International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 10, October 2026)

# AI Deepfake Detection in Digital Forensics

Shashikiran V

*Department of Criminology & Forensic Science, Acharya Institute of Graduate Studies, Bengaluru, India*

***Abstract*-- Artificial intelligence (AI) enables the generation and manipulation of highly realistic images, videos, and audio recordings, commonly referred to as deepfakes. Such media may be misused for impersonation, fraud, misinformation, harassment, and the fabrication of apparent evidence. These developments create challenges for digital forensic investigators, who must assess authenticity while preserving evidence integrity and explaining the limitations of analytical methods. This paper presents a technical narrative review and a forensic-oriented evaluation framework for AI-based deepfake detection. It discusses major detection approaches, benchmark datasets, cross-dataset generalisation, robustness testing, and statistical performance measures. Published studies report that detector performance can decline substantially under distribution shifts; one cross-dataset comparison reported Xception AUC values of 96.30% on FaceForensics++ HQ, 36.19% on Celeb-DF, and 48.98% on DFDC. These are study-specific published results and are not experiments conducted for this paper. The manuscript proposes an evaluation protocol involving within-dataset, cross-dataset, and transformation-based testing, together with confidence intervals and paired statistical comparisons. It also presents a digital forensic workflow integrating evidence acquisition, integrity preservation, automated screening, independent analysis, corroboration, and reporting. The paper argues that detector outputs should be treated as technical indicators rather than conclusive proof of authenticity or manipulation. The framework is intended to support transparent and reproducible forensic evaluation.**

***Keywords*-- artificial intelligence; deepfake detection; digital forensics; multimedia forensics; cross-dataset generalisation; digital evidence; robustness; statistical evaluation.**

## I. Introduction

Digital images, videos, and audio recordings are increasingly used in criminal investigations, legal proceedings, public communications, and online platforms. Generative artificial intelligence has made it possible to create synthetic media or alter existing recordings to imitate a person's appearance, voice, or actions. These technologies have legitimate applications in entertainment, accessibility, education, and creative production. However, they can also be misused to create fabricated statements, impersonate individuals, facilitate financial fraud, and distribute misleading content.

Deepfake detection has consequently become an important area within digital and multimedia forensics. Detection systems attempt to identify traces of manipulation through spatial artefacts, frequency-domain patterns, temporal inconsistencies, biological-signal anomalies, audio-visual mismatches, or learned representations. Their performance, however, may depend heavily on the datasets and conditions used for training and testing.

A central forensic concern is whether a model's performance on a benchmark dataset can be generalised to media encountered in an investigation. Compression, resizing, platform transcoding, lighting, recording devices, unseen generation methods, and dataset composition can affect detector reliability. Moreover, false-positive and false-negative classifications may have significant consequences when media is considered evidentiary material.

This paper reviews deepfake detection from a digital forensic perspective and proposes a quantitative evaluation framework that emphasises cross-dataset testing, robustness, statistical uncertainty, reproducibility, and evidentiary interpretation.

## II. Problem Statement

Deepfake detectors may achieve strong performance on benchmark datasets but perform less reliably on media from unseen sources or manipulation methods. Evaluations that rely only on within-dataset accuracy can therefore overstate practical reliability. A forensic evaluation should test performance under independent conditions, report uncertainty, and distinguish an automated classification from a complete expert conclusion.

## III. Aim, Objectives, And Research Questions

### *3.1 Aim*

To examine AI-based deepfake detection methods and propose a statistically grounded evaluation framework for digital forensic use.

### *3.2 Objectives*

- Review major technical approaches used in deepfake detection.
- Explain evaluation metrics relevant to forensic screening.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 10, October 2026)

- Examine cross-dataset generalisation and robustness testing.
- Propose a reproducible quantitative evaluation protocol.
- Develop a forensic workflow for integrating automated detection with conventional digital examination.
- Discuss evidentiary, ethical, and operational limitations.

