

The Detection Arms Race: A Multimodal Survey of Deepfake Forensics from Pixels to Phonemes

Dr. Chayadevi M. L.¹; Harsha Bharadwaj²; Guru Raj Patil³; Keerthana K.⁴

^{1;2;3;4}Department of Computer Science and Engineering, B.N.M. Institute of Technology, Bengaluru, India

Publication Date: 2026/09/30

Abstract: Recent progress in generative deep learning—progressing from early autoencoders and GANs to modern latent diffusion models and neural vocoders—have made synthetic media generation accessible, realistic, and increasingly difficult to identify. While these tools offer productive uses in creative production and accessibility, they create serious security risks through identity fraud, automated disinformation, and unauthorized voice and video impersonation. In response, forensic detection has developed from simple pixel-level artifact checks into complex multimodal pipelines. This survey provides an organized, critical review of current deepfake detection techniques across image, video, audio, and hybrid modalities. We examine the core architectures used in detection—such as Convolutional Neural Networks, Recurrent Networks, Vision Transformers, graph-based phoneme models, and acoustic reverberation estimators—and analyze their behavior on standard benchmarks including FaceForensics++, Celeb-DF, DFDC, ASVspoof, and FakeAVCeleb. We give particular attention to cross-dataset generalization failures, practical latency constraints for on-device deployment, and the growing need for multimodal fusion to maintain detection reliability against next-generation generative models.

Keywords: Deepfake Detection; Digital Media Forensics; Convolutional Neural Networks; Vision Transformers; Audio Forensics; Speech Synthesis Detection; Cross-Dataset Generalization; Direct-to-Reverberant Ratio; Multimodal Fusion.

How to Cite: Dr. Chayadevi M. L.; Harsha Bharadwaj; Guru Raj Patil; Keerthana K. (2026) The Detection Arms Race: A Multimodal Survey of Deepfake Forensics from Pixels to Phonemes. *International Journal of Innovative Science and Research Technology*, 11(9), 1905-1913. <https://doi.org/10.38124/ijisrt/26sep1010>

I. INTRODUCTION

Over the past several years, generative deep learning has evolved from producing low-resolution, visibly flawed synthetic faces to generating photorealistic video sequences and highly convincing cloned voices. Early deepfakes relied heavily on autoencoder-based face replacement pipelines like DeepFaceLab and FaceSwap. Today, the field is dominated by high-capacity Generative Adversarial Networks (such as StyleGAN variants) and Latent Diffusion Models (LDMs) that can generate seamless full-frame synthetic scenes, alter specific facial attributes, or clone a speaker's voice from just few seconds of reference audio. While these technological breakthroughs foster compelling creative applications across cinema, digital avatars, and accessibility aids, their misuse in political deception, financial fraud, impersonation scams, and non-consensual media creation presents an urgent challenge to digital forensics.

The dynamic between media generators and forensic detectors has created an ongoing arms race. Early forensic detectors focused on obvious visual flaws: unnatural eye-blinking rates, irregular facial boundary blending, asymmetric teeth, or color mismatches between facial cutouts and backgrounds. However, as generative architectures improved

and integrated explicit loss penalties for these flaws, simple visual heuristics quickly lost effectiveness. When tested on uncompressed datasets, standard CNN detectors often report accuracy above 95%; yet when evaluated on social media platforms that introduce lossy compression, resizing, and noise, their accuracy routinely drops to near chance levels [1], [2].

To overcome these limitations, recent forensic research has shifted in three key directions. First, researchers have expanded beyond face-only crops to analyze whole-frame context, looking for subtle spatial-temporal contradictions across background objects, lighting angles, and motion trajectories. Second, audio forensics has moved from basic spectral energy checks to analyzing phonetic co-articulation dynamics and physical room reverberation cues. Third, hybrid multimodal frameworks now jointly examine visual, acoustic, and physiological signals to eliminate single-point-of-failure vulnerabilities.

This survey provides an organized assessment of deepfake detection techniques, structuring current methods across image, video, audio, and multimodal pipelines. We review the architectural foundations of existing detectors, examine empirical performance benchmarks, and highlight

practical constraints in real-world deployment. The paper is organized as follows: Section II establishes baseline generation mechanisms, forensic signal spaces, evaluation metrics, and benchmark datasets. Section III reviews foundational surveys and taxonomies [1]–[6]. Section IV details recent methodological advancements [7]–[12], focusing on phoneme modeling, whole-frame transformer detectors, dual-channel reverberation analysis, and mobile probing. Section V discusses unresolved challenges and open research directions [13]–[16], and Section VI concludes the paper.

II. BACKGROUND AND PRELIMINARIES

Understanding how modern deepfake detectors work requires examining the mathematical principles behind the generative models they aim to expose. Generative Adversarial Networks (GANs) formulate media synthesis as a two-player zero-sum game between a generator G and a discriminator D . The generator maps random latent vectors to synthetic images, while D learns to distinguish real samples from synthetic ones. In Variational Autoencoders (VAEs), an encoder compresses an input image into a probabilistic latent distribution, and a decoder reconstructs the image. More recently, Denoising Diffusion Probabilistic Models (DDPMs) generate synthetic data through a parameterized reverse Markov process that step-by-step removes noise from a Gaussian prior, producing exceptional visual realism without the mode-collapse problems common in GAN architectures [13].

