International Journal of Advances in Data and Information Systems
Vol. 7, No. 3, December 2026, pp. 1054~1068
ISSN: 2721-3056, DOI: 10.59395/ijadis.v7i3.1702

Blockchain and Cryptocurrency Forensics for Cybercrime Investigation: From Trace to Proof, A New Investigative Frontier

Rahman Noviyanto[1], Ruki Harwahyu[2]
[1,2]Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia, Depok, Indonesia

Article Info

Article history:

Received Aug 13, 2026
Revised Oct 09, 2026
Accepted Oct 20, 2026

Keywords:

Blockchain forensics
Cryptocurrency investigation
Cybercrime
Digital forensics
Systematic literature review

ABSTRACT

Blockchain and cryptocurrency transactions leave persistent, inspectable traces that increasingly empower cybercrime investigations, yet their evidentiary value hinges on data provenance, analytical assumptions, reproducibility, and corroboration. This review deploys a structured review protocol and transparent search reporting. Primary searches in Scopus, IEEE Xplore, ScienceDirect, and Springer Nature Link were conducted between 6 and 8 August 2026 and identified 473 records. After 118 duplicates were removed, 355 records were screened; 174 reports were sought, 102 were retrieved and assessed, four full-text reports were excluded, and 98 studies were synthesized. Quality appraisal classified 83 studies as High and 15 as Moderate, with a mean score of 8.54/10. The synthesis identified transaction tracing, graph analytics, machine learning, clustering, analysis of privacy-enhancing and obfuscation mechanisms, wallet and device forensics, and cross-chain methods as complementary investigative capabilities. Sensitivity analyses restricted to High-quality studies preserved the ranking of technique, reliability, and research-gap themes. Critically, this work delivers an investigation-oriented synthesis linking evidence sources, analytical techniques, investigative functions, targets or contexts, and evidentiary outcomes. A key finding is that technical traceability alone did not establish human attribution; robust conclusions required auditable methods, explicit uncertainty, and off-chain corroboration, paving the way toward a new evidentiary standard for decentralized crime scenes.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Ruki Harwahyu,
Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia,
Depok, Indonesia.
Email: ruki.h@ui.ac.id

1. INTRODUCTION

The Blockchain and cryptocurrency ecosystems have transformed the evidentiary landscape of cybercrime. Public ledgers preserve timestamped and independently inspectable records of value transfers, token movements, smart-contract interactions, and address relationships. These traces can remain available after the originating intrusion, extortion, scam, or theft has ended, making cryptocurrency transactions valuable artifacts for post-incident reconstruction and financial investigation [1], [2].

The central forensic challenge is attribution. A blockchain address is a pseudonymous identifier rather than direct proof of a person, device, organization, or account owner. Investigated fund flows may span multiple wallets, exchanges and custodians, CoinJoin or mixer transactions,

decentralized exchanges, cross-chain bridges, and privacy-enhancing assets or mechanisms. Consequently, blockchain analysis must be interpreted as a sequence of observations and inferences whose reliability depends on assumptions, corroborating evidence, and the transparency of the analytical method.

Existing reviews provide broad taxonomies of blockchain forensics and discuss common challenges [1], [2]. However, the recent literature has rapidly expanded to encompass dynamic graph learning, explainable artificial intelligence, cross-chain prioritization, wallet and memory forensics, analysis of privacy-enhancing and obfuscation mechanisms, and evidentiary reasoning. A more focused synthesis is therefore needed to connect technical methods to cybercrime-specific investigative functions and to distinguish predictive performance from forensic strength. In this review, blockchain and cryptocurrency forensics is operationally defined as the systematic acquisition, preservation, analysis, tracing, reconstruction, attribution-support, or evidentiary assessment of blockchain-, cryptocurrency-, wallet-, smart-contract-, or associated off-chain artifacts for an investigative purpose. An analytical output is therefore considered forensic only when it contributes to an identifiable investigative function rather than predictive classification alone. Detection or triage is distinguished from forensic inference and actor attribution; actor attribution represents the strongest evidentiary step and requires corroborating evidence to support linkage to a real-world person or organization.

This review makes four contributions. First, it synthesizes 98 full-text primary studies published between 2020 and 2026, based on a literature search updated through 8 August 2026, and maps their forensic methods, evidence sources, investigative functions, targets, and evidentiary outcomes. Second, it evaluates methodological quality using an auditable ten-item quality-assessment (QA) rubric and reports study-level QA1–QA10 scores. Third, it analyzes reliability, reproducibility, attribution uncertainty, and investigative and evidentiary usefulness rather than treating predictive accuracy as forensic strength. Fourth, it proposes a hierarchical synthesis linking evidence sources, analytical techniques, investigative functions, investigative targets or contexts, and evidentiary outcomes, thereby distinguishing methods, tools, obfuscation targets, and acquisition domains that are often conflated in flat taxonomies.

Accordingly, the novelty of this review lies in the investigation-oriented synthesis and the five-level hierarchical framework derived from the 98-study corpus, rather than in proposing or combining a new machine-learning classification algorithm.

The review is guided by five research questions (RQs): RQ1 examines the techniques used to trace, classify, reconstruct, and attribute suspicious cryptocurrency activities; RQ2 examines the cybercrime categories investigated using blockchain and cryptocurrency forensic methods; RQ3 examines the underlying data, benchmark datasets, acquisition mechanisms, tools, and evidence sources; RQ4 examines the limitations affecting the reliability, reproducibility, attribution confidence, and evidentiary usefulness of blockchain forensic findings; and RQ5 identifies research gaps and future directions.

2. RESEARCH METHOD

2.1. Review Protocol and Information Sources

The review followed a PRISMA-oriented study-selection process [3], adopted systematic-review principles from evidence-based software engineering [4], and used PRISMA-S reporting items to improve the transparency of the literature search [25]. Primary database searches were conducted in Scopus, IEEE Xplore, ScienceDirect, and Springer Nature Link. The Scopus searches were conducted on 6 August 2026, while the final IEEE Xplore, ScienceDirect, and Springer Nature Link searches were completed by 8 August 2026. Accordingly, 2026 represents a partial publication year through 8 August 2026. Google Scholar was used only for supplementary citation chasing and was not included in the 473-record primary corpus. A separate Scopus prior-review query was also excluded from the primary-study pool. Complete search strings, filters, dates, and database-specific result counts are provided in Supplementary Appendix B. Table 1 summarizes the primary database yields used to construct the screening pool.

Table 1. Primary database search results

Database	Raw records	Search role
Scopus	178	Core forensics + tracing/attribution/illicit queries
IEEE Xplore	95	Core forensics + tracing/attribution/illicit queries
ScienceDirect	20	Core forensics + targeted tracing/AML and ransomware/phishing/scam queries
Springer Nature Link	180	Core forensics + tracing/attribution/illicit queries
Total	473	Primary records before deduplication

This study is a systematic literature review and does not train, tune, or evaluate a predictive model; machine-learning preprocessing procedures, train-test splits, and performance metrics discussed in later sections refer to the included primary studies rather than to an experimental pipeline conducted in this review.

2.2. Search Strategy and Eligibility

For the purposes of eligibility, blockchain and cryptocurrency forensics was operationally defined as the use of blockchain-, cryptocurrency-, wallet-, smart-contract-, or associated off-chain evidence to support at least one investigative function: evidence acquisition or preservation, transaction or fund tracing, event reconstruction, investigative detection or triage, forensic inference, attribution-support, corroboration, or evidentiary assessment and reporting. A study was considered forensically relevant only when its method or output was explicitly connected to one or more of these functions.

