International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 5, Sep-Oct-2026
, ISSN (Online): 2395-566X

Deepfake Detection Using AI: A Review Of Audio, Visual And Audio–Visual Methods For Digital Forensics

Kokane Sanika Santosh[1]; Prof. M. C. Kolse[2]; Prof. S. K. Said[3]
Department Of Artificial Intelligence And Data Science Jaihind College Of Engineering Pune, Maharashtra, India

Abstract—Deepfakes are realistic manipulated images, videos and audio created with deep learning and generative models that can imitate a person's face, voice, expressions or movements. Although useful in entertainment and education, they are misused for misinformation, fake identities, fraud and reputational harm. This paper reviews AI-based deepfake detection across both visual and audio modalities. It explains generation methods, then examines visual detection from facial texture, lighting, eye and lip movement and inter-frame differences using convolutional and attention-based networks; generalizable detectors for fully synthetic images; audio detection from spectral and prosodic voice characteristics using anti-spoofing models; and audio–visual methods that detect mismatches between speech and lip motion. Benchmarks including FaceForensics++, ASVspoof and FakeAVCeleb are discussed. The detection process of collecting and preprocessing media, extracting features and applying machine learning or deep learning models is formalized in a five-layer architecture. Research gaps, objectives, a comparison of modalities and challenges such as generalization, compression, adversarial evasion and explainability for forensic use are presented. The review concludes that combining audio and visual evidence gives the most reliable detection for social media, journalism, banking, law enforcement and digital forensics.

Index Terms—Deepfake detection, audio deepfakes, voice anti-spoofing, audio–visual analysis, digital forensics, deep learning, synthetic media

I. INTRODUCTION

1.1 Motivation

Synthetic media have reached the point where many people cannot tell a fake video or cloned voice from the real thing. Criminals have used cloned voices to impersonate relatives or executives and request money transfers, and fake videos of public figures spread rapidly online. Detecting such content is now important for banks, newsrooms, courts and social platforms.

Mirsky and Lee [1] surveyed both the creation and detection of deepfakes, highlighting that video and voice manipulations require different detection techniques. Many early detectors focused only on faces, but real attacks increasingly combine synthetic video with synthetic speech.

This paper reviews visual, audio and audio–visual deepfake detection and proposes a forensic detection architecture.

1.2 Objectives Of The Review

- **O1:** To review visual and audio deepfake generation and detection methods.
- **O2:** To examine audio–visual consistency approaches.
- **O3:** To propose a forensic deepfake detection architecture.
- **O4:** To compare modalities and methods.
- **O5:** To identify challenges and future research directions.

II. OVERVIEW OF KEY CONCEPTS

2.1 How Deepfakes Are Made

Visual deepfakes use autoencoders, generative adversarial networks [2] and style-based generators [3] to swap or synthesize faces. Audio deepfakes use text-to-speech and voice-conversion models to imitate a target speaker. Fig. 1 summarizes manipulation types and detection modalities.

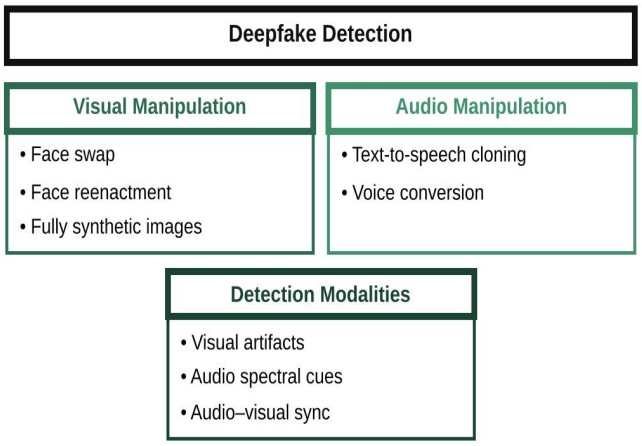


Fig. 1. Deepfake manipulation types and detection modalities.

© 2026 IJSRET

2.2 Detection As Classification

Detection is usually posed as binary classification of real versus fake, at the level of a frame, an utterance or a whole video, sometimes with localization of the manipulated region or time segment.

