International Journal of Research in Advanced Engineering and Technology
www.allengineeringjournal.in
ISSN: 2455-0876
Received: 21-07-2026, Accepted: 20-08-2026, Published: 04-09-2026
Volume 12, Issue 3, 2026, Page No. 7-12

Detection of AI-generated deepfake images and videos: Methods, evaluation, and research challenges

Ritu Dahiya
Assistant Professor, Department of Computer Science Chhotu Ram Arya College Sonipat, Haryana, India
DOI: https://doi.org/10.66856/ijraet.2026.12.3.12023

Abstract
The rapid development of generative artificial intelligence has enabled the creation of highly realistic synthetic images and videos. Although these technologies support beneficial applications in entertainment, education, accessibility, and digital content production, they also enable impersonation, misinformation, fraud, and non-consensual synthetic media. Reliable detection of AI-generated deepfake content is therefore an important problem in computer vision, multimedia forensics, and information security. This paper reviews the principal techniques used to detect manipulated images and videos, including spatial artifact analysis, frequency-domain analysis, physiological signals, temporal inconsistency modeling, multimodal verification, and provenance-based methods. It also examines representative datasets, evaluation metrics, generalization problems, adversarial attacks, and emerging detection architectures. The analysis indicates that current detectors can achieve high performance on familiar datasets but often degrade substantially when evaluated on unseen generators, compression levels, identities, and post-processing operations. Future systems should combine visual analysis with temporal reasoning, provenance metadata, robust uncertainty estimation, and continuous adaptation to new generative models.

Keywords: Deepfake detection, synthetic media, computer vision, multimedia forensics, generative artificial intelligence, video authentication

Introduction

Generative models can now synthesize photorealistic faces, alter facial expressions, replace identities, and generate complete video sequences. Deepfake content is commonly produced using generative adversarial networks, autoencoders, diffusion models, neural rendering systems, and face-swapping pipelines. The resulting media may be visually convincing to human observers, particularly after resizing, compression, noise addition, or social-media processing.

The detection problem can be formulated as a binary classification task. Given an image or video x, a detector estimates whether it is authentic or AI-generated:

$$f(x) = \begin{cases} 0, & \text{authentic content,} \\ 1, & \text{synthetic or manipulated content.} \end{cases}$$

However, deepfake detection is not simply a matter of identifying visible defects. Modern generators increasingly remove obvious artifacts, while manipulation pipelines may include several stages of editing. A reliable system must therefore identify subtle inconsistencies in texture, geometry, lighting, frequency statistics, motion, biological signals, or content provenance.

This paper addresses five central questions:

1. Which visual and temporal features are useful for detecting AI-generated media?
2. How do current machine-learning architectures perform on images and videos?
3. Which datasets and metrics support meaningful evaluation?
4. Why do detectors fail when confronted with unseen generators and real-world distortions?
5. What research directions are most promising for robust deployment?

Deepfake Generation and Threat Model

1. Categories of synthetic media

Deepfakes can be divided into several broad categories:

- **Face swapping:** Replacing one person's identity with another person's identity.
- **Face reenactment:** Transferring facial expressions, head motion, or mouth movements.
- **Attribute manipulation:** Changing age, hairstyle, skin appearance, facial structure, or accessories.
- **Entire-face synthesis:** Generating a face that does not correspond to a real person.
- **Lip synchronization:** Modifying mouth movements to match a target speech track.
- **Full-scene generation:** Producing an image or video frame using a generative model rather than modifying a real face.
- **Image-to-video synthesis:** Generating motion from a still image or text prompt.

2. Attacker capabilities

A practical threat model assumes that an attacker may:

- Generate content using a model unavailable to the detector developer.

- Apply resizing, cropping, sharpening, color adjustment, or denoising.
- Re-encode the content using a social-media video codec.
- Remove or alter metadata.
- Add subtitles, logos, borders, or other distractions.
- Produce short clips containing only a few manipulated frames.
- Use adversarial perturbations designed to reduce detector confidence.

