



# International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

*(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)*



**Impact Factor: 9.935**

**Volume 15, Issue 8, August 2026**

# Artificial Intelligence in Criminal Investigation and Digital Forensics: A Survey of Explainability, Retrieval-Augmented Reasoning, Graph Intelligence, and Evaluation Practices

Samar Quadri<sup>1</sup>, Mohan Kumar Patel<sup>2</sup>

M. Tech Scholar, Department of Computer Science and Engineering, Patel College of Science and Technology,  
Bhopal, Madhya Pradesh, India<sup>1</sup>

Assistant Professor, Department of Computer Science and Engineering, Patel College of Science and Technology,  
Bhopal, Madhya Pradesh, India<sup>2</sup>

**ABSTRACT:** The application of artificial intelligence (AI) to criminal investigation and digital forensics has expanded rapidly, spanning predictive policing, automated evidence triage, knowledge-graph criminal-network analysis, large language model (LLM)-based investigative reasoning, deepfake and synthetic-media forensics, and blockchain-based evidence custody. This growth has produced a fragmented literature: individual studies advance retrieval, explainability, graph intelligence, content authenticity, or evidence integrity in isolation, using inconsistent evaluation practices that make cross-study comparison difficult, and largely without reference to the regulatory frameworks that increasingly govern their deployment. This survey organizes the field along seven axes: (i) AI in criminal justice and predictive policing, with attention to documented bias and fairness concerns; (ii) multimodal and graph-based investigative frameworks; (iii) large language models and retrieval-augmented generation (RAG) for legal and crime analytics, including hallucination risk and its mitigation; (iv) explainable AI and forensic-compliance mechanisms required for judicial admissibility; (v) deepfake and synthetic-media forensics; (vi) blockchain-based chain-of-custody systems; and (vii) the legal and regulatory frameworks, chiefly the European Union's Artificial Intelligence Act, that now classify several of these use cases as high-risk or prohibit them outright. An eighth axis, evaluation methodology and reproducibility, is treated as a first-class, cross-cutting concern rather than an afterthought, reflecting a broader reproducibility literature that this survey argues has not yet been adequately absorbed by AI-assisted criminal-justice research specifically. A comparative analysis of twelve representative systems across seven dimensions identifies a consistent gap: no reviewed system combines proactive hypothesis generation, explainability, graph intelligence, forensic-compliance tooling, a fully disclosed evaluation methodology, and explicit regulatory-compliance awareness within a single architecture. The survey concludes with open challenges and research directions, including standardized benchmarks for cross-case retrieval, integrated content-authenticity and chain-of-custody pipelines, disclosure norms for AI-assisted development, comparative regulatory analysis beyond the EU AI Act, and human-subject evaluation of investigator workflow impact.

**KEYWORDS:** Survey, Artificial Intelligence, Criminal Investigation, Digital Forensics, Explainable AI, Retrieval-Augmented Generation, Knowledge Graphs, Predictive Policing, Reproducibility, Deepfake Detection, Blockchain, Chain of Custody, EU AI Act, AI Regulation.

## I. INTRODUCTION

The volume and heterogeneity of digital evidence generated by contemporary criminal investigations, spanning witness statements, surveillance footage, mobile-device logs, financial records, and open-source intelligence, has outpaced manual, keyword-driven investigative workflows. Artificial intelligence, and large language models (LLMs) in particular, have been proposed as a remedy across a wide range of investigative tasks: evidence triage, cross-case pattern matching, hypothesis generation, and risk assessment.

This growth has, however, produced a fragmented body of work. Studies on graph-based criminal-network analysis, LLM-based investigative reasoning, explainable digital forensics, and algorithmic fairness in predictive policing have largely developed in parallel literatures, with limited cross-referencing and, more importantly, no shared evaluation standard. A separate but directly relevant literature on reproducibility in machine-learning research has begun to document exactly this kind of fragmentation as a general problem in AI/ML research [1], [2]; this survey argues that criminal-justice AI is a particularly high-stakes instance of it, given that investigative and judicial consequences follow from these systems' outputs.

### **A. Scope and Contribution of this Survey**

This survey covers peer-reviewed and preprint literature from approximately 2017 to 2026, with an emphasis on 2024-2026 publications, spanning: (1) AI and predictive policing in criminal justice, including documented bias and fairness concerns; (2) multimodal and graph-based investigative frameworks; (3) LLMs and retrieval-augmented generation for legal and crime analytics, including hallucination risk; (4) explainable AI and forensic-compliance mechanisms; (5) deepfake and synthetic-media forensics; (6) blockchain-based chain-of-custody systems; and (7) the legal and regulatory frameworks, chiefly the EU AI Act, that now govern several of the preceding categories directly. It contributes a seven-axis taxonomy (Section III), a twelve-system comparative analysis (Section XII), and an explicit treatment of evaluation methodology and reproducibility (Section XI) as a cross-cutting concern rather than a footnote, motivated by the broader ML-reproducibility literature [1], [2].

### **B. Search Methodology and Inclusion Criteria**

This survey began as background reading for a separate project—building a prototype investigative-support system that combined retrieval, explainability, and forensic-integrity checks. As we read across the relevant literatures, a pattern became hard to ignore. The people writing about predictive policing rarely cited the people writing about knowledge graphs for criminal networks. The deepfake-detection community published rigorous benchmarks, and the blockchain community published chain-of-custody architectures, but neither group referenced the other, even though a tamper-proof evidence log and a synthetic-media detector address two halves of the same evidentiary trust problem. Papers on large language models for legal reasoning mentioned retrieval-augmented generation but said little about explainability or courtroom admissibility. Each subfield was advancing on its own track, using its own vocabulary and its own evaluation habits, with almost no cross-referencing.

The fragmentation itself became worth documenting. What started as a literature review for a system paper grew into the survey presented here. The companion system, SCISS, is described separately [cite SCISS paper]; its prototype evaluation shares methodology with this survey (both use a disclosed synthetic benchmark and a disclosed evaluation harness) but the system paper reports algorithmic results while this survey maps the broader research landscape that made those results meaningful.

We identified an initial set of seed papers that we already knew to be central within each axis: Lewis et al. [6] for retrieval-augmented generation, Vaswani et al. [3] for the transformer architecture, Billah [9] and Solanke [33] for explainable forensic AI, Ji et al. [11] for knowledge graphs, the predictive-policing bias report [16], the EU AI Act text [14], Semmelrock and Kowald [2] for reproducibility, and Amerini et al. [13] for deepfake forensics. From these seeds we worked outward in two directions. First, we searched arXiv, IEEE Xplore, ScienceDirect, and the general web between March and May 2026, using combinations of "crime investigation," "digital forensics," "predictive policing," "large language model," "retrieval-augmented generation," "knowledge graph," "explainable AI," "deepfake," "chain of custody," "blockchain," and "AI regulation." We adjusted the search terms as we learned each subfield's internal vocabulary; one community's "case matching" is another's "precedent retrieval." Second, we followed reference trails backward from papers already in our working set. Several important sources, including the hallucination-mitigation taxonomies [4, 5, 7] and the legal-retrieval baseline study [51], surfaced through citation chasing rather than keyword search.