Forensic detection systems exploit the subtle physical and statistical traces left behind by these generative processes. In the spatial domain, detectors look for abnormal pixel correlations, color filter array (CFA) interpolation inconsistencies, and residual patterns left by upsampling and deconvolution layers. In the temporal domain, detectors examine optical flow continuity between consecutive frames, eye-blink duration statistics, and remote photoplethysmography (rPPG) pulse signals extracted from subtle skin color fluctuations. In the audio domain, detectors analyze spectral flux, formant transitions, and acoustic reverberation decay profiles that neural voice conversion models fail to recreate.

Detector performance in the literature is typically evaluated using three core metrics. Classification Accuracy (ACC) measures the overall proportion of correct binary decisions across real and fake samples:

$$\text{ACC} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (1)$$

Where TP, TN, FP, and FN represent True Positives, True Negatives, False Positives, and False Negatives, respectively. Because operational deployment requires choosing a decision threshold that balances security against user friction, biometric and speech verification systems use the Equal Error Rate (EER), which marks the threshold θ^* where False Acceptance Rate (FAR) equals False Rejection Rate (FRR):

$$\text{EER} = \text{FAR}(\theta^*) = \text{FRR}(\theta^*) \quad (2)$$

To evaluate performance independently of any single operating threshold, researchers rely on the Area Under the Receiver Operating Characteristic Curve (AUC):

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(t)) \, dt \quad (3)$$

A critical factor in evaluating detection research is the benchmark dataset used. Commonly used video benchmarks include FaceForensics++ (which provides standardized face swaps and reenactments across raw, c23, and c40 compression levels), Celeb-DF (v1 and v2, known for higher visual quality and fewer obvious boundary artifacts), the Deepfake Detection Challenge (DFDC, featuring diverse lighting, camera angles, and demographics), and WildDeepfake [15] (gathered directly from internet sources). For speech and multimodal detection, ASVspoof (2019/2021), FakeAVCeleb [14], and the Sports Press Conference (SPC) dataset provide diverse acoustic environments and conversational speaking styles.

III. OVERVIEW OF STATE-OF-THE-ART TECHNIQUES AND TAXONOMIES

In their comprehensive survey, Heidari et al. [1] establish a four-part taxonomy dividing detection methods into image-based, video-based, audio-based, and hybrid multimodal categories. Image-based detectors primarily use CNN backbones (such as ResNet and EfficientNet) coupled with specialized feature extraction modules like Adaptive Manipulation Trace Extraction Networks (AMTEN) and Error Level Analysis (ELA) to expose local compression discrepancies and boundary artifacts. Video-based detectors leverage recurrent architectures (RNNs, LSTMs) and Temporal Convolutional Networks (TCNs) to capture inter-frame incoherence, unnatural head pose jitter, and desynchronized facial movements. Audio methods target voice cloning artifacts through spectral and waveform analysis, while hybrid approaches fuse multiple streams to minimize false alarms. Heidari et al. also emphasize the role of transfer learning on pre-trained vision models and forensic trace extraction (e.g., Expectation-Maximization and convolutional trace analysis) in improving baseline detection performance.