Detection-only, anti-money-laundering (AML) monitoring, or classification studies were eligible only when their outputs were explicitly applicable to an investigative function; generic predictive monitoring without such linkage was excluded. Detection or triage was treated as the identification and prioritization of suspicious activity, whereas forensic inference referred to reconstruction or hypothesis generation from the available evidence. Actor attribution required a further evidentiary step linking an on-chain entity or inferred cluster to a real-world person or organization through corroborating evidence. Thus, predictive classification was not treated as equivalent to forensic inference or human attribution. Cryptocurrency price prediction, blockchain-performance research, and generic blockchain-security analysis without a forensic or investigative contribution were outside the review scope. Searches were limited to English-language peer-reviewed journal articles and conference papers published from 2020 to 2026. Open-access status was not an eligibility criterion. Table 2 summarizes the operational eligibility and full-text decision rules.

Table 2. Eligibility and full-text decision rules

Stage	Include	Exclude
Title/abstract	Direct cryptocurrency/blockchain forensic or investigative relevance; method, data, tool, workflow, case analysis, or evaluation identifiable.	Generic blockchain security, non-crypto evidence-integrity use, secondary review, legal/social-only work without forensic contribution, wrong period/type/language.
Full text	Direct crypto/blockchain forensic contribution, cybercrime/investigative relevance, and extractable method/tool/dataset/workflow/evidence.	FT-E1: no direct crypto-forensic contribution; FT-E2: no investigative or evidentiary relevance; FT-E3: no identifiable method, tool, data, or workflow; FT-E4: secondary study; FT-E5: legal/social-only study; FT-E6: insufficient extractable evidence; FT-E7: duplicate companion publication; FT-E8: publication type outside the inclusion criteria.
Retrieval	Publisher, institutional access, legitimate repository, or author manuscript/preprint accepted as access routes.	Open Access (OA) status was not an eligibility criterion. Unretrieved reports were recorded as reports not retrieved rather than scientific exclusions.

2.3. Deduplication, Screening, and Full-text Retrieval

The four primary databases yielded 473 raw records. Deduplication removed 118 duplicate records, using normalized DOI as the primary matching key and normalized title and publication year when DOI information was unavailable. Records with different non-empty DOIs were not automatically merged. The resulting deduplicated pool contained 355 unique records. After title and abstract screening, 174 reports were sought for full-text retrieval. Full texts were sought through publisher websites, DOI resolver links, institutional access, and legitimate repositories. A final retrieval attempt was conducted before the review corpus was finalized; 72 reports remained unavailable, leaving 102 reports for full-text assessment, corresponding to a retrieval rate of 58.6%. Four full-text reports were excluded: two lacked a direct cryptocurrency or blockchain-forensics

contribution (FT-E1), one provided insufficient methodological detail or extractable evidence (FT-E6), and one was a companion publication without a substantial independent contribution (FT-E7). Consequently, 98 studies were included in the final synthesis. The 72 unavailable reports were classified as reports not retrieved, rather than full-text exclusions, and their titles, identifiers, sources, and retrieval attempts are documented in Supplementary Appendix C. Figure 1 summarizes the identification, screening, retrieval, eligibility assessment, and final inclusion process.

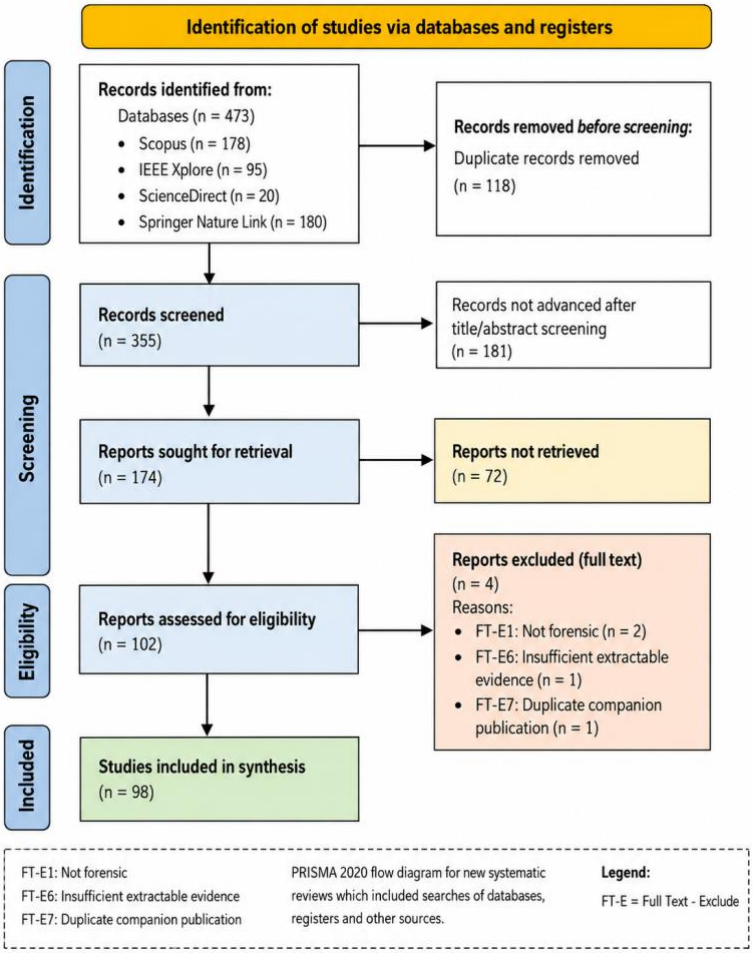


Figure 1. PRISMA flow diagram of the systematic review process

2.4. Quality Assessment and Data Extraction

Each included study was appraised using ten quality-assessment (QA) criteria scored as 0 (not met), 0.5 (partly met), or 1 (met). QA1 assessed objective clarity; QA2 assessed data, case, or dataset provenance; QA3 assessed methodological detail; QA4 assessed the appropriateness of evaluation or validation; QA5 assessed whether results were supported by reported evidence; QA6 assessed the reporting of limitations, assumptions, or uncertainty; and QA7 assessed whether conclusions were supported by the study results. QA8–QA10 were study-type specific. Machine-learning and graph-based studies were assessed for data handling, comparator and metric rigor, and generalization or reproducibility; tool and framework studies were assessed for architecture or workflow, case or dataset demonstration, and implementation reproducibility; case- and wallet-based studies were assessed for acquisition or tracing provenance, corroboration, and uncertainty; legal and evidentiary studies were assessed for explicit premises, reliability or probative reasoning, and case application; and analytical or heuristic studies were assessed for procedural specification, error or comparative validation, and reproducibility. Total scores ≥8 were classified as High, scores of 6–7.5 as Moderate, and scores <6 as Low. These cut-offs were defined as transparent descriptive bands rather than validated psychometric thresholds. QA was used to contextualize evidence strength rather

than as an additional post-hoc exclusion criterion. The complete study-level QA1–QA10 matrix for S001–S098 is provided in Supplementary Appendix D.

Data extraction was performed at the study level using structured fields for publication year, platform, study type, cybercrime context, analytical method, forensic function, tool or framework, underlying data, benchmark or labeled dataset, acquisition or API mechanism, on-chain evidence, off-chain evidence, evaluation metrics, findings, reproducibility indicators, investigative and evidentiary usefulness, limitations, future work, reliability issues, and RQ1–RQ5 mapping. Thematic codes were assigned only when an explicit statement or directly supported methodological implication matched the operational codebook; inclusion cues, exclusion cues, and supporting Study IDs are reported in Supplementary Appendix E. For the cross-study hierarchical synthesis presented in Figure 2, five synthesis dimensions were derived from the extracted fields: evidence source, analytical technique, investigative function, investigated target or context, and evidentiary outcome. Study-level terms within these dimensions were normalized into comparable coding labels, and semantically overlapping labels were consolidated into broader categories while preserving the distinction between evidence, analytical technique, investigative purpose, application context, and evidentiary interpretation. The resulting five-level framework was therefore derived from the coded extraction matrix of the 98 included studies rather than defined independently of the reviewed corpus. Quantitative synthesis used descriptive counts and percentages. To evaluate whether study quality materially affected the synthesis, thematic rankings based on all included studies were compared with rankings restricted to the 83 studies classified as High quality. Table 3 presents the quality-assessment rubric and aggregate outcomes.