A detector that succeeds only under laboratory conditions is insufficient for deployment. The primary objective is therefore cross-domain robustness, meaning reliable performance on media created by unseen systems and processed through realistic distribution channels.

Taxonomy of Detection Techniques

1. Spatial artifact analysis

Early deepfake detectors focused on artifacts in individual frames. These artifacts may include:

- Irregular facial boundaries.
- Inconsistent skin texture.
- Unnatural teeth, eyes, or hair.
- Asymmetric reflections.
- Incorrect shadow geometry.
- Blending errors around the face.
- Inconsistent color transitions between the manipulated region and the surrounding scene.

Convolutional neural networks can learn these patterns directly from pixel data. A typical image detector extracts a face region, passes it through a feature extractor, and produces a probability score:

$$p(y = 1 \mid x) = \sigma(W^{\top}\phi(x) + b)$$

where $\phi(x)$ is the learned feature representation, W is a classifier weight vector, b is a bias term, and σ is the sigmoid function.

Spatial detectors are computationally efficient and useful for frame-level analysis. Their primary weakness is dependence on the visual artifacts of the training data. If a new generator produces different artifacts, performance may decline sharply.

2. Frequency-domain analysis

Synthetic images may contain abnormal frequency patterns caused by upsampling, interpolation, convolutional architectures, or diffusion-model sampling. A two-dimensional Fourier transform represents an image in terms of spatial frequencies:

$$F(u,v) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} I(x,y)\exp\left[-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)\right]$$

The magnitude spectrum $|F(u,v)|$ can reveal periodic structures that are difficult to observe in the spatial domain. High-frequency inconsistencies may occur around facial boundaries, hair, eyes, or generated backgrounds.

Frequency-based methods are attractive because certain generator fingerprints can persist after moderate image processing. Nevertheless, they are vulnerable to strong compression, resizing, denoising, and changes in the generator architecture.

3. Physiological signal analysis

Real videos contain weak biological signals associated with blood-volume changes, eye motion, pulse-related skin variation, and natural breathing. A manipulated face may fail to preserve these signals consistently across frames.

For a facial region of interest, a temporal color signal can be approximated as

$$s_c(t) = \frac{1}{N}\sum_{i=1}^{N} I_{i,c}(t)$$

where c identifies a color channel, N is the number of pixels in the region, and $I_{i,c}(t)$ is the intensity of pixel i at time t.

The detector can analyze periodicity, spatial consistency, and correlation between multiple facial regions. Physiological methods are valuable because they examine temporal properties that are not limited to blending artifacts. However, they require sufficient video quality and duration. Low frame rates, poor illumination, motion blur, and aggressive compression can make the signal unreliable.

4. Temporal inconsistency analysis

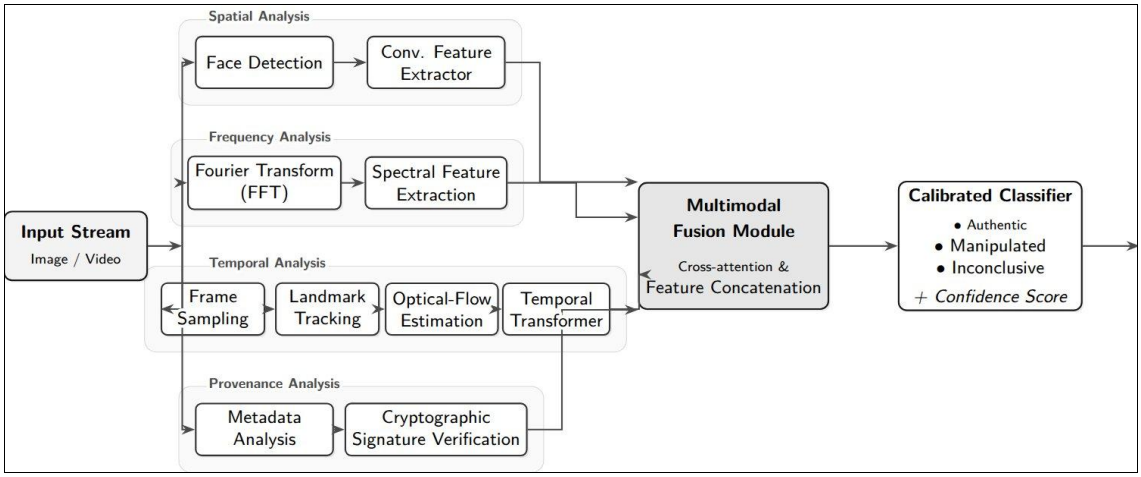
A video detector should examine whether facial appearance and motion evolve naturally over time. Important cues include:

- Abrupt changes in identity features.
- Inconsistent head pose.
- Flickering textures.
- Irregular eye motion.
- Unstable facial landmarks.
- Mismatch between lip motion and speech.
- Inconsistent illumination across frames.
- Unnatural object motion in generated backgrounds.

Let z_t denote the visual feature extracted from frame t. A temporal model may estimate the sequence probability using

$$P(z_1, z_2, \ldots, z_T) = \prod_{t=1}^{T} P(z_t \mid z_1, \ldots, z_{t-1})$$

Recurrent neural networks, temporal convolutional networks, transformers, and optical-flow models can learn these dependencies. Temporal analysis generally improves video-level detection, although it introduces additional computational cost and requires appropriately labeled sequence



5. Audio-visual consistency

Video authenticity can also be assessed by comparing speech audio with mouth motion. A detector may measure whether phonetic units align with visible lip movements. Let a_t represent an audio feature and v_t represent a visual feature. A cross-modal similarity score can be written as

$$S_{\text{av}} = \frac{1}{T}\sum_{t=1}^{T} \text{sim}\,(a_t, v_t)$$

where sim is a similarity function and T is the number of synchronized time steps?

Audio-visual methods are particularly useful for detecting lip-sync manipulation, but they are not sufficient for silent videos or content with dubbed audio. Natural variations in speech, accents, camera delay, and editing may also produce false positives.

6. Provenance and watermark verification

Content analysis attempts to determine whether media is synthetic by examining the media itself. Provenance methods instead verify how the content was created and modified. These methods may use:

- Cryptographic signatures.
- Secure capture devices.
- Robust watermarks.
- Content credentials.
- Generation-time metadata.
- Chain-of-custody records.

Provenance is complementary to forensic detection. It can provide strong evidence when trustworthy metadata exists, but metadata may be stripped, forged, or unavailable. Consequently, provenance should be combined with content-based analysis rather than treated as a complete replacement.

Detection Architecture

A practical detector can be organized into five stages.

1. Preprocessing

The system first decodes the media, samples video frames, detects faces, aligns facial regions, and records technical properties such as resolution, frame rate, and codec.

2. Feature extraction

Different feature extractors identify different classes of evidence:

- A spatial encoder analyzes pixel-level and semantic visual information.
- A spectral encoder analyzes frequency-domain statistics.
- A landmark and optical-flow module analyzes motion.
- An audio encoder analyzes speech characteristics.
- A provenance module verifies metadata and signatures.

3. Feature fusion

The feature vectors may be concatenated:

$$h = \left[h_{\text{spatial}}; h_{\text{frequency}}; h_{\text{temporal}}; h_{\text{audio}}; h_{\text{provenance}}\right]$$

A learned fusion layer then produces the final representation. Early fusion combines features before classification, whereas late fusion combines independent detector scores. Late fusion can be more robust when one modality is missing.

4. Confidence calibration

A detector should not produce only a binary decision. It should estimate uncertainty and identify cases outside its training distribution. Calibration seeks agreement between predicted confidence and observed accuracy. The expected calibration error can be estimated by partitioning predictions into confidence bins:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|$$

where B_m is confidence bin m, n is the total number of predictions, and M is the number of bins.

