*Page 1938 - 1950*


# INTERNATIONAL JOURNAL OF LAW MANAGEMENT & HUMANITIES

## [ISSN 2581-5369]

**Volume 9 | Issue 3**

**2026**

# Challenges in Deepfake Authentication: Harmonizing International Rules for Criminal Evidence

HASRAT BOPARAI[*], JATINDER MAAN[**] AND GURJINDER SINGH[***]

## ABSTRACT

*The legitimacy of all digital evidence used in criminal proceedings is under an unprecedented threat due to the rise of deepfakes. Video and audio files that cannot be verified by current evidentiary standards now appear in courts all over the world. The majority of current AI detection techniques offer a workable substitute; however, pre-AI statutes dictate the legal standing of these AI-based detection systems with respect to their constraints, resilience against adversaries, and stability. This paper compares the regulatory environment for detection systems in four major jurisdictions, the United States, the European Union, the United Kingdom, and India, with the objective of identifying important gaps between the different types of regulation. The specific gaps identified include the judiciary's inconsistent gatekeeping requirements for AI-derived forensic evidence, the lack of a mandatory requirement for forensic certification of AI detection systems, the lack of public disclosure of the accuracy or error rates of AI forensic detection, and the limited cooperation between jurisdictions on detection-tool certification. A new international framework is recommended in order to preserve jurisdictional integrity, safeguard an individual's right to a fair trial, and rebuild trust in digital evidence. This paper suggests that, as part of this new framework, AI detection tools should, at the very least, obtain international certification; the admissibility of AI forensic evidence should require a modified gatekeeping analysis; the parties should give advance proof of the accuracy and error rate of an AI detection system; and the new framework should include certification of detection tools in compliance with the Budapest Convention.*

***Keywords:*** *Deepfakes, Criminal Justice System, Detection, Admissibility, Authenticity, Digital Evidence.*

[*] Author is a Research Scholar at Panjab University, Chandigarh, India.
[**] Author is an Assistant Professor (Part-Time Faculty) at Panjab University, Chandigarh, India.
[***] Author is an Assistant Professor (Part-Time Faculty) at Panjab University, Chandigarh, India.

 [ISSN 2581-5369]

## I. INTRODUCTION

The emergence of highly realistic deepfake technology, synthetic media generated by generative adversarial networks (GANs) and diffusion models, has created a profound crisis of trust in digital evidence. By the end of 2025, deepfakes had been exploited in criminal activities, including the creation of fake alibis, extortion, interference with elections for political reasons, and the production of revenge pornography. When courts receive audio, video, or photographic evidence, they must now determine whether the evidence is authentic or whether it has been altered. However, most of the laws governing the admissibility of digital evidence were established long before the advent of generative AI. Generally, the current laws concentrate on ensuring a proper chain of custody, establishing relevance, and determining whether the evidence satisfies the hearsay rule; they therefore provide little direction on how to evaluate the effectiveness of the AI-based detection tools used to validate or refute the authenticity of digital media. The absence of standard forensic certification for the reliability of AI-generated content is made even more complicated because courts have handled algorithmic error rates inconsistently and international collaboration is scarce. In 2024, a deepfake attempt was occurring approximately once every five minutes on a global basis,[1] with a reported average financial loss of more than $500,000 per deepfake-related fraud incident.[2] Projections from Deloitte's Center for Financial Services indicate that generative AI-related fraud losses are anticipated to exceed $40 billion in the U.S. by 2027, mainly due to the growing prevalence of synthetic media in fraudulent activities.[3]

This paper presents a comparative analysis of four major jurisdictions: the USA, the EU, the UK, and India. The objective is to identify systemic regulatory gaps and recommend a harmonised framework that balances technological realities with fair-trial guarantees.

## II. RESEARCH METHODOLOGY

Adopting an interdisciplinary approach spanning AI, law, and forensic science, this paper employs a doctrinal and comparative legal research methodology. The analysis draws on primary legal sources, including legislation from the USA, the EU, the UK, and India.

[1] Ken Kadet, *Deepfake Attempts Occur Every Five Minutes Amid 244% Surge in Digital Document Forgeries*, ENTRUST (Nov. 19, 2024), https://www.entrust.com/company/newsroom/deepfake-attacks-strike-every-five-minutes-amid-244-surge-in-digital-document-forgeries.