A source was included if it directly addressed AI or algorithmic methods within criminal investigation, digital forensics, or a closely adjacent evidentiary context, and if it appeared in 2017 or later—the year of the transformer architecture [3] that anchors the LLM-based systems reviewed in Section VI. A small number of general sources on fairness, reproducibility, or regulation were included where a specific argument in the survey depended on them, but we defaulted to domain-specific citations wherever possible.

This is not a PRISMA systematic review. We did not use a single pre-registered query string, dual-reviewer screening, or formal inter-rater agreement checks. The seven axes we cover use such different terminology and indexing conventions that forcing them into one search protocol would have implied census-completeness where structured sampling is more appropriate. The reference list should therefore be read as a structured sample of the field's major currents along each axis, not an exhaustive census. We state this limitation directly because one of the arguments this survey makes, in Section XI, is that AI-and-criminal-justice research should disclose its methodology rather than assume it is self-evident.

### C. Organization

Section II introduces background concepts. Section III presents the taxonomy. Sections IV-VII and VIII-X review the literature along each axis. Section XI treats evaluation methodology and reproducibility as a cross-cutting concern. Section XII presents the comparative analysis. Section XIII synthesizes cross-axis themes. Section XIV discusses open challenges, Section XV proposes future research directions, and Section XVI concludes.

## II. BACKGROUND AND PRELIMINARIES

### A. The Digital Investigative Workflow

A typical digital investigation proceeds through evidence acquisition, pre-processing (OCR, transcription, metadata extraction), analysis (pattern detection, cross-case correlation), hypothesis formation, and, ultimately, presentation of findings that may need to satisfy evidentiary admissibility standards. Each stage has attracted AI-based tooling, reviewed in Sections IV-X.

### B. Large Language Models and Retrieval-Augmented Generation

LLMs built on transformer architectures [3] generate fluent text conditioned on a prompt, but are prone to hallucination: plausible-sounding but factually unsupported output [4], [5]. Retrieval-Augmented Generation (RAG) [6] mitigates this by conditioning generation on retrieved, source-traceable documents rather than parametric knowledge alone, and has become the default architecture for domain-specific, high-stakes LLM applications, including legal and forensic ones [7], [8].

### C. Explainable AI (XAI)

XAI techniques such as SHAP and LIME [9] aim to make model decisions interpretable to a human reviewer, a requirement made explicit by judicial admissibility standards that require expert testimony to withstand scrutiny [10]. Explainability is generally treated as a spectrum, from post-hoc local explanations of individual predictions to structurally interpretable models designed for transparency from the outset.

### D. Knowledge Graphs

Knowledge graphs represent entities (persons, locations, organizations) and their relationships as nodes and edges, supporting multi-hop relational queries that text or tabular representations do not directly support [11], [12]. In criminal and legal contexts, they have been used for citation networks, criminal-network analysis, and, increasingly, as a structural constraint on LLM generation ("graph-constrained generation") [8].

### E. Content Authenticity and Chain of Custody

Two distinct evidentiary properties recur throughout Sections VII-IX and are easy to conflate. Chain of custody concerns whether a piece of evidence, once collected, has been altered, and by whom; it is addressed by tamper-evident logging, whether centralized (Section VII) or distributed via blockchain (Section IX). Content authenticity concerns whether the evidence itself, at the moment of collection, depicted a real, unaltered event, a distinct problem that chain-of-custody logging does not address and that deepfake-detection research (Section VIII) targets directly [13]. A system can maintain perfect chain-of-custody integrity over content that was synthetic from the moment of ingestion; this survey treats the two properties as complementary rather than substitutable throughout.

### F. Risk-Tiered AI Regulation

Regulatory instruments reviewed in Section X, principally the EU AI Act, classify AI systems by risk tier (unacceptable, high, limited, minimal) rather than by technical architecture, meaning two systems built with identical techniques can face different obligations depending on their deployment context [14]. This risk-tiered structure is a background concept this survey applies when characterizing the systems reviewed in Sections IV-IX, several of which

would plausibly fall into the high-risk or prohibited categories described in Section X had they been deployed within the EU's jurisdiction.

### III. TAXONOMY OF AI-ASSISTED CRIMINAL INVESTIGATION RESEARCH

This survey classifies reviewed work along seven primary axes, summarized in Table I, with evaluation methodology treated as a cross-cutting eighth concern applicable to all seven.

**TABLE I. TAXONOMY AXES USED IN THIS SURVEY**

| Axis                                         | Representative Concerns                                                     |
|----------------------------------------------|-----------------------------------------------------------------------------|
| AI in Criminal Justice & Predictive Policing | Risk scoring, resource allocation, algorithmic bias/fairness                |
| Multimodal & Graph-Based Frameworks          | Cross-case correlation, entity/relationship extraction, network analysis    |
| LLMs & RAG for Legal/Crime Analytics         | Hypothesis generation, hallucination risk and mitigation                    |
| Explainable AI & Forensic Compliance         | Interpretability, admissibility, chain-of-custody, audit                    |
| Deepfake & Synthetic-Media Forensics         | Detection, attribution, benchmark realism, legal admissibility              |
| Blockchain-Based Chain of Custody            | Tamper-evidence, provenance, distributed-ledger evidence management         |
| Legal & Regulatory Frameworks                | Risk-tiered AI regulation, prohibited practices, law-enforcement exemptions |
| Evaluation & Reproducibility (cross-cutting) | Benchmark realism, disclosed methodology, replicability                     |

Table I.1 summarizes the approximate publication-year distribution of the reviewed sources by axis, as a rough indicator of each axis's maturity rather than a precise bibliometric count (this survey's inclusion criteria, Section I.B, are targeted rather than exhaustive, so these counts should not be read as census-complete).

