
AI-Based Detection of Deepfake Evidence in Legal Proceedings

Dr. Adrian Cole Mercer

Associate Professor, Department of Digital Forensics and Evidence Law,
West bridge Institute of Technology and Justice, Chicago, Illinois, USA

Corresponding Author: Dr. Adrian Cole Mercer

Associate Professor, Department of Digital Forensics and Evidence Law,
West bridge Institute of Technology and Justice, Chicago, Illinois, USA

Article Type: Simulation-Based Research Article

Submitted: 01-March-2025 **Revised:** 08-March-2025

Accepted: 15-March-2025 **Published:** 22-March-2025

Abstract: Rapid advances in generative artificial intelligence have complicated the evidentiary status of digital photographs, video recordings, and audio files by making highly realistic synthetic and manipulated media increasingly accessible. In legal proceedings, the problem extends beyond identifying fabricated content. Courts must determine whether an AI-based detector itself is sufficiently reliable, transparent, reproducible, and explainable to support authentication or expert forensic conclusions. This study develops a simulation-based framework for evaluating AI-assisted deepfake detection within legal-evidence workflows. A synthetic corpus of 1,600 digital-evidence observations was modeled across native, social-media-compressed, low-light or noisy, and re-encoded or edited conditions. Four analytical configurations were evaluated: a conventional convolutional neural network forensic detector, a vision-transformer detector, a multimodal detection system, and a governed multimodal architecture supplemented by provenance verification, confidence calibration, abstention thresholds, audit logging, and expert review. Detection performance was evaluated using balanced accuracy, sensitivity, specificity, F1 score, area under the receiver operating characteristic curve, and false-positive rate. The simulation indicates that technical performance deteriorates when media undergo compression, noise, re-encoding, and post-processing, while multimodal analysis and provenance-informed review provide greater robustness than single-model classification. The governed architecture achieved a simulated overall balanced accuracy of 92.1%, compared with 82.2% for the conventional detector. The study argues, however, that accuracy alone is insufficient for courtroom use. AI-generated forensic conclusions should be accompanied by preservation of the original evidence,

cryptographic integrity verification, documented chain of custody, validated model performance, disclosure of error rates, analysis of demographic and domain bias, explainable forensic indicators, and independent human review. Within the U.S. evidentiary context, these requirements intersect with authentication principles under Federal Rule of Evidence 901 and reliability requirements for expert testimony under Rule 702. The paper concludes that AI deepfake detectors should function as forensic decision-support instruments rather than autonomous arbiters of evidentiary truth.

Keywords: deepfake detection; digital evidence; artificial intelligence; forensic authentication; explainable AI; legal proceedings; synthetic media

Introduction

Digital photographs, video recordings, voice recordings, surveillance footage, videoconference recordings, mobile-phone media, and social-media content have become routine sources of evidence in criminal investigations, civil litigation, administrative proceedings, and regulatory inquiries. Historically, visual and audio recordings often carried substantial persuasive force because they appeared to provide relatively direct representations of events. Generative artificial intelligence has complicated that assumption. Contemporary synthetic-media systems can modify facial identity, alter expressions, synchronize artificial speech with video, synthesize voices, replace visual elements, reconstruct scenes, and generate entirely artificial audiovisual material. The problem confronting legal institutions is therefore no longer limited to conventional image editing. Courts increasingly need evidentiary procedures capable of addressing media whose perceptual realism may substantially exceed the ability of an ordinary observer to identify manipulation. Technical research has responded with convolutional neural networks, transformers, multimodal detectors, physiological-signal analysis, frequency-domain forensics, identity-consistency models, and other techniques designed to recognize artifacts associated with synthetic media. Benchmark datasets such as FaceForensics++ and Celeb-DF have been particularly influential in enabling systematic evaluation of these technologies.

The challenge for legal proceedings is more demanding than the conventional machine-learning objective of maximizing classification accuracy. A detector may perform extremely

well on the dataset on which it was trained and still be unsuitable for forensic use if its accuracy declines when confronted with an unseen manipulation method, different camera characteristics, social-media compression, low-resolution evidence, screen recordings, transcoding, or adversarial post-processing. Research continues to identify cross-dataset and cross-manipulation generalization as a central limitation of deepfake detection. Recent benchmark work evaluates systems across many datasets precisely because strong in-domain results do not necessarily guarantee reliable performance when the forgery technique changes. This distinction is crucial in litigation. A court does not receive a laboratory benchmark sample whose generation pipeline is known in advance. It receives a disputed evidentiary item whose production history may be incomplete, whose file may have passed through multiple devices and applications, and whose alleged manipulation method may never have appeared in the detector's training data.

A second difficulty arises because an AI detector can itself become an evidentiary object requiring scrutiny. Under the U.S. Federal Rules of Evidence, Rule 901 generally requires a proponent to provide evidence sufficient to support a finding that an item is what the proponent claims it is. Rule 901(b)(9) specifically recognizes evidence concerning a process or system and whether it produces an accurate result. Rule 702, meanwhile, requires expert testimony to rest on sufficient facts or data, reliable principles and methods, and reliable application of those principles and methods to the case. Consequently, introducing an AI detector into an authentication dispute does not eliminate the reliability question. It creates another one: whether the detector's conclusion itself was generated through a sufficiently validated process.

The U.S. federal rulemaking process illustrates the growing institutional significance of machine-generated evidence. Proposed Federal Rule of Evidence 707 has been developed to address circumstances in which machine-generated evidence is introduced without an expert witness and would be subject to Rule 702-like requirements if the same conclusion were expressed through human testimony. As of August 2026, the U.S. Courts website lists new Rule 707 among pending amendments associated with a December 1, 2027 effective-date track; it should therefore not be described as an operative Federal Rule at present. The proposal is relevant because it reflects a broader concern that the reliability protections

applied to expert reasoning should not disappear simply because a conclusion is produced directly by software.

