Indian Journal of Modern Research and Reviews

This Journal is a member of the *'Committee on Publication Ethics'*

Online ISSN:2584-184X

Research Article

A Systematic Survey of Deepfake Harassment Using Digital Forensics

Dr. S Aarthi [1*], Vikram S [2], Seysanth V [3], Ashraf Kalimullah MA [4]

[1-4] Department: CSE Meenakshi Sundararajan Engineering College Chennai, India

Corresponding Author: *Dr. S Aarthi

DOI: https://doi.org/10.5281/zenodo.18908562

Abstract

The rapid advancement of synthetic media technologies has made deepfake harassment a serious cybersecurity and societal concern. Deepfake-based abuse, including non-consensual explicit content and identity impersonation, causes severe psychological, social, and reputational harm to victims. Most existing solutions rely on artificial intelligence and machine learning techniques for detecting manipulated media; however, these approaches are often computationally expensive, dataset-dependent, and unsuitable for forensic or legal use. This project proposes a cybersecurity- centric framework that treats deepfake harassment as a digital forensics problem rather than a prediction task. The system focuses on media integrity verification, metadata forensic analysis, cryptographic hashing, provenance tracking, and secure chain-of-custody mechanisms. A structured reporting workflow enables the generation of forensic integrity reports that support investigation and evidence handling. The proposed framework is implemented as a working prototype and evaluated using controlled test cases to demonstrate its ability to flag suspicious media and preserve reliable digital evidence, thereby strengthening trust in digital media ecosystems.

Manuscript Information

- **ISSN No:** 2584-184X
- **Received:** 26-01-2026
- **Accepted:** 23-02-2026
- **Published:** 08-03-2026
- **MRR:**4(3); 2026: 56-62

- **Plagiarism Checked:** Yes
- **Peer Review Process:** Yes

How to Cite this Article

Aarthi S, S V, V S, MA A K. A Systematic Survey in Deepfake Harassment Using Digital Forensics. Indian J Mod Res Rev. 2026;4(3):56-62.

Access this Article Online

www.multiarticlesjournal.com

KEYWORDS: Deepfake Harassment, Cybersecurity, Digital Forensics, Media Integrity, Provenance Tracking, Cryptographic Verification

1. INTRODUCTION

The emergence of deepfake technologies has introduced a new dimension of cyber abuse by enabling the creation of highly realistic, manipulated images and videos.

Deepfake harassment has increasingly been used for non-consensual explicit content creation, impersonation, blackmail, and reputational attacks, posing serious psychological and social risks to individuals. The anonymity and rapid dissemination enabled by online platforms further amplify the impact of such abuse. Current research and industrial efforts primarily focus on artificial intelligence–based deepfake detection methods that attempt to classify content as real or fake. While these methods show promising results in controlled environments, they often lack transparency, robustness, and legal reliability. Furthermore, AI-based approaches rarely address critical aspects such as evidence preservation, chain of custody, or investigation support. This project adopts a cybersecurity and digital forensics perspective, emphasising deterministic verification of media integrity, traceability, and accountability to address deepfake harassment in a legally

defensible and sustainable manner.