**TABLE I.1 APPROXIMATE PUBLICATION-YEAR DISTRIBUTION BY AXIS (REVIEWED SOURCES)**

| Axis                                   | 2017-2021 | 2022-2023 | 2024-2026 |
|----------------------------------------|-----------|-----------|-----------|
| Criminal Justice & Predictive Policing | 3         | 2         | 5         |
| Multimodal & Graph-Based Frameworks    | 0         | 1         | 6         |
| LLMs & RAG for Legal/Crime Analytics   | 1         | 0         | 8         |
| Explainable AI & Forensic Compliance   | 1         | 1         | 6         |
| Deepfake & Synthetic-Media Forensics   | 0         | 0         | 4         |
| Blockchain-Based Chain of Custody      | 1         | 0         | 3         |
| Legal & Regulatory Frameworks          | 0         | 0         | 3         |
| Evaluation & Reproducibility           | 0         | 0         | 5         |

The concentration of sources in 2024-2026 across every axis reflects both this survey's deliberate recency emphasis (Section I.B) and a genuine acceleration of publication activity following the wider availability of capable LLMs from approximately 2023 onward; the evaluation-and-reproducibility axis in particular has almost no pre-2024 presence in this review's source set, consistent with it being a comparatively recent, LLM-era concern rather than a longstanding one.

### IV. AI IN CRIMINAL JUSTICE AND PREDICTIVE POLICING

Adler et al. examined the ethical implications of AI in criminal-justice systems, emphasizing that computational intelligence must not reinforce societal bias [15]. Predictive policing tools, which allocate police resources based on historical crime data, have drawn sustained scrutiny on exactly this point: because such tools are trained on historical arrest and enforcement data, they risk encoding and amplifying pre-existing patterns of over-policing in specific communities, a feedback-loop concern documented across multiple independent analyses [16].

A systematic literature review of 30 studies on algorithmic bias in criminal justice found bias-mitigation techniques concentrated on racial bias, with comparatively little attention to age, gender, or socio-economic status [17]. Separately, a systematic review of fairness strategies in predictive policing identified domain knowledge injection, likelihood-function penalties, counterfactual reasoning, and demographic segmentation as the primary mitigation approaches in use, while noting that no strategy simultaneously satisfies all standard fairness criteria and predictive accuracy [18]. Formal analysis has gone further, showing that several distinct, individually reasonable fairness criteria are mathematically incompatible with one another in general, meaning a system cannot satisfy all of them simultaneously except in special cases [19]. Patel and Singh similarly identified lack of transparency and accountability as structural risks of AI-based policing rather than incidental implementation defects [20], and the Alan Turing Institute's analysis of AI against serious online crime concluded that AI-based detection is beneficial only when paired with human-centered governance, not as a substitute for it [21].

The best-documented single case illustrating this literature's abstract fairness concerns is COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), a proprietary recidivism-risk tool used across U.S. courts for sentencing, parole, and bail recommendations. ProPublica's 2016 investigation found the tool nearly twice as likely to misclassify Black defendants as high-risk who did not subsequently reoffend, compared to white defendants (45% vs. 23% false-positive rates) [22]. COMPAS's vendor countered that the tool was equally accurate across racial groups when accuracy was measured as predictive parity rather than error-rate balance; both claims were subsequently shown to be simultaneously true and mathematically incompatible with each other, since a predictor cannot generally satisfy both equalized odds and predictive parity when base rates differ between groups [19]. This is not a flaw specific to COMPAS but a formal property of the fairness-criteria trade-off; COMPAS is simply the case in which it became empirically visible and publicly litigated, and it remains the reference case against which newer risk-assessment and case-matching systems, including several reviewed in Sections V-VI, are implicitly compared whenever a fairness claim is made without specifying which fairness criterion is meant.

The consistent thread across this literature is that bias in predictive policing is a property of the training data and deployment context at least as much as of model architecture, meaning technical mitigation alone is unlikely to resolve it without governance and oversight mechanisms, a point echoed in policy guidance from criminal-justice practitioner bodies [23].

## V. MULTIMODAL AND GRAPH-BASED INVESTIGATIVE FRAMEWORKS

Fischer et al. proposed MULTI-CASE, a transformer-based multimodal framework incorporating provenance tracking, but oriented toward retrospective case review rather than proactive, cross-case hypothesis generation, and without graph-based relational analysis [24]. Kim et al. introduced LAPIS, enabling natural-language querying of criminal databases with effective retrieval for structured queries, but without a systematic explainability layer or the cross-modal evidence fusion required for unstructured evidence [25]. Sharma and Lee proposed rule-based reasoning over structured criminal records, which does not generalize to unstructured evidence or narratively distinct but operationally related cases [26].

Graph-centric approaches address part of this gap directly. The ROXANNE project's Autocrime platform builds knowledge graphs from lawfully intercepted phone and speaker networks to support organized-crime investigation [27]. Pozzi et al. combine knowledge graphs with NLP to analyze instant-messaging evidence in criminal investigations [28]. Outside the criminal domain but methodologically adjacent, GNN-ERE applies graph neural networks to entity-relation extraction for legal case reasoning, reporting substantially improved entity-extraction accuracy and improved interpretability of case-similarity matching [12]; and Wiratunga et al.'s CBR-RAG integrates case-based reasoning into retrieval-augmented generation for legal question answering, structuring retrieval around precedent cases rather than unstructured document chunks [8]. Ji et al.'s broader survey of knowledge-graph representation, acquisition, and application provides the general technical foundation these domain-specific systems build on [11].

Across this literature, graph-based methods consistently improve relational and cross-case pattern detection relative to text-only retrieval, but, with the partial exception of CBR-RAG, are rarely combined with explicit hypothesis-generation or decomposed confidence-scoring mechanisms — a gap Section XII examines directly.

## VI. LARGE LANGUAGE MODELS AND RAG FOR LEGAL AND CRIME ANALYTICS

LLMs such as GPT-4 and Gemini can process witness statements and unstructured evidence to generate hypotheses, identify contradictions, and construct timelines. Reza et al.'s CriX showed explainability-aware LLM frameworks can support demographic-sensitive crime analysis with improved transparency [29], and Farber found AI decision-support in forensic image analysis can approach expert-level performance when grounded in domain-specific evidence [30]. A 2024 IEEE study proposed a RAG-based LLM trained on criminology data for rapid investigative insight [31].

Hallucination remains the central risk for standalone LLM deployment in judicial contexts. Tonmoy et al. provide a comprehensive taxonomy of hallucination-mitigation techniques, categorizing approaches by whether they intervene during retrieval, generation, or post-hoc verification [5]. Huang et al.'s survey similarly organizes hallucination research around principles, taxonomy, and open questions, and notes that RAG reduces but does not eliminate hallucination, since retrieval failures or misleading retrieved context can themselves induce hallucinated output [4]. A focused review of hallucination mitigation specifically within retrieval-augmented LLMs reaches the same qualified conclusion: RAG shifts the failure mode from pure fabrication toward retrieval-context misuse, rather than eliminating factual risk outright [7]. This distinction matters directly for legal and forensic applications, where documented cases of AI-generated false case citations leading to lawyer sanctions illustrate the practical consequences of treating RAG as a complete solution rather than a partial mitigation [32].

