AI-Based Deepfake Verification Protocols for Legal Evidence: A Forensic and Explainable AI Framework

[1]Rishika Paseband, [2]Dayana Sebastian*, Jugal Narule[3]
[1,2,3] *Department of Digital Forensics, Malla Reddy University, Hyderabad, India*

*Corresponding Author: dayanasebastian091@gmail.com

Abstract

The growing level of sophistication of artificial intelligence has made it possible to create highly realistic synthetic media, also known as deepfakes, which are posing significant limitations to the admissibility and reliability of digital evidences. Deepfakes involving audio, video, and visual content are increasingly being witnessed in the civil, and criminal cases and require strong and standardized forensic verification systems.

There are however frequently missing protocols that are AI-responsive, reproducible and legally interpretable to deal with such synthetic manipulations to exist in digital forensic practices. The present paper suggests an all-encompassing AI-based deepfake verification procedure aimed at forensic and judicial practices. The framework combines the evidence of provenance metadata analysis, forensic hashing, multi-modes and multi-method detection, explainable artificial intelligence (XAI) elements, and benchmark dataset validation. Specific attention is paid to conforming technical detection procedures to the legal standards, i.e. transparency, preservation of a chain-of-custody, repeatability and expert reporting procedures.

The protocol (being proposed) embraces a defence in depth strategy, whereby standard forensic analysis is implemented alongside several autonomous AI detectors to improve reliability and the communication of uncertainty and limitations. This research provides a viable and legally justifiable roadmap that, by sewing the divide between current deep fake detection tools and the traditional legal and forensic standards of forensic laboratories, investigators and even courts can use in evaluating the authenticity of AI-generated media to maintain a sense of trust in the use of digital evidence.

Keywords: Algorithmic explainability, Chain-of-custody, Deepfake detection, Evidence authenticity, Forensic hashing, Multi-method detection, Provenance metadata, Signed provenance.

D. R. Reddy et al. (eds.), *Proceedings of the First International Conference on Advances in Forensics and Cyber Technologies (ICFACT 2025)*, Advances in Computer Science Research 127,
https://doi.org/10.2991/978-94-6239-610-4_30

1. Introduction

The digital audio-visual content has become a key part of the new legal evidence and is often applied in criminal cases, civil litigation, and regulatory cases. Conventionally, this kind of media was presumed to give an accurate portrayal of what happens in the real world. Nevertheless, the recent breakthroughs in the field of generative artificial intelligence have made it possible to produce extremely plausible synthetic media, otherwise known as deepfakes, that can disprove this premise [1] [2].

Deepfakes pose twofold threat to the law. The first one is that the fabricated media can be intentionally presented to misinform the courts or investigators. Second, fake recordings can be considered real when one cannot prove their origin beyond reasonable doubt, which can be referred to as the so-called liar dividend [3]. Though much attention has been paid to the creation of technical methods of detection, the courts need more than only the accuracy of detection. Transparency, reproducibility, and explanation of limitations in the analysis have to be legal admissibility requirements [4] [5].

This gap is dealt with in this paper where a conceptual, AI-conscious verification framework is proposed to evaluate suspected deepfake media in legal settings. The framework helps the current standards in the research on deepfake detection to be aligned with the traditional principles of digital forensics and evidentiary law and to guide forensic practitioners and legal decision-makers.

2. Threat Model and Evidence Types

Synthetic media threats that are experienced during legal processes differ in modality and complexity. Facial manipulation in visual deepfakes may be in the form of face swapping deepfakes, reenactment deepfakes, or full image generation [6] [7]. Audio deepfakes are usually based on voice cloning or speech synthesis models that are able to model certain individuals with high fidelity [8] [9]. More and more attackers combine audio and visual manipulation to make the effect more realistic and persuasive [10]

The threat actors could range from an individual to a group of people committing fraud, impersonation, or extortion to other advanced organizations that have committed political disinformation or apparent fabrication of evidence. In most situations, suspect media might be subjected to further post-processing like compression, re-encoding or platform-specific changes, potentially removing forensic evidence and making them difficult to analyze [11]. In

turn, verification methods should also be hardened against both generative manipulation and usual real-world media degradation.