### *3.3 Research questions*

1. Which statistical measures provide a meaningful assessment of deepfake detector performance?
2. How does performance change when detectors are evaluated on independent datasets?
3. Which testing conditions are important for forensic reliability?
4. How should automated detection results be incorporated into digital forensic reporting?

## IV. Review Methodology

This manuscript is a technical narrative review and proposed evaluation framework. It is not a completed systematic review and does not report a new experiment. To convert it into a systematic review, the author should conduct and document a reproducible search of scholarly databases, record the exact search strings and date, define inclusion and exclusion criteria, remove duplicates, screen records, and report the number of studies included.

### *4.1 Suggested databases and search terms*

Suggested databases include Scopus, Web of Science, IEEE Xplore, ACM Digital Library, and Google Scholar, subject to access. Suggested search concepts include "deepfake detection," "digital media forensics," "cross-dataset generalisation," "video manipulation detection," "audio deepfake detection," "robustness evaluation," and "deepfake statistical analysis."

### *4.2 Proposed inclusion criteria*

- Studies evaluate detection of manipulated or synthetic image, video, or audio content.
- Studies report quantitative performance metrics or a clearly described evaluation protocol.
- Dataset composition and test conditions are described sufficiently for interpretation.
- Peer-reviewed papers or established benchmark publications are used.
- The original publication is accessible for bibliographic and methodological verification.

Studies focused exclusively on media generation without evaluating detection should be excluded from quantitative detector comparisons. Search counts and screening decisions must be based on the actual search and must not be estimated.

## V. Technical Background

### *5.1 Deepfake generation and manipulation*

Deepfakes may be produced through face swapping, facial reenactment, generative adversarial networks, diffusion models, voice cloning, or combinations of these methods. Manipulation may affect an entire file or selected frames, regions, or audio segments. Post-processing and recompression can weaken or remove traces that detection systems rely on.

### *5.2 Detection approaches*

Spatial-feature methods examine image-level patterns, texture inconsistencies, and blending artefacts. Frequency-domain methods analyse spectral characteristics. Temporal methods examine inconsistencies across frames. Biological-signal methods investigate physiological cues, while audio-based methods examine spectral, prosodic, and phase-related properties of speech. Multimodal methods compare information across audio and visual streams, such as lip movement and speech timing.

Convolutional neural networks (CNNs), transformer architectures, and hybrid systems are commonly used to learn discriminative features. Their performance depends on training data, model architecture, preprocessing, decision thresholds, and evaluation conditions.

### *5.3 Generalisation and robustness*

Generalisation is the ability of a detector to perform on data drawn from a distribution different from its training data. Robustness refers to maintaining performance under transformations or attacks. Common distribution shifts include new generation methods, different identities, compression, resizing, platform transcoding, lighting, camera differences, cropping, blur, and noise. Strong within-dataset results alone do not establish reliability across these conditions.

## VI. Datasets And Benchmarking

Benchmark datasets support comparative research, but their differences must be considered when interpreting results. Dataset versions, sample counts, class balance, manipulation methods, split procedures, and preprocessing should be reported. Researchers should also check for identity overlap or near-duplicate leakage across training and test partitions.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 10, October 2026)

| Dataset | Content / purpose | Typical role | Important consideration |
|---|---|---|---|
| FaceForensics++ | Manipulated facial videos | Training and within-dataset testing | Multiple manipulation methods and compression settings |
| Celeb-DF | High-quality manipulated videos | Cross-dataset evaluation | Different visual quality and generation characteristics |
| DFDC | Large-scale deepfake video benchmark | Training and testing | Document exact split and preprocessing |
| DFFD | Diverse facial manipulation data | Cross-dataset testing | Multiple manipulation types |
| UADFV | Early deepfake video dataset | Benchmark comparison | Relatively limited scale |
| WildDeepfake | In-the-wild manipulated videos | Real-world generalisation | Collection and quality conditions differ from controlled data |
| ASVspoof | Synthetic and spoofed speech | Audio detection evaluation | Audio-specific metrics and protocols |
| WaveFake | Synthetic speech | Audio cross-dataset testing | Generator and recording conditions affect results |