[2] Brightside Team, *Deepfake CEO Fraud: $50M Voice Cloning Threat CFOs*, BRIGHTSIDE AI BLOG (Oct. 19, 2025), https://www.brside.com/blog/deepfake-ceo-fraud-50m-voice-cloning-threat-cfos.

[3] Satish Lalchand, Val Srinivas, Brendan Maggiore & Joshua Henderson, *Generative AI Is Expected to Magnify the Risk of Deepfakes and Other Fraud in Banking*, DELOITTE CTR. FOR FIN. SERVS. (May 29, 2024), https://www.deloitte.com/us/en/insights/industry/financial-services/deepfake-banking-fraud-risk-on-the-rise.html.

 [ISSN 2581-5369]

Secondary sources include peer-reviewed literature from IEEE, ACM, Elsevier, and Springer, as well as authoritative technical reports from NIST, ISO/IEC, Sensity AI, and the Deepfake Incident Database. The web-based search was conducted using combinations of keywords such as "deepfakes", "deepfake detection tools", and "deepfakes and the criminal justice system". The technical discussion of deepfake detection is grounded in recent forensic studies, ensuring that the legal arguments reflect current scientific realities. The study is limited to criminal proceedings in selected common-law and civil-law jurisdictions and does not extend to civil or administrative contexts.

## III. TECHNICAL CHALLENGES OF DEEPFAKE DETECTION IN FORENSIC CONTEXTS

Deepfake detection tools play a crucial role in determining the reliability of digital evidence in criminal proceedings, but these instruments have serious technical limitations that make them unsuitable for use as forensic evidence.

### A. Principal Approaches to Deepfake Detection

Current deepfake detection technologies can be grouped into three primary categories.

**Artifact-based detection:** These techniques rely on identifying visual, audio, or temporal inconsistencies produced during the creation of media content. Examples include abnormal eye-blinking, mismatched lips and speech, irregular micro-expressions, temporal flickering, and inconsistent audio-video synchronisation.[4] The production of specific manual features indicating that video content was not authentic was demonstrated in earlier studies performed by UC Berkeley and the State University of New York, Albany.[5] However, as the technology becomes more sophisticated, it is becoming easier for generative models to create content with fewer visible inconsistencies.

**Statistical and deep-learning-based detection:** The primary methods used today to detect deepfakes are machine learning and deep neural networks, including CNNs, vision transformers, and multimodal networks.[6] The models used today, such as MesoNet, XceptionNet, and EfficientNet classifiers, are trained on large datasets of real and artificial media, allowing them to learn statistical differences between real and synthetic media. The

[4] A. Raza, A. Basit, A. Amin, Z.A. Arfeen, M.I. Masud, U. Fayyaz & T.A. Jumani, *A Comprehensive Review of Deepfake Detection Techniques: From Traditional Machine Learning to Advanced Deep Learning Architectures*, 7 AI 68 (2026), https://doi.org/10.3390/ai7020068.

[5] S. Harris, H.J. Hadi, N. Ahmad & M.A. Alshara, *Fake News Detection Revisited: An Extensive Review of Theoretical Frameworks, Dataset Assessments, Model Constraints, and Forward-Looking Research Agendas*, 12 TECHNOLOGIES 222 (2024), https://doi.org/10.3390/technologies12110222.

[6] Mohd Tahir Irfan, Bhavna Arora, Neha Sandotra & Abrar Ahmed Raza, *On Machine Learning and Deep Learning Based Deepfake Generation and Detection*, 259 PROCEDIA COMPUT. SCI. 1927, 1936 (2025).

 [ISSN 2581-5369]

newest approaches utilise transformer-based architectures and multimodal networks, such as those inspired by CLIP, to examine relationships among audio, text, and visual characteristics across different media types. While these methods yield high accuracy in laboratory testing, they often break down when faced with new manipulations, changes in domain, or adversarial threats to the model.[7]

**Metadata and provenance-based analysis:** Authentication of digital media can also be accomplished through metadata and provenance-based analysis, which verifies a medium's origin, integrity, and modification history rather than merely examining its content. The Coalition for Content Provenance and Authenticity (C2PA) has developed standards that combine cryptographic signatures with provenance metadata at the time of capture, during each stage of editing, and across multiple platforms, as a means of establishing an unbroken chain of custody for all content.[8] Further proposals present blockchain-based tracking as a method for building immutable audit trails. These methods are promising, but to be as effective as possible they must be widely embraced by manufacturers, platforms, and creators.