Evaluation metrics were extracted only as reported characteristics of the included primary studies; they were not recomputed, pooled into a common confusion matrix, or used for model selection in this review.

Table 3. Review-level quality-assessment rubric and aggregate study-quality outcomes

QA	Operational Criterion	Scoring / result
QA1	Objective or research aim is explicit	1 explicit; 0.5 inferable; 0 absent.
QA2	Data source, case, or dataset provenance is clear	1 named + collection detail; 0.5 partial; 0 unclear.
QA3	Method/workflow is sufficiently described	1 detailed; 0.5 partial; 0 unclear.
QA4	Evaluation/validation is appropriate	1 experiment/benchmark/real case; 0.5 limited; 0 none
QA5	Results are supported by reported evidence	1 quantitative/concrete trace; 0.5 qualitative; 0 unsupported
QA6	Limitations, assumptions, or uncertainty are discussed	1 explicit; 0.5 partial; 0 absent
QA7	Conclusions are supported by study results	1 clear link; 0.5 weak; 0 unsupported
QA8	Type-specific methodological rigor	Rule varies by study type; see Supplementary Appendix D
QA9	Type-specific validation/corroboration	Rule varies by study type; see Supplementary Appendix D
QA10	Type-specific reproducibility/uncertainty	Rule varies by study type; see Supplementary Appendix D
Outcome	High >=8; Moderate 6-7.5; Low <6	83 High; 15 Moderate; 0 Low; mean 8.54/10

2.5. Reviewer Workflow and Coding Auditability

Study selection, full-text eligibility assessment, quality appraisal, and thematic coding were conducted by a single reviewer using predefined eligibility criteria, documented full-text decision rules, a structured QA rubric, and an operational thematic codebook. No independent second-reviewer screening or coding assessment was performed; therefore, inter-rater agreement statistics such as Cohen's kappa or percentage agreement were not calculated, and no claim of inter-rater reliability is made. To reduce decision opacity and facilitate independent scrutiny, the review provides the complete database search log (Supplementary Appendix B), report-retrieval records and availability-bias analysis (Supplementary Appendix C), the study-level QA1–QA10 matrix (Supplementary Appendix D), and operational thematic definitions with supporting Study IDs (Supplementary Appendix E). Study IDs S001–S098 provide consistent traceability across the included-study register, QA assessment, thematic coding, and bibliography. These measures improve transparency and auditability of the review process but do not eliminate the potential for reviewer subjectivity or substitute for independent reviewer agreement. The single-reviewer design is therefore explicitly retained as a methodological limitation.

2.6. Availability-bias and QA Sensitivity Analyses

Because 72 of 174 reports were not retrieved, potential availability bias was evaluated descriptively by publication year, database provenance, and overlapping title-keyword topic proxies.

In addition, a limited quality-based rank-stability analysis was conducted by repeating the primary-technique, RQ4 reliability-theme, and RQ5 research-gap rankings using only the 83 studies classified as High quality. Because this subset represented 84.7% of the included corpus and excluded only 15 Moderate-quality studies, the analysis was interpreted conservatively as a check of whether those studies materially altered the relative rank ordering of the synthesis categories, rather than as comprehensive evidence of robustness to study-quality effects. These analyses were not used to impute missing reports, reweight the corpus, or assess inter-rater reliability.

3. RESULTS AND DISCUSSION

3.1. Corpus Characteristics

The final corpus contained 98 studies. Each included study was assigned a fixed Study ID (S001–S098), which is used consistently throughout the Supplementary Material. Supplementary Appendix A provides the included-study register and primary synthesis coding, while Supplementary Appendix F provides the corresponding complete bibliographic entry for each Study ID. The same Study IDs are also used in the study-level QA matrix (Supplementary Appendix D) and thematic coding with supporting-study lists (Supplementary Appendix E), enabling one-to-one study-level traceability across the review outputs. These Study IDs are independent of the IEEE citation numbering used for selected representative references in the main manuscript. Publication counts were 4 in 2020, 6 in 2021, 10 in 2022, 9 in 2023, 19 in 2024, 27 in 2025, and 23 in 2026. The 2026 count is partial through the final primary search update on 8 August 2026 and should not be interpreted as a complete calendar-year count. Sixty-nine studies (70.4%) were published between 2024 and the partial 2026 period, indicating that recent research dominates the included evidence while avoiding interpretation of 2026 as a complete annual trend.

Bitcoin was the primary platform in 42 studies (42.9%), Ethereum in 25 (25.5%), general or multi-asset settings in 18 (18.4%), Bitcoin plus Ethereum in 7 (7.1%), cross-chain or multi-chain environments in 5 (5.1%), and Monero in 1 (1.0%). The corpus therefore remains Bitcoin-heavy, while Ethereum and cross-chain investigations constitute a substantial component of the included literature.

3.2. RQ1 – Forensic Techniques

For descriptive counting, each study was assigned one primary technique code: machine learning/artificial intelligence (ML/AI) detection (18; 18.4%), forensic tools/platforms (12; 12.2%), graph analytics/graph learning (10; 10.2%), transaction tracing/fund flow (9; 9.2%), CoinJoin/mixer analysis (9; 9.2%), address clustering/heuristics (8; 8.2%), attribution/deanonymization (6; 6.1%), statistical/heuristic analysis (6; 6.1%), evidentiary/forensic reasoning (5; 5.1%), wallet/device forensics (5; 5.1%), general forensic methodology (5; 5.1%), smart-contract/DeFi forensics (3; 3.1%), and cross-chain forensics (2; 2.0%). These flat labels are mutually exclusive counting codes rather than a conceptual taxonomy. To avoid conflating methods, tools, targets, and evidence domains, Figure 2 reorganizes the synthesis hierarchically across five levels: evidence source, analytical technique, investigative function, target or context, and evidentiary outcome. The hierarchy was derived from the structured study-level extraction matrix rather than defined independently of the corpus. Across S001–S098, extracted fields describing evidence sources, analytical methods, investigative functions, investigated targets or contexts, and reported investigative or evidentiary outcomes were normalized into comparable coding terms. Semantically overlapping labels were consolidated into broader categories while preserving distinctions between what evidence was examined, how it was analyzed, what investigative function the analysis served, the context in which it was applied, and the type of evidentiary conclusion supported. The resulting five-level framework therefore represents a cross-study synthesis of the coded extraction fields from the 98 included studies rather than a taxonomy defined independently of the reviewed corpus.

The operational categories represented in Figure 2 are mapped to their supporting Study IDs in Supplementary Appendix G, enabling each hierarchy level and category to be traced directly to the coded study-level records S001–S098. Because the hierarchy represents a cross-study synthesis rather than mutually exclusive primary counting codes, an individual study may support more than

one hierarchy category; Appendix G provides the corresponding operational mapping rules and Study IDs for each category.

