International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Deepfake Detection Using Artificial Intelligence: A Review of Generation Techniques, Visual and Audio Forensics, Multimodal Analysis, Benchmark Datasets, and Deployable Detection Systems

Gaurav Belgamwar
Dept. of Computer Science and Engineering
Guru Nanak Institute of Technology, Nagpur, India
gauravbelgamwar8@gmail.com

Himanshu Zol
Dept. of Computer Science and Engineering
Guru Nanak Institute of Technology, Nagpur, India
himanshuzol063@gmail.com

Mauli Bade
Dept. of Computer Science and Engineering
Guru Nanak Institute of Technology, Nagpur, India
bademauli21@gmail.com

Jagruti Ghatole
Project Guide
Dept. of Computer Science and Engineering
Guru Nanak Institute of Technology, Nagpur, India
jagrutighatole@gmail.com

Dr. Bharati Sayankar
Head of Department
Dept. of Computer Science and Engineering
Guru Nanak Institute of Technology, Nagpur, India
bharati.sayankar@gmail.com

Dr. Sudhir N. Shelke
Principal
Guru Nanak Institute of Technology, Nagpur, India
gni.gnitprincipal@gmail.com

Abstract—Deepfakes, which are fake images, videos, and audio created with generative artificial intelligence (AI), can now make a person appear to say or do something they never did, and they are increasingly misused for fraud, misinformation, and identity theft. This review surveys research on AI-based deepfake detection. We organize the literature into eight themes: generation techniques, visual artifact-based detection, frequency-domain and low-quality media analysis, generalization and adversarial robustness, audio and speech deepfake detection, multimodal audio-visual detection, benchmark datasets, and deployment and governance. For each of them we summarize representative work, the techniques used, and the limitations reported by the original authors. We then compare detection approaches and datasets, and use a proposed web-based detection system as a reference case that shows how the themes combine in one pipeline. The review identifies open challenges, including generalization to unseen and diffusion-based forgeries, robustness to compression and adversarial perturbation, in-the-wild audio, joint audio-visual benchmarks, and lightweight real-time deployment. We conclude with research directions for trustworthy and accessible deepfake detection.

Index Terms—Deepfake detection, artificial intelligence, deep learning, computer vision, audio spoofing, multimodal forensics, generalization, benchmark datasets, survey.

I. Introduction

A deepfake is a fake image, video, or audio recording created with AI that makes a person appear to say or do something they never actually said or did. Since generative adversarial networks (GANs) were introduced, the quality of synthetic media has improved, and the tools needed to produce it have become widely available. Deepfakes can be used for entertainment, but they are also misused for fraud, fake news, cybercrime, and the spread of misinformation. Detecting them with the human eye is becoming very difficult, which motivates automatic detection based on machine learning and deep learning that analyzes facial expressions, eye movements, lip synchronization, voice patterns, and image quality.

Research on deepfake detection is extensive but spread across computer vision, speech processing, multimedia forensics, and security. Earlier surveys describe generation and detection in general, but there is little consolidated guidance that links the limitations reported in individual studies to the design of an end-user detection system. This paper reviews that body of work with three objectives: (i) to organize the relevant research into a taxonomy of detection approaches; (ii) to compare approaches and datasets and their trade-offs; and (iii) to identify open problems. We use a proposed web-based AI detection system as a running reference case.

A. Review Methodology

The review follows the planning, conducting, and reporting stages recommended for structured literature reviews. Sources were drawn from peer-reviewed journals and conference proceedings (including IEEE Access, IEEE Transactions on Artificial Intelligence, and the major computer vision conferences), dataset papers, and the NIST AI Risk Management Framework. Inclusion criteria were relevance to at least one theme in the taxonomy and a verifiable published source. Limitations quoted for each study are those stated or implied by the original authors. The review is therefore integrative rather than a quantitative meta-analysis.

B. Contributions

1) An eight-theme taxonomy of AI-based deepfake detection (Fig. 1).
2) A thematic synthesis of representative work for each theme, with the limitations and future scope reported by the authors (Table I).
3) Comparative analyses of detection approaches (Table II) and benchmark datasets (Table III).
4) A mapping of the themes to a proposed reference system (Fig. 2, Table IV) and a set of open challenges and research directions.

II. Background

A. Deepfakes and Generative Models

Deepfakes are produced by generative models, most commonly GANs and autoencoders, which learn to synthesize or replace a

face or a voice. Common manipulation types are face swapping, in which one identity replaces another, and face reenactment, in which expressions or lip movements are transferred. More recent diffusion models can produce images with clean spatial features that lack the fingerprints of earlier GAN-based tools, which makes detection harder.

B. Detection as a Classification Problem

Detection is usually posed as binary classification of an image, a video, or an audio clip as real or fake, using supervised machine learning and deep neural networks. Classifiers can use spatial cues within a frame, temporal cues across frames, frequency-domain cues, audio features, or cross-modal consistency between sound and lip movement.