### B. Forensic Fragility of Deepfake Detection Technologies

Even though deepfake detection research has advanced significantly, shortcomings in dependability, admissibility, and probative value are revealed when these tools are employed in the criminal justice system.

One limitation is that these tools exhibit very low levels of robustness when tested under real-world conditions.[9] The forensic literature consistently demonstrates that even small changes to a piece of digital media, often too minor for a human viewer to recognise, can have a very large impact on the ability to accurately analyse the media. Independent evaluations, including those conducted under the U.S. National Institute of Standards and Technology (NIST) Face Analysis Technology Evaluation (FATE) programme, confirm significant performance deterioration when detectors are exposed to altered or evasive inputs. The future tendency of criminals to employ 'adversarial' or misleading forms of evidence creates a legitimate legal concern regarding a person's ability to appeal a conviction when deepfake technology can be used to create false reports from genuine sources.

---

[7] Felix Viktor Jedrzejewski, Lukas Thode, Jannik Fischbach, Tony Gorschek, Daniel Mendez & Niklas Lavesson, *Adversarial Machine Learning in Industry: A Systematic Literature Review*, 145 COMPUTERS & SEC. 103988 (2024).

[8] Enis Golaszewski, Neal Krawetz, Alan T. Sherman, Edward Zieglar, Sai K. Matukumalli, Roberto Yus, Carson L. Kegley, Michael Barthel, William Bowman, Bharg Barot & Kaur Kullman, *Verifying Provenance of Digital Media: Why the C2PA Specifications Fall Short*, ARXIV (Apr. 27, 2026), https://arxiv.org/html/2604.24890v1.

[9] Felipe Romero-Moreno, *Deepfake Detection in Generative AI: A Legal Framework Proposal to Protect Human Rights*, 58 COMPUT. L. & SEC. REV. 106162 (2025), https://doi.org/10.1016/j.clsr.2025.106162.

Another related issue is the challenge of domain shifting and the limited generalisability of models that are specific to narrow benchmark datasets. The empirical evidence shows that detection models trained on one dataset or manipulation method are often ineffective when applied to new synthesis methods or to real-world sources that have undergone video compression or resizing.[10]

Reported performance metrics further illustrate these limitations. In the academic literature, the area-under-the-curve (AUC) value of the best-performing models on datasets such as FaceForensics++, Celeb-DF v2, and DFDC is frequently reported as greater than 0.95, while commercial detection systems such as Intel's FakeCatcher and Microsoft's Video Authenticator have been documented as achieving detection rates over 90% in curated environments.[11] However, this does not account for the fact that such data typically comes from non-standardised assessments conducted under optimal conditions (high-quality media, ideal lighting, frontal facial orientation, and matched training and testing distributions). In contrast, independent evaluations such as those of NIST consistently show that accuracy is diminished under cross-dataset evaluation, with degraded media quality, and when the system is exposed to unseen methods of manipulation.[12]

Another limitation is the lack of transparency and explainability in detection systems. Most high-performing detection systems utilise deep neural networks, which result in black-box classifiers that provide probabilistic results but no interpretable link between features of the media and the classification outcome. Forensic assessments rely on an understanding of both the rationale and the methods used to reach expert opinions before courts can accept those opinions; the lack of transparency in the method therefore renders the expert opinion of little or no evidentiary value.[13]

Furthermore, detection systems are particularly affected by compression and transcoding, making it increasingly difficult for them to perform reliably on most forms of media.[14] Much of

[10] Ramcharan Ramanaharan, Deepani B. Guruge & Johnson I. Agbinya, *DeepFake Video Detection: Insights into Model Generalisation – A Systematic Review*, 9 DATA & INFO. MGMT. 100099 (2025), https://doi.org/10.1016/j.dim.2025.100099.

[11] Ruben Tolosana, Sergio Romero-Tapiador, Ruben Vera-Rodriguez, Ester Gonzalez-Sosa & Julian Fierrez, *DeepFakes Detection Across Generations: Analysis of Facial Regions, Fusion, and Performance Evaluation*, 110 ENG'G APPLICATIONS OF ARTIFICIAL INTELLIGENCE 104673 (2022), https://doi.org/10.1016/j.engappai.2022.104673.

[12] NAT'L INST. OF STANDARDS & TECH., *GenAI: Deepfakes 2026, A Novel Methodology for Evaluating Deepfake Detection Systems in Forensics*, NIST INFO. TECH. LAB., https://ai-challenges.nist.gov/forensics.