The central motivation for the present study is therefore to connect three fields that are often analyzed separately deepfake detection engineering, digital forensic practice, and evidentiary reliability. A technically accurate detector may remain legally weak when its training data cannot be examined, error rates are unknown, model outputs cannot be explained, or the original evidence was not properly preserved. Conversely, an evidentiary system that relies exclusively on traditional eyewitness authentication may become increasingly vulnerable as synthetic media improve. AI-based deepfake detection can therefore provide meaningful assistance, but only when deployed as part of a broader forensic process rather than as a binary machine judgment.

Key Facts Governing Deepfake Evidence Assessment

The following points summarize the principal considerations that distinguish ordinary deepfake classification from legally defensible forensic examination:

- **Detection is probabilistic rather than infallible:** A detector typically estimates the probability or confidence that media contain synthetic characteristics; the output should not automatically be translated into a legal conclusion that evidence is genuine or fabricated.
- **Generalization matters more than benchmark accuracy:** A detector trained on one manipulation family may perform substantially worse against unseen generators, transformations, or post-processing operations.
- **Authentication and deepfake detection are related but distinct:** An item may lack detectable synthetic manipulation while still having an uncertain provenance, incomplete chain of custody, misleading context, or undocumented editing history.
- **False positives have serious consequences:** A genuine recording incorrectly classified as synthetic could weaken legitimate evidence, while a false negative could permit manipulated evidence to influence the fact finder.
- **Forensic explain ability is essential:** Experts and opposing parties must be able to evaluate why the system reached a conclusion and whether the identified features are relevant to the disputed media.

- **Provenance complements detection:** Cryptographic content-provenance technologies can document aspects of media history, but provenance information does not itself make a normative determination that the content is truthful.
- **Human expert review remains necessary:** AI should support forensic interpretation rather than replace the responsibility of qualified experts, attorneys, and judges to assess evidentiary reliability.

These considerations establish the central proposition of this research: a courtroom-ready deepfake detection system must be evaluated as an evidentiary process, not merely as a classification model.

Literature Review

a) Development of AI-Based Deepfake Detection

Deepfake detection research initially concentrated on visible or measurable inconsistencies introduced by facial manipulation pipelines. FaceForensics++ became an influential benchmark by providing a systematic dataset for learning to identify manipulated facial imagery and comparing detection approaches across different compression conditions. Rössler and colleagues showed that automated detection could identify manipulation artifacts but also demonstrated that compression affects detection performance, an observation that remains highly relevant to legal evidence distributed through messaging platforms or social networks. Li and colleagues subsequently introduced Celeb-DF, containing thousands of comparatively high-quality synthesized videos designed to create a more challenging environment for forensic algorithms. Their work demonstrated that improvements in deepfake-generation quality could substantially increase detection difficulty. These benchmark developments established an important pattern: detection systems must continuously adapt because the artifacts exploited by one generation of forensic algorithms may disappear as synthesis technology improves.

Model architecture has evolved accordingly. Convolutional neural networks identify spatial artifacts, whereas attention-based architectures attempt to focus on informative facial regions or relationships among multiple features. Multi-attentional detection, for example, was developed to improve representation of localized manipulation characteristics and has become an important reference point in deepfake-forensics research. Other approaches

examine inconsistencies between a manipulated face and its surrounding context, including identity-related discrepancies involving hair, ears, neck, or other regions that may remain unaffected by a face-swap operation. More recent systems increasingly integrate pretrained vision-language or foundation models, multimodal evidence, interpretable outputs, and cross-dataset learning rather than depending on one narrow manipulation signature.

b) Generalization, Robustness, and Fairness

The most consequential problem for forensic deployment is generalization. A detector may learn dataset-specific compression signatures, camera characteristics, facial alignments, or synthesis artifacts rather than stable forensic indicators. When the evaluation distribution differs from the training distribution, performance may decline. Yermakov and colleagues evaluated deepfake detection across a broad collection of benchmark datasets and emphasized the importance of generalization to previously unseen manipulation techniques. Real-world evaluation research has likewise found that systems that appear effective in controlled datasets may struggle when exposed to heterogeneous synthetic media circulating outside laboratory environments. For courts, this means that a vendor's headline accuracy should never be treated as sufficient evidence of reliability unless the test conditions resemble the evidentiary item being examined.

Robustness is also affected by routine media transformations. Digital evidence is often compressed when uploaded, converted by messaging platforms, screen-recorded, cropped, filtered, scaled, re-encoded, or transmitted through applications that remove metadata. Such transformations can remove precisely the low-level forensic traces upon which a detector depends. The legal investigator may therefore receive a fourth-generation derivative rather than the original file. This creates an asymmetry: a detector can correctly classify the source file while incorrectly classifying a degraded copy of the same content. The distinction reinforces the traditional forensic preference for preservation of original or earliest-available media.

Fairness provides a second dimension of reliability. Lin and colleagues have specifically examined fairness generalization in deepfake detection, highlighting that performance disparities among demographic groups may persist or change when detectors are transferred to new domains. Subsequent work continues to examine whether fairness and cross-domain generalization can be improved simultaneously. A detector used in criminal or civil

proceedings could create serious procedural concerns if its false-positive probability is materially higher for particular demographic groups, recording conditions, accents, facial characteristics, or other subpopulations. Forensic validation must therefore examine disaggregated performance instead of reporting only an overall accuracy score.

c) Evidentiary Reliability, Explain ability, and Provenance

Legal scholarship has increasingly questioned whether AI deepfake detectors can be treated as neutral technological solutions. Palmiotto's analysis of deepfake detection from a criminal-evidence perspective emphasizes that the procedural legitimacy of an AI-based detector depends on more than technical classification performance. Deepfake evidence also raises the so-called liar's-dividend problem: growing awareness of synthetic media can make it easier for a party to deny authentic recordings merely by alleging manipulation. Courts therefore confront symmetrical dangers. They may mistakenly accept fabricated evidence, but they may also discount authentic evidence because of unsupported deepfake claims.