C. Evaluation and Generalization

Detectors are commonly reported with accuracy, area under the ROC curve, and false acceptance or rejection rates. A recurring finding is that performance falls sharply when a detector is tested on a dataset or generator that it was not trained on. Cross-dataset testing is therefore essential for judging real-world usefulness.

III. Taxonomy of Detection Approaches

Fig. 1 presents the taxonomy used to structure this review. The eight themes cover how deepfakes are made (generation techniques), how they are detected in each modality (visual artifacts, frequency and compression cues, audio, and multimodal analysis), what limits detectors in practice (generalization and robustness), how they are evaluated (datasets and benchmarks), and how they are put to use (deployment and governance).

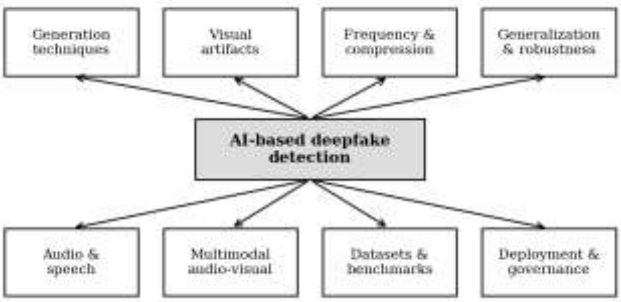


Fig. 1. Taxonomy of AI-based deepfake detection.

IV. Thematic Review

A. Deepfake Generation Techniques

Face swapping and reenactment pipelines typically detect a face, synthesize a replacement, and blend it into the original frame. The FaceForensics++ benchmark, for example, covers four manipulation methods applied to the same source videos. Rana *et al.* review the field systematically and report that existing datasets and detectors were largely optimized for legacy GAN architectures, which makes them fragile against newer blending techniques. Mirsky and Lee organize generation methods by modality and by how well they generalize across identities.

B. Visual Artifact-Based Detection

Many detectors look for traces left by the synthesis pipeline. Li and Lyu observe that the face region is warped to match the target, which leaves a resolution mismatch between the face and its surroundings, and they train convolutional networks to find this warping artifact. Convolutional backbones such as XceptionNet have been used as strong baselines on FaceForensics++. These methods are efficient and fairly interpretable, but their dependence on generator-specific artifacts limits them when new generators appear.

C. Frequency-Domain and Low-Quality Media

Media shared online is usually resized and compressed, which weakens spatial clues. Chhabra *et al.* address low-quality deepfakes by identifying unseen hidden artifacts with a visibility matrix approach and find that heavily compressed fakes leave distinct frequency-domain and visibility-based anomalies. Their results also show that performance drops significantly under extreme downsampling or multi-layer compression, which motivates compression-resilient and multi-scale feature extractors.

D. Generalization and Adversarial Robustness

Tran *et al.* use meta-learning so that a detector adapts quickly to unseen forgery types with few fine-tuning samples; the cost is high computational overhead during meta-optimization, which limits prototyping and real-time use. He *et al.* report that combining spatial-frequency prompt learning with relation learning improves the identification of unseen generative manipulations, while detectors still struggle with rapidly evolving diffusion models. Rana *et al.* note high vulnerability to adversarial perturbations and heavy cross-dataset degradation on unseen forgery pipelines.

E. Audio and Speech Deepfake Detection

Synthetic and cloned voices are a growing source of fraud. Jung *et al.* introduce SpoofCeleb, a large-scale dataset for speech deepfake detection and speaker verification built from real-world noisy audio, and show that training on in-the-wild data sharply reduces false acceptance compared with clean studio-recorded data. Models trained only on clean data can fail in reverberant environments with background noise, so robust acoustic features are needed.

F. Multimodal Audio-Visual Detection

A forged video may manipulate the face, the voice, or both. Salvi *et al.* combine text-to-speech generation with dynamic time warping to build multimodal forgeries, which exposes weaknesses in detectors that examine only one modality. Because most benchmarks cover a single modality, checks of audio-visual synchronization remain underdeveloped, and joint benchmarks for streaming media are still needed.

G. Benchmark Datasets

Dataset quality shapes what detectors learn. Early datasets were small, such as Deepfake-TIMIT; later datasets are larger and more varied, including FaceForensics++, Celeb-DF, the DFDC dataset, DeeperForensics-1.0, and FFIW-10K for videos with multiple faces. Recent surveys also catalogue these resources: one lists Deepfake-TIMIT, DeeperForensics-1.0, Vox-Deepfake, and FFIW-10K, and another covers FaceForensics++ and Celeb-DF as common benchmarks. Table III summarizes the main datasets.