[13] Brian MacKenzie, *Part Three: AI on Trial – Admissibility of AI-Generated Evidence*, JUSTICE SPEAKERS INST. (Oct. 14, 2025), https://justicespeakersinstitute.com/ai-generated-evidence-admissibility-on-trial/.

[14] Tanfeng Sun, Xiao Han, Qiang Xu, Xing Yan & Yueneng Wang, *A Review of Double Compression Detection for Digital Multimedia*, 651 NEUROCOMPUTING 130983 (2025), https://doi.org/10.1016/j.neucom.2025.130983.

 [ISSN 2581-5369]

the digital media used in digital forensics is sourced from social-networking and mobile-messaging platforms, whose aggressive compression and re-encoding procedures can degrade or remove the artefacts that many systems rely upon for successful detection. This has resulted in significantly lower accuracy for such systems. In addition, advances in generative AI architecture routinely outstrip the development and validation of associated detection methods, and this ongoing 'arms race' has created challenges in establishing stable error rates and determining the long-term forensic validity of detection systems.

Finally, there is an institutional gap in the absence of a standardised set of guidelines for performing digital forensic performance evaluations. There are currently no universally accepted frameworks that can accurately evaluate detection capabilities in real-world forensic scenarios and that also account for the cross-platform nature of many forensic situations.[15] The inability to establish a standardised framework for comparing detection tools also hinders the ability of courts to judge the validity and reliability of forensic methods consistently.

## IV. Consequences for Criminal Justice and Evidentiary Decision-Making

The limitations discussed above have both direct and systemic consequences for the operation of criminal justice systems with respect to the use, appraisal, and weighing of digital evidence. In a criminal trial, where a person's liberty and right to due process are at stake, reliance on unreliable detection methods poses significant risks. An elevated incidence of false positives from an AI detection method can result in the erroneous exclusion of legitimate evidence and the discrediting of items that would otherwise support a prosecution or a defence.[16] Conversely, a false negative could lead to the introduction of tampered or fabricated audio-video material into the courtroom under the pretence that it is genuine.[17]

This presents even greater challenges where evidence is presented in adversarial scenarios, since both parties may be able to exploit established weaknesses in the various detection methods.

Additionally, the issue of judicial 'gatekeeping' of AI-based evidence adds a further layer of complexity for the courts. Courts will need to evaluate not only the technical results of an expert

[15] James R. Lyle, Barbara Guttman, John M. Butler, Kelly Sauerwein, Christina Reed & Corrine E. Lloyd, *Digital Investigation Techniques: A NIST Scientific Foundation Review*, NIST Internal Rep. 8354 (2021), https://doi.org/10.6028/NIST.IR.8354.

[16] Francesca Palmiotto, *Detecting Deep Fake Evidence with Artificial Intelligence: A Critical Look from a Criminal Law Perspective* 15 (Mar. 10, 2023) (unpublished manuscript), https://ssrn.com/abstract=4384122.

[17] Cyber Centaurs Team, *Detecting Deepfakes in Legal Cases*, Cyber Centaurs Digital Forensics Res. (Oct. 8, 2024), https://cybercentaurs.com/blog/detecting-deepfakes-in-legal-cases/.

 [ISSN 2581-5369]

witness's work, but also the trustworthiness of the methodology and reasoning.[18] The presence of black-box systems that deliver probabilistic conclusions without clear, understandable explanations significantly impedes meaningful cross-examination and reduces the scope for effective judicial oversight of the systems used to assess the reliability of digital evidence.[19] The adversarial examination of evidence may therefore ultimately be compromised.

**Evidentiary Standards**

These concerns intersect directly with established evidentiary standards. Under the Daubert framework in the United States, courts assess testability, known or potential error rates, peer review, and general acceptance within the relevant scientific community.[20] A similar approach to examining the reliability of techniques has been adopted in England and Wales, as set out in Part 19 of the Criminal Procedure Rules, and in India, as stated in Section 63 of the Bharatiya Sakshya Adhiniyam, 2023 (BSA), where the admission of evidence depends on a demonstration of reliability and transparent methodology.