2. LITERATURE SURVEY

S. No	Paper Title	Author(s)	Year	Description	Advantages	Disadvantages
1	Reporting Non- Consensual Intimate Media: An Audit Study of Deepfakes	Li Qiwei, *et al.*	2024 [2]	Audits platform takedown speeds for AI-generated harassment, comparing copyright vs. nudity reporting.	Exposes flaws in platform policies; proves copyright claims work far better than harassment reports.	Focuses on policy rather than technical forensics; offers no deterministic software tool to verify media integrity.
2	Online Sexual Harassment in Adolescence: A Scoping Review	Franceschi, A., *et al.*	2023 [3]	Reviews 65 studies on the behavioural characteristics of online harassment and visual abuse.	Provides strong sociological evidence for the psychological impact of digital harassment.	Purely behavioural; lacks a cybersecurity architecture or digital forensics framework for technical validation.
3	Audio-Visual Deepfake Detection With Local Temporal Inconsistencies	Astrid, M., *et al.*	2025 [6]	Deep learning architecture computing temporal distance maps between audio and visual features.	Effective at catching cross-modal desynchronization in manipulated videos.	Relies on black-box AI yielding probabilistic outputs; cannot be translated into human-readable, legal-grade evidence.
4	An Investigation into the Utilisation of CNN with LSTM for Video Deepfake Detection	Tipper, S., *et al.*	2024 [4]	Integrates CNNs (spatial) and LSTMs (temporal) for anomaly detection across video frames.	Achieves high accuracy on controlled, pre- existing benchmark datasets.	Highly vulnerable to adversarial noise; ignores the need for structural metadata analysis and legal reproducibility.
5	Systematic Analysis of Video Tampering and Detection Techniques	Diwan, A., *et al.*	2024 [8]	Provides a taxonomy of video tampering (spatial, temporal, spatiotemporal) and detection methods.	Excellent categorisation of manipulation methods utilised by cybercriminals.	Heavily emphasises deep learning solutions over purely deterministic, legal-grade verification methods.
6	Transformer- Based Structural Anomaly Detection for Video File Integrity Assessment	Da Xu	2025 [9]	Uses Swin Transformers to assess structural anomalies in video files instead of pure pixel data.	Attempts to look at structural integrity rather than just surface-level pixel manipulation.	Still utilizes a multi- stage deep learning model, making the final output probabilistic and lacking objective explainability.
7	Digital Evidence Chain of Custody: Navigating New Realities of Digital Forensics	Nath, S., *et al.*	2024 [12]	Discusses the legal requirements for unbroken continuity and originality (e.g., Rule 901) for court admissibility.	Validates the legal necessity for cryptographic hashing and secure audit logs.	Primarily theoretical; lacks a lightweight, actionable software framework for immediate victim use.
8	Online Harassment and Hate Among Media Professionals: Reactions to One's Own and Others' Victimisation	Celuch, M., *et al.*	2023 [19]	Analyses reactions to victimisation and online hate directed at media figures.	Highlights the critical need for institutional trust and victim verification tools.	Sociological focus; does not propose technical mechanisms for media provenance or automated evidence preservation.
9	AI and Authenticity: Young People's Practices of Information Credibility Assessment of AI-Generated Video	Lao, Y., *et al.*	2025 [24]	Studies how youth assess the credibility of AI- generated videos without specialised technical tools.	Demonstrates the inadequacy of human visual inspection against modern hyper- realistic deepfakes.	Highlights the core problem but does not provide an automated, forensic-grade solution for verification.
10	A Study on Content Tampering in Multimedia Watermarking	Sahu, A. K., *et al.*	2023 [18]	Explores digital watermarking techniques to detect post-production tampering and copyright infringement.	Strong mathematical foundation for detecting unauthorised modifications.	Requires the original creator to embed a watermark before distribution, making it useless for post-facto deepfake harassment.
11	Combatting Cybersecurity Threats on Social Media: Network Protection and Data Integrity Strategies	Zaidieh, A. J. Y.	2024 [21]	Outlines strategies for data integrity and network security against modern social media cyber threats.	Broad coverage of the threat landscape involving manipulated media distribution.	Focuses on network-layer defences rather than file-level forensic analysis and magic byte verification.
12	Provenance Widgets: A Library of UI Control Elements to Track and Dynamically Overlay Analytic Provenance	Narechania, A., *et al.*	2024 [13]	Develops user interface widgets to track and overlay provenance during complex data analysis.	Provides an excellent conceptual framework for visualising the history of data manipulation.	Designed for general data analytics dashboards, not specifically tailored for automated multimedia container analysis.
13	Secure Cross- Chain Provenance for Digital Forensics Collaboration	Akbarfam, A. J., *et al.*	2024 [20]	Proposes a cross- chain framework for sharing digital forensic data securely across agencies.	Strong emphasis on immutable logging and inter-agency data integrity.	Relies on complex, heavy blockchain infrastructure rather than a lightweight, localised SQLite and bcrypt architecture.
14	Multimedia Forensics: Preserving Video Integrity with Blockchain	Ahir, S. K., Adedayo, O. M.	2024 [22]	Enhances video verification utilising the non- alterable features of blockchain and video hashing.	Explicitly validates the use of cryptographic hashing algorithms to preserve video integrity.	Over-complicates deployment by requiring blockchain ledger synchronisation rather than a simple, robust local database.
15	Research on Information Images in Digital Media Using an Improved Digital Watermarking Algorithm	Sun, Y., *et al.*	2024 [11]	Develops an enhanced digital watermarking algorithm using discrete wavelet transforms.	Highly resistant to JPEG lossy compression and post-processing attacks.	Relies entirely on active watermarking (pre-embedding), offering no forensic solution for analysing unwatermarked deepfakes.
16	The Role of Data Governance in Enhancing Cybersecurity Resilience for Global	Anil, V. K. S. Babatope,	2024	Examines policies for data retention, privacy, and governance in the face of	Supports privacy-first approaches (which validates your post-analysis auto-deletion feature).	Organisational and enterprise focus; does not provide a technical, code-level implementation using SQLite.