Tonmoy et al.'s taxonomy is useful for characterizing where in the pipeline a given investigative system intervenes against hallucination [5]. Retrieval-stage interventions, the category RAG itself belongs to, constrain what context the model sees but not how faithfully it uses that context. Generation-stage interventions, such as constrained decoding or self-consistency sampling (generating multiple candidate outputs and retaining only claims that agree across samples), address the second half of that gap but are rarely discussed in the criminal-justice-specific literature reviewed in this survey; none of the systems in Table II report using them. Post-hoc verification, the third category, checks generated claims against source documents after generation rather than constraining generation itself, and is the closest existing-technique analogue to the decomposed, source-cited confidence scoring reviewed in Section VII, though the reviewed forensic-XAI literature does not frame it in these terms. This survey did not identify any criminal-justice-specific investigative system that combines interventions from all three categories, which is a specific, checkable form of the general hallucination-mitigation gap this section identifies.

Graph-constrained generation has been proposed as a stronger intervention than retrieval alone: rather than using a knowledge graph merely to improve what a model retrieves, some recent legal-AI systems use the graph to constrain what the model is permitted to claim at generation time [8]. This distinction, between improving retrieval and constraining generation, is a useful organizing principle for evaluating how seriously a given RAG-based investigative system treats hallucination risk.

## VII. EXPLAINABLE AI AND FORENSIC COMPLIANCE

Digital forensics has a specific and demanding explainability requirement: outputs must be defensible under expert-testimony standards such as Daubert or Frye, not merely intuitively plausible [10]. Billah developed an XAI system for digital forensics using SHAP and LIME to improve the transparency of evidence flagging [9], building on Solanke's earlier framework for explainable digital-forensics AI aimed at mitigating analyst distrust in AI-based analysis [33]. A 2025 practitioner study involving digital-forensics experts found that, despite a range of proposed XAI solutions, existing work focuses primarily on theoretical analysis while overlooking the practical feasibility constraints practitioners actually face [34]. A separate line of work addresses explainability specifically for unsupervised learning in forensics, converting unsupervised models into functionally equivalent, explainable representations, motivated by the observation that real-world forensic data is predominantly unlabeled, unlike the labeled benchmarks most XAI techniques assume [35].

Durán argued that AI forensic outputs require computational-reliabilism principles: epistemic justification through traceable, reproducible computational processes, not merely accurate output [36]. A preliminary robustness study of AI-driven forensic tools reinforces this, arguing cross-validation and explainability are both necessary to establish transparency in how a tool's outputs relate to specific evidentiary labels [37]. Security concerns compound the explainability requirement: a 2025 secure-by-design generative-AI framework for cloud forensics highlights prompt-

injection as an underexplored attack surface specific to LLM-based forensic tools, distinct from the accuracy and bias concerns discussed elsewhere in this survey [38].

### VIII. DEEFAKE AND SYNTHETIC-MEDIA FORENSICS

Generative models capable of producing convincing synthetic images, audio, and video have introduced a distinct evidentiary problem: not whether evidence has been tampered with after collection (the chain-of-custody concern of Section VII and Section IX), but whether the evidentiary content itself was ever real. Amerini et al. survey the current state of deepfake media forensics, characterizing detection approaches by modality (image, audio, video) and by whether they target low-level pixel/signal artifacts or higher-level semantic inconsistencies, and identify generalization to unseen generation methods as the field's central unsolved problem [13]. This generalization gap is not merely theoretical: Deepfake-Eval-2024, a multi-modal benchmark built from deepfakes actually circulated in the wild during 2024, found that detector performance on in-the-wild content is substantially lower than on the curated academic benchmarks such detectors are typically evaluated against, a finding directly analogous to the benchmark-realism concerns raised for case-matching evaluation in Section XI [39].

A separate integrative review situates deepfake detection alongside broader multimedia-forensics and cybersecurity concerns, arguing that detection tooling alone is insufficient without parallel advances in provenance tracking and platform-level content authentication [40]. From a law-enforcement operations perspective, Europol's assessment of deepfakes as an investigative challenge reaches a similar conclusion from the practitioner side: it identifies deepfakes as a growing threat to the reliability of digital evidence and to witness/victim credibility assessments specifically, and calls for law-enforcement-specific detection tooling and training rather than reliance on general-purpose detectors developed for social-media content moderation [41].

For an investigative-support system such as the SCISS prototype discussed in Section XII, the implication is that evidence authenticity cannot be assumed once evidence integrity is verified: a hash-chained audit log (Section VII) proves an image file has not been altered since ingestion, but says nothing about whether the image content depicts a real, unaltered event. Deepfake detection and chain-of-custody integrity are complementary, not substitutable, safeguards, a distinction this survey has not seen made explicit in the reviewed investigative-AI literature (Section V), which is one of the specific gaps identified in Section XIV.

### IX. BLOCKCHAIN-BASED CHAIN OF CUSTODY

Section VII described hash-chained audit logging as a mechanism for tamper-evident record-keeping within a single system. A parallel and increasingly active literature proposes distributed-ledger (blockchain) architectures for the same purpose, with the additional property that no single custodian, including the investigating agency itself, can unilaterally alter historical records without detection by other participants in the ledger. B-CoC, an early Ethereum-based implementation, established the basic pattern: evidence content is hashed and the hash, together with custody-transfer events, is recorded on-chain, while the evidence itself remains off-chain [42]. More recent systems have moved toward permissioned ledgers: a Hyperledger Fabric-based design logs the full evidence lifecycle as immutable, permissioned ledger entries with cryptographically timestamped custody events, arguing that permissioned (rather than public) blockchains better match law-enforcement governance requirements around participant vetting and access control [43]. Loffi et al.'s systematic literature review extends this design space to include self-sovereign identity (SSI) systems, in which custodians (investigators, forensic analysts, couriers) hold cryptographically verifiable digital identities independent of any single central authority, and argue that combining SSI with blockchain-based custody logging addresses an access-control gap in earlier blockchain-CoC proposals: the identity of who performed a logged action, not merely the fact that some action occurred, is itself made tamper-evident [44]. A study of blockchain-based evidence management in a post-conflict prosecution context makes the institutional case for this design explicitly, arguing that permissioned blockchain systems are well suited to settings of institutional fragmentation, where no single sovereign custodian can be assumed to exist or to remain trustworthy throughout a prosecution's timeline, treating evidentiary integrity as something demonstrated through distributed, structured records rather than through trust in any one party [45].

Evaluation practice within this literature is itself instructive for the reproducibility concern of Section XI: blockchain chain-of-custody systems are typically evaluated on ledger-level metrics such as transaction throughput, write latency,

and storage overhead relative to a non-blockchain baseline, rather than on whether the system detects tampering when actually tested against an adversarial attempt to alter a record [43]. This favors a performance-engineering framing over a security-validation one; a system can report excellent throughput while never having been shown, empirically, to detect a real tampering attempt. This survey did not identify a blockchain chain-of-custody study in its source set that reports an adversarial tamper-detection test comparable in kind to a live penetration test, a gap directly relevant to the forensic-reliability requirements discussed in Section VII.