Against these standards, current deepfake detection technologies are best regarded as investigative or corroborative tools rather than independent forensic evidentiary tools. Reliance on such technology, without sufficient validation under realistic conditions, standardised assessments, and clear and consistent reporting of error rates, could create procedural unfairness and evidence-integrity issues. Responses from regulators and courts regarding the use of deepfake evidence should therefore proceed with caution during this developmental period. When integrating AI-assisted detection into the criminal justice system, there should be strict criteria for evidence validation, mandatory disclosure of limitations, and active judicial supervision. Without such standards, the application of deepfake detection technologies will likely damage, rather than enhance, trust in digital evidence and in the criminal adjudication process.

## V. COMPARATIVE ANALYSIS OF EVIDENTIARY FRAMEWORKS

The rules of admissibility for digital evidence, including cases in which authenticity is disputed by way of alleged deepfake manipulation, are determined by jurisdiction-specific rules of evidence and criminal procedure. As these rules were generally written before the emergence

[18] Tapesh Meghwal, *Admissibility of AI-Generated Forensic Evidence: Legal Standards, Ethical Challenges, and Comparative Jurisprudential Analysis* (July 17, 2025) (unpublished manuscript), https://ssrn.com/abstract=5998017.
[19] Francesca Palmiotto, *The Black Box on Trial: The Impact of Algorithmic Opacity on Fair Trial Rights in Criminal Proceedings* (Apr. 16, 2019) (unpublished manuscript), https://ssrn.com/abstract=6588138.
[20] Righton Smith, *The Daubert Standard: Why the Standard Needs to Be Strengthened*, VT. L. REV., https://lawreview.vermontlaw.edu/the-daubert-standard-why-the-standard-needs-to-be-strengthened/.

 [ISSN 2581-5369]

of generative AI, they do not contain standards specifically tailored to evaluating AI-based detection methods, raising new questions concerning reliability, explainability, error rates, and technological change.

In the USA, the admissibility of expert evidence is determined by Federal Rule of Evidence 702, which establishes a number of criteria that courts must consider when evaluating expert opinion evidence.[21] The Daubert standard requires a court to determine whether an expert's opinion is based on empirical testing, has been subject to peer review, has a low error rate, follows standard operating procedures, and is generally accepted.[22] A small number of states continue to apply the Frye standard, which turns solely on whether the expert opinion is generally accepted within the relevant scientific community.[23] These standards are capable of governing the admissibility of AI-based deepfake detection, but because of the lack of standardised validation procedures, the limited transparency of the datasets on which the AI models were trained, and the uncertainty surrounding non-laboratory performance, decisions will vary from case to case.

In the EU, the AI Act classifies certain AI systems as 'high-risk' based on their use in law enforcement or judicial settings, placing additional obligations on AI developers concerning risk management, transparency, and human oversight.[24] In England and Wales, expert evidence is governed by Part 19 of the Criminal Procedure Rules, which requires that expert evidence be relevant, reliable, and of assistance to the court.[25] As there is no specific guidance on evaluating deepfake technology as a type of evidence, there has been repeated emphasis on the need for validation, methodological transparency, and reliability in relation to new forensic technologies. Guidance from both the Forensic Science Regulator and the Law Commission reviews has identified deficiencies in the current system regarding emerging forensic technology; however, no enforceable rules have been introduced regarding media authentication using AI.[26]

---

[21] Timothy L. O'Brien, *Beyond Reliable: Challenging and Deciding Expert Admissibility in U.S. Civil Courts*, 17 LAW, PROBABILITY & RISK 29, 44 (2018), https://doi.org/10.1093/lpr/mgx010.

[22] J.E. Kurtz & E.M. Pintarelli, *The Daubert Standards for Admissibility of Evidence Based on the Personality Assessment Inventory*, 17 PSYCHOL. INJ. & L. 105, 116 (2024), https://doi.org/10.1007/s12207-024-09508-5.

[23] Anjelica Cappellino, *Daubert vs. Frye: Navigating the Standards of Admissibility for Expert Testimony*, EXPERT INST., https://www.expertinstitute.com/resources/insights/daubert-vs-frye-navigating-the-standards-of-admissibility-for-expert-testimony/.

[24] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), 2024 O.J. (L 1689) 1.

[25] The Criminal Procedure Rules 2025, S.I. 2025/909 (U.K.).

[26] CROWN PROSECUTION SERV., *Forensic Science Regulator Act 2021 and the Forensic Science Regulator's Code of Practice 2023* (Oct. 2, 2023), https://www.cps.gov.uk/prosecution-guidance/forensic-science-regulator-act-2021-and-forensic-science-regulators-code.