## VII. Quantitative Evaluation Metrics

For binary classification, true positives (TP) are manipulated samples correctly classified as manipulated; true negatives (TN) are authentic samples correctly classified as authentic; false positives (FP) are authentic samples incorrectly flagged as manipulated; and false negatives (FN) are manipulated samples incorrectly classified as authentic.

| Metric | Formula | Forensic interpretation |
|---|---|---|
| Accuracy | (TP + TN) / (TP + TN + FP + FN) | Overall proportion of correct classifications |
| Precision | TP / (TP + FP) | Proportion of flagged items that are manipulated |
| Recall / Sensitivity | TP / (TP + FN) | Proportion of manipulated items detected |
| Specificity | TN / (TN + FP) | Proportion of authentic items correctly identified |
| False Positive Rate | FP / (FP + TN) | Rate at which authentic media is incorrectly flagged |
| F1-score | 2 × (Precision × Recall) / (Precision + Recall) | Balance between precision and recall |
| ROC-AUC | Area under ROC curve | Discrimination across thresholds |
| Equal Error Rate | Point where FPR equals FNR | Often used in biometric and audio detection |

Accuracy may be misleading when class distributions are imbalanced. For forensic screening, specificity and false-positive rate are particularly important because an incorrect manipulation claim can have serious consequences.

ROC-AUC should be supplemented by threshold-specific metrics and, where appropriate, precision-recall analysis.

## VIII. Published Quantitative Evidence

Published evaluations demonstrate that detector performance can change substantially across datasets. One cross-dataset comparison reported Xception AUC values of 96.30% on FaceForensics++ HQ, 36.19% on Celeb-DF, and 48.98% on DFDC.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 10, October 2026)

These values illustrate a specific model and protocol; they should not be generalised to all detectors or treated as results from the present manuscript.

| Training dataset | Test dataset | Model | Reported AUC |
|---|---|---|---|
| FaceForensics++ HQ | FaceForensics++ | Xception | 96.30% |
| FaceForensics++ HQ | Celeb-DF | Xception | 36.19% |
| FaceForensics++ HQ | DFDC | Xception | 48.98% |
| FaceForensics++ HQ | DFD | Xception | 87.86% |
| FaceForensics++ HQ | Celeb-DF | Study's proposed method | 82.16% |
| FaceForensics++ HQ | DFDC | Study's proposed method | 80.35% |
| FaceForensics++ HQ | DFD | Study's proposed method | 94.33% |

The reported Xception AUC difference between FaceForensics++ HQ and Celeb-DF is 60.11 percentage points (96.30 − 36.19). The difference between FaceForensics++ HQ and DFDC is 47.32 percentage points (96.30 − 48.98). These are arithmetic comparisons of published values, not experiments conducted for this paper.

A 2026 cross-modality survey reported evaluation across image, video, and audio, including 10 datasets, 13 unimodal detectors, 33 generalisation methods, and six methods tested for robustness.

It reported approximately 10–15% performance degradation in out-of-distribution settings and white-box attack success rates above 80% in the evaluated conditions. These findings are study-specific and do not represent universal detector performance.

## IX. Cross-Dataset Performance Analysis

The absolute performance gap can be calculated as:

$$\Delta abs = Pwithin - Pcross$$

The relative degradation can be calculated as:

$$\Delta rel = [(Pwithin - Pcross) / Pwithin] \times 100$$

| Measure | Illustrative value |
|---|---|
| Within-dataset accuracy | 96% |
| Cross-dataset accuracy | 81% |
| Absolute gap | 15 percentage points |
| Relative degradation | 15.63% |

The values in Table 5 are hypothetical and are included only to demonstrate the calculation. They are not empirical results from this paper and must not be presented as findings.

