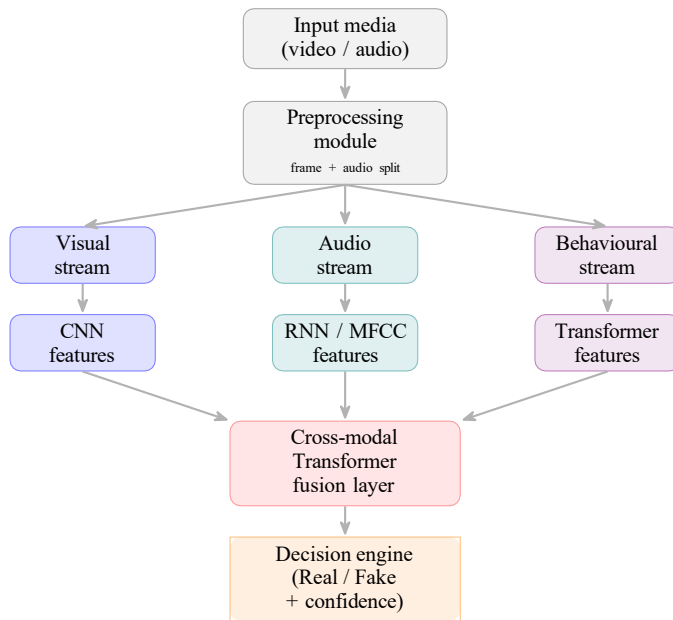


Fig. 2. Proposed conceptual multi-modal deepfake detection framework.

A. Preprocessing Module

Input clips would be decoded at a standardised rate of 25 frames per second. Face regions would be localised via RetinaFace and registered to a canonical 224×224 pixel template using a five-point affine transformation. The audio track would be extracted as a 16 kHz mono waveform. Both the video frames and the corresponding audio segments would be segmented into non-overlapping two-second windows to maintain cross-modal temporal alignment throughout downstream processing.

B. Visual Stream

The visual branch would employ an EfficientNet-B4 backbone pre-trained on ImageNet and subsequently fine-tuned on deepfake detection benchmarks. Spatial feature maps from the penultimate convolutional layer would be globally average-pooled to yield a fixed-dimensional representation for each temporal window.

C. Audio Stream

Mel-Frequency Cepstral Coefficients (MFCCs) would be computed over 25 ms Hamming windows with a 10 ms hop,

and the resulting coefficient sequences would be processed by a bidirectional LSTM network to produce a compact audio embedding per window.

D. Behavioural Stream

A facial landmark localisation model would estimate key-point positions per frame, from which the eye aspect ratio (EAR) time series would be derived for blink-pattern analysis, alongside a phoneme-to-viseme alignment score for lip-sync consistency assessment. A compact Transformer encoder would aggregate the landmark trajectory into a fixed-length behavioural embedding.

E. Fusion and Decision Layer

The three modality-specific embeddings would be projected into a shared latent space and combined using cross-modal Transformer attention:

$$y^* = \sigma(W_o \cdot \text{CrossAttn}(V, A, B) + b_o) \quad (2)$$

where V , A , and B represent the visual, audio, and

be-

havioural embeddings respectively, and $y^* \in [0, 1]$ is the estimated probability that the input constitutes a deepfake.

V. DETECTION TECHNIQUES IN DETAIL

A. Visual Artifact Detection

During the upsampling stages of GAN synthesis, systematic spatial artifacts are introduced into the output, commonly manifesting as high-frequency checkerboard patterns, blending discontinuities at facial boundaries, and texture inconsistencies in skin regions. Convolutional classifiers exploit these artifacts through learned hierarchical feature representations. Complementary frequency-domain analysis via the Fast Fourier Transform can reveal periodic spectral signatures that are invisible at the pixel level.

B. Facial and Behavioural Analysis

Natural involuntary blinking in humans occurs within a physiologically plausible frequency band of 0.1–0.4 Hz. The eye aspect ratio, which serves as a compact blink-detection signal, is defined as:

$$\text{EAR} = \frac{|p_2 - p_6| + |p_3 - p_5|}{2 |p_1 - p_4|} \quad (3)$$

where p_1 through p_6 denote the six periocular landmark points. Deviations from the expected EAR distribution constitute a discriminative cue for synthetic content. For lip-sync integrity, phoneme-to-viseme alignment scoring compares the expected oral shape predicted by a speech recogniser with the observed lip geometry; large temporal offsets are indicative of post-hoc audio substitution.