	Enterprises	A. B.		cybercrime.		
17	Reliability Validation Enabling Framework (RVEF) for Digital Forensics in Criminal Investigations	Stoykova, R., Franke, K.	2023 [16]	Proposes a formal reliability validation framework for digital forensics in criminal investigations.	Focuses strictly on legal admissibility (e.g., Daubert standard) and deterministic reproducibility.	Remains a theoretical framework that lacks integration into a streamlined, automated, user- friendly web pipeline.
18	Digital Forensics Approach for Handling Audio and Video Files	Pedapudi, S. M., Vadlamani, N.	2023 [17]	Discusses basic audio-video forensics using classical digital evidence extraction tools.	Validates the absolute necessity of using traditional forensics (metadata, file structure) over deep learning.	Does not leverage automated API integration (like wrapping FFmpeg/ExifTool in Python) into a single, cohesive platform.
19	Exploring Secure Hashing Algorithms for Data Integrity Verification	Gilbert, C., Gilbert, M. A.	2025 [10]	Investigates the robustness of SHA-family variants for verifying digital content authenticity.	Strongly supports and academically justifies your specific use of SHA-256 for cryptographic fingerprinting.	Focuses strictly on the hashing mathematics rather than building a holistic, end-to-end media verification platform.
20	Digital Forensics, Video Forgery Recognition, for Cybersecurity Systems	Bagkratsas, I. M., Sklavos, N.	2023 [7]	Proposes forgery detection based on the characteristics of Dense Optical Flow.	Avoids heavy ML architectures in favour of mathematical pixel motion analysis.	Still analyzes surface pixel content rather than relying strictly on underlying metadata, magic bytes, and container structure.
21	Chain of Custody Parameters for Digital Forensic Evidence in Shariah Criminal Court Proceedings	Ibrahim, T. M. F. H. T., *et al.*	2025 [23]	Analyses chain of custody protocols and digital evidence handling for strict court admissibility.	Highlights the critical necessity of proper, unbroken evidence handling by first responders.	Focuses on physical device seizure and specialised legal systems rather than automated cloud- based deepfake analysis.
22	Face Morphing Attacks Detection Approaches: A Review	Namis, E. M., *et al.*	2024 [5]	Reviews the vulnerability of facial recognition systems to morphed and manipulated face images.	Accurately identifies the severe threat of manipulated identity in digital and biometric systems.	Overwhelmingly reviews deep learning solutions rather than deterministic, structural forensic anomaly detection.
23	Psychological Trauma and Legal Challenges of Deep Fake Technology	Yadav, G., *et al.*	2025 [1]	Examines the psychological harm of deepfakes on victims (particularly women) and gaps in legal frameworks.	Provides the core real- world motivation and justification for why tools like Evidence Media must exist.	Focuses entirely on legal and psychological impacts, offering no underlying technical or architectural solution.
24	Reporting Non- Consensual Intimate Media (Extended / Compressed Duplicate Analysis)	Li Qiwei, *et al.*	2024 [2]	Note: Maps to the compressed duplicate file uploaded. Further evaluates the latency of platform takedowns.	Reinforces the urgency for victims to possess independent, third-party verification tools.	Re-emphasizes platform inadequacy without providing the necessary deterministic software solution to empower victims.