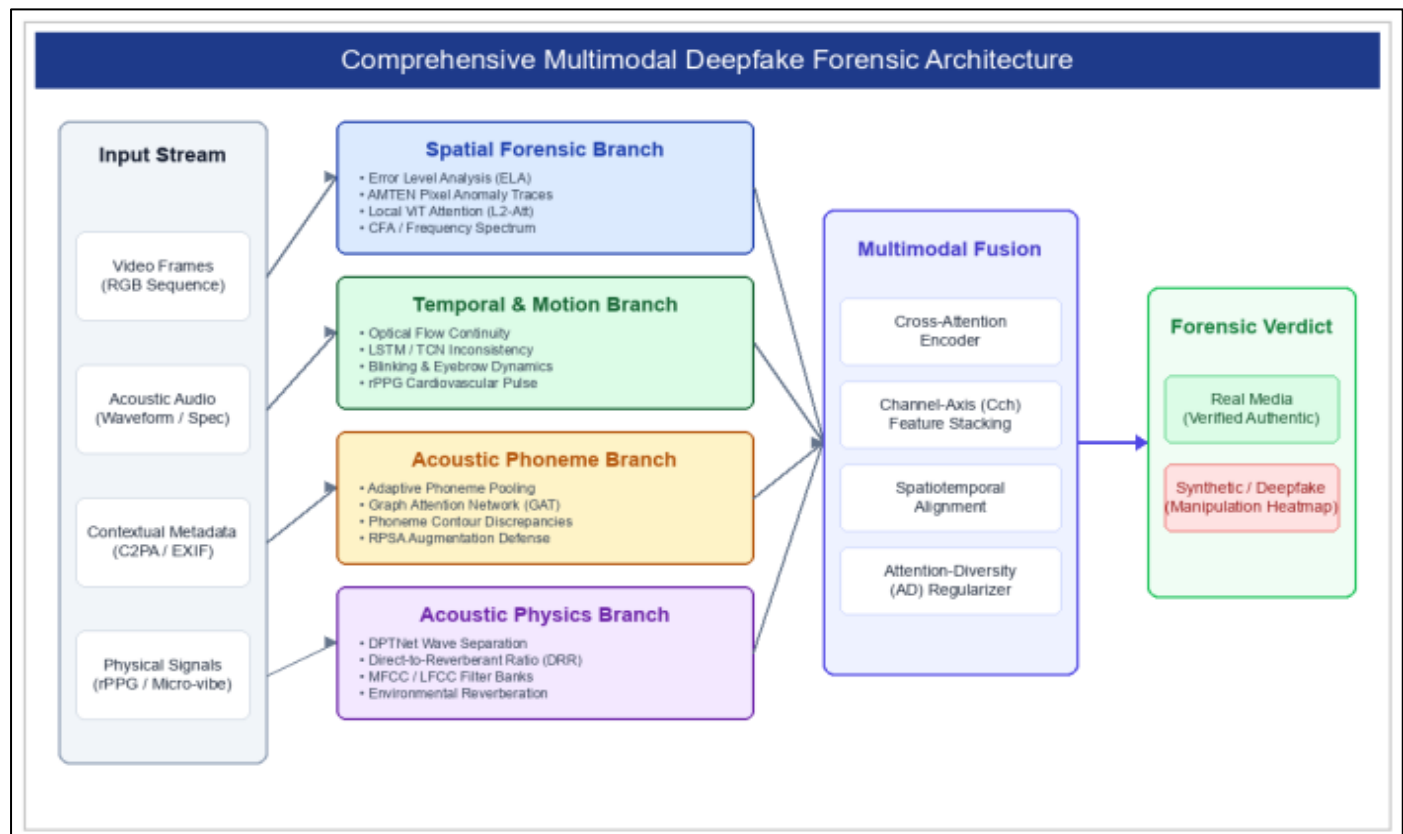


Fig 1 System Architecture of Multimodal Deepfake Forensics, Integrating Visual Spatial, Temporal Motion, Acoustic Phoneme, and Acoustic Physics Branches into a Unified Fusion Classifier.

Pei et al. [2] frame deepfake generation and detection as dual sides of a shared problem: generation as conditioned content synthesis and detection as fine-grained image/video classification. They classify generation methods into four distinct tasks: face swapping (replacing identity while keeping expression), face reenactment (driving target facial motion from source video), talking face generation (animating still portraits from audio speech), and facial attribute editing (modifying age, hair, or expression). On the detection side, they trace the shift from handcrafted spatial and frequency features toward deep end-to-end models, showing that while frequency features (e.g., discrete cosine and Fourier transforms) efficiently spot upsampling artifacts, spatial-temporal deep networks hold up significantly better under heavy lossy compression.

Khan et al. [3] provide an empirical evaluation comparing supervised CNNs against self-supervised foundation models (such as DINO and CLIP). By evaluating models across FakeAVCeleb [14], Celeb-DF-v2, DFDC, and FaceForensics++, they uncover a persistent generalization bottleneck: models achieving near-perfect in-distribution accuracy (AUC > 0.98) frequently experience severe drops (AUC falling below 0.65) when tested on unseen generative tools or datasets with different compression standards. Importantly, their experiments show that simply scaling up model parameters does not fix out-of-distribution vulnerability without explicit domain-invariant training objectives.

Physiological signal analysis offers an alternative detection strategy. Patil et al. [4] survey biological signal inconsistencies in manipulated video, examining involuntary eye-blink dynamics, eyebrow movement trajectories, and remote photoplethysmography (rPPG). Because deep generative networks typically process video on a frame-by-frame basis or across very short temporal windows, they struggle to maintain the periodic, subtle blood volume pulse signals present across real facial skin. Extracting these pulse signals in the frequency domain provides a physically grounded defense that is difficult for standard GAN generators to bypass.

Nguyen et al. [5] review deep autoencoder architectures, GAN variants, and transformer detectors, outlining the dual-use nature of synthetic media across entertainment and cybersecurity. Hashmi et al. [6] focus specifically on audiovisual deepfakes, noting that unimodal detectors are easily fooled when only one modality is manipulated, whereas joint audiovisual models detect subtle timing mismatches between lip movements and spoken phonemes. This observation is further supported by recent multi-scale spatio-temporal fusion architectures [16], which demonstrate that combining visual motion cues with acoustic features produces consistently higher detection stability across diverse compression levels.

IV. ADVANCEMENTS IN DEEPPFAKE DETECTION TECHNIQUES

➤ Phoneme-Level Speech Forensics

Recent breakthroughs in neural Text-to-Speech (TTS) and Voice Conversion (VC) produce synthetic speech with remarkably natural prosody and timbre. However, Zhang et al. [7] point out that modern neural vocoders (such as HiFi-GAN and WaveGlow) and diffusion-based acoustic synthesizers still struggle to accurately capture rapid, fine-grained phonetic co-articulation. During natural human speech, vocal tract dynamics create smooth formant transitions (F1, F2, F3) and consistent glottal pulse phases between adjacent phonemes. Synthetic speech generators often introduce micro-discontinuities and phase inconsistencies at these boundaries. Rather than using static frame-level spectrograms that blur these fleeting transitions, Zhang et al. propose a phoneme-level modeling pipeline.

The pipeline first aligns the speech audio with phonetic boundaries and applies Adaptive Phoneme Pooling, aggregating frame feature vectors x_t over the duration T_p of each phoneme token p :

$$h_p = (1 / |T_p|) \sum_{t \in T_p} x_t \quad (4)$$

This pooling step produces a compact sequence of phoneme representations $\{h_1, h_2, \dots, h_M\}$ that preserves temporal transition dynamics while filtering out frame-level noise. Next, a Graph Attention Network (GAT) models transition relationships across the entire phoneme sequence. In this graph, nodes correspond to phoneme tokens, and edges connect sequential and co-occurring phonemes. The multi-

head attention weight α_{ij} between phonemes i and j is computed as:

$$\alpha_{ij} = \exp(\text{LeakyReLU}(a^T [W h_i \parallel W h_j])) / \sum_{k \in N_i} \exp(\text{LeakyReLU}(a^T [W h_i \parallel W h_k])) \quad (5)$$

Where W is a shared weight matrix, a is a learnable attention parameter vector, and N_i represents the phonetic neighborhood of node i . To prevent the model from simply memorizing speaker-specific vocal traits, Zhang et al. introduce Random Phoneme Substitution Augmentation (RPSA). By randomly swapping a small fraction of phoneme nodes during training with acoustic tokens from other contexts, RPSA forces the graph network to focus on structural transition artifacts rather than individual vocal timbres, significantly improving generalization on unseen voice synthesis engines.