The reviewed blockchain chain-of-custody literature evaluates systems along dimensions largely distinct from those used elsewhere in this survey: ledger write throughput and confirmation latency (how quickly a custody event becomes immutably recorded), storage overhead of on-chain hash commitments versus off-chain evidence content, and, for permissioned ledgers, the number of participating nodes required before Byzantine fault tolerance guarantees hold. None of the reviewed blockchain-CoC studies report these metrics using a shared benchmark or a shared hardware/network baseline, so comparative throughput claims across the systems reviewed here should be treated with the same caution this survey applies to unqualified accuracy figures elsewhere (Section XI): a throughput number without a stated baseline is not directly comparable to another system's throughput number.

The hash-chained audit log design reviewed in Section VII and evaluated elsewhere in this research program is a single-custodian analogue of this literature's core idea: detecting tampering via recomputed hash verification. The distributed-ledger literature reviewed here generalizes that idea to multi-party settings, at the cost of materially greater deployment complexity (ledger governance, participant onboarding, off-chain storage integration) that single-system audit logging avoids. Which design is appropriate depends on whether a single investigating agency, or multiple mutually distrusting parties (e.g., cross-jurisdictional task forces, defense and prosecution both requiring independent verification), must be able to audit the chain of custody, a distinction the reviewed literature does not always make explicit.

## X. LEGAL AND REGULATORY FRAMEWORKS FOR AI IN CRIMINAL JUSTICE

Every technical capability reviewed in Sections IV-IX now operates within an increasingly explicit regulatory perimeter. The European Union's Artificial Intelligence Act (Regulation (EU) 2024/1689) is the most comprehensive example to date: it establishes a four-tier risk classification and, notably for this survey's scope, classifies a specific and broad set of law-enforcement AI use cases as high-risk by default, including systems used to assess the risk of an individual offending or re-offending, to profile individuals in the course of criminal detection or investigation, and to assess the reliability of evidence [14]. High-risk classification triggers substantive obligations: conformity assessment prior to deployment, human-oversight requirements, and registration in an EU-wide database, obligations that apply directly to systems structurally similar to the predictive-policing and case-matching systems reviewed in Section IV and Section V [46].

The AI Act additionally prohibits outright several practices with direct relevance to investigative AI: untargeted scraping of facial images from the internet or CCTV footage to build facial-recognition databases, and risk-assessment of an individual's likelihood of committing a criminal offense based solely on profiling or personality-trait analysis, without objective, verifiable facts directly linked to criminal activity [46]. This second prohibition bears directly on the predictive-policing bias literature reviewed in Section IV: it does not resolve the fairness-criteria incompatibility result discussed there [19], but it does establish a legal floor beneath which profiling-only risk assessment is not merely inadvisable but impermissible in EU jurisdictions, regardless of a system's measured fairness properties.

Real-time remote biometric identification in publicly accessible spaces for law-enforcement purposes is likewise prohibited by default, with narrow, court-approved exceptions for a limited set of serious-crime scenarios. This is a stricter default position than several of the systems reviewed in Section V and Section VI implicitly assume, none of which discuss biometric-identification legality as a design constraint.

Outside the EU, comparably comprehensive statutory frameworks are scarce. In the United States, the White House Office of Science and Technology Policy published a "Blueprint for an AI Bill of Rights" in October 2022, articulating five principles, including protection from algorithmic discrimination and a right to human alternatives to automated decisions, that closely parallel several EU AI Act obligations in substance [47]. The Blueprint, however, is explicitly non-binding: it creates no enforceable obligations, imposes no conformity-assessment requirement analogous to

Section X's high-risk tier, and a subsequent legislative effort to codify comparable protections, the 2022 Algorithmic Accountability Act, did not advance out of Congress. This leaves U.S. law-enforcement AI governed primarily by pre-existing, non-AI-specific civil-rights statutes rather than by a purpose-built, risk-tiered regime.

Two subsequent developments partially qualify this picture. First, NIST's AI Risk Management Framework, though voluntary and non-regulatory in form, has become a de facto reference standard: federal contractors face procurement-linked expectations to demonstrate NIST-aligned governance, and several sectoral regulators reference its GOVERN-MAP-MEASURE-MANAGE structure in enforcement guidance, despite NIST having no sanctioning authority of its own [48]. Second, state-level legislation has begun to supply the binding obligations the federal Blueprint lacks: Colorado's SB 24-205, effective February 2026, imposes impact-assessment requirements on developers and deployers of high-risk AI systems, defined by reference to consequential decisions including several categories that would map onto this survey's Sections IV-IX [49]. The resulting U.S. picture is therefore not simply "unregulated" relative to the EU, but jurisdictionally fragmented in a different way: voluntary federal technical guidance, binding but state-specific legislation with differing thresholds and enforcement mechanisms, and no single national conformity regime analogous to the EU AI Act's high-risk tier. The practical consequence for this survey's scope is that the same investigative-AI architecture may face binding conformity obligations in one jurisdiction, binding but narrower obligations in a second, and only voluntary guidance in a third; this asymmetry, discussed further in Section XIV, means compliance analysis cannot be done once at the architecture level but must be redone per deployment jurisdiction, a burden the reviewed technical literature does not discuss.

## **XI. EVALUATION METHODOLOGY AND REPRODUCIBILITY**

This axis is treated as cross-cutting because it concerns how the claims in Sections IV-X should be weighed, rather than a separate technical capability. Semmelrock and Kowald document persistent reproducibility barriers across ML-based research generally, including underspecified methodology, unreported evaluation protocols, and selective result reporting [2]. Desai and Padalkar argue the field lacks shared, precise definitions of repeatability, reproducibility, and replicability, which in turn makes it difficult to even state what "reproducing" a claimed result would require [1].

Within criminal-justice AI specifically, Singh identifies the absence of standardized validation protocols as a central, specific obstacle to the legal admissibility of AI-generated forensic evidence, distinct from the general ML reproducibility problem in that it has direct legal consequences [32]. Practitioner guidance has begun to respond: the Council on Criminal Justice's AI Task Force calls for rigorous, independent validation of criminal-justice AI tools rather than reliance on vendor or developer claims [23].