III. REVIEW OF EXISTING WORK

Visual. Rössler *et al.* [4] provided FaceForensics++ and strong CNN baselines. Zhao *et al.* [5] proposed multi-attentional detection that focuses on several local regions to capture subtle artifacts. Haliassos *et al.* [6] used lip-movement features learned from lip-reading to build detectors that generalize better and resist compression. Wang *et al.* [7] showed that a classifier trained on images from one generator can detect images from others with proper augmentation, and Ojha *et al.* [8] used features from a large pre-trained vision–language model to build a more universal fake image detector.

Audio. The ASVspoof 2019 challenge [9] and database [10] provided synthesized, converted and replayed speech for training anti-spoofing systems. Jung *et al.* [11] proposed AASIST, which uses graph attention over spectral and temporal features and performed strongly on these benchmarks.

Audio–visual. Chugh *et al.* [12] detected deepfakes by measuring dissonance between audio and visual streams, and Khalid *et al.* [13] released FakeAVCeleb, a dataset in which video, audio or both are manipulated. Tolosana *et al.* [14] reviewed face manipulation detection. Table I summarizes these works.

Sr. No.	Ref.	Year	Methodology	Findings	Research Gap
1	[9]	2019	ASVspoof 2019 challenge	Benchmark for audio spoofing	English speech
2	[12]	2020	Audio–visual dissonance	Detects mismatched streams	Needs speech in video
3	[1]	2021	Survey of creation and detection	Covers video and audio	Survey only
4	[5]	2021	Multi-attentional detection	Captures subtle local artifacts	Compute cost
5	[6]	2021	Lip-reading features (LipForensics)	Better generalization and robustness	Needs mouth visible
6	[13]	2021	FakeAVCeleb dataset	Joint audio–video fakes	Limited languages
7	[11]	2022	AASIST graph attention	Strong audio anti-spoofing	Unseen attacks remain hard
8	[8]	2023	Universal fake image detector	Generalizes across generators	Images, not video

TABLE I Summary of Literature Survey

IV. IDENTIFIED RESEARCH GAPS

1. 1) Detectors often fail on new generators and on content recompressed by social media.

2. 2) Audio deepfakes in Indian languages are under-represented in datasets.

3. 3) Joint audio–visual detection is less mature than single-modality detection.

4. 4) Forensic use requires explanations and evidence standards that most detectors lack.

5. 5) Attackers can apply adversarial perturbations to evade detection.

V. SYSTEM ARCHITECTURE

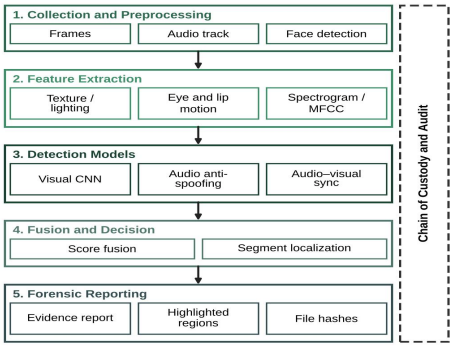


Fig. 2 shows the proposed architecture.

Fig. 2. Proposed forensic deepfake detection architecture.

© 2026 IJSRET

5.1 Collection And Preprocessing Layer

Media files are collected with metadata; videos are split into frames and audio tracks; faces are detected and audio is resampled.

B. Feature Extraction Layer

Visual features capture skin texture, lighting, eye and lip movement and inter-frame differences; audio features include spectrograms, cepstral coefficients and prosody.

5.2 Detection Model Layer

A visual network such as EfficientNet [15] scores frames, an anti-spoofing network scores audio and an audio–visual model measures lip–speech synchronization.

5.3 Fusion And Decision Layer

Scores are combined into a final decision with confidence, and suspicious time segments are localized.

5.4 Forensic Reporting Layer

Reports include scores, highlighted regions and segments, processing details and hashes of the analysed files for chain of custody.