➤ Universal Whole-Frame & Synthetic Video Detection

Most existing deepfake detectors were designed specifically for face-swapping scenarios, assuming the manipulation is confined to a cropped facial bounding box. However, modern generative video tools (such as Sora, Runway Gen-2, and Stable Video Diffusion) can generate entire video scenes from text prompts or edit complex scene backgrounds. In these cases, facial features may look natural or may not be present at all, while forgery traces appear in background geometry, unnatural object physics, shadow inconsistencies, or texture anomalies. Kundu et al. [8] address this gap with UNITE (Universal Network for Video Manipulation Detection), a model designed to detect facial, background, and fully synthetic video manipulations across the entire frame.

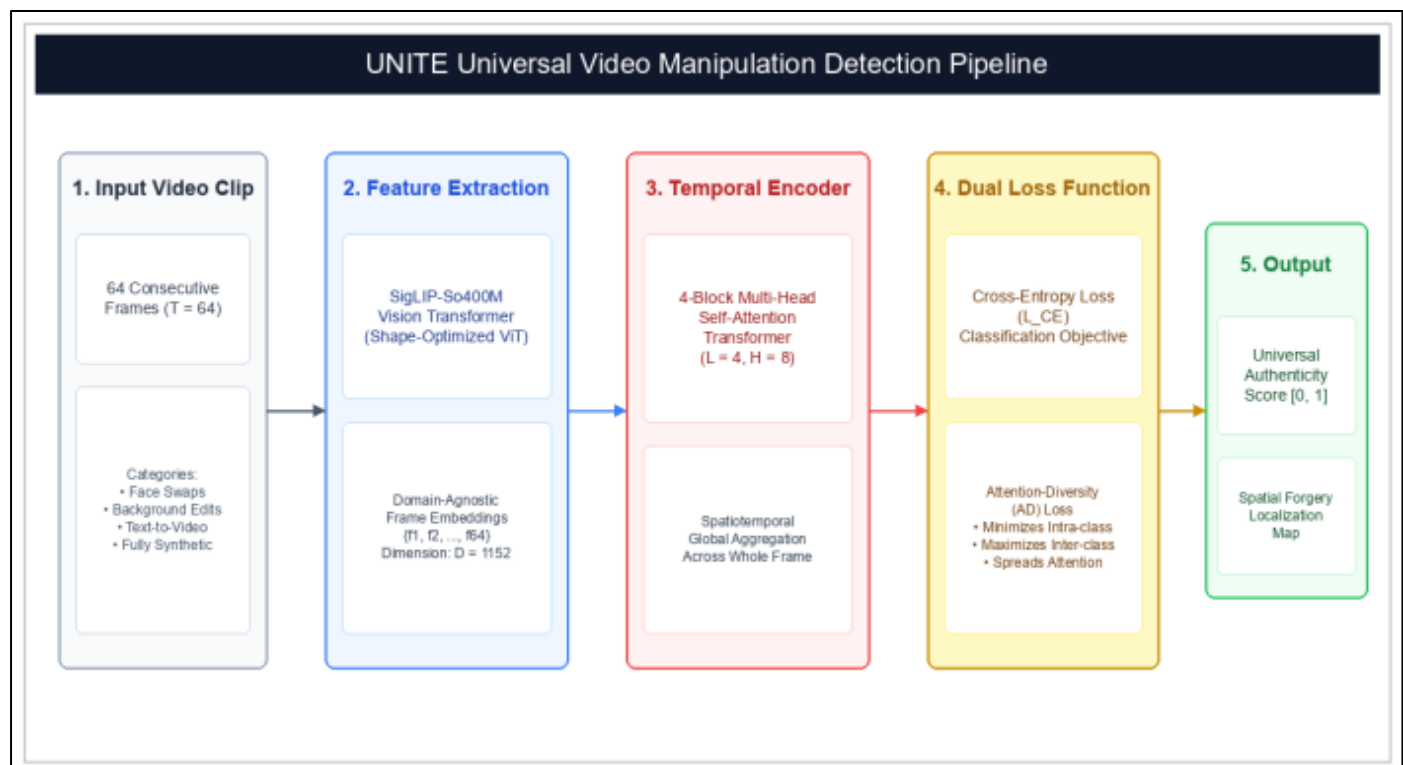


Fig 2 Architecture of the UNITE Universal Video Manipulation Detector, Combining SigLIP ViT Frame Embeddings, Temporal Transformer Encoding, and Attention-Diversity Loss [8].

As shown in Fig. 2, UNITE processes clips of 64 consecutive video frames. Each frame is passed through a pre-trained SigLIP-So400M Vision Transformer (ViT) backbone to extract a 1,152-dimensional domain-invariant feature embedding. Unlike standard CLIP models trained with softmax loss over batch samples, SigLIP uses a pairwise sigmoid loss that avoids cross-sample normalization artifacts, producing feature representations that transfer more reliably across distinct video domains. A learnable 4-block multi-head self-attention transformer encoder then models temporal relationships across the frame sequence.