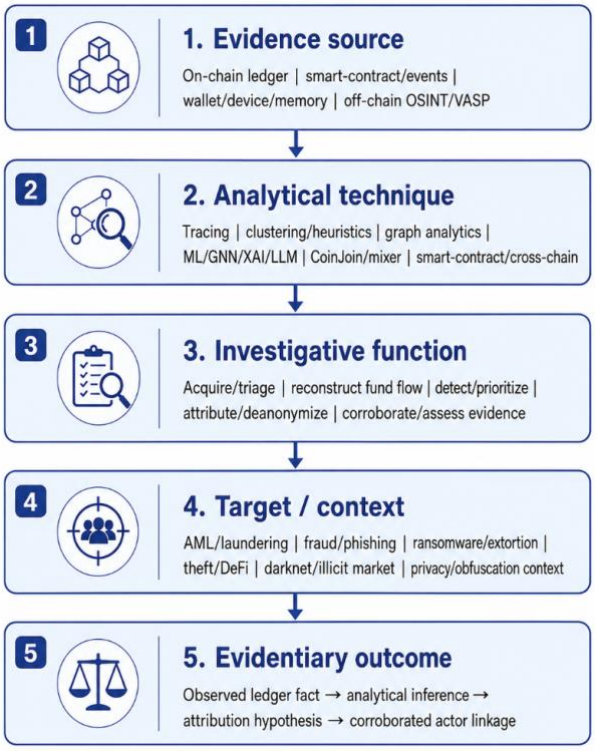


Figure 2. Hierarchical synthesis of the blockchain-forensics investigation chain.

Table 4. Synthesis of primary forensic technique categories

Primary counting code	All studies / High-quality studies	Investigative function	Main limitation
ML / AI Detection	18 / 15	Detect/prioritize suspicious entities	Ground truth, imbalance, drift, explainability
Forensic tool / platform	12 / 9	Operationalize acquisition/analysis/reporting	Coverage, validation, maintainability
Graph analytics / graph learning	10 / 9	Model relational and temporal behavior	Scale, dynamic graphs, interpretability
Transaction tracing / fund flow	9 / 8	Reconstruct movement of value	Service, bridge, and mixer ambiguity
CoinJoin / Mixer Analysis	9 / 8	Detect or reduce obfuscation	Adaptive privacy mechanisms; mapping uncertainty
Address clustering / heuristics	8 / 6	Infer common control / entity structure	Heuristic error and CoinJoin/custodial effects
Attribution / deanonymization	6 / 4	Generate identity/service hypotheses	Human identity requires corroboration
Statistical / heuristic analysis	6 / 6	Detect anomalous patterns / quantify behavior	Assumptions and transferability
Evidentiary / forensic reasoning	5 / 5	Assess reliability/probative interpretation	Jurisdiction and premise sensitivity
Wallet / device forensics	5 / 4	Acquire wallet/device/memory artifacts	Volatility, device diversity, preservation
General forensic methodology	5 / 5	Structure investigation and evidence handling	Needs validation and standardization
Smart-contract / DeFi Forensics	3 / 3	Reconstruct contract/exploit behavior	Complex calls, proxies, protocol composition
Cross-chain Forensics	2 / 1	Prioritize/trace multi-chain movement	Bridge semantics and data interoperability

Across these categories, machine-learning and graph-based methods complemented rather than replaced traditional forensic reasoning. To clarify their investigative roles, this review distinguishes three stages of investigative interpretation. Detection or triage refers to the identification and prioritization of suspicious transactions, addresses, or accounts for further examination. Forensic inference refers to the reconstruction of transaction flows, relationships, behavioral patterns, or common-control hypotheses from the available evidence. Actor attribution

represents a stronger evidentiary step in which an on-chain entity or inferred cluster is linked to a real-world person or organization and therefore requires corroborating evidence beyond predictive, clustering, or tracing outputs alone. Accordingly, predictive methods primarily support detection and triage, whereas tracing, clustering, and related analytical methods can support reconstruction and forensic inference; actor attribution requires additional corroboration. The limited High-quality-only rank-stability analysis produced a Spearman rank correlation of $\rho = 0.96$ across the 13 primary technique categories. ML/AI remained the largest category among the 83 High-quality studies ($n = 15$), while graph analytics, forensic tools/platforms, transaction tracing, and privacy-related techniques retained broadly similar relative positions. Because the High-quality subset comprised 83 of the 98 included studies (84.7%) and excluded only 15 Moderate-quality studies, this comparison should be interpreted conservatively. Within this limited sensitivity analysis, removal of the Moderate-quality studies did not materially alter the relative rank ordering of the RQ1 technique categories; it should not be interpreted as comprehensive evidence that the synthesis is insensitive to study-quality effects. Joindetect [8], for example, explicitly addresses class imbalance and concept drift in CoinJoin detection, while temporal-logic querying [5] illustrates an intrinsically interpretable alternative for investigative pattern testing. Representative studies further illustrate the methodological breadth of the corpus, including temporal-logic investigation, explainable machine learning, transaction-network analysis, CoinJoin detection, fund-flow tracing, deanonymization reasoning, wallet forensics, illicit-transaction detection, heuristic clustering, and graph-based analysis [5]-[13], [19], [21], [24].

3.3. RQ2 – Cybercrime Investigation Use Cases

For descriptive synthesis, each study was assigned one primary cybercrime category. The money laundering/anti-money laundering and counter-financing of terrorism (AML/CFT) category was the most frequent, with 20 studies (20.4%), followed by general illicit or investigative activity with 18 (18.4%), fraud/scams with 15 (15.3%), ransomware/extortion with 12 (12.2%), theft/exploit/hack with 9 (9.2%), Privacy / Identity-Related Investigative Context with 9 (9.2%), phishing with 6 (6.1%), darknet/illicit markets with 5 (5.1%), and Ponzi/investment fraud and APT/state-sponsored activity with 2 studies each (2.0%). These categories represent the primary investigative use case assigned for synthesis and do not exhaust all cybercrime contexts discussed within individual studies. Representative applications included sextortion, mixer investigations, pig-butchering complaints, ransom-payment analysis, darknet markets, smart-contract fraud, and illicit-market detection [14]-[17].

The Privacy / Identity-Related Investigative Context category denotes studies examining investigative challenges involving privacy-enhancing, identity-obscuring, deanonymization-related, or similar mechanisms. This category does not imply that privacy-enhancing technologies, CoinJoin, mixers, privacy coins, or their use are inherently illicit. Such studies were included only when these mechanisms were examined within an explicit forensic or investigative context.

3.4. RQ3 - Tools, Datasets, and Evidence Sources

The evidence base combined blockchain-native artifacts with external investigative context. On-chain evidence included transaction hashes, addresses, unspent transaction output (UTXO) relationships, token transfers, timestamps, contract calls and events, gas values, and graph topology. Off-chain evidence included exchange or virtual asset service provider (VASP) records, open-source intelligence (OSINT), scam websites and social-media content, complaints, seized devices, mobile-wallet artifacts, memory images, communications, and other lawfully obtained metadata. The strongest attribution-oriented studies combined these evidence layers rather than inferring an individual solely from an address cluster.

For RQ3, the extracted information was organized into three conceptually distinct dimensions. Underlying data described the origin of the evidence, such as the Bitcoin ledger, Ethereum account or contract data, darknet or market data, social-media or OSINT data, and device or memory artifacts. Benchmark or labeled datasets represented packaged experimental corpora, including the Elliptic and Elliptic++ datasets, Kaggle-derived datasets, synthetic data, and proprietary labeled analytics. Acquisition mechanisms described how the data were accessed,

including blockchain APIs, Etherscan or explorer interfaces, archive nodes, and custom parsers. These dimensions were intentionally non-mutually exclusive because a single study could use multiple evidence sources and acquisition mechanisms. Studies that described their data only narratively without identifying a named dataset or acquisition mechanism were treated as having a data-provenance reporting limitation rather than as a distinct data-source category. Twenty studies exhibited this limitation, indicating insufficient specificity for direct replication.

Most included studies used custom research implementations rather than standardized forensic products. Named platforms and tools included blockchain explorers and APIs, BlockSci, GraphSense, purpose-built investigation tools, SQL- or graph-based analysis systems, wallet or device forensic tools, and recent AI-assisted frameworks. Commercial intelligence platforms can improve service tagging and operational coverage, but proprietary labels and clustering logic can limit scientific auditability. Conversely, academic implementations are often more transparent but may lack continuously updated intelligence and production-scale validation. Representative studies of acquisition, preservation, and supporting forensic workflows include [11], [18], [20], [21], and [23].