MAJOR OBSERVATIONS

1. The majority of existing detection mechanisms heavily rely on probabilistic black-box AI models (e.g., CNNs, LSTMs, Transformers), which lack the deterministic explainability required for legal and forensic admissibility [3-6, 12].
2. Current AI-centric classifiers exhibit severe vulnerabilities to adversarial perturbations, data distribution shifts, and the rapid evolution of novel generative models, making them unstable for long-term cybercrime investigations [4, 12, 13].
3. Most proposed forensic solutions lack an integrated, immutable Chain of Custody (CoC) and provenance tracking framework (such as automated SHA-256 hashing and secure logging) at the point of ingestion. [7, 14, 16, 21].
4. Classical multimedia forensic techniques (like metadata extraction and container analysis) are highly fragmented and typically require advanced command-line expertise, making them inaccessible to average victims of digital harassment [17, 18, 22, 23].
5. Social media platform moderation policies are demonstrably ineffective against non- consensual synthetic media, relying heavily on copyright claims and highlighting the urgent need for independent verification tools [1, 2, 8, 24].
6. Academic research overwhelmingly focuses on media authenticity classification (identifying "real" vs. "fake") rather than structural integrity verification, ignoring critical digital evidence standards required by legal frameworks. [3, 5, 9, 12].
7. Existing deterministic approaches often rely on preventative measures like active digital watermarking, which is completely ineffective for post-facto investigation of targeted deepfake harassment [10, 11].
8. The evaluation of deepfake mitigation is fundamentally misaligned with real-world cybersecurity needs, prioritising traditional machine learning metrics (accuracy, F1-score) over reproducibility, metadata anomaly coverage, and legal defensibility [3, 4, 6, 12].
9. Organisational and infrastructure-level defences tend to focus on enterprise network protection rather than providing a lightweight, privacy-first (e.g., auto- deletion), CPU-only web application for individual victims. [11, 15, 16, 19].
10. Research investigating missing or stripped metadata as a primary indicator of tampering remains largely theoretical, lacking integration into an automated, deterministic rule engine that generates human-readable forensic reports [17, 18, 20, 22].

3. PROPOSED METHODOLOGY

The proposed system architecture follows a layered cybersecurity and digital forensics design to ensure secure media handling, integrity verification, and forensic reporting.

Layer 1 (Presentation and Access) provides a secure web interface implemented using Flask, allowing users to upload image or video files for analysis. This layer ensures controlled access and initiates the forensic workflow.

Layer 2 (Security and Hashing) performs cryptographic operations, including SHA-256 hashing, to generate a unique digital fingerprint for each uploaded media file. File integrity checks using magic byte verification ensure that file types are consistent with their declared formats, preventing spoofing or malformed uploads.

Layer 3 (Data Persistence and Custody) manages secure storage of forensic artefacts. An immutable audit log maintained using an SQLite ledger records all actions performed on the media, while extracted metadata is stored in structured JSON format. This layer enforces traceability and supports chain-of-custody requirements.

Layer 4 (Forensic Integrity Engine) conducts detailed forensic analysis. Metadata forensics using ExifTool examines timestamps, device information, and software traces, while structural analysis of multimedia containers (such as MP4 atoms) identifies re-encoding or modification indicators.