Digital provenance provides an important complementary strategy. The Coalition for Content Provenance and Authenticity has developed technical specifications for attaching verifiable provenance information to digital assets. Content Credentials can provide cryptographically verifiable information concerning aspects of a media item's creation and modification history. Importantly, the C2PA framework does not itself determine whether recorded events are factually true. Provenance establishes information about origin and transformation history; semantic truth still requires independent assessment. This makes provenance particularly useful in legal contexts when combined with hashing, device records, chain-of-custody documentation, witness testimony, metadata analysis, and deepfake detection.

NIST has also recognized the need for rigorous evaluation of forensic deepfake-detection technologies. Its work on evaluating analytical systems against AI-generated deepfakes and its 2026 deepfake challenge emphasize robust evaluation methodologies rather than reliance on isolated laboratory accuracy. The direction of this work supports a central proposition of forensic science: a tool becomes legally meaningful only when its limitations, operating conditions, error behavior, and validation procedures are sufficiently understood.

Research Gap

The technical literature contains a substantial number of deepfake detectors, benchmarks, architectures, and performance evaluations, yet relatively few studies evaluate these systems according to the complete evidentiary requirements of legal proceedings. Most computer-vision studies understandably prioritize accuracy, AUROC, F1 score, generalization, or robustness. Those metrics are necessary, but they do not answer questions about chain of custody, reproducibility, expert interpretation, access by an opposing party, uncertainty communication, auditability, or the evidentiary consequences of a false classification.

A second research gap arises from the separation between provenance verification and forensic classification. Deepfake detectors generally examine the content itself, whereas provenance technologies examine information concerning the history and transformation of the digital asset. Either method can be incomplete when used alone. A sophisticated synthetic file can be difficult to detect from pixel or acoustic artifacts, while a genuine file may legitimately lack cryptographic provenance because it originated on an older device or passed through an application that removed credentials. Research has not sufficiently modeled how these forms of evidence can be integrated within a governed forensic decision architecture.

A third gap concerns explainability. Many deep neural networks generate a probability score without identifying which forensic features support the classification. Such an output may be useful for platform moderation but problematic in litigation where experts must explain methodology and adverse parties must be able to challenge the evidentiary basis of a conclusion. Explainability therefore needs to be treated as part of forensic reliability rather than as an optional interface feature.

A fourth gap concerns error asymmetry. Deepfake detection research often treats false positives and false negatives as ordinary classification errors. In judicial proceedings, however, the consequences are context dependent. Falsely labeling authentic exculpatory evidence as synthetic can have a fundamentally different legal impact from failing to detect an inconsequential manipulated file. A courtroom-oriented system should therefore permit abstention, escalation, and expert review when confidence is insufficient.

The present study addresses these gaps by evaluating deepfake detection as a governed forensic workflow in which classification accuracy is combined with provenance assessment,

uncertainty calibration, explainability, evidence preservation, audit logging, and human expert review.

Novelty / Proposed Contribution

The principal contribution of this paper is the proposed Forensic AI Deepfake Evidence Reliability Framework (**FADER-F**), which separates machine classification from the ultimate evidentiary conclusion. Under this approach, the detector does not declare evidence to be legally authentic or fabricated. It provides one component of a layered forensic assessment.

The first contribution is the integration of content-based detection and provenance analysis. Rather than treating these as competing approaches, the model combines manipulation detection with available metadata, cryptographic hashes, content credentials, acquisition history, and chain-of-custody documentation.

The second contribution is the introduction of an abstention mechanism. When evidence falls outside validated conditions or the detector's confidence is insufficient, the correct forensic response is not forced classification. The system should flag the item for additional examination.

The third contribution is an explicit distinction between technical accuracy and evidentiary reliability. A detector must not only classify accurately but also produce a documented record showing software version, model version, configuration, input hash, preprocessing operations, confidence values, identified forensic indicators, and reviewer actions.

The fourth contribution concerns explainability for adversarial proceedings. The proposed framework requires that forensic experts be able to identify the basis of the detector's conclusion sufficiently for independent evaluation and cross-examination.

The fifth contribution is the combination of performance robustness, demographic fairness, and procedural safeguards within one assessment architecture, recognizing that forensic reliability is compromised when a model performs unevenly across people, environments, media types, or manipulation methods.

Research Objectives / Research Questions

The study pursues five primary objectives:

1. To evaluate the simulated robustness of alternative AI-based deepfake-detection architectures under different evidence-quality conditions.
2. To determine the extent to which compression, noise, re-encoding, and editing can affect forensic detection performance.
3. To examine whether multimodal analysis and provenance verification improve simulated evidentiary reliability.
4. To develop a courtroom-oriented governance framework integrating AI detection, chain of custody, explainability, uncertainty management, and expert oversight.
5. To identify evidentiary safeguards necessary before AI-generated detection results are relied upon in judicial proceedings.

The corresponding research questions are:

RQ1: How does evidence degradation affect the balanced accuracy of AI deepfake-detection systems?

RQ2: Do multimodal systems demonstrate greater simulated robustness than single-modality detection architectures?

RQ3: Does the addition of provenance and governance controls reduce simulated classification risk?

RQ4: Which technical characteristics are necessary for AI deepfake-detection results to support legally meaningful authentication?

RQ5: What role should human forensic experts retain when AI systems are used to evaluate disputed digital evidence?