5. Decision policy

For high-risk applications, uncertain cases should be referred for human or expert review. A three-class decision policy is preferable to forced binary classification:

$$\hat{y} = \begin{cases} \text{authentic}, & p < \tau_0, \\ \text{inconclusive}, & \tau_0 \le p \le \tau_1, \\ \text{manipulated}, & p > \tau_1. \end{cases}$$

The thresholds τ_0 and τ_1 should be selected according to the operational cost of false positives and false negatives.

Datasets and Benchmarking

Representative datasets include FaceForensics++, Celeb-DF, DeepFakeDetection, DFDC, DeeperForensics-1.0, ForgeryNet, and recent benchmark collections such as DeepFakeBench. These datasets vary in identity distribution, manipulation method, resolution, compression, and demographic coverage.

A valid evaluation protocol should separate identities and source videos across training and testing partitions. Randomly splitting frames from the same video can produce inflated performance because nearly identical frames may appear in both partitions.

1. Evaluation metrics

The most common metrics are:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

and

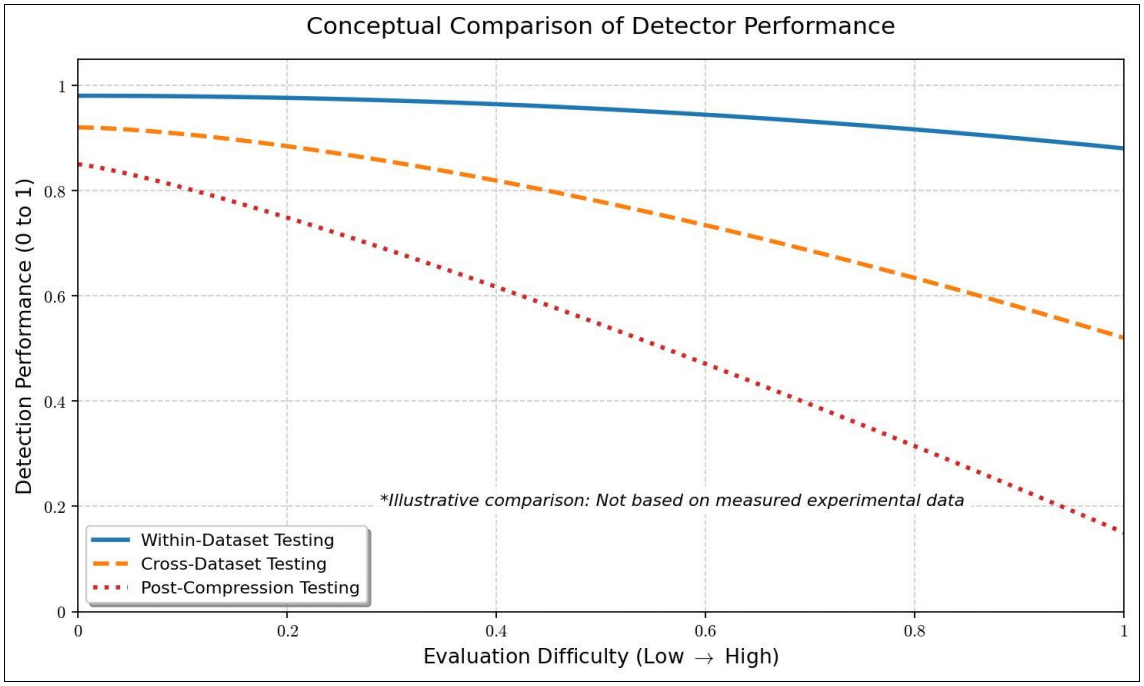
$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

For imbalanced datasets, accuracy can be misleading. Area under the receiver operating characteristic curve, area under the precision-recall curve, balanced accuracy, and equal error rate should also be reported.

A particularly important measure is cross-dataset generalization. A detector may obtain a high within-dataset score while performing poorly on an unseen dataset.