H. Deployment, Explainability, and Governance

A detector is useful only if people can use it. Practical systems wrap trained models in a web interface and API, return a clear risk score or flag, and keep logs for review. Risk scoring has a long history in statistical fraud detection. The NIST AI Risk Management Framework lists characteristics of trustworthy AI, including validity and reliability, security and resilience, transparency, explainability, and privacy, which apply directly to detection tools that may affect a person's reputation. Lightweight

backbones and efficient optimizers are needed for edge deployment.

TABLE I
SUMMARY OF KEY STUDIES AND REPORTED LIMITATIONS

Study	Focus	Reported limitation	Future scope
Rana et al.	Systematic review of deepfake detection	Fragile on unseen forgeries; vulnerable to adversarial attacks	Generalized, attack-robust detectors
Chhabra et al.	Low-quality deepfakes via hidden artifacts	Drops under extreme downsampling or compression	Compression-resilient, multi-scale features
Tran et al.	Meta-learning for generalization	High meta-optimization overhead	Lightweight meta-optimizers for edge use
He et al.	Spatial-frequency prompt and relation learning	Struggles with diffusion-model outputs	Dynamic prompt tuning; graph-based relations
Salvi et al.	Multimodal forgery data and evaluation	Single-modality benchmarks; no sync checks	Joint audio-visual benchmarks
Jung et al.	SpoofCeleb: in-the-wild speech	Clean-data models fail in noisy rooms	Noise-robust acoustic features

V. Comparative Analysis

Table II compares the main families of detection approaches discussed in Section IV, and Table III lists the benchmark datasets. The comparison is qualitative and reflects the limitations reported by the original authors rather than a new experiment.

TABLE II
COMPARISON OF DETECTION APPROACHES

Approach	Strengths	Weaknesses
Spatial artifact CNN	Efficient; strong on known generators	Tied to generator-specific artifacts
Frequency / compression-aware	Works on low-quality, compressed media	Degrades under extreme downsampling
Meta-learning	Adapts to unseen forgeries with few samples	Heavy training overhead
Prompt / relation learning	Better on unseen manipulations	Struggles with diffusion outputs
Audio-only	Targets voice cloning directly	Weak in noisy rooms if trained on clean data
Audio-visual	Checks cross-modal consistency	Few joint benchmarks

TABLE III
BENCHMARK DATASETS FOR DEEPFAKE DETECTION

Dataset	Modality	Notes
Deepfake-TIMIT	Video	640 fake videos at low and high quality
FaceForensics++	Video	Four manipulation methods on shared sources
Celeb-DF	Video	Celebrity face swaps with improved visual quality
DFDC	Video	Large-scale challenge dataset
DeeperForensics-1.0	Video	60,000 videos, 17.6 million frames
FFIW-10K	Video	Multiple faces per frame, in the wild
TIMIT-TTS	Audio-visual	Text-to-speech forgeries with video
SpoofCeleb	Audio	Real-world noisy speech

VI. Reference Case: Proposed Detection System

The reference case is a proposed web-based system that detects whether an image, video, or audio clip is real or fake, identifies manipulated facial expressions, voice, and lip movements, and presents the result to non-expert users. Its method follows the stages of problem identification, literature review, data collection, data pre-processing, AI model development, training, testing, performance evaluation, web application development, result analysis, and conclusion with future scope. Fig. 2 shows the processing pipeline, and Table IV maps its modules to the themes of this review.

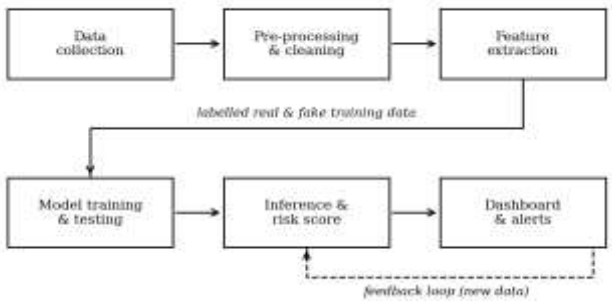


Fig. 2. Processing pipeline of the proposed detection system.

TABLE IV
MAPPING OF REVIEW THEMES TO THE PROPOSED SYSTEM

Theme	System module	Key technique
Visual artifacts	Frame extraction; face analysis	OpenCV, FaceNet; CNN classifiers
Generalization	Model development	TensorFlow / PyTorch; cross-dataset testing
Compression	Pre-processing	Resizing, normalization, multi-scale features
Audio	Audio analysis	Denoising and acoustic features
Text and fraud	NLP module	Hugging Face BERT / RoBERTa
Datasets	Data ingestion and storage	Pandas, NumPy; MySQL

Theme	System module	Key technique
Deployment	API and dashboard	Node.js API; React.js interface
Governance	Risk score and flags	Logging; feedback loop for retraining

The intended outcomes are automated, near real-time scanning of multi-modal data, an easy-to-use dashboard that shows risk scores and flags for quick decisions, models that learn from new data patterns over time, and protection of users from identity theft, phishing links, and fake media before they spread. This paper reports no measured results for the system; its performance must be established by controlled evaluation.