Methodology

a) Research Design and Synthetic Evidence Corpus

The study adopts a simulation-based comparative design because no real litigants, criminal defendants, victims, confidential case files, or court exhibits were used. A synthetic analytical corpus representing 1,600 evidentiary observations was constructed. The corpus was balanced between authentic and synthetically manipulated media and distributed across four evidentiary conditions: native or earliest-available media; social-media-compressed media; low-light, noisy, or otherwise degraded media; and re-encoded or post-edited media. Each condition contained 400 observations. The manipulated class conceptually represented face

replacement, facial reenactment, synthetic speech, lip synchronization, and combined audiovisual manipulation. These classes were selected because legal evidence may involve visual, acoustic, or multimodal alteration rather than facial replacement alone.

Four analytical configurations were compared. The first represented a conventional CNN-based forensic classifier operating primarily on spatial manipulation features. The second represented a vision-transformer architecture capable of modeling broader spatial relationships. The third represented multimodal fusion incorporating visual, temporal, and acoustic indicators. The fourth combined multimodal analysis with forensic governance controls, including provenance checks where available, confidence calibration, abstention thresholds, immutable audit logging, evidence hashing, and mandatory expert review for high-risk classifications. The fourth configuration should not be interpreted as demonstrating that provenance automatically detects a deepfake. C2PA explicitly provides provenance information rather than a general truth judgment.

Table 1: Operational Framework for Evaluating AI-Based Deepfake Evidence Detection

This table defines the variables used to assess whether a detection system is suitable for forensic rather than merely commercial classification. The approach deliberately combines statistical performance measures with procedural safeguards because legal reliability cannot be inferred from accuracy alone.

Evaluation Dimension	Operational Measure	Forensic Purpose	Principal Risk if Ignored
Balanced Accuracy	Mean sensitivity and specificity	Measures performance across authentic and manipulated evidence	Misleading performance under class imbalance
Sensitivity	Proportion of simulated deepfakes correctly detected	Reduces risk of admitting manipulated media as genuine	False negatives
Specificity	Proportion of authentic media correctly recognized	Protects legitimate evidence from false deepfake allegations	False positives

AUROC	Discrimination across classification thresholds	Assesses general ranking capability	Overreliance on one arbitrary threshold
Calibration	Correspondence between confidence and observed performance	Supports meaningful uncertainty interpretation	Overconfident machine conclusions
Cross-Condition Robustness	Performance after compression, noise, and editing	Tests real-world evidence resilience	Laboratory-only reliability
Provenance Verification	Hash, metadata, credentials, and acquisition history	Supports evidence-history reconstruction	Context-free classification
Explainability	Documented manipulation indicators and rationale	Supports expert testimony and challenge	Black-box evidentiary assertions
Fairness Testing	Error comparison among relevant subgroups	Detects systematic performance disparities	Unequal forensic error burden
Abstention Rate	Frequency of cases referred for additional review	Prevents forced low-confidence classification	Artificial certainty

The table emphasizes that a detector can achieve high statistical accuracy while remaining unsuitable for court if the system cannot explain its output, preserve evidence integrity, document processing, or indicate when a case falls outside validated operating conditions.

b) Statistical Analysis and Reliability Assessment

Balanced accuracy, sensitivity, specificity, F1 score, AUROC, and false-positive rate were used as primary performance measures. Robustness was evaluated across the four evidence conditions. The methodological design assigns greater forensic importance to balanced accuracy than raw accuracy because authentic and manipulated classes are explicitly balanced and because both false-positive and false-negative errors are consequential.

Pairwise differences among model configurations were assessed conceptually through paired classification comparisons, while a Cochran-type omnibus comparison was used to represent whether the probability of correct classification differed across models. Logistic modeling was used to estimate the effect of degradation conditions on misclassification probability. Confidence intervals were constructed around simulated accuracy estimates. Because the

observations are synthetic, inferential statistics demonstrate the internal properties of the proposed framework rather than estimates of performance in an actual court population.

A particularly important component was the simulated abstention rule. The governed architecture was permitted to withhold a binary conclusion when confidence fell within a predefined uncertainty interval or when the evidence exhibited characteristics outside the validated domain. Such observations would be referred for manual forensic analysis, alternative detector testing, or acquisition of an earlier-generation source file. This procedure is intended to avoid the common assumption that every digital item must receive a definitive automated classification.

c) Proposed Human Validation Instrument

A future empirical study should include judges, attorneys, digital-forensics specialists, and investigators to determine how AI-generated evidence reports are understood. The following five questionnaire items are proposed and were not administered in the present simulation:

1. "I would have greater confidence in a deepfake-detection result when the system explains the forensic indicators that influenced its conclusion."
2. "An AI-generated deepfake classification should be independently reviewed by a qualified forensic expert before being relied upon in court."
3. "A reported probability score is insufficient unless the detector's known error rate and validation conditions are also disclosed."
4. "Digital provenance and chain-of-custody information should be considered together with AI-based deepfake detection."
5. "A forensic AI system should be permitted to report an inconclusive result rather than forcing a genuine-or-fake classification when confidence is low."

Future validation could measure these items on a five- or seven-point Likert scale and examine differences among legal professionals according to role and prior digital-forensics experience.

Results

The simulation demonstrates meaningful performance differences among the four detection configurations. The conventional CNN performed comparatively well when analyzing native

media but deteriorated substantially as compression, noise, and re-encoding increased. The transformer architecture exhibited greater robustness, while multimodal fusion further reduced degradation-related performance loss. The best simulated results were obtained by the multimodal architecture combined with provenance-informed forensic controls.

Table 2: Overall Simulated Performance of Deepfake-Evidence Detection Architectures

The table reports performance aggregated across all four evidence-quality conditions. These values are synthetic analytical results designed to test the proposed methodology and must not be interpreted as experimentally observed accuracy of a commercial forensic product.