Evaluation should therefore include:

- Unseen manipulation methods.
- Unseen identities.
- Unseen compression levels.
- Unseen resolutions.
- Unseen demographic groups.
- Real-world social-media transformations.
- Generators released after model training.



2. Experimental protocol

A reproducible experiment should specify:

- Dataset versions and licensing conditions.
- Training, validation, and test partitions.
- Face detection and alignment procedures.
- Frame sampling rate.
- Image resolution.
- Data augmentation operations.
- Hardware and software versions.
- Optimization algorithm and stopping criterion.
- Statistical confidence intervals.
- Threshold-selection procedure.

Results should be reported separately for images, short videos, long videos, high-quality content, and heavily compressed content.

Major Challenges

1. Generalization to unseen generators

The most important limitation is that detectors often learn generator-specific fingerprints. A model trained on one family of face-swapping systems may not recognize images produced by diffusion models or neural rendering systems.

This is a form of distribution shift

$$P_{\text{train}}(x, y) \neq P_{\text{test}}(x, y)$$

Robust systems should learn manipulation-invariant evidence rather than memorizing dataset-specific artifacts.

2. Post-processing and compression
Social platforms commonly resize, crop, re-encode, sharpen, and recompress media. These operations can remove high-frequency artifacts while also introducing new artifacts that confuse detectors. Robust training should include realistic transformations, but excessive augmentation may remove useful forensic evidence.

3. Adversarial adaptation
Once the attacker knows the detection strategy, the generation process can be optimized to evade it. This creates an adversarial relationship between synthesis and detection. A detector should therefore be evaluated against adaptive attacks rather than only static benchmark samples.

4. Bias and demographic fairness
Performance may vary according to skin tone, age, gender presentation, facial hair, lighting, camera type, and image quality. If a dataset underrepresents particular demographic groups, the detector may produce unequal false-positive or false-negative rates.

Fairness analysis should report group-conditional metrics such as

$$\Delta_{\mathrm{FPR}} = \max_{g} \mathrm{FPR}_g - \min_{g} \mathrm{FPR}_g$$

where g indexes demographic or contextual groups.

5. Explainability
A binary output is insufficient for journalism, law enforcement, platform moderation, and academic verification. The detector should identify evidence such as:

- Regions with abnormal texture.
- Inconsistent facial boundaries.
- Unstable landmarks.
- Spectral anomalies.
- Audio-visual synchronization errors.
- Missing or invalid provenance information.

Explanations should be treated as supporting evidence rather than absolute proof, because heat maps and saliency methods can be unstable.

6. Open-world detection
Real-world systems encounter media types that were not present during training. Open-world detectors should be able to output "unknown" or "inconclusive" rather than assigning excessive confidence to an unfamiliar sample. This requires out-of-distribution detection, uncertainty estimation, and continual evaluation.

Recommended Hybrid Framework
A robust future system should combine independent evidence sources. The proposed framework consists of:

1. Frame-level visual analysis using a convolutional or transformer-based encoder.
2. Frequency analysis using spectral statistics and learned frequency features.
3. Landmark and motion analysis for temporal coherence.
4. Physiological signal analysis for pulse-related and blink-related consistency.
5. Audio-visual alignment for speech-containing videos.
6. Provenance verification using signed metadata and watermarks.
7. Uncertainty estimation with an inconclusive output.
8. Human review for high-impact decisions.

The final score can be represented as

$$S = \alpha S_{\text{spatial}} + \beta S_{\text{frequency}} + \gamma S_{\text{temporal}} + \delta S_{\text{audio}} + \epsilon S_{\text{provenance}}$$

where the weights are learned or selected according to validation performance and the availability of each modality.
This architecture should be trained with a combination of supervised and self-supervised objectives. Supervised learning provides manipulation labels, while self-supervised learning encourages representations that remain stable under benign transformations and sensitive to meaningful inconsistencies.