A narrower but instructive result concerns baseline reporting specifically: Rosa et al. show that BM25 [50] remains a difficult baseline for dense or embedding-based retrieval methods to outperform in legal case-retrieval tasks specifically [51]. This is relevant to evaluation practice generally, since a retrieval-based system's advantage over BM25 is a more informative and more easily verified claim than an unqualified accuracy figure reported without a comparable baseline. Foundational retrieval and embedding methods commonly used as components in the systems reviewed above, BM25 itself [50], BERT [52], Sentence-BERT [53], and the original RAG formulation [6], remain the shared technical vocabulary against which newer systems' claimed improvements should be, but are not always, benchmarked.

## **XII. COMPARATIVE ANALYSIS**

Table II compares twelve representative systems and frameworks discussed in this survey across seven dimensions: whether the system performs proactive hypothesis generation (as opposed to retrospective review only); whether it includes an explicit explainability mechanism; whether it incorporates graph-based relational intelligence; whether it includes forensic chain-of-custody or audit-log compliance mechanisms; whether its evaluation methodology is disclosed with sufficient detail to be independently assessed; and whether the system or its authors explicitly address regulatory compliance (e.g., EU AI Act risk-tier obligations, Section X).

TABLE II. COMPARATIVE ANALYSIS OF REPRESENTATIVE SYSTEMS

| System                            | Hypothesis Gen.       | Explain.         | Graph Intel.        | Forensic Compl.                                  | Disclosed Eval.                | Regulatory Awareness |
|-----------------------------------|-----------------------|------------------|---------------------|--------------------------------------------------|--------------------------------|----------------------|
| MULTI-CASE [24]                   | No                    | Partial          | No                  | Provenance only                                  | Partial                        | No                   |
| LAPIS [25]                        | No                    | No               | No                  | No                                               | Partial                        | No                   |
| Autocrime [27]                    | No                    | No               | Yes                 | No                                               | Partial                        | No                   |
| Pozzi et al. [28]                 | No                    | No               | Yes                 | No                                               | Partial                        | No                   |
| GNN-ERE [12]                      | Partial               | Partial          | Yes                 | No                                               | Yes                            | No                   |
| CBR-RAG [8]                       | Yes                   | Partial          | Partial             | No                                               | Partial                        | No                   |
| CriX [29]                         | Partial               | Yes              | No                  | No                                               | Partial                        | No                   |
| IEEE RAG-criminology [31]         | Yes                   | No               | No                  | No                                               | No                             | No                   |
| Billah XAI-DF [9]                 | No                    | Yes              | No                  | No                                               | Yes                            | No                   |
| Deepfake-Eval-2024 [39]           | No                    | Partial          | No                  | No                                               | Yes                            | No                   |
| Blockchain CoC (Hyperledger) [43] | No                    | No               | No                  | Yes (distributed)                                | Partial                        | Partial              |
| prototype SCISS                   | Rule-based + live LLM | Decomposed score | Partial (flat tags) | SHA-256 + hash-chained log, adversarially tested | Yes (disclosed dataset & code) | No                   |

This row describes the SCISS prototype, reported in full in a companion paper currently under review. Its evaluation methodology and results are detailed in that companion work. It is included here for completeness in the comparative analysis.

No reviewed system combines proactive hypothesis generation, an explicit multi-dimensional explainability mechanism, graph-based relational intelligence, forensic-compliance tooling, a fully disclosed evaluation methodology, and explicit regulatory-compliance awareness. Systems oriented toward relational analysis (Autocrime, Pozzi et al., GNN-ERE) generally lack hypothesis generation and compliance tooling; systems oriented toward LLM-based reasoning (CriX, the IEEE RAG-criminology study) generally lack graph intelligence and, in the latter case, disclosed evaluation methodology. The deepfake-detection and blockchain chain-of-custody entries, added to reflect Section VIII and Section IX, illustrate a further pattern: systems in these two categories are typically evaluated rigorously within their own narrow technical scope (detection accuracy; ledger throughput and tamper-resistance) but do not integrate with the hypothesis-generation or graph-reasoning capabilities reviewed in Sections IV-VI, and none of the twelve systems reviewed here explicitly addresses regulatory-compliance obligations as a design input rather than an afterthought. The final row reports the SCISS prototype, described in a companion paper; it is included as a reference point, not as a peer-reviewed system within this survey.

The seven-dimension completeness standard used in Table II is a survey-organizing device, not a design requirement that every investigative-support system should satisfy. Some combinations of capabilities may be unnecessary or architecturally incoherent for specific deployment contexts, and a system that omits, for example, blockchain-based custody logging may have made that choice deliberately rather than through oversight. The pattern Table II reveals is not that individual systems are deficient, but that the literature's separate research programs have advanced without reference to one another, and that capabilities developed in one subfield (explainability, graph reasoning, custody logging, content-authenticity detection) have not yet been brought together into a single architecture by any published system. This is consistent with the research-gap framing throughout Sections IV-X: the literature has advanced along each axis largely independently.

**XIII. DISCUSSION: CROSS-AXIS SYNTHESIS****A. Evidentiary Trust as a Composite Property**

Sections IV-X, read individually, each address a distinct technical or regulatory concern. Read together, they suggest that what an investigator, prosecutor, or court ultimately needs — confidence that a given piece of AI-assisted evidence or analysis is trustworthy — is not a single property that any one reviewed axis supplies on its own, but a composite of at least three separable claims: that the underlying content is authentic (Section VIII), that its custody has been unbroken and tamper-evident since collection (Section VII, Section IX), and that any AI-generated inference drawn from it is explainable and evidentially grounded rather than hallucinated (Section VI, Section VII). No system reviewed in Section XII's comparative table addresses all three simultaneously, which this survey takes as its central synthesis rather than merely an item in Section XIV's challenge list.

**B. Regulation as a Convergence Point**

Section X's review of the EU AI Act is placed deliberately after the six preceding technical axes rather than at the survey's outset, because its practical significance only becomes visible once those axes are in view: nearly every system category reviewed in Sections IV-IX (predictive-risk scoring, criminal profiling, evidence-reliability assessment, biometric identification) maps onto a category the AI Act treats as high-risk or prohibits outright [46]. This suggests regulation is not a peripheral compliance concern for this literature but a convergence point that technical design in every preceding axis will increasingly need to treat as a first-class constraint, in the same way forensic admissibility standards already constrain the explainability design choices reviewed in Section VII.

**C. Reproducibility as the Connective Tissue**

Section XI's reproducibility concern is not merely one axis among eight; it is what would allow a claim made under any of the other seven axes to be checked. A deepfake-detection accuracy figure (Section VIII), a blockchain ledger's throughput claim (Section IX), a predictive-policing fairness audit (Section IV), and a case-matching F1 score (Section V) are all, structurally, the same kind of claim: a number obtained by running some code on some data under some protocol. The reproducibility literature reviewed in Section XI applies identically to all of them [1], [2], which is why this survey treats it as connective tissue across the taxonomy of Table I rather than as a peer axis to be weighed against the others.