Two observations follow the case. First, a single pipeline can serve several media types, but each modality needs its own pre-processing and features, and the dashboard must combine their outputs into one understandable score. Second, features that look simple at the interface, such as a single risk score, depend on careful engineering beneath them: models must be tested across datasets and compression levels, because a detector that scores well on its training data can still fail on unseen forgeries.

VII. Open Challenges and Research Directions

- **Generalization to unseen generators.** Detectors that remain accurate on diffusion-based and other new forgeries, validated across datasets.
- **Compression resilience.** Multi-scale and frequency-aware features for low-resolution and heavily compressed streaming media.
- **Adversarial robustness.** Feature extraction that resists deliberate perturbation.
- **In-the-wild audio.** Acoustic feature extractors that tolerate reverberation, background noise, and channel distortion.
- **Joint audio-visual benchmarks.** Larger multimodal datasets with cross-modal synchronization checks.
- **Lightweight deployment.** Efficient backbones and optimizers for real-time and edge inference.
- **Trustworthy use.** Explainable scores, privacy-aware handling of uploaded media, and human review of flagged content.
- **Empirical evaluation.** Controlled studies of integrated detection systems, including usability, detection speed, and false-alarm cost.

VIII. Conclusion

Deepfake detection sits at the intersection of computer vision, speech processing, machine learning, and digital security. This review organized the literature into eight themes, compared detection approaches and datasets, and illustrated their integration through a proposed web-based detection system. The analysis shows that visual, frequency-based, meta-learning, audio, and multimodal methods each solve part of the problem, but that generalization to unseen generators, robustness to compression and adversarial attack, and noise-robust audio analysis remain unresolved. Addressing these challenges, together with joint audio-visual benchmarks and lightweight deployment, will be essential for detection systems that the public can trust and use.

References

1. I. Goodfellow *et al.*, “Generative adversarial nets,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014, pp. 2672–2680.

2. T. T. Nguyen *et al.*, “Deep learning for deepfakes creation and detection: A survey,” *Comput. Vis. Image Underst.*, vol. 223, Art. no. 103525, 2022.

3. Y. Mirsky and W. Lee, “The creation and detection of deepfakes: A survey,” *ACM Comput. Surv.*, vol. 54, no. 1, Art. no. 7, 2021.

4. M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, “Deepfake detection: A systematic literature review,” *IEEE Access*, vol. 10, pp. 25494–25513, 2022.

5. B. Kitchenham, “Procedures for performing systematic reviews,” Keele Univ., Keele, U.K., Tech. Rep. TR/SE-0401, 2004.

6. R. He *et al.*, *IEEE Trans. Inf. Forensics Security*, 2023.

7. T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.

8. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

9. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1–11.

10. Y. Li and S. Lyu, “Exposing DeepFake videos by detecting face warping artifacts,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2019.

11. F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1251–1258.

12. S. Chhabra *et al.*, “Low-quality deepfake detection via unseen artifacts,” *IEEE Trans. Artif. Intell.*, vol. 5, no. 4, 2024.

13. V.-N. Tran, S.-H. Lee, H.-S. Le, and K.-R. Kwon, “Generalization of forgery detection with meta deepfake detection model,” *IEEE Access*, vol. 11, pp. 535–546, 2023.

14. H. Jung *et al.*, “SpoofCeleb: Speech deepfake detection and SASV in the wild,” *IEEE Open J. Signal Process.*, 2025.

15. D. Salvi *et al.*, “A robust approach to multimodal deepfake detection,” *J. Imaging*, vol. 9, no. 6, Art. no. 122, 2023.

16. P. Korshunov and S. Marcel, “Deepfakes: A new threat to face recognition? Assessment and detection,” arXiv:1812.08685, 2018.

17. 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 Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 3207–3216.

18. B. Dolhansky *et al.*, “The DeepFake Detection Challenge (DFDC) dataset,” arXiv:2006.07397, 2020.

19. L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, “DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 2889–2898.

20. T. Zhou, W. Wang, Z. Liang, and J. Shen, “Face forensics in the wild,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 5778–5788.

21. B. C. Soundarya and H. L. Gururaj, 2025.

22. M. Al-Fehani, A. Al-Baseer, and S. Al-Kuwari, *IEEE Access*, vol. 14, 2026, doi: 10.1109/ACCESS.2026.3660914.

23. R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Statist. Sci.*, vol. 17, no. 3, pp. 235–255, 2002.

24. National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, Gaithersburg, MD, USA, Jan. 2023.