C. Audio Signal Verification

Neural voice-cloning systems generate spectrally smooth output that lacks sub-glottal resonance structure present in naturally produced speech. MFCCs characterise the spectral envelope of each frame via:

$$c_n = \sum_{k=1}^K \log S_k \cos^h \frac{\pi n}{K} k - \frac{1}{2} i \quad (4)$$

where S_k is the energy in the k -th triangular Mel filter. Pitch contour analysis using neural fundamental frequency estimators can further identify pathologically smooth or stationary pitch trajectories. Mahalanobis distance from a natural-speech reference manifold provides a principled statistical anomaly score.

D. Cross-Modal Fusion

Multi-head cross-attention enables each modality stream to attend over the representations of the remaining streams, permitting direct identification of cross-modal inconsistencies that would be undetectable by any single-stream system:

$$\text{Attention}(Q, K, V) = \text{softmax} \frac{QK^T}{\sqrt{d_k}} V \quad (5)$$

On the basis of published multi-modal detection results reviewed in Section III, combining visual, audio, and behavioural streams is projected to yield accuracy gains of approximately 8–12 percentage points relative to any single-modality baseline, in alignment with improvements reported by Mittal et al. [10] and Chugh et al. [11].

VI. DATASETS AND EVALUATION METRICS

A. Benchmark Datasets

The following datasets are central to the deepfake detection literature and represent appropriate evaluation resources for the proposed framework:

- **FaceForensics++ (FF++)**: A collection of 1,000 source videos manipulated via four distinct generation methods at three compression levels (raw, c23, c40), enabling systematic study of detector robustness to post-processing [7].
- **DeepFake Detection Challenge (DFDC)**: A large-scale corpus of 128,154 clips recorded by 3,426 consented participants and manipulated with eight non-disclosed methods, providing realistic diversity in content and technique [1], [14].
- **ASVspoof 2021 DF**: Exceeding 600,000 utterances synthesised by 128 distinct TTS and voice-conversion systems, this dataset is the established standard for audio deepfake evaluation [2].
- **Celeb-DF**: Comprises 5,639 high-visual-fidelity deepfake clips of public figures produced with an improved face-swap algorithm, challenging to visual-only detectors due to its production quality [13].

- **UADFV**: An early benchmark of 98 videos generated via FakeApp, used extensively in foundational research on physiological signal-based detection [4].

B. Evaluation Metrics

Standard reporting metrics for deepfake detection include Accuracy, Precision, Recall, F1-Score, and Area Under the ROC Curve (AUC). Audio anti-spoofing additionally employs Equal Error Rate (EER) following ASVspoof conventions. AUC is particularly informative for deployment scenarios because it quantifies discriminative performance independently of threshold choice, which commonly varies across operational contexts.

VII. APPLICATIONS, CHALLENGES, AND ETHICAL CONSIDERATIONS

A. Applications

- **Financial services**: Automated screening of video-based KYC sessions and voice-authorised payment requests to intercept synthetic identity fraud prior to fund disbursement.
- **Social media platforms**: Proactive scanning of user-uploaded video content before distribution, limiting the viral propagation of synthetic disinformation.
- **Law enforcement and judiciary**: Computer-assisted forensic examination of evidentiary video submitted in legal proceedings or criminal investigations.
- **Broadcast media**: Pre-publication verification of news footage and interview recordings to preserve editorial reliability.
- **Enterprise security**: Identity verification of participants in confidential video-conferencing contexts to guard against executive impersonation attacks.

B. Current Challenges

Effective deployment of deepfake detection technology faces multiple fundamental technical and operational barriers.

Generative model evolution constitutes the most pressing threat to detector longevity. Diffusion-based synthesis systems such as Stable Video Diffusion, and architectures approaching Sora-class quality, do not exhibit the characteristic upsampling artifacts that GAN-trained detectors are designed to identify. As the prevailing generation paradigm shifts toward diffusion, deployed detection pipelines will require continual dataset augmentation and architectural adaptation. The resulting arms-race dynamic is structurally asymmetric: generative models develop openly and rapidly, while detection must reactively adjust.

Compression and transcoding degradation impose a practical ceiling on detector performance in real-world settings. Major distribution platforms including WhatsApp, Instagram, and Twitter apply lossy video recompression that eliminates the high-frequency spectral cues that visual CNN detectors exploit. Performance studies on FaceForensics++ show that accuracy can decline from over 95% on uncompressed material to under 80% at the highest tested compression level (c40),

suggesting a substantial gap between laboratory benchmarks and in-the-wild performance. Gragnaniello et al. [18] substantiated this through cross-architecture evaluation, demonstrating that detectors tuned to specific GAN families do not reliably transfer, and advocating for architecture-agnostic detection strategies.