VI. IMPLEMENTATION METHODOLOGY

Fig. 3 describes the workflow. Audio anti-spoofing systems are evaluated with the equal error rate (EER), the operating point where the false acceptance rate equals the false rejection rate:

$$FAR(\tau^*) = FRR(\tau^*) = EER \quad (1)$$

and the audio–visual dissonance score is the distance between audio and visual embeddings,

$$D = 1 - cos(e_audio, e_visual) \quad (2)$$

with high D indicating mismatch.

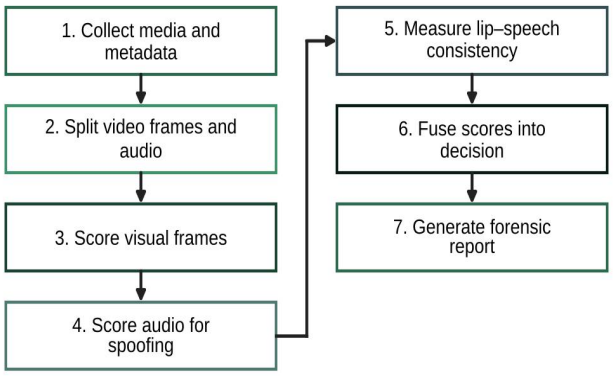


Fig. 3. Audio–visual deepfake detection workflow.

VII. PERFORMANCE COMPARISON

Table II compares modalities. Visual detectors are mature, audio detectors are essential for voice fraud, and audio–visual methods catch mismatches that single modalities miss.

Modality	Detects	Maturity	Limitation
Visual only	Face manipulation	High	Misses voice fakes
Audio only	Voice cloning	Medium–High	Misses video fakes
Audio–visual	Mismatched streams	Medium	Needs both streams
Fusion of all	Most manipulation types	Emerging	Complexity

TABLE II Comparison of Detection Modalities

VIII. OPEN ISSUES AND FUTURE DIRECTIONS

Table III summarizes challenges. Future work includes multilingual audio deepfake datasets, content provenance and watermarking, real-time detection in video calls and standards for presenting AI evidence in court.

Challenge	Mitigation
New generators	Generalizable features; continual training
Compression	Train on recompressed media
Indian languages	Multilingual audio datasets
Adversarial evasion	Robust training; ensembles
Forensic acceptance	Explainable reports; standards

TABLE III Challenges and Mitigation

IX. CONCLUSION

Deepfakes now affect both what we see and what we hear. The reviewed literature shows effective visual detectors, strong audio anti-spoofing models and promising audio–visual consistency methods. Combining these modalities in a forensic pipeline that explains its findings gives the most reliable defence for social media monitoring, journalism, banking, law enforcement and digital forensics.

9.1 Acknowledgment

© 2026 IJSRET

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Jaihind College of Engineering, Pune, for providing the facilities, guidance and encouragement needed to carry out this work. The authors also thank the faculty members and colleagues whose valuable suggestions helped improve the quality of this review.

REFERENCES

1. Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," ACM Comput. Surv., vol. 54, no. 1, Art. no. 7, 2021.
2. I. Goodfellow et al., "Generative adversarial nets," in Proc. NeurIPS, vol. 27, 2014.
3. 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.
4. A. Rössler, 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.
5. H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in Proc. IEEE/CVF CVPR, 2021, pp. 2185–2194.
6. A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in Proc. IEEE/CVF CVPR, 2021, pp. 5039–5049.
7. S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in Proc. IEEE/CVF CVPR, 2020, pp. 8695–8704.
8. U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in Proc. IEEE/CVF CVPR, 2023, pp. 24480–24489.
9. M. Todisco et al., "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in Proc. Interspeech, 2019, pp. 1008–1012.
10. X. Wang et al., "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," Comput. Speech Lang., vol. 64, Art. no. 101114, 2020.
11. J.-w. Jung et al., "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in Proc. IEEE ICASSP, 2022, pp. 6367–6371.
12. K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other—Audio-visual dissonance-based deepfake detection and localization," in Proc. 28th ACM Int. Conf. Multimedia, 2020, pp. 439–447.
13. H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," in Proc. NeurIPS Datasets and Benchmarks Track, 2021.
14. R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and beyond: A survey of face manipulation and fake detection," Inf. Fusion, vol. 64, pp. 131–148, 2020.
15. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105–6114.

© 2026 IJSRET