**XIV. OPEN CHALLENGES**

- Hallucination risk in judicial contexts: RAG reduces but does not eliminate hallucination [7], [4], [5], and the specific consequence profile in legal settings (sanctionable false citations) differs from the consequence profile in general-purpose chatbot use, warranting domain-specific mitigation standards rather than generic ones.
- Fragmented, non-comparable evaluation practices: benchmark datasets, evaluation protocols, and reported metrics vary widely across the reviewed literature, with few studies disclosing enough methodological detail for independent replication [1], [2], [32].
- Algorithmic bias and fairness in predictive components: documented bias-amplification feedback loops [16] and the formal impossibility of satisfying all fairness criteria simultaneously [19] mean bias mitigation must be treated as an ongoing governance requirement rather than a solved technical preprocessing step.
- Legal admissibility standards remain unsettled: explainability techniques developed for general ML (SHAP, LIME) were not designed against Daubert/Frye-style admissibility standards specifically, and their sufficiency for that purpose remains an open legal, not just technical, question [10].
- Practical feasibility gap: practitioner-facing studies consistently report that proposed XAI and AI-forensics solutions remain theory-heavy relative to practitioners' operational constraints [34].
- Deepfake-detection generalization gap: detectors evaluated on curated academic benchmarks perform substantially worse on in-the-wild content [39], and no reviewed investigative-AI system integrates deepfake detection with its evidence-integrity pipeline (Section VIII), leaving content authenticity and record tamper-evidence as separate, unconnected concerns in practice.
- Chain-of-custody architecture mismatch: single-custodian hash-chained logs (Section VII) and multi-party blockchain ledgers (Section IX) address different trust models, but the reviewed literature rarely states which trust model a given deployment context actually requires before selecting an architecture.
- Regulatory fragmentation: the EU AI Act's risk-tiered obligations [14], [46] have no close international equivalent identified within this review's scope, meaning multinational investigative-AI deployments face materially different

legal obligations, or outright prohibitions, by jurisdiction, a compliance dimension largely absent from the technical literature reviewed in Sections IV-IX.

- Jurisdiction and language coverage: the great majority of reviewed systems assume English-language, single-jurisdiction deployment, with cross-border and multilingual investigative scenarios comparatively underserved.

## **XV. FUTURE RESEARCH DIRECTIONS**

1. Standardized, disclosed benchmarks for cross-case retrieval and hypothesis-generation tasks, ideally spanning multiple jurisdictions and crime categories, to allow the kind of cross-study comparison Section XII shows is currently difficult.
2. Explicit disclosure norms for AI-assisted development and evaluation in this specific research area, extending the general reproducibility literature [1], [2] to address the additional legal-admissibility stakes identified by [32].
3. Deeper integration of graph-constrained generation [8] as a stronger hallucination intervention than retrieval-only RAG, evaluated specifically for the judicial admissibility consequences of residual hallucination.
4. Human-subject evaluation of investigator workflow impact, which remains almost entirely absent from the reviewed literature despite being the ultimate justification for deploying these systems operationally.
5. Fairness-aware predictive components evaluated jointly with accuracy, rather than as a separate post-hoc audit, given the documented incompatibility of simultaneously satisfying multiple fairness criteria [19].
6. Integrated content-authenticity and chain-of-custody pipelines that combine deepfake/synthetic-media detection (Section VIII) with tamper-evident evidence logging (Section VII, Section IX), rather than treating authenticity and integrity as separately solved problems.
7. Comparative regulatory analysis extending Section X beyond the EU AI Act to other jurisdictions as comparable frameworks mature, to establish whether risk-tiered, prohibited-practice-based regulation of investigative AI generalizes or requires jurisdiction-specific redesign.

## **XVI. CONCLUSION**

This survey organized AI-assisted criminal-investigation and digital-forensics research along seven substantive axes, plus a cross-cutting evaluation-and-reproducibility concern, and compared twelve representative systems across seven dimensions. The consistent finding is one of fragmentation rather than a single dominant paradigm: graph-based, LLM-based, explainability-focused, deepfake-forensics, and blockchain-custody research programs have each advanced largely independently of one another and of the regulatory frameworks, chiefly the EU AI Act, that now govern several of them directly. No reviewed system combines all the reviewed capabilities within one architecture, evaluation methodology is disclosed unevenly across the literature, and regulatory-compliance awareness is essentially absent as an explicit design input across every system reviewed. The most consequential open problem may not be a missing algorithmic capability, but the absence of shared, disclosed evaluation standards, integrated authenticity/integrity pipelines, and jurisdiction-aware compliance design robust enough to support the judicial and operational trust these systems are ultimately meant to earn.

## **REFERENCES**

- [1] A. Desai and N. Padalkar, "What is Reproducibility in Artificial Intelligence and Machine Learning Research?," AI Magazine, vol. 46, e70004, 2025.
- [2] R. Semmelrock and D. Kowald, "Reproducibility in Machine-Learning-Based Research: Overview, Barriers, and Drivers," AI Magazine, vol. 46, e70002, 2025.
- [3] A. Vaswani et al., "Attention Is All You Need," NeurIPS, vol. 30, pp. 5998–6008, 2017.
- [4] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, and B. Qin, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," ACM Trans. Information Systems, vol. 43, no. 2, 2025.
- [5] S. Tonmoy, S. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das, "A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models," arXiv:2401.01313, 2024.
- [6] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," NeurIPS, vol. 33, pp. 9459–9474, 2020.
- [7] "Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review," Mathematics (MDPI), vol. 13, no. 5, p. 856, 2025.