3.5. RQ4 - Reliability, Reproducibility, and Investigative/Evidentiary Usefulness

The main RQ4 finding is that predictive performance is not equivalent to forensic reliability. Consistent with the distinction introduced in Section 3.2, detection or triage identifies and prioritizes potentially suspicious transactions, addresses, or accounts, whereas forensic inference seeks to reconstruct relationships, transaction flows, behavioral patterns, or common-control hypotheses from the available evidence. Actor attribution requires a further evidentiary step in which an on-chain entity or inferred cluster is linked to a real-world person or organization through sufficiently reliable corroborating evidence. Consequently, a high-performing classifier may provide a useful investigative lead without independently establishing forensic inference or human attribution. High benchmark accuracy can coexist with weak evidentiary support when labels are uncertain, temporal leakage is possible, methods are not reproducible, or attribution depends on undocumented heuristics. Reliability was therefore interpreted across data provenance, analytical assumptions, error behavior, reproducibility, explainability, corroboration, and auditability. In this review, investigative and evidentiary usefulness refers to the extent to which an analytical output can support case prioritization, reconstruction, attribution hypotheses, evidence interpretation, auditability, or probative reasoning. It does not mean that this review determines legal admissibility, which is jurisdiction-specific.

Using the operational definitions provided in Supplementary Appendix E, cross-chain or service-boundary ambiguity was identified in 64 studies, attribution uncertainty in 45, reproducibility or code/data availability concerns in 39, explainability or black-box risk in 29, legal or evidentiary reliability concerns in 23, class imbalance in 19, clustering or heuristic error in 17, ground-truth scarcity or uncertain labels in 12, concept drift in 6, and dataset opacity or provenance limitations in 3. These themes were non-mutually exclusive. Attribution uncertainty primarily reflected three recurring problems represented in the operational coding: uncertainty over who controls an address or cluster, ambiguity between service-level and person-level attribution, and the need for corroborating off-chain evidence such as VASP/KYC records, device evidence, or OSINT before linking an on-chain entity to a human actor.

Within the limited High-quality-only rank-stability analysis, the corresponding RQ4 counts were 54, 40, 36, 24, 21, 18, 17, 11, 6, and 2, respectively, and their relative rank ordering remained unchanged (Spearman $\rho = 1.00$). Because the High-quality subset comprised 83 of the 98 included studies (84.7%) and excluded only 15 Moderate-quality studies, this result should be interpreted conservatively. It indicates that removal of the Moderate-quality studies did not alter the relative ordering of the RQ4 reliability themes in this comparison; it should not be interpreted as comprehensive evidence that the findings are insensitive to study-quality effects. This analysis also does not address reviewer-dependent coding subjectivity or establish inter-rater reliability. Table 5 summarizes the coded reliability challenges and the corresponding High-quality comparison counts.

Table 5. Reliability challenges identified across the 98 studies

Reliability theme	All studies / High-quality studies	Operational meaning
Cross-chain / service boundary ambiguity	64 / 54	Uncertainty after exchange, mixer, bridge, DEX, custodian, or chain boundary
Attribution uncertainty	45 / 40	Address ownership uncertainty, service-versus-person ambiguity, and insufficient off-chain corroboration can prevent reliable human attribution
Reproducibility / code-data availability	39 / 36	Code, data, parameters, or proprietary labels limit independent verification
Explainability / black-box risk	29 / 24	Technical model rationale and/or the forensic reasoning chain is difficult to interpret, explain, or audit
Legal/evidentiary reliability	23 / 21	Probative reliability, auditability, expert interpretation, or evidence-handling concern
Class imbalance	19 / 18	Illicit minority class can distort accuracy and error rates
Clustering/heuristic error	17 / 17	Assumption-sensitive false links or missed links in clustering/deanonymization
Ground-truth scarcity / uncertain labels	12 / 11	Incomplete or uncertain malicious/entity labels
Concept drift / temporal change	6 / 6	Behavior changes over time; temporal validation/retraining is needed
Dataset opacity / provenance	3 / 2	Dataset/label source is insufficiently transparent

Investigative and evidentiary usefulness is particularly sensitive to the distinction between directly verifiable facts and analytical inference. Transaction hashes, amounts, timestamps, and contract events can be independently verified against a ledger, whereas common control, service attribution, laundering intent, and human identity remain inferential. This review therefore distinguishes technical explainability from forensic explainability. Technical explainability concerns whether the operation or output of an analytical method can be interpreted, for example through understandable features, model rationale, graph relationships, or decision logic. Forensic explainability concerns whether an investigator can reconstruct and communicate an auditable chain from source evidence, through analytical transformations and assumptions, to the resulting investigative inference, including uncertainty and corroboration. The existing explainability/black-box theme was coded as an aggregate reliability theme; these two dimensions are distinguished conceptually in the revised interpretation rather than retrospectively assigned separate frequency counts. The deanonymization literature represented in the corpus [10], [22] emphasizes that probative value depends on making analytical assumptions explicit and assessing their reliability. Accordingly, this review discusses legal relevance only in terms of evidentiary reliability, explainability, auditability, chain-of-custody and reporting support, and probative reasoning; it does not make cross-jurisdictional admissibility determinations.

3.6. RQ5 - Research Gaps and Future Directions

The research-gap synthesis shows that the field is moving from single-chain analytics toward multi-source, multi-protocol investigations, while validation and operational standardization remain uneven. Explainability and evidentiary value constituted the most frequently coded research-gap theme, appearing in 67 studies, followed by standardization and operational integration in 56, scalability and real-time analysis in 47, privacy-enhancing and obfuscation technologies in 43, temporal validation and concept drift in 42, on-chain and off-chain attribution in 40, reproducible benchmarks and open data in 23, cross-chain tracing in 22, and standardized ground truth in 18. Within the aggregate explainability/evidentiary-value theme, two conceptually distinct needs are relevant. Technical explainability concerns making model behavior, features, graph relations, or analytical outputs interpretable, whereas forensic explainability concerns producing an auditable reasoning chain from source evidence through analytical assumptions and transformations to an investigative inference, including uncertainty and corroboration. Because the original thematic codebook treated explainability/evidentiary value as a single aggregate theme, these dimensions are distinguished conceptually in the revised synthesis rather than assigned retrospective separate frequency counts. The codebook in Supplementary Appendix E defines the operational criteria for each theme and lists the supporting Study IDs. These directions are consistent with recent work on explainable models, drift-aware privacy analysis, privacy-preserving wallet forensics,

benchmarking, lawful metadata correlation, and graph learning [6], [8], [20], [23], [24]. Table 6 summarizes the coded research-gap themes and their corresponding future directions.