Standard cross-entropy loss often causes vision transformers to focus almost entirely on high-contrast facial regions. To prevent this attention collapse, Kundu et al. introduce the Attention-Diversity (AD) Loss. The AD loss penalizes feature dispersion around class centroids $c_{\{y_i\}}$

while enforcing an explicit geometric margin m between real and fake class centers c_j and c_k :

$$L_{AD} = (1/B) \sum_{i=1}^B \|z_i - c_{\{y_i\}}\|^2 + \lambda \sum_{j \neq k} \max(0, m - \|c_j - c_k\|_2) \quad (6)$$

$$L_{total} = L_{CE} + \beta \cdot L_{AD} \quad (7)$$

Where B is batch size, z_i is the pooled video embedding, L_{CE} is cross-entropy loss, and β , λ are weighting hyperparameters. By training on a combined dataset of face manipulations (FaceForensics++) and synthetic video engine footage (GTA-V, SAIL-VOS-3D), UNITE achieves strong zero-shot detection performance on unseen video generators. Table 1 summarizes the primary differences between UNITE and conventional face-centric detectors.

Table 1 Comparison Between Unite and Traditional Deepfake Detectors

Feature / Dimension	UNITE Framework [8]	Traditional Deepfake Detectors
Scope of Detection	Detects face, background, and fully synthetic video content	Predominantly face-centric bounding box analysis
Core Architecture	Vision Transformer (ViT) with temporal attention encoding	Standard CNNs (ResNet, EfficientNet) and RNNs/LSTMs
Feature Extraction	Domain-agnostic foundation embeddings (SigLIP-So400M)	Supervised face-specific feature representations
Robustness	Resilient to fully AI-generated and partially manipulated scenes	Struggles with full-frame synthesis and scene edits
Dataset Coverage	Trained on real, face-swap, and synthetic sets (FF++, GTA-V)	Mostly trained on face-swapped datasets (FF++, Celeb-DF)
Generalization	Superior zero-shot adaptation to unseen generative models	Vulnerable to cross-dataset performance drops
Loss Function	Attention-Diversity (AD) Loss + Standard Cross-Entropy	Standard Cross-Entropy Loss only
Deployment Outlook	Optimizable for universal video stream verification	Constrained to face-tracking forensic pipelines

➤ Dual-Channel Acoustic and Reverberation Forensics

A largely overlooked cue in audio forensics is the physical acoustics of the recording environment. When genuine speech is recorded in an enclosed room, sound waves reflect off physical boundaries (walls, floor, ceiling) following the Room Impulse Response (RIR). The recorded signal contains both direct line-of-sight sound and a characteristic pattern of early reflections followed by late reverberant decay governed by room geometry and material absorption. In contrast, neural text-to-speech and voice conversion models generate dry anechoic speech or apply simplistic artificial reverberation filters that violate physical energy decay relationships across frequency bands.

Lee et al. [9] exploit these physical room acoustics by designing a dual-channel forensic framework. They model a single-channel recording $y(t)$ as the sum of direct sound $s_{direct}(t)$, reverberant sound $s_{reverb}(t)$, and background noise $n(t)$:

$$y(t) = s_{direct}(t) + s_{reverb}(t) + n(t) \quad (8)$$

The Direct-to-Reverberant Ratio (DRR), calculated as the ratio of direct signal energy within arrival window τ_d to late reverberant energy, serves as a physically grounded indicator of acoustic authenticity:

$$DRR = 10 \log_{10} \left[\int_0^{\tau_d} s_{direct}^2(t) dt / \int_{\tau_d}^{\infty} s_{reverb}^2(t) dt \right] \quad (9)$$

To separate direct and reverberant components from single-microphone recordings without special hardware, Lee et al. deploy a Dual Path Transformer Network (DPTNet). DPTNet uses alternating intra-chunk and inter-chunk self-attention blocks to separate the raw waveform into estimated direct and reverberant streams, as outlined in Fig. 3.