 [ISSN 2581-5369]

In India, the BSA, 2023 governs the admissibility of electronic records. It confirms three main aspects of admissibility for electronic records, namely certification, integrity, and continuity of the chain of custody, under Section 63.[27] The standards for establishing authenticity frequently rely on cryptographic hash verification, most commonly the SHA-256 method, which validates the integrity of a file after it has been seized. However, verifying a hash only confirms file integrity; verifying the authenticity of the file's content lies outside the realm of cryptographic hashes. In the case of a deepfake allegation, a matching hash neither supports nor denies that content was synthetically created before it was received by the investigating agency. Neither the BSA nor any allied law contains supported testing procedures or reliability validation for AI-based detection instruments. As such, courts are left to rely on general principles of expert evidence in determining the validity and reliability of AI-based detection methods.

Because the jurisdictions surveyed presently use evidentiary frameworks that prioritise the procedural authenticity of a record rather than substantively assessing the risk of synthetic media, the comparison emphasises the need to establish clearer standards for judicial guidance and forensic methodology concerning the authentication of AI-assisted media in criminal adjudication.

## VI. SYSTEMIC REGULATORY GAPS

The assessment methodologies for expert testimony were created for traditional technology platforms that had a known level of stability and transparency, as well as a predictable degree of error, none of which applies to generative AI. The existing technical deficiencies are not merely the result of a technological limitation; they create systemic regulatory lacunae, thereby creating barriers to reliability, consistency, and fairness in the determination of criminal guilt or innocence with respect to disputed digital media.

### A. Absence of Mandatory Forensic Certification

None of the jurisdictions surveyed require fully independent, standardised forensic certification of deepfake detection tools as a condition of their use at trial. The EU AI Act identifies certain AI tools as high-risk and imposes requirements for risk management, but it does not address whether the tools are admissible in court. Likewise, there are no statutory or procedural processes in the USA, the UK, or India that require independent certification; judges must therefore individually decide whether the tools are reliable on a case-by-case basis, using inconsistent benchmarks and without certified records or audits. This ad hoc methodology may

[27] The Bharatiya Sakshya Adhiniyam, 2023, No. 47, Acts of Parliament, 2023 (India).

 [ISSN 2581-5369]

lead either to the admission of unreliable tools or to the unwarranted exclusion of evidence that should be admitted.

### B. Inadequate and Inconsistent Judicial Gatekeeping

The distinctive features of artificial-intelligence-based detection systems are not sufficiently reflected in current judicial gatekeeping standards, which were created for conventional scientific and forensic methods. As a result, courts often lack an organised framework for assessing real-world error rates, dataset bias, model drift, and the resilience of these technologies to the rapid advances in generative AI. Inconsistent admissibility rulings and decreased legal certainty across jurisdictions have resulted from the lack of specific judicial guidance.

### C. Limited Disclosure of Algorithmic Limitations

Parties using AI-based detection systems are typically not required by the procedural rules governing expert and technical evidence to reveal important details about system limitations, such as false-positive and false-negative rates, dataset representativeness, demographic bias, or susceptibility to adversarial manipulation. In the absence of such transparency, courts and opposing parties are unable to thoroughly examine the basis for claims of evidentiary reliability. Effective cross-examination and expert challenge are hampered by this asymmetry, which compromises the integrity of procedural decision-making and the fairness of the adversarial process.

### D. The Uncertainty of the Burden of Proof

The evidentiary framework provides very little guidance on how to allocate the burden of proof where the authenticity of evidence is challenged on the basis of deepfake manipulation. In many instances the systems in place require one side to meet the burden of proof first, while in others the burden shifts as the challenge becomes more plausible. This lack of clarity increases the potential for inconsistent results or strategic exploitation of the uncertainty surrounding the burden of proof.

### E. Fragmented Expert Accreditation and Forensic Capacity

No internationally recognised accreditation exists for experts who work on the detection of deepfakes or the authentication of media using AI. Most courts depend on qualifications that are not uniform and are largely based on self-reporting. The disparity of resources between regions, especially in many parts of India, amplifies this situation, as public forensic laboratories lack sufficient computing resources, specialised expertise, and access to the best detection tools

 [ISSN 2581-5369]

on the market. They are therefore unable to apply international standards equally across jurisdictions.