Table 6. Research gaps and future directions

Research gap	All studies / High-quality studies	Future direction
Explainability / evidentiary value	67 / 58	Improve technical model interpretability and forensic explainability through auditable reasoning, explicit uncertainty/error reporting, and human-readable evidence-to-inference links.
Standardization / operational integration	56 / 47	Standardize workflows, evidence handling, interfaces, and organizational integration
Scalability / real-time analysis	47 / 43	Evaluate large/dynamic graphs and operational latency/cost
Privacy-enhancing / obfuscation technologies	43 / 37	Develop context-aware, assumption-aware forensic methods for investigations involving mixers, CoinJoin, privacy coins, and other privacy-enhancing or obfuscation mechanisms.
Temporal validation / concept drift	42 / 36	Use temporal holdouts, drift monitoring, and retraining/evolution analysis
On-chain + off-chain attribution	40 / 35	Combine ledger evidence with VASP/KYC, OSINT, devices, and communications
Reproducible benchmarks / open data	23 / 21	Publish versioned datasets, label provenance, code, and parameters
Cross-chain tracing	22 / 19	Model bridges, assets, timestamps, and chain-independent graph relations
Standardized ground truth	18 / 17	Create transparent, uncertainty-aware labeling and corroborated identity/service ground truth

Within the limited High-quality-only rank-stability analysis, the corresponding RQ5 counts were 58, 47, 43, 37, 36, 35, 21, 19, and 17, respectively, and the relative rank ordering remained unchanged (Spearman $\rho = 1.00$). Because the High-quality subset comprised 83 of the 98 included studies (84.7%) and excluded only 15 Moderate-quality studies, this comparison should be interpreted conservatively. It indicates that removal of the Moderate-quality studies did not alter the relative ordering of the RQ5 research-gap themes in this comparison; it should not be interpreted as comprehensive evidence that the gap synthesis is insensitive to study-quality effects. Importantly, thematic frequency is used here only as a descriptive measure of recurrence across the included literature and does not establish research priority, urgency, or scientific importance. A less frequently coded issue may still be consequential for forensic practice or future research. Accordingly, the future directions in Table 6 are presented as evidence-informed areas for further investigation rather than as a priority ranking. Future work should strengthen label provenance and uncertainty reporting, temporal and cross-chain validation, reproducible benchmarks, and both technical and forensic explainability so that investigators can distinguish directly observed ledger facts from analytical inference.

3.7. Cross-RQ Synthesis and Practical Investigative Framework

The synthesis supports a five-link operational model connecting analytical techniques, investigative functions, evidence, reliability considerations, and investigative and evidentiary usefulness. The hierarchical synthesis in Figure 2 adds an upstream evidence-source layer and distinguishes investigative targets or contexts from analytical techniques. Transaction tracing supports timeline and fund-flow reconstruction but becomes ambiguous at service boundaries. Address clustering and deanonymization support hypothesis generation but remain probabilistic and assumption-sensitive. Machine-learning and graph-based methods support large-scale detection, triage, and prioritization but depend on ground-truth quality, temporal validity, and explainability. Wallet and device forensics provide control-related artifacts, while off-chain correlation can strengthen attribution. The strongest investigative outcome is therefore not a score or visual graph alone, but a transparent evidentiary chain documenting which facts were directly observed, which inferences were made, and what corroborating evidence supports actor linkage. Figure 3 translates this cross-RQ synthesis into an operational evidence-handling and investigative workflow.

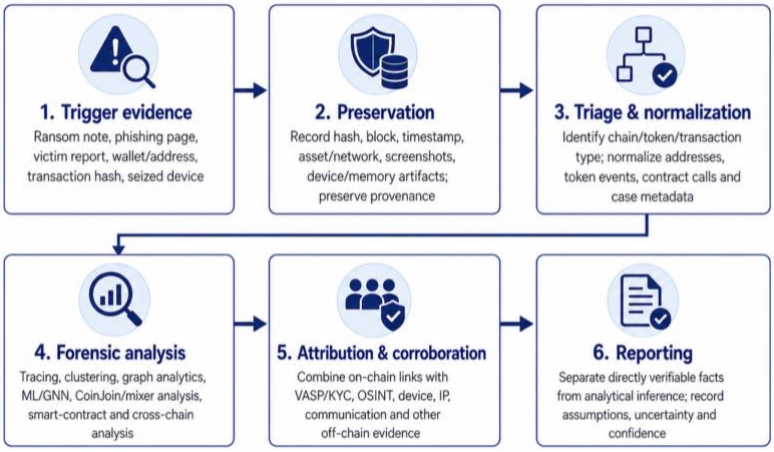


Figure 3. Cryptocurrency evidence handling and investigative workflow derived from the synthesis.

For incident-response and ICT engineering management, the workflow begins with preservation of trigger evidence, including ransom communications, phishing pages, victim reports, transaction hashes, addresses, and relevant device artifacts. Triage then identifies the network, asset, transaction type, smart-contract or bridge involvement, and preservation requirements. Analytical methods should be selected according to the investigative question, followed by attribution and corroboration using VASP/KYC information, OSINT, device evidence, network evidence, or communication records where lawfully available. The final report should retain an audit trail of tools, parameters, hashes, data sources, analytical assumptions, uncertainty, and the evidentiary basis for each conclusion.

3.8. Comparison with Prior Reviews and Novelty

Prior systematic reviews established blockchain forensics as a multidisciplinary field and catalogued its techniques, applications, challenges, and future directions [1], [2]. The present review extends this perspective in three ways. First, it covers studies published from 2020 to 2026, with the literature search updated through 8 August 2026, and explicitly maps methods, evidence sources, crime contexts, quality-assessment results, reliability issues, and research gaps across 98 full-text studies. Second, it distinguishes investigative and evidentiary usefulness from predictive performance and evaluates the stability of major findings through sensitivity analyses restricted to High-quality studies. Third, it organizes the field hierarchically from evidence sources through analytical techniques and investigative functions to investigative targets or contexts and evidentiary outcomes. This framing links technical blockchain analytics to evidence acquisition, reconstruction, attribution confidence, corroboration, reporting, and ICT engineering management without claiming jurisdiction-independent legal admissibility.

3.9. Threats to Validity

This review has four principal limitations. First, 72 of 174 reports sought for retrieval remained unavailable after the final documented retrieval pass (58.6% retrieval), creating a risk of availability bias, particularly for records originating from Springer Nature Link. Supplementary Appendix C provides source-, year-, and topic-level diagnostics and the complete list of reports not retrieved. Second, study screening, full-text eligibility assessment, QA scoring, and thematic coding were performed by a single reviewer. This design creates potential for reviewer-dependent selection, appraisal, and coding judgments, particularly where eligibility boundaries or thematic interpretations require qualitative assessment. Predefined eligibility criteria, structured decision rules, the QA rubric, operational coding definitions, and retained study-level records were used to reduce decision opacity and improve auditability. However, these safeguards do not establish inter-rater reliability or eliminate reviewer subjectivity; consequently, the absence of an independent second-reviewer assessment remains an important methodological limitation. Third, assigning one primary technique code per study simplifies multi-method papers, while some thematic counts necessarily depend on structured qualitative interpretation. Finally, coverage for 2026 is partial through 8 August 2026, and

exact prevalence estimates should therefore be interpreted cautiously. The High-quality-only sensitivity analyses indicated that the relative rank ordering of the principal technique, reliability, and research-gap categories remained broadly stable. However, because the High-quality subset comprised 83 of the 98 included studies (84.7%) and excluded only 15 Moderate-quality studies, these analyses should be interpreted as limited rank-stability checks rather than comprehensive evidence that the synthesis is insensitive to study-quality effects.

4. CONCLUSION

This systematic literature review synthesized 98 primary studies on blockchain and cryptocurrency forensics for cybercrime investigation. The evidence shows that ML/AI detection, forensic tools, graph analytics, transaction tracing, CoinJoin and mixer analysis, clustering, wallet and device forensics, and cross-chain methods support distinct but complementary investigative functions. Money laundering/AML-CFT, fraud, ransomware/extortion, theft and exploitation, phishing, darknet activity, and privacy- or identity-related investigative contexts were among the investigated contexts.

The central conclusion is that blockchain transparency provides technical traceability but does not independently establish human attribution or legal admissibility. Evidentiary usefulness depends on ground-truth quality, label provenance, heuristic error, temporal validity, reproducibility, explainability, and corroborating off-chain evidence. Sensitivity analyses restricted to High-quality studies preserved the principal rankings, but the 41.4% report non-retrieval rate and single-reviewer workflow remain important limitations. Future research should emphasize transparent ground truth, versioned benchmarks, temporal and cross-chain validation, privacy-aware analysis, auditable explanations, and stronger integration with conventional digital forensics.