## X. Proposed Statistical Evaluation Framework

An original experiment should evaluate detectors under three conditions:

(1) within-dataset testing using a held-out partition; (2) cross-dataset testing on independent data without retraining; and (3) robustness testing under controlled transformations such as recompression, resizing, blur, cropping, and noise.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 10, October 2026)

| Comparison | Recommended analysis | Purpose |
|---|---|---|
| Same model across two conditions | Paired bootstrap confidence intervals; McNemar's test for paired binary predictions | Quantify change and compare paired classification outcomes |
| Metric estimates | 95% bootstrap confidence intervals | Estimate uncertainty |
| Multiple models on same test samples | Paired bootstrap or suitable repeated-measures method | Account for shared test cases |
| Multiple datasets | Per-dataset reporting and macro-average | Prevent largest dataset dominating summary |
| Transformation levels | Repeated-measures or mixed-effects analysis where appropriate | Test performance changes across conditions |
| Threshold-dependent performance | ROC and precision-recall curves | Show trade-offs across thresholds |

*10.1 Proposed hypotheses*

- H1: Detector performance is lower on independent datasets than on within-dataset test data.
- H2: Common media transformations significantly affect detector performance.
- H3: Models with higher within-dataset performance do not necessarily achieve the highest cross-dataset performance.
- H4: Multiple metrics provide a more informative assessment than accuracy alone.

*10.2 Reporting requirements*

- Confusion matrix and sample counts for each dataset and condition.
- Accuracy, precision, recall, specificity, F1-score, ROC-AUC, and false-positive rate.
- Threshold-selection procedure and whether it was fixed before testing.
- 95% confidence intervals where feasible.
- Dataset and model versions, preprocessing, software versions, and seeds.
- Separate frame-level and video-level results; do not mix these units.

## XI. Proposed Experimental Design

| Component | Proposed specification |
|---|---|
| Training dataset | FaceForensics++ |
| Independent test datasets | Celeb-DF, DFDC, DFD |
| Baseline models | Xception, EfficientNet, transformer-based detector |
| Primary metric | Video-level ROC-AUC |
| Secondary metrics | Accuracy, precision, recall, specificity, F1-score, FPR |
| Robustness tests | Recompression, resizing, blur, noise, cropping |
| Repeated runs | At least 3–5 seeds where training is stochastic |
| Uncertainty | 95% bootstrap confidence intervals |
| Model comparison | Paired bootstrap or suitable paired statistical test |
| Reproducibility | Record dataset versions, code, seeds, preprocessing, thresholds |

This is a proposed design, not a completed experiment. Results must be generated by implementing the protocol and should not be inferred from published studies unless clearly presented as a separate literature comparison.

## XII. Results Reporting Template

Table 8 is intentionally blank. Populate it only after conducting the experiment or accurately extracting comparable results from published studies.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 10, October 2026)

| Model | Test condition | Accuracy | Precision | Recall | Specificity | F1 | AUC |
|---|---|---|---|---|---|---|---|
| Model A | Within-dataset | — | — | — | — | — | — |
| Model A | Cross-dataset | — | — | — | — | — | — |
| Model B | Within-dataset | — | — | — | — | — | — |
| Model B | Cross-dataset | — | — | — | — | — | — |
| Model C | Within-dataset | — | — | — | — | — | — |
| Model C | Cross-dataset | — | — | — | — | — | — |

## XIII. Proposed Digital Forensic Workflow

Figure 1. Deepfake examination workflow:

**Evidence acquisition → Integrity preservation → Metadata and provenance examination → Automated detection screening → Independent forensic analysis → Corroboration → Reporting and disclosure**

### *13.1 Acquisition and preservation*

Obtain media through lawful and documented procedures. Preserve the original where possible, create a forensic working copy, and calculate a cryptographic hash such as SHA-256. Record acquisition method, date, source, and transfer history.

### *13.2 Initial examination*

Record file format, size, timestamps, metadata, encoding information, source information, and known conversion history. Metadata may be incomplete or altered and should not be treated as conclusive by itself.

### *13.3 Automated screening*

Use detection tools as screening aids. Document software version, model, preprocessing, threshold, and output. Where feasible, compare independent methods and retain logs.

### *13.4 Independent analysis and corroboration*

Examine relevant frames, temporal inconsistencies, audio-visual synchronisation, compression artefacts, and provenance information. Consider alternative explanations such as ordinary editing or platform transcoding. Compare technical findings with independent evidence.