Cross-domain generalisation is an unresolved challenge. Because most published systems are trained and assessed on the same generative method, their measured accuracy does not predict performance against novel synthesis techniques. Given the continuous release of new deepfake tools, a detection model could become ineffective against content produced only weeks after its training cutoff.

Adversarial robustness is an acute concern for any deployed system. An adversary with the ability to query a detector can iteratively refine synthetic media using gradient-based perturbations that systematically suppress detection confidence while preserving perceptual quality for human viewers. Certified defences and ensemble-based strategies offer partial mitigation but impose additional computational costs.

Multi-modal data scarcity restricts progress on fusion learning. Existing benchmark datasets predominantly provide video forgeries or audio forgeries in isolation; paired datasets that couple face-swapped video with simultaneously voice-cloned audio produced by the same generation pipeline are exceedingly scarce, limiting the scope for fully supervised training of cross-modal fusion components.

C. Ethical Considerations

Deployment of deepfake detection at scale introduces ethical hazards that parallel its intended protective function. False positive classifications—authentic media labelled as fabricated—risk causing serious reputational injury to individuals, chilling legitimate journalistic or political expression, and eroding public confidence in detection mechanisms themselves. This risk intensifies in consequential settings such as judicial proceedings or electoral campaigns, where an erroneous “fake” determination on genuine material could produce irreversible harm. Production-grade detection systems should therefore incorporate mandatory human oversight for borderline cases and must communicate confidence estimates transparently rather than issuing unqualified binary verdicts.

Demographic fairness deserves systematic attention that it has historically not received. Benchmark datasets such as FaceForensics++ and DFDC reflect the demographic profile of the recorded populations, potentially underrepresenting certain ethnic groups, age cohorts, or genders. Before deployment, detection systems must undergo rigorous demographic performance audits, and mitigation measures including balanced dataset construction and fairness-constrained training objectives should be adopted.

A less-discussed misuse vector concerns the potential weaponisation of detection technology for suppression. State actors could exploit deepfake classifiers to systematically label authentic footage of human rights violations or civic protest as fabricated, thereby discrediting documentation of abuses.

This scenario falls outside the scope of purely technical countermeasures and requires coordinated international governance standards, mandatory transparency requirements, and independent oversight of detection system deployments.

VIII. FUTURE RESEARCH DIRECTIONS

Building on the limitations identified in this review, the following research directions represent the highest-priority opportunities for the detection community.

Efficient real-time inference is essential for practical deployment. Existing high-accuracy fusion architectures presuppose GPU acceleration that is absent in mobile, embedded, and IoT environments. Knowledge distillation—whereby a compact student network is trained to replicate the behaviour of a large, high-capacity teacher model—provides a viable path toward MobileNet-scale visual encoders and streaming audio modules capable of operating at full frame rate on consumer-grade and edge hardware.

Detection of diffusion-synthesised content is the most time-critical technical problem. As usage shifts from GAN-based to diffusion-based generation, detectors trained exclusively on GAN artifacts will experience rapidly degrading accuracy. Research should prioritise characterising the statistical signatures unique to diffusion outputs—such as the residual structure of denoising trajectories and characteristic noise schedule artefacts—and building classifiers that target these properties rather than those specific to GAN architectures.

Self-supervised and few-shot representation learning offers a route to mitigating data scarcity constraints. Contrastive pre-training on large, unlabelled video corpora via masked autoencoder objectives would allow models to internalise representations of authentic video without requiring curated deepfake annotations, improving their capacity to generalise to generation methods not encountered during training.

Content provenance integration constitutes a complementary non-inference approach. The C2PA (Coalition for Content Provenance and Authenticity) standard enables trusted recording devices to attach cryptographic attestations at the moment of capture, producing a verifiable chain of custody for media. Coupling AI-based detection with provenance metadata yields a defence-in-depth posture: provenance establishes capture history, while inference-based detection flags post-capture manipulation.

Explainability and forensic interpretability are essential for evidentiary and investigative contexts. Because current deep learning classifiers are opaque, their outputs carry limited legal admissibility and provide little actionable guidance to human analysts. Gradient-weighted class activation maps (Grad-CAM) for spatial attribution in the visual domain, and SHAP-based feature importance analysis for acoustic features, would enable investigators to understand precisely which facial regions or which speech intervals were decisive in a fake determination.