SUPPLEMENTARY MATERIAL

A separate Supplementary Material file provides Appendices A-G, including the included-study register and QA summary, reproducible search log, reports not retrieved and availability-bias analysis, complete QA1-QA10 matrix with sensitivity analyses, thematic coding codebook with supporting Study IDs, the complete included-study bibliography, and the Figure 2 hierarchy-to-study traceability mapping.

ACKNOWLEDGEMENTS

The authors have no acknowledgements to declare.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY

The study-level register, database search log, report-retrieval record, quality-assessment matrix, sensitivity analysis, thematic coding codebook, complete included-study bibliography, and Figure 2 hierarchy-to-study traceability mapping are provided in the Supplementary Material accompanying this article.

AUTHOR CONTRIBUTIONS

Rahman Noviyanto: Conceptualization; Methodology; Investigation; Data Curation; Formal Analysis; Visualization; Writing - Original Draft. Ruki Harwahyu: Conceptualization; Methodology; Supervision; Validation; Writing - Review & Editing.

REFERENCES

[1] H. F. Atlam, N. Ekuri, M. A. Azad, and H. S. Lallie, "Blockchain Forensics: A Systematic Literature Review of Techniques, Applications, Challenges, and Future Directions," *Electronics*, vol. 13, no. 17, p. 3568, Sep. 2024, doi: 10.3390/electronics13173568.

[2] S. Dudani, I. Baggili, D. Raymond, and R. Marchany, "The current state of cryptocurrency forensics," *Forensic Sci. Int. Digit. Investig.*, vol. 46, p. 301576, Sep. 2023, doi: 10.1016/j.fsidi.2023.301576.

[3] M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.

[4] Barbara Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele University and Durham University, EBSE Technical Report, 2007.
[5] M. Blanchini, M. Cerreta, D. Di Monda, M. Fabbri, M. Raciti, H. S. Ahmad, and G. Costa, "Supporting Criminal Investigations on the Blockchain: A Temporal Logic-based Approach," in Proceedings of the 8th Italian Conference on Cyber Security (ITASEC 2024), CEUR Workshop Proceedings, vol. 3731, paper 33, Salerno, Italy, Apr. 8–12, 2024. [Online]. Available: https://ceur-ws.org/Vol-3731/paper33.pdf.
[6] S. A. Bokade, V. K. Sharma, and B. Manjre, "Design of an Integrated Modern Explainable Machine Learning Framework for Blockchain Forensic Analysis," *Sci. Technol. Asia*, vol. 30, no. 4, pp. 270–289, 2025.
[7] Y. Wu *et al.*, "A Bitcoin Transaction Network Analytic Method for Future Blockchain Forensic Investigation," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 2, pp. 1230–1241, Apr. 2021, doi: 10.1109/TNSE.2020.2970113.
[8] J. W. O'Meara, M. Taneja, J. Nicholls, N. Kothale, S. Flinter, and A. D. Jurcut, "JoinDetect: A Data-Led Machine Learning Approach to Detection of Coinjoin Transactions," in *Proceedings of the 2024 International Conference on Information Technology for Social Good*, Bremen Germany: ACM, Sep. 2024, pp. 5–13. doi: 10.1145/3677525.3678633.
[9] G. Cola, M. Mazza, and M. Tesconi, "From Tweet to Theft: Tracing the Flow of Stolen Cryptocurrency," in Proceedings of the Italian Conference on Cyber Security (ITASEC 2023), CEUR Workshop Proceedings, vol. 3488, paper 11, Bari, Italy, May 2–5, 2023. [Online]. Available: https://ceur-ws.org/Vol-3488/paper11.pdf.
[10] D. Deuber, J. Gruber, M. Humml, V. Ronge, and N. Scheler, "Argumentation Schemes for Blockchain Deanonymisation," *FinTech*, vol. 3, no. 2, pp. 236–248, Mar. 2024, doi: 10.3390/fintech3020014.
[11] A. Bhattarai *et al.*, "Cryptocurrency forensics automation: a deep learning and NLP-based approach for mobile platforms," *Discov. Comput.*, vol. 28, no. 1, p. 123, Jun. 2025, doi: 10.1007/s10791-025-09595-1.
[12] F. M. Ayoob, M. Jose, and S. U. Kumar, "Forensic Analysis and Detection of Illicit Transactions in Bitcoin Network," in *2024 5th International Conference on Smart Electronics and Communication (ICOSEC)*, Trichy, India: IEEE, Sep. 2024, pp. 673–679. doi: 10.1109/ICOSEC61587.2024.10722590.
[13] Y. Gong, K. P. Chow, and S. M. Yiu, "Improved Bitcoin simulation model and address heuristic method," *Forensic Sci. Int. Digit. Investig.*, vol. 53, p. 301935, Jul. 2025, doi: 10.1016/j.fsidi.2025.301935.
[14] K.-S. Choi, M. Chawki, and S. Basu, "Uncovering Cryptocurrency-Enabled Sextortion: A Blockchain Forensic Analysis of Transactions and Offender Laundering Tactics," *Information*, vol. 17, no. 3, p. 236, Mar. 2026, doi: 10.3390/info17030236.
[15] P. Tippe and C. Deckers, "Unmixing the mix: Patterns and challenges in Bitcoin mixer investigations," *Forensic Sci. Int. Digit. Investig.*, vol. 52, p. 301876, Mar. 2025, doi: 10.1016/j.fsidi.2025.301876.
[16] S. Ko, J. Kim, J. Y. Jang, and G. Kim, "Human-supervised LLM triage of pig butchering complaints: A validation study of multi-model identifier extraction," J. Forensic Sci., 2026, doi: 10.1111/1556-4029.70411.
[17] C. Patsakis, E. Politou, E. Alepis, and J. Hernandez-Castro, "Cashing out crypto: state of practice in ransom payments," *Int. J. Inf. Secur.*, vol. 23, no. 2, pp. 699–712, Apr. 2024, doi: 10.1007/s10207-023-00766-z.
[18] S. J. Aquilina, F. Casino, M. Vella, J. Ellul, and C. Patsakis, "EtherClue: Digital investigation of attacks on Ethereum smart contracts," *Blockchain Res. Appl.*, vol. 2, no. 4, p. 100028, Dec. 2021, doi: 10.1016/j.bcra.2021.100028.
[19] M. Dragošević, D. Topolčić, K. Hausknecht, and D. Delija, "Enhancing Digital Forensics Through Machine Learning Algorithms for Detecting Illicit Bitcoin Transactions in Blockchain Analysis," in *2025 MIPRO 48th ICT and Electronics Convention*, Opatija, Croatia: IEEE, Jun. 2025, pp. 831–836. doi: 10.1109/MIPRO65660.2025.11131873.
[20] H.-H. Kao, J.-M. Feng, C.-Y. Li, and C.-H. Lin, "A Local Hierarchical LLM Framework for Privacy-Preserving Memory Forensics of Cryptocurrency Wallets," *IEEE Access*, pp. 1–1, 2026, doi: 10.1109/ACCESS.2026.3682641.
[21] S. Taylor, S. H. Kim, K. A. Zainol Ariffin, and S. N. H. Sheikh Abdullah, "A comprehensive forensic preservation methodology for crypto wallets," *Forensic Sci. Int. Digit. Investig.*, vol. 42–43, p. 301477, Oct. 2022, doi: 10.1016/j.fsidi.2022.301477.
[22] M. Fröwis, T. Gottschalk, B. Haslhofer, C. Rückert, and P. Pesch, "Safeguarding the evidential value of forensic cryptocurrency investigations," *Forensic Sci. Int. Digit. Investig.*, vol. 33, p. 200902, Jun. 2020, doi: 10.1016/j.fsidi.2019.200902.
[23] W.-H. Lin and C.-Y. Wen, "A Lawful Metadata-Driven Framework for Linking Encrypted Communication Behavior and Cryptocurrency Wallet Activity in Digital Investigations," *Appl. Syst. Innov.*, vol. 9, no. 4, p. 73, Mar. 2026, doi: 10.3390/asi9040073.
[24] N. Nimmagadda and M. M, "Graph Neural Networks for Illicit Transaction Detection in the Elliptic Bitcoin Dataset," in *2025 5th International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN)*, Coimbatore, India: IEEE, Nov. 2025, pp. 1121–1127. doi: 10.1109/ICECMSN68058.2025.11382645.
[25] M. L. Rethlefsen *et al.*, "PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews," *Syst. Rev.*, vol. 10, no. 1, p. 39, Jan. 2021, doi: 10.1186/s13643-020-01542-z.
[26] M. Šafář, J. Pluskal, V. Veselý, and O. Ryšavý, "The enemy of reproducibility is opacity: What's inside the Elliptic bitcoin dataset (and why it is wrong)," *Forensic Sci. Int. Digit. Investig.*, vol. 57, p. 302124, Jun. 2026, doi: 10.1016/j.fsidi.2026.302124.
[27] Z. Wu, J. Liu, J. Wu, Z. Zheng, and T. Chen, "TRacer: Scalable Graph-Based Transaction Tracing for Account-Based Blockchain Trading Systems," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 2609–2621, 2023, doi: 10.1109/TIFS.2023.3266162.
[28] T. Leng *et al.*, "A Conflict Resolution Method for Multi-source Labels of Bitcoin Addresses," *Blockchain Res. Appl.*, p. 100486, Apr. 2026, doi: 10.1016/j.bcra.2026.100486.
[29] Z. Che *et al.*, "Across-Platform Detection of Malicious Cryptocurrency Accounts via Interaction Feature Learning," *IEEE Trans. Inf. Forensics Secur.*, vol. 20, pp. 4783–4798, 2025, doi: 10.1109/TIFS.2025.3565130.