### *13.5 Reporting*

A forensic report should distinguish observed artefacts, detector classifications, expert interpretation, limitations, uncertainty, and conclusions supported by the available evidence.

## XIV. Evidentiary And Indian Legal Considerations

Electronic and digital records in India raise questions concerning authenticity, integrity, admissibility, and the procedures used to produce or certify them. The Bharatiya Sakshya Adhiniyam, 2023 is relevant to the treatment of electronic and digital records. Investigators and researchers should consult the current statutory text and qualified legal advice for case-specific questions.

An automated detector's output alone should not be treated as conclusive proof that a recording is fake or authentic. The evidentiary value of an analysis depends on the complete examination, documentation, method limitations, and applicable legal requirements. This section is general academic discussion and not legal advice.

## XV. Ethical Considerations And Limitations

Deepfake research and casework should consider privacy, consent, dataset licensing, demographic differences in performance, and the risk of false accusations. Sensitive media should be handled using appropriate access controls and retention practices.

This manuscript is a narrative review and proposed framework, not a completed systematic review or original experiment. Some quantitative comparisons are taken from published studies and are protocol-specific.

The hypothetical example is not an empirical finding. The literature search and bibliographic details require final verification. Detector performance can change as generation methods evolve, and no single model or metric can be assumed to generalise to every source or manipulation type.

## XVI. Recommendations

1. Use independent cross-dataset testing rather than relying only on within-dataset accuracy.
2. Report multiple metrics, particularly recall, specificity, and false-positive rate.
3. Provide confidence intervals and sample counts.
4. Document dataset versions, model versions, preprocessing, and thresholds.
5. Test robustness under realistic media transformations.
6. Combine automated screening with metadata, provenance, file-integrity checks, and expert examination.
7. Avoid claiming that a tool definitively proves manipulation unless the complete analysis supports that conclusion.
8. Complete a reproducible literature search and verify references before journal submission.

## XVII. Conclusion

AI-generated deepfakes create significant challenges for digital forensic examination because realistic manipulation can be difficult to identify and automated detectors may not generalise across datasets. Published evaluations report meaningful performance degradation under out-of-distribution conditions and vulnerabilities to adversarial attacks, reinforcing the need for testing beyond conventional benchmark accuracy.

A reliable evaluation framework should include within-dataset testing, independent cross-dataset evaluation, robustness analysis, appropriate statistical measures, and transparent documentation. In practical investigations, automated detection should be treated as one component of a broader forensic process that preserves digital evidence, considers alternative explanations, and communicates uncertainty.

Future research should implement the proposed framework using verified datasets and documented models, report reproducible results, and evaluate performance under realistic forensic conditions.

## REFERENCES

[1] Bharatiya Sakshya Adhiniyam, 2023. Government of India. India Code. https://www.indiacode.nic.in/

[2] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). https://doi.org/10.1109/ICCV.2019.00009

[3] Nguyen-Le, H.-H., Tran, V.-T., Nguyen, D.-T., et al. (2026). Deepfake detection across image, video, and audio: A comprehensive survey with empirical evaluation of generalization and robustness. Artificial Intelligence Review, 59, Article 203. https://doi.org/10.1007/s10462-026-11608-4.

[4] A Survey on Speech Deepfake Detection. (2025). ACM Computing Surveys. https://doi.org/10.1145/3714458.

[5] Refining Localized Attention Features with Multi-Scale Relationships for Enhanced Deepfake Detection in Spatial-Frequency Domain. (2024). Electronics, 13(9), 1749. https://www.mdpi.com/2079-9292/13/9/1749.

[6] Guan, H., Horan, J., & Zhang, A. (2025). Guardians of Forensic Evidence: Evaluating Analytic Systems Against AI-Generated Deepfakes. Forensics@NIST 2024. https://www.nist.gov/publications/guardians-forensic-evidence-evaluating-analytic-systems-against-ai-generated-deepfakes.