### F. Absence of International Coordination

With the increasing prevalence of deepfakes throughout the world, there is a significant lack of international coordination in the recognition of forensic certifications, certification standards, and AI detection results. The existing instruments of the Budapest Convention do not address AI-generated evidence or its validation by means of forensic technology. This prevents effective cross-border cooperation in the investigation and prosecution of deepfake crimes.

## VII. Proposed harmonised framework

### A. International Certification of Detection Tools

An independent evaluation of the instruments used to validate or dispute digital content is required to establish reliable decision-making. It is therefore proposed that all deepfake detection instruments be certified against standards set by an internationally recognised validating authority. Certification should be issued by an authoritative agency, and the basis for certification should include the evaluation of how an instrument responds to various datasets and tests of its reliability in each detection scenario. A certification mark would permit courts to accept that certified tools meet a defined standard without needing to evaluate the technical validity of the instrument on each occasion of use.

### B. Tailored Judicial Gatekeeping

In determining the admissibility of AI-detected or AI-validated evidence, courts should evaluate a range of reliability factors, such as real-world and adversarial error rates; the presence of peer-reviewed validation; robustness and external validity, especially with regard to performance under compression and domain shift across various forensic contexts; the degree of explainability of detection outputs; and the extent to which the relevant techniques have gained acceptance within the forensic science community. The standard outlined by such a framework would function as a doctrinal overlay upon, rather than a replacement for, current admissibility regimes, even though an international certification scheme could be a powerful, if debatable, indicator of reliability.

### C. Procedural Safeguards

Improved procedural safeguards are required for parties using AI detection tools, in order to guarantee equity and transparency in the presentation of AI-based identification evidence. During the pre-trial phase, such parties should be required to reveal the name of the AI tool

 [ISSN 2581-5369]

used, its certification status for forensic identification purposes, and any known limitations. Where limitations are identified, the disclosure should include a summary of the training data and any potential biases affecting system performance.

To maintain the integrity of the adversarial process, opposing parties must be given meaningful opportunities to examine AI-derived evidence independently through expert testimony. The framework would allow courts to recognise a rebuttable presumption of authenticity when media is authenticated using a certified AI detection tool. While preserving judicial discretion to determine admissibility and evidentiary weight, this presumption would operate only at the level of authentication, shifting the burden from the proponent, who would otherwise be required to establish reliability, to the opposing party.

The establishment of accredited expert registers that can support well-informed gatekeeping decisions, together with ongoing judicial training, is also necessary for the effective implementation of these safeguards.

### D. International Cooperation and Mutual Recognition

Because of the transnational nature of deepfake-related offences and digital evidence, this harmonisation cannot be accomplished through domestic reform alone. The framework therefore envisages methods for mutually recognising forensic validation results and certified detection tools across jurisdictions.

Information sharing, expert assistance, and standardised forensic reporting would be facilitated by targeted amendments to existing international instruments, such as the Budapest Convention on Cybercrime, and by the creation of a global coordination mechanism through organisations such as INTERPOL or UNODC. Cross-border enforcement capabilities should be strengthened, and duplication of effort minimised, through coordinated national and regional initiatives.

## VIII. Conclusion

The rise of deepfake technology poses a serious challenge to the integrity of criminal justice systems worldwide. Courts increasingly face complex questions concerning the authenticity of digital media, as sophisticated manipulation techniques undermine traditional assumptions about evidentiary reliability. Existing evidentiary frameworks were not designed for technologies that evolve rapidly, operate opaquely, and exhibit variable error rates. The absence of mandated forensic certification, inconsistent judicial gatekeeping, limited disclosure of algorithmic limitations, unclear allocation of the burden of proof, fragmented expert

 [ISSN 2581-5369]

accreditation, and weak international cooperation together create a regulatory gap that threatens both the reliability of digital evidence and the fairness of criminal adjudication.

This paper has argued for a harmonised, technology-specific response capable of addressing these challenges at a global level. The proposed framework integrates international certification of AI-based detection tools, a tailored judicial gatekeeping standard, enhanced procedural safeguards, and strengthened cross-border cooperation. These measures aim to restore evidentiary trust while remaining consistent with existing admissibility doctrines and due-process guarantees.

Without such reforms, courts risk either admitting fabricated evidence or excluding reliable and probative material, outcomes that directly implicate the right to a fair trial and undermine confidence in the criminal justice system. As the pace of generative AI development continues to outstrip legal and forensic adaptation, timely institutional intervention is essential to bridge the growing gap between technological reality and evidentiary principle.

*****

 [ISSN 2581-5369]