[30] A. Wahrstätter, A. Taudes, and D. Svetinovic, "Reducing Privacy of CoinJoin Transactions: Quantitative Bitcoin Network Analysis," *IEEE Trans. Dependable Secure Comput.*, vol. 21, no. 5, pp. 4543–4558, Sep. 2024, doi: 10.1109/TDSC.2024.3353803.
[31] F. Liu *et al.*, "Bitcoin Address Clustering Based on Change Address Improvement," *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 6, pp. 8094–8105, Dec. 2024, doi: 10.1109/TCSS.2023.3239031.
[32] N. Pocher, M. Zichichi, F. Merizzi, M. Z. Shafiq, and S. Ferretti, "Detecting anomalous cryptocurrency transactions: An AML/CFT application of machine learning-based forensics," *Electron. Mark.*, vol. 33, no. 1, p. 37, Dec. 2023, doi: 10.1007/s12525-023-00654-3.
[33] H. Du *et al.*, "Fine-Grained and Class-Incremental Malicious Account Detection in Ethereum via Dynamic Graph Learning," *IEEE Trans. Inf. Forensics Secur.*, vol. 20, pp. 10130–10145, 2025, doi: 10.1109/TIFS.2025.3612194.
[34] A. Brechlin, J. Schäfer, and F. Armknecht, "Buy Crypto, Sell Privacy: An Extended Investigation of the Cryptocurrency Exchange Evonax," *Int. J. Netw. Manag.*, vol. 35, no. 1, p. e2325, Jan. 2025, doi: 10.1002/nem.2325.
[35] X. Liu, Z. Tang, P. Li, S. Guo, X. Fan, and J. Zhang, "A Graph Learning Based Approach for Identity Inference in DApp Platform Blockchain," *IEEE Trans. Emerg. Top. Comput.*, vol. 10, no. 1, pp. 438–449, Jan. 2022, doi: 10.1109/TETC.2020.3027309.
[36] T. Thomas, T. Edwards, and I. Baggili, "BlockQuery: Toward forensically sound cryptocurrency investigation," *Forensic Sci. Int. Digit. Investig.*, vol. 40, p. 301340, Apr. 2022, doi: 10.1016/j.fsidi.2022.301340.
[37] D. Lin, J. Wu, Q. Fu, Z. Zheng, and T. Chen, "RiskProp: Account Risk Rating on Ethereum via De-anonymous Score and Network Propagation," *IEEE Trans. Dependable Secure Comput.*, vol. 22, no. 3, pp. 2132–2145, May 2025, doi: 10.1109/TDSC.2024.3475825.
[38] Y. Wu *et al.*, "Unveiling Ethereum Mixing Services Using Enhanced Graph Structure Learning," *IEEE Trans. Dependable Secure Comput.*, vol. 23, no. 2, pp. 3845–3863, Mar. 2026, doi: 10.1109/TDSC.2025.3640941.
[39] Y. Wu *et al.*, "Fine-Grained Detection for Bitcoin Mixing Services Using Heterogeneous Representation Learning," *IEEE Trans. Netw. Sci. Eng.*, vol. 13, pp. 7069–7083, 2026, doi: 10.1109/TNSE.2026.3664284.
[40] J. Lu, B. Liao, Y. Liu, and K. E. Yao, "ETHIAD: A novel explainable model for detecting illicit accounts on Ethereum," *PLOS One*, vol. 20, no. 12, p. e0338366, Dec. 2025, doi: 10.1371/journal.pone.0338366.
[41] R. Shevchuk, B. Adamyk, and V. Benson, "Threshold-Based Private Set Intersection Protocol for Secure Deconfliction in Multi-Jurisdictional Blockchain Investigations," *Electronics*, vol. 15, no. 12, p. 2709, Jun. 2026, doi: 10.3390/electronics15122709.
[42] M. Kim, J. Lee, H. Kwon, and J. Hur, "Get off of Chain: Unveiling Dark Web Using Multilayer Bitcoin Address Clustering," *IEEE Access*, vol. 10, pp. 70078–70091, 2022, doi: 10.1109/ACCESS.2022.3187210.
[43] H.-H. Kao, "Accelerating Multilingual Cryptocurrency Forensics: An NLP-Driven Approach for Efficient Mnemonic Identification," *IEEE Access*, vol. 13, pp. 10513–10526, 2025, doi: 10.1109/ACCESS.2025.3528829.
[44] S. Chen, Y. Liu, Q. Zhang, Z. Shao, and Z. Wang, "Multi-Distance Spatial-Temporal Graph Neural Network for Anomaly Detection in Blockchain Transactions," *Adv. Intell. Syst.*, vol. 7, no. 8, p. 2400898, Aug. 2025, doi: 10.1002/aisy.202400898.
[45] L. M. Shekhtman, A. Sela, and S. Havlin, "Percolation framework reveals limits of privacy in conspiracy, dark web, and blockchain networks," *EPJ Data Sci.*, vol. 12, no. 1, p. 16, May 2023, doi: 10.1140/epjds/s13688-023-00392-8.
[46] E. Pettersson Ruiz and J. Angelis, "Combating money laundering with machine learning – applicability of supervised-learning algorithms at cryptocurrency exchanges," *J. Money Laund. Control*, vol. 25, no. 4, pp. 766–778, Oct. 2022, doi: 10.1108/JMLC-09-2021-0106.
[47] F. Ertam, "Near Real-Time Ethereum Fraud Detection Using Explainable AI in Blockchain Networks," *Appl. Sci.*, vol. 15, no. 19, p. 10841, Oct. 2025, doi: 10.3390/app151910841.
[48] G. Min Baek, Y. Ah Shin, G. Hyun Jang, and I. R. Jeong, "Block Number-Based Address Clustering for Bitcoin Taproot Upgrade," *IEEE Access*, vol. 13, pp. 94523–94537, 2025, doi: 10.1109/ACCESS.2025.3570459.