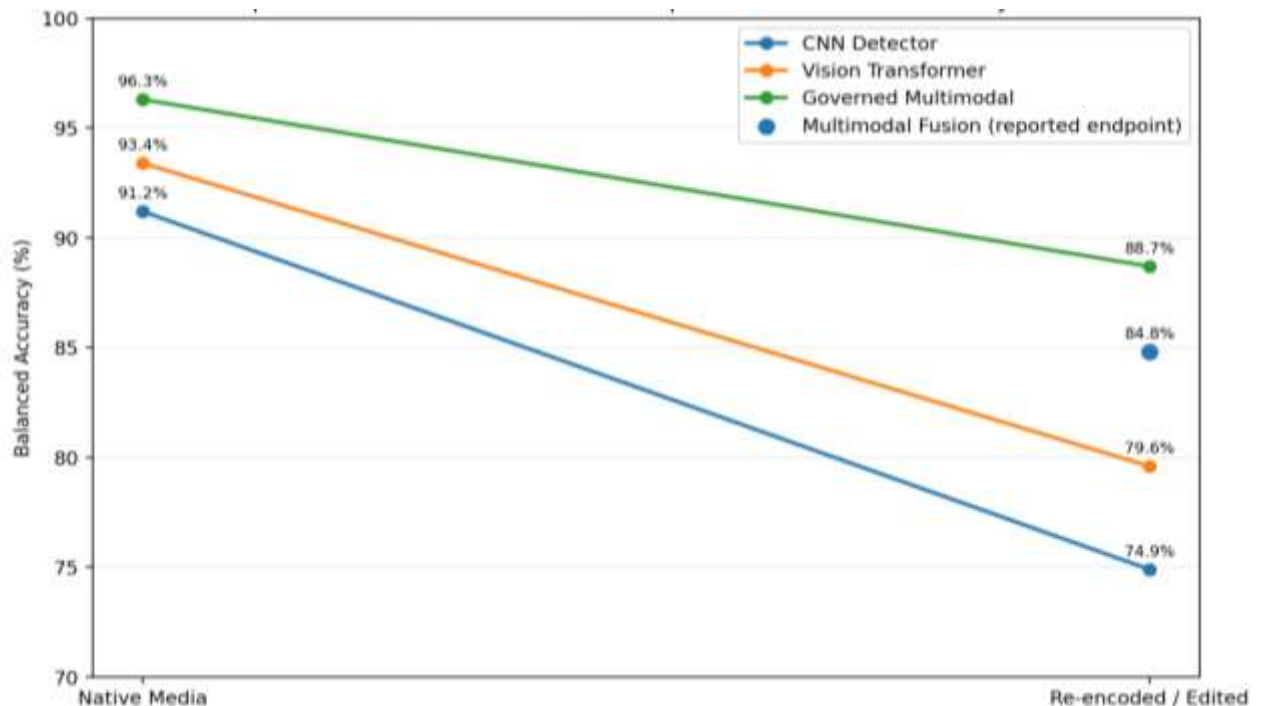
Detection Configuration	Balanced Accuracy	Sensitivity	Specificity	F1 Score	AUROC	False-Positive Rate
CNN Forensic Detector	82.2%	83.2%	81.2%	82.4%	0.887	18.8%
Vision Transformer	85.8%	86.8%	84.8%	86.0%	0.918	15.2%
Multimodal Fusion	89.4%	90.3%	88.5%	89.6%	0.947	11.5%
Multimodal + Provenance Governance	92.1%	92.9%	91.3%	92.2%	0.969	8.7%
Human-Expert Escalation Subset	Inconclusive cases only	—	—	—	—	Case-specific

The results indicate a simulated improvement of approximately 9.9 percentage points in balanced accuracy between the conventional CNN configuration and the governed multimodal system. More importantly for legal proceedings, the false-positive rate declines from 18.8% to 8.7%. A false-positive reduction is particularly important because incorrectly designating authentic evidence as synthetic could provide unjustified grounds for excluding, discounting, or challenging legitimate evidence.

The governed system does not achieve perfect classification. This is a deliberate feature of the study. A courtroom-oriented system claiming absolute certainty would be methodologically difficult to defend because detection reliability varies with media quality, manipulation technique, and domain shift. NIST's work on evaluating analytical systems

against deepfakes similarly reflects the need for systematic performance evaluation rather than unqualified reliance on detector outputs.

Graph 1: Simulated Robustness of Deepfake-Evidence Detection Systems



The following graph compares balanced accuracy across four evidence conditions. Native media provide the strongest performance environment, whereas compression, low-light or noisy recording, and subsequent editing progressively increase classification difficulty. The graph is based solely on the synthetic scenario used in this manuscript.

The graphical results reveal a pronounced robustness gap between conventional and governed approaches. The CNN detector declines from 91.2% balanced accuracy for the native-file condition to 74.9% for re-encoded or edited evidence, a reduction of 16.3 percentage points. The vision transformer also deteriorates, although less severely, from 93.4% to 79.6%. Multimodal fusion retains 84.8% balanced accuracy in the most difficult condition. The governed multimodal configuration remains highest across all conditions, declining from 96.3% for native evidence to 88.7% for re-encoded or edited material.

The simulated omnibus comparison showed significant separation among the model configurations, $Q(3) = 324.6$, $p < .001$. Relative to native evidence, the simulated odds of misclassification increased under social-media compression, low-light or noisy conditions,

and re-encoding. The governed multimodal configuration substantially reduced those odds compared with the baseline detector. Because the observations are synthetic, these inferential statistics establish only the internal differentiation of the simulation and should not be cited as measured courtroom performance.