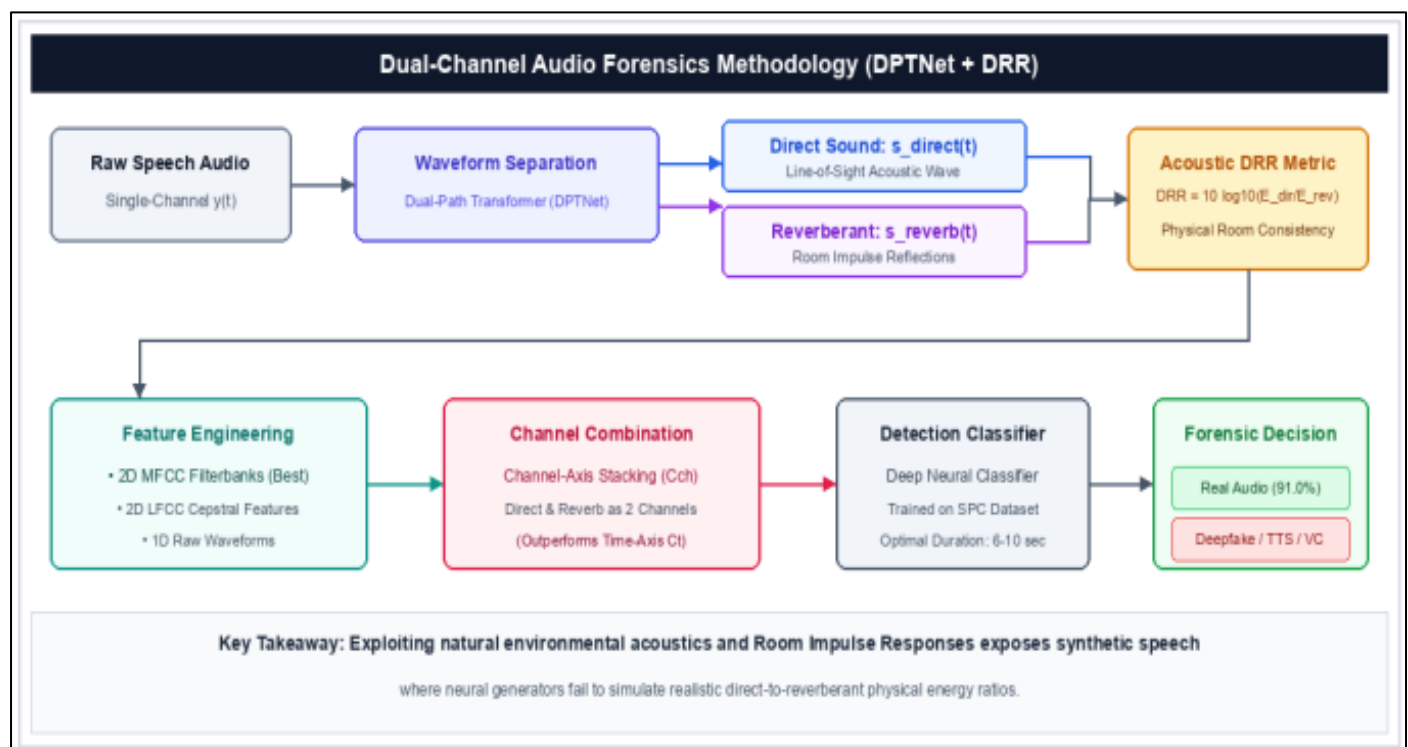


Fig 3 Workflow of the Dual-Channel Audio Deepfake Detector, Using DPTNet Waveform Separation and Channel-Axis Feature Combination [9].

The separated direct and reverberant waveforms are transformed into feature representations: raw 1D waveforms, 2D Mel-Frequency Cepstral Coefficients (MFCCs), and 2D Linear-Frequency Cepstral Coefficients (LFCCs). The features are then combined using either time-axis concatenation (C_t) or channel-axis stacking (C_{ch}). On the challenging Sports Press Conference (SPC) dataset, channel-

axis stacking (C_{ch}) with 2D MFCC features reached the highest accuracy (91.0%), outperforming 1D waveforms (82.4%) and LFCC features (86.1%). Audio duration ablation tests showed that 6 to 10 seconds of speech provides the best operational balance between stable reverberation estimation and low processing latency.

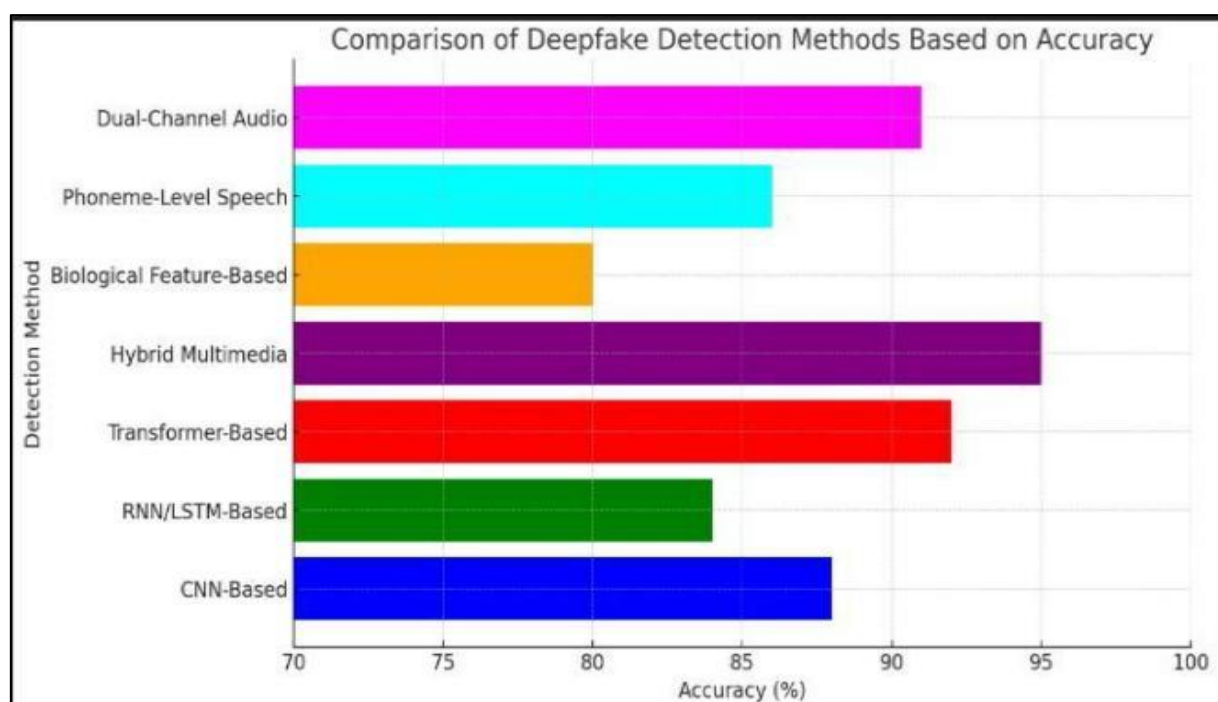


Fig 4 Detection Accuracy Comparison Across Different Forensic Paradigms, Showing Performance Advantages for Hybrid Multimedia and Acoustic Reverberation Models.