- [8] N. Wiratunga, R. Abeyratne, L. Jayawardena, K. Martin, S. Massie, I. Nkisi-Orji, R. Weerasinghe, A. Liret, and B. Fleisch, "CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering," *Proc. Int. Conf. Case-Based Reasoning (ICCBR)*, Springer, 2024, pp. 445–460.
- [9] M. Billah, "Explainable AI for Digital Forensics: Ensuring Transparency in Legal Evidence Analysis," *J. Forensic Science Research*, vol. 9, no. 2, pp. 109–116, 2025.
- [10] S. W. Hall, A. Sakzad, and S. Minagar, "A Proof of Concept Implementation of Explainable Artificial Intelligence (XAI) in Digital Forensics," *Proc. Int. Conf. Network and System Security*, Springer, 2022, pp. 66–85.
- [11] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications," *IEEE Trans. Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 494–514, 2022.
- [12] "GNN-ERE: A Graph Neural Network Framework for Entity-Relation Extraction and Intelligent Legal Case Reasoning," *Informatica*, vol. 49, pp. 431–448, 2025.
- [13] I. Amerini, M. Barni, S. Battiato, G. Bestagini, G. Boato, T. S. Bonaventura, V. Bruni, R. Caldelli, F. De Natale, R. De Nicola, L. Guarnera, S. Mandelli, T. Marchioni, D. Milani, R. Piva, A. Pistocchi, A. Rocha, A. Roman, D. Vazquez-Padin, and M. Zampoglou, "Deepfake Media Forensics: Status and Future Challenges," *Journal of Imaging*, vol. 11, no. 3, p. 73, 2025.
- [14] European Parliament and Council, "Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)," *Official Journal of the European Union*, 2024.
- [15] J. Adler, J. Antoine, and L. Al-Saadoon, "Minding the Machines: On Values and AI in the Criminal Legal System," *Centre for Justice Innovation Report*, 2024.
- [16] "Algorithmic Bias and the Erosion of Procedural Fairness in Predictive Policing," *Human Rights Research Center Report*, 2026.
- [17] "A Literature Review: Bias Detection and Mitigation in Criminal Justice," *Eng. Proc.*, vol. 107, no. 1, p. 72, 2025.
- [18] "Algorithmic Fairness in Predictive Policing," *AI and Ethics*, Springer, 2024.
- [19] R. Binns, "On the Apparent Conflict Between Individual and Group Fairness," *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAT\*)*, 2020, pp. 514–524.
- [20] R. Patel and M. Singh, "Legal and Ethical Impact of AI in Criminal Justice," *IJNRD*, vol. 9, no. 8, pp. 261–271, 2024.
- [21] Centre for Emerging Technology and Security, "AI and Serious Online Crime," *The Alan Turing Institute Report*, 2025.
- [22] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias: There's Software Used Across the Country to Predict Future Criminals, and It's Biased Against Blacks," *ProPublica*, May 23, 2016.
- [23] Council on Criminal Justice, Task Force on Artificial Intelligence, "Assessing AI for Criminal Justice: A User Decision Framework," *Council on Criminal Justice Report*, 2026.
- [24] M. T. Fischer, Y. Metz, L. Joos, M. Miller, and D. A. Keim, "MULTI-CASE: A Transformer-based Ethics-aware Multimodal Investigative Intelligence Framework," *arXiv:2401.01955*, 2024.
- [25] J. Kim, S. Choi, and H. Park, "LAPIS: Language Model-Augmented Police Investigation System," *arXiv:2407.20248*, 2024.
- [26] A. Sharma and D. Lee, "An AI Knowledge-Based System for Police Assistance in Criminal Investigation," *Expert Systems*, vol. 41, no. 1, e13524, 2024.
- [27] P. Motlicek et al., "Autocrime: An AI-Powered Platform for Criminal Network Analysis from Lawfully Intercepted Communications," *Forensic Science International: Digital Investigation*, vol. 54, p. 301937, 2025.
- [28] R. Pozzi et al., "Combining Knowledge Graphs and NLP to Analyze Instant Messaging Data in Criminal Investigations," *Proc. WISE 2024 / arXiv:2509.26487*, 2025.
- [29] S. Reza, T. Banerjee, and R. Kumar, "CriX: Intersection of Crime, Demographics and Explainable AI," *Proc. ICAART*, 2025, pp. 412–420.
- [30] S. Farber, "AI as a Decision Support Tool in Forensic Image Analysis," *J. Forensic Sciences*, vol. 70, no. 4, pp. 1024–1035, 2025.
- [31] "Large Language Models in Modern Forensic Investigations," *Proc. IEEE Conference, IEEE Xplore*, 2024.
- [32] S. Singh, "Reliability and Admissibility of AI-Generated Forensic Evidence in Criminal Trials," *arXiv:2601.06048*, 2025.
- [33] A. A. Solanke, "Explainable Digital Forensics AI: Towards Mitigating Distrust in AI-Based Digital Forensics Analysis Using Interpretable Models," *Forensic Science International: Digital Investigation*, vol. 42, p. 301403, 2022.

- [34] "Practitioner-Driven Framework for AI Adoption in Digital Forensics," Forensic Science International: Digital Investigation, 2025.
- [35] "Bridging Knowledge Gaps in Digital Forensics Using Unsupervised Explainable AI," Forensic Science International: Digital Investigation, vol. 53, p. 301924, 2025.
- [36] J. M. Durán, "Computational Reliabilism for AI-Based Forensic Evidence Evaluation," Patterns, vol. 5, no. 6, p. 100013, 2024.
- [37] "Exploring the Robustness of AI-Driven Tools in Digital Forensics: A Preliminary Study," arXiv:2412.01363, 2024.
- [38] "Automating Cloud Security and Forensics Through a Secure-by-Design Generative AI Framework," Proc. ICDF2C, 2025.
- [39] S. Chandra et al., "Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024," arXiv:2503.02857, 2025.
- [40] "Unmasking Digital Deceptions: An Integrative Review of Deepfake Detection, Multimedia Forensics, and Cybersecurity Challenges," Frontiers/PMC, 2025.
- [41] Europol Innovation Lab, "Facing Reality? Law Enforcement and the Challenge of Deepfakes," Europol Report, 2022.
- [42] A. Bonomi, M. Casini, and C. Vagata, "B-CoC: A Blockchain-Based Chain of Custody for Evidences Management in Digital Forensics," arXiv:1807.10359, 2018.
- [43] "A Blockchain-Based Chain of Custody for Digital Evidence: Design and Evaluation," Hyperledger Fabric implementation study, 2025.
- [44] L. Loffi, G. L. Camillo, C. A. Souza, and C. B. Westphall, "Management of the Chain of Custody of Digital Evidence Using Blockchain and Self-Sovereign Identities: A Systematic Literature Review," 2025.
- [45] "Blockchain Based Chain of Custody and Digital Evidence Legality in Post-Conflict Prosecutions," Frontiers in Blockchain, 2026.
- [46] European Commission, "High-Level Summary and High-Risk Classification Guidance Under the AI Act," Directorate-General for Communications Networks, Content and Technology, 2025.
- [47] White House Office of Science and Technology Policy, "Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People," Oct. 2022.
- [48] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, Jan. 2023.
- [49] Colorado General Assembly, "Concerning Consumer Protections in Interactions with Artificial Intelligence Systems (SB 24-205)," effective Feb. 2026.
- [50] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009.
- [51] G. M. Rosa, R. C. Rodrigues, R. Lotufo, and R. Nogueira, "Yes, BM25 is a Strong Baseline for Legal Case Retrieval," arXiv:2105.05686, 2021.
- [52] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers," Proc. NAACL-HLT, 2019, pp. 4171–4186.
- [53] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Proc. EMNLP-IJCNLP, 2019, pp. 3982–3992.



INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  [ijirset@gmail.com](mailto:ijirset@gmail.com)



[www.ijirset.com](http://www.ijirset.com)

Scan to save the contact details