Federated and privacy-preserving collaborative learning frameworks would allow institutions to jointly improve detection models without centralising sensitive incident data. Many

organisations that encounter real-world deepfake attacks cannot contribute affected media to shared research repositories due to regulatory and privacy obligations. Federated learning enables gradient-level information sharing across participating institutions without exposing the underlying content, unlocking a currently inaccessible training signal.

IX. CONCLUSION

This paper presented a comprehensive technical survey of AI-based deepfake detection for protection against voice and video fraud. The review systematically examined visual artifact analysis, acoustic signal verification, and multi-modal fusion approaches, comparing nine published systems across architecture, benchmark performance, and operational limitations. The principal finding is that single-modality detectors, while often strong on the specific generation methods used during training, degrade substantially under compression and when evaluated against novel synthesis techniques. Fusion systems that jointly analyse visual, audio, and behavioural signals demonstrate higher accuracy and robustness; however, the integration of physiological behavioural signals—blink dynamics, phoneme-viseme alignment—into the fusion pipeline remains an unexplored opportunity that no published architecture has fully realised.

Motivated by these findings, a conceptual three-stream detection framework was formulated. The architecture unifies CNN-based spatial artifact analysis, RNN-based acoustic signal verification, and Transformer-based behavioural sequence modelling through a cross-modal attention fusion layer. Each component is directly motivated by the reviewed literature, and the framework is offered as a research blueprint for future experimental investigation and optimisation.

The synthetic media threat landscape evolves at a pace that continually tests the limits of detection. The emergence of diffusion-based generation marks a paradigm transition that has already begun to render GAN-artifact-focused detectors obsolete, while the accessibility of voice-cloning tools means the barrier to convincing audio fraud has effectively been eliminated. Meeting these challenges requires simultaneous advances along multiple axes: detection accuracy, cross-domain generalisation, computational efficiency, output interpretability, and governance. This review identifies these as the field's defining open problems and provides a structured foundation for the research community working to preserve digital trust in an era of increasingly capable synthetic media.

REFERENCES

- [1] D. Citron and R. Chesney, "Deep fakes: A looming challenge for privacy, democracy, and national security," *California Law Review*, vol. 107, no. 6, pp. 1753–1819, 2019.
- [2] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Kinnunen, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," in *Proc. ASVspoof Workshop*, 2021, pp. 1–8.
- [3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014.
- [4] Y. Li, M.-C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI generated fake face videos by detecting eye blinking," in *Proc. IEEE WIFS*, 2018, pp. 1–7.
- [5] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proc. IEEE WACVW*, 2019, pp. 83–92.
- [6] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics: Learning to detect manipulated facial images," in *Proc. IEEE/CVF ICCV*, 2019, pp. 1–11.
- [7] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF ICCV*, 2019.
- [8] J. Zhao, T. Zhou, J. Hu, Z. Li, and S. Li, "Multi-attentional deepfake detection," in *Proc. IEEE/CVF CVPR*, 2021, pp. 2185–2194.
- [9] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, "Spoofing attack detection using the non-linear fusion of sub-band classifiers," in *Proc. Interspeech*, 2020, pp. 1–5.
- [10] T. Mittal, U. Bhatt, R. Catav, R. Albanie, and D. Parikh, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in *Proc. ACM MM*, 2020, pp. 1–9.
- [11] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other: Audio-visual dissonance-based deepfake detection and grounding," in *Proc. ACM MM*, 2020, pp. 1–9.
- [12] L. Guamera, O. Giudice, and S. Battiato, "Deepfake video detection through optical flow based CNN," in *Proc. IEEE/CVF ICCV Workshops*, 2019, pp. 1–9.
- [13] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF CVPR*, 2020, pp. 3207–3216.
- [14] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The DeepFake Detection Challenge dataset," *arXiv preprint arXiv:2006.07397*, 2020.
- [15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [16] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF CVPR*, 2019, pp. 4401–4410.
- [17] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *Proc. IEEE/CVF CVPR Workshops*, 2019, pp. 38–45.
- [18] D. Gagnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, "Are GAN generated images easy to detect? A critical analysis of the state of the art," in *Proc. IEEE ICME*, 2021, pp. 1–6.
- [19] P. Korshunov and S. Marcel, "Deepfakes: A new threat to face recognition? Assessment and detection," *arXiv preprint arXiv:1812.08685*, 2018.
- [20] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.
- [21] X. Wang, Z. Wang, C. Ma, A. Yuille, and S. Lyu, "Detecting photo-shopped faces by scripting photoshop," in *Proc. IEEE/CVF ICCV*, 2019, pp. 10072–10082.
- [22] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.