➤ Edge Probing and Spatial-Temporal Attention Architectures

To address deepfakes in practical deployment environments, researchers have also explored active hardware probing and localized attention mechanisms. Xie and Luo [10] develop SFake, an active mobile defense system. During live video recording, SFake briefly triggers the smartphone's linear resonant vibration motor to introduce subtle, controlled physical micro-vibrations. Because software-based deepfake filters operate post-capture and cannot anticipate these physical hardware perturbations, they fail to reproduce the resulting camera motion blur synchronously across frames, exposing the manipulation in real time with minimal computational cost.

For passive spatial detection, Nguyen et al. [11] present FakeFormer. Standard Vision Transformers tend to focus heavily on global semantic context, often missing subtle local boundary flaws. FakeFormer addresses this by introducing a Local Attention (L^2 -Att) module that calculates vulnerability weights for image patches, directing attention toward subtle Poisson blending boundaries and color interpolation seams. In parallel, Chen et al. [12] combine spatial attention over texture features (using a boosted ResNet backbone) with temporal attention over optical flow fields, turning fine-grained spatio-temporal inconsistencies into reliable

detection signals. Fig. 4 compares detection accuracy across these different approaches.

➤ Real-Time and On-Device Deployment Paradigms

Moving forensic algorithms from research servers to edge devices (smartphones, teleconferencing apps, and IoT cameras) requires optimizing for strict latency, memory, and power limits. In live audio streaming (such as telephone fraud prevention), systems must process streaming sliding windows rather than complete files. Implementing recursive DRR calculations over 250 ms sub-segments keeps acoustic inference latency under 20 ms.

For video models, compression techniques are essential. Structured channel pruning removes redundant attention heads and convolutional channels, while INT8 Post-Training Quantization (PTQ) reduces memory footprints by approximately 75% with minimal accuracy loss ($< 0.8\%$). Combined with teacher-student knowledge distillation from heavy Vision Transformer ensembles to lightweight MobileNetV4 or Tiny-ViT architectures, edge forensic models can sustain 30+ FPS real-time processing on embedded Neural Processing Units (NPUs). Table 2 provides an overview of the primary methodological paradigms, core architectures, and key detection cues across all forensic modalities.

Table 2 Methods and Key Features of Deepfake Detection Modalities

Detection Modality	Methodological Paradigms & Core Models	Primary Detection Artifacts & Cues
Image-Based Detection	CNNs, AMTEN, Error Level Analysis (ELA), ResNet Backbones	Pixel anomalies, unnatural skin textures, CFA noise, blending seams
Temporal Video Detection	RNNs, LSTMs, Temporal Convolutions (TCNs), Optical Flow	Blinking irregularities, head pose jitter, inter-frame discontinuity
Audio-Based Detection	Spectral Analysis, Dual-Path Transformers (DPTNet), DRR	Pitch variations, spectral envelope anomalies, reverberation decay
Hybrid Multimedia	Cross-Modal Attention Networks, Audiovisual Concurrency	Lip-sync discrepancies, audiovisual temporal desynchronization
Phoneme-Level Detection	Adaptive Phoneme Pooling, Graph Attention Networks (GAT)	Phoneme co-articulation failure, synthetic acoustic transitions

V. CHALLENGES AND FUTURE SCOPE

Despite significant progress across unimodal and multimodal detection, several fundamental challenges remain unresolved for practical forensic deployment:

- **Generalization to Unseen Generative Paradigms:** Detectors trained on specific GAN or autoencoder datasets often experience sharp performance drops when exposed to images generated by Latent Diffusion Models (LDMs) or autoregressive vision models [13]. Because diffusion models operate through iterative denoising rather than convolutional upsampling, they leave behind different high-frequency frequency traces. Future research must focus on meta-learning, domain generalization, and self-supervised representations that isolate generative-model-agnostic inconsistencies.
- **Low-Latency Streaming on Constrained Edge Hardware:** High-performing spatio-temporal and transformer models remain computationally intensive, making them difficult to run directly on consumer hardware during live video

calls or continuous broadcast monitoring. Developing lightweight, recursive streaming architectures that operate within sub-50 ms latency budgets remains an active engineering priority.

- **Deep Cross-Modal & Audiovisual Biometric Fusion:** While visual and acoustic detectors are often developed independently, advanced synthetic media tools now modify audio and video simultaneously [14], [16]. Developing cross-attention architectures that evaluate visual biometric signals alongside acoustic phoneme contours provides stronger protection against single-modality deception.
- **Robustness to Anti-Forensic Filtering and Adversarial Perturbations:** Deep learning detectors remain susceptible to imperceptible adversarial noise and routine post-processing transformations (such as heavy social media re-compression, aggressive downsampling, and gaussian blur). Incorporating adversarial training, randomized smoothing, and compression-invariant objective functions is essential to maintain detection reliability.

- **Dataset Diversity and Standardized Benchmarks:** Many popular training datasets have limited demographic, lighting, and acoustic diversity [15]. Developing standardized, open benchmark datasets that span diverse age groups, skin tones, accents, and recording environments is critical to ensure fair and reliable forensic tools.
- **Digital Provenance and Proactive Watermarking:** Passive detection alone cannot fully solve media authentication. Combining passive forensic detectors with proactive provenance technologies—such as C2PA cryptographic metadata credentials and robust latent watermarks embedded during generative model training—offers a more comprehensive defense strategy.