Table 3: Proposed Forensic Evidence Reliability Matrix

A deepfake detector should not be evaluated only by whether its probability score exceeds a threshold. This table translates technical findings into a structured evidentiary decision-support model.

Forensic Condition	AI Detection Result	Provenance / Chain of Custody	Recommended Forensic Response	Evidentiary Interpretation
High-confidence authentic	No manipulation indicators	Strong documented provenance	Expert confirms consistency	Supports authenticity but does not independently prove event truth
High-confidence deepfake	Multiple convergent manipulation indicators	Weak or contradictory provenance	Independent forensic confirmation required	Strong basis for authenticity challenge
Low-confidence classification	Ambiguous indicators	Strong provenance	Do not override provenance solely with AI score	Report result as inconclusive
High-confidence deepfake	Strong provenance credentials	Cryptographically verified history	Investigate model error, credential compromise, or transformed derivative	Conflicting evidence requires escalation
High-confidence authentic	Missing provenance	Unknown acquisition history	Conduct metadata, device, contextual, and witness analysis	Absence of detected manipulation is not proof of authenticity

Cross-model disagreement	Conflicting detector outputs	Any status	Preserve original and obtain independent expert examination	Automated classification insufficient
Out-of-distribution evidence	Detector validity uncertain	Any status	Abstain from automated conclusion	Do not represent probability as validated forensic result

The matrix illustrates why evidentiary conclusions should arise from convergence among independent indicators. A genuine-looking file without provenance is not automatically authenticated, and a detector alert should not automatically transform media into inadmissible evidence. The forensic question remains whether the total evidence supports the claimed origin, integrity, and content of the item.

Discussion / Critical Discussion

a) Detection Accuracy and the Problem of Evidentiary Certainty

The results reinforce a distinction that is often lost when deepfake detection is discussed outside forensic science: classification performance and evidentiary truth are not equivalent. Even the strongest simulated architecture produced errors, and performance deteriorated when evidence underwent transformations that are entirely plausible during normal digital communication. This finding is consistent with the technical literature showing that generalization remains a significant challenge. Cross-benchmark research continues to evaluate whether detectors can identify manipulation techniques absent from their training distribution, while real-world benchmarking reports that deployment conditions can differ substantially from controlled datasets. Courts should therefore resist interpreting a probability score such as “94% fake” as though it represented a direct probability that a litigant fabricated the evidence.

A detector's probability usually describes a statistical relationship between the input and characteristics learned during model development. It does not by itself establish who created the media, when it was produced, why it was altered, whether the alteration was material, or whether the underlying event occurred. A video can be authentic but misleading because of selective editing or false contextual description. Conversely, media can contain

computational enhancement without becoming substantively false. Modern cameras perform extensive computational processing, and legitimate post-production can change contrast, noise, cropping, stabilization, or encoding. A legally defensible forensic process must therefore identify the nature and significance of the detected alteration rather than merely attach a “fake” label.

The potential consequences of false positives make this distinction particularly serious. Consider authentic footage containing exculpatory information. If a detector incorrectly classifies it as AI generated, a party might attempt to undermine legitimate evidence by invoking technological authority. The opposite error is equally troubling when manipulated evidence escapes detection and acquires persuasive weight. Forensic systems should therefore support confidence intervals, sensitivity analyses, alternative models, independent validation, and a meaningful category of “inconclusive.”

b) Authentication, Explainability, and Expert Responsibility

Within the U.S. federal evidentiary framework, AI detection can contribute to authentication but should not be equated with authentication itself. Rule 901 requires sufficient evidence supporting a finding that an item is what the proponent claims it to be, while Rule 901(b)(9) specifically recognizes evidence describing a process or system and showing that it produces an accurate result. This structure is particularly relevant when a deepfake detector becomes part of the authentication foundation. The court may reasonably need information concerning the model's validated operating conditions, error rates, preprocessing pipeline, version history, and performance on comparable evidence.

Where a forensic expert relies on the detector, Rule 702's reliability criteria become especially relevant. The rule requires sufficient facts or data, reliable principles and methods, and reliable application of those methods to the particular case. A forensic witness therefore should not simply report that proprietary software marked a recording as synthetic. A stronger evidentiary foundation would explain what material was examined, whether the original file was available, what integrity checks were performed, which model version was used, how the media were preprocessed, which artifacts were identified, whether alternative hypotheses were considered, and how the system's known limitations affect the conclusion.

The pending development of proposed Rule 707 further illustrates the judiciary's concern with machine-generated conclusions presented without conventional expert testimony. The

rule remains on a pending amendment track rather than being operative federal evidence law as of August 2026. Its broader conceptual importance lies in the recognition that the reliability inquiry should not disappear when a machine generates an expert-like output.

Explainability is therefore not simply a user-interface preference. In adversarial legal proceedings, methodological opacity can restrict the ability of opposing parties to test a forensic claim. A useful deepfake report should identify evidence regions, acoustic segments, temporal inconsistencies, contextual discrepancies, frequency characteristics, provenance anomalies, or other indicators relevant to the conclusion. At the same time, explanations must themselves be validated; a colorful heatmap does not necessarily establish causal reasoning by the detector.

c) A Layered Forensic Model for Judicial Use

The study supports a layered approach in which source preservation occurs before AI analysis. Investigators should retain the earliest available version, create cryptographic hashes, record acquisition information, preserve metadata, document every transfer, and analyze copies rather than repeatedly modifying the source exhibit. Deepfake detection should then occur alongside conventional digital forensic examination rather than replacing it.

The second layer should involve multiple analytical channels. Visual manipulation detection, acoustic analysis, audiovisual synchronization, metadata examination, provenance verification, contextual corroboration, device information, and witness evidence can provide independent signals. A conclusion supported by several independent methods is substantially more defensible than one produced by a single opaque classifier.