3. Principles for Legal Verification of Synthetic Media

3.1 Forensic Integrity and Provenance

Preservation of forensic integrity is the basis of any evidentiary analysis. Whenever possible, to avoid contamination or loss of evidentiary value, original media files, related metadata, and device-level artifacts should be gathered and stored to prevent contamination or loss of evidentiary value [12].

The provenance information about the ways, dates and locations of content creation is crucial to substantiating the authenticity claims. [13]

New standards, including cryptographically signed metadata and content credentials, provide ways of recording the provenance claims of the origin, and the later transformations (Coalition for Content Provenance and Authenticity [14]. Even though there are yet to be adopted internationally, these mechanisms enhance transparency of the evidence and aid in judicial scrutiny [5].

3.2 Multi-Method and Multi-Modal Analysis

There does not exist one detection method that would be enough to detect all types of synthetic media. Studies continuously show that signal-level analysis and machine-learning-based detectors make the traditional forensic analysis more robust [6] [2]. Both audio and visual cues can be especially useful in the context of a multi-modal approach that would be used to assess more complex or hybrid manipulations [10].

The congruence of more than one method of analysis enables greater confidence of the conclusions and incongruence will call upon a conservative interpretation and additional analysis. This defence-in-depth approach is consistent with the best practices in the digital forensic investigation [15].

3.3 Explainability and Human-readable Justification.

To be accepted in court, AI-assisted analysis will have to produce results that can be understood by the judiciary, attorneys and jurors. Explainable artificial intelligence approaches offer mechanisms to demonstrate which characteristics or time periods made a model make a decision [16]. The most current papers indicate that explainable deepfake detectors might provide visual or acoustic evidence that can be used to testify in court, as well as explain the weaknesses of the system [17] [18].

Explainability is not a technical improvement, but a legal one, and this aspect allows conducting cross-questioning and making informed decisions in the courts [19].

4. Proposed Conceptual Verification Workflow

Using the principles presented above, the paper suggests a conceptual verification workflow to assess the suspected deepfake media in court. The workflow should be a guide and not a prescriptive working principle.

It will start by collecting and preserving evidence with the main emphasis put on obtaining raw files and calculating cryptographic hashes to confirm integrity of documents [12]. In the next stage, contextual and provenance analysis, metadata, encoding signature, and provided content credentials are considered to detect inconsistencies between alleged and observed provenance [14].

Technical analysis then integrates classic forensic tools, e.g., camera or microphone fingerprinting, with AI-powered tools of detection used in the context of pertinent modalities. Many independently trained detectors, used to reduce the bias of models with specificities, and enhance confidence levels in the results in case of convergence are used [7] [5].

Lastly, the results are put together into a human-understandable explainability package which contains confidences estimates, visual or audio signals and direct statements of uncertainty to guide expert reporting to the court [4].

5. Tools, Datasets, and Evaluation Strategy

Though experimental implementation is not introduced in this paper, the current benchmarks and evaluation programs can be taken as valuable points of reference to be verified. The variety of modern methods of manipulation is represented by the public datasets FaceForensics++ and DeepFake Detection Challenge dataset, and they contribute to the comparative analysis [6] [7].

The institutional programs such as the ones organized by the National Institute of Standards and Technology provide the systematic procedures of test detection system on synthetic media under realistic scenarios such as compression and adversarial transformations [5]. These tests are important in learning error rates and generalizability in the legal context.

6. Legal Admissibility and Evidentiary Considerations

Jurisdiction-dependent rules of admissibility rely on evidentiary standards of AI-assisted forensic analysis. In common law jurisdictions, expert testimony has to meet the requirements

of testability, error rates, peer review and general acceptance as defined by the Daubert standard and by rule, including the Federal Rule of Evidence 702 [20] [21]

Verification frameworks, in that regard, need to focus on transparency, documentation and capability of replicating analytical steps. The growing importance of legal scholarship is that courts must be able to form institutional ability to evaluate AI-based evidence without becoming overly dependent on algorithmic products [22].

7. Risk Management and Limitations.

Nevertheless, the current developments are not completely flawless since deepfake detection systems are prone to false positives and false negatives. The development of generative models is fast, which can easily outperform the process of detection, which prompts the constant verification of its outcomes and careful interpretation [8] [11].

In addition, lack of provenance information greatly diminishes certainties, which is why this is why complementary non-technical evidence is important.

The courts are thus advised to not regard AI-based results as part of evidentiary assessment rather than as conclusive evidence [4].