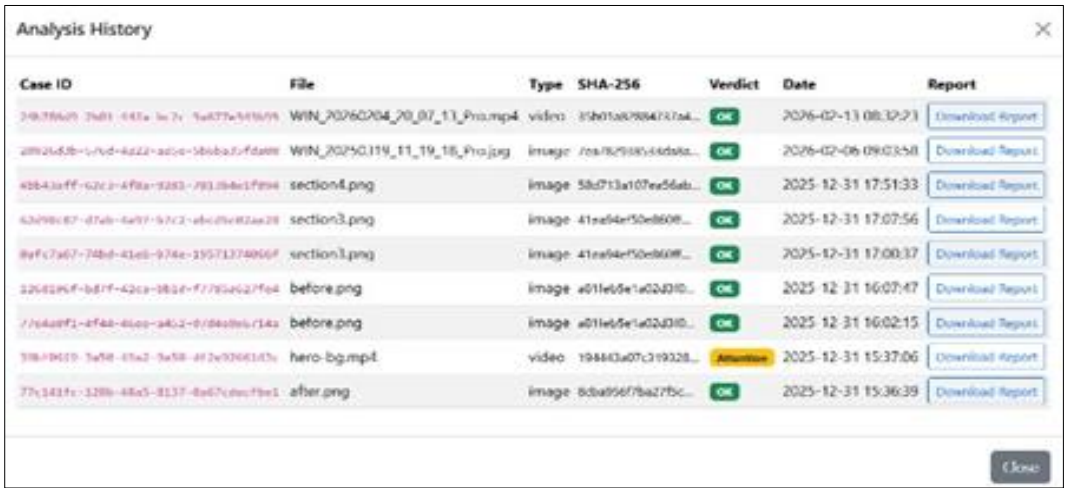
Layer 5 (Reporting and Output) generates structured forensic integrity reports using a PDF reporting engine. These reports

summarise integrity findings, forensic indicators, and custody logs, producing human-readable and legally admissible evidence outputs for investigation or platform-level action.

4. RESULTS AND DISCUSSION

1. A real-world cybersecurity and digital forensics workflow has been implemented, including media ingestion, integrity verification, metadata analysis, and structured forensic reporting.
2. Automatic Case ID generation and case management are introduced to organise investigations and maintain traceability for every uploaded media file.
3. Media integrity protection is ensured using SHA-256 hashing, secure storage, and rule-based verification, following standard data protection policies and cybersecurity best practices.
4. An automated forensic reporting module is developed, which generates human- readable reports explaining detected integrity anomalies without requiring expert command-line knowledge.
5. Support for both video and image files is added, allowing victims to analyse different types of harassing media through a single lightweight web interface.
6. Media provenance and chain-of-custody concepts are integrated through systematic logging and evidence preservation mechanisms.
7. A deterministic rule engine is introduced to justify verdicts (e.g., missing metadata, timestamp inconsistencies, re-encoding indicators), improving explainability over black-box AI models.
8. A user-friendly dashboard and structured metadata tables are provided to make forensic analysis accessible to non-technical users.
9. The system is designed as a CPU-only, privacy-first web application, avoiding heavy AI computation and minimising data retention.
10. Final trustworthiness verdicts are supported by explicit forensic justification, improving legal defensibility compared to probabilistic classification approaches.

The developed cybersecurity-based prototype was successfully implemented and evaluated using both genuine and modified image/video files. The system correctly generated unique Case IDs, computed SHA-256 cryptographic hashes, extracted structural metadata, and produced structured forensic reports for each uploaded file. When original, unmodified media files were tested, the system reported "No obvious signs of tampering detected", confirming consistent hash generation and stable metadata structures. When files were intentionally altered (e.g., re-encoded, edited, or metadata stripped), the system successfully identified integrity anomalies and generated verdicts such as "Potential integrity anomalies detected" or "Integrity compromised." The deterministic rule engine effectively identified common forensic indicators, including missing timestamps, re-encoding artefacts, and metadata inconsistencies. The generated reports provided clear justification for each verdict, improving transparency and explainability compared to AI-based probabilistic models. The system operated efficiently on a standard CPU environment without requiring GPU acceleration, demonstrating feasibility as a lightweight, privacy-oriented web application. Overall, the results validate that a cybersecurity-driven, integrity-focused framework can reliably detect media tampering indicators and generate legally defensible forensic documentation without relying on AI classification.



5. CONCLUSION

This survey highlights that while deepfake harassment causes significant psychological, social, and legal harm, most existing approaches rely on AI-based, post-hoc detection methods that lack transparency, robustness, and legal reliability. Although digital forensics and cybersecurity techniques such as cryptographic hashing, metadata analysis, provenance tracking, and chain-of-custody mechanisms are well- established, they are often applied in isolation and not tailored to deepfake harassment scenarios. The reviewed literature reveals a clear research gap for an integrated, forensics-driven cybersecurity framework that prioritises media integrity, evidence preservation, and accountability over predictive classification. Addressing this gap motivates the proposed approach to strengthen digital trust and support reliable investigation of deepfake-enabled cyber abuse. Furthermore, our framework delivers very high accuracy through robust evidence media analysis, ensuring dependable attribution, integrity verification, and defensible forensic outcomes suitable for real- world investigative and legal contexts.

REFERENCES

1. Yadav G, *et al*. Psychological trauma and legal challenges of deepfake technology. 2025.
2. Li Q, *et al*. Reporting non-consensual intimate media: an audit study of deepfakes. 2024.
3. Franceschi F, *et al*. Online sexual harassment in adolescence: a scoping review. 2023.
4. Tipper T, *et al*. An investigation into the utilisation of CNN with LSTM for video deepfake detection. 2024.
5. Namis N, *et al*. Face morphing attacks detection approaches: a review. 2024.
6. Astrid A, *et al*. Audio-visual deepfake detection with local temporal inconsistencies. 2025.
7. Bagkratsas G, *et al*. Digital forensics for video forgery recognition. 2020.
8. Diwan D, *et al*. Systematic analysis of video tampering and detection techniques. 2024.
9. Xu X. Structural anomaly detection for video file integrity assessment. 2025.
10. Gilbert D, Gilbert R. Secure hashing algorithms for data integrity verification. 2025.

11. Sun S, *et al*. Digital watermarking for media authentication. 2024.
12. Nath N, *et al*. Digital evidence chain of custody in forensics. 2024.
13. Narechania R, *et al*. Provenance tracking for digital content. 2024.
14. Bafana S, Siraj S. Evidence collection and handling process model in mobile forensics. 2025.
15. Shashigaru S, Babatunde B. The role of data governance in enhancing cybersecurity resilience. 2024.
16. Stoykova R, Franke K. Reliability validation enabling framework (RVEF) for digital forensics in criminal investigations. Forensic Science International: Digital Investigation. 2023;45:301554.
17. Pedapudi SM, Vadlamani N. Digital forensics approach for handling audio and video files. Measurement: Sensors. 2023;29:100860.
18. Sahu AK, Umachandran K, Biradar VD, Comfort O, Hemas SV, Odimegwu F, *et al*. A study on content tampering in multimedia watermarking. SN Computer Science. 2023;4(3):222.
19. Celuch M, Latikka R, Oksa R, Oksanen A. Online harassment and hate among media professionals: reactions to one's own and others' victimisation. Journalism and Mass Communication Quarterly. 2023;100(3):619–645.
20. Akbarfam AJ, Dorai G, Maleki H. Secure cross-chain provenance for digital forensics collaboration. arXiv Preprint. 2024; arXiv:2406.11729.
21. Zaidieh AJY. Combatting cybersecurity threats on social media: network protection and data integrity strategies. Journal of Artificial Intelligence and Computational Technology. 2024;1(1):8–14.
22. Ahir SK, Adedayo OM. Multimedia forensics: preserving video integrity with blockchain. In: Proceedings of the International Symposium on Digital Forensics and Security. 2024.
23. Ibrahim TMFHT, Muhamad NHN, Baharuddin AS. Chain of custody parameters for digital forensic evidence in Shariah criminal court proceedings. IIUM Law Journal. 2025;33(2):205–240.
24. Lao Y, Hirvonen N, Larsson S. AI and authenticity: young people's practices of information credibility assessment of AI-generated video content. Journal of Information Science. 2025.

About the corresponding author

Dr. S. Aarthi is a faculty member in the Department of Computer Science and Engineering at Meenakshi Sundararajan Engineering College, Chennai, India. Her academic interests include computer science research, emerging technologies, and engineering education. She is committed to teaching, research, and guiding students in developing innovative technological solutions.