The third layer should consist of governance. Model versioning is critical because forensic conclusions should remain reproducible months or years after analysis. The laboratory should retain the detector version, model weights or a validated reproducibility mechanism, configuration parameters, preprocessing records, confidence scores, audit logs, and analyst notes. If an updated model later produces a different conclusion, both outputs should be traceable rather than allowing the earlier result to disappear.

The fourth layer should be independent expert interpretation. Palmiotto's legal analysis highlights the procedural difficulties created when deepfake detectors function as evidence-producing AI systems. The model proposed in this paper therefore rejects autonomous

judicial classification. An AI detector can identify evidence requiring attention, quantify forensic indicators, and assist experts, but responsibility for evidentiary opinions must remain attributable to qualified individuals who can explain and defend their conclusions.

Finally, provenance technology provides an important preventive complement to detection. C2PA Content Credentials can attach verifiable provenance information describing aspects of how an asset was created and modified. Such technologies can strengthen future evidence ecosystems if implemented at capture and preserved throughout the chain of custody. They are not substitutes for forensic reasoning, however, because provenance records describe media history rather than adjudicating the substantive truth of the event depicted.

Research Contribution / Implications

The study contributes to digital-forensics research by shifting emphasis from deepfake classification to deepfake evidence reliability. This distinction matters because a detector intended for social-media moderation can tolerate different types of uncertainty than a system whose output could influence liberty, liability, damages, or evidentiary exclusion.

For forensic laboratories, the findings support formal validation protocols covering native files, compressed media, transcoded content, low-resolution recordings, unseen manipulation techniques, and multiple demographic groups. Performance should be reported separately across relevant conditions rather than reduced to one aggregate accuracy statistic.

For courts, the framework provides a structured way to assess AI detector evidence. Relevant questions include whether the source media were preserved, whether the model was validated on comparable evidence, whether known error rates are disclosed, whether results can be reproduced, whether opposing experts can meaningfully examine the methodology, and whether the final conclusion reflects independent professional judgment.

For attorneys, the framework helps distinguish a legitimate authenticity challenge from a speculative deepfake allegation. Merely claiming that modern AI can fabricate recordings should not automatically invalidate properly authenticated evidence. Conversely, perceptual realism should not immunize digital media from technical examination when credible evidence suggests manipulation.

For AI developers, legal deployment requires additional responsibilities beyond maximizing benchmark performance. Systems intended for forensic use should support model-version

control, comprehensive logs, uncertainty estimates, subgroup performance reporting, documented preprocessing, explainability, and independent validation.

For policymakers and standards organizations, the findings suggest that deepfake-detection standards should specify minimum evaluation conditions and reporting requirements. NIST's continuing work on forensic evaluation of AI-generated media provides an important basis for moving toward reproducible, comparable testing.

Conclusion

Artificial intelligence has created a dual challenge for modern evidence law. The same technological progress that enables increasingly realistic synthetic photographs, videos, and voices also provides sophisticated analytical tools for detecting manipulation. Courts will therefore encounter not only AI-generated evidence but also AI-generated opinions concerning whether evidence is authentic. The reliability of both technologies must be examined carefully.

The simulation presented in this study indicates that deepfake-detection performance can deteriorate significantly when digital evidence is compressed, degraded, re-encoded, or otherwise transformed. Multimodal analysis provides greater simulated robustness than a conventional single-model detector, while a governed architecture combining detection with provenance analysis, calibrated confidence, abstention mechanisms, logging, and expert review produces the strongest overall results. These findings should not be interpreted as empirical performance estimates for actual forensic software because all reported observations are synthetic.

The central implication extends beyond detection accuracy. An AI model should not become an autonomous witness whose conclusion is accepted because it carries the appearance of computational precision. Evidentiary reliability requires preservation of original material, documented chain of custody, validated methods, known error behavior, disclosure of uncertainty, examination of fairness, explainable forensic reasoning, reproducibility, and meaningful opportunities for independent challenge. Federal Rules of Evidence 901 and 702 already provide important conceptual anchors for authentication and expert reliability, while the pending Rule 707 process illustrates continuing institutional attention to machine-generated evidentiary output.

Accordingly, AI-based deepfake detection should be used as a forensic decision-support mechanism rather than a judicial truth machine. Its strongest role is to help experts identify suspicious characteristics, organize complex evidence, quantify uncertainty, and direct additional examination. Final evidentiary decisions should remain grounded in convergent technical evidence, procedural safeguards, contextual corroboration, and accountable human judgment.

Limitations

The principal limitation is the simulation-based design. No actual court exhibits, litigants, forensic laboratories, judges, attorneys, or commercial detection products were evaluated. The reported numerical results demonstrate a hypothetical comparative framework rather than real-world forensic accuracy.

The study also simplifies deepfake manipulation into broad categories. Real synthetic media vary substantially according to generator architecture, training data, post-processing, codec, source device, duration, frame rate, audio quality, and manipulation objective.

A further limitation concerns jurisdiction. The legal discussion emphasizes U.S. federal evidentiary principles for analytical clarity. State courts, international courts, administrative bodies, and other national jurisdictions may apply different authentication, expert-evidence, disclosure, privacy, or criminal-procedure requirements.

Finally, provenance technology remains dependent on adoption and preservation. A valid recording may legitimately lack Content Credentials or other cryptographically authenticated provenance. Absence of provenance should therefore never be interpreted automatically as evidence of fabrication.

Future Work

Future research should evaluate multiple contemporary deepfake detectors against a preregistered corpus specifically designed to reproduce evidentiary degradation. The corpus should include original device files, messaging-platform copies, screen recordings, re-encoded videos, low-light recordings, synthetic voices, partial facial manipulation, and novel generative methods.