Ethical, Legal, and Operational Considerations
Deepfake detection systems can protect individuals and institutions, but incorrect classifications can also cause harm. A false accusation may damage a person's reputation, while a missed deepfake may enable fraud or misinformation. Therefore:

- Detector outputs should not be treated as definitive proof without corroborating evidence.
- High-impact decisions should include human review.
- Training data should respect privacy, consent, and licensing requirements.
- Evaluation should measure demographic and contextual disparities.
- Researchers should disclose limitations and known failure cases.
- Detectors should be updated as generative models evolve.
- Evidence and decision logs should be preserved for later auditing.

The distinction between "AI-generated" and "false" is also important. Authentic footage may contain fictional or staged events, while synthetic content may accurately represent a fictional scene. Detection systems should report the nature of the manipulation rather than infer intent or truthfulness automatically.

Conclusion
The detection of AI-generated deepfake images and videos is a continuously evolving problem. Spatial artifacts, spectral inconsistencies, temporal instability, physiological signals, audio-visual mismatch, and provenance metadata each provide useful evidence, but no individual method is reliable under all conditions. Current detectors commonly perform well on familiar benchmark distributions and degrade when facing unseen generators, compression, adversarial manipulation, or demographic variation.
The most promising direction is a multimodal and uncertainty-aware framework that combines content-based forensic analysis with provenance verification. Future research should emphasize cross-generator generalization, open-world evaluation, fair performance across demographic groups, robust calibration, adversarial testing, and interpretable evidence. Deepfake detection should ultimately be viewed not as a single classifier but as a

broader authentication system integrating algorithms, metadata, platform infrastructure, and expert judgment.

References

1. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. FaceForensics++: Learning to Detect Manipulated Facial Images. Proceedings of the IEEE International Conference on Computer Vision, 2019.
2. Li Y, Yang X, Sun P, Qi H, Lyu S. Celeb-DF: A New Dataset for DeepFake Forensics. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2020.
3. Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer CC. The DeepFake Detection Challenge Dataset. arXiv preprint arXiv, 2006, 07397, 2020.
4. Jiang L, Li R, Wu W, Qian C, Loy CC. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.
5. He Y, Li N, Ding X, *et al*. ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021.
6. Verdoliva L. Media Forensics and DeepFakes: An Overview. IEEE Journal of Selected Topics in Signal Processing,2020:14(5):910–932.
7. Li Y, Chang MC, Lyu S. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. IEEE International Workshop on Information Forensics and Security, 2018.
8. Guera D, Delp EJ. Deepfake Video Detection Using Recurrent Neural Networks. IEEE International Conference on Advanced Video and Signal Based Surveillance, 2018.
9. Li Y, Sun P, Qi H, Lyu S. Exposing DeepFake Videos by Detecting Face Warping Artifacts. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.
10. Durall R, Keuper M, Keuper J. Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.
11. Wang S-Y, Wang O, Zhang R, Owens A, Efros AA. CNN-Generated Images Are Surprisingly Easy to Spot… for Now. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.
12. Wang S-Y, Wang O, Zhang R, Owens A, Efros AA. CNN-Generated Images Are Surprisingly Easy to Spot… for Now. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.
13. Yan Z, Zhang Y, Yuan X, Lyu S, Wu B. DeepFakeBench: A Comprehensive Benchmark of Deepfake Detection. Advances in Neural Information Processing Systems, 2023.
14. Frank J, Eisenhofer T, Schönherr L, Fischer A, Kolossiatis M, Holz T, *et al*. Leveraging Frequency Analysis for Deep Fake Image Recognition. Proceedings of the International Conference on Machine Learning, 2020.
15. C2PA. C2PA Technical Specification: Content Credentials. Coalition for Content Provenance and Authenticity, current specification.
16. National Institute of Standards and Technology. Face Recognition Technology Evaluation: Ongoing Face Recognition Vendor Test. NIST, relevant evaluation reports.