VI. CONCLUSION

This survey has reviewed current developments in deepfake media forensics, tracing how detection strategies have evolved from localized pixel-level checks to whole-frame Vision Transformers, phoneme-level speech modeling, and dual-channel acoustic reverberation analysis. By examining foundational models [1]–[6], recent architectural innovations [7]–[12], standard evaluation metrics, and edge optimization techniques, we have mapped out key trends and highlighted persistent cross-dataset generalization challenges. As generative models continue to improve in visual realism and acoustic fidelity, future forensic systems will need to move beyond isolated unimodal heuristics toward unified, real-time, adversarial-resilient multimodal pipelines paired with cryptographic media provenance.

ACKNOWLEDGMENT

The authors thank the Department of Computer Science and Engineering at B.N.M. Institute of Technology (BNMIT), Bengaluru, for providing institutional resources, computing infrastructure, and academic guidance that supported this work.

➤ Author Contributions

Dr. Chayadevi M L: research supervision, methodological review, and manuscript validation. Harsha Bharadwaj: taxonomy formulation, multimodal synthesis, mathematical modeling, and manuscript drafting. Guru Raj Patil: visual and temporal forensic analysis, metric evaluations, and comparative tables. Keerthana K.: acoustic forensics review, on-device deployment analysis, and reference curation. All authors reviewed and approved the final manuscript.

➤ Data Availability Statement

All benchmark datasets and reference studies cited in this survey are accessible through public scholarly repositories and challenge platforms (FaceForensics++, Celeb-DF, DFDC, ASVspoof, FakeAVCeleb, WildDeepfake, SPC). Summary tables and taxonomy data are available from the authors upon reasonable academic request.

REFERENCES

- [1]. A. Heidari, N. J. Navimipour, H. Dag, and M. Unal, "Deepfake detection using deep learning methods: A systematic and comprehensive review," *Arch. Comput. Methods Eng.*, vol. 31, no. 3, pp. 1903–1936, 2024.
- [2]. G. Pei, J. Zhang, M. Hu, Z. Zhang, C. Wang, Y. Wu, G. Zhai, J. Yang, C. Shen, and D. Tao, "Deepfake generation and detection: A benchmark and survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, 2025.
- [3]. S. A. Khan and D.-T. Dang-Nguyen, "Deepfake detection: A comparative analysis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 20, no. 3, Art. no. 75, pp. 1–28, 2024.
- [4]. K. Patil, S. Kale, J. Dhokey, and A. Gulhane, "Deepfake detection using biological features: A survey," in *Proc. Int. Conf. Advances in Computing, Communication and Control (ICAC3)*, 2023, pp. 1–7.
- [5]. T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, "Deep learning for deepfake creation and detection: A survey," *Comput. Vis. Image Underst.*, vol. 223, Art. no. 103525, 2022.
- [6]. A. Hashmi, S. A. Shahzad, C.-W. Lin, Y. Tsao, and H.-M. Wang, "Understanding audiovisual deepfake detection: Techniques, challenges, and human factors," *IEEE Trans. Multimedia*, vol. 26, pp. 10582–10599, 2024.
- [7]. K. Zhang, Z. Hua, R. Lan, Y. Zhang, and Y. Guo, "Phoneme-level feature discrepancies: A key to detecting sophisticated speech deepfakes," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 4150–4154.
- [8]. R. Kundu, H. Xiong, V. Mohanty, A. Balachandran, and A. K. Roy-Chowdhury, "Towards a universal synthetic video detector: From face or background manipulations to fully AI-generated content," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 1–11.
- [9]. G. Lee, J. Lee, M. Jung, J. Lee, K. Hong, S. Jung, and Y. Han, "Dual-channel deepfake audio detection: Leveraging direct and reverberant waveforms," *IEEE Access*, vol. 12, pp. 142105–142118, 2024.
- [10]. Z. Xie and J. Luo, "Shaking the fake: Detecting deepfake videos in real time via active probes," in *Proc. 32nd ACM Int. Conf. Multimedia (MM '24)*, 2024, pp. 1095–1104.
- [11]. D. Nguyen, M. Astrid, E. Ghorbel, and D. Aouada, "FakeFormer: Efficient vulnerability-driven transformers for generalisable deepfake detection," in *Proc. European Conf. Computer Vision (ECCV)*, 2024, pp. 1–18.
- [12]. Y. Chen, N. Akhtar, A. Mian, and N. A. H. Haldar, "Deepfake detection with spatio-temporal consistency and attention," *IEEE Trans. Inf. Forensics Security*, early access, 2025.
- [13]. R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," in

- Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.
- [14]. H. Khalid, S. Tariq, M. Kim, and S. S. Woo, “FakeAVCeleb: A novel audio-video multimodal deepfake dataset,” in Proc. 35th Conf. Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track, 2021, pp. 1–12.
- [15]. B. Zi, M. Peng, J. Wang, B. Chen, X. Wang, and B. Xiao, “WildDeepfake: A challenging real-world dataset for deepfake detection,” in Proc. 28th ACM Int. Conf. Multimedia (MM ’20), 2020, pp. 2382–2390.
- [16]. Y. Ju, S. Jia, J. Ke, Y. Yao, and D. Patel, “Fusing multi-scale spatio-temporal features for audio-visual deepfake detection,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 6, pp. 4812–4825, 2024.