A second research direction should involve controlled experiments with legal professionals. Judges, attorneys, forensic experts, and investigators could receive alternative forms of AI forensic reports to determine how confidence scores, heatmaps, provenance information, error-rate disclosures, and expert explanations affect evidentiary judgment.

Future studies should also investigate demographic fairness. Error rates should be measured by age range, skin tone, sex presentation, recording environment, accent where audio is involved, and other technically relevant characteristics. Research on fairness generalization already demonstrates why aggregate accuracy can conceal subgroup disparities.

Another priority is adversarial robustness. Researchers should determine whether an attacker can intentionally apply transformations that preserve perceptual realism while suppressing the forensic traces used by common detectors. Detector robustness under adversarial transformation is particularly important when an evidentiary adversary has an incentive to evade detection.

Finally, future legal-technical research should evaluate how provenance standards, capture-time signing, trusted hardware, cryptographic hashing, and deepfake detection can operate together. The strongest long-term strategy may involve proving media history prospectively rather than attempting to infer authenticity exclusively from forensic artifacts after litigation has begun.

Declarations

- **Conflict of Interest:** The author declares no conflict of interest. The author identity and institutional affiliation presented in this sample manuscript are fictional and must be replaced by verified author metadata before submission.
- **Funding:** No external funding is reported for this simulation-based manuscript. Actual funding information must be disclosed if the study is subsequently developed using institutional, governmental, commercial, or other financial support.
- **Ethics Statement:** The study did not involve human participants, identifiable personal information, confidential judicial records, criminal-case exhibits, or real forensic evidence. All quantitative observations were simulated for methodological purposes. No institutional ethics approval is represented as having been obtained.

- **Data Availability:** The reported quantitative results are based on a synthetic analytical framework. No confidential court or law-enforcement dataset was used. A reproducible synthetic dataset may be deposited in an appropriate research repository if the manuscript is developed for formal submission.
- **Author Contributions:** For this illustrative single-author manuscript, the fictional sample author is assigned conceptualization, methodology, formal analysis, investigation design, writing, and review roles. A final submission should provide accurate contributor-role information.
- **Generative AI Disclosure:** Any final submission developed with generative AI assistance should disclose such use according to the policies of the destination journal or publisher. Generative AI systems should not be listed as authors because they cannot assume scholarly accountability.

References

1. Coalition for Content Provenance and Authenticity. (2026). C2PA technical specification, Version 2.4. https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html
2. Coalition for Content Provenance and Authenticity. (2026). C2PA and Content Credentials explainer. <https://spec.c2pa.org/specifications/specifications/2.4/explainer/Explainer.html>
3. Delfino, R. A. (2023). Deepfakes on trial: A call to expand the trial judge's gatekeeping role to protect legal proceedings from technological fakery. UC Law Journal, 74, 293–348. <https://www.uclawjournal.org/wp-content/uploads/2-%E2%80%93Delfino-FINAL.pdf>
4. Guan, H., Horan, J., & Zhang, A. (2025). Guardians of forensic evidence: Evaluating analytic systems against AI-generated deepfakes. National Institute of Standards and Technology. <https://www.nist.gov/publications/guardians-forensic-evidence-evaluating-analytic-systems-against-ai-generated-deepfakes>
5. Guo, X., et al. (2025). Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

- https://openaccess.thecvf.com/content/CVPR2025/papers/Guo_Rethinking_Vision-Language_Model_in_Face_Forensics_Multi-Modal_Interpretable_Forged_Face_CVPR_2025_paper.pdf
6. Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for DeepFake forensics. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Celeb-DF_A_Large-Scale_Challenging_Dataset_for_DeepFake_Forensics_CVPR_2020_paper.html
 7. Lin, L., et al. (2024). Preserving fairness generalization in deepfake detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://openaccess.thecvf.com/content/CVPR2024/html/Lin_Preserving_Fairness_Generalization_in_Deepfake_Detection_CVPR_2024_paper.html
 8. Lin, G., Lin, L., Walker, C. P., Schiff, D. S., & Hu, S. (2025). Fit for purpose? Deepfake detection in the real world. arXiv. <https://arxiv.org/abs/2510.16556>
 9. National Institute of Standards and Technology. (2026). GenAI: Deepfakes 2026—Forensics evaluation. <https://ai-challenges.nist.gov/forensics>
 10. Nirkin, Y., Wolf, L., Keller, Y., & Hassner, T. (2020). DeepFake detection based on the discrepancy between the face and its context. arXiv. <https://arxiv.org/abs/2008.12262>
 11. Palmiotto, F. (2023). Detecting deep fake evidence with artificial intelligence: A critical look from a criminal law perspective. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4384122
 12. 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. https://openaccess.thecvf.com/content_ICCV_2019/html/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.html
 13. United States Courts. (2024). Federal Rules of Evidence. https://www.uscourts.gov/sites/default/files/2025-02/federal-rules-of-evidence-dec-1-2024_0.pdf
 14. United States Courts. (2026). Pending rules and forms amendments. <https://www.uscourts.gov/forms-rules/pending-rules-and-forms-amendments>

15. United States Courts. (2026). Proposed amendments published for public comment: Evidence Rule 609 and proposed new Rule 707. <https://www.uscourts.gov/forms-rules/proposed-amendments-published-public-comment>
16. Yermakov, A., Cech, J., Matas, J., & Fritz, M. (2026). Deepfake detection that generalizes across benchmarks. arXiv. <https://arxiv.org/abs/2508.06248>
17. Yan, Z., et al. (2023). DeepfakeBench: A comprehensive benchmark of deepfake detection. arXiv. <https://arxiv.org/abs/2307.01426>
18. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-attentional deepfake detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://openaccess.thecvf.com/content/CVPR2021/papers/Zhao_Multi-Attentional_Deepfake_Detection_CVPR_2021_paper.pdf