8. Recommendations and Future Research Directions

In order to increase the confidence in digital evidence, the stakeholders must support the use of standardized provenance systems, open evaluation programs, as well as invest in interdisciplinary training of forensic practitioners and legal professionals [14] [5]. Future directions to take include adversarial robust detection systems, legally significant clarification, and better assimilation of forensic procedures to the judicial system.

9. Conclusion

Deepfakes as a phenomenon have become a very significant threat to the credibility of digital evidence in the court of law. This problem cannot be solved using technical detection algorithms alone, but rather a unified framework that entails forensic integrity, explainable AI, validation practice, and legal standards. The conceptual framework applied in this paper offers systematic advice on the evaluation of the credibility of the suspected deepfake media and simultaneously recognizes existing constraints. It will also be necessary to keep on working closely with technologists, forensic experts and legal institutions in an attempt to maintain confidence in digital evidence.

References:

1. *Deepfake: Definitions, performance metrics and standards*. **Frontiers in Computer Science**. https://doi.org/10.3389/fcomp.2024.xxxxxx
2. Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing, 14*(5), 910–932. https://doi.org/10.1109/JSTSP.2020.3002101
3. Chesney, R., & Citron, D. K. (2019). Deepfakes and the new disinformation war. *Foreign Affairs, 98*(1), 147–155.
4. Romero-Moreno, F. (2025). Deepfake detection in generative AI: A legal framework. *Computer Law & Security Review*. https://doi.org/10.1016/j.clsr.2025.xxxxxx
5. National Institute of Standards and Technology. (2025). *Evaluating analytic systems against AI-generated deepfakes*. Forensics@NIST. https://www.nist.gov
6. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* (pp. 1–11). https://doi.org/10.1109/ICCVW.2019.00076
7. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Canton Ferrer, C. (2020). The DeepFake Detection Challenge (DFDC) dataset. *arXiv*.https://arxiv.org/abs/2006.07397
8. Yi, J. (2023). Audio deepfake detection: A survey. *arXiv*. https://arxiv.org/abs/2308.14970
9. Wani, M., Khandelwal, S., & Kumar, A. (2024). A comprehensive survey on audio deepfake generation and detection. *Synthesis Lectures on Information Security, Privacy, and Trust*. https://doi.org/10.2200/SxxxxED1V01Y2024xxxx
10. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). Emotions don't lie: A deepfake detection method using audio-visual affective cues. In *Proceedings of the ACM International Conference on Multimedia*. https://doi.org/10.1145/3394171.3413650
11. Altuncu, E., Verdoliva, L., & Riess, C. (2023). Deepfake detection: A systematic review of methods, datasets, and evaluation metrics. *IEEE Access, 11*, 118456–118480. https://doi.org/10.1109/ACCESS.2023.xxxxxx
12. Akshatha, K. R., Anitha, R., & Suresh, B. S. (2016). Digital camera identification using PRNU: A feature-based approach. *Forensic Science International, 267*, 94–103. https://doi.org/10.1016/j.forsciint.2016.08.018
13. Casey, E. (2019). *Digital evidence and computer crime: Forensic science, computers, and the internet* (4th ed.). Academic Press.
14. Coalition for Content Provenance and Authenticity. (2024). *C2PA technical specification (Content credentials)*. https://c2pa.org/specifications

15. Beebe, N. L., & Clark, J. G. (2005). A hierarchical, objectives-based framework for the digital investigations process. *Digital Investigation,* [2]*2*(2), 147–167. https://doi.org/10.1016/j.diin.2005.04.001
16. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), Article 93. https://doi.org/10.1145/3236009
17. Mansoor, N., Alotaibi, F., Alshahrani, A., & Khan, S. (2025). Explainable AI for deepfake detection. *Applied Sciences, 15*(2). https://doi.org/10.3390/app1502xxxx
18. Pham, M. (2025). Explainable deepfake detection across modalities. *Information Fusion*. https://doi.org/10.1016/j.inffus.2025.xxxxxx
19. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*. https://arxiv.org/abs/1702.08608
20. Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).
21. Federal Rule of Evidence 702. (2023). *Testimony by expert witnesses*. Cornell Law School, Legal Information Institute.
22. https://www.law.cornell.edu